Mathematical Statistics
Stockholm University

Improving the Calculation of Statistical Significance in Genome-Wide Scans

Lars Ängquist and Ola Hössjer

Research Report 2003:3

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 **2003:3**,
http://www.math.su.se/matstat

Improving the Calculation of Statistical Significance in Genome-Wide Scans

Lars Ängquist*and Ola Hössjer†

May 2003

Abstract

This article deals with some topics regarding linkage analysis and significance. Imagine that one has found a maximum NPL-score in a (complete/partial) genome scan, then the next step is to calculate the significance (p-value) of the result in a satisfactory way- simple and reliable. This calculation may be perfomed by simulation or by theoretical approximation, with or without the assumption of perfect marker information. Here we will concentrate on the context of theoretical approximation with the further assumption of fully informative data (perfect marker information). Our starting point is the asymptotic approximation formula presented by Lander and Kruglyak (1995) which is based on extreme value theory for Gaussian processes (cf. e.g. Lander and Botstein 1989). The major focus and possible importance of this article will then be the suggestions of two distinct improvements to this formula.

Firstly, we present a formula for calculating the crossover rate ρ for a pedigree of a general family structure. These values may then be weighted into an overall crossover rate which finally may be used in the significance calculations using the original approximation formula.

*Department of Mathematical Statistics, Lund University, Lund and Wallenberg Laboratory, Department of Endocrinology, Malmö University Hospital, Lund University, Malmö.

†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.

Secondly, the existing p-value formulas are based on the assumption of a normally distributed NPL score and the implication (conservative/anticonservative p-values) of this proposition is depending on the pedigree structure. Here we are using the following approach to adjust for non-normality. The first step is to calculate the marginal distribution of the NPL score under the null hypothesis of no linkage with an arbitrarily small error. Then the NPL score is transformed to have a marginal standard normal distribution. The transformed maximal NPL score may, together with a slightly corrected value of the overall crossover rate, be inserted into the Lander and Kruglyak formula when performing p-value calculations.

We have used pedigrees of seven different structures to compare the performance of the adjusted approximation formula and the traditional approximation formula with respect to results found by simulation. We have also performed the same comparisons applied to two real data sets- e.g. the BOTNIA study data set (cf. Parker et al. 2001; Lindgren et al. 2002). The result is that our suggested improvements, in general, seem to strongly improve the correctness of the p-value calculations, especially for pedigree sets which correspond to distributions of obvious nonnormality.

Contents

1 Introduction

The techniques of *linkage analysis* are usually used to try to identify possible chromosomal regions that are linked to an interesting phenotype (disease). This article will deal with nonparametric linkage analysis (cf. e.g. Kruglyak et al. 1996) and the issue of calculating the statistical significance of the maximal NPL-score when performing a genome-wide scan. For a brief overview of different approaches to gene mapping cf. e.g. Ott (2000) and different p-value methods based on permutation tests is described for quantitative traits and animal breeding models in e.g. Doerge and Churchill (1996), Abney et al. (2002) and Lystig (2002).

When calculating the statistical significance (p-values) it is possible to use two distinctly different approaches - simulation or theoretical approximation. In this work we use the latter approach and we are primarily interested in trying to improve the performance of the asymptotic approximation formula given in Lander and Kruglyak (1995), which is based on extreme value theory and normal approximations of distributions and processes (cf. e.g. Lander and Botstein 1989; Feingold et al. 1993). Here two different improvements are suggested. Firstly, we give a formula for calculating the crossover rate ρ for a general collection of pedigrees (pedigree set) based on the theory presented in Hössjer (2002). Secondly, we approximate the distribution function for the NPL score of a pedigree set, F_ϵ, with arbitrary good accuracy and then we transform F_ϵ to the cumulative distribution function, Φ, for the standard normal distribution to improve the performance of the p-value approximation in Lander and Kruglyak (1995).

2 Basic Linkage Analysis

2.1 Definitions

In this section we give some basic definitions. More details can be found in Kruglyak et al. (1996).

A *pedigree* is a set of relatives. The members of a pedigree may be divided into two subgroups: *founders* and *nonfounders*. The distinction between these groups is based on the fact that the parents of the founders are not included in the pedigree. An individual's genotype at a specific marker locus consists of two alleles. One allele is inherited from the mother (maternal

allele) and the second allele is inherited from the father (paternal allele). Further, n is the number of individuals in the pedigree, f the number of founders, $n-f$ the number of non-founders and $m = 2(n-f)$ is the number of meioses present in the pedigree. The alleles of the founders may be described as

$$g = (g_1, g_2, \ldots, g_{2f}). \tag{1}$$

and the inheritance process of interest may then be seen as the distribution of the founder's alleles among the present nonfounders. When performing linkage analysis one is ordinarily using information from a whole *pedigree set* consisting of N distinct families.

Two individuals are said to share an allele *IBD* (identical-by-descent) if they have inherited exactly the same founder allele (g_i). If the alleles are of the same kind but not of identical origin the present individuals are said to share an allele *IBS* (identical-by-state). Obviously sharing an allele IBD implies sharing the allele IBS, but the reverse implication is not true.

At each marker locus, x, the whole inheritance process for a given pedigree may be described by the *inheritance vector*, $v(x)$, defined as

$$v(x) = (p_1, m_1, p_2, m_2, \ldots, p_{n-f}, m_{n-f}) \tag{2}$$

In (2) p_i and m_i are defined to equal 0 if the i:th nonfounder's, at position x, paternal and maternal allele respectively originate from a grandfather and to equal 1 if they originate from a grandmother.

In nonparametric linkage analysis (NPL-analysis) one defines *genetic linkage* to a locus as deviations from random inheritance of the founder alleles among certain subsets (i.e. the affected members) of the included pedigree members. If one wishes to discover possible linkage to a locus within a certain chromosomal region, then one has to test for linkage against a marker map (collection of present markers) that are covering the whole region of interest.

Usually there is incomplete *information* in the data set. Data may be lost, only a subset of the pedigree members may be genotyped and the markers may be non-polymorphic with a large map distance between adjacent loci. For example, if an individual is homozygous at a locus, it may be impossible to find out if it was the maternal or paternal allele that was inherited by the present offspring. Several different measures of the information contained in the data set at a locus exist. The information measure is usually ranging from 0 (no information) to 1 (full information). One commonly used entropy-based information measure was introduced in Kruglyak et al. (1996).

If one only uses the information contained in the data set at each marker locus separately one performs a *singlepoint* linkage analysis. The natural extension to that kind of approach is called *multipoint linkage analysis* and then one is using the information from several markers (at the same chromosome) surrounding a given locus. Such calculations are usually performed using some available computer program, for example GENEHUNTER (Kruglyak et al. 1996) or ALLEGRO (Gudbjartsson et al. 2000). A common way to perform the multipoint calculations is to use the theory of Hidden Markov Models (HMM), known recombination fractions between markers and marker allele frequencies to gain information from the neighbouring markers about the inheritance vector at each locus, cf. e.g. Lander and Green (1987) and Kruglyak et al. (1996).

2.2 Score Functions and the NPL-Score

To statistically quantify the evidence of linkage at a marker locus one has to introduce some kind of *score function*. The choice of score function is not trivial. If the genetic model is known it may be possible to find a function that is optimal in terms of power to detect linkage. When considering, for example, complex diseases with unknown mode of inheritance it may be preferable to choose among score functions that are robust, w.r.t. variations in the genetic model. The function that is used throughout this work is called S_{all}. It was introduced by Whittemore and Halpern (1994) and is possibly the most commonly used score function in NPL analysis. It is defined as

$$S_{all}(v) = 2^{-a} \sum_{h} \prod_{i=1}^{2f} b_i(h)! \tag{3}$$

where a is the number of affected individuals in the pedigree, h is a selection that picks one of the two present alleles for each affected individual, $b_i(h)$ is the number of times founder allele number i ($i = 1 \ldots 2f$) appears in h (given v) and the sum includes terms related to all the possible 2^a ways of choosing h. The score may be seen as the average number of permutations that preserve a collection obtained by choosing one allele from each of the affected individuals in the pedigree (cf. Kruglyak et al. 1996). A lot of articles about the performance of different suggested score functions applied to situations of different models of inheritance have been published. The performance of S_{all} is usually, especially when the information content is

reasonably high, quite robust with respect to variations in the genetic models (Whittemore and Halpern 1994; Kruglyak et al 1996; Davis and Weeks 1997; McPeek 1999; Feingold et al. 2000; Sengul et al. 2001).

To get a proper test statistic one usually standardizes the score function as

$$Z(v) = \frac{S(v) - \mu}{\sigma}, \tag{4}$$

where $\mu = \sum_w S(w)p_v(w)$, $\sigma^2 = \sum_w S(w)^2 p_v(w) - \mu^2$ and $p_v(w) = 2^{-m}$ is the probability distribution of the inheritance vector under the null hypothesis H_0 of *no linkage*. It is then an easy task to show that (under H_0) $E(Z_1(v)) = 0$ and $V(Z_1(v)) = 1$.

The NPL score for one family can now be written as

$$\bar{Z}(x) = \sum_w P(v(x) = w)Z(w), \tag{5}$$

where $P(v(x) = w)$ is the probability function (at position x) for the inheritance vector $v(x)$ given the marker data. With perfect information contained in the marker data at position x we get $\bar{Z}(x) = Z(v(x))$.

For a collection of N pedigrees, the *NPL score* (cf. Kruglyak et al. 1996; Kong and Cox 1997; Gudbjartsson et al. 2000) is defined as

$$\bar{Z}(x) = \frac{\sum_{k=1}^{N} \gamma_k \bar{Z}_k(x)}{\sqrt{(\sum_{k=1}^{N} \gamma_k^2)}}, \tag{6}$$

where $\bar{Z}_k(x)$ is the NPL score assigned to the k:th pedigree and γ_k its corresponding weight. With full information from the markers $E(\bar{Z}(x)) = 0$ and $V(\bar{Z}(x)) = 1$. If the number of families is large (how large depends on the variablility among the present pedigree structures and the different given weights), then $\bar{Z}(x)$ will be approximately *standard normally distributed; N(0,1)*. This property is extensively used in the existing theory of linkage analysis. In the above case the score is a weighted combination of the different pedigree scores. Here the weights might be related to (functions of) the individual pedigree size, the pedigree information, the pedigree structure or the family scores at different loci located on other chromosomes. This last *conditional two-locus NPL-analysis* approach has been used, for instance, in Cox (1999), Kämpe (2001) and Ängquist (2001).

For testing the presence of any disease locus one uses the maximum NPL defined as

$$\bar{Z}_{max} = sup\{\bar{Z}(x),\ x \in \Omega\}, \tag{7}$$

where Ω is the chromosomal region(s) of interest in the study.

2.3 Significance of the Results

To be able to probabilistically interpret the significance of a certain observed maximal linkage score $\bar{z}_{max}$ we need to calculate the appropriate p-value

$$\alpha(\bar{z}_{max}) = P(\bar{Z}_{max} \geq \bar{z}_{max}|H_0), \tag{8}$$

which tells us how likely it is to find a maximal NPL score at least as large as $\bar{z}_{max}$ under the *null hypothesis* H_0 that no $x \in \Omega$ is linked to the disease locus.

The genome-wide p-value, $\alpha(\bar{z}_{max})$, may sometimes, if the sizes of the pedigrees and the number of markers are small enough, be computed exactly. Otherwise, one has to approximate the p-value using simulation techniques or some available asymptotic approximation formula. Often an assumption of normally distributed linkage scores is made.

An interesting thing to notice is that $E(\bar{Z}(x)) = 0$ but $V(\bar{Z}(x)) < 1$ when the markers are not fully informative at x. This makes the p-value based on the *perfect data approximation* $V(\bar{Z}(x)) = 1$ statistically conservative. Kruglyak et al. (1996) also points out that though the p-value is conservative it sacrifices relatively little power except for situations where the information content is low.

An asymptotic approximation formula regarding the probability $\alpha(\bar{z}_{max})$ is given by Lander and Kruglyak (1995):

$$\alpha(\bar{z}_{max}) \approx 1 - e^{-\mu(\bar{z}_{max})}. \tag{9}$$

Above, $\mu(z)$ is an approximation for the mean number of regions where the linkage score process exceeds the threshold z. This essential parameter may be more explicitly defined as

$$\mu(z) = [C + 2\rho G z^2]\alpha_{pt}(z), \tag{10}$$

where C is the number of chromosomes included in the scan and G is the total genetic (in this case sex-averaged) length of these chromosomes. According to Collins et al. (1996) and Ott (1999) $G = 35.75$ Morgans if all the 22 autosomes are included in the scan. Moreover, the variable z refers to the used numerical threshold, $\alpha_{pt}(z)$ is the pointwise significance level of

exceeding z (i.e. $\alpha_{pt}(z) = 1 - \Phi(z)$, since a normally distributed linkage score is assumed). Finally, under H_0 the crossover rate ρ is related to the autocorrelation function $r_{\bar{Z}}(\cdot)$ of the score function (Lander and Kruglyak 1995), where the stationary process $\bar{Z}(x)$ is assumed to be observed under perfect marker information. The crossover rate is defined as

$$\rho = -\frac{r'_{\bar{Z}}(0)}{2}, \tag{11}$$

where the derivative is taken from above. It measures the amount of fluctuation of the NPL score statistic $\bar{Z}(\cdot)$ and therefore depends both on the family structures in the pedigree set as well as on the choice of score function.

One example of a process satisfying (11) is the *Ornstein-Uhlenbeck (OU)* process with mean zero and covariance function $r_{\bar{Z}}(\tau) = e^{-2\rho|\tau|}$. However, the approximation formula (9) does not require $\bar{Z}$ to be an OU-process.

For further discussion about the derivation of the formula above and its validity cf. Appendix A and for complementary reading in this area cf. Lander and Botstein (1989) or Feingold et al. (1993). These articles contain further references to extreme value theory for Gaussian processes forming the basis of (9)-(10).

3 Methods

In this chapter we first describe how to calculate the crossover rate, ρ, for a general pedigree set of an arbitrary structure and how to use Monte Carlo simulations when exact calculations are too time consuming to perform. Next, we adjust the approximation formula (9) for p-value calculations when the marginal distribution of the linkage score is not normal. A more detailed and technical treatment of corresponding issues will be given in Appendices B-D. Last, the topic of calculations of p-values using straightforward Monte Carlo simulations in the context of nonparametric linkage analysis is discussed.

3.1 Calculating the Crossover Rate

If using (11) as a starting point it is possible to show that the following equality holds for a single pedigree of a completely general structure:

$$\rho = \frac{1}{4} \cdot \lambda \cdot 2^{-m} \sum_{w} \sum_{j=1}^{m} (Z(w) - Z(w + e_j))^2, \tag{12}$$

where λ equals 1 whenever the genetic length is given in Morgans (0.01 if given in centiMorgans), w ranges over all 2^m binary vectors of length m and e_j is a binary vector with 1 in position j and zeros elsewhere. For more details cf. Appendix B and Hössjer (2002). One may notice that this calculation is easily implemented in a computer program by, for instance, using binary to decimal conversion for the different inheritance vectors and an index matrix reflecting all vector transformations $w \to w + e_j$.

Considering a pedigree set consisting of N pedigrees it is then possible to weight the different ρ-values into an overall value

$$\rho = \sum_{i=1}^{N} \psi_i^2 \rho_i, \tag{13}$$

where ρ_i is the crossover rate and $\psi_i = \gamma_i / \sqrt{\sum_{k=1}^{N} \gamma_k^2}$ the normalized weight of the ith pedigree. By definition we have $\sum_{i=1}^{N} \psi_i^2 = 1$.

3.2 Calculating the Crossover Rate Using Monte Carlo Simulation

The computational complexity $O(2^m)$ in (12) can be reduced to $O(2^{m-f})$ using founder phase symmetry (Kruglyak et al. 1996; Gudbjartsson et al. 2000). Still, for large pedigrees, the computational complexity gets so high that the calculations can't be performed within a reasonable time limit. Under these circumstances an option is to use Monte Carlo simulation to approximate the ρ-value.

First notice that the expression of ρ may be reformulated as

$$\rho = \frac{1}{4} \cdot \lambda \cdot E(f(v)), \tag{14}$$

where $f(w) = \sum_{j=1}^{m} (Z(w + e_j) - Z(w))^2$ and v is random with $P(v = w) = 2^{-m}$ for each w.

A proper Monte Carlo approximation of ρ is thus:

$$\hat{\rho} = \frac{\lambda}{4} \frac{1}{K} \sum_{i=1}^{K} f(v_i), \tag{15}$$

where $\{v_i\}_{i=1}^{K}$ are independent and identically distributed (i.i.d.) with the same distribution as v and K is the number of simulated inheritance vectors.

3.3 Adjusting the Approximation Formula

In this section the theory of the non-normality adjusted asymptotic approximation formula will be discussed. Only the main results will be given, for more details consider Appendix C.

Consider the NPL-score $\bar{Z}(x) = \sum_{k=1}^{N} \psi_k \bar{Z}_k(x)$ with $\sum_{k=1}^{N} \psi_k^2 = 1$. The marginal distribution function of $\bar{Z}(\cdot)$ under H_0 is by definition

$$F(z) = P(\bar{Z}(x) \leq z | H_0). \tag{16}$$

Under perfect marker assumption is $F(\cdot)$ independent of x. In the original asymptotic approximation formula $F = \Phi$ is assumed. A possible approximate correction of the distributional function, $F(\cdot)$, with respect to deviations from the standard normal distribution, is to use Edgeworth expansions (see McCune and Gray 1982) but in this case it is actually possible to calculate $F(\cdot)$ exactly with arbitrarily small errors.

Define

$$Y(x) = g^{-1}(\bar{Z}(x)), \tag{17}$$

where $g = (F^{-1} \circ \Phi)$. Since F is the distribution function of a discrete random variable, g does not have a unique inverse. We put $g^{-1} = (\Phi^{-1} \circ \tilde{F})$, where $\tilde{F}$ is the version of F which at points of discontinuity takes values halfway between the right-hand and the left-hand limits, cf. Appendix C. Here Y is a stationary process, whose marginal distribution converges to Φ as F tends to a continuous function. Now it is possible to approximate the unknown p-value by the following formula

$$\alpha(z) = P(\bar{Z}_{max} \geq z | H_0) = P(Y_{max} \geq g^{-1}(z) | H_0). \tag{18}$$

In Appendix D it is shown how $\rho_Y = -r_Y'(0)/2$ may be derived from $\rho = \rho_{\bar{Z}}$. By combining (9) and (18) and replacing ρ with ρ_Y we finally end up with

$$\alpha(z) \approx 1 - e^{-\mu_{adj}(z)}, \tag{19}$$

where

$$\mu_{adj}(z) = [C + 2 \cdot \rho_Y \cdot G \cdot g^{-1}(z)^2] \cdot \alpha_{pt}(g^{-1}(z)). \tag{20}$$

3.4 Computing the p-value by Monte Carlo Simulations

An obvious way of approximating the p-value $\alpha(z)$ is to generate i.i.d. replicates $\bar{Z}^1_{max}, \ldots, \bar{Z}^S_{max}$ of $\bar{Z}_{max}$ (under H_0) and then put

$$\alpha(z) \approx \frac{1}{S} \sum_{i=1}^{S} I(\bar{Z}^i_{max} \geq z), \tag{21}$$

where $I(A)$ is the indicator function for the event A. We will use the Monte Carlo approximation (21) in order to check the validity of the approximative p-value formulas (9) and (19).

We generate each $\bar{Z}^i_{max}$ under H_0 and the perfect marker information assumption. This strongly reduces the complexity of the $\bar{Z}^i_{max}$-computation, since no hidden Markov algorithm is needed for computing the family scores corresponding to $\bar{Z}^i_{max}$. Assuming no chiasma interference and considering the inheritance on different chromosomes as independent, we simply progress the components of the inheritance vector as independent and stationary Markov processes with two states - 0 and 1 - and intensity matrices

$$\begin{pmatrix} -\lambda & \lambda \\ \lambda & -\lambda \end{pmatrix}. \tag{22}$$

This means that jumps (crossovers) occur according to a Poisson process with intensity λ. Since each family score is a deterministic function of the inheritance vector at the same locus the total NPL score and hence $\bar{Z}^i_{max}$ is easily computed from (6) once all inheritance vector processes have been simulated. Further results on simulation of linkage scores can be found in e.g. Boehnke (1986), Ploughman and Boehnke (1989), Ott (1989) and Terwilliger et al. (1993).

4 Results

In the following section we have applied the theory discusssed in the last section to different kinds of pedigree sets. Here the main results will be given and briefly discussed. All the necessary calculations have been made in *MATLAB*.

4.1 Calculating the Significance Using the Crossover Rate

As noted above the score function used in this work is S_{all}. First we investigated the performance of the crossover rate calculation (12) for pedigrees of seven different structures. The results of the calculations are shown in table 1 and the appropriate pedigree structures are graphically displayed in figure 1 and 2. Pedigree examples number 1-4 replicate the results found by Lander and Kruglyak (1995). Pedigrees number 6-7 are both present in the second BOTNIA-study, which is further described in the next subsection.

As pointed out above the ρ-value measures the fluctuation rate of the score function $\bar{Z}(x)$. This means, loosely speaking, that a high value of ρ will make it easier for the process $\bar{Z}(x)$ to attain extreme values at some locus and therefore, with a higher probability, exceed a given threshold value z. The p-value will in this case turn out to be higher than for a low value of ρ.

4.2 Further Properties of the Crossover Rate

To get a clearer view of the distribution of the crossover rate we calculated the parameter ρ for the whole pedigree set included in the second BOTNIA-study (see Parker 2001; Lindgren 2002). This set consisted of 337 different pedigrees, which included individuals originating from Finland and Sweden. Exact calculations were performed for the 324 families which had $m \leq 20$ meioses. The remaining crossover rates (13 families) were estimated using the Monte Carlo simulation technique (15) with K=1000. In table 2 a summary of the outcomes of the ρ-calculations, the exact and the Monte Carlo simulated, is given.

A graphical presentation of the ρ-values (total case) may be seen in figure 3. The figure shows plots of ρ against both the variables m (number of meioses) and n (number of individuals). There is some positive correlation between the variables in both of these cases (cf. table 3).

The reason why the ρ-values for some pedigrees equals 0 is that in these specific cases the value of the score function is independent of the inheritance vector. Using (12) or the formulas in Appendix B one may easily find that ρ in this case is zero. A simple example of a pedigree with this property is a family consisting only of two unaffected parents and one affected child.

4.3 Calculating the p-value Using Simulations

To be able to calculate appropriate p-values we made a few assumptions. The following definitions is valid throughout the rest of this section:

- The total sex-averaged genome length G is measured in Morgans and, according to Collins et al. (1996), equals 35.75 Morgans.

- We used equal weighting for the N different pedigrees included in the pedigree set (i.e. $\gamma_k = 1;\ k = 1, ..., N$).

We applied the simulation to the seven pedigrees displayed in figures 1-2. The number of simulations used (in all cases) was $S = 10000$ and we compared the performance of the traditional asymptotic approximation formula (9)-(10), with ρ as in (12), to the simulation results. Homogenous (i.e. all of identical structure) pedigree sets of sizes $N = 60$ and $N = 180$ respectively were used for all the seven examples. All the results may be seen in figure 4 where the p-values of exceeding thresholds $2.0, 2.1, \ldots, 6.0$ are shown.

The distribution of the simulated process is depending on both the number of families, their pedigree structure and the choice of scoring function. The performance of the standard approximation formula depends on how well the discrete distribution of $\bar{Z}$ is approximated by the standard normal distribution (see the next subsection) and an increase of the number of pedigrees improves the performance of the approximation formula by the *central limit theorem*.

4.4 Calculating the p-value Using the Adjusted Approximation Formula

We tested the accuracy of the adjusted p-value approximation (19). The number of pedigrees was set to $N = 60$ throughout.

As examples we once more used the seven family structures of figures 1-2. Comparisons, with respect to their performance, between the original asymptotic approximation formula, the Monte Carlo simulation technique and the adjusted asymptotic approximation formula is made in each case. The results are given in figure 5.

The performance of the traditional Lander/Kruglyak formula is to a large extent depending on how well the distribution of $\bar{Z}$ is approximated by a standard normal distribution. Loosely speaking, skew distributions with a

thick (narrow) right tail will give anticonservative (conservative) results for extreme z-values (valid for p-values corresponding to that region of the tail for $F_{\bar{Z}}$) and symmetric distributions with truncated tails, for instance because of very few present meioses (m small), will give conservative p-values for large thresholds z. This is formalized by using the theory of Edgeworth expansions (see e.g. McCune and Gray 1982). In our case we calculated the first four *cumulants* of the distributions (table 4) and continued with calculating

$$\begin{aligned} F_{diff}(x) &= F_{\bar{Z}}(x) - \Phi(x) \\ &\approx -\frac{k_3}{6}(x^2-1)\Phi'(x) - \frac{k_4}{4}(x^3-3x)\Phi'(x) \\ &\quad -\frac{k_3^2}{72}(x^5-10x^3+15x)\Phi'(x), \end{aligned} \tag{23}$$

(figure 6) where k_3 and k_4 are the third and fourth cumulant respectively and F_{diff} measures the numerical difference, at value x, between the distribution for $\bar{Z}$ and a standard normal variable. The cumulants may be calculated recursively using the moments of the given discrete distributional function (cf. e.g. Lehmann and Casella 1998).

Some further information regarding the skewness of these seven distributions may be seen in table 5, where some numerical properties of these discrete distributions are shown for $N = 60$. In table 5 column 1 is related to the previous enumeration of the seven pedigrees, column 2-4 includes the minimal, maximal and median normalized score for $\bar{Z}_i$. Column 5-6 shows the minimal and maximal values of $\bar{Z}$ for the given pedigree set, i.e. the most extreme values with probablity mass greater than zero. In our case these values are $\sqrt{60}$ times the column 2-3 values as we assume equal weighting of the included pedigrees.

Looking at figures 5-6 one may see that the Edgeworth expansions explain the bias of the normality based p-value approximations (9)-(10). Pedigrees 1-3 and 7 have positive deviations $F_{diff}(z)$, for large z, from the normal distribution and the traditional asymptotic approximation formula is therefore, under these circumstances, conservative. As expected, the opposite implication is true for pedigree 4-6. A negative deviation F_{diff} implies anticonservativeness. Further, the adjusted approximation formula seems to perform better than the original formula in all of the seven cases.

4.5 Two Real Data Sets

We have also tried the new method on two different real data set examples.

First, we used pedigree structure data from the previously discussed BOTNIA-study, but in this case we only used the 266 pedigrees with both the number of meioses $m \leq 12$ (to decrease the computational complexity) and which also had a non-constant outcome of the score-function ($\sigma > 0$ in (4) and $\rho > 0$). These families had crossover rates in the interval $[1.0; 2.4309]$ and the corresponding mean value was 2.0390. The results of the p-value estimates are displayed in figure 7. The same general conclusions as in the previous cases seem to be valid in this case and the results confirm, with satisfactory accuracy, the theory. (An Edgeworth expansion not shown here gives negative deviation F_{diff}, implying anticonservative tests).

Next, we investigated the performance of the adjusted approximation formula when applied to a second pedigree example. Here the pedigree set consists of only one single and large pedigree with an extremely skew NPL distribution. A presentation of the pedigree is given in figure 8. The structure of the pedigree forced us to strongly increase the value of the number of influential Hermite polynomials l, see Appendix D. The reason for this is that $\sum_{k=1}^{5} \alpha_k^2 = 0.8795$ which is far from 1; see (30). In order not to get very conservative significance results, cf. (31), we had to increase l all the way up to 100 which gave us $\sum_{k=1}^{100} \alpha_k^2 = 0.9963$. We also found that $\rho = 4.1288$ and $\sum_{k=1}^{100} k\alpha_k^2 = 8.9602$ which forced the final crossover rate (31) to equal $\rho_Y = 0.4608$. This pedigree illustrates the importance of adjusting for non-normality, see figure 9 where the results are displayed. The reason for this successful behaviour is depending on the large distributional deviation of F from normality.

4.6 Properties and Performance

It is clear that the adjustment for non-normality (9)-(10) will gain importance and be more accurate than (19)-(20) the more F departs from Φ. This may be one reason behind the fact that the formula seems to be less superior and preferable in the BOTNIA-case. When F is close to normal the two methods are virtually equal.

The results in figures 5,7 and 9 indicate that formula (19) has a tendency of being slightly conservative. There may be several reasons for this. First, one may choose the parameter l (cf. Appendix D) to be too small, which

then forces ρ_Y in (31)-(32) to be too large which in turn implies that the p-values are overestimated. Secondly, the Lander/Kruglyak formula (9)-(10) is most accurate for OU-processes. However, the true covariance function of $\bar{Z}$ differs from $e^{-2\rho|t|}$, and this may cause some conservativeness of (9)-(10) in the normal case and of (19)-(20) otherwise. A third possible source of conservative p-values may be the choice of clumping rate and the $L = 0$ correction (25) in Appendix A.

One may notice that when $N = 1$ it is possible to use another method when calculating the adjusted crossover rate ρ_Y. This approach is based on the substitution of Z with $g^{-1}(Z)$ in the original crossover formula (12). In the second real data example this method suggests a substantially larger value of ρ_Y, but the performance of the p-value estimation is in this case not as accurate as when using the first method. In other cases this method may be preferable.

Further discussions of related issues, e.g. concerning distributions and their correspondance with skewness, conservativeness and deviations from the standard normal distribution, may be found in Kong and Cox (1997), Nicolae et al. (1998) and Sengul et al. (2001). In these articles some of the problems mentioned above are, for instance, discussed in contexts of different scoring functions and different pedigree structures (mainly nuclear families).

5 Discussion

5.1 General Comments

In this article we have mainly described two suggested improvements of traditional significance calulation through the approximation formula given in Lander and Kruglyak (1995).

A method of calculating the crossover rate ρ for a general pedigree set is given. This parameter affects the p-values in such a way that a large ρ-value will increase the corresponding p-value because of the increased fluctuation in the NPL scores. The p-value in that sense depends on the number of present chromosomes, their corresponding genome length, the appropriate linkage threshold, the collection of the pedigree structures and the crossover rate ρ.

In the traditional formula an assumption of normally distributed NPL scores is made. The causality if this assumption fails is that the estimated

p-values might be either conservative or anticonservative. In this work we present an approach which takes the pedigree set structure into consideration in such a way that the p-values will be corrected for deviations from the normal distribution of the NPL score. We compared the performance of the traditional and the adjusted approximation approach and found that the latter one seems to be successful, not only because of the better (less biased) numerical results but also because the p-value curve, as a function of thresholds, more closely follows the discrete behaviour of the NPL score. The use of the adjusted approximation formula also forces us to update the crossover rate with respect to the distribution function $F_{\bar{Z}}$. This correction may be described using the Hermite polynomials for the distribution function for the NPL score. However, one may notice that deviations from the normal distribution is to a large extent described by cumulants (see table 4 and figure 6).

5.2 Assumptions

Throughout this article we have, to gain simplicity, assumed that some approximations are valid. For instance, we assume that there is no difference between the genetic genome lengths with respect to the genders and we therefore use sex-averaged values. One may note that it is at least no theoretical problem in generalizing (12) to sex-specific genetic lengths. If we assume λ_f and λ_m to equal the genetic map length in Morgans per unit of measurement (x) for females and males respectively we get

$$\rho = \frac{1}{4} \cdot 2^{-m} \cdot \sum_{w} \sum_{j=1}^{m} \lambda_j \cdot (Z(w) - Z(w + e_j))^2, \tag{24}$$

where λ_j equals λ_f or λ_m depending on whether the j:th meioses corresponds to the creation of an egg or sperm cell.

Moreover, we describe the intensity of crossovers to be independent of previous crossovers i.e. no chiasma interference is assumed. Further, we assume that we have fully informative data (perfect marker approximations) and this makes real data significance calculations conservative. The effects of all these approximations deserve further study.

Further, other types of linkage analysis (or other theoretical approaches) may be interesting in this context. For instance, one might consider the possibility of several interacting disease loci and perform a conditional linkage

analysis (cf. e.g. Cox et al. 1999) where weighted quantitative linkage information from the multiple loci is used.

6 Acknowledgments

This research is sponsored by the Swedish Research Council, under the contracts 6152-8013 and 621-2001-3288.

We wish to thank the Department of Endocrinology, Malmö University Hospital, in particular professor Leif C. Groop, for the permission to use the Botnia data set in the analyses.

A Extreme Value Formulas for Gaussian Processes

In this appendix we will assume that $\bar{Z}$ is a zero mean stationary Gaussian process with covariance function $r_{\bar{Z}}(t) = 1 - 2\rho|t| + o(|t|)$. Let

$$\alpha = P(max\ Z(x) \geq T \mid H_0), \quad 0 \leq x \leq L,$$

be the appropriate p-value with respect to a single chromosome of length L. If τ is the position where $\bar{Z}$ exceeds T for the first time (starting at position 0) we can rewrite α as

$$\alpha = P(\tau \leq L \mid H_0).$$

According to Aldous (1989), section D, we may approximate τ with an exponentially distributed stochastic variable with *clump intensity* $\xi = \xi(T)$ which will give us

$$\alpha \approx 1 - exp(-\xi L).$$

In this context a clump is defined as a genome section which begins with an upcrossing of T and then continues until the first downcrossing of some $T_0 < T$.

One may observe that $\alpha \to 0$ when $L \to 0$, which give us a nonconservative test. To get an exactly correct value for $L = 0$ one may use the following adjustment

$$\alpha \approx 1 - \Phi(T) \cdot exp(-\xi L). \tag{25}$$

According to (D10d) in Aldous (1989) an approximation of the clump intensity is

$$\xi(T) \approx 2\rho \cdot T \cdot \phi(T). \tag{26}$$

Next we consider the general case with C distinct chromosomes, where the ith one has length L_i. If α_i is the p-value corresponding to the ith chromosome we then derive

$$1 - \alpha = \prod_{i=1}^{C}(1 - \alpha_i) \approx \Phi(T)^C \cdot exp(-\xi G), \tag{27}$$

where $G = \sum_{i=1}^{C} L_i$ is the total genetic length.

The formula presented by Lander and Kruglyak (1995) is based on the substitution $\Phi(T) \rightarrow exp(\Phi(T) - 1)$ in (25) and (27), which makes little difference for large or moderately large thresholds. Moreover they use $\xi(T) = 2\rho \cdot T^2 \cdot (1 - \Phi(T))$ instead of (26), which is similar since $T(1 - \Phi(T)) \approx \phi(T)$ for large values of T. For some further information cf. e.g. Leadbetter et al. (1983), chapter 12, and Feingold et al. (1993).

B The Crossover Rate for a Pedigree of Arbitrary Structure

Let $\bar{Z}(x) = Z(v(x))$ be the linkage score for one pedigree under perfect marker information data. It is proved in Hössjer (2002) that

$$Var(\Delta(h)) = Var(\bar{Z}(x + h) - \bar{Z}(x)) = \sigma^2 \cdot h + o(h) \tag{28}$$

as $h \rightarrow 0$. Under H_0, $\sigma^2 = 4\rho$, and so (12) can be derived from the general formula for σ^2 given in Hössjer (2002).

C Approximating the Marginal Distribution of the NPL Score

We now describe how to compute the distribution function F in (16) with arbitrarily good accuracy. Our method is similar to the one implemented in the Genehunter program (Kruglyak et al. 1996), but in our case we used an approximation (simple and complexity efficient) of the exact distribution based on *linear binning* and *mixtures of uniform distributions*.

Choose ϵ to be a small number and let p_k represent the probability function for a discrete distribution, that approximates the distribution of $\psi_k \bar{Z}_k(x)$, with probablity mass only at the grid $\{\ldots, -2\epsilon, -\epsilon, 0, \epsilon, 2\epsilon, \ldots\}$.

To do this we note that the random variable $\psi_k \bar{Z}_k(x)$ may, with equal probability 2^{-m}, be anyone out of the $M = 2^m$ values $x_1, x_2, \ldots, x_M$ corresponding to all possible inheritance vectors. These values do not have to be numerically distinct. In fact, they never are and therefore the computational complexity when implementing this method may be considerably reduced by using different kinds of data symmetry according to the pedigree structure (cf. e.g. Gudbjartsson et al. 2000). If $l\epsilon \leq x_j < (l+1)\epsilon$ then we increase the point mass p_k only for the values $l\epsilon$ and $(l+1)\epsilon$ according to a linear binning updating procedure:

$$\begin{aligned} p_k(l\epsilon) &\leftarrow p_k(l\epsilon) + 2^{-m}\frac{((l+1)\epsilon - x_j)}{\epsilon} \\ p_k((l+1)\epsilon) &\leftarrow p_k((l+1)\epsilon) + 2^{-m}\frac{(x_j - l\epsilon)}{\epsilon}. \end{aligned}$$

The next step will then be to calculate the convolution of the N present probability distributions p_k as

$$p = p_1 * p_2 * \ldots * p_N,$$

which will be a discrete approximation of the continuous original distribution. To make the distribution for p continuous as well, we notice that $p(l\epsilon) = 0$ whenever $l < K_1$ and $l > K_2$ for some integers $K_1 < 0 < K_2$. Then we substitute the probability mass at $l\epsilon$ with a uniform distribution on the interval $[(l - \frac{1}{2})\epsilon, (l + \frac{1}{2})\epsilon]$. Explicitly, the new distribution may be written as $X + \delta$, where $X \sim p$, $\delta \sim U(-\epsilon/2, \epsilon/2)$ and δ is independent of X. The cumulative distribution function F_ϵ may then be formed as

$$F_\epsilon = \{l.i.\ \Big((l + \frac{1}{2})\epsilon, \sum_{j=K_1}^{l} p(j\epsilon)\Big);\ l = K_1 - 1, \ldots, K_2\},$$

where *l.i.* means *linear interpolation* which in each case (for each x) may be seen as being performed between the appropriate pair of values in the expression of F_ϵ above that is surrounding the value x. In other words F_ϵ is a piecewise linear function and this, in itself, forces F_ϵ^{-1} to be a piecewise linear function. Now $F_\epsilon \to F$ as $\epsilon \to 0$, so F may be computed with arbitrarily good accuracy.

D Adjusting the Crossover Rate

In this section we derive an approximation for the crossover rate ρ_Y of the process Y appearing in (20). To this end, we will use the Hermite polynomials $\{H_k\}_{k=0}^{\infty}$ (see Taqqu 1975) which form a complete orthogonal system of functions in $\mathcal{L}^2(\mathcal{R}, \varphi(x)dx)$ using the scalar product $(f,h) = \int_{-\infty}^{\infty} f(x)h(x)\varphi(x)dx$, where $\varphi = \Phi'$. The standardized Hermite polynomials may be defined recursively (see Taqqu 1975; McCune and Gray 1982) by

$$H_i(x) = \begin{cases} 1 & , \ i = 0 \\ x & , \ i = 1 \\ \frac{x \cdot \sqrt{(i-1)!} H_{i-1}(x) - (i-1)\sqrt{(i-2)!} \cdot H_{i-2}(x)}{\sqrt{i!}} & , \ i \geq 2. \end{cases}$$

If g is expanded as $g = \sum_{k=0}^{\infty} \alpha_k H_k$, with $\alpha_k =< g, H_k >$, the following identities hold:

$$0 \ = \ E(\bar{Z}(t)) \ = \ \int_{-\infty}^{\infty} g(x)\varphi(x)dx \ =< g, H_0 > = \ \alpha_0, \tag{29}$$

$$1 \ = \ E(\bar{Z}(t)^2) \ = \ \int_{-\infty}^{\infty} g(x)^2\varphi(x)dx \ = \ \sum_{k=0}^{\infty} \alpha_k^2. \tag{30}$$

Although Y has marginal distribution (close to) Φ it may not be a Gaussian process. In this article we make the approximate assumption that Y in fact is a Gaussian process, and in particular that its bivariate distributions are Gaussian. We will utilize the fundamental covariance formula $Cov(H_k(Y_1), H_l(Y_2)) = \delta_{kl} \cdot Cov(Y_1, Y_2)^k$ which holds if (Y_1, Y_2) is bivariate normal with standard normal marginals and $\delta_{kl} = 1$ if $k = l$ and 0 otherwise (cf. Taqqu 1975). Hence

$$\begin{aligned} r_{\bar{Z}}(\tau) &= \sum_{j=1}^{\infty}\sum_{k=1}^{\infty} \alpha_j\alpha_k r_Y(H_j(Y(x)), H_k(Y(x+\tau))) \\ &\approx \sum_{k=1}^{\infty} \alpha_k^2 r_Y(H_k(Y(x)), H_k(Y(x+\tau))) \\ &\approx \sum_{k=1}^{\infty} \alpha_k^2 r_Y(\tau)^k \\ &= \sum_{k=1}^{\infty} \alpha_k^2 (1 - 2\rho_Y \mid \tau \mid)^k + o(\mid \tau \mid) \\ &= 1 - 2\bar{\rho} \mid \tau \mid + o(\mid \tau \mid) \end{aligned}$$

as $\tau \to 0$, where $\bar{\rho} = \rho_Y \sum_{k=1}^{\infty} k\alpha_k^2$. Hence we may approximate $\rho_{\bar{Z}} = \rho$ in (12) by $\bar{\rho}$. This yields

$$\rho_Y \approx \frac{\rho}{\sum_{k=1}^{\infty} k\alpha_k^2}, \tag{31}$$

which is inserted into (20).

In reality one may approximate $\sum_{k=1}^{\infty} k\alpha_k^2$ by $\sum_{k=1}^{l} k\alpha_k^2$ using only the first l terms in the sum. The choice of l must be made in the light of the actual structure of the pedigree set. Often, since we ordinarily expect F to be quite close to Φ and hence g to be well approximated by the identity function, it is sufficient with a quite small value of l (e.g. 3-10). This follows from the property that in the case of an almost-normal distribution α_1 is close to 1 and α_k is close to 0 for $k > 1$. For this reason we then expect $k \cdot \alpha_k^2$ to be small for all but the first few k:s. An important exception was given in the second real pedigree set example ($l = 100$). For theoretical consistency, cf. (30), one may use the adjustment

$$\rho_Y = \frac{\rho}{max(\sum_{k=1}^{l} k\alpha_k^2, 1)}, \tag{32}$$

to rule out the possibility of $\sum_{k=1}^{l} k\alpha_k^2$ being smaller than one.

References

[1] Abney, M., Ober, C. and McPeek, M.S. (2002) Quantitative-Trait Homozygosity and Association Mapping and Empirical Genomewide Significance in Large, Complex Pedigrees: Fasting Serum-Insulin Level in the Hutterites. *American Journal of Human Genetics 70*: 920-934.

[2] Aldous, D. (1989) *Probability Approximations via the Poisson Clumping Heuristic*. Springer-Verlag.

[3] Boehnke, M. (1986) Estimating the Power of a Proposed Linkage Study: A Practical Computer Simulation Approach. *American Journal of Human Genetics 39*: 513-527.

[4] Collins, A., Frezal, J., Teague, J. and Morton, N.E. (1996) A Metric Map of Humans: 23 500 Loci in 850 Bands. *Proc.Natl.Acad.Sci. USA 93*: 14771-14775.

[5] Cox, N.J., Frigge, M., Nicolae, D.L., Concannon, P., Hanis, C.L., Bell, G.I. and Kong, A. (1999) Loci on Chromosomes 2 (NIDDM1) and 15 Interact to Increase Susceptibility to Diabetes in Mexican Americans. *Nature Genetics 21*: 213-215.

[6] Davis, S. and Weeks, D.E. (1997) Comparisons of Nonparametric Statistics for Detection of Linkage in Nuclear Families: Single-Marker Evaluations. *American Journal of Human Genetics 61*: 1431-1444.

[7] Doerge, R.W. and Churchill, G.A. (1996) Permutation Tests for Multiple Loci Affecting a Quantitative Character. *Genetics 142*: 285-294.

[8] Feingold, E., Brown, P.O. and Siegmund, D. (1993) Gaussian Models for Genetic Linkage Analysis Using Complete High-Resolution Maps of Identity by Descent. *American Journal of Human Genetics 53*: 234-251.

[9] Feingold, E., Song, K.K. and Weeks, D.E. (2000) Comparisons of Allele-Sharing Statistics for General Pedigrees. *Genetic Epidemiology 19(Suppl.1)*: S92-S98.

[10] Gudbjartsson, D.F., Jonasson, K., Frigge, M. and Kong, A. (2000) Allegro, a New Computer Program for Multipoint Linkage Analysis. *Nature Genetics 25*: 12-13.

[11] Hössjer, O. (2002) Asymptotic Estimation Theory of Multipoint Linkage Analysis Under Perfect Marker Information. *To appear in Annals of Statistics*.

[12] Kong, A. and Cox, N. (1997) Allele-Sharing Models: LOD Scores and Accurate Linkage Tests. *American Journal of Human Genetics 61*: 1179-1188.

[13] Kruglyak, L., Daly, M.J., Reeve-Daly, M.P. and Lander, E.S. (1996) Parametric and Nonparametric Linkage Analysis: A Unified Multipoint Approach. *American Journal of Human Genetics 58*: 1347-1363.

[14] Kämpe, M. (2001) Two-Locus Nonparametric Linkage Analysis for Complex Diseases. *Master's Thesis E4, Mathematical Statistics*, Lund Institute of Technology, Lund University.

[15] Lander, E.S. and Green, P. (1987) Construction of Multilocus Genetic Linkage Maps in Humans. *Proceedings of the National Academy of Sciences of the United States of America 84, issue 8*: 2363-2367.

[16] Lander, E.S. and Botstein, D. (1989) Mapping Mendelian Factors Underlying Quantitative Traits Using RFLP Linkage Maps. *Genetics 121*:185-199.

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

[18] Leadbetter, R., Lindgren, G. and Rootzen, H. (1983) *Extremes and Related Properties of Random Sequences and Processes.* Springer-Verlag.

[19] Lehmann, E.L. and Casella, G. (1998) *Theory of Point Estimation*, 2:nd edition. Springer.

[20] Lindgren, C.M., Mahtani, M.M., Widén, E., McCarthy, M.I., Daly, M.J., Kirby, A., Reeve, M.P., Kruglyak, L., Parker, A., Meyer, J., Almgren, P., Lehto, M., Kanninen, T., Tuomi, T., Groop, L.C., and Lander, E.S. (2002) Genomewide Search for Type 2 Diabetes Mellitus Susceptibility Loci in Finnish Families: The Botnia Study. *American Journal of Human Genetics 70*: 509-516.

[21] Lystig, T.C. (2002) Using Adjusted p-values to Summarize Genome Wide Scans. *Manuscript; lystig@math.chalmers.se.*

[22] McCune, E.D. and Gray, H.L. (1982) Cornish-Fisher and Edgeworth Expansions. In: Kotz, S. and Johnson, N.L. (eds) *Encyclopedia in Statistical Sciences: volume 2*: pp 188-193, John Wiley & Sons.

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

[24] Nicolae, D.L., Frigge, M.L., Cox, N.J. and Kong, A. (1998) Discussion. *Biometrics 54*: 1271-1274.

[25] Ott, J. (1989) Computer-Simulation Methods in Human Linkage Analysis. *Proceedings of National Academy of Sciences of the United States of America 86, issue 11*: 4175-4178.

[26] Ott, J. (1999) *Analysis of Human Genetic Linkage*, 3:rd edition. The John Hopkins University Press.

[27] Ott, J. and Hoh, J. (2000) Statistical Approaches to Gene Mapping. *American Journal of Human Genetics 67*: 289-294.

[28] Parker, A., Meyer, J., Lewitzky, S., Rennich, J.S., Chan, G., Thomas, J.D., Orho-Melander, M., Lehtovirta, M., Forsblom, C., Hyrkkö, A., Carlsson, M., Lindgren, C. and Groop, L.C. (2001) A Gene Conferring Susceptibility to Type 2 Diabetes in Conjunction With Obesity Is Located on Chromosome 18p11. *Diabetes 50*: 675-680.

[29] Ploughman, L.M. and Boehnke, M. (1989) Estimating the Power of a Proposed Linkage Study for a Complex Genetic Trait. *American Journal of Human Genetics 44*: 543-551.

[30] Sengul, H., Weeks, D.E. and Feingold, E. (2001) A Survey of Affected-Sibship Statistics for Nonparametric Linkage Analysis. *American Journal of Human Genetics 69*: 179-190.

[31] Taqqu, M.S. (1975) Weak Convergence of Fractional Brownian Motion and to the Rosenblatt Process. *Z. Wahrscheinlichkeitstheorie verw. Gebiete 31*: 287-302.

[32] Terwilliger, J.D., Speer, M. and Ott, J. (1993) Chromosome-Based Method for Rapid Computer Simulation in Human Genetic Linkage Analysis. *Genetic Epidemiology 10*: 217-224.

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

[34] Ängquist, L. (2001) Conditional Two-Locus NPL-Analyses: Theory and Applications. *Master's Thesis E22, Mathematical Statistics*, Lund University.

List of Tables

Table 1: Crossover rate for pedigrees of different structure.

Ex.	Pedigree structure	ρ-value (S_{all})
1	sib-pairs	2.0000
2	grandparent/grandchild	1.0000
3	uncle/nephew	2.5000
4	first cousins	2.6667
5	five affected siblings	2.0847
6	Botnia data	2.5053
7	Botnia data	2.1880

Table 2: Crossover rate; exact or Monte Carlo-approximated.

Calc.	Nr of ped.	ρ-max	ρ-min(>0)	ρ-min	mean	std
Exact	324	3.1069	1.0000	0.0000	2.0246	0.4031
Monte Carlo	13	3.5620	2.0348	2.0348	2.6208	0.4868
Total	337	3.5620	1.0000	0.0000	2.0476	0.4218

Table 3: Correlation of ρ with m and n respectively.

Var.1	Var.2	Correlation
ρ	m	0.4115
ρ	n	0.4140
m	n	0.9738

Table 4: The first four cumulants for pedigree 1-7.

Ped.	k_1	k_2	k_3	k_4
1	0	1	0	-1
2-3	0	1	0	-2
4	0	1	1.1547	-0.6667
5	0	1	1.9972	4.1044
6	0	1	2.2722	6.6118
7	0	1	0	-1.64

Table 5: Distributional properties of the normalized score functions (using S_{all}) under H_0.

	$N = 1$			$N = 60$	
Ped.	min	max	median	min	max
1	-1.4142	1.4142	0	-10.9545	10.9545
2-3	-1	1	0	-7.7460	7.7460
4	-0.5774	1.7321	-0.5774	-4.4721	13.4164
5	-0.7711	5.3142	-0.1876	-5.9728	41.1638
6	-0.9814	4.6366	-0.3541	-7.6017	35.9152
7	-1.2649	1.2649	0	-9.7980	9.7980

List of Figures

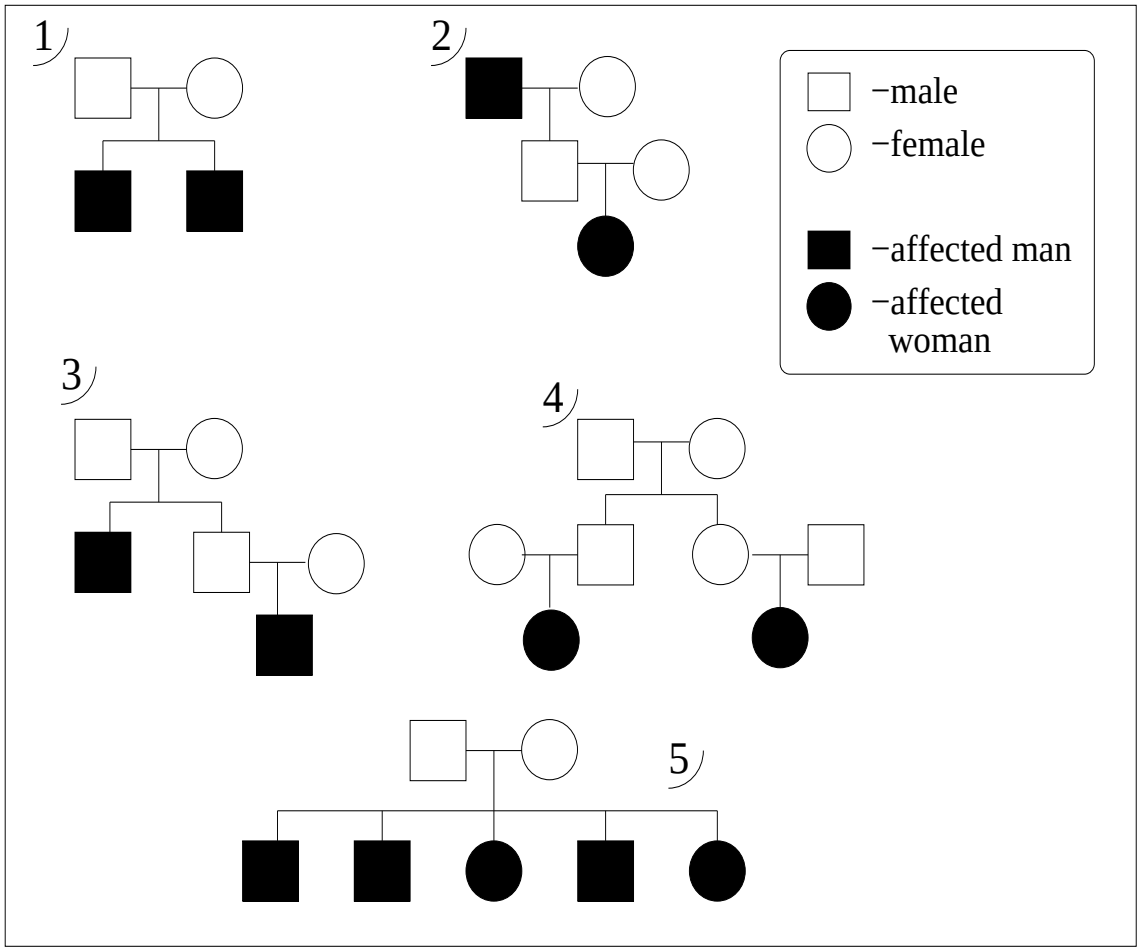

Figure 1: Pedigrees corresponding to the first five ρ-values given in Table 1.

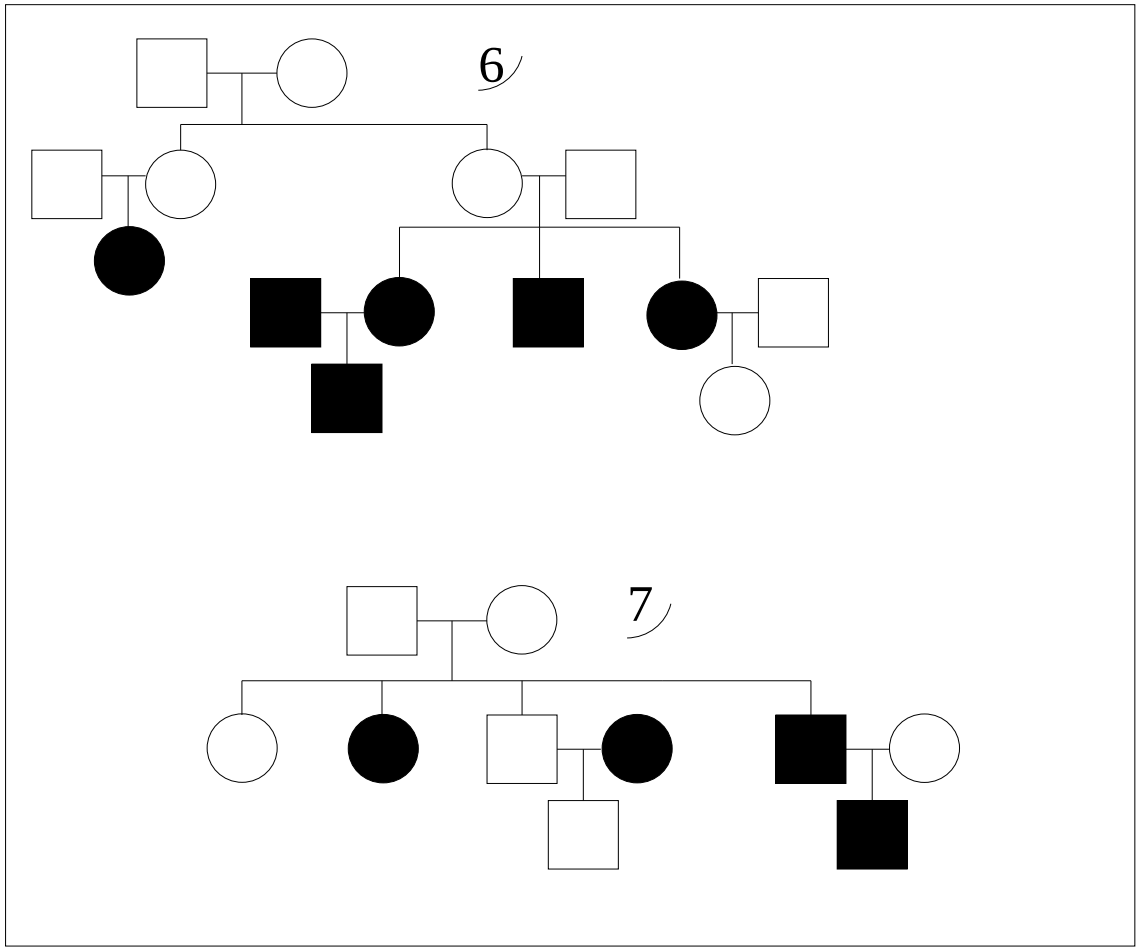

Figure 2: Pedigrees corresponding to the last two ρ-values given in Table 1.

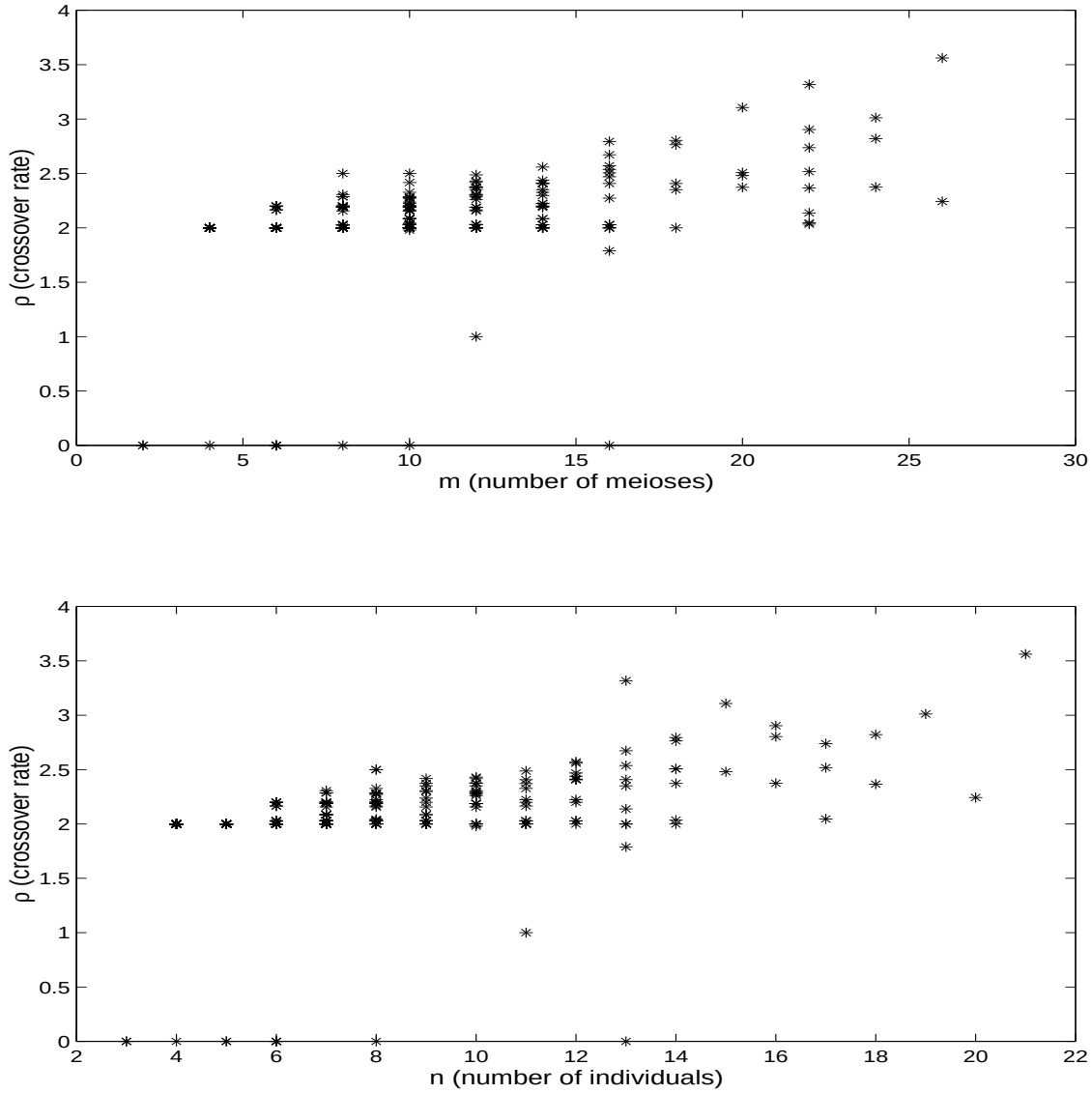

Figure 3: Plotting ρ against m and n.

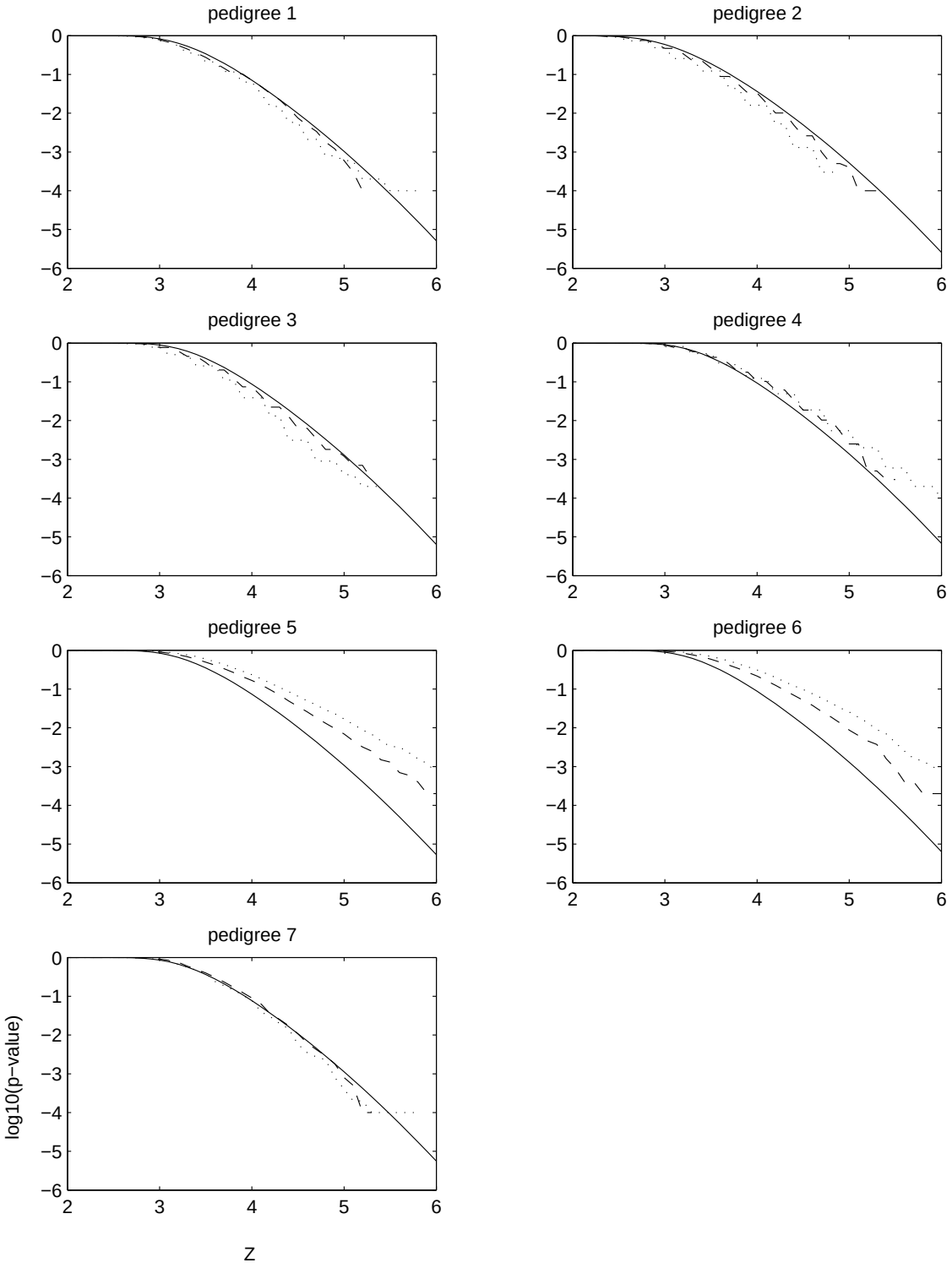

Figure 4: Comparisons between the p-values for the normal approximation (9), the simulation procedure (21) with $N = 60$ ($\cdots$) and $N = 180$ (- -).

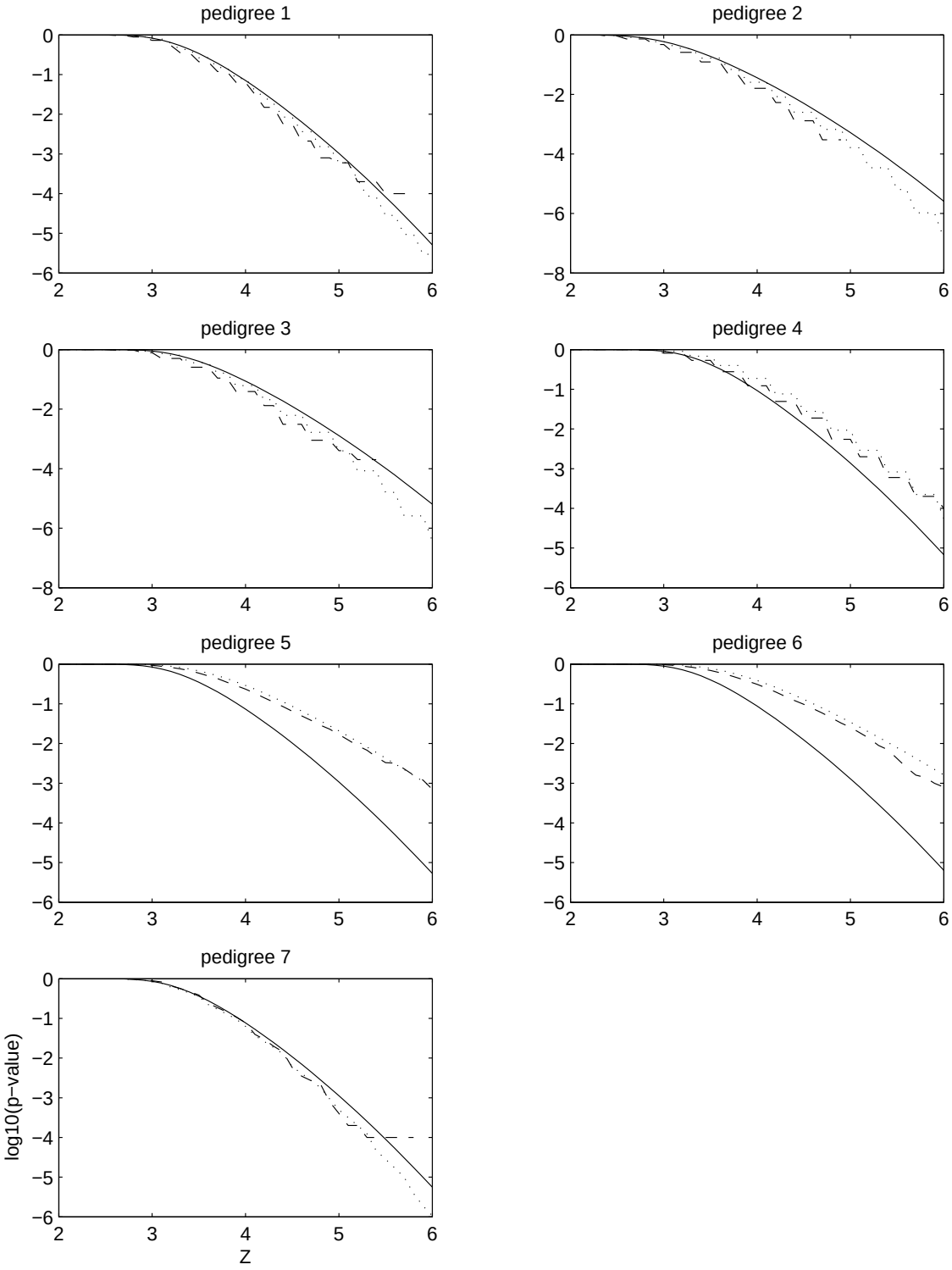

Figure 5: Comparisons between the p-values for the normal approximations (9), the simulation procedure (21) with $N = 60$ (- -) and the adjusted normal approximation (19) ($\cdots$). Further assumptions are that (see Appendix C-D) the accuracy parameter for the probability distribution approximation is set to $\epsilon = 0.001$ and that $l = 5$.

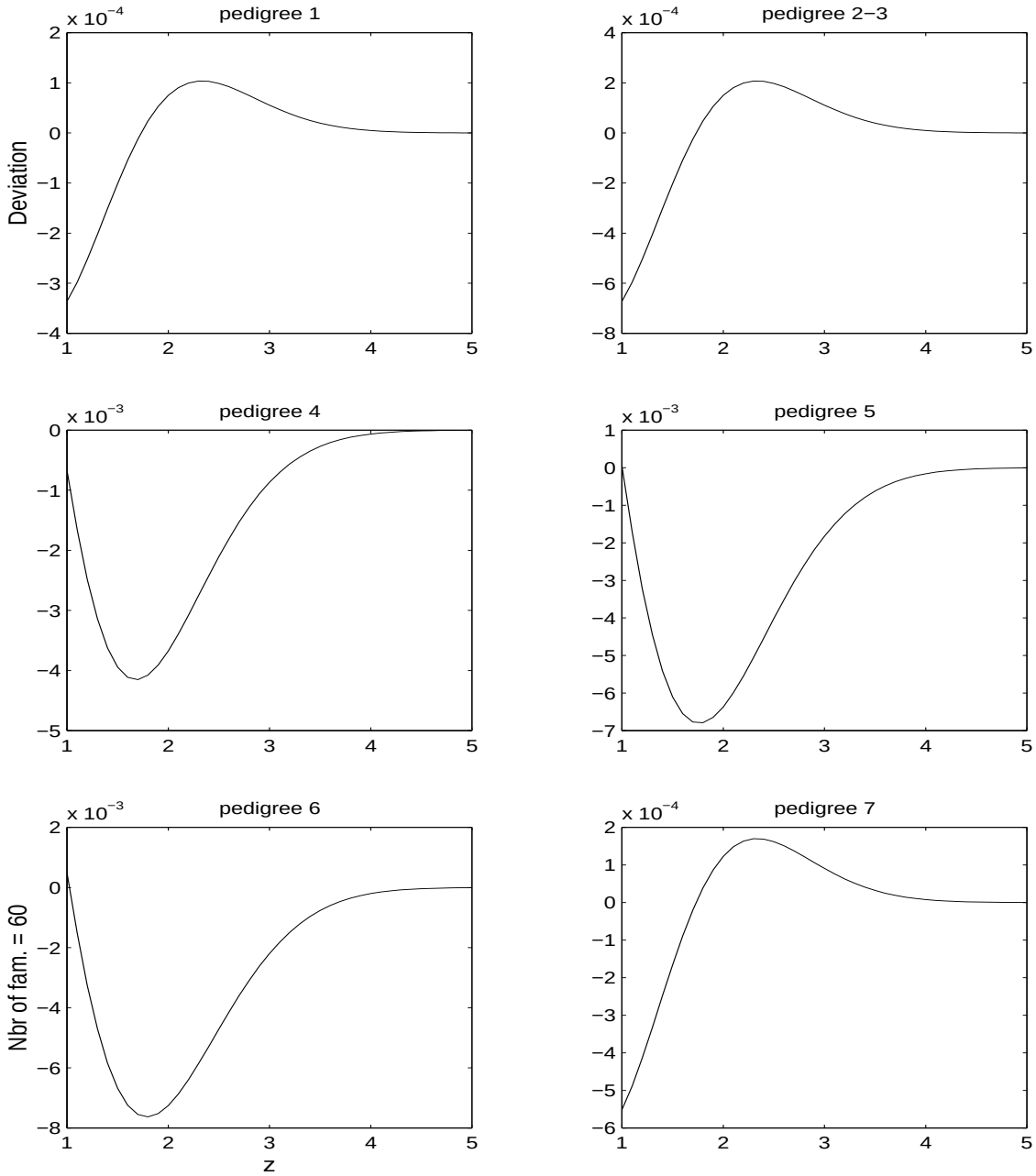

Figure 6: The distance between $F_{\bar{Z}}$ and the cumulative distribution function Φ for the standard normal distribution using Edgeworth expansions.

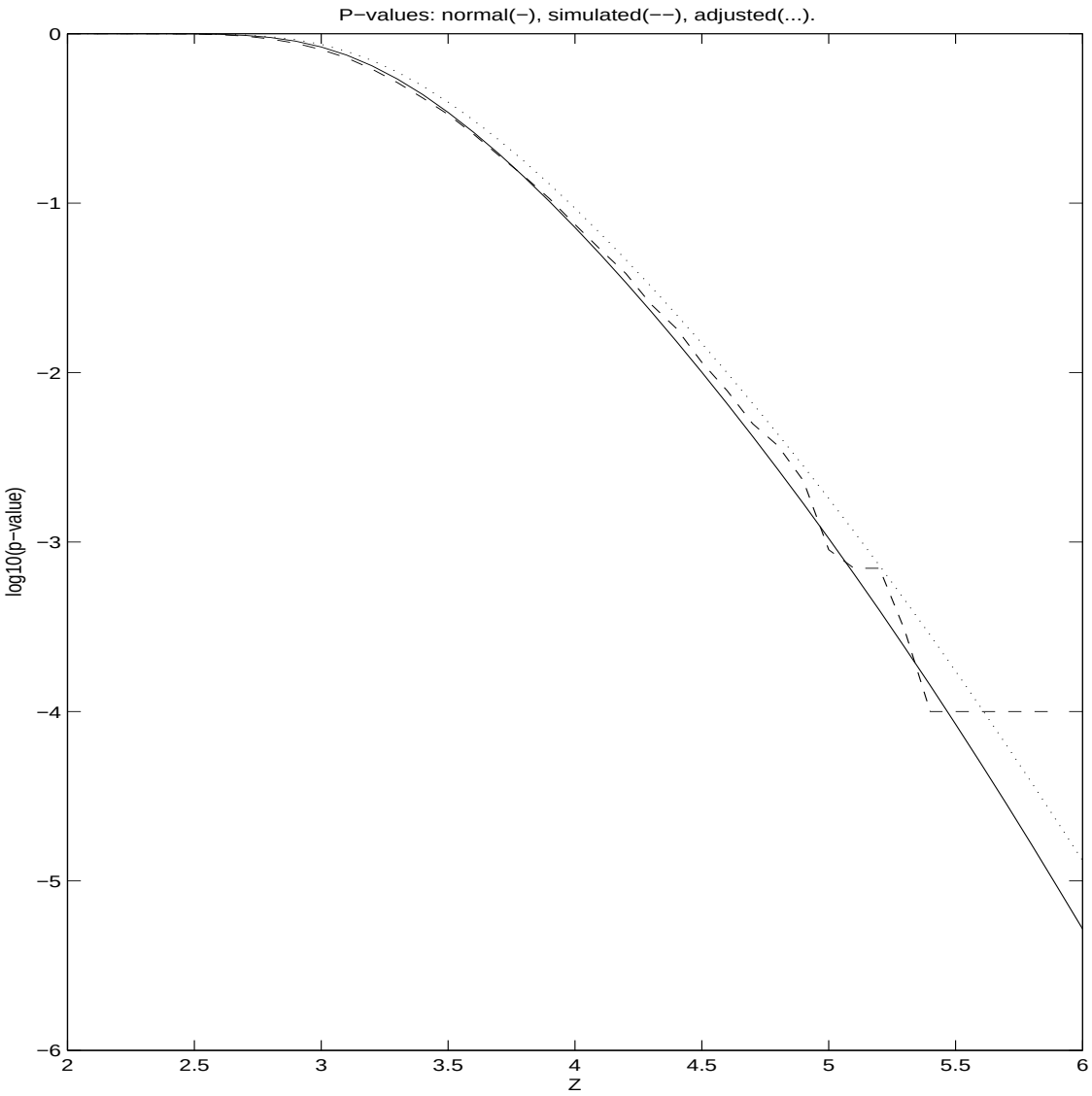

Figure 7: Comparisons, for the Botnia data set, between the p-values for the normal approximations (9), the simulation procedure (- -) given by (21) and the adjusted normal approximation (19) ($\cdots$). Further, $\epsilon = 0.0005$ and $S = 10000$.

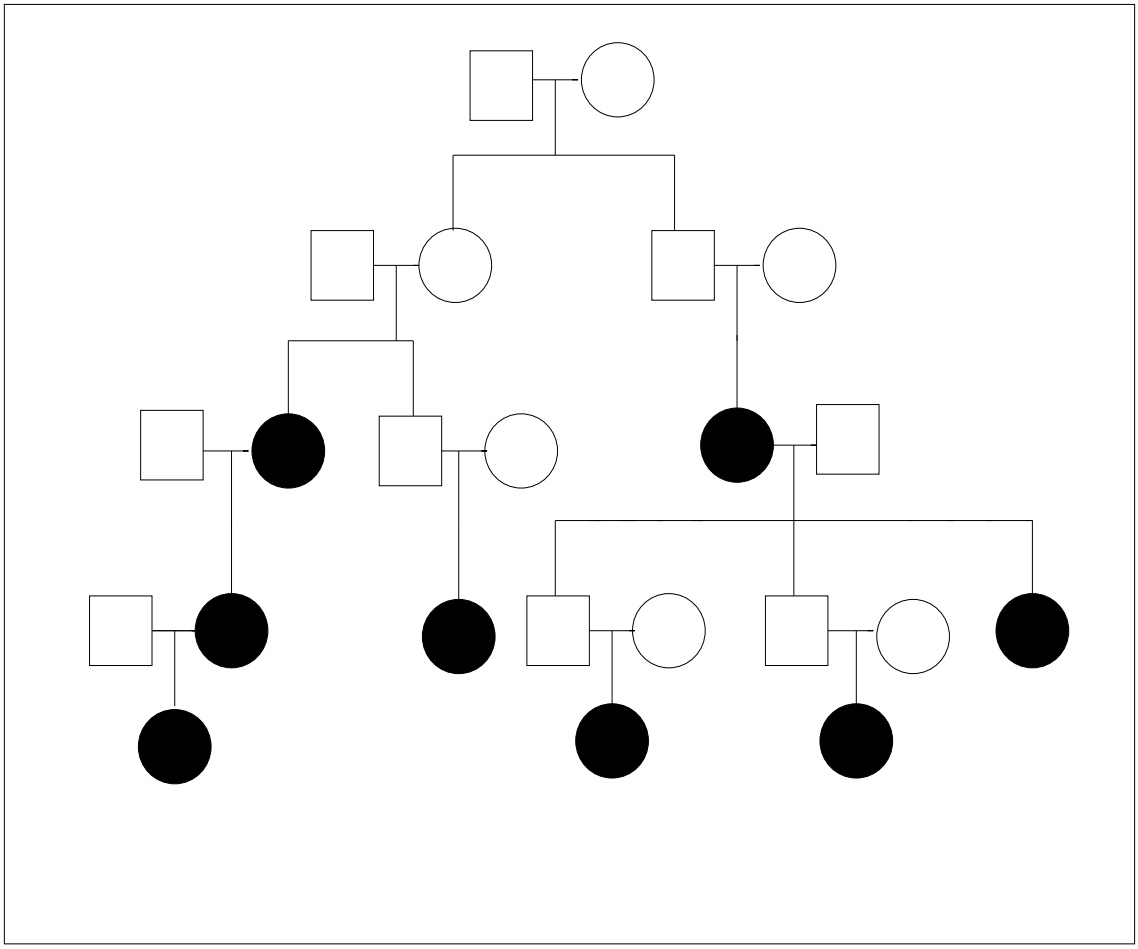

Figure 8: The pedigree structure corresponding to the second real data set.

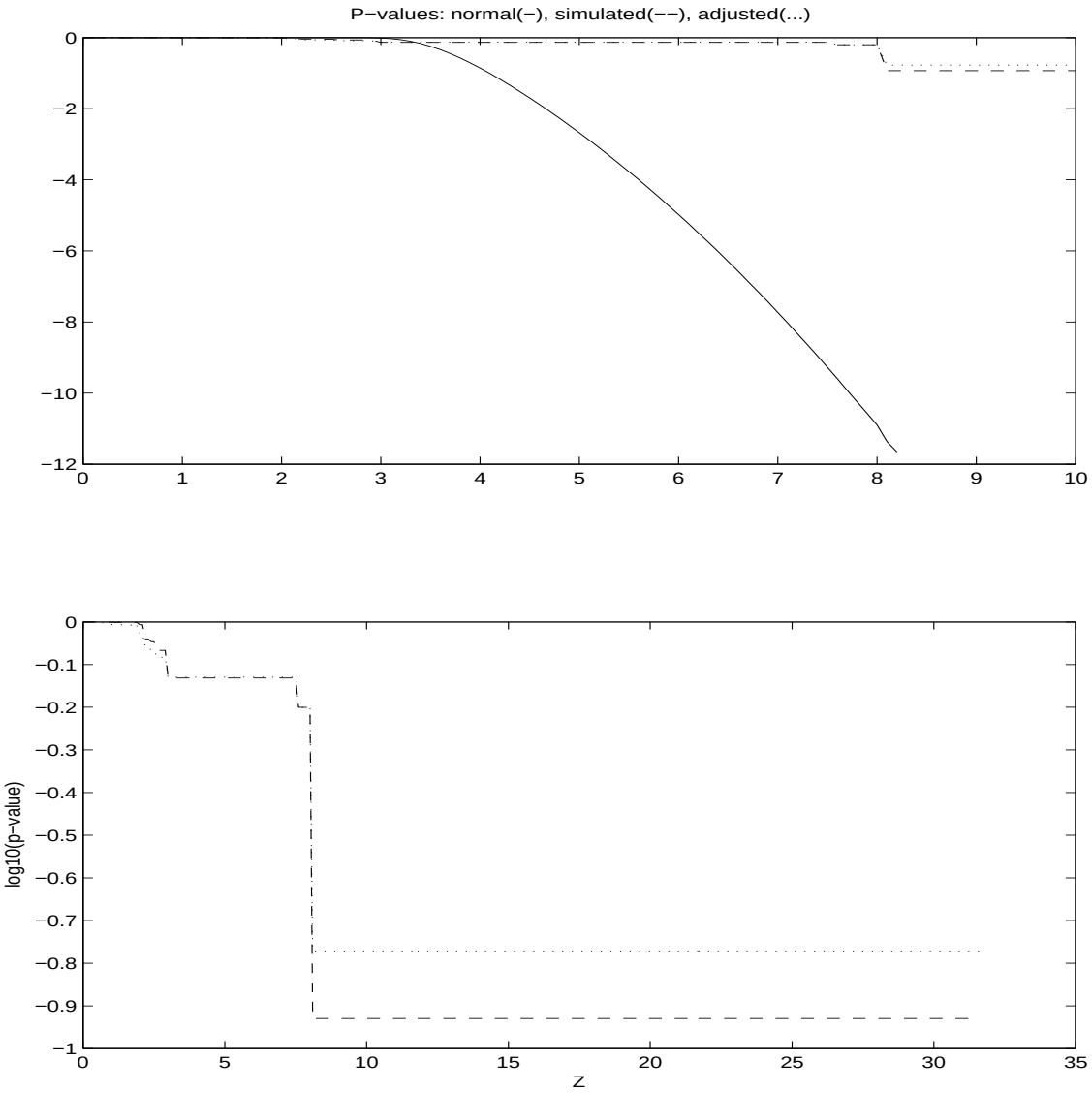

Figure 9: Comparisons, for the second real data set, between the p-values for the normal approximations (9), the simulation procedure (- -) given by (21) and the adjusted normal approximation (19) ($\cdots$). Moreover, $\epsilon = 0.001$ and $S = 100000$. The two largest values the normalized score function (4) can attain are 31.6115 and 8.0830. The largest value corresponds to the case where all the affected individuals share an allele IBD, and this possibility is the reason behind the extreme skewness of F in this case.