MAXIMUM LIKELIHOOD ESTIMATION OF CHROMOSOME FREQUENCIES FROM FAMILY DATA ON MNS BLOOD GROUPS

By RANAJIT CHAKRABORTY and D. C. RAO*

Indian Statistical Institute

SUMMARY. Following Adhikari, Chakraborty and Sarma (1971) maximum likelihood estimates of chromosome frequencies are obtained here from family data on MNS blood groups. The method is illustrated in the last section.

1. Introduction

The Hardy-Weinberg law (HWL) also called the Hardy-Weinberg-Castle Law (Li, 1967), is so powerful a tool that many population geneticists have exploited it thoroughly in many useful applications. A great application indeed is in the estimation of gene frequencies of Mendelian characters. Though the assumption of HWL greatly simplifies the labour involved in estimating the frequencies, it seems reasonable, however, that one should relax this assumption to the extent possible since the HWL does not strictly hold good in any natural population. An attempt to this effect is made in Adhikari, Chakraborty and Sarma (1971). In their work they considered the problem of estimating gene frequencies from family data by the method of maximum likelihood, by slightly relaxing the assumption of HWL. They start with a general set-up at the phenotypic level and make use of the HWL only at the level of dividing the general phenotypic mating frequencies into the corresponding genotypic mating frequencies (as done in Table 1 here). They call such a mating system 'Restricted Random Mating' (RRM). The main purpose of this paper is to present the estimation of chromosome frequencies of MNS-blood groups from family data by the maximum likelihood method under RRM.

One might wonder as to why at all the present method (involving family data) is desirable in view of the simplicity of estimation methods for random samples of individuals. Firstly, it is to be recalled that family data offers better estimates even under HWL. Secondly, it will be interesting to examine how much we really gain by appealing to RRM rather than the celebrated HWL. It is for these two reasons mainly that the present investigation is carried out. However, a comparative treatment will be presented in a subsequent paper.

2. Estimation of parameters

The six phenotypes in MNS-blood group system are denoted by MS, M, MNS, MN, NS and N. Standard notations (e.g., Boyd, 1955) are used throughout this paper which are given below for ready reference :

m : frequency of the M-gene

n : $1-m$: frequency of the N-gene

m_s : frequency of the Ms-chromosome

m_S : $m-m_s$: frequency of the MS-chromosome

n_s : frequency of the Ns-chromosome

n_S $= n-n_s$: frequency of the NS-chromosome

g $= m_s/m$

and d $= n_s/n$

* Present address : Dept. of Prob. and Stat. University of Sheffield, U.K.

The 21 phenotypic mating types, their frequencies λ_i's and the conditional probabilities of the offspring phenotypes are shown in Table 1. The table also includes data from a random sample of G families (and their offspring). The conditional probabilities in Table 1 are constructed as follows :

Consider the mating type $MS \times MS$. This phenotypic mating is split up into the three corresponding genotypic matings in the proportions as shown below.

genotypic mating type given the phenotypic mating $MS \times MS$.	Probability
$MS/MS \times MS/MS$	$(1-g)^2$
$MS/MS \times MS/Ms$	$2g(1-g)$
$MS/Ms \times MS/Ms$	g^2

Observe that the first two genotypic matings give offspring of the phenotype MS with probability 1. The last mating gives two types of offspring MS and M with probabilities 3/4 and 1/4 respectively. Therefore, we get

P(an offspring is of phenotype MS/the phenotypic mating is $MS \times MS$).

$$= (1-g)^2+2g(1-g)+(3/4)g^2$$
$$= 1-g^2/4.$$

Similarly the other probabilities are computed.

TABLE 1. PHENOTYPIC MATING TYPES, THEIR FREQUENCIES AND THE CONDITIONAL PROBABILITIES OF THE OFFSPRING PHENOTYPES (SAMPLE FREQUENCIES ARE SHOWN WITHIN PARENTHESES)

mating		offspring conditional probabilities and observed frequencies						
type	frequency	MS	M	MNS	MN	NS	N	total frequency
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$MS \times MS$	λ_1	$(4-g^2)/4$	$g^2/4$	0	0	0	0	
	$G_1(6)$	$n_1(15)$	$n_2(0)$	$n_3(0)$	$n_4(0)$	$n_5(0)$	$n_6(0)$	$R_1(15)$
$MS \times M$	λ_2	$(2-g)/2$	$g/2$	0	0	0	0	
	$G_2(3)$	$n_7(7)$	$n_8(1)$	$n_9(0)$	$n_{10}(0)$	$n_{11}(0)$	$n_{12}(0)$	$R_2(8)$
$MS \times MNS$	λ_3	$(2-g^2)/4$	$g^2/4$	$(2-gd)/4$	$gd/4$	0	0	
	$G_3(17)$	$n_{13}(21)$	$n_{14}(1)$	$n_{15}(9)$	$n_{16}(3)$	$n_{17}(0)$	$n_{18}(0)$	$R_3(34)$
$MS \times MN$	λ_4	$(2-g)/4$	$g/4$	$(2-g)/4$	$g/4$	0	0	
	$G_4(10)$	$n_{19}(11)$	$n_{20}(3)$	$n_{21}(10)$	$n_{22}(1)$	$n_{23}(0)$	$n_{24}(0)$	$R_4(25)$
$MS \times NS$	λ_5	0	0	$(4-gd)/4$	$gd/4$	0	0	
	$G_5(6)$	$n_{25}(0)$	$n_{26}(0)$	$n_{27}(12)$	$n_{28}(2)$	$n_{29}(0)$	$n_{30}(0)$	$R_5(14)$
$MS \times N$	λ_6	0	0	$(2-g)/2$	$g/2$	0	0	
	$G_6(6)$	$n_{31}(0)$	$n_{32}(0)$	$n_{33}(6)$	$n_{34}(6)$	$n_{35}(0)$	$n_{36}(0)$	$R_6(12)$

TABLE 1 (*contd.*). PHENOTYPIC MATING TYPES, THEIR FREQUENCIES AND THE CONDITIONAL PROBABILITIES OF THE OFFSPRING PHENOTYPES (SAMPLE FREQUENCIES ARE SHOWN WITHIN PARENTHESES)

mating		offspring conditional probabilities and observed frequencies						
type	frequency	*MS*	*M*	*MNS*	*MN*	*NS*	*N*	total frequency
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$M \times M$	λ_7	0	1	0	0	0	0	
	$G_7(0)$	$n_{37}(0)$	$n_{38}(0)$	$n_{39}(0)$	$n_{40}(0)$	$n_{41}(0)$	$n_{42}(0)$	$R_7(0)$
$M \times MNS$	λ_8	$(1-g)/2$	$g/2$	$(1-d)/2$	$d/2$	0	0	
	$G_8(4)$	$n_{43}(0)$	$n_{44}(0)$	$n_{45}(0)$	$n_{46}(9)$	$n_{47}(0)$	$n_{48}(0)$	$R_8(19)$
$M \times MN$	λ_9	0	$1/2$	0	$1/2$	0	0	
	$G_9(7)$	$n_{49}(0)$	$n_{50}(7)$	$n_{51}(0)$	$n_{52}(9)$	$n_{53}(0)$	$n_{54}(0)$	$R_9(16)$
$M \times NS$	λ_{10}	0	0	$(2-d)/2$	$d/2$	0	0	
	$G_{10}(1)$	$n_{55}(0)$	$n_{56}(0)$	$n_{57}(1)$	$n_{58}(0)$	$n_{59}(0)$	$n_{60}(0)$	$R_{10}(1)$
$M \times N$	λ_{11}	0	0	0	1	0	0	
	$G_{11}(5)$	$n_{61}(0)$	$n_{62}(0)$	$n_{63}(0)$	$n_{64}(13)$	$n_{65}(0)$	$n_{66}(0)$	$R_{11}(13)$
$MNS \times MNS$	λ_{12}	$(1-g^2)/4$	$g^2/4$	$(1-gd)/2$	$gd/2$	$(1-d^2)/4$	$d^2/4$	
	$G_{12}(12)$	$n_{67}(8)$	$n_{68}(0)$	$n_{69}(16)$	$n_{70}(2)$	$n_{71}(4)$	$n_{72}(3)$	$R_{12}(33)$
$MNS \times MN$	λ_{13}	$(1-g)/4$	$g/4$	$(2-g-d)/4$	$(g+d)/4$	$(1-d)/4$	$d/4$	
	$G_{13}(15)$	$n_{73}(5)$	$n_{74}(1)$	$n_{75}(15)$	$n_{76}(6)$	$n_{77}(1)$	$n_{78}(7)$	$R_{13}(35)$
$MNS \times NS$	λ_{14}	0	0	$(2-gd)/4$	$gd/4$	$(2-d^2)/4$	$d^2/4$	
	$G_{14}(3)$	$n_{79}(0)$	$n_{80}(0)$	$n_{81}(5)$	$n_{82}(0)$	$n_{83}(2)$	$n_{84}(1)$	$R_{14}(8)$
$MNS \times N$	λ_{15}	0	0	$(1-g)/2$	$g/2$	$(1-d)/2$	$d/2$	
	$G_{15}(10)$	$n_{85}(0)$	$n_{86}(0)$	$n_{87}(6)$	$n_{88}(4)$	$n_{89}(5)$	$n_{90}(7)$	$R_{15}(22)$
$MN \times MN$	λ_{16}	0	$1/4$	0	$1/2$	0	$1/4$	
	$G_{16}(7)$	$n_{91}(0)$	$n_{92}(2)$	$n_{93}(0)$	$n_{94}(11)$	$n_{95}(0)$	$n_{96}(3)$	$R_{16}(16)$
$MN \times NS$	λ_{17}	0	0	$(2-d)/4$	$d/4$	$(2-d)/4$	$d/4$	
	$G_{17}(4)$	$n_{97}(0)$	$n_{98}(0)$	$n_{99}(2)$	$n_{100}(2)$	$n_{101}(1)$	$n_{102}(1)$	$R_{17}(6)$
$MN \times N$	λ_{18}	0	0	0	$1/2$	0	$1/2$	
	$G_{18}(2)$	$n_{103}(0)$	$n_{104}(0)$	$n_{105}(0)$	$n_{106}(2)$	$n_{107}(0)$	$n_{108}(3)$	$R_{18}(5)$
$NS \times NS$	λ_{19}	0	0	0	0	$(4-d^2)/4$	$d^2/4$	
	$G_{19}(0)$	$n_{109}(0)$	$n_{110}(0)$	$n_{111}(0)$	$n_{112}(0)$	$n_{113}(0)$	$n_{114}(0)$	$R_{19}(0)$
$NS \times N$	λ_{20}	0	0	0	0	$(2-d)/2$	$d/2$	
	$G_{20}(4)$	$n_{115}(0)$	$n_{116}(0)$	$n_{117}(0)$	$n_{118}(0)$	$n_{119}(4)$	$n_{120}(4)$	$R_{20}(8)$
$N \times N$	λ_{21}	0	0	0	0	0	1	
	$G_{21}(1)$	$n_{121}(0)$	$n_{122}(0)$	$n_{123}(0)$	$n_{124}(0)$	$n_{125}(0)$	$n_{126}(3)$	$R_{21}(3)$
totals	1							
	$G(123)$	$C_1(77)$	$C_2(15)$	$C_3(82)$	$C_4(70)$	$C_5(17)$	$C_6(32)$	$T(293)$

Observe that the parameters to be estimated are λ_i's ($i = 1, \ldots, 21$), m, g and d, from where we can obtain the chromosome frequencies.

The log-likelihood of the sample (parents and offspring) is easily seen to be

$$\log L = \text{constant} + \log L_1 + \log L_2$$

where,

$$\log L_1 = \sum_{i=1}^{21} G_i \log \lambda_i$$

$$\begin{aligned} \log L_2 = {} & A \log g + B \log d + C \log(1-g) + n_{67} \log(1+g) \\ & + D \log(2-g) + n_1 \log(2+g) + E \log(1-d) \\ & + n_{71} \log(1+d) + F \log(2-d) + n_{113} \log(2+d) \\ & + n_{113} \log(2-g^2) + (n_{15} + n_{81}) \log(2-gd) \\ & + n_{27} \log(4-gd) + n_{69} \log(1-gd) + n_{83} \log(2-d)^2 \\ & + n_{76} \log(g+d) + n_{75} \log(2-g-d). \end{aligned}$$

where

$$\begin{aligned} A &= 2n_2 + n_8 + 2n_{14} + n_{16} + n_{20} + n_{22} + n_{28} + n_{34} + n_{44} + 2n_{68} + n_{70} + n_{74} + n_{62} + n_{88} \\ B &= n_{16} + n_{28} + n_{46} + n_{58} + n_{70} + 2n_{72} + n_{78} + n_{82} + 2n_{84} + n_{90} + n_{100} + n_{102} \\ & \quad + 2n_{114} + n_{120} \\ C &= n_{43} + n_{67} + n_{73} + n_{87} \\ D &= n_1 + n_7 + n_{19} + n_{21} + n_{33} \\ E &= n_{45} + n_{71} + n_{77} + n_{89} \end{aligned}$$

and

$$F = n_{57} + n_{99} + n_{101} + n_{113} + n_{119}.$$

It is to be remembered that the parameters of interest, the λ_i's, g and d are involved only in L_1 and L_2. However the complete expression for L also involves the distribution of the number of children for each mating type and this part will get absorbed in the constant term while taking logarithms. We also assume identical family size distribution for different mating types.

It may be noted that the above partition of L into L_1 and L_2 simplifies the estimation procedure since λ_i's are confined only in L_1 and, g and d are confined only in L_2. Thus, one obtains the maximum likelihood estimates of λ_i's as

$$\hat{\lambda}_i = G_i/G$$

and

$$\hat{V}(\hat{\lambda}_i) = G_i(G-G_i)/G^2 \qquad \ldots \quad (1)$$

$$\widehat{\text{cov}}(\hat{\lambda}_i, \hat{\lambda}_j) = -G_iG_j/G^2 \quad \text{for} \quad i \neq j$$

$$i, j = 1, \ldots, 21.$$

Now turning to the estimation of m, by gene-counting method one obtains

$$m = (M)+(MS)+1/2[(MN)+(MNS)] \qquad \dots \quad (2)$$

where, (M), (MS) etc. are the expected phenotypic proportions among the offspring, which are obtained from Table 1. One can also show that (2) is the maximum likelihood estimate of m. After substituting the expressions for (M), (MS) etc., (2) reduces to

$$\hat{m} = 1/4 \,.\, (4\hat{\alpha}_1+3\hat{\alpha}_2+2\hat{\alpha}_3+\hat{\alpha}_4) \qquad \dots \quad (3)$$

where

$$\begin{aligned} \hat{\alpha}_1 &= \hat{\lambda}_1+\hat{\lambda}_2+\hat{\lambda}_7 \\ \hat{\alpha}_2 &= \hat{\lambda}_3+\hat{\lambda}_4+\hat{\lambda}_8+\hat{\lambda}_9 \\ \hat{\alpha}_3 &= \hat{\lambda}_5+\hat{\lambda}_6+\hat{\lambda}_{10}+\hat{\lambda}_{11}+\hat{\lambda}_{12}+\hat{\lambda}_{13}+\hat{\lambda}_{16} \end{aligned}$$

and

$$\hat{\alpha}_4 = \hat{\lambda}_{14}+\hat{\lambda}_{15}+\hat{\lambda}_{17}+\hat{\lambda}_{18}.$$

$\hat{V}(\hat{m})$ can be obtained from the variance-covariance matrix of $\hat{\lambda}_i$'s given in (1). Note that $\hat{V}(\hat{n}) = \hat{V}(1-\hat{m})$ is the same as $\hat{V}(\hat{m})$.

It remains now to estimate g and d for which it is enough to consider L_2. The estimation is carried out through the well-known iterative method.

In what follows we first start with the initial estimates of g and d, say g_0 and d_0, given by (DeGroot, 1956).

$$g_0 = \sqrt{\overline{C}_2/(\overline{C}_1+\overline{C}_2)}$$

and

$$d_0 = \sqrt{\overline{C}_6/(\overline{C}_5+\overline{C}_6)} \qquad \dots \quad (4)$$

where, $\overline{C}_i$'s are the observed phenotypic proportions of the offspring in the sample of G families (Table 1)

In order to obtain the correction factor to the initial estimates, one obtains

$$L_g = \frac{\partial \log L_2}{\partial g} = \frac{A}{g}-\frac{C}{1-g}+\frac{n_{67}}{1+g}-\frac{D}{2-g}+\frac{n_1}{2+g}-\frac{2gn_{13}}{2-g^2}-\frac{d(n_{15}+n_{81})}{2-gd}$$

$$-\frac{dn_{27}}{4-gd}-\frac{dn_{69}}{1-gd}+\frac{n_{76}}{g+d}-\frac{n_{75}}{2-g-d} \qquad \dots \quad (5)$$

$$\mathrm{L}_d = \frac{\partial \log L_2}{\partial d} = \frac{B}{d}-\frac{E}{1-d}+\frac{n_{71}}{1+d}-\frac{F}{2-d}+\frac{n_{112}}{2+d}-\frac{2dn_{83}}{2-d^2}-\frac{g(n_{15}+n_{81})}{2-gd}$$

$$-\frac{gn_{27}}{4-gd}-\frac{gn_{69}}{1-gd}+\frac{n_{76}}{g+d}-\frac{n_{75}}{2-g-d}$$

where A, B, C, D, E & F are as defined earlier and the information matrix $I = ((I_{ij}))$; $i, j = 1, 2$.

$$I_{11} = \frac{R_2+R_4+R_6}{g(2-g)}+\frac{2R_8+2R_{15}+R_{13}}{4g(1-g)}+\frac{4R_1}{4-g^2}+\frac{2R_3}{2-g^2}+\frac{R_{12}}{1-g^2}+\frac{df}{g}+e, \quad \dots \quad (6)$$

$$I_{12} = f+e,$$

and $$I_{22} = \frac{R_{10}+R_{17}+R_{20}}{d(2-d)}+\frac{2R_{8}+2R_{15}+R_{13}}{4d(1-d)}+\frac{R_{12}}{1-d^2}+\frac{2R_{14}}{2-d^2}+\frac{4R_{19}}{4-d^2}+\frac{gf}{d}+e$$

where $$f = \frac{R_3+R_{14}}{2(2-gd)}+\frac{R_5}{4-gd}+\frac{R_{12}}{2(1-gd)}$$

and $$e = \frac{R_{13}}{4(2-g-d)}+\frac{R_{13}}{4(g+d)}$$

and, R_i's are the totals of offspring for each type of phenotypic mating (Table 1).

By substituting the initial values g_0 and d_0 one can evaluate the likelihood derivatives L_g and L_d and the information matrix, from where the correction factors are obtained as

$$\delta g_0 = V_{11}L_g+V_{12}L_d$$

and $$\delta d_0 = V_{21}L_g+V_{22}L_d$$

where $V = ((V_{ij}))$ is the inverse of I.

Hence the improved estimates (are written as)

$$g_1 = g_0+\delta g_0$$

and $$d_1 = d_0+\delta d_0.$$

This process should be repeated till we obtain stable estimates of g and d. Note that once we get the stable estimates of g and d, say $\hat{g}$ and $\hat{d}$, by then we will have obtained their respective variances V_{11} and V_{22} and also their covariance V_{12} in the last iteration.

Having estimated all the parameters, we now give the maximum likelihood estimates of the chromosome frequencies and their variances (DeGroot, 1956).

$$\begin{aligned}
\hat{m}_s &= \hat{g}\,\hat{m}\\
\hat{n}_s &= \hat{d}\,\hat{n}\\
\hat{m}_S &= (1-\hat{g})\hat{m}\\
\hat{n}_S &= (1-\hat{d})\hat{n}\\
\hat{V}(\hat{m}_s) &= \hat{g}^2\hat{V}(\hat{m})+\hat{m}^2\hat{V}(\hat{g}) \qquad \ldots \quad (7)\\
\hat{V}(\hat{n}_S) &= \hat{d}^2\hat{V}(\hat{n})+\hat{n}^2\hat{V}(\hat{d})\\
\hat{V}(\hat{m}_S) &= (1-\hat{g})^2\hat{V}(\hat{m})+\hat{m}^2\hat{V}(\hat{g})\\
\hat{V}(\hat{n}_S) &= (1-\hat{d})^2\hat{V}(\hat{n})+\hat{n}^2\hat{V}(\hat{d})
\end{aligned}$$

3. Illustration

We shall illustrate here the methods developed in the previous section with the data presented in Table 1. The data were extracted from a series of two earlier publications (Sanger *et al.*, 1948 and Race *et al.*, 1949). However, the estimate of λ_i's and their variance-covariance matrix are not presented here. From (3) we obtain

$$\hat{m} = 0.554878$$

$$\hat{n} = 1-\hat{m} = 0.445122$$

and $$\hat{V}(\hat{m}) = \hat{V}(\hat{n}) = 0.000412.$$

The initial estimates of g and d are given by (4)

$$g_0 = 0.403786$$

and $$d_0 = 0.808122.$$

By using these estimates one evaluates, from (5)

$$L_g = -57.391564$$

and $$L_d = -31.700293$$

and hence,

$$\delta g_0 = -0.139217$$

and $$\delta d_0 = -0.067030.$$

Since the correction factors are quite large, we get the following improved estimates after the completion of one cycle

$$g_1 = 0.264469$$

and $$d_1 = 0.741092.$$

Repeating these computations for five cycles altogether we obtain the final improved estimates, which are presented in Table 2 along with their standard errors. Table 2 also presents the estimates and standard errors of several other parameters of interest.

TABLE 2. MAXIMUM LIKELIHOOD ESTIMATES AND THEIR STANDARD ERRORS

parameter	estimate	standard error
m	0.554878	0.020293
n	0.445122	0.020293
g	0.269478	0.047864
d	0.735158	0.060473
m_s	0.149527	0.027111
m_S	0.405351	0.030414
n_s	0.327235	0.030773
n_S	0.117917	0.027441

ACKNOWLEDGEMENT

We are greatly indebted to Professor C. R. Rao, F.R.S., and Professor B. P. Adhikari for their constant encouragement.

References

Adhikari, B. P., Chakraborty, R. and Sarma, Y. R. (1971) : Estimation of ABO gene frequencies under an assumption of 'restricted random mating,' *Excerpta Medica*. International Congress Series No. 233, 13.

Boyd, W. C. (1955) : Letters to the editor. *Amer. J. Hem. Genet.*, 7, 444-445.

DeGroot, M. H. (1956) : The covariance structure of maximum likelihood gene frequency estimates for the MNS system. *Amer. J. Hum. Genet.*, 8, 229-235.

Li, C. C. (1967) : Castle's early work on selection and equilibrium. *Amer. J. Hum. Genet.*, 19, 70-74.

Race, R. R., Sanger, R., Lawler, S. D. and Bertinshaw, D. (1949) : The inheritance of the MNS blood groups : A second series of families. *Heredity*, 3, 205-213.

Sanger, R., Race, R. R., Walsh, R. J. and Montgomery, C. (1948) : An antibody which subdivides the human MN blood groups. *Heredity*, 2. 131-139.

Paper received : July, 1971.