Boundary Crossing Probabilities in Linkage Analysis

Josée Dupuis
Genome Therapeutics Corporation, Waltham, MA

David Siegmund
Department of Statistics, Stanford University, Stanford, CA

Abstract

Two novel problems of boundary crossing probabilities that arise in genetic linkage analysis based on sib pairs are addressed by modifications of techniques developed to solve problems of sequential analysis.

1 Introduction.

Genome scans in linkage analysis lead to problems involving boundary crossing probabilities (cf. Feingold, Brown and Siegmund, 1993), which can be addressed using methods developed in sequential analysis during the 1970's. In this paper we discuss two problems where genetically natural conditions lead to novel variations.

The goal of linkage analysis is to identify regions of the genome harboring genes affecting particular traits. In humans these are often genes that increase susceptibility to particular diseases, and it is convenient to speak of "disease" genes, although other traits affected by an individual's genetic makeup can be studied similarly.

A convenient unit for the linkage analysis of human diseases is an affected sib pair. Given $N \geq 1$ unrelated sib pairs, we let $X_{i,t}^{(N)}$ denote the number of pairs that share i alleles identical by descent ($i = 2, 1, 0$) at locus t, and let $X_t^{(N)} = (X_{0,t}^{(N)}, X_{1,t}^{(N)}, X_{0,t}^{(N)})$. (An allele is shared identical by descent by two relatives if it is inherited from a common ancestor.) With probability $1/2$ a sibling pair can inherit zero or one allele identical by descent from their mother and similarly from their father. These events are independent, so the probability that two siblings share i alleles identical by descent at locus t is given by $EX_{2,t}^{(1)} = 1/4, EX_{1,t}^{(1)} = 1/2, EX_{0,t}^{(1)} = 1/4$.

For an *affected* sib pair, on a chromosome containing a disease locus at τ the (conditional on being affected) distribution of alleles shared identical

by descent at τ is of the form

$$EX_{2,\tau}^{(1)} = (1+\alpha+2\delta)/4, \quad EX_{1,\tau}^{(1)} = (1-\delta)/2, \quad EX_{0,\tau}^{(1)} = (1-\alpha)/4, \tag{1}$$

where α and δ can be expressed in terms of a genetic model for the transmission of the disease and under reasonably general conditions satisfy the constraint

$$0 \leq \delta < \alpha < 1 \tag{2}$$

(cf. Risch, 1990a,b or the Appendix). The extreme $\delta \approx \alpha$ corresponds to a rare recessive disease, while the case $\delta \approx 0$ occurs for a dominantly inherited trait.

A null hypothesis of interest is that τ is in fact unlinked to the disease, i.e., that $\alpha = \delta = 0$. Holmans (1993) calls the likelihood ratio test of this hypothesis based on $X_\tau^{(N)} = (X_{0,\tau}^{(N)}, X_{1,\tau}^{(N)}, X_{2,\tau}^{(N)})$ the "possible triangle test" because the constraints (2) force the vector consisting of the first two elements of (1), say, to lie in a triangular subregion of the unconstrained set of values. See also Faraway (1993).

In experimental genetics, which may involve agriculturally important species or animal models for human traits, one is usually concerned with quantitative phenotypes, hence linkage analysis of so-called quantitative trait loci (QTLs). In an intercross two inbred strains having differing genotypes AA and aa at each locus are crossed to produce individuals who are uniformly Aa. These are bred to one another to produce individuals with genotypes AA, Aa and aa in the expected ratio 1:2:1. A simple regression model expressing the phenotype as a linear combination of the number of A alleles at a QTL and a random error leads via a large sample approximation to a similar problem to that described above, but without the constraints indicated in (2) (e.g., Lander and Botstein, 1989, Dupuis and Siegmund, 1998). Teng (1996) in her analysis of the Haseman-Elston (1972) method for detecting linkage of QTL's in human genetics obtains a parameterization of that problem that again contains the constraints (2).

For the unconstrained problem twice the log likelihood ratio statistic is asymptotically distributed as χ^2 with two degrees of freedom under the null hypothesis of no linkage. For the problem constrained by (2) this asymptotic distribution, by a classical result of Chernoff (1954) (see also Self and Liang, 1987), is a mixture of χ_1^2 and χ_2^2 distributions. Specifically, in a form that will prove useful below, the probability that the square root of twice the log likelihood ratio statistic exceeds a threshold b is approximately

$$1 - \Phi(b) + (2\pi)^{-1}(\pi/2 - \tan^{-1} 2^{1/2}) \exp(-b^2/2). \tag{3}$$

In the preceding paragraph we have assumed that data are available on a single marker, which is either a disease locus τ itself or is unlinked to the

disease. Since τ has an unknown position on the genome, one uses an array of markers spread throughout the genome, hence makes many simultaneous tests, which lead to a problem of multiple comparisons. In Section 2 of this paper we give approximations to the significance level and power of the likelihood ratio test constrained by (2) by adapting methods developed to study sequential hypothesis tests (cf. Woodroofe, 1976, Lai and Siegmund, 1977, Siegmund, 1985 and for classical background material in sequential analysis Ferguson, 1967).

It is also of interest to estimate the location τ of a trait locus by a confidence region. This problem is described in more detail in Section 3.

2 Approximations for significance level and power.

To obtain the joint distribution of $X_t^{(1)}$ for different values of t, we assume crossovers occur along each chromosome according to a Poisson process, which by a change in the "time" scale (actually distance along the chromosome) can be assumed to be homogeneous. This is the "no interference" model suggested by Haldane in 1919, which is known not to be correct, but is still commonly used as a simple, reasonably robust model. It implies that $X_t^{(1)}$ is a three state Markov chain which changes its state at a constant rate, say $\beta/2$. The parameter β depends on the units chosen for t to describe genetic distance along the chromosome. If we use centimorgans (cM), which are defined by the property that in one unit of genetic distance there is an expectation of 0.01 crossovers per meiosis, then $\beta = 0.04$, and the human chromosomes average about 140 cM in length. From the state (0,1,0) the chain moves to each of the other states with probability 1/2. From (1,0,0) or (0,0,1) it moves with certainty to (0,1,0). See Feingold (1993) for a more detailed description of this basic model.

¿From the model of the preceding paragraph together with (1), one can show by conditioning on $X_\tau^{(1)}$ and reasonably straightforward but tedious calculations that at an arbitrary marker t the probability of sharing 2, 1, or 0 alleles identical by descent is $[1 + (\alpha + \delta)\exp(-\beta|t - \tau|) + \delta\exp(-2\beta|t - \tau|)]/4$, $[1-\delta\exp(-2\beta|t-\tau|)]/2$, and $[1-(\alpha+\delta)\exp(-\beta|t-\tau|)+\delta\exp(-2\beta|t-\tau|)]/4$. If t is on a different chromosome from τ, we set $t = \infty$, to obtain the null hypothesis probabilities.

To obtain a Gaussian process approximation to the likelihood ratio process, we introduce the notation

$$Z_{1,t} = -2(X_{1,t}^{(N)} - N/2)/N^{1/2}, \quad Z_{2,t} = 2^{1/2}(X_{2,t}^{(N)} - X_{0,t}^{(N)})/N^{1/2}.$$

Calculations based on the results stated above show that on unlinked chromosomes (i.e., for $\alpha = \delta = 0$) these two processes have expectation 0, vari-

ance 1, and are uncorrelated. Also

$$\mathrm{Cov}[\mathrm{Z}_{1,\mathrm{s}}, \mathrm{Z}_{1,\mathrm{t}}] = \exp[-\beta_1|\mathrm{t}-\mathrm{s}|], \quad \mathrm{Cov}[\mathrm{Z}_{2,\mathrm{s}}, \mathrm{Z}_{2,\mathrm{t}}] = \exp[-\beta_2|\mathrm{t}-\mathrm{s}|],$$

where $\beta_1 = 2\beta = 0.08, \beta_2 = \beta = 0.04$ (cf. Feingold, Brown and Siegmund, 1993, for details). Thus $Z = (Z_1, Z_2)$ is approximately for large N a two-dimensional Ornstein-Uhlenbeck process. On a linked chromosome having a single trait locus at τ,

$$EZ_{1,t} = \mu_1 \exp[-\beta_1|t-\tau|], \quad EZ_{2,t} = \mu_2 \exp[-\beta_2|t-\tau|], \tag{4}$$

where $\mu_1 = N^{1/2}\delta, \mu_2 = (N/2)^{1/2}(\alpha+\delta)$. The genetic constraint (2) is equivalent to $0 \le 2^{1/2}\mu_1 \le \mu_2$. Thus the triangular constraint noted by Holmans becomes the constraint that $\mu = (\mu_1, \mu_2)$ lie in a wedge in the first quadrant of the xy plane, which is defined by the lines $y = 2^{1/2}x$ and $x = 0$. The log likelihood function for the limiting Gaussian process under contiguous alternatives is

$$\mu_1 Z_{1,\tau} + \mu_2 Z_{2,\tau} - ||\mu||^2/2. \tag{5}$$

Note that although we observe the entire process indexed by t, the likelihood function depends only on the process at τ, which is itself an unknown parameter (Feingold, Brown and Siegmund, 1993).

The following two extreme cases are of interest. If we assume $\delta = 0$, which defines an additive model (or approximately a dominant model) of inheritance (Risch 1990b), then $\mu_1 = 0$, so the likelihood ratio test to detect linkage is asymptotically equivalent to the maximum over all marker loci t of

$$Z_{2,t}. \tag{6}$$

For a rare recessive trait, where $\delta \approx \alpha$, the appropriate test is based on the maximum over t of

$$[Z_{1,t} + 2^{1/2}Z_{2,t}]/3^{1/2} = 4[X_{2,t}^{(N)} - N/4]/(3N)^{1/2}. \tag{7}$$

The statistic (7) is the projection of the vector $(Z_{1,t}, Z_{2,t})$ along the line $y = 2^{1/2}x$, making an angle $\tan^{-1} 2^{1/2}$ with the positive x axis in the xy plane. The statistic (6) is obviously the projection of $(Z_{1,t}, Z_{2,t})$ along the y axis, which makes an angle of $\pi/2$ with the positive x axis.

If there were no constraints on μ_1, μ_2, the likelihood ratio statistic at a putative disease locus t would be obtained by maximizing (5) with respect to μ_1, μ_2. This yields

$$||Z_t|| = [Z_{1,t}^2 + Z_{2,t}^2]^{1/2}, \tag{8}$$

which would in turn be maximized over all marker loci t to search the genome for the disease gene. To incorporate the constraints we use (8) if the point $(Z_{1,t}, Z_{2,t})$ lies in the wedge defined by the lines $y = 2^{1/2}x$ and $x = 0$. If

the point does not lie in this wedge, we use the larger of (6) and (7). Let $||\tilde{Z}_t||$ denote the statistic so obtained. In effect the likelihood ratio test incorporating the constraints is based on (8) unless the data tell us that the mode of inheritance appears to be purely additive or purely recessive; in these extreme cases we use the statistic appropriate for the apparent mode of inheritance. (These geometric observations lead to a simple direct demonstration of (3), which is the marginal distribution of the statistic described in this paragraph at each fixed t.)

The false positive rate of the likelihood ratio test is the probability, computed under the assumption $\alpha = \delta = 0$, that the statistic described above exceeds the detection threshold b at some locus t in the genome. This probability can be evaluated approximately by adapting arguments developed to study sequential hypothesis tests (e.g., Woodroofe, 1976, Lai and Siegmund, 1977). Slightly more generally, we suppose that (fully informative) markers are placed at constant intermarker distances Δ. Recall the special function ν defined by Siegmund (1985, p. 82). For numerical purposes, for $0 < x < 2$, $\nu(x) \approx \exp(-0.583x)$; for larger x the first four terms of the defining infinite series provide a satisfactory numerical evaluation. Suppose $b \to \infty$ and $\Delta \to 0$ in such a way that $b\Delta^{1/2}$ is bounded away from 0 and ∞. Then for a single chromosome of length ℓ,

$$P\{\max_i ||\tilde{Z}_{i\Delta}|| > b\} = \ell \exp(-b^2/2)\{C_1 b^2/(2\pi) + C_2 b/(2\pi)^{1/2} + o(b)\}, \quad (9)$$

where

$$C_1 = \int_{\tan^{-1} 2^{1/2}}^{\pi/2} (\beta_2 \sin^2\omega + \beta_1 \cos^2\omega)\nu\{b[2\Delta(\beta_1 \cos^2\omega + \beta_2 \sin^2\omega)^{1/2}]\}d\omega, \quad (10)$$

and

$$C_2 = 6^{-1}(\beta_1 + 2\beta_2)\nu\{b[2\Delta(\beta_1/3 + 2\beta_2/3]^{1/2}\} + 2^{-1}\beta_2\nu[b(2\Delta\beta_2)^{1/2}].$$

In (9) the term involving b^2 accounts for the probability that $||\tilde{Z}_{i\Delta}||$ exceeds b at a point where $Z_{i\Delta}$ lies inside the wedge, while the terms involving b account for the probability that $Z_{2,i\Delta}$ or $[Z_{1,i\Delta} + 2^{1/2}Z_{2,i\Delta}]/3^{1/2}$ exceeds b for some value of $i\Delta$ where the two dimensional process is outside the wedge. For a search involving the entire genome, we can use the independent assortment of chromosomes to obtain from (9) the Poisson approximation

$$P\{\max_i ||\tilde{Z}_{i\Delta}|| > b\} \approx 1 - \exp[-L\exp(-b^2/2)\{C_1 b^2/(2\pi) + C_2 b/(2\pi)^{1/2}], \quad (11)$$

where L is the total length of the genome (approximately 3400 cM).

A slightly better approximation is presumably obtained by adding (3) to (9) as an edge correction to account for the initial marker on the chromosome.

In the special case $\Delta = 0$, the integral (10) equals $(\beta_1 + \beta_2)(\pi/2 - \tan^{-1} 2^{1/2})/2 - (\beta_1 - \beta_2)/(3 \times 2^{1/2}) = 0.0275$ for $\beta_1 = 0.08, \beta_2 = 0.04$. As $\Delta \to \infty$, the asymptotic relation $\nu(x) \sim 2/x^2$ shows that the exponent in (11) is asymptotic to L/Δ times (3), so the $Z_{i\Delta}$ for different i are treated as independent, as they should be when Δ is large.

For a human genome of 23 chromosomes averaging 140 cM in length, the edge corrected (11) yields 0.05 level thresholds of 4.30, 4.11, 3.91 and 3.78 for $\Delta = 0$, 1, 5, and 10 cM, respectively. The threshold for $\Delta = 0$ has been cited by Lander and Schork (1995). An often recommended one degree of freedom statistic is (6), which is the score statistic when δ is assumed equal to 0. Corresponding thresholds are 4.08, 3.92, 3.73 and 3.6 (Feingold, Brown and Siegmund, 1993).

The power of the likelihood ratio test is the probability under the alternative that $||\tilde{Z}_{i\Delta}||$ exceeds the threshold b at some marker near to the true disease locus. To approximate the power we let $\xi = (\mu_1^2 + \mu_2^2)^{1/2}$ denote the distance of the point $\mu = (\mu_1, \mu_2)$ from the origin. We assume for simplicity that the disease locus is exactly at one of the markers, which is not near the end of its chromosome. The point μ lies in the wedge defined by the angles $\tan^{-1} 2^{1/2}$ and $\pi/2$ measured from the positive x axis. A slightly different approximation to the power is appropriate depending on whether the point is strictly inside the wedge or on one of the edges. In the former case the power is approximately

$$1 - \Phi(b - \xi) + \phi(b - \xi)\{1/(2\xi) + (b/\xi)^{1/2}[2\nu/\xi - \nu^2/(b + \xi)]\}, \qquad (12)$$

while in the latter it is

$$1 - \Phi(b - \xi) + \phi(b - \xi)\{1/4\xi + [(b/\xi)^{1/2} + 1][\nu/\xi - \nu^2/(2(b + \xi))]\}, \quad (13)$$

where $\nu = \nu\{b[2\Delta(\beta_1\mu_1^2 + \beta_2\mu_2^2)/(\mu_1^2 + \mu_2^2)]^{1/2}\}$. These approximations are based on decomposing the event in question according as $||Z_\tau|| > b$, or the contrary and $||Z_{i\Delta}|| > b$ at some nearby marker. The details can be obtained by arguments similar to those used by Siegmund (1985) and Feingold *et al.* (1993). In the case where the disease locus is between marker loci one must condition on the value of (Z_1, Z_2) at the two flanking markers (cf. Dupuis, 1994).

Remarks. (i) It is a straightforward consequence of methods in the cited literature to prove the leading term in (9). A rigorous proof of the result stated would be substantially more difficult. If there were no constraints on μ_1 and μ_2, the error in the leading term would be expected to be of order $\exp(-b^2/2)$ (cf. Woodroofe and Takahashi, 1982, Siegmund, 1985), so it is reasonable to conjecture that the term of order $b\exp(-b^2/2)$ arising from the constraints is the correct second order term for this problem. (ii)

Several of our hypotheses can easily be weakened. For example, some simple modifications are mathematically appropriate if the markers are not equally spaced, but the approximations are not especially sensitive to this change, so for practical purposes it usually suffices to treat markers as equally spaced and use an average intermarker distance. (iii) For these approximations to hold it is not necessary that the process Z_t be exactly an Ornstein-Uhlenbeck process, but only that its covariance function behave like that of an Ornstein-Uhlenbeck process near 0. A consequence is that one can use other models for the recombination process instead of the no interference model we have assumed, since all lead to the same small time behavior for the covariance function. (iv) To achieve the best numerical accuracy from (9), one should apply it to the square root of twice the log likelihood ratio statistic. To apply it directly to the score statistic one should correct for skewness, e.g., along the lines of Tu and Siegmund (1998).

3 Confidence regions.

From (4) we see that the expectation of $Z_{i,t}$ is increasing for $t < \tau$ and decreasing for $t > \tau$. Hence τ behaves like a change-point, and estimation of τ by a confidence region is closely related to the problem of a confidence region for a change-point (cf. Siegmund, 1989). For simplicity we consider the case of mapping QTLs based on an intercross. As noted above, in this case the nuisance parameters μ_1 and μ_2 are unconstrained, hence are arbitrary real numbers. The covariance parameters can be shown to be $\beta_1 = 0.04$ and $\beta_2 = 0.02$. We also assume that the trait locus τ is a marker locus. Since one often types additional markers near a suspected locus, this hypothesis is often approximately true.

It follows from the form of the likelihood function given in (5) that the likelihood ratio statistic for testing that a true QTL is τ against the alternative that it lies somewhere else on the chromosome has as its acceptance region the event

$$A_\tau = \{\max_i ||Z_{i\Delta}||^2 - ||Z_\tau||^2 \le x\}. \tag{14}$$

Moreover, for each τ, we see from (5) that Z_τ is sufficient for μ. Hence if $x = x(Z_\tau)$ is chosen to satisfy $P(A_\tau|Z_\tau) = \gamma$, then the set of all τ such that the event A_τ occurs is a γ level confidence region. To evaluate the required probability we have the following approximation, which plays an important role in the comparative analysis of different confidence regions given by Dupuis and Siegmund (1998).

Proposition. *Let $Z_t = (Z_{1,t}, Z_{2,t})$ where $Z_{1,t}$ and $Z_{2,t}$ are independent Gaussian processes with covariance functions satisfying*

$$R_i(t) = 1 - \beta_i|t| + o(|t|) \qquad \text{as} \quad t \to 0.$$

Assume $b \to \infty, \Delta \to 0$ *and* $b\Delta^{1/2}$ *is bounded away from 0 and* ∞. *Let* $0 < ||z||^2 < b^2$ *and define* t^*, w^* *to be the solution of*

$$\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} b\, R_1(t^*)\, \cos w^* \\ b\, R_2(t^*)\, \sin w^* \end{pmatrix}.$$

Assume t^* *is contained in* $(0, l)$ *and is bounded away from the upper endpoint* $(l > 0)$. *Then*

$$P\Big\{ \max_{0 \le i\Delta \le l} ||Z_{i\Delta}|| \ge b \,\Big|\, Z_0 = z \Big\}$$

$$\sim \frac{\beta \exp[-\frac{1}{2}(b^2 - ||z||^2)]}{\left| \dot{R}_1(t^*) R_2(t^*) \cos^2 w^* + R_1(t^*) \dot{R}_2(t^*) \sin^2 w^* \right|} \nu\big(b\,(2\beta\Delta)^{1/2}\big],$$

where $\dot{R}_i(t) = dR_i(t)/dt$ *and* $\beta = \beta_1 \cos^2(w^*) + \beta_2 \sin^2(w^*)$.

For our particular application, $R_i(t) = \exp(-\beta_i |t|)$ with $\beta_1 = 2\beta_2$, so the equations defining t^*, w^* are quadratic and can be solved analytically. In this case one can obtain a completely explicit approximation, albeit involving some complicated expressions. For a maximum over markers on both sides of τ, as indicated in (14), one should double the approximation given in the Proposition.

Derivation of the Proposition: We first condition on the position where the process last exceeded the value b, how far above b it reached at that position and the angle between $Z_{1,t}$ and $Z_{2,t}$.

Define $D_i = \{j : j \ge 1, (i+j)\Delta \le l\}$, where l and Δ are fixed. Let $\omega_{i\Delta}$ be the angle between $Z_{1,i\Delta}$ and $Z_{2,i\Delta}$. Then, we can write

$$\begin{aligned} &P\Big\{\max_{0 \le i\Delta \le l} ||Z_{i\Delta}|| > b \,\Big|\, Z_0 = z\Big\} \\ &\quad = \sum_{i=0}^{l/\Delta} \int_{-\pi}^{\pi} \int_{0}^{\infty} P\Big\{||Z_{i\Delta}|| \in b + dy, \omega_{i\Delta} \in dw \,\Big|\, Z_0 = z\Big\} \\ &\qquad\qquad * P\Big\{||Z_{(i+j)\Delta}|| < b \;\forall j \in D_i \,\Big|\, ||Z_{i\Delta}|| = b + y, \omega_{i\Delta} = w, Z_0 = z\Big\} \end{aligned} \tag{15}$$

The fact that $Z_{1,t}$ and $Z_{2,t}$ are independent and normally distributed for fixed t yields

$$P\Big\{||Z_{i\Delta}|| \in b + dy, \omega_{i\Delta} \in dw \,\Big|\, Z_0 = z\Big\}$$

$$= \phi\left(\frac{(b+y)\cos w - x_1 R_1(i\Delta)}{[1 - R_1^2(i\Delta)]^{1/2}}\right) \phi\left(\frac{(b+y)\sin w - x_2 R_2(i\Delta)}{[1 - R_2^2(i\Delta)]^{1/2}}\right)$$

$$\frac{(b+y)\,dy\,dw}{[1-R_1^2(i\Delta)]^{1/2}[1-R_2^2(i\Delta)]^{1/2}}$$

If we expand the above around t^* and w^* where t^* and w^* are defined in the statement of the Proposition, we get

$$P\Big\{||Z_{i\Delta}|| \in b+dy, \omega_{i\Delta} \in dw \,\Big|\, Z_0 = z\Big\}$$

$$\approx \frac{be^{-\frac{1}{2}[b^2-||x||^2]}e^{-\frac{1}{2}b^2[(i\Delta-t^*)^2a_1+(w-w^*)^2a_2+2(w-w^*)(i\Delta-t^*)^2a_3]}e^{-by}dydw}{2\pi[1-R_1^2(i\Delta)]^{1/2}[1-R_2^2(i\Delta)]^{1/2}}, \tag{16}$$

where

$$a_1 = \frac{\dot{R}_1^2(t^*)\cos^2 w^*}{1-R_1^2(t^*)} + \frac{\dot{R}_2^2(t^*)\sin^2 w^*}{1-R_2^2(t^*)},$$

$$a_2 = \frac{\sin^2 w^*}{1-R_1^2(t^*)} + \frac{\cos^2 w^*}{1-R_2^2(t^*)} - 1,$$

$$a_3 = \cos w^* \sin w^* \left[\frac{\dot{R}_2^2(t^*)R_2^2(t^*)}{1-R_2^2(t^*)} + \frac{\dot{R}_1^2(t^*)R_1^2(t^*)}{1-R_1^2(t^*)}\right].$$

Using an argument similar to Siegmund (1985), p. 202, we see that

$$P\Big\{||Z_{(i+j)\Delta}|| < b \;\forall j \in D_i \,\Big|\, ||Z_{i\Delta}|| = b+y, \omega_{i\Delta} = w, Z_0 = z\Big\} \to P_{-\mu,\sigma}\Big\{\max_{j>0} S_j < -y\Big\}, \tag{17}$$

where S_j is the sum of j independent normal random variables with mean and variance $-\Delta b(\beta_1 \cos^2 w + \beta_2 \sin^2 w)$ and $2\Delta(\beta_1 \cos^2 w + \beta_2 \sin^2 w)$, respectively.

Substituting (16) and (17) into (15) we obtain

$$P\Big\{\max_t ||Z_t|| \geq b \,\Big|\, Z_0 = z\Big\}$$

$$\approx \sum_{i=0}^{l/\Delta} \int_{-\pi}^{\pi} \frac{be^{-\frac{1}{2}[b^2-||x||^2]}e^{-\frac{1}{2}b^2[(i\Delta-t^*)^2a_1+(w-w^*)^2a_2+2(w-w^*)(i\Delta-t^*)^2a_3]}}{2\pi[1-R_1^2(i\Delta)]^{1/2}[1-R_2^2(i\Delta)]^{1/2}}$$

$$* \int_0^\infty e^{-by} P_{\mu,\sigma}\Big\{\min_{j>0} S_j > y\Big\}\, dydw$$

$$\approx \frac{\beta \exp[-\frac{1}{2}(b^2 - ||x||^2)]}{\dot{R}_1(t^*)R_2(t^*)\cos^2 w^* + R_1(t^*)\dot{R}_2(t^*)\sin^2 w^*}\, \nu\Big(b\,[2\beta\Delta]^{\frac{1}{2}}\Big).$$

To obtain the last line we use Corollary 8.45 of Siegmund (1985) to evaluate the inner integral. The summation on i, is approximately an integral in the variable $i\Delta$; and since $b >> 0$ the bivariate normal density in $i\Delta$ and w behaves like a delta function concentrating at t^*, w^*.

4 Appendix.

For completeness we give in this appendix a simple derivation of (1) and (2) for a disease having a single trait predisposing locus. The derivation follows Risch (1990a,b).

Let K denote the probability that a random individual has the disease. Let φ be the indicator of an individual's phenotype, i.e., $\varphi = 1$ or 0 according as the individual is affected or not, so $K = E\varphi$. We consider only a monogenic disease and let $G = \{a, b\}$ denote an individual's genotype at the disease locus. Assuming that the population is random mating, so genotype frequencies are in Hardy-Weinberg equilibrium, we can by an analysis of variance decomposition write

$$E(\varphi|G) = K + f_a + f_b + d_{ab},$$

where $f_a(f_b)$ is the additive effect of allele $a(b)$ and $d_{a,b}$ is the interaction (dominance deviation). Hence $\Sigma f_a p_a = 0$ and $\Sigma_a d_{a,b} p_a = \Sigma_b d_{a,b} p_b = 0$, where p_a is the frequency of allele a in the population.

A basic assumption is that the phenotypes of two individuals are conditionally independent given their genotypes, i.e.,

$$E(\varphi_1\varphi_2|G_1, G_2) = E(\varphi_1|G_1)E(\varphi_2|G_2).$$

Then the probability that two relatives are both affected is

$$E(\varphi_1\varphi_2) = K^2 + (\frac{1}{2}V_A)e_{12} + V_D u_{12},$$

where $V_A = 2\Sigma p_a f_a^2$ is the additive variance of the penetrances, $V_D = \Sigma p_a p_b d_{ab}^2$ is the dominance variance of the penetrances, e_{12} is the expected number of alleles shared identical-by-descent by individuals 1, 2 and u_{12} is the probability that both alleles are shared identical-by-descent. For siblings $e_{12} = 1$ and $u_{12} = 1/4$. For the same calculation, *conditional* on the event that the siblings share 2, 1, or 0 alleles identical by descent, we have $e_{12} =$ 2, 1, or 0 and $u_{12} = 1$, 0 or 0, respectively.

Hence by Bayes' theorem the conditional probability that siblings inherit two alleles identical by descent, given that both are affected, is

$$\frac{1}{4}[K^2 + V_A + V_D]/[K^2 + V_A/2 + V_D/4].$$

Also the probability they inherit one or no allele identical by descent is respectively

$$\frac{1}{2}[K^2 + V_A/2]/[K^2 + V_A/2 + V_D/4].$$

and

$$\frac{1}{4}K^2/[K^2 + V_A/2 + V_D/4].$$

By simple algebra, we see that these probabilities can be rewritten in the form of display (1) with $\alpha = [V_A/2 + V_D/4]/[K^2 + V_A/2 + V_D/4]$ and $\delta = [V_D/4]/[K^2 + V_A/2 + V_D/4]$, which satisfy the constraints (2).

ACKNOWLEDGMENT

This research was partly supported by NSF Grant DMS-9704324 and by NIH Grant 5 R01 HG00898.

LITERATURE CITED

Chernoff, H. (1954). On the distribution of the likelihood ratio statistic, *Ann. Math. Statist.* **25** 573-578.

Dupuis J (1994) Statistical problems associated with mapping complex and quantitative traits from genomic mismatch scanning data, Ph. D. Thesis, Stanford University.

Dupuis J. and Siegmund, D. (1998). Statistical methods for mapping quantitative trait loci from a dense set of markers, submitted for publication.

Faraway, J. (1993). Improved sib pair linkage test for disease susceptibility loci, *Genet. Epidemiol.* **10**, 225-233.

Feingold, E. (1993) Markov processes for modeling and analyzing a new genetic mapping method, *J. Appl. Probab.* **30:** 766-779.

Feingold, E., P. O. Brown, and D. Siegmund (1993) Gaussian models for genetic linkage analysis using complete high resolution maps of identity-by-descent, *Am. J. Hum. Genet.* **53:** 234-251.

Ferguson, T. (1967). *Mathematical Statistics: A Decision Theoretic Approach*, Academic Press, New York and London.

Haseman, J.K. and Elston, R.C. (1972). The investigation of linkage between a quantitative trait and a marker locus, *Behavior Genetics* **2**, 3-19.

Lai, T. L. and D. Siegmund (1977). A non-linear renewal theory with applications to sequential analysis I,*Ann. Statist.* **5** (1977), 946–954.

Lander, E.S. and Botstein, D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps, *Genetics* **121**, 185-199.

Lander, E. S. and N. J. Schork (1994) Genetic Dissection of complex traits, *Science* **265:** 2037-2048.

Risch, N. (1990a,b,c) Linkage strategies for genetically complex traits I, II, III. The power of affected relative pairs, *Am. J. Hum. Genet.* **46:** 222-228, 229-241, 242-253.

Self, S. G. and K.-Y. Liang (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions, *Jour. Amer. Statist. Assoc.* **82**, 605-610.

Siegmund, D. (1985). *Sequential Analysis: Tests and Confidence Intervals.* Springer–Verlag, New York.

Siegmund, D. (1989). Confidence sets in change-point problems, *International Statistical Review* **56**, 31-48.

Teng, J. (1996). Statistical methods in linkage analysis, Stanford University Ph. D. thesis.

Tu, I.-P. and Siegmund, D. (1998). The maximum of a function of a Markov chain and application to linkage analysis, submitted for publication.

Woodroofe, M. and H. Takahashi (1982). Asymptotic expansions for the error probabilities of some repeated significance tests, *Ann. Statist.* **10**, 895-908.

Woodroofe, M. (1976). Frequentist properties of Bayesian sequential tests, *Biometrika* **63**, 101–110.