Environmental Health Perspectives
Vol. 87, pp. 149–155, 1990

Estimation of Gene Frequency and Test for Hardy-Weinberg Equilibrium in the HLA System

by Masaaki Matsuura* and Shinto Eguchi†

This paper concerns the testing for Hardy-Weinberg equilibrium and the estimation of gene frequency in the human leukocyte antigens (HLA) system. An extensive simulation study for both testing and estimation is given for investigating the performance of the projection method by Eguchi and Matsuura, which has a closed form, and the method is asymptotically equivalent to the maximum likelihood method. We compare our projection test statistic with the likelihood ratio test and the single degree of freedom chi-square test suggested by Nam and Gart. Actual mean square errors of the projection estimator of gene frequency under the Hardy-Weinberg equilibrium are compared with the maximum likelihood estimator and some other estimators recently discussed by Nam.

Introduction

The human leukocyte antigen (HLA) system has been observed, not only in human genetics and anthropology but also in biostatistics. Farewell and Dahlberg (*1*) investigated some statistical methodology including the analyses for associating particular diseases with some genotypes in the HLA system. For these statistical analyses or gene frequency estimations, one of the most fundamental assumptions is the Hardy-Weinberg law. Eguchi and Matsuura (*2*) proposed a projection method for the testing and estimation problem. The method is associated with a geometric interpretation similar to least-square method in the linear regression model. The key idea is to construct the regression plane, which is tangent to the (gene frequency) parameter space in the Hardy-Weinberg equilibrium. Thus, the method is based on the projection of sufficient statistics onto the regression plane. The test statistic can be regarded as the residual sum of squares and the estimator for gene frequency regarded as the least-square estimator.

This paper describes the structure of HLA data used in the statistical analysis and gives brief reviews on the methods used to test the Hardy-Weinberg equilibrium and estimate gene frequencies in the HLA system. Furthermore, the test statistics and the gene frequency estimator given by Eguchi and Matsuura (*2*) are compared with those either recently proposed or those formally used in the practical fields in a simulation study. The test statistics that are compared include the ordinal likelihood ratio statistics and the single degree of freedom test statistic proposed by Nam and Gart (*3*). The simple Bernstein's method extended by Yasuda and Kimura (*4*); the gene-counting method developed by Smith et al. (*5–7*) and extended by Yasuda and Kimura (*4*); and the bias reduced method proposed by Nam (*8*) are included for comparing the gene frequency estimators.

Structure of the HLA Data and Model under the Hardy-Weinberg Law

The HLA system consists of several linked loci on chromosome 6. A large number of alleles in the population are at each locus, but the complete set of alleles at a particular locus has not yet been identified. New loci and alleles have been recognized at the International HLA Workshop and their names are decided by the World Health Organization (WHO) Nomenclature Committee. By 1984 the loci A, B, C, D, DR, DP, and DQ were recognized [see Albert et al. (*9*) for all alleles found at each locus]. Throughout this paper, we consider the alleles at a fixed locus and treat the phenotypic data.

Using the notation of Yasuda and Kimura (*4*) and Nam and Gart (*3*), $m - 1$ known alleles (or antigens) at a given locus are denoted as $A_1, A_2, \ldots, A_{m-1}$, and a pool of unidentified alleles denoted by O, which is considered a recessive allele in the generalized ABO-like blood system. Since chromosomes are paired, an individual will

*Department of Epidemiology and Social Medicine, Research Institute for Nuclear Medicine and Biology, Hiroshima University, Hiroshima 734, Japan.

†Department of Mathematics, Shimane University, Matsue 690, Japan.

Address reprint requests to M. Matsuura, Department of Epidemiology and Social Medicine, Research Institute for Nuclear Medicine and Biology, Hiroshima University, Hiroshima 734, Japan.

have two alleles. Thus, the possible combinations of genotypes are A_iA_i, A_iO, A_jA_k, and OO, $1 \leq i \leq m - 1$, $1 \leq j \leq k \leq m - 1$, and the total genotype is $m(m + 1)/2$. The indices i, j, and k are run, as previously mentioned, and we sometimes omit their ranges.

For the serologically defined antigens (for example HLA-A, -B, -C, and -DR), typing panels—where the known antibody corresponding to the antigen A_i for some i is included—are used for detecting the HLA type of an individual, which is decided by an antigen-antibody reaction. The mixed lymphocyte reaction is used for HLA-D locus. The alleles $A_1, A_2, \ldots, A_{m-1}$ are codominant, therefore, if an individual possesses two distinct antigens A_j and A_k, the phenotype is denoted by A_jA_k, which is identical to the genotype. However, if only one antigen A_i is detected for a person, it is not possible to distinguish the genotype A_iA_i from the genotype A_iO without typing other family members, so we denote phenotype of this person as A_i. If no antigen is detected, the person can be considered as having two unknown alleles that are different from the known alleles A_i, $i = 1, 2, \ldots, m - 1$. In this case the genotype is referred to as OO and the phenotype denoted by O. We sometimes call the phenotype O double blanks. Accordingly, the total phenotype becomes $(m^2 - m + 2)/2$.

There are two possibilities in constructing the genotype A_jA_k, $1 \leq j < k \leq m - 1$: one assumes that the allele A_j comes from the mother and A_k, from the father; the other is the reverse with the allele A_k coming from mother and vice versa. According to the Hardy-Weinberg equilibrium or random mating, the probabilities of the genotype A_jA_k in the population are given by $p_jp_k + p_kp_j = 2p_jp_k$, $1 \leq j < k \leq m - 1$, where p_i is the gene frequency of the codominant allele A_i, $1 \leq i \leq m - 1$. Here the frequency of a pool of unknown alleles is denoted by r, so that $\sum_{i=1}^{m-1} p_i + r = 1$. Similarly, probabilities of genotypes A_iA_i, A_iO, and OO are given by p_i^2, $2p_ir$ and r^2, respectively, where $i = 1, 2, \ldots, m - 1$. For the phenotypic data the probability of the phenotype A_i can be taken by merging the probabilities of the genotypes A_iA_i and A_iO, thus we obtain $p_i^2 + 2p_ir$ for $i = 1, 2, \ldots, m - 1$. Probabilities of the phenotypes A_jA_k and O are the same as those genotypes A_jA_k and OO, respectively. In the population of sample size N, the observed numbers of phenotypes A_i, A_jA_k, and O are written by n_i, n_{jk}, and n_{OO}, respectively. Define $n_{kj} = n_{jk}$, $1 \leq j < k \leq m - 1$. For the convenience notation, let G_i be the sum of the observations with the phenotype A_i, that is

$$G_i = n_i + \sum_{j \neq i}^{m-1} n_{ij} \text{ for } i = 1,2, \ldots, m - 1.$$

The next section investigates the testing problem based on the phenotypic data.

Testing the Hardy-Weinberg Equilibrium

As noted in the "Introduction," the notion of gene frequency is reasonable only when the population is subject to the Hardy-Weinberg law. Therefore, if this assumption does not hold, resulting estimates of gene frequencies will be misleading whatever method is used. Thus, a check of this assumption should be included in the analysis of gene frequencies.

In a population of sample size N, the counts n_i, n_{jk}, and n_{OO}, $1 \leq i \leq m - 1$, $1 \leq j < k \leq m - 1$, respectively, have a multinomial distribution with cell parameter vector

$$\boldsymbol{\pi} = (\pi_i, \pi_{jk}, \pi_0)_{1 \leq i \leq m-1, 1 \leq j < k \leq m-1},$$

where $\pi_O + \Sigma \pi_i + \Sigma \pi_{jk} = 1$. If the population is subject to the Hardy-Weinberg law, then the vector $\boldsymbol{\pi}$ is in the subsurface

$$\begin{aligned}\Pi_{HW} = \{&\boldsymbol{\pi}_p = \\ &(p_i^2 + 2p_ir, 2p_jp_k, r^2)_{1 \leq i \leq m-1, 1 \leq j < k \leq m-1}: \\ &0 < p_i < 1 \ (1 \leq i \leq m - 1), \\ &0 < r < 1, \sum_{i=1}^{m-1} p_i + r = 1\}\end{aligned}$$

which is the subsurface in the space Π of cell parameter vectors. Thus the null hypothesis that the population is under the Hardy-Weinberg law is expressed as H: $\boldsymbol{\pi} \in \Pi_{HW}$ and the alternatives as K: $\boldsymbol{\pi} \in \Pi\text{-}\Pi_{HW}$.

Testing the hypothesis H is usually accomplished by means of a goodness-of-fit likelihood ratio or Pearson chi-square statistic with a null distribution characterized by chi-square distribution with $(m -)(m - 2)/2$ degrees of freedom. Let $L(\boldsymbol{\pi})$ be the likelihood of $\boldsymbol{\pi}$, then the likelihood ratio statistic is given by

$$\chi^2_{LR} = 2\{\log L(\hat{\boldsymbol{\pi}}) - \log L(\boldsymbol{\pi}_{p^*})\}$$

where $\hat{\boldsymbol{\pi}}$ is the full maximum likelihood estimator (MLE) on $\boldsymbol{\pi}$, and $\mathbf{p}^*$ is the MLE of $\boldsymbol{p} = (p_1, p_2, \ldots p_{m-1})^T$ under the Hardy-Weinberg law. Here, it is intractable to obtain the MLE $\mathbf{p}^*$, which may need an iteration method. Note that the minimum chi-square statistic also requires such an iteration technique. To obtain the MLE iteratively, Yasuda and Kimura (*4*) extended the gene counting method to the generalized ABO-like system. The gene counting method was devised by Ceppellini, Siniscalco and Smith (*5*) and develop by Smith (*6,7*). Using the gene counting method the estimates at the kth step are given by

$$p_i^{(k)} = \frac{G_i}{2N} + \frac{p_i^{(k-1)}}{p_i^{(k-1)} + 2r^{(k-1)}} \left(\frac{n_i}{2N}\right) \quad (1)$$

for $i = 1, 2, \ldots, m - 1$.

and

$$r^{(k-1)} = 1 - \sum_{i=1}^{m-1} p_i^{(k-1)}.$$

Here, the Bernstein estimator extended by Yasuda and Kimura (*4*) may be used for the first step, which is given by

$$\hat{p}_i = 1 - \sqrt{1 - (G_i / N)} \quad \text{for } i = 1, 2, \ldots, m - 1. \tag{2}$$

and

$$\hat{r} = \sqrt{n_{00}/N}$$

Nam and Gart (*3*) suggested another chi-square test statistic with a single degree of freedom

$$\chi_D^2 = \boldsymbol{D}^2 / V(\boldsymbol{D})$$

where

$$\boldsymbol{D} = 1 - \left(\sum_{i=1}^{m-1} \hat{p}_i + \hat{r} \right) \tag{3}$$

and $V(\boldsymbol{D})$ is the variance of $\boldsymbol{D}$ and is given by

$$V(\boldsymbol{D}) = \frac{1}{4N} \left[\left(\sum \frac{\hat{p}_i}{1 - \hat{p}_i} \right)^2 - \sum \left(\frac{\hat{p}_i}{1 - \hat{p}_i} \right)^2 \right]$$

When $m = 3$, or the ABO system, χ_D^2 reduces to the statistic given by Stevens (*10*). Though the statistic ${\chi^2}_D$ does not require the iterative manner, Eguchi and Matsuura (*2*) theoretically pointed out that the statistic ${\chi^2}_D$ essentially causes an acceptance region outside Π_{HW} whatever significance level one chooses when $m > 3$. This point will be observed in the simulation study here.

Eguchi and Matsuura (*2*) introduced a closed form chi square statistic

$$\chi_{PR}^2 = N\hat{\boldsymbol{\theta}}^T \, \mathcal{I}_{\hat{\theta}} \hat{\boldsymbol{\theta}}$$

where $\hat{\boldsymbol{\theta}} = (\theta_{12}, \theta_{13}, \ldots, \theta_{m-2,\, m-1})^T$ with $\theta_{jk} = n_{jk}/2N - \hat{p}_j\hat{p}_k$ $1 \leq j < k \leq m - 1$, and $\mathcal{I}_{\hat{\theta}}$ is the Fisher information matrix of $\boldsymbol{\theta}$ when evaluated at $\hat{\boldsymbol{\theta}}$, and is given in Appendix 1. The test statistic ${\chi^2}_{PR}$ has a closed representation of a residual sum of squares by a projection mapping onto the estimated regression plane. The derivation is based on the maximum likelihood method and closely related to a standard regression theory.

Estimation of Gene Frequencies

The simple Bernstein's estimator obtained by Eq. (2) has been widely used for the generalized ABO-like system. However, the sum of these estimates (i.e., $\Sigma \hat{p}_i + \hat{r}$) is not necessarily equal to one. Also, according to the simple Bernsteins's method, the estimate of the recessive gene frequency has the negative bias, (*11*). Furthermore, the simple Bernstein's estimator is generally inefficient when $m \geq 3$.

Smith (*7*) gave an alternative estimator of the recessive gene frequency defined by

$$\hat{r}_0 = 1 - \sum_{i=1}^{m-1} \hat{p}_i,$$

in which case the sum of estimates is equal to one and $V(\hat{r}_0) \leq V(\hat{r})$, but $\hat{r}_0$ may yield negative values. Yasuda and Kimura (*4*) proposed the so-called adjusted Bernstein's estimator

$$\hat{p}_i^* = \hat{p}_i(1 + \boldsymbol{D}/2) \text{ for } i = 1, 2, \ldots, m - 1,$$

and

$$\hat{r}^* = (\hat{r} + \boldsymbol{D}/2)(1 + \boldsymbol{D}/2),$$

where $\boldsymbol{D}$ is defined in Eq. (3). Here $\Sigma \hat{p}_i^* + \hat{r}^* = 1 - \boldsymbol{D}^2/4$, and $\hat{r}^*$ may also yield negative values. Nam and Gart (*3*) suggested the so-called modified Bernstein's estimator

$$\hat{p}'_i = \hat{p}_i / (1 - \boldsymbol{D}/2) \text{ for } i = 1, 2, \ldots, m - 1$$

and

$$\hat{r}' = (\hat{r} + \boldsymbol{D}/2) / (1 - \boldsymbol{D}/2)$$

and showed that the adjusted and modified estimators are not only inefficient, but also both of them have asymptotic variances larger than the simple Bernstein's estimator. They also investigated the problem that the method with a single gene-counting iteration, defined in Eq. (1) using the modified Bernstein's estimator as an initial step, leads to a nearly efficient estimator.

Haldane (*11*) recognized that the Bernstein's recessive gene estimator is negatively biased and suggested the corrected version:

$$\bar{r}^* = \sqrt{(n_{00} + 0.25) / (N + 0.25)}$$

Also, Nam (*8*) recently investigated the bias of the simple Bernstein's method and considered the reduction of its bias in the generalized ABO-like system. Removing the major bias term, he obtained a bias reduced Bernstein estimator as

$$\bar{p}_i^\dagger = 1 - \sqrt{1 - G_i / (N + 0.25)} \quad \text{for } i = 1, 2, \ldots, m - 1$$

for the codominant allele frequencies. For the recessive allele frequency, the form of the estimator is identical to Haldane's estimator. Nam (*8*) also suggested another estimator called the bias reduced Smith's estimator, which is given by the equation

$$\bar{r}^\dagger = 1 - \sum_{i=1}^{m-1} \bar{p}_i^\dagger$$

and he showed that this estimator is best used in achieving the smallest absolute bias among the simple Bernstein's, Smith's and Haldane's estimators.

Regarding the inference procedure based on the likelihood function, Fisher (*12*) introduced the joint maximum likelihood estimation of gene frequencies; Rao (*13*) also discussed the estimation. Farewell (*14*) gave the details of its application to HLA data and used some convenient reparametrizations to improve the rate of convergence [see also Farewell and Dahlberg (*1*)]. Furthermore, Gart and Nam (*15*) showed the detailed description for obtaining the MLEs using the scoring method discussed by Rao (*16*). They also showed that the MLEs of codominant allele frequencies are given by

$$\hat{p}_i^{**} = (G_i + n_i) / 2N \text{ for } i = 1, 2, \ldots, m - 1,$$

under the assumption that $r = 0$. As noted in the previous section, by using the gene-counting method defined in Eq. (1), the estimates obtained that are well converged are equivalent to the MLEs, which are fully efficient estimators, but the estimates require the iterative calculations.

Eguchi and Matsuura (*2*) suggested the noniterative estimator using the projection mapping and showed that the projection estimates are asymptotically equivalent to the MLEs. The projection estimates are obtained by

$$\tilde{\boldsymbol{p}} = (\boldsymbol{X}_{\hat{p}}^T \boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{p}}) \boldsymbol{X}_{\hat{p}})^{-1} \boldsymbol{X}_{\hat{p}}^T \boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{p}}) \hat{\boldsymbol{\tau}},$$

where $\tilde{\boldsymbol{p}} = (\tilde{p}_1, \tilde{p}_2, \ldots \tilde{p}_{m-1})^T$, and $\tilde{r} = 1 - \Sigma \tilde{p}_i$. Here $\boldsymbol{X}_{\hat{p}}$, $\boldsymbol{\Sigma}^{-1}(\hat{p})$ and $\hat{\tau}$ are given in Appendix 2. From the form of the projection estimator, $\tilde{p}$ can be regarded as the least square estimator in the regression model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{p} + \boldsymbol{\epsilon}$$

with $\boldsymbol{Y} = \hat{\mathcal{T}}_\tau^{1/2} \hat{\tau}$, $\boldsymbol{X} = \hat{\mathcal{T}}_\tau^{1/2} \boldsymbol{X}_{\hat{p}}$ and independent standard normal errors $\boldsymbol{\epsilon}$, where $\hat{\mathcal{T}}_\tau$ is given in Appendix 1.

Simulation Study

Simulation Procedure

For the test statistics χ^2_{PR}, χ^2_{LR} and χ^2_D, the actual Type I error (the size of test) and the powers are compared in a simulation study. We also investigate the actual squared errors of the projection method, gene counting method, bias reduced method and the simple Bernstein's method.

In the simulation for fixed N and m, our procedure for the data generating process under the Hardy-Weinberg equilibrium is taken as follows:

a) To determine the true gene frequencies, $m - 1$ random numbers are taken from the uniform distribution $U(0,1)$ and are arranged in ascending of magnitude such that $u_1 < u_2 < \ldots < u_{m-1}$. Define $r = u_1$, $p_i = u_{i+1} - u_i$ for $i = 1, 2, \ldots, m - 1$, and $p_{m-1} = 1 - u_{m-1}$. We regard the set $(r, p_1, \ldots, p_{m-1})$ as the true gene frequencies. Note that we omit the frequency set if $r > 0.5$, since such a situation is rate in practice; we also omit the one with $p_i < 0.005$ for some i, because the G_i is likely to be zero.

b) To determine the genotype of an individual, we generate a pair of random numbers (u_1^*, u_2^*) from U(0,1) and define $u_m = 1$. If u_1^* lies in a interval between 0 and u_1, namely $u_1^* \in [0, u_1]$, we regard the individual as having the recessive allele O, and if $u_1^* \in (u_i, u_{i+1})$, then we consider that the individual has the codominant allele A_i for some $i = 1, 2, \ldots, m - 1$. Note that the length of the interval (u_i, u_{i+1}) is equivalent to the true gene frequency p_i. Similarly, we determine the other phenotype of this individual using u_2^*. Therefore, we can decide the phenotype of this subject according to the genotype.

c) For fixed sample size N, step *b* is iterated N times and we count the observed numbers n_{00}, n_i, n_{jk} for $1 \leq i \leq m - 1$ and $1 \leq i < j \leq m - 1$. If $G_i = 0$ for some i, we omit the data set from the simulation.

d) We repeat the process from *a* to *c* 5000 times for fixed N and m.

For the situations where the Hardy-Weinberg law equilibrium does not hold, we change the steps *a* and *b* as follows:

a) Since the possible number of phenotypes is $1 + m(m - 1)/2$, $m(m - 1)/2$ random numbers are taken from U(0,1) and arranged in ascending of magnitude such that $v_1 < v_2 <, \ldots, < v_{m(m-1)/2}$. We define $\pi_0 = v_1$, $\pi_1 = v_2 - v_1, \ldots, \pi_{m-3,m-1} = v_{m(m-1)/2-1} - v_{m(m-1)/2}$ and $\pi_{m-2,m-1} = 1 - v_{m(m-1)/2}$. We regard these π_0, π_i, $1 \leq i \leq m - 1$, π_{jk}, $1 \leq j \leq k \leq m - 1$ as the cell probabilities of the multinomial distribution of sample size N.

b) If a random number v^* following U(0,1) lies in a interval $[0, v_1]$, we assume that the subject has the phenotype O. If $v^* \in (v_{i+1} - v_i)$, the subject has phenotype A_i. Similarly, the subject has the phenotype A_{jk}, if v^* falls in the corresponding interval of π_{jk}, $1 \leq j < k \leq m - 1$. The remainder is the same as the step (*b*). Note that true gene frequencies and true cell probabilities are not fixed, and these are changed in each replication in order to examine the numerous situations.

Results of Testing the Hardy-Weinberg Equilibrium

The results on sizes, based on the percentage of times the computed test statistic exceeded the 0.05 level critical values, are presented in Table 1 for $m - 1 = 5$ and 10. The results based on the 0.10 critical value are also included in Table 1.

For the test statistic based on the projection method, the actual Type I errors are less than nominal size 0.05 for $m - 1 = 5$ and for $m - 1 = 10$ with $N > 500$; the test statistic becomes conservative as the sample size becomes large. The likelihood ratio test seems to be somewhat liberal for the range of N between 250 and 2500, which is the usual sample size in practical fields. And this test is too conservative for $m - 1 = 10$ and $N < 75$. For the one degree of freedom test statistic, the actual Type I error is larger than 0.05. Therefore this test seems to be liberal, regardless of sample size.

To investigate the situation with no double blanks, that is $n_{00} = 0$, under the Hardy-Weinberg law, we

Table 1. Percentage of times that H was rejected at $\alpha = 0.05$ and 0.10 level when the H was correct ($\pi \epsilon \Pi_{HW}$); 5000 replicates.

		$\alpha = 0.05$			$\alpha = 0.10$		
$m-1$	N	χ^2_{PR}	χ^2_{LR}	χ^2_D	χ^2_{PR}	χ^2_{LR}	χ^2_D
5	20	4.0	3.8	7.0	7.1	9.4	11.3
	50	4.6	5.7	7.3	8.3	11.4	12.1
	100	4.4	5.5	6.7	8.3	11.2	12.0
	250	4.0	5.6	6.0	8.1	10.8	11.0
	500	4.7	6.2	6.0	9.0	11.6	10.8
	1,000	3.9	5.0	5.8	7.3	9.9	11.2
	2,500	3.9	5.2	5.8	8.5	10.3	10.9
	5,000	3.8	5.2	5.4	7.9	10.0	10.7
	10,000	3.8	4.9	5.3	7.7	9.4	10.6
10	50	6.3	1.0	5.1	10.0	2.8	9.5
	75	6.7	2.4	5.3	10.4	6.1	9.9
	100	6.3	3.9	6.4	10.1	8.1	11.3
	250	6.2	6.8	6.1	9.9	13.4	10.8
	500	4.9	6.6	6.2	8.8	13.2	11.7
	1,000	4.5	6.3	6.1	8.1	12.5	11.4
	2,500	3.7	6.1	6.0	8.1	11.5	12.1
	5,000	3.6	5.0	5.7	7.6	10.2	10.3
	10,000	3.5	4.6	5.7	7.4	9.4	10.6

Table 2. Percentage of times that H was rejected at $\alpha = 0.05$ and 0.10 when the H was correct ($\pi \epsilon \Pi_{HW}$) and $n_{OO} = 0$; 5000 replicates.

		$\alpha = 0.05$			$\alpha = 0.10$		
$m-1$	N	χ^2_{PR}	χ^2_{LR}	χ^2_D	χ^2_{PR}	χ^2_{LR}	χ^2_D
5	100	2.8	3.9	1.4	5.5	8.6	3.2
	250	3.6	4.1	4.0	6.3	8.6	8.7
	500	3.6	5.0	10.1	6.7	10.0	17.3
	1000	3.1	5.9	25.7	6.7	12.0	35.0
10	100	6.0	3.3	0.5	9.5	7.1	1.8
	250	5.9	5.7	2.4	9.2	11.8	6.7
	500	4.6	7.1	9.8	8.6	12.9	18.5
	1000	3.5	6.8	27.4	7.2	13.3	37.3

Table 3. Percentage of times that H was rejected at $\alpha = 0.05$ when the H was not correct ($\pi \epsilon \Pi$); 5000 replicates.

$m-1$	N	χ^2_{PR}	χ^2_{LR}	χ^2_D
5	100	98.0	99.0	56.0
	250	100.0	100.0	71.4
	500	100.0	100.0	80.4
	1000	100.0	100.0	86.1
10	100	98.9	99.8	27.9
	250	100.0	100.0	52.1
	500	100.0	100.0	64.8
	1000	100.0	100.0	74.6

restrict the true recessive gene frequency to be less than 0.05 in step (*b*) and picked up only the data set with $n_{00} = 0$ in step (*c*) until the number of data set attains 5000. This situation sometimes holds, as noted by Smouse and Williams (*17*), Gart and Nam (*18*), and Nam and Gart (*19,20*). The results are given in Table 2 for $m - 1 = 5$ and 10 and $N = 100$, 250, 500, and 1000. In this restricted situation theoretical behavior has not been examined. But the simulation shows that the actual Type I error decreases for the projection method, as compared to Table 1 with the same sample size. The similar behavior can be seen for the likelihood ratio test when $m - 1 = 5$ and $N < 500$, and $m - 1 = 10$ and $N < 250$; conversely, the Type I error tends to increase for a moderate large sample size. For test statistics with one degree of freedom, such behavior like the likelihood ratio test becomes too strong.

Table 3 presents the percentage of times that the null hypothesis H was rejected at the nominal 0.05 level when a data set follows a multinomial distribution with cell probabilities $\pi \epsilon \Pi$. Results for the projection method and the likelihood ratio method are almost evenly matched. However, the chi-square test χ^2_D with a single degree of freedom is likely to accept the null hypothesis. Thus, the power of this test can be considered as less powerful than the other two tests. This result may confirm the theoretical behavior of χ^2_D that was investigated by Eguchi and Matsuura (*2*). Nam and Gart (*20*) also showed the inefficiency of the statistic χ^2_D in another testing problem.

Results of Estimation of Gene Frequency

To compare several estimators we use the measure of actual mean square effort (MSE) defined by

$$\text{MSE} = \frac{1}{S} \sum_{s=1}^{S} (\hat{\boldsymbol{p}}(s) - \boldsymbol{p}^{(s)})^T (\hat{\boldsymbol{p}}(s) - \boldsymbol{p}^{(s)}),$$

Table 4. Actual mean square error under the Hardy-Weinberg law; 5000 replicates.

$m-1$	N	$(\tilde{p}_i, \tilde{r})$	$(p_i^{(k)}, r^{(k)})$	$(\hat{p}_i, \hat{r})$	$(\hat{p}_i, \hat{r}_o)$	$(\hat{p}_i^{\dagger}, \hat{r}^{*})$	$(\hat{p}_i^{\dagger}, \hat{r}_o^{\dagger})$
5	100	0.00466	0.00445	0.00644	0.00498	0.00561	0.00494
	250	0.00183	0.00175	0.00253	0.00195	0.00226	0.00194
	500	0.00092	0.00089	0.00127	0.00098	0.00116	0.00098
	1,000	0.00047	0.00045	0.00064	0.00050	0.00059	0.00050
	2,500	0.00019	0.00018	0.00025	0.00020	0.00024	0.00020
10	100	0.00483	0.00460	0.00677	0.00499	0.00593	0.00496
	250	0.00193	0.00186	0.00275	0.00199	0.00244	0.00198
	500	0.00096	0.00092	0.00140	0.00099	0.00125	0.00099
	1,000	0.00048	0.00046	0.00070	0.00050	0.00064	0.00050
	2,500	0.00019	0.00019	0.00028	0.00020	0.00026	0.00020

Table 5. Actual mean square error under the Hardy-Weinberg law with $n_{OO} = 0$; 5000 replicates.

$m-1$	N	$(\tilde{p}_i, \tilde{r})$	$(p_i^{(k)}, r^{(k)})$	$(\hat{p}_i, \hat{r})$	$(\hat{p}_i, \hat{r}_o)$	$(\hat{p}_i^{\dagger}, \hat{r}^{*})$	$(\hat{p}_i^{\dagger}, \hat{r}_o^{\dagger})$
5	100	0.00478	0.00400	0.00495	0.00535	0.00497	0.00527
	250	0.00192	0.00165	0.00244	0.00210	0.00189	0.00209
	500	0.00097	0.00085	0.00161	0.00104	0.00102	0.00104
	1,000	0.00049	0.00044	0.00119	0.00052	0.00069	0.00052
10	100	0.00482	0.00445	0.00518	0.00505	.00535	0.00501
	250	0.00193	0.00180	0.00251	0.00202	0.00205	0.00201
	500	0.00098	0.00092	0.00162	0.00102	0.00110	0.00101
	1,000	0.00049	0.00046	0.00119	0.00050	0.00071	0.00050

where $\tilde{\boldsymbol{p}}^{(s)}$ is the vector of certain estimates of true gene frequencies $\tilde{\boldsymbol{p}}^{(s)} = (p_1^{(s)}, p_2^{(s)}, \ldots, p_{m-1}^{(s)}, r^{(s)})$ in the sth simulation with replications $S = 5000$. Table 4 presents the comparison of MSE based on the several methods of estimation under the Hardy-Weinberg equilibrium. For the situation with no double blanks under the Hardy-Weinberg law, the results are shown in Table 5. Estimates based on the gene-counting method have small MSE compared to other estimates for all sample sizes and for the number of alleles in Tables 4 and 5. The MSE based on the projection method is slightly larger than that of MLE, but the difference in MSE becomes smaller as the sample size becomes larger.

Appendix 1

The Fisher information matrix of $\boldsymbol{\theta}$ evaluated at $\hat{\boldsymbol{\theta}}$, $\mathcal{I}_{\hat{\theta}}$, is given by

$$\mathcal{I}_{\hat{\theta}} = \mathcal{I}_{22} - \mathcal{I}_{12}^{\mathrm{T}}\, \mathcal{I}_{11}^{-1}\, \mathcal{I}_{12}.$$

where $\mathcal{I}_{11}$, $\mathcal{I}_{12}$, and $\mathcal{I}_{22}$ are submatrices of the Fisher information matrix of $(\hat{\boldsymbol{p}}, \hat{\theta})$, $\mathcal{I}_{(\hat{p},\hat{\theta})}$, written by

$$\mathcal{I}_{(\hat{p},\hat{\theta})} = \begin{pmatrix} \mathcal{I}_{12}\,\mathcal{I}_{12} \\ \mathcal{I}_{12}\,\mathcal{I}_{22} \end{pmatrix} = \boldsymbol{J}_{\psi}^{T}\, \boldsymbol{J}_{\varphi}^{T}\, \mathcal{I}_{\underline{\pi}}\, \boldsymbol{J}_{\varphi}\, \boldsymbol{J}_{\psi} \quad (= \boldsymbol{J}_{\psi}^{T}\, \mathcal{I}_{\tau}\, \boldsymbol{J}_{\psi})$$

Here $\boldsymbol{J}_{\varphi}$ is the inverse of Jacobi matrix of a mapping φ of $\underline{\boldsymbol{\pi}} = (\pi_i, \pi_{jk})$ into $\boldsymbol{\tau}$ by

$$\boldsymbol{\tau} = \varphi(\underline{\boldsymbol{\pi}}) = (1 - \sqrt{1 - G_i/N},\ \tfrac{1}{2}\pi_{jk})_{i,j,k}.$$

and is given by

$$\boldsymbol{J}_{\varphi} = \left(\frac{\partial \varphi_*}{\partial \underline{\pi}}\right)^{-1} = \begin{pmatrix} \boldsymbol{D}^{-1} & -2\boldsymbol{A}^{T} \\ \boldsymbol{O} & 2\boldsymbol{I}_f \end{pmatrix}.$$

where $f = (m-2)(m-1)/2$,

$$\boldsymbol{A}^{T} = \left(\begin{array}{c|c|c|c} 1, \ldots, 1 & 0, \ldots, 0 & \multirow{2}{*}{$\ldots$} & 0 \\ \boldsymbol{I}_{m-2} & 1, \ldots, 1 & & 1 \\ & \boldsymbol{I}_{m-3} & & 1 \end{array}\right)$$

and

$$\boldsymbol{D} = \operatorname{diag}\left(\frac{1}{2\sqrt{(1 - G_i/N)}}\right)_{1 \le i \le m-1}$$

$\boldsymbol{J}_{\psi}$ is the Jacobi matrix of a mapping ψ of $(\boldsymbol{p}, \boldsymbol{\theta})$ into $\boldsymbol{\tau}$ by

$$\psi(\boldsymbol{p}, \boldsymbol{\theta}) = (p_i,\ p_j p_k)_{i,j,k}.$$

and is given by

$$\boldsymbol{J}_{\psi} = \left(\frac{\partial \psi}{\partial(\boldsymbol{p}, \boldsymbol{\theta})}\right) = \begin{pmatrix} \boldsymbol{I}_{m-1} & \boldsymbol{O} \\ \boldsymbol{A}_p & \boldsymbol{I}_f \end{pmatrix}$$

where $f = (m-1)(m-2)/2$ and

$$\boldsymbol{A}_p^{T} = \frac{1}{2}\left(\begin{array}{c|c|c|c} p_2, p_3, \ldots, p_{m-1} & 0, 0, \ldots, 0 & \multirow{2}{*}{$\ldots$} & \begin{matrix} 0 & \\ p_{m-2} & p_{m-1} \end{matrix} \\ p_1 & p_3, p_4, \ldots, p_{m-1} & & \begin{matrix} p_{m-3} & 0 \\ 0 & p_{m-3} \end{matrix} \\ \begin{matrix} p_1 & \\ & \ddots \\ & p_1 \end{matrix} & \begin{matrix} p_2 & \\ & \ddots \\ & p_2 \end{matrix} & & \begin{matrix} 0 \\ p_{m-1} \\ p_{m-2} \end{matrix} \end{array}\right)$$

And $\mathcal{I}_{\pi}$ is the information matrix of π given by

$$\mathscr{I}_{\bar{\pi}} = (\text{diag } \bar{\pi})^{-1} + \frac{1}{\pi_0} \boldsymbol{l}\boldsymbol{l}^T$$

where $\boldsymbol{l} = (1, \ldots, 1)^T$.

Here we use the consistent estimator π_c and π_{0c} instead of $\bar{\pi}$ and π_0, respectively, to prevent the some elements of $\mathscr{I}_{\bar{\pi}}$ having infinite values. The consistent estimator $\boldsymbol{\pi}_c$ is given by

$$\boldsymbol{\pi}_c = (\hat{p}_i^2 + 2\hat{p}_i\hat{r}_c, 2\hat{p}_j\hat{p}_k)_{i,j,k}$$

and

$$\hat{r}_c = \begin{cases} \hat{r}_0 & , \text{ if } \Sigma\, p_i < 1, \\ \dfrac{\hat{r}_0}{2} + \left(\dfrac{\hat{r}_0^2}{4} + \dfrac{1}{N}\right)^{1/2} & (= r_{0c}), \text{ otherwise.} \end{cases}$$

Here $\hat{r}_0 = 1 - \sum_{i=1}^{m-1} \hat{p}_i$, the Smith's estimator. The derivation of $\hat{r}_c$ is based on the maximization of $\rho(r)$ under $\hat{r}_0 < 0$, defined by

$$\rho(r) = 2r \exp\left[- \frac{N}{2} (r - \hat{r}_0) \right].$$

Here $2r$ can be considered as prior density of r. And the $\hat{r}_{0c}$ have the properties as follows: (a) $\hat{r}_{0c} > 0$ and $\hat{r}_{0c} > \hat{r}_0$ for any $\hat{r}_0$, (b) if $\hat{r}_0$ is fixed and $N \to \infty$, then $\hat{r}_{0c} = \hat{r}_0$ when $\Sigma\, \hat{p}_i < 1$, and $\hat{r}_{0c} = 0$ when $\sum_{i=1}^{m-1} p_i \geqslant 1$, and (c) $\hat{r}_{0c}$ and $\hat{r}_0$ have the same asymptotic distribution. Therefore, the $\hat{r}_{0c}$ is consistent since $\hat{r}_0$ is consistent. Thus, we use the consistent estimator π_{0c} defined by $\hat{r}_c^2$.

Appendix 2

Using the notation in Appendix 1, $\boldsymbol{X}_{\hat{p}}$ is defined by

$$\boldsymbol{X}_{\hat{p}} = \begin{pmatrix} \boldsymbol{I}_{m-1} \\ \boldsymbol{A}_{\hat{p}} \end{pmatrix}$$

Here $\mathbf{A}_{\hat{p}}$ is same as $\mathbf{A}_p$ with the elements $\hat{p}_i$ instead of p_i. And $\hat{\tau}_p = (\hat{p}_i, \tfrac{1}{2}\hat{p}_j\hat{p}_k)_{i,j,k}$. Considering the

$$\hat{\boldsymbol{\tau}} - \boldsymbol{X}_{\hat{p}}\boldsymbol{p} = \begin{pmatrix} \hat{\boldsymbol{p}} - \boldsymbol{p} \\ \dfrac{1}{2}\hat{\boldsymbol{\pi}}_2 - \boldsymbol{A}_p\hat{\boldsymbol{p}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{I}_{m-1} & \boldsymbol{O} \\ -\boldsymbol{A}_p & \boldsymbol{I}_f \end{pmatrix} (\hat{\boldsymbol{\tau}} - \boldsymbol{X}_p\boldsymbol{p})$$

then we can obtain $\Sigma^{-1}(\hat{\boldsymbol{p}})$

$$\sum^{-1}(\hat{\boldsymbol{p}}) = \begin{pmatrix} \boldsymbol{I}_{m-1} & \boldsymbol{A}_{\hat{p}}^T \\ \boldsymbol{O} & \mathrm{I}_f \end{pmatrix} \mathscr{I}_{\hat{\tau}} \begin{pmatrix} \boldsymbol{I}_{m-1} & \boldsymbol{O} \\ \boldsymbol{A}_{\hat{p}} & \boldsymbol{I}_f \end{pmatrix}.$$

REFERENCES

1. Farewell, V. T., and Dahlberg, S. Some statistical methodology for the analysis of HLA data. Biometrics 40: 547–560 (1984).
2. Eguchi, S., and Matsuura, M. Testing the Hardy-Weinberg equilibrium in the HLA system. Submitted.
3. Nam, J., and Gart, J. J. Bernstein's and gene-counting methods in generalized ABO-like systems. Ann. Hum. Genet. 39: 361–373 (1976).
4. Yasuda, N., and Kimura, M. A gene-counting method of maximum likelihood for estimating gene frequencies in ABO and ABO-like systems. Ann. Hum. Genet. 31: 409–420 (1968).
5. Ceppellini, R., Siniscalco, M., and Smith, C. A. B. The estimation of gene frequencies in a random-mating population. Ann. Hum. Genet. 20: 97–115 (1955).
6. Smith, C. A. B. Counting methods in genetical statistics. Ann. Hum. Genet. 21: 254–276 (1957).
7. Smith, C. A. B. Notes on gene frequency estimation with multiple alleles. Ann. Hum. Genet. 31: 99–107 (1967).
8. Nam, J. Bias correction for Bernstein's estimator in generalized ABO-like system. In: Proceedings of the Tenth Korea Symposium on Technology, Seoul, Korea. Math. Stat. 9187, pp. 31–35.
9. Albert, E. D., Bauer, M. P., and Mayr, W. R. Histocompatibility Testing in 1984. Springer-Verlag, Berlin, 1984.
10. Stevens, W. L. Statistical analysis of the A-B-O blood groups. Hum. Biol. 22: 191–217 (1950).
11. Haldane, J. B. S. Almost unbiased estimates of functions of frequencies. Sankhyā 17: 201–208 (1956).
12. Fisher, R. A. Statistical Methods for Research Workers, 14th ed. Hafner, New York, 1970.
13. Rao, C. R. Advanced Statistical Methods in Biometric Research. John Wiley and Sons, New York, 1952.
14. Farewell, V. T. The estimation of gene frequency based on a particular mixed leukocyte culture experiment. Biometrics 38: 769–775 (1982).
15. Gart, J. J., and Nam, J. Statistical methods for genetic studies of HLA and cancer. In: Statistical Methods for Cancer Studies (R. G. Cornell, Ed.), Marcel Dekker, New York, 1984, pp. 229–266.
16. Rao, C. R. Linear Statistical Inference and Its Application. John Wiley and Sons, New York, 1973.
17. Smouse, P. E., and Williams, R. C. Multivariate analysis of HLA-disease associations. Biometrics 38: 757–768 (1982).
18. Gart, J. J., and Nam, J. A score test for the possible presence of recessive allels in generalized ABO-like systems. Biometrics 40: 887–894 (1984).
19. Nam, J., and Gart, J. J. The ML estimation and testing of generalized ABO-like data with no observed double recessives. Biometrics 41: 455–466, (1985).
20. Nam, J., and Gart, J. J. On two tests of fit for HLA data with no double blanks. Am. J. Hum. Genet. 41: 71–76 (1987).