Heredity **77** (1996) 396–399 *Received 21 November 1995*

Small sample properties for estimators of cytonuclear disequilibria

ROB DEAN & JONATHAN ARNOLD*
Department of Genetics, University of Georgia, Athens, GA 30602

Unbiased and maximum likelihood estimators for the cytonuclear disequilibrium measures D_1, D_2, and D_3 are given, and their exact variances are determined with the use of indicator variables. Conditions under which the exact variances should be used include when the genotypic disequilibria or the cytoplasmic allele frequency are extreme. These unbiased estimators are shown to have high efficiency by comparison to the Cramer–Rao lower bound on variances of unbiased estimators of the disequilibria. The maximum likelihood estimators are recommended on the basis of small sample properties.

Keywords: cytonuclear disequilibria, hybrid zone, linkage disequilibria.

Introduction

The estimation of linkage disequilibria between genes and genotypes is a central problem in population genetics (Hill, 1974; Weir & Cockerham, 1979; Weir, 1990). For example, disequilibria between cytoplasmic markers, as in mitochondrial DNA, and nuclear markers have come to play a central role in recent analyses of hybrid zones (Arnold, 1993). To date, studies of estimators of cytonuclear disequilibria have focused on their large-sample properties (Asmussen & Basten, 1994). Here, for the first time, we examine the small-sample properties of estimators of cytonuclear disequilibria. These results are particularly important to studies with nuclear molecular markers and cytoplasmic markers because sample sizes are generally quite restricted.

Model and analysis

Calculations of the expectations of the various nuclear and cytoplasmic allele frequencies are standard, and being based on the same principles as the calculations of the cytonuclear disequilibrium measures, are not shown here.

The nuclear genotypes are represented by AA, Aa and aa. The cytoplasmic genotypes are represented by M and m. The frequencies of the AA, Aa and aa genotypes are u, v and w, respectively. The frequency of the cytotype M is x. Table 1 shows the probability model.

*Correspondence.

Departures from a no association hypothesis are measured by the cytonuclear disequilibria, D_1, D_2 and D_3. If a random sample of size N is taken according to the probabilities in Table 1, then the model specification is multinomial. The likelihood is shown in eqn 1:

$$L = \left(\frac{N!}{N_1!N_2!N_3!N_4!N_5!N_6!}\right) u_1^{N_1} u_2^{N_4} v_1^{N_2} v_2^{N_5} w_1^{N_3} w_2^{N_6}, \quad (1)$$

where $N_1, \ldots, N_6$ are the individual cell counts in Table 1. There are five independent parameters: u, v, x, D_1 and D_2. The cytonuclear disequilibrium measures to be estimated from the likelihood in eqn 1 are D_1, D_2 and D_3.

To calculate the moments of a disequilibrium estimator we will focus here on the measure D_1, as the calculations are similar for all three disequilibrium estimators. The only assumption we will make is that a random sample of individuals $i = 1, \ldots, N$ is obtained to estimate D_1, D_2 and D_3.

The indicator variables that we will use are defined as follows:

Let $X_i^{AA/M} = 1$, if individual i has AA/M cytonuclear genotype;
$= 0$, otherwise.
Let $X_j^{AA} = 1$, if individual j has AA nuclear genotype;
$= 0$, otherwise
Let $X_k^{M} = 1$, if individual k has M cytotype;
$= 0$, otherwise.
From Table 1 we see that

$$D_1 = u_1 - ux. \quad (2)$$

Table 1 Genotypic disequilibria

	AA	*Aa*	*aa*	Total
M	$u_1 = ux + D_1$	$v_1 = vx + D_2$	$w_1 = wx + D_3$	x
m	$u_2 = u(1-x) - D_1$	$v_2 = v(1-x) - D_2$	$w_2 = w(1-x) - D_3$	y
Total	u	v	w	1

The maximum likelihood estimator (MLE) of D_1 (Asmussen *et al.*, 1987) can be written in terms of these indicator variables under the hypothesis $D_1 \neq 0$ and $D_2 \neq 0$:

$$\hat{D}_1 = \frac{1}{N}\sum_{i=1}^{N} X_i^{AA/M} - \frac{1}{N^2}\left(\sum_{i=1}^{N} X_i^{AA}\right)\left(\sum_{j=1}^{N} X_j^{M}\right), \text{ or}$$

$$\hat{D}_1 = \frac{1}{N}\sum_{i=1}^{N} X_i^{AA/M} - \frac{1}{N^2}\left(\sum_{i=1}^{N}\sum_{j=1}^{N} X_i^{AA} X_j^{M}\right). \quad (3)$$

Substituting the expectations from Table 2 into the expectation of eqn 3 we obtain:

$$\mathrm{E}(\hat{D}_1) = \frac{1}{N} N u_1 - \frac{1}{N^2}(N u_1 + N(N-1)ux),$$

which reduces to:

$$\mathrm{E}(\hat{D}_1) = \frac{N-1}{N} D_1. \quad (4)$$

The same approach can be used to calculate the variance,

$$\mathrm{VAR}(\hat{D}_1) = \mathrm{E}(\hat{D}_1^2) - \mathrm{E}^2(\hat{D}_1). \quad (5)$$

From the above calculation for $\mathrm{E}(\hat{D}_1)$ we have:

$$\mathrm{E}^2(\hat{D}_1) = \left(\frac{N-1}{N}(u_1 - ux)\right)^2. \quad (6)$$

Now we need $\mathrm{E}(\hat{D}_1^2)$ to complete the variance calculation. Expanding the square of the estimator in terms of the indicator variables yields:

$$\hat{D}_1^2 = \frac{1}{N^2}\left(\sum_{i=1}^{N} X_i^{AA/M}\right)^2 - 2\frac{1}{N}\frac{1}{N^2}\sum_{i=1}^{N} X_i^{AA/M}\sum_{j=1}^{N} X_j^{AA}\sum_{k=1}^{N} X_k^{M} + \frac{1}{N^4}\sum_{j=1}^{N} X_j^{AA}\sum_{k=1}^{N} X_k^{M}\sum_{j=1}^{N} X_j^{AA}\sum_{k=1}^{N} X_k^{M}. \quad (7)$$

For the three terms in eqn 7 we need several expectations from Table 2.

When we put the corresponding expressions in Table 2 into eqn 7 and simplify, we obtain the variance for the maximum likelihood estimator $\hat{D}_1$:

$$\begin{aligned}\mathrm{VAR}(\hat{D}_1) = {} & 1/N(-D_1^2 + D_1(1-2u)(1-2x) \\ & + u(1-u)x(1-x)) \\ & + 1/N^2(3D_1^2 - 2D_1(1-2u)(1-2x) \\ & - u(1-u)x(1-x)) \\ & + 1/N^3(-2D_1^2 + D_1(1-2u)(1-2x)), \end{aligned} \quad (8)$$

where $D_1 = u_1 - ux$.

The first term is the Cramer–Rao Lower Bound (CRLB), as shown by inverting the information matrix for u, v, x, D_1 and D_2 calculated from the likelihood in eqn 1.

Table 2 Expectations of indicator variables and their products for calculating $\mathrm{E}(\hat{D}_1)$ and $\mathrm{VAR}(\hat{D}_1)$

Equation	Term	Product	Expectation
3	1	$X_i^{AA/M}$	u_1
	2	$X_i^{AA} X_j^{M}$	$u_1, i = j$
			$ux, i \neq j$
7	1	$X_i^{AA/M} X_j^{AA/M}$	$u_1, i = j$
			$u_1^2, i \neq j$
	2	$X_i^{AA/M} X_j^{AA} X_k^{M}$	$u_1, i = j = k$
			$u_1 x, i = j \neq k$
			$u_1 u, i = k \neq j$
			$u_1^2, j = k \neq i$
			$u_1 ux, i \neq j \neq k$
	3	$X_i^{AA} X_j^{M} X_k^{AA} X_l^{M}$	$u_1, i = j = k = l$
			$u_1 x, i = j = k \neq l$
			$u_1 u, i = j = l \neq k$
			$u_1 x, i = k = l \neq j$
			$u_1 u, j = k = l \neq i$
			$u_1^2, i = j \neq k = l$
			$ux, i = k \neq j = l$
			$u_1^2, i = l \neq j = k$
			$u_1 ux, i = j \neq k \neq l$
			$ux^2, i = k \neq j \neq l$
			$u_1 ux, i = l \neq j \neq k$
			$u_1 ux, j = k \neq i \neq l$
			$u^2 x, j = l \neq i \neq k$
			$u_1 ux, k = l \neq i \neq j$
			$u^2 x^2, i \neq j \neq k \neq l$

There are two cases for which the exact variances in eqn 8 are to be preferred over the asymptotic variance provided by the CRLB even with a relatively large sample size. In one case D_1 (or D_2 or D_3) is large in magnitude (at least $+0.19$) and the $1/N^2$ term can be substantial. This is the most frequent case. For example, when the sample size is 100 and the cytonuclear frequencies for AA/M, Aa/M,...,aa/m are 0.25, 0.29, 0.01, 0.0, 0.01 and 0.44, respectively, D_3 is -0.2375 and the ratio of CRLB/VAR(D_3) is 0.876. In the second case, where the cytoplasmic allele frequency (x) is either below 0.10 or above 0.90, the exact variance is again preferable (Table 3). Degradation of the large sample approximation of the CRLB in the second case was associated with extreme values in both x and the relevant nuclear genotypic frequency (e.g. u).

We now introduce a closely related unbiased estimator, which can be thought of as the maximum likelihood estimator (MLE) with bias correction from eqn 4,

$$\hat{D}_{1,u} = \frac{N}{N-1}\hat{D}_1. \tag{9}$$

The exact variance of this unbiased estimator can be obtained by correcting for the bias and can be written as:

$$\mathrm{VAR}(\hat{D}_{1,u}) = \left(\frac{N}{N-1}\right)^2 \mathrm{VAR}(\hat{D}_1). \tag{10}$$

It is worth noting that although the bias correction reduces bias, the trade-off is a slightly larger variance relative to the maximum likelihood estimator.

Expectations and variances for $\hat{D}_{2,u}$ and $\hat{D}_{3,u}$ are the same as those for $\hat{D}_1$, with D_2 (D_3) replacing D_1 and v (w) replacing u.

To determine the efficiency of the unbiased estimators, we compared the variance of the unbiased estimator with the asymptotic variance of the MLE ($\hat{D}_1$) as provided by the CRLB. These results are summarized in Table 4. The wide ranges of efficiencies are primarily caused by the two cases previously mentioned in which an exact variance would be preferred.

It is natural also to ask how the unbiased estimator performs relative to the maximum likelihood estimator in small samples on the basis of variance. The efficiency in a variance sense can be computed from eqs 8 and 10 and is quite high even in a sample as small as size $N = 10$. A related question is how both estimators perform when bias and variance are both taken into consideration. We answer that question by examining the mean square error of each estimator (MSE($\hat{\theta}$)), which has these two components:

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{VAR}(\hat{\theta}) + (\mathrm{E}(\hat{\theta}) - \theta)^2.$$

To examine the MSE of the estimators in small samples we divided the MSE of the biased MLE by the MSE of the bias-corrected MLE to yield an efficiency. In Table 4 these results for varying sample sizes are summarized. To insure that $\hat{D}_2$ yielded the same results the above calculations were carried out using D_2 in place of D_1, and the results were exactly the same as we found for $\hat{D}_1$. In conclusion, even though the maximum likelihood estimators are more efficient than the unbiased estimators in small samples, the unbiased estimators are still highly efficient.

Table 3 Minimum of ratio of asymptotic variance(CRLB) to the exact variance from eqn 8

x	$N = 25$	$N = 50$	$N = 75$	$N = 100$
0.01 or 0.99	0.65	0.78	0.84	0.88
0.02 or 0.98	0.69	0.81	0.86	0.89
0.03 or 0.97	0.83	0.88		
0.04 or 0.96	0.85	0.89		
0.05 or 0.95	0.76	0.86		

These ratios were obtained by a grid search in increments of 0.01 for u_1, u, v_1, v and x. Ratios greater than 0.90 are not reported.

Example

The cytonuclear genotypic counts in Table 5 are from a classic hybrid zone study (Harrison & Arnold, 1982) between the crickets *Gryllus pennsylvanicus* and *G. firmus*. Alleles diagnostic for *G.*

Table 4 Range of efficiencies of the unbiased estimator $\hat{D}_{1,u}$

N	$\frac{\mathrm{CRLB}}{\mathrm{VAR}(\hat{D}_{1,u})}$	$\frac{\mathrm{VAR}(\hat{D}_1)}{\mathrm{VAR}(\hat{D}_{1,u})}$	$\frac{\mathrm{MSE}(\hat{D}_1)}{\mathrm{MSE}(\hat{D}_{1,u})}$
10	0.27–0.97	0.81	0.81–0.94
25	0.49–0.99	0.92	0.92–0.99
50	0.66–0.99	0.96	0.96–1.00
100	0.80–1.00	0.98	0.98–1.00
500	0.95–1.00	1.00	1.00–1.00
1000	0.99–1.00	1.00	1.00–1.00

We obtained the table by setting up a grid in increments of 0.01 for u_1, u, v_1, v and x at a set sample size N.

Table 5 Counts of cytonuclear genotypes

	FF	FP	PP	Total
P	1	6	7	14
F	11	0	0	11
Total	12	6	7	25

From R. G. Harrison *et al.* (unpublished data).

pennsylvanicus are denoted by a 'P', and alleles diagnostic for *G. firmus* by an 'F'. An exact test of the no cytonuclear disequilibrium hypothesis is significant at the 0.001 level.
From Table 1, $\hat{D}_1 = -0.2288$.
From eqn 8 we see that variance $(\mathrm{VAR}(\hat{D}_1)) = 0.0005527$.
The standard error (SE) = 0.0235
Replacing D_1 and u in eqn 8 with D_2 and v we see that:

$\hat{D}_2 = 0.1056$
$\mathrm{SE}(\hat{D}_2) = 0.0330$.

Replacing D_1 and u in eqn 8 with D_3 and w we see that:

$\hat{D}_3 = 0.1232$
$\mathrm{SE}(\hat{D}_3) = 0.0336$.

To examine the accuracy of these estimates we compared them to the CRLB, which gave a standard error of 0.0202 for D_1, 0.0330 for D_2, and 0.0335 for D_3. This translates into a ratio of $\mathrm{VAR}(\hat{D}_i)/\mathrm{CRLB}$ of 74.2 per cent, 99.9 per cent and 99.6 per cent, respectively, for $i = 1, 2, 3$. This would indicate that for a relatively small sample size (as characteristic of molecular studies) these exact variances will be useful.

Discussion

As an analytical tool efficient estimates of cytonuclear disequilibrium measures have broad utility. These measures can be useful when examining: (i) the directionality of crosses between conspecifics; (ii) levels of gene flow; (iii) degrees of assortative mating between conspecifics; (iv) age of reproductive barriers; and (v) mechanisms of selection (Arnold, 1993).

By comparing two classes of cytonuclear disequilibrium estimators it is demonstrated that using unbiased estimators of cytonuclear disequilibria can be highly efficient relative to the maximum likelihood estimators. Exact variances for the maximum likelihoood and bias-corrected maximum likelihood estimators are given. For instance, in eqn 8 the variance of $\hat{D}_1$ is split into three terms with increasing powers in sample size. The first term is the asymptotic variance of the maximum likelihood estimator (i.e. the CRLB). The differences between the two types of estimators can be seen in the N^2 and N^3 terms.

For sample sizes ($N > 100$) the asymptotic variances are within 20 per cent of the exact variances and are adequate (Table 4). When the genotypic disequilibria are large in magnitude or the cytoplasmic allele frequency is extreme (below 0.05 or above 0.95), then the exact variances should be used.

For smaller sample sizes ($N > 50$) the variance for the unbiased estimator is within 4 per cent of the biased maximum likelihood estimator (Table 4). There is a slight advantage with respect to the mean squared error, usually in small samples, to using the maximum likelihood estimators. This characteristic and also the fact that the estimators and their variances are in a relatively simple closed form makes these estimators attractive and easy to use, without resorting to the use of a computer.

Acknowledgements

We thank Maria Sanchez for her assistance in checking the mathematics and statistics. We are also deeply appreciative to the two reviewers whose comments substantially improved this manuscript.

References

ARNOLD, J. 1993. Cytonuclear disequilibria in hybrid zones. *Ann. Rev. Ecol. Syst.*, **24**, 521–554.

ARNOLD, J., ASMUSSEN, M. A. AND AVISE, J. C. 1988. An epistatic mating system model can produce permanent cytonuclear disequilibria in a hybrid zone. *Proc. Natl. Acad. Sci. U.S.A.*, **85**, 1893–1896.

ASMUSSEN, M. A. AND BASTEN, C. J. 1994. Sampling theory for cytonuclear disequilibria. *Genetics*, **138**, 1351–1363.

ASMUSSEN, M. A., ARNOLD, J. AND AVISE, J. C. 1987. Definition and properties of disequilibrium statistics for associations between nuclear and cytonuclear genotypes. *Genetics*, **115**, 755–768.

HARRISON, R. G. AND ARNOLD, J. 1982. A narrow hybrid zone between closely related cricket species. *Evolution*, **36**, 535–552.

HILL, W. G. 1974. Estimation of linkage disequilibrium in randomly mating populations. *Heredity*, **32**, 229–239.

RAO, C. R. 1973. *Linear Statistical Inference and Its Applications*. John Wiley, New York.

WEIR, B. S. 1990. *Genetic Data Analysis*. Sinauer, Sunderland, MA.

WEIR, B. S. AND COCKERHAM, C. C. 1979. Estimation of linkage disequilibrium in randomly mating populations. *Heredity*, **42**, 105–111.