ESTIMATION OF LINKAGE DISEQUILIBRIUM IN RANDOMLY MATING POPULATIONS*

B. S. WEIR and C. CLARK COCKERHAM
Department of Statistics, North Carolina State University, Raleigh, North Carolina 27650

Received 2.viii.78

SUMMARY

The maximum likelihood method for estimating linkage disequilibrium from genotypic data for randomly mating populations is studied. Instead of iterative methods for finding a root of the cubic equation for one of the gametic frequencies (Hill, 1974), it is recommended that the cubic be solved completely. For data with some missing genotypic classes, it is further recommended that explicit solutions for the cubic be used.

1. INTRODUCTION

THE degree of linkage disequilibrium in a randomly mating population can be estimated from genotypic frequencies in a sample of individuals taken from the population. It is appropriate to use maximum likelihood (ML) estimation, and a comprehensive review of the methodology was given by Hill (1974). In the case of two codominant loci and where coupling and repulsion double heterozygotes cannot be distinguished, Hill provides a cubic equation for the ML estimate of one of the gametic frequencies. He suggests that a solution to this equation be found by numerical iteration. This note presents two comments on the iterative technique.

In the first place, it is probably better to solve the cubic completely and examine the likelihoods for all valid roots found. This will prevent any problems of non-convergence, or of convergence to a valid root that does not maximise the likelihood. Secondly, it is often the case that samples of moderate sizes have some of the nine genotypic classes missing. It is then often possible to provide analytic solutions to the cubic and further reduce the dependence on numerical algorithms.

2. NOTATION

The notation of Hill (1974) is retained. The first locus has codominant alleles A, a with frequencies p, $1-p$, while the second locus has codominant alleles B, b with frequencies q, $1-q$. Gametic types AB, Ab, aB, ab have frequencies f_{11}, f_{12}, f_{21}, f_{22}, respectively, and the usual measure of linkage disequilibrium is

$$D = f_{11} - pq.$$

The notation for observed genotypic numbers as well as the genotypic

* Paper No. 5628 of the Journal Series of the North Carolina Agricultural Experiment Station, Raleigh, North Carolina. This investigation was supported in part by NIH Research Grant No. GM 11546 from the National Institute of General Medical Sciences.

TABLE 1

Observed numbers and expected frequencies

	Observed numbers			
	BB	*Bb*	*bb*	Total
AA	N_{11}	N_{12}	N_{13}	$N_{1.}$
Aa	N_{21}	N_{22}	N_{23}	$N_{2.}$
aa	N_{31}	N_{32}	N_{33}	$N_{3.}$
Total	$N_{.1}$	$N_{.2}$	$N_{.3}$	N

	Expected frequencies			
	BB	*Bb*	*bb*	Total
AA	f_{11}^2	$2f_{11}f_{12}$	f_{12}^2	p^2
Aa	$2f_{11}f_{21}$	$2f_{11}f_{22}+2f_{12}f_{21}$	$2f_{11}f_{12}$	$2p(1-p)$
aa	f_{21}^2	$2f_{21}f_{22}$	f_{22}^2	$(1-p)^2$
Total	q^2	$2q(1-q)$	$(1-q)^2$	1

$$X_{11} = 2N_{11}+N_{12}+N_{21} \qquad X_{12} = 2N_{13}+N_{12}+N_{23}$$
$$X_{21} = 2N_{31}+N_{21}+N_{32} \qquad X_{22} = 2N_{33}+N_{23}+N_{32}$$

frequencies expected under random mating are displayed in table 1, and some summary measures X_{ij} are also defined there.

It can be helpful to clarify the nature of linkage disequilibrium by partitioning it into components for within individuals, D_w, and between individuals, D_b (Cockerham and Weir, 1977). If g_{11} denotes the frequency with which genes A, B are found on different gametes within the same individual, then

$$D_w = f_{11} - g_{11}, \quad D_b = g_{11} - pq.$$

In an obvious notation,

$$f_{11} = f(AABB) + \tfrac{1}{2}f(AABb) + \tfrac{1}{2}f(AaBB) + \tfrac{1}{2}f(AB/ab)$$
$$g_{11} = f(AABB) + \tfrac{1}{2}f(AABb) + \tfrac{1}{2}f(AaBB) + \tfrac{1}{2}f(Ab/aB).$$

The within-individuals component is also equal to half the difference in frequencies of coupling (AB/ab) and repulsion (Ab/aB) double heterozygotes, while the between-individuals component is a measure of the non-randomness of gametic union. If $D_b = 0$, as assumed in this note, then $D_w = D$. Without this assumption, the separate estimation of the two components from genotypic data requires that the two types of double heterozygotes can be distinguished. When the assumption is true, the expected frequencies in table 1 are appropriate and the frequency f_{11} can be estimated by maximising the likelihood L where (Hill, 1974)

$$\log L = C + \sum_{i,j} X_{ij} \log f_{ij} + N_{22} \log (f_{11}f_{22}+f_{12}f_{21}),$$

and C is a constant.

3. Cubic Likelihood Equation

Using a " chromosome counting " method, Hill (1974) provided the following cubic equation for the ML estimate $\hat{f}_{11}$ of f_{11}:

$$\hat{f}_{11} = \{X_{11}+N_{22}\hat{f}_{11}(1-\hat{p}-\hat{q}+\hat{f}_{11})/[\hat{f}_{11}(1-\hat{p}-\hat{q}+\hat{f}_{11})+(\hat{p}-\hat{f}_{11})(\hat{q}-\hat{f}_{11})]\}/2N \tag{1}$$

where

$$\hat{p} = (X_{11}+X_{12}+N_{22})/2N = (N_{1.}+\tfrac{1}{2}N_{2.})/N$$
$$\hat{q} = (X_{11}+X_{21}+N_{22})/2N = (N_{.1}+\tfrac{1}{2}N_{.2})/N$$

are the usual ML estimates of gene frequencies. Hill suggests that an initial value

$$\tilde{f}_{11} = (X_{11}-X_{12}-X_{21}+X_{22})/4N+\tfrac{1}{2}-(1-\hat{p})\ (1-\hat{q})$$
$$= (2X_{11}+N_{22})/2N-\hat{p}\hat{q}$$

be substituted into the right-hand side of (1) and the resulting expression regarded as an improved estimate and itself substituted into the right-hand side of (1). This iteration procedure is continued until stability is reached. The final value $\hat{f}_{11}$ is the ML estimate, and the ML estimate of D is

$$\hat{D} = \hat{f}_{11}-\hat{p}\hat{q}.$$

The disequilibrium $\tilde{D}$ corresponding to Hill's initial value $\tilde{f}_{11}$ is, apart from a factor of $N/(N-1)$, the estimate suggested by Burrows (Cockerham and Weir, 1977) for the composite measure $\Delta = D_w+2D_b$. Burrows' estimator is $\hat{\Delta} = N\tilde{D}/(N-1)$. When $D_b = 0$ and there is random union of gametes, Burrows' estimator is unbiased for $\Delta = D$, so that Hill's initial value should be close to the ML estimate of D. If $D_b \neq 0$, it cannot be said what is being estimated by the solution to (1).

As numerical methods are already being employed for the iterative procedure of Hill, it seems preferable to use a numerical algorithm to solve the cubic equation (1) and determine three roots $f_{11}^*(i)$, $i = 1, 2, 3$. Algorithms to find the roots of polynomials are readily available in computer subroutine libraries. The roots will be termed valid if they are real and satisfy

$$\max\ (0, \hat{p}+\hat{q}-1) \leqq f_{11}^*\ (i) \leqq \min\ (\hat{p}, \hat{q}). \qquad (2)$$

The ML estimate of f_{11} is that valid root that maximises L. Note that the initial value $\tilde{f}_{11}$ can be invalid but such samples are unlikely. Unlike the iterative method, the procedure of solving the cubic and examining all roots cannot have problems of non-convergence or of convergence to the wrong root.

One way in which non-convergence could result is when the quadratic in the denominator of the left-hand side of equation (1) is zero. Suppose F_{11} is a gametic frequency that would cause such a discontinuity in the iterative procedure. We can show that for F_{11} to be real, it is necessary that

$$(2\hat{p}-1)^2+(2\hat{q}-1)^2 \geqq 1$$

but that F_{11} does not lie in the range of validity (2). Problems with such discontinuities are likely to arise then only if the initial value for the iteration procedure is not chosen carefully. The situation is illustrated by the frequencies shown in table 2. These numbers provide two coincident discontinuities and the iterative procedure leads to the ML value f_{11}^* (3) when $\tilde{f}_{11}$ is used. An initial value of $f_{11} = 0{\cdot}4$ leads immediately to the root f_{11}^* (2) and the iterations stop, as they would if f_{11}^* (1) or F_{11} were ever reached. Because of the discontinuity, it is often found that intermediate iterates are outside the range of validity (2).

It will be noticed that the marginal one-locus frequencies in table 2 are not significantly different from Hardy-Weinberg proportions. Problems with multiple valid roots, and convergence to the wrong root, seem to arise when the marginals depart from these random mating frequencies. In

TABLE 2

Frequencies to illustrate discontinuities

	BB	*Bb*	*bb*
AA	154	81	8
Aa	37	14	3
aa	1	1	1

$f_{11}^{*}(1) = 0{\cdot}5239$ $f_{11}^{*}(2) = 0{\cdot}6750$ $f_{11}^{*}(3) = 0{\cdot}7277$ $\tilde{f}_{11} = 0{\cdot}7233$

$\hat{p} = 0{\cdot}9$ $\hat{q} = 0{\cdot}8$ $F_{11} = 0{\cdot}6$

Range of validity $0{\cdot}7 \leqq f_{11} \leqq 0{\cdot}8$

such situations, of course, there is evidence that $D_b \neq 0$ and methods based on Hill's likelihood should not be used. Blind application of the iterative method however can lead to surprising results. The two sets of frequencies in table 3 both provide three valid roots. For table 3a, the initial value $\tilde{f}_{11} = \hat{q}/2$ is also a root f_{11}^{*} (2) of the cubic and will be presented as the solution by the iterative technique. This root is not a stable equilibrium for the iterative method, however, and any initial value other than $\tilde{f}_{11}$ will give convergence to one of the two stable roots $(\hat{q} \pm \sqrt{\hat{q}^2 - X_{11}/N})/2$. These two roots have equal likelihoods that exceed the likelihood of f_{11}^{*} (2). A different situation is provided by the frequencies in table 3b. The iterative scheme leads to the ML root f_{11}^{*} (3) provided the initial value is greater than f_{11}^{*} (2). Initial values less than f_{11}^{*} (2) lead to the root f_{11}^{*} (1).

The log-likelihoods log $L(i)$ are also shown in table 3 for each of the valid solutions $f^{*}(i)$ and it is clear that there are very small differences between the likelihoods for these alternative solutions. Small differences in likelihoods can exist even when the differences between corresponding estimates of

TABLE 3

Frequencies to illustrate multiple valid roots

(a)

	BB	*Bb*	*bb*
AA	12	3	3
Aa	3	54	3
aa	12	3	3

$f_{11}^{*}(1) = 0{\cdot}1968$ $\tilde{f}_{11} = f_{11}^{*}(2) = 0{\cdot}2969$ $f_{11}^{*}(3) = 0{\cdot}3969$

$\log L(1) = C - 224{\cdot}0017$ $\log L(2) = C - 225{\cdot}3435$ $\log L(3) = C - 224{\cdot}0017$

$\hat{p} = 0{\cdot}5$ $\hat{q} = 0{\cdot}5938$

Range of validity $0 \leqq f_{11} \leqq 0{\cdot}5$

(b)

	BB	*Bb*	*bb*
AA	12	3	3
Aa	3	51	3
aa	12	6	3

$f_{11}^{*}(1) = 0{\cdot}2202$ $f_{11}^{*}(2) = 0{\cdot}2726$ $f_{11}^{*}(3) = 0{\cdot}3745$ $\tilde{f}_{11} = 0{\cdot}2905$

$\log L(1) = C - 227{\cdot}2005$ $\log L(2) = C - 227{\cdot}3635$ $\log L(3) = C - 226{\cdot}3820$

$\hat{p} = 0{\cdot}4844$ $\hat{q} = 0{\cdot}5938$

Range of validity $0 \leqq f_{11} \leqq 0{\cdot}4844$

linkage disequilibrium are quite large. In table 3b, for example, f_{11}^{*} (3) provides an estimate of $\hat{D} = 0{\cdot}0869$ that is significantly greater than zero, while f_{11}^{*} (1) provides an estimate $\hat{D} = -0{\cdot}0674$ that is significantly less than zero.

4. Special cases

With samples of moderate size, such as 100 individuals, several genotypic classes may be missing and in many cases the cubic equation (1) can then be solved analytically. In 104 data sets for which D could be estimated (*i.e.* both loci polymorphic), Laurie-Ahlberg and Weir (1979) found 10 cases in which N_{22} was zero, 20 cases in which one X_{ij} was zero and eight cases in which two of the X_{ij} were zero. These data sets had between 91 and 134 *Drosophila melanogaster* individuals and represented pairs from 10 loci in samples from nine laboratory populations. In almost every case, the zero classes arose because at least one of the two loci had an allele with a frequency of greater than 0·85. With such frequencies, one homozygote at that locus has an expected count of only 2·25 in a sample of size 100 and will often be absent. Missing classes can even arise with less extreme frequencies. In one case, the one-locus numbers were $N_{1.} = 53$, $N_{2.} = 43$, $N_{3.} = 3$, $N_{.1} = 44$, $N_{.2} = 47$, $N_{.3} = 8$, yet X_{22} was zero. The data set did exhibit general agreement with Hardy-Weinberg frequencies at all loci.

Each of the three situations of zero classes will now be considered.

(i) $N_{22} = 0$

Clearly gametic frequencies can be estimated directly when the sample contains no double heterozygotes. Either by inspection, or from (1),

$$\hat{f}_{11} = X_{11}/2N,\ \hat{D} = X_{11}/2N - \hat{p}\hat{q}.$$

When gametic frequencies are available, Cockerham and Weir (1977) showed how D_w and D_b may be estimated and tests for the hypotheses $H{:}D_w = 0$, $H{:}D_b = 0$ were established. In the present case, $\hat{D}_w = 0$ and $\hat{D}_b = \hat{D}$ [apart from a bias correction term $N/(N-1)$] so that the assumption $D_b = 0$ may be tested.

(ii) $X_{ij} = 0$

When one of the summary measures X_{ij} is zero, the cubic (1) has one root which may be found analytically. The other two roots follow from a quadratic. These roots and quadratics are as follows:

$$X_{11} = 0{:}f_{11}^{*} = 0, \qquad f_{11}^{*2} + (W-Y)f_{11}^{*} + \left[Z + \frac{N_{22}}{4N}(\hat{p}+\hat{q}-1)\right] = 0$$

$$X_{12} = 0{:}f_{11}^{*} = \hat{p}, \qquad f_{11}^{*2} + (W+Y)f_{11}^{*} + \left(Z - \frac{N_{22}}{4N}\hat{q}\right) = 0$$

$$X_{21} = 0{:}f_{11}^{*} = \hat{q}, \qquad f_{11}^{*2} + (W+Y)f_{11}^{*} + \left(Z - \frac{N_{22}}{4N}\hat{p}\right) = 0$$

$$X_{22} = 0{:}f_{11}^{*} = \hat{p}+\hat{q}-1, f_{11}^{*2} + (W-Y)f_{11}^{*} + Z = 0$$

where

$$W = (1-2\hat{p}-2\hat{q})/2, \quad Y = N_{22}/4N, \quad Z = \hat{p}\hat{q}/2.$$

There is no way of predicting, in general, if the quadratics will provide valid roots, or if they do, which of the valid roots is the ML estimate of f_{11}. The likelihoods must still be evaluated for each valid root.

(iii) $X_{ij}, X_{kl} = 0$

One root and a quadratic also arise in the following two cases:

$$X_{11} = X_{22} = 0\colon\ f_{11}^* = 0,\ f_{11}^{*2} - \tfrac{1}{2}\left(1 + \frac{N_{22}}{2N}\right) f_{11}^* + \tfrac{1}{2}\hat{p}\hat{q} = 0$$

$$X_{12} = X_{21} = 0\colon\ f_{11}^* = \hat{p},\ f_{11}^{*2} + \tfrac{1}{2}\left(1 - 4\hat{p} + \frac{N_{22}}{2N}\right) f_{11}^* + \tfrac{1}{2}\hat{p}\left(\hat{p} - \frac{N_{22}}{2N}\right) = 0,$$

but further consideration of these situations shows that disequilibrium must be extreme. For $X_{11} = X_{22} = 0$, $\hat{p} = 1 - \hat{q}$ and $\hat{f}_{11} = 0$, $\hat{D} = -\hat{p}(1 - \hat{p})$ while for $X_{12} = X_{21} = 0$, $\hat{p} = \hat{q}$ and $\hat{f}_{11} = \hat{p}$, $\hat{D} = \hat{p}(1 - \hat{p})$. Otherwise, two zero X_{ij} values provide three analytic roots:

$$X_{11} = X_{12} = 0\colon\ f_{11}^* = 0, \hat{p}, \hat{q} - \tfrac{1}{2}(1 - \hat{p})$$

$$X_{11} = X_{21} = 0\colon\ f_{11}^* = 0, \hat{q}, \hat{p} - \tfrac{1}{2}(1 - \hat{q})$$

$$X_{12} = X_{22} = 0\colon\ f_{11}^* = \hat{p}, \hat{p} + \hat{q} - 1, \tfrac{1}{2}\hat{q}$$

$$X_{21} = X_{22} = 0\colon\ f_{11}^* = \hat{q}, \hat{p} + \hat{q} - 1, \tfrac{1}{2}\hat{p}$$

5. Discussion

Maximum likelihood estimation of linkage disequilibrium from genotypic data is currently of interest to population geneticists, and efficient estimation techniques are necessary. For very large samples, clearly from Hardy-Weinberg populations, the iterative technique suggested by Hill (1974) will probably be adequate, but it seems preferable to remove any doubts by solving the cubic and examining all valid roots. The computing required to solve a cubic is comparable to that for the iterations and is not dependent on features such as the choice of an initial value. The computing may even be reduced in cases when one gene has a high frequency and several genotypic classes are missing.

It must be stressed, however, that ML estimation of linkage disequilibrium from data in which the two classes of double heterozygotes are not distinguished rests on the assumption of random mating. If there is evidence of non-random mating, such as departures from Hardy-Weinberg frequencies at either locus, the method should not be used. Instead, composite measures such as $\triangle$ can be estimated. We would even recommend that experimenters consider estimating $\triangle$ as a routine procedure. If there is random mating, then $\triangle$ is unbiased for the usual measure, D, of linkage disequilibrium, and it can be found directly from the data without need for numerical computations. If there is not random mating, then $\triangle$ is the only measure that can be estimated. It need hardly be added that the likelihood established from table 1 also assumes the absence of disturbing forces such as selection.

Acknowledgment.—Dr W. G. Hill made several valuable comments on a draft of this paper.

6. References

COCKERHAM, C. CLARK, AND WEIR, B. S. 1977. Digenic descent measures for finite populations. *Genetical Research*, *30*, 121-147.

HILL, W. G. 1974. Estimation of linkage disequilibrium in randomly mating populations. *Heredity*, *33*, 229-239.

LAURIE-AHLBERG, C. C., AND WEIR, B. S. 1979. Allozymic variation and linkage disequilibrium in some laboratory populations of *Drosophila melanogaster*. *Genetics* (submitted).