An asymptotic test for quantitative gene detection

Jean-Marc Azaïs [1] Christine Cierco-Ayrolles[1,2]

December 19, 2001

[1]Laboratoire de Statistique et Probabilités, U.M.R. C5583, Université Paul Sabatier, 118 Route de Narbonne, 31062 Toulouse Cedex 4, France.

[2]Institut National de la Recherche Agronomique, Unité de Biométrie et Intelligence Artificielle, B.P. 27, Chemin de Borde-Rouge, 31326 Castanet-Tolosan Cedex, France.

email: azais@cict.fr ; cierco@toulouse.inra.fr

Abstract

The problem of detecting the presence of a quantitative gene using a great number of markers in a backcross genetic scheme is addressed.

An asymptotic test based on the maximum of a differentiable stochastic process is constructed. Bounds for threshold and power calculation are presented. Simulations and numerical experiments illustrate the convergence towards the asymptotic distribution and the sharpness of the bounds.

Key words: Genetic markers, Haldane mapping function, Ornstein-Uhlenbeck process, QTL, Rice formulae

1 Introduction

This paper presents a statistical test applied to a genetic problem that leads to a non-standard situation.

Detecting genes involved in the variation of quantitative traits is a very important issue for animal and plant breeders. In this perspective, a mapping can be achieved by the use of genetic markers, which inform about the genotype of an individual for any position on any chromosome. The distribution of a quantitative trait among the different possible genotypes of a genetic marker gives the possibility to know whether there is some gene influencing the quantitative trait. Such a gene is called a QTL (Quantitative Trait Loci) by geneticists.

In this paper we suppose to know on on which chromosome is located the QTL. This chromosome will be modelled by the segment $[0, L]$. References and details on the way of building a distance on the chromosome will be given in Section 2. In general the statistics for testing the null hypothesis H_0 "there is no QTL" against the alternative $H_1^{d_0}$ "there is a QTL at location d_0 on the chromosome", depends on d_0. When testing now H_0 against H_1 "there is a QTL at some location d on the chromosome" we are led to make the position d vary on $[0, L]$: we get " the detection test process". As detailed further, this process depends on the information provided by genetic markers that permit in some sense to "read" the genoma at certain locations.

Classical methods study the limit of the detection test process as the number n of observations tends to infinity for a fixed number M of markers. Under certain conditions and after normalization, the limit is a χ^2 process and its distribution depends on the various locations of the markers.

On given species, the number of genetic markers becomes greater and greater, thus a new asymptotic framework is worth being studied where the number of genetic markers tends to infinity as well as the number of observations. This has been done by Feingold, Brown & Siegmund (1993), or Cierco (1998). In a particular theoretical framework, it is proved that the detection test process (more precisely, the square root of Lander and Botstein's process) converges in distribution to an Ornstein-Uhlenbeck process under H_0 and to the same Ornstein-Uhlenbeck process superimposed on a mean value function under H_1. By Ornstein-Uhlenbeck process we mean a centred stationary Gaussian process with a covariance function of the form $r(t) = \sigma^2 \exp(-\beta |t|)$, $\beta > 0$ ($\beta = 2$ in our case).

By analogy with Lander & Botstein's procedure, one way to build the test is to use the maximum of the absolute value of the detection test process. For a practical use, one of the major difficulty is computing thresholds and then powers in both approaches (M finite, M infinite). In the first case, these quantities are obtained by the Davies' method (Davies, 1977, 1987) that gives upper bounds. However, when the number of markers is large, these bounds are not sharp, and moreover, since both thresholds and powers depend on markers locations, the calculation must be adapted to every particular situation and it is impossible to construct tables.

In the second case, when the number of markers tends to infinity, the limit process is not smooth, thus the Davies' method cannot be applied. For some given chromosome lengths, the thresholds can be derived exactly from DeLong's results (DeLong, 1981). Unfortunately, this method cannot be applied to a non centred process, which corresponds to power calculation. Feingold, Brown & Siegmund, studying the Ornstein-Uhlenbeck limit process, gave approximations for threshold and power determinations that are not sharp.

Our contribution is to use a smoothing procedure on the detection test process. It yields the following improvements:

- first, we find a value of the smoothing parameter that does not diminish the power of the test and that improves the convergence towards the asymptotic distribution,
- second, for the same value of the smoothing parameter, we propose a new bound for

the distribution of the supremum of a stochastic process which is, in some sense, an improvement of Davies' method (see Section 3). In Section 4, this bound is shown to provide a rather good approximation of thresholds and powers.

The model is detailed in Section 2. Section 3 presents the method. Finally, in Section 4, a simulation study proves that:

- for a number of markers near to the one available in some species the asymptotic approximation is valid in the smoothed case,
- smoothing does not diminish the power,
- the proposed bounds for thresholds and powers are nearly exact.

2 Model, notations and assumptions

Here, we take up the model described by Cierco (1998). Consider two inbred lines, A and B, that are homozygous. From A and B, a backcross population of size n is derived using the crossing scheme $(A \times B) \times A$. Each individual of this population has two chromosomes, one purely inherited from the A line and the other, called the "recombined" one, with some part issued from the A line and other part from the B line, due to genetical recombinations during meiosis. We will use the Haldane distance (1919) which consists in modelling the recombination by a standard Poisson process. L will denote the chromosome length. Since the number of genetic markers M_n tends to infinity with the number n of data, their locations also depend on n. For given n, let $\left\{(d_{i,n})_{i=1,\ldots,M_n}\,; d_{1,n} < d_{2,n} < \ldots < d_{M_n,n}\right\}$ be the markers locations. The markers are assumed to make it possible to distinguish between the two situations:

- the individual is homozygous for the marker: the allele at location $d_{i,n}$ of the recombined chromosome is issued from the A line,
- the individual is heterozygous for the marker: the allele at location $d_{i,n}$ of the recombined chromosome is issued from the B line.

This is the case, for example, for codominant markers that allow to distinguish between the three cases: homozygous identical to the A line, heterozygous and homozygous identical to the B one (note that this last case is excluded by our crossing scheme).

Since we use Haldane's mapping function (i.e. intensity 1 Poisson crossing-over), the recombination probability between two loci at genetic distances d and d' from the origin of the chromosome is equal to the probability for a parameter $|d - d'|$ Poisson variable to be even:

$$R(d, d') = \frac{1 - \exp(-2\,|d - d'|)}{2}.$$

The general model, that will be precised further, assumes that there is a QTL at the location $d_0 \in [0, L]$ influencing the trait of interest. The phenotypic observation Y_k of the k^{th} individual is modelled by a classical analysis of variance model (Knapp, Bridges & Birkes (1990), Haley & Knott (1992)):

$$Y_k = \mu + X_k(d_0)\, a/2 + \varepsilon_k,$$

where:

- μ is the global mean,
- a is the gene effect,
- $X_k(d_0) = \begin{cases} 1 & \text{if the } k^{th} \text{ individual is homozygous for the QTL} \\ -1 & \text{if the } k^{th} \text{ individual is heterozygous for the QTL} \end{cases}$
- the error terms $(\varepsilon_k)_{1 \le k \le n}$ are independent and identically distributed with mean 0 and finite variance σ^2.

We assume the following local asymptotic framework (F):

- the number n of individuals tends to infinity,
- the number of genetic markers tends to infinity and $d_n := \inf_{1 \le i \le M_n} d_{i+1,n} - d_{i,n} \ge C n^{-\gamma}$ for some positive constants C and γ. This condition is satisfied for example if the markers are equally spaced and if M_n is comparable to any power of n.
- the QTL effect is small:

$$a = a_n = \delta n^{-1/2} \text{ for some parameter } \delta \in \mathbb{R},$$

The true observation of the problem is

$$\{\big(Y_k, X_k(d_{1,n}), ..., X_k(d_{M_n,n})\big), k = 1, ..., n\}.$$

So we test $\mathrm{H}_\delta = \big\{a_n = \delta n^{-1/2},\ \text{QTL located at some unknown } d_0 \in [0, L]\big\}$ against H_0. Some of these assumptions are also used by Feingold, Brown & Siegmund (1993) and Mangin, Goffinet & Rebaï (1994).

The test statistic is classically based on an estimator of a determined at any position d on the chromosome. If we first suppose that the location d of the QTL is known and that this location coincide with a marker position, then the a natural estimator for a is the difference between the empirical means of homozygous and heterozygous individuals; it is the analysis of variance estimator:

$$S_{n,1}(d) := \frac{\sum_{k=1}^{n} Y_k \mathbb{1}_{[X_k(d)=1]}}{\sum_{k=1}^{n} \mathbb{1}_{[X_k(d)=1]}} - \frac{\sum_{k=1}^{n} Y_k \mathbb{1}_{[X_k(d)=-1]}}{\sum_{k=1}^{n} \mathbb{1}_{[X_k(d)=-1]}},$$

where $\mathbf{1}_{[X_k(d)=\cdot]}$ is the indicator function of the event $\{X_k(d) = \cdot\}$.

However, this estimator is inconvenient because the denominators are random variables. By the Law of Large Numbers, they are close to $\frac{n}{2}$. Moreover simulation results have shown that this estimator could have some bias for small samples in the case when the Y_k's are not centred. Thus we consider:

$$S_{n,2}(d) := \frac{2}{n}\sum_{k=1}^{n}\left(Y_k - \overline{Y}_n\right)\mathbb{1}_{[X_k(d)=1]} - \frac{2}{n}\sum_{k=1}^{n}\left(Y_k - \overline{Y}_n\right)\mathbb{1}_{[X_k(d)=-1]}.$$

When the QTL is located at d which is not a marker position, $X_k(d)$ is not observable and we must use interpolation. Let d_n^- and d_n^+ be defined as follows:

$$d_n^- = \sup\left\{d_{i,n} : d_{i,n} < d; i = 1, \ldots, M_n\right\},$$

$$d_n^+ = \inf\left\{d_{i,n} : d_{i,n} > d; i = 1, \ldots, M_n\right\}.$$

An estimator for a is obtained using simply a linear interpolation between $S_{n,2}(d_n^-)$ and $S_{n,2}(d_n^+)$. For $d \leq d_{1,n}$ or $d \geq d_{M_n,n}$, $S_{n,2}$ is prolonged by a constant. The final process will be denoted $S_{n,2}(d)$ again. The process $(S_{n,2}(d))_{d\in[0,L]}$ is the detection test process mentioned in the introduction.

Remark: In dense map condition, this estimator has the same properties as the one obtained by least-squares interpolation between $S_{n,2}(d_n^-)$ and $S_{n,2}(d_n^+)$ found in Mangin, Goffinet & Rebaï (1994) or in Cierco (1998) and equivalent to the Gaussian maximum likelihood estimator of Lander & Botstein's (Rebaï, Goffinet & Mangin (1995)).

Under hypotheses (F), Cierco (1998) studied the normalized detection test process

$$X_n(d) := S_{n,2}(d)\left(\widehat{Var}(S_{n,2}(d))\right)^{-\frac{1}{2}},$$

where $\widehat{Var}$ is obtained from the estimator $\hat{\sigma}^2$ function of the residual sum of square. She proved that this process converges in distribution to a Gaussian process $(X(d))_{d\in[0,L]}$ with:

- covariance function $r(t) := cov\left(X(d), X(d+t)\right) = \exp(-2|t|)$,
- mean value function $m(d) = \dfrac{\delta}{2\sigma}\exp(-2|d_0 - d|)$.

3 Process smoothing

The classical approach would be to use the test statistic $T_n = \sup\limits_{d\in[0,L]} |X_n(d)|$ which corresponds to a likelihood ratio test in the case of Gaussian observations. This is inconvenient for two reasons.

1. The limit process has irregular sample paths, the distribution of its supremum is known when $\delta = 0$ and for some given lengths of the chromosome (DeLong, 1981) but is unknown when $\delta \neq 0$. So powers can only be attained by bounds that are not very sharp (Feingold, Brown & Siegmund, 1993) or Monte Carlo methods calculations that are long and imprecise.

2. It does not take into account that the presence of a QTL at d_0 modifies the expectation of the limit process in a neighbourhood of d_0.

For these two reasons, we have decided to smooth the detection test process $(X_n(d))_{d\in[0,L]}$. For calculations simplicity, we used a centred Gaussian kernel of varying variance ε^2 denoted φ_ε. Let $(X_n^\varepsilon(d))_{d\in[0,L]}$ be the smoothed process.

$$X_n^\varepsilon(d) = \int_{\mathbb{R}} X_n(u)\varphi_\varepsilon(d-u)du = (X_n * \varphi_\varepsilon)(d).$$

We considered the following test statistic $T_n^\varepsilon = \sup_{d\in[0,L]} |X_n^\varepsilon(d)|$. Property of weak convergence of processes (Billingsley, 1968) implies that the limit of $(X_n^\varepsilon(d))_{d\in[0,L]}$ is the smoothed version of the limit process, that is the process $(X^\varepsilon(d))_{d\in[0,L]}$ with

- mean value function $m^\varepsilon(d) = (m * \varphi_\varepsilon)(d) =$

$$\frac{\delta}{2\sigma}\left\{\exp\left[2(-d_0+\varepsilon^2+d)\right]\Phi\left(\frac{d_0-2\varepsilon^2-d}{\varepsilon}\right) + \exp\left[2(d_0+\varepsilon^2-d)\right]\overline{\Phi}\left(\frac{d_0+2\varepsilon^2-d}{\varepsilon}\right)\right\}$$

- covariance function r^ε given by

$$r^\varepsilon(d) = \int\int_{\mathbb{R}^2} \varphi_\varepsilon(-u)\varphi_\varepsilon(d-v)r(u-v)dudv = \int_{\mathbb{R}} \varphi_{\sqrt{2}\varepsilon}(d-v)\, r(v)\, dv$$

$$= \exp(2(2\varepsilon^2-d))\Phi\left(\frac{d-4\varepsilon^2}{(2\varepsilon^2)^{\frac{1}{2}}}\right) + \exp(2(2\varepsilon^2+d))\overline{\Phi}\left(\frac{d+4\varepsilon^2}{(2\varepsilon^2)^{\frac{1}{2}}}\right),$$

where Φ is the standard gaussian distribution function, $\Phi(t) = \int_{-\infty}^{t} \varphi_1(u)du$, and $\overline{\Phi} = 1-\Phi$.

Threshold and power calculation

Bounds are described for a generic process that will be denoted $(Y(d))_{d\in[0,L]}$. In practice, this process is the limit process $(X^\varepsilon(d))_{d\in[0,L]}$. Note that since we work on asymptotic distribution, our results are free from the markers locations.

So we consider a Gaussian process $(Y(d))_{d\in[0,L]}$ with $\mathcal{C}^1$ sample paths and we assume that for every $t_1, t_2; s_1, s_2 \in [0, L]$, the distribution of $Y(t_1), Y(t_2); Y'(s_1), Y'(s_2)$ is non

degenerate, (Y' is the derivative) and we denote by $p_{t_1,t_2;s_1,s_2}$ their joint density. In our particular case, this condition is met because the spectrum of $(Y(d) - \mathbf{E}(Y(d)))_{d\in[0,L]}$ has a continuous component (Cramér & Leadbetter, 1967).

For threshold or power calculations, we are interested in the distribution function of the random variable $|Y|^\star = \sup_{d\in[0,L]} |Y(d)|$. We use the following event equality which is a particular case of the general method described by Azaïs & Wschebor (1997, 1999, 2000, 2001) and Azaïs, Cierco-Ayrolles & Croquette (1999):

$$\forall u \geq 0, \quad \mathbb{P}\left(\{|Y|^\star > u\}\right) = \mathbb{P}\left(\{|Y(0)| > u\} \cup \{|Y(0)| \leq u\,;\, (U_u + D_{-u}) \geq 1\}\right), \qquad (1)$$

where ";" denotes the intersection, U_u and D_{-u} are respectively the number of upcrossings of u and of downcrossings of $-u$ by the process Y on the interval $[0, L]$, defined as

$$U_u = \#\left\{d \in [0, L]\,; Y(d) = u; Y'(d) > 0\right\}$$

$$D_{-u} = \#\left\{d \in [0, L]\,; Y(d) = -u; Y'(d) < 0\right\}.$$

Our method is based on the double inequality below. If ξ is a random variable with non-negative integer values, then:

$$\mathbf{E}\left(\xi\right) - \frac{1}{2}\left[\mathbf{E}\left(\xi(\xi-1)\right)\right] \;\leq\; \mathbb{P}\left(\xi \geq 1\right) \;\leq\; \mathbf{E}\left(\xi\right). \qquad (2)$$

Applying (2) with $\xi = (U_u + D_{-u})\,\mathbf{1}_{|Y(0)|\leq u}$, we get

$$\mathbf{E}\left((U_u + D_{-u})\mathbf{1}_{|Y(0)|\leq u}\right) - \frac{\mathbf{E}[(U_u + D_{-u})(U_u + D_{-u} - 1)]}{2} \leq \mathbb{P}\left(|Y|^* > u\right) \leq \mathbf{E}\left((U_u + D_{-u})\,\mathbf{1}_{|Y(0)|\leq u}\right) \quad (3)$$

Remark that $\mathbf{E}\left[(U_u + D_{-u})\,(U_u + D_{-u} - 1)\right] = \mathbf{E}\left(U_u(U_u - 1)\right) + \mathbf{E}\left(D_{-u}(D_{-u} - 1)\right) + 2\mathbf{E}\left(U_u D_{-u}\right)$.

Expectations involved in the above inequality can be evaluated by Rice's formulae (Rice, 1944 and 1945, Cramér and Leadbetter, 1967, Marcus, 1977, Wschebor, 1985):

- $\mathbf{E}\left(U_u \mathbf{1}_{|Y(0)|\leq u}\right) = I_1(u) = \displaystyle\int_0^L dt \int_{-u}^{u} dx \int_0^{+\infty} y p_{0,t;t}(x, u; y) dy,$
- $\mathbf{E}\left(D_{-u} \mathbf{1}_{|Y(0)|\leq u}\right) = I_2(u) = \displaystyle\int_0^L dt \int_{-u}^{u} dx \int_{-\infty}^{0} |y| p_{0,t;t}(x, -u; y) dy,$
- $\mathbf{E}\left(U_u(U_u - 1)\right) = I_3(u) = \displaystyle\int_0^L \int_0^L dt_1 dt_2 \int_0^{+\infty} \int_0^{+\infty} y_1 y_2 p_{t_1,t_2;t_1,t_2}(u, u; y_1, y_2) dy_1 dy_2,$
- $\mathbf{E}\left(D_{-u}(D_{-u} - 1)\right) = I_4(u) = \displaystyle\int_0^L \int_0^L dt_1 dt_2 \int_{-\infty}^{0} \int_{-\infty}^{0} |y_1 y_2|\, p_{t_1,t_2;t_1,t_2}(-u, -u; y_1, y_2) dy_1 dy_2,$

- $\mathbb{E}\left(U_u D_{-u}\right) = I_5(u) = \displaystyle\int_0^L \int_0^L dt_1 dt_2 \int_0^{+\infty} \int_{-\infty}^0 |y_1 y_2|\, p_{t_1,t_2;t_1,t_2}(u,-u;y_1,y_2) dy_1 dy_2,$

where $p_{0,t;t}$ is the joint density function of $(Y(0), Y(t), Y'(t))$. Expressions of these integrals more adapted to numerical computation may be found in Azaïs, Cierco-Ayrolles & Croquette (1999). Hence we obtain the fundamental inequality:

$$\mathbb{P}\left(|Y(0)| > u\right) + I_1(u) + I_2(u) - \frac{I_3(u) + I_4(u) + 2I_5(u)}{2} \leq \mathbb{P}\left(|Y|^\star > u\right) \leq \mathbb{P}\left(|Y(0)| > u\right) + I_1(u) + I_2(u). \quad (4)$$

Remarks:

- Relation (4) is a refinement of Davies' method (Davies, 1977). Davies worked with the random variable $Y^\star = \sup_{d\in[0,L]} Y(d)$ and, instead of (3), he used the relation:

$$\mathbb{P}\left(Y^\star > u\right) \leq \mathbb{P}\left(Y(0) > u\right) + \mathbb{P}\left(U_u \geq 1\right) \leq \mathbb{P}\left(Y(0) > u\right) + \mathbb{E}\left(U_u\right).$$

  Besides the fact that we work with $|Y|^\star$ instead of $Y^\star$, the upper bound is very similar to Davies' one, except for the small improvement due to the event $\{|Y(0)| \leq u\}$.

- By simulation, it has been verified that for centred process and rather large values of u, the lower bound is more accurate than the upper one. This is because of the use of the second order factorial moment. So, in the following, for threshold calculations, we will use the lower bound.

- This inequality cannot be applied directly to the original limit process for it has non differentiable sample paths.

- Under the null hypothesis, the process $(Y(d))_{d\in[0,L]}$ is centred, Y and $-Y$ have the same distribution and relation (4) has a simpler form:

$$2\left[\mathbb{P}(Y(0) > u) + I_1(u)\right] - I_3(u) - I_5(u) \leq \mathbb{P}\left(|Y|^\star > u\right) \leq 2\left[\mathbb{P}(Y(0) > u) + I_1(u)\right]. \quad (5)$$

4 Simulation study

This section presents the results of a Monte-Carlo experiment to evaluate the quality of the proposed method under a variety of conditions. Our aim was to study

1. the relationship between the value of the smoothing parameter and the validity of the asymptotic approximation for reasonable numbers of markers and individuals,

2. the sharpness of the bounds given by the "fundamental inequality" for various values of the smoothing parameter.

Length of the chromosome = 0.98 Morgan

	nominal level of the test								
	10%			5%			1%		
5% confidence interval for the emp. level	9.41-10.59			4.57-5.43			0.80-1.19		
threshold $\varepsilon = 0$	2.74			3.01			3.55		
threshold $\varepsilon^2 = 10^{-2}$	2.019			2.276			2.785		
threshold $\varepsilon^2 = 10^{-3}$	2.321			2.593			3.128		
marker density	1 cM	2 cM	7 cM	1 cM	2 cM	7 cM	1 cM	2 cM	7 cM
empirical level for $\varepsilon = 0$	7.37	6.67	4.99	3.91	3.42	2.4	0.77	0.67	0.43
empirical level for$\varepsilon^2 = 10^{-2}$	12.17	12.17	11.82	6.75	6.69	6.53	1.76	1.72	1.77
empirical level for$\varepsilon^2 = 10^{-3}$	10.84	10.66	9.71	5.63	5.55	5.02	1.34	1.32	1.04

Table 1

Threshold and empirical level(in %) of test using the unsmoothed detection test process ($\varepsilon = 0$) $(X_n(d))_{d\in[0,L]}$ and the smoothed detection process $(X_n^{\varepsilon}(d))_{d\in[0,L]}$. The chromosome length is equal to 0.98 M, and the number of individuals is equal to 500. The second line of the table gives a confidence interval for the empirical proportion related to the nominal level over 10^4 simulations.

Tables 1 displays empirical levels for smoothed and unsmoothed procedures with thresholds calculated under the asymptotic distribution.

- For the unsmoothed process, the threshold is calculated using Table II of DeLong's (1981). For this reason, the chromosome length, 0.98 M (Morgan) has been chosen to correspond to an entry of DeLong's table, and to be to length encountered for several vegetal species.
- For the smoothed process, we used the lower bound in the "fundamental inequality" and the S^+ program developed by Cierco-Ayrolles, Croquette and Delmas (2000) following formulae given in Section 3.

Simulations have been performed for two values of the smoothing parameter and three markers densities. The number of individuals have been chosen equal to 500. The crossing-overs were simulated according to a Poisson process with intensity 1. We performed 10000 simulations, so that the 5% confidence interval for the empirical levels associated to the theoretical ones are indicated.

Table 2 presents the power associated to the detection test in the case of a gene of size $\delta = 6$ located at the position $d_0 = 0.4$. The length of the chromosome is 1 M, calculations are made under the asymptotic distribution, using a test with nominal level equal to 5%.

- For the unsmoothed detection test process, the threshold is calculated via DeLong's table and the power by the only possible method which is a Monte-Carlo method. 10^4 simulations have been used.

- For the smoothed process, the threshold is calculated as above using the lower bound and the power is calculated by three manners: using the upper bound in the "fundamental inequality", using the lower bound in the "fundamental inequality", by a Monte-Carlo method. A full detail of the terms involved in the "fundamental inequality" is given.

	$\varepsilon^2 = 10^{-2}$	$\varepsilon^2 = 10^{-3}$	unsmoothed process
5% level	2.281	2.599	3.02
$\mathbb{P}(\lvert Y(0)\rvert > u)$	15.70	9.81	-
$\mathbf{E}\left((U_u + D_{-u})\mathbb{1}_{\lvert Y(0)\rvert \leq u}\right)$	55.57	72.30	-
$\frac{\nu_2}{2}$	1.43	13.06	-
$\frac{\mathbf{E}(U_u(U_u-1))}{2}$	1.43	13.06	-
$\frac{\mathbf{E}(D_{-u}(D_{-u}-1))}{2}$	$7.84\ 10^{-6}$	$3.40\ 10^{-4}$	-
$\mathbf{E}(U_u D_{-u})$	$2.54\ 10^{-5}$	$4.10\ 10^{-3}$	-
lower bound	69.84	69.05	-
upper bound	71.27	82.11	-
empirical power	71.37 ± 0.88	72.53 ± 0.87	68.99 ± 0.91

Table 2

Power in % associated to the detection test in the case of a gene of size $\delta = 6$, located at a distance $d_0 = 0.4$ from the origin of a chromosome of length 1 M. The value of σ is equal to 1. In the smoothed procedure, the 5% level, the upper and lower bounds are calculated using the S^+ program previously mentioned. For the unsmoothed process, the 5% level is given by DeLong's table. The empirical powers are calculated over 10^4 simulations and the corresponding 95 % confidence intervals are given.

Discussion

Tables 1 clearly indicate that the unsmoothed procedure is very conservative. We have checked by Monte-Carlo that this is not due to a typo in DeLong's table.

The empirical level given by the smoothed procedure is close to the nominal value. For $\varepsilon^2 = 10^{-3}$, it is nearly inside the confidence interval.

The comparison of powers given by the unsmoothed and smoothed procedures on conditions of Tables 1 would not be fair since the unsmoothed procedure is very conservative, which implies low power. Table 2 shows clearly that smoothing at size $\varepsilon^2 = 10^{-2}, 10^{-3}$ does not diminish power on the asymptotic distribution.

It is also clear in Table 2 that at the size $\varepsilon^2 = 10^{-2}, 10^{-3}$, the lower bound is almost exact.

5 Conclusion

In conclusion, the procedure we advocate is the use of the asymptotic test after smoothing with $\varepsilon^2 = 10^{-3}$ and with thresholds and powers calculated using the lower bound in the "fundamental inequality". The corresponding thresholds are given in Table 4 for a level α of 1%, 5 % and 10% and for certain chromosome lengths. For the other cases and for power calculations, the S^+ program developed by Cierco-Ayrolles, Croquette & Delmas (2000) can be used.

	chromosome length in Morgans					
	0.75	1	1.5	2	2.5	3
1% level	3.059	3.133	3.239	3.315	3.375	3.423
5 % level	2.516	2.599	2.721	2.809	2.878	2.934
10% level	2.239	2.328	2.458	2.553	2.626	2.687

Table 3

Thresholds calculated using the lower bound of the "fundamental inequality" for different values of level α and various chromosome lengths. The smoothing parameter is fixed to 10^{-3}.

The validity of this procedure has been justified, in conditions near to practice, by the following statements that have been illustrated by a numerical Monte-Carlo experiment:

1. smoothing improves the convergence to the asymptotic distribution,
2. for a reasonable number of markers and individuals, the asymptotic behaviour is met,
3. when working with the asymptotic distribution, smoothing at size $\varepsilon^2 = 10^{-2}, 10^{-3}$ does not diminish power on the asymptotic distribution,
4. at the size $\varepsilon^2 = 10^{-2}, 10^{-3}$, the lower bound is almost exact.

ACKNOWLEDGMENT

We thank C. Delmas for computational assistance.

6 References

Azaïs, J.-M. and Wschebor M. (1997). Une formule pour calculer la distribution du maximum d'un processus stochastique. *C.R. Acad. Sci. Paris,* t. 324, série I, 225-230.

Azaïs, J.-M. and Wschebor M. (1999). Régularité de la loi du maximum de processus gaussiens réguliers. *C.R. Acad. Sci. Paris,* t. 328, série I, 333-336.

Azaïs, J.-M. and Wschebor M. (2000). Computing the Distribution of the Maximum of a Gaussian Process. *submitted.*

Azaïs, J.-M. and Wschebor M. (2001). On the Regularity of the Distribution of the Maximum of one-parameter Gaussian Processes. *Probability Theory and Related Fields*119,70-98.

Azaïs, J.-M., Cierco-Ayrolles, C. and Croquette, A. (1999). Bounds and asymptotic expansions for the distribution of the Maximum of a smooth stationary Gaussian process. *ESAIM: Probability and Statistics*, **3**, 107-129.

Billingsley, P. (1968). *Convergence of Probability Measures.* Wiley, New York.

Cierco, C. (1996). *Problèmes statistiques liés à la détection et à la localisation d'un gène à effet quantitatif.* PHD dissertation. University of Toulouse, France.

Cierco, C. (1998). Asymptotic Distribution of the Maximum Likelihood Ratio Test for Gene Detection. *Statistics* **31**, **3**, 261-285.

Cierco-Ayrolles, C., Croquette, A. and Delmas, C. (2000). Computing the Distribution of the Maximum of Regular Gaussian Processes. *Preprint.*

Cramér, H. et Leadbetter, M.R. (1967). *Stationary and Related Stochastic Processes.* Wiley, New York.

Davies, R.B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* **64**, 247-254.

Davies, R.B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* **74**, 33-43.

DeLong, D.M. (1981). Crossing Probabilities for a Square Root Boundary by a Bessel Process. *Communications in Statistics-Theory and Methods* **A10**, 2197-2213.

Feingold, E., Brown, P.O. et 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.

Haldane, J.B.S. (1919). The Combination of Linkage Values and the Calculation of Distances Between the Loci of Linked Factors. *Journal of Genetics* **8**, 299-309.

Haley, C. et Knott, S.A. (1992). A Simple Regression Method for Mapping Quantitative Trait Loci by using Molecular Markers. *Heredity* **69**, 315-324.

Knapp, S.J., Bridges, W.C. and Birkes, D. (1990). Mapping Quantitative Trait Lociusing Molecular Marker Linkage Maps. *Theoretical and Applied Genetics*, **79**, 583-592.

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

Leadbetter, M.R., Lindgren, G. et Rootzén, H. (1983). *Extremes and Related Properties of Random Sequences and Processes.* Springer-Verlag, New York.

Mangin, B., Goffinet, B. and Rebaï, A. (1994). Constructing Confidence Intervals for QTL Location. *Genetics*, **138**, 1301-1308.

Marcus, M.B. (1977). Level Crossings of a Stochastic Process with Absolutely Continuous Sample Paths. *The Annals of Probability* **5**, 52-71.

Rebaï, A., Goffinet, B. and Mangin, B. (1995). Comparing Power of Different Methods for QTL Detection. *Biometrics*, **51**, 87-99.

Rice, S.O. (1944). The Mathematical Analysis of Random Noise. *Bell Syst. Tech. Journal* **23**, 282-332.

Rice, S.O. (1945). The Mathematical Analysis of Random Noise. *Bell Syst. Tech. Journal* **24**, 46-156.

SPLUS, 1993. Statistical Sciences, *S-Plus Programmer's Manual, Version 3.2*, Seattle: StatSci, a division of MathSoft, Inc., 1993.

Wschebor, M. (1985). Surfaces Aléatoires - Mesure Géométrique des Ensembles de Niveau. *Lecture Notes in Mathematics* **1147**. Springer-Verlag, New-York.