A Note on the Estimation of Individual Admixture

by

R. Dennis Cook and Sanford Weisberg

Technical Report No. 207

April 1973

Department of Applied Statistics
University of Minnesota, St. Paul, Minnesota 55101

Introduction

In this note we consider the problem of estimating the proportion (individual admixture) of genes contributed by one of two distinct ancestral populations to an individual who is a member of a distinct hybrid population. The hybrid population is assumed to have been formed by migrants drawn randomly from the ancestral populations over a long period of time. For example, one might wish to estimate the proportion of Caucasian genes for a given member of the American Black or American Indian population.

The problem of estimating the proportion of genes in the entire hybrid population (group admixture) has been considered (see, for example, Pollitzer [1972] and Reed [1969]). Elston [1971] considered least-squares and maximum likelihood (ML) estimates of group admixture. It appears, however, that little has been done directly concerning the problem of obtaining estimates of individual admixture.

Here we discuss a method of estimating individual admixture based on maximum likelihood. The method requires that for each of a number, L, of loci (1) the allelic forms are known, (2) each possible genotype can be classified according to phenotype, at least two of which must be distinguishable, (3) the frequency of each allele in both ancestral populations is known, and (4) no mutation or selection occurs in the populations. The method will be outlined, first under the assumption that no parental information is available and later extended to take parental information into account.

Estimation Using Progeny Data

We first consider the case in which all genotypes at the L loci to be sampled are distinguishable. Let n_ℓ represent the number of alleles at locus ℓ, $\ell = 1,2,\ldots,L$. The $n_\ell(n_\ell+1)/2$ possible genotypes will be represented as $g(i,j|\ell)$ where $(i,j) \in E_{n_\ell} = \{(i,j)\ ,\ i \leq j \leq n_\ell\}$ represents the observed allelic pair at the locus. We assume that the L loci are statistically independent, and that the Hardy-Weinberg proportions hold in the ancestral populations.

Let

$P(\ell,i|k)$ = frequency of allele i at locus ℓ in population k (k = 1,2)

M_m = proportion of alleles in the male parent from population 1 (male admixture)

M_f = female admixture

$M = (M_m + M_f)/2$ = progeny admixture

$P(\ell,i|k)$ can also be interpreted as the conditional probability that, at locus ℓ, a parent contributes allele i given that the contributed allele is attributable to population k. The $P(\ell,i|k)$ are assumed to be known.

Using the preceeding and the rules of conditional probability, we have

$$\Pr(g(i,i|\ell)) = [M_m P(\ell,i|1) + (1-M_m)P(\ell,i|2)][M_f P(\ell,i|1) + (1-M_f)P(\ell,i|2)] \quad (1)$$

and

$$\begin{aligned}\Pr(g(i,j|\ell)) = {} & [M_m P(\ell,i|1) + (1-M_m)P(\ell,i|2)][M_f P(\ell,j|1) + (1-M_f)P(\ell,j|2)] \\ & + [M_m P(\ell,j|1) + (1-M_m)P(\ell,j|2)][M_f P(\ell,i|1) + (1-M_f)P(\ell,i|2)]\end{aligned} \quad (2)$$

where in (2), $i \neq j$.

The assumption of statistically independent loci in the parents implies that the loci in the progeny are statistically independent. Therefore, the likelihood function of the data (phenotypes) is simply

$$\mathcal{L}(M_m, M_f) = \prod_{\ell=1}^{L} \Pr(g(i,j|\ell)) \qquad (3)$$

where $(i,j) \in E_{n_\ell}$ is the observed phenotype at locus ℓ and n_ℓ is the number of alleles at this locus.

We now wish to maximize (3) with respect to M_m and M_f. From (1), (2) and (3), for any observed phenotypes in the offspring, the likelihood (3) will be identical if M_m and M_f are interchanged. Hence, unless additional information on one or both parents is also available, we will have $\hat{M}_f = \hat{M}_m = \hat{M}$, say. This follows from our inability to tell which allele came from which parent. Substituting $\hat{M}$ for $\hat{M}_f$ and $\hat{M}_m$, the likelihood becomes proportional to

$$\mathcal{L}(M) \propto \prod_{\ell=1}^{L} [MP(\ell,i|1) + (1-M)P(\ell,i|2)][MP(\ell,j|1) + (1-M)P(\ell,j|2)]$$

with $(i,j) \in E_{n_\ell}$. Taking logarithms and differentiating, the ML estimate will be the solution to

$$\sum_{\ell=1}^{L} \left[\frac{P(\ell,i|1)-P(\ell,i|2)}{\hat{M}[P(\ell,i|1)-P(\ell,i|2)] + P(\ell,i|2)} + \frac{P(\ell,j|1)-P(\ell,j|2)}{\hat{M}[P(\ell,j|1)-P(\ell,j|2)] + P(\ell,j|2)} \right] = 0. \qquad (4)$$

This solution is a special case of that obtained by Elston [1971] for the similar problem of estimating group admixture given the average gene frequencies in a hybrid population. Approximate variances and a discussion of the solution when genotypes are indistinguishable at multi-allelic loci are given by Elston and will not be repeated here. It is noteworthy that, whenever $P(\ell,i|1) = P(\ell,i|2)$, the locus in question furnishes no information (i.e. the ancestral populations are indistinguishable at locus ℓ); therefore, those loci at which the ancestral allelic frequencies are the same can be ignored.

Some insight into the above estimation procedure can be gained by considering the extreme case in which, for each of L bi-allelic loci, we have $|P(\ell,i|1) - P(\ell,i|2)| = 1$; that is, each allele can be classified with certainty as to its

ancestral population. In this case, as expected,

$$\hat{M} = (2L_1 + L_2)/2L$$

and

$$\hat{Var}(\hat{M}) = (2L_1 + L_2)(L_2 + 2L_3)/8L^3 = \hat{M}(1-\hat{M})/2L \qquad (5)$$

where L_1 (L_3) is the number of observed loci that are homozygous for an allele in ancestral population 1 (2), L is the total number of loci sampled, and $L_2 = L - L_1 - L_3$. Thus, as one might expect, the variance of $\hat{M}$ decreases at a rate that is proportional to the inverse of the number of loci sampled. If an estimate, M_g, of group admixture is available then the above expression can be used, by setting $\hat{M} = M_g$, to determine the number of loci required to achieve a given variance. For example, if $M_g = 0.1$ then about 18 perfect loci are required to achieve a standard deviation of 0.05.

In Table 1, we give the number of these perfectly discriminatory loci necessary to achieve a standard error of .05 and .1 for M = .1,.2,.3,.4,.5. If s = .05 is desired (giving a bound of $\pm$ 2s $\simeq \pm$ .1) it is clear that a prohibitively large number of loci will be required. (Of course, loci are not perfectly discriminatory, so that the number of loci required in practice will be considerably larger than the values given in Table 1.) We are led to seek other information to reduce variability.

Table 1. Number of perfect bi-allelic loci needed to achieve a specified standard error (s) of $\hat{M}$.

	M_g				
s	0.1	0.2	0.3	0.4	0.5
.05	18	32	42	48	50
.10	4	8	10	12	12

Estimation Using Parental Data

The maximum likelihood estimate, $\hat{M}$, of individual admixture has been considered in the context of estimating the admixture in the progeny. However, either parent may be considered as a progeny. Thus, this procedure can be used to estimate the female admixture, say, by assuming the admixtures in the grandparents are the same. In this way an estimate, $\hat{M}_f$, of M_f can be obtained and used in expressions (1), (2) and (3). By regarding M_f as known, the likelihood becomes a function of M_m only. The progeny data can now be used to estimate M_m and an estimate of the progeny admixture obtained as $\hat{M} = (\hat{M}_f + \hat{M}_m)/2$. Alternatively, both M_m and M_f can be estimated using data on the respective parents and the progeny admixture can again be estimated using the average of the parental estimates.

As an example, suppose we observe L perfect bi-allelic loci (all phenotypes distinguishable) for both parents, and estimate $\hat{M}_f$ and $\hat{M}_m$. It then follows that the gene configuration of the offspring <u>given</u> the gene configurations in the parents will be independent of M_m and M_f (this will not be so if the loci are not perfect) so that

$$\hat{M} = (\hat{M}_f + \hat{M}_m)/2$$

and

$$\hat{Var}(\hat{M}) = [\hat{M}_f(1-\hat{M}_f) + \hat{M}_m(1-\hat{M}_m)]/8L \ .$$

In the least favorable case of $M_f = M_m$, the variance of $\hat{M}$ will be reduced by 50% when compared to (5); if M_f and M_m are different, the reduction in variance will be larger.

If only one parent, say the female, is observed and M_f is estimated, again assuming L perfect bi-allelic loci, then the ML estimate, $\hat{M}_m$, of M_m from the

offspring data reduces to

$$\hat{M}_m = (L_1^* + L_2^*)/(L - L_3^*)$$

where

L_1^* = number of homozygous loci in the offspring from population 1

L_2^* = number of heterozygous loci in offspring given that the locus in the female parent is homozygous from population 2

L_3^* = number of loci that are heterozygous in the parent and progeny.

We find

$$\hat{Var}(\hat{M}_m) = \hat{M}_m(1-\hat{M}_m)/(L-L_3^*)$$

so that, setting $\hat{M} = (\hat{M}_m + \hat{M}_f)/2$,

$$\hat{Var}(\hat{M}) = \tfrac{1}{4}\left\{\frac{\hat{M}_m(1-\hat{M}_m)}{L-L_3^*} + \frac{\hat{M}_f(1-\hat{M}_f)}{2L}\right\} .$$

Assuming that $M_m = M_f$ and $L_3^* = 0$, the reduction in variance if one parent is measured as $\frac{1}{8}$; differences in M_m and M_f will decrease the variance, while increases in L_3^* will increase the variance.

In the more realistic situation (not all phenotypes distinguishable or non-perfect loci), the estimate of $\hat{M}_m$ given $\hat{M}_f$ will depend on $\hat{M}_f$ and the variance of the estimate will be larger than those given here. Generally, using (1) and (2), the procedure discussed by Elston can be amended to furnish the ML estimate of M_m in this case.

The essential point in this section is that more information about admixture in the offspring (that is, smaller variance) can be obtained by measuring both parents than by measuring the offspring; if only one parent and the offspring are measured, the resulting variance will be intermediary between the two alternatives.

Comments

The general ML estimation procedure for estimating group admixture discussed by Elston can also be used to obtain ML estimates of individual admixture when the only information available is on the individual in question. When parental information is available the general likelihood equation given by Elston can be easily amended using (1) and (2) to furnish ML estimates of individual admixture.

In general, because of the large number of loci required (see Table 1) to achieve a reasonable degree of precision, it is doubtful that useful estimates of individual admixture can be obtained without parental information. The most desirable situation is to have estimates of individual admixture on both parents and then estimate the admixture of the person in question by averaging the parental estimates. The variance of the estimate in this case is roughly 50% of that using only data on the individual. The variance can be reduced by roughly 12% if information on one parent is available.

References

Elston, R.C. (1971). The estimation of admixture in racial hybrids. Annals of Human Genetics 35, 9-17.

Pollitzer, W.S. (1972). Problems in admixture estimates from different genetic loci. Haematologia 6, 193-198.

Reed, T.E. (1969). Caucasian genes in American Negroes. Science 165, 762-768.