A Class of Population Genetic Questions Formulated as the Generalized Occupancy Problem

Ranajit Chakraborty

Center for Demographic and Population Genetics, University of Texas Graduate School of Biomedical Sciences, Houston, Texas 77225

Manuscript received November 12, 1992
Accepted for publication March 12, 1993

ABSTRACT

In categorical genetic data analysis when the sampling units are classified into an arbitrary number of distinct classes, sometimes the sample size may not be large enough to apply large sample approximations for hypothesis testing purposes. Exact sampling distributions of several statistics are derived here, using combinatorial approaches parallel to the classical occupancy problem to help overcome this difficulty. Since the multinomial probabilities can be unequal, this situation is described as a *generalized occupancy problem*. The sampling properties derived are used to examine nonrandomness of occurrence of mutagen-induced mutations across loci, to devise tests of Hardy-Weinberg proportions of genotype frequencies in the presence of a large number of alleles, and to provide a global test of gametic phase disequilibrium of several restriction site polymorphisms.

ANALYSIS of categorical observations constitutes one of the principal ways of handling genetic data (WEIR 1990). In such analyses, hypothesis testing regarding specific genetic models generally proceeds by comparing the observed frequencies in samples with the ones that are expected under the model. The computation of expected frequencies requires estimation of parameters. Therefore, the choice of estimation procedures and consideration of appropriate statistics both play significant roles in hypothesis testing involving categorical data. Tests of nonrandomness of mutagen-induced mutations across loci (HANASH *et al.* 1988), tests of Hardy-Weinberg equilibrium in the presence of a large number of alleles in samples of comparatively moderate sizes (LEVENE 1949; GUO and THOMPSON 1992) and global tests of gametic phase disequilibria within a defined DNA segment (BLANTON and CHAKRAVARTI 1987) are examples of population genetic questions where such situations arise.

Generally Monte Carlo methods of simulation are adopted to ameliorate the problems of inadequacy of sample sizes, as evidenced in these publications. In this paper, a more complete solution is given using a combinatorial approach. Since the approach used here emerges from the combinatorics of the classical occupancy problem (FELLER 1968), I call this class of problems *generalized occupancy problems* in population genetics.

FORMULATION OF THE GENERALIZED OCCUPANCY PROBLEM

We consider a multinomial distribution with K classes (K can be arbitrarily large), with class probabilities represented by the vector $\pi' = (\pi_1, \pi_2, \ldots, \pi_K)$. Obviously, the elements of π form a simplex on a K-dimensional space satisfying

$$0 < \pi_i < 1, \sum_{i=1}^{K} \pi_i = 1. \tag{1}$$

When a random sample (with replacement) of size n is drawn from such a distribution, several characteristics of the sample size are of interest; *e.g.*, the number of classes represented in the sample, the number of classes represented with a specified number of sampling units in each of these classes, and the sampling distributions of such variables. Denote the observed number of classes that are represented in the sample by X. Before attempting to solve the problem, we note two interesting properties of the random variable X.

First, when all π_i's are equal ($\pi_i = 1/K$ for all i), but the number of classes, K, is unknown, ARNOLD and BEAVER (1988) showed that X is a sufficient statistic for the parameter K, and its sampling distribution is given by

$$P_{[m]} = \text{Prob.}(X = m) = \binom{K}{m} \frac{m!\, S_n^{(m)}}{K^n}, \tag{2}$$

for $m = 1, 2, \ldots, (K,n)$, where $S_n^{(m)}$ is a Stirling number of the second kind (ABRAMOWITZ and STEGUN 1965). This is the case of the classical occupancy problem.

Second, in the general case of unknown unequal π_i's, but the number of classes (K) is known, the expectation and variance of X can be derived using the indicator variable approach (EMIGH 1983, CHAKRABORTY, SMOUSE and NEEL 1988), without specifying the complete distribution of X. For this we define K

indicator variables, $Y_1, Y_2, \ldots, Y_K$, so that

$$Y_i = \begin{cases} 1 \text{ if the } i\text{th class is unobserved in the sample,} \\ 0 \text{ otherwise.} \end{cases}$$

With $Y = \sum_{i=1}^{K} Y_i$, we have $X = K - Y$, and hence, the expectation of X is

$$\begin{aligned} E(X) &= K - \sum_{i=1}^{K} E(Y_i) \\ &= K - \sum_{i=1}^{K} (1 - \pi_i)^n. \end{aligned} \tag{3}$$

Furthermore, the variance of X is given by

$$V(X) = V(Y) = \sum_{i=1}^{K} V(Y_i) + \sum_{i \neq j=1}^{K\,K} \text{Cov}(Y_i, Y_j). \tag{4}$$

Since Y_i's are Bernoulli variables,

$$V(Y_i) = (1 - \pi_i)^n[1 - (1 - \pi_i)^n], \tag{5}$$

and

$$\begin{aligned} \text{Cov}(Y_i, Y_j) &= (1 - \pi_i - \pi_j)^n \\ &\quad - (1 - \pi_i)^n(1 - \pi_j)^n. \end{aligned} \tag{6}$$

Substituting into Equation 4, we obtain

$$\begin{aligned} V(X) &= \sum_{i=1}^{K} (1 - \pi_i)^n \cdot [1 - \textstyle\sum (1 - \pi_i)^n] \\ &\quad + \sum_{i \neq j=1}^{K\,K} (1 - \pi_i - \pi_j)^n. \end{aligned} \tag{7}$$

When n is large, and each π_i small, we may approximate each term of the formulas (3) and (7) by $(1 - \pi_i)^n \approx e^{-n\pi_i}$ and $(1 - \pi_i - \pi_j)^n \approx e^{-n(\pi_i + \pi_j)}$, so that the variance of X can be approximated by

$$V(X) \approx \sum_{i=1}^{K} e^{-n\pi_i}(1 - e^{-n\pi_i}), \tag{8}$$

as shown in CHAKRABORTY, SMOUSE and NEEL (1988). Formula (7) is, however, exact, and not difficult to compute numerically even if the number of classes is large.

COMBINATORIAL SOLUTION OF THE PROBABILITY FUNCTION OF X

Instead of working with the variable X, the observed number of nonempty classes, it is easier to work with its complement, $Y = K - X$, the number of classes not represented in the sample. Let A_i be the event that the ith class ($i = 1, 2, \ldots, K$) is not represented in the sample. The A_i's are not exclusive of each other, since there can be sample configurations where more than one class frequencies can be simultaneously zero.

For sampling with replacement, the following formulas hold:

$$p_i = \text{Prob.}(A_i) = (1 - \pi_i)^n, \tag{9a}$$

$$p_{ij} = \text{Prob.}(A_iA_j) = (1 - \pi_i - \pi_j)^n, \tag{9b}$$

$$p_{ijk} = \text{Prob.}(A_iA_jA_k) = (1 - \pi_i - \pi_j - \pi_k)^n, \tag{9c}$$

etc., for all $i \neq j \neq k = 1, 2, \ldots, K$.

Following FELLER (1968, p. 99), we define a sequence of summations $\{T_1, T_2, \ldots, T_K\}$ where

$$T_1 = \sum_i p_i, \quad T_2 = \sum_i \sum_j p_{ij}, \quad T_3 = \sum_i \sum_j \sum_k p_{ijk}, \text{ etc.}$$

where the summations are taken such that $1 \leq i < j < k < \ldots \leq K$, so that each combination appears once and only once; hence, the summation $T_r (1 \leq r \leq K)$ contains $\binom{K}{2}$ terms. The last term, T_K reduces to only one term,

$$T_K = \text{Prob.}(A_1A_2 \ldots A_K) = P_{123\ldots K},$$

which is the probability of simultaneous occurrences of all K events A_1 through A_K. Clearly, $T_K = 0$, and furthermore,

$$T_{K-1} = \sum_{i=1}^{K} \pi_i^n, \tag{10a}$$

$$T_{K-2} = \sum_{j<1=1}^{K\,K} (\pi_i + \pi_j)^n, \tag{10b}$$

$$T_{K-3} = \sum_{k>j>i=1}^{K\,K\,K} (\pi_i + \pi_j + \pi_k)^n, \tag{10c}$$

etc.

Applying FELLER's theorem (FELLER 1968, p. 106), we obtain

$$\begin{aligned} P_{[K-m]} &= \text{Prob.}(X = K - m) = \text{Prob.}(Y = m) \\ &= \sum_{i=m}^{K} (-1)^{i-m} \binom{i}{m} T_i, \end{aligned} \tag{11}$$

for $K - \min(K,n) \leq m \leq K - 1$, giving the sampling distribution of X, the number of nonempty classes in a sample of size n. Note that in Equation 11, T_0 is conventionally defined as unity [see also FELLER (1968)].

Algebraic simplifications of the standard binomial expansion establish that the Equations 3 and 7 are compatible with the probability function (11), resulting in $E(X) = K - T_1$, and $V(X) = T_1(1 - T_1) + T_2$. Similar computational logic also yields the rth factorial moment of Y, $\mu_{[r]}(Y)$, given by

$$\mu_{[r]}(Y) = E[Y(Y - 1) \ldots (Y - r + 1)] = r! \cdot T_r, \tag{12}$$

for any $r \geq 1$, giving the complete characterization of the probability function (11) through its moments.

When all π_i's are equal (*i.e.*, $\pi_i = 1/K$ for all i),

$$T_1 = K[(K - 1)/K]^n,$$

and

$$T_2 = K(K - 1)\cdot[(K - 2)/K]^n/2,$$

and hence

$$E(X) = K[1 - \{(K - 1)/K\}^n] \tag{13}$$

and

$$V(X) = K\cdot\{(K - 1)/K\}^n\cdot[1 - K\cdot\{(K - 1)/K\}^n] + K(K - 1)\cdot\{(K - 2)/K\}^n, \tag{14}$$

which are derived in ARNOLD and BEAVER (1988). Further,

$$T_i = \binom{K}{i}\{(K - i)/K\}^n,$$

so that the probability function (11) reduces to

$$P_{[K-m]} = \sum_{i=m}^{K} (-1)^{i-m}\binom{i}{m}\binom{K}{i}(K - i)^n/K^n = \binom{K}{K - m}\cdot\frac{m!\, S_n^{(K-m)}}{K^n},$$

invoking the definition of a Stirling number of the second kind (ABRAMOWITZ and STEGUN 1965, p. 824).

Therefore, the above derivations show that the sampling distribution of the number of observed classes in a finite sample can be analytically specified for any arbitrary multinomial distribution. This generalizes the solution of ARNOLD and BEAVER (1988). The algebraic solutions of other relevant random variables (*e.g.*, the number of classes containing a specified number of sampling units within each of them) are also similar.

APPLICATIONS

Three applications of this generalized occupancy problem may be considered to show the utility of this theory.

Are mutagen-induced mutations equally likely to occur at all loci? HANASH *et al.* (1988) demonstrated that somatic cell gene mutations altering protein structure do not occur with equal probability at all loci when cultured human lymphoblastoid cell lines are treated with mutagens like ethylnitrosourea. To show this, they used the technique of two-dimensional polyacrylamide gel electrophoresis, and found 65 mutants occurring at 49 of the 263 loci scored in their experiments. The locus-specific distributions of the mutation frequencies in their work were: three mutants observed at each of five loci ($n_1 = \ldots = n_5 = 3$), two mutants at each of six loci ($n_6 = \ldots = n_{11} = 2$), and one mutant at each of 38 loci ($n_{12} = \ldots = n_{49} = 1$). No mutation was detected at each of the remaining 214 loci ($n_{50} = \ldots = n_{263} = 0$). The null hypothesis to be tested is H_0: {π_i = the probability of mutation occurring at the ith locus = $1/K = 1/263$, for all i}. In their work, the authors defined the concept of "repeat" mutations (R), noting that 16 mutations occurred at loci each of which contained already one mutation (*i.e.*, $R = n - X$, where X is the number of loci containing at least one mutant). Under the null hypothesis of equiprobable mutation frequencies across loci, the number of "repeat" mutations should be small, since $K = 263$ is much larger than the sample size $n = 65$. Through simulations, they determined that the probability of 16 or more "repeat" mutants is below 0.0005, and hence they conclude that mutagen-induced mutations are not equally likely to occur at all loci.

The theory described above provides a complete analytical solution. With $K = 263$, $n = 65$, from the probability function (11), we have $P(1 \leq X \leq 49) = 0.0001$, under the null hypothesis H_0, suggesting that the qualitative conclusions of HANASH *et al.* (1988) is the same as the one obtained by the present analytical solution.

Since under the null hypothesis H_0, X is a sufficient statistic for K, it is of interest to check if the exact distribution (11) of X can be approximated by a normal distribution. Using Equations 3 and 7, under H_0, we have $E(X) = 57.687$ and $V(X) = 5.289$, yielding a normal deviate $z = -3.78$ for $X = 49$. Although, $P(X \leq 49)$ becomes 0.00008, somewhat smaller than the analytical solution, the normal approximation for the distribution of X appears adequate for the sample size of this experiment.

Test of Hardy-Weinberg expectation based on the observed numbers of distinct genotypes in a finite sample: The present approach can also be used to design conditional tests for Hardy-Weinberg expectation (HWE) of genotypic distribution given the allele frequencies in a sample. Generally, this is done by either a likelihood ratio test or a goodness of fit chi-square test, contrasting the observed and expected frequencies of all possible genotypes (WEIR 1990; GUO and THOMSON 1992). However, there are occasions when the number of alleles is so large that many of the genotypes are either not observed in a sample, or the observed frequencies of several genotypes are so small that the large sample approximation of these test statistics is unwarranted. The recently discovered VNTR polymorphisms provide examples of this nature, where the number of possible alleles (K) are often so large that no reasonably sized survey can encompass all possible genotypes [$K(K + 1)/2$] in any given sample. One might ask, what would be the conditional distribution of the numbers of distinct genotypes (of homozygote and heterozygote types,

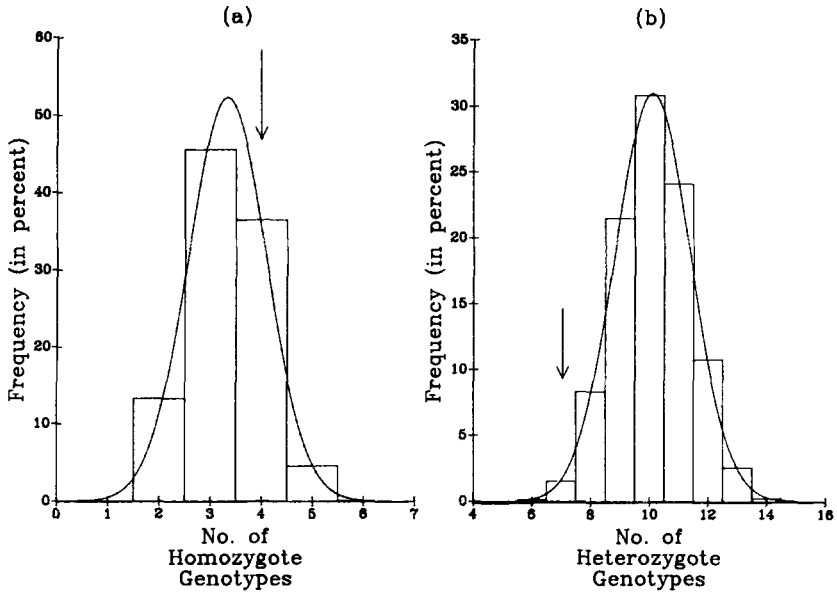

FIGURE 1.—The sampling distributions of the number of distinct homozygote (a) and heterozygote (b) genotypes at the D1S76 VNTR locus in a sample of 35 individuals from Papua New Guinea (DEKA, CHAKRABORTY and FERRELL 1990) under the assumption of Hardy-Weinberg equilibrium frequencies of genotypic proportions. The histograms are exact computations (Equation 11) conditioned on the observed allele frequencies, and the smooth curves are the normal approximations based on mean and variance, given in Equations 3 and 7. The arrows indicate the observed numbers of distinct homozygote (4) and heterozygote genotypes (7) found in the sample.

separately) given the observed allele frequencies in a sample of n individuals? Under HWE the genotypic probabilities are: p_i^2 for homozygotes and $2p_ip_j$ for heterozygotes, where p_i represents the allele frequencies in the population, and hence, we can use the above analytical formulation to compute the exact distributions of the distinct numbers of homozygote and heterozygote genotypes seen in a sample.

Figure 1 shows a numerical example of such computations. DEKA, CHAKRABORTY and FERRELL (1991) recently surveyed the New Guinea population for VNTR polymorphisms at six loci. At the D1S76 locus, they discovered 6 alleles in a sample of 35 individuals. Gene counting showed that in the sample of 70 genes at this locus, the allele counts of these 6 alleles are 1, 3, 7, 9, 25 and 25. In total they observed 20 heterozygous individuals (consisting of 7 distinct genotypes). However, under the HWE assumption, the expected frequency of heterozygotes from the above allele counts is 25.4, showing a significant deficiency of heterozygotes ($P < 0.05$, by the traditional chi-square test with 1 d.f.). Since the observed numbers of distinct homozygote and heterozygote genotypes in their sample were 4 and 7, respectively, we can ask if these observations deviate from their respective expectations under the HWE assumption, conditional upon the observed allele frequency distribution in the sample. Figure 1a shows the exact distribution of the observed number of distinct homozygote genotypes (drawn as a histogram) and Figure 1b gives the same for the observed number of distinct heterozygote genotypes, under the HWE assumption conditioned on the observed allele frequencies. The arrows represent the observed statistics. Clearly, the observed number of distinct homozygote genotypes ($m = 4$) is not at variance with the HWE assumption, since the probability of observing 4 or more distinct homozygote genotypes is 0.411. Under HWE the probability of observing 7 or less distinct heterozygote genotypes is 0.017, suggesting that a significant deficiency is observed in the total number of heterozygotes as well as in the number of distinct heterozygote genotypes.

From Equations 3 and 7, the mean and variance of the number of distinct genotypes were computed as 3.329 and 0.578 for the homozygotes, and 10.106 and 1.652 for the heterozygote genotypes, respectively. The expected distributions under the normality approximation are also shown by the smooth curves in both panels of Figure 1. The normal approximation is fairly adequate for the distribution of distinct heterozygote genotypes, while it is not so for the homozygotes because of the narrow range of variation in the number of distinct homozygote genotypes. Under the normality approximation, the normal deviate corresponding to observing 7 or less distinct heterozygote genotypes is $z = -2.41$, with a P value of 0.008, which is smaller than the exact P value, but does not alter the qualitative inference with regard to the validity of HWE.

Global test of disequilibrium based on multiple-locus haplotype data: Consider the haplotype frequency data surveyed by WAINSCOAT *et al.* (1986) at the β-globin gene cluster detected by five polymorphic restriction sites, at each of which there are two segregating alleles. This results in $2^5 = 32$ possible haplotypes at this gene region, but in a sample of 55 chromosomes sampled from a Polynesian population, these authors found only 5 observed haplotypes (see Table 1 of WAINSCOAT *et al.* 1986). One might ask, what is the expected distribution of the number of haplotypes given that these five sites are independently segregating? Figure 2 shows the exact distribu-

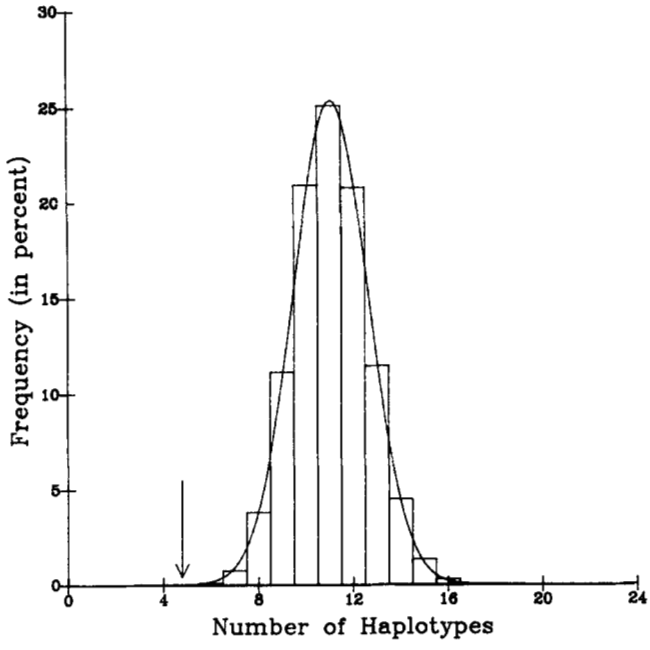

FIGURE 2.—The sampling distribution of the number of DNA haplotypes at the β-globin gene cluster, defined by 5 restriction site polymorphisms (WAINSCOAT *et al.* 1986), in a sample of 55 chromosomes from a Polynesian population under the assumption of complete linkage equilibrium. The histogram is the exact distribution based on Equation 11 and the smooth curve is its normal approximation based on mean and variance given by Equations 3 and 7. The arrow indicates the observed number of 5 different haplotypes found in the sample.

tion (represented by a histogram), following Equation 11, where the expected haplotype frequencies are assumed to follow the independent segregation rule. Clearly, almost the entire distribution is to the right of the observed number ($m = 5$) of haplotypes, giving a small probability of observing 5 or less haplotypes ($P < 10^{-5}$), suggesting that the observed number of haplotypes is incompatible with the assumption of independent segregation. Under independent segregation, the expected mean and variance of the observed number of haplotypes in a sample of 55 chromosomes are 11.059 and 2.477, respectively. The normal approximation of the sampling distribution is again shown by the smooth curve of Figure 2. The normal approximation appears satisfactory; the normal deviate corresponding to the observed number 5 is $z = -3.85$, giving a P value of 0.00006, somewhat larger than the exact P value, but does not change the inference regarding gametic phase disequilibrium.

DISCUSSION

The analytical theory presented here, along with the specific applications, indicate that the generalized occupancy problem has a number of possible applications in population genetics. This is particularly true in the context of sparse data, where by the very nature of the problem, the exact sampling distribution must be evaluated and no adequate large sample approximation is available. This theory enables comparison of occurrences of several biological endpoints in cross survey comparisons, adjusting for sample size differences, as shown in CHAKRABORTY, SMOUSE and NEEL (1988a).

The consideration of observed number of classes raise the possibility of some loss of information, since the frequencies of the different observed categories do not enter into the present analysis. GUO and THOMPSON (1992), for example, developed a simulation-based exact test procedure for testing HWE that utilizes frequencies of all genotypes observed. However, when the number of categories is large compared to the sample size, most of the observed categories are likely to have one or a few sample points in them and such loss of information may not be critical.

The application of the theory presented here (see *e.g.*, formulas 3, 7 and 11) requires the knowledge of the parameters (π_i's), and all numerical results presented are based on sample estimates of π_is from the data. In this sense, the tests conducted are all conditional. For example, in the context of testing the Hardy-Weinberg assumption (Figure 1), the six allele frequencies used are actually the ones observed in the sample. The question that we truly asked: is the genotype data consistent with the hypothesis that all of these six alleles (with the given frequencies) randomly combined to form genotypes? While unconditional test procedures for similar situations have been described in the literature (BINDER and HEERMANN 1988), the conditional aspect of the sampling distribution presented here (*i.e.*, when formula (11) is computed from estimated π_i values) is no different from those of the test procedures suggested in GUO and THOMPSON (1992), WEIR (1992), or CHAKRABORTY, SRINIVASAN and DE ANDRADE (1993).

In closing, another important application of this methodology may be cited that has relevance to mating behavior studies in human and animal populations. With the advent of hypervariable polymorphic loci, it is now possible to assert parentage from allele sharing with a small number of hypervariable loci, or even by using DNA fingerprints obtained from a multi-locus probe (JEFFREYS, BROOKFIELD and SEMENOFF 1985). However, since mutation rates at such loci are not negligible (JEFFREYS *et al.* 1988), it becomes important to know the sampling distribution of the number of loci with reference to which a randomly accused man could be excluded if this man is not the father of a child born to a specific mother. Analytical solution of this problem, suggested in CHAKRABORTY and SCHULL (1976), or that of the distribution of the proportions of offspring in a population with unambiguous parentage (CHAKRABORTY, MEAGHER and SMOUSE 1988)

relies on combinatorial approaches that are similar to the theory described here.

This research is partially supported by grants GM41399 and GM45861 from the National Institutes of Health and grant 92-IJ-CX-K024 from the National Institute of Justice. Comments from the associate editor and two anonymous reviewers have greatly improved the presentation.

LITERATURE CITED

ABRAMOWITZ, M., and I. A. STEGUN, 1965 *Handbook of Mathematical Functions*. Dover, New York.

ARNOLD, B. C., and R. J. BEAVER, 1988 Estimation of the number of classes in a population. Biometrical J. **30:** 413–424.

BINDER, K., and D. W. HEERMANN, 1988 *Monte Carlo Methods in Statistical Physics*. Springer-Verlag, Berlin.

BLANTON, S. H., and A. CHAKRAVARTI, 1987 A global test of linkage disequilibrium. Am. J. Hum. Genet. **41:** A250.

CHAKRABORTY, R., T. R. MEAGHER and P. E. SMOUSE, 1988 Parentage analysis with genetic markers in natural populations. I. The expected proportion of offspring with unambiguous paternity. Genetics **118:** 527–536.

CHAKRABORTY, R., and W. J. SCHULL, 1976 A note on the distribution of the number of exclusions to be expected in paternity testing. Am. J. Hum. Genet. **28:** 615–618.

CHAKRABORTY, R., P. E. SMOUSE and J. V. NEEL, 1988 Population amalgamation and genetic variation: observations on artificially agglomerated tribal populations of Central and South America. Am. J. Hum. Genet. **43:** 709–725.

CHAKRABORTY, R., M. R. SRINIVASAN and M. DE ANDRADE, 1993 Intraclass and interclass correlations of allele sizes within and between loci in DNA typing data. Genetics **133:** 411–419.

DEKA, R., R. CHAKRABORTY and R. E. FERRELL, 1991 A population genetic study of six VNTR loci in three ethnically defined populations. Genomics **11:** 83–92.

EMIGH, T. H., 1983 On the number of observed classes from a multinomial distribution. Biometrics **39:** 485–491.

FELLER, W., 1968 *An Introduction to Probability Theory and Its Applications*. Wiley, New York.

GUO, S. W., and E. A. THOMPSON, 1992 Performing the exact test of Hardy-Weinberg proportions for multiple alleles. Biometrics **48:** 361–372.

HANASH, S. M., M. BOEHNKE, E. H. Y. CHU, J. V. NEEL and R. D. KUICK, 1988 Nonrandom distribution of structural mutants in ethylnitrosourea treatment of cultured human lymphoblastoid cells. Proc. Natl. Acad. Sci. USA **85:** 165–169.

JEFFREYS, A. J., J. F. Y. BROOKFIELD and R. SEMENOFF, 1985 Positive identification of an immigration test-case using human DNA fingerprints. Nature **314:** 818–819.

JEFFREYS, A. J., N. J. ROYLE, V. WILSON and Z. WONG, 1988 Spontaneous mutation rates in new length alleles at tandem-repetitive hypervariable loci in human DNA. Nature **332:** 278–281.

LEVENE, H., 1949 On a matching problem arising in genetics. Ann. Math. Stat. **20:** 91–94.

WAINSCOAT, J. S., A. V. S. HILL, A. L. BOYCE, J. FLINT, M. HERNANDEZ, S. L. THEIN, J. M. OLD, J. R. LYNCH, A. G. FALUSI, D. J. WEATHERALL and J. B. CLEGG, 1986 Evolutionary relationships of human populations from an analysis of nuclear DNA polymorphisms. Nature **319:** 491–493.

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

WEIR, B. S., 1992 Independence of VNTR alleles defined as fixed bins. Genetics **130:** 873–887.

Communicating editor: E. THOMPSON