Theoretical and Applied Genetics 45, 52–58 (1974)

Likelihood Analysis of Geographic Variation in Allelic Frequencies

II. The Logit Model and an Extension to Multiple Loci[1]

Peter E. SMOUSE

Department of Human Genetics, University of Michigan Medical School, Ann Arbor, Michigan (U.S.A.)

Summary. Likelihood estimation and testing procedures are described for treating geographic variation in allelic frequencies as a logistic function of environmental variables. The basic response curve is sigmoidal, and avoids the necessity of invoking environmental thresholds imposed by a linear response model. The one-locus two-allele, one-locus three-alleles, and two-locus two-allele cases are explicitly treated, and the extensions to multiple alleles and loci are indicated. Three cases of geographic variation in gametic frequencies are analyzed to illustrate the utility of these techniques. A biological rationale is given for a sigmoidal response curve, and the utility of the logit model for univariate and multivariate "analysis of variance" is indicated.

Introduction

Likelihood estimation and testing procedures have recently been developed for the analysis of geographic variation in allelic frequencies by Smouse and Kojima (1972), who were concerned with testing the hypothesis that genetic frequencies were correlated with the environment. The particular form of association postulated in that paper was specified by the regression equation:

$$P_i = \beta_0 Z_{0i} + \beta_1 Z_{1i} + \cdots + \beta_K Z_{Ki} \qquad (1)$$

where P_i is the frequency of an allele in the i-th population, the Z's are a set of environmental measures of interest (Z_{0i} is a dummy regression variable of 1 for all populations), and the β's are the usual sort of regression coefficients. The linear model given by (1) is not the only possible choice of functional relationship, but forms a familiar and convenient point of departure for the analysis of pattern in geographic variation (Kojima, *et al.*, 1972). Equation (1) suffers, however, from two limitations.

The first of these is that P_i, as formulated, is not bounded by (0, 1), as must be the case for a probability. This fact necessitates the imposition of thresholds, and the relationship takes the form shown in Fig. 1a. The necessity for thresholds becomes particularly important when several populations exhibit observed frequencies of 1 or 0 (*c. f.* Table 3 of Kojima, *et al.*, 1972). Observed fixation may often arise solely as a result of finite sampling, however, and the imposition of a threshold in such cases is an artifice at best.

The second limitation is that the model is inconvenient for multiple-locus analysis. If the two loci are segregating independently, the likelihood function (except for a combinatorial constant) is:

$$L(\boldsymbol{P}|\boldsymbol{X}) \propto \prod_{i=1}^{I} \prod_{j=1}^{J} \prod_{k=1}^{K} [P_{ij.} P_{i\cdot k}]^{X_{ijk}} \qquad (2)$$

where:

$$P_{ijk} = [P_{ij.} P_{i\cdot k}] = (\boldsymbol{B}_j' \boldsymbol{Z}_i)\,(\boldsymbol{B}_k' \boldsymbol{Z}_i) \qquad (3)$$

is a quadratic equation in the Z's. The index (i) refers to the population; the indices (j) and (k) reference alleles at the A and B loci, respectively. The analysis degenerates to the sum of its single-locus components. If the two loci are not segregating independently, the likelihood function takes the form:

$$L(\boldsymbol{P}|\boldsymbol{X}) \propto \prod_{i=1}^{I} \prod_{j=1}^{J} \prod_{k=1}^{K} P_{ijk}^{X_{ijk}} \qquad (4)$$

where:

$$P_{ijk} = \boldsymbol{B}_{jk}' \boldsymbol{Z}_i \qquad (5)$$

is a linear function in the Z's. Although an appropriate test criterion to distinguish between (2) and (4) is easily constructed, it is difficult to relate the test to any meaningful statement about the β-coefficients. The transition between the two models is rather forced.

The objective of this paper is to suggest an alternative to (1), and to show how it overcomes both of the above difficulties. This alternative is the logistic (or logit) model of Fisher (1935) and Finney (1952).

[1] Supported by AEC AT(11-1)-1552.

The Logit Model

One Locus

Let us define a regression hypothesis, not on P, but rather on lnP:

$$\begin{aligned} lnP_i &= \beta_{10}Z_{0i} + \beta_{11}Z_{1i} + \cdots + \beta_{1k}Z_{ki} = \boldsymbol{B}_1'\boldsymbol{Z}_i \\ ln\ (1 - P_i) &= \beta_{20}Z_{0i} + \beta_{21}Z_{1i} + \cdots + \beta_{2k}Z_{ki} = \boldsymbol{B}_2'\boldsymbol{Z}_i \end{aligned} \tag{6}$$

which may be rewritten either as:

$$ln\left(\frac{P_i}{1 - P_i}\right) = \boldsymbol{A}'\boldsymbol{Z}_i \tag{7}$$

where $\boldsymbol{A} = \boldsymbol{B}_1 - \boldsymbol{B}_2$ or as:

$$\begin{aligned} P_i &= \alpha \cdot \exp\{\boldsymbol{A}'\boldsymbol{Z}_i\}; \\ (1 - P_i) &= \alpha = [1 + \exp\{\boldsymbol{A}'\boldsymbol{Z}_i\}]^{-1}. \end{aligned} \tag{8}$$

This formulation has the advantage that for all real values of the A's and Z's, P_i is bounded by:

$$0 \leqq P_i \leqq 1. \tag{9}$$

The sigmoid form of (7) is shown in Fig. 1b, and has the same general shape as (1), while avoiding arbitrarily sharp thresholds. The logistic model has received much attention in bioassay, and the reader interested in more detail is referred to Cox (1970).

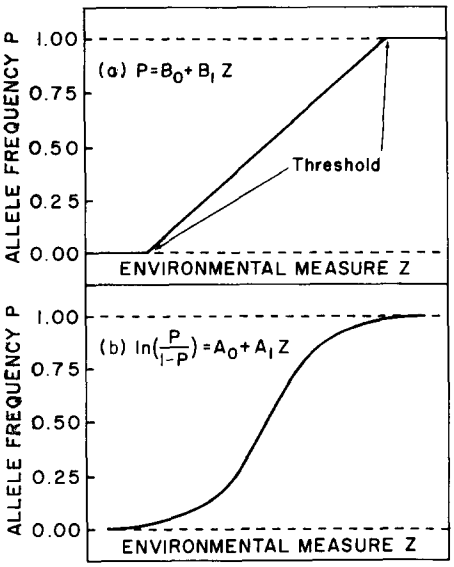

Fig. 1. Regression of Allelic Frequency P on Environmental Measure Z; (a) Linear Model $P = B_0 + B_1Z$; (b) Logistic Model $ln[P/(1 - P)] = A_0 + A_1Z$

The logistic formulation may be extended to multiple-allelic loci. Consider the three-allele case, where the counterpart of (6) is:

$$\begin{aligned} ln\ P_{1i} &= \boldsymbol{B}_1'\boldsymbol{Z}_i;\ ln\ P_{2i} = \boldsymbol{B}_2'\boldsymbol{Z}_i; \\ ln\ (1 - P_{1i} - P_{2i}) &= \boldsymbol{B}_3'\boldsymbol{Z}_i; \end{aligned} \tag{10}$$

which may be rewritten either as:

$$\begin{aligned} ln\left(\frac{P_{1i}}{1 - P_{1i} - P_{2i}}\right) &= \boldsymbol{A}_1'\boldsymbol{Z}_i;\ ln\left(\frac{P_{2i}}{1 - P_{1i} - P_{2i}}\right) \\ &= \boldsymbol{A}_2'\boldsymbol{Z}_i\ ; \end{aligned} \tag{11}$$

where $\boldsymbol{A}_1 = \boldsymbol{B}_1 - \boldsymbol{B}_3$ and $\boldsymbol{A}_2 = \boldsymbol{B}_2 - \boldsymbol{B}_3$, or as:

$$\begin{aligned} P_{1i} &= \alpha \cdot \exp\{\boldsymbol{A}_1'\boldsymbol{Z}_i\};\ P_{2i} = \alpha \cdot \exp\{\boldsymbol{A}_2'\boldsymbol{Z}_i\}; \\ (1 - P_{1i} - P_{2i}) &= \alpha; \end{aligned} \tag{12}$$

with

$$\alpha = [1 + \exp\{\boldsymbol{A}_1'\boldsymbol{Z}_i\} + \exp\{\boldsymbol{A}_2'\boldsymbol{Z}_i\}]^{-1}.$$

To obtain estimates of the A-coefficients, one differentiates the logarithm of the likelihood function with respect to each parameter (A_j), sets each of the resulting equations equal to zero, and solves for the coefficients. For the two allele case, one has:

$$\left\{\frac{\delta lnL}{\delta A_j}\right\}_{\hat{\boldsymbol{A}}} = \sum_{i=1}^{I}{}' \left(\frac{X_i}{P_i} - \frac{(N_i - X_i)}{(1 - P_i)}\right)\frac{\delta P_i}{\delta A_j} = 0 \tag{13}$$

where

$$\left\{\frac{\delta P_i}{\delta A_j}\right\}_{\hat{\boldsymbol{A}}} = Z_{ji}\hat{P}_i\,(1 - \hat{P}_i)\,. \tag{14}$$

The resulting equations may be written in matrix form:

$$\boldsymbol{Z}'\boldsymbol{U}\,[\tilde{\boldsymbol{P}} - \hat{\boldsymbol{P}}] = \boldsymbol{0}\,. \tag{15}$$

The matrix $\boldsymbol{Z}'$ is $(K + 1) \times I$, and contains the independent regression variables for all I populations. $\boldsymbol{U} = \text{diag}\{N_i\}$; $\hat{\boldsymbol{P}}$ is the $I \times 1$ vector of estimated frequencies; $\tilde{\boldsymbol{P}}$ is the $I \times 1$ vector of observed frequencies $\{X_i/N_i\}$; and $\boldsymbol{0}$ is the $I \times 1$ vector of zeroes. Equation (15) is not explicitly solvable in terms of the A_js and one must iterate to a solution.

It can be shown that the matrix of second partial derivatives, evaluated at $\hat{\boldsymbol{P}}$ is:

$$\left\{\frac{\delta^2 lnL}{\delta \boldsymbol{A}\,\delta \boldsymbol{A}'}\right\}_{\hat{\boldsymbol{A}}} = -\,(\boldsymbol{Z}'\hat{\boldsymbol{W}}\boldsymbol{Z}) \tag{16}$$

where $\hat{\boldsymbol{W}} = \text{diag}\{N_i\hat{P}_i\,(1 - \hat{P}_i)\}$. The solution is obtained by iterating

$$\begin{aligned} \hat{\boldsymbol{A}}_{(r+1)} &= \hat{\boldsymbol{A}}_{(r)} + [\boldsymbol{Z}'\hat{\boldsymbol{W}}_{(r)}\boldsymbol{Z}]^{-1}\,[\boldsymbol{Z}'\boldsymbol{U}\,(\tilde{\boldsymbol{P}} - \hat{\boldsymbol{P}}_{(r)})] \\ r &= 0,\ 1,\ 2,\ \ldots\ . \end{aligned} \tag{17}$$

This is the standard Gauss-Newton procedure, and converges to the correct solution, provided $\hat{\boldsymbol{A}}$ is finite, which will usually be the case. Since (16) is strictly negative definite for all $\hat{\boldsymbol{A}}$, there is only one relative extremum, and the solution to (15) is unique.

To obtain estimates of the vectors $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ of the three allele case, one differentiates $ln\,L$ as before, and obtains the following matrix equation:

$$\boldsymbol{Z}'^*\boldsymbol{U}^*\,[\tilde{\boldsymbol{P}}^* - \hat{\boldsymbol{P}}^*] = \boldsymbol{0} \tag{18}$$

where:

$$\boldsymbol{Z}'^* = \begin{bmatrix} \boldsymbol{Z}' & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{Z}' \end{bmatrix}\begin{matrix}(K+1) \\ (K+1)\end{matrix} \qquad \begin{matrix} \text{I} & \text{I} \end{matrix} \tag{19}$$

and:

$$\boldsymbol{U}^* = \begin{bmatrix} \boldsymbol{U} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{U} \end{bmatrix}\begin{matrix}\text{I} \\ \text{I}\end{matrix} \qquad \begin{matrix} \text{I} & \text{I} \end{matrix} \tag{20}$$

with $\boldsymbol{U} = \text{diag}\ \{N_i\}$, and $\tilde{\boldsymbol{P}}^*$ and $\hat{\boldsymbol{P}}^*$ given, respectively, by:

$$\tilde{\boldsymbol{P}}^* = \begin{bmatrix} \tilde{\boldsymbol{P}}_1 \\ \tilde{\boldsymbol{P}}_2 \end{bmatrix} \quad \text{and} \quad \hat{\boldsymbol{P}}^* = \begin{bmatrix} \hat{\boldsymbol{P}}_1 \\ \hat{\boldsymbol{P}}_2 \end{bmatrix} \tag{21}$$

Direct estimates of $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ are not available from (18) and one must iterate to a solution.

The matrix of second partial derivatives for the three allele case is easily shown to be:

$$\left\{\frac{\delta^2 lnL}{\delta \boldsymbol{A}^* \delta \boldsymbol{A}'^*}\right\}_{\hat{\boldsymbol{A}}^*} = -(\boldsymbol{Z}'^* \hat{\boldsymbol{W}}^* \boldsymbol{Z}^*) \tag{22}$$

where $\hat{\boldsymbol{W}}^*$ is given by:

$$\hat{\boldsymbol{W}}^* = \begin{bmatrix} \hat{\boldsymbol{W}}_{11} & \hat{\boldsymbol{W}}_{12} \\ \hat{\boldsymbol{W}}_{21} & \hat{\boldsymbol{W}}_{22} \end{bmatrix} \tag{23}$$

and $\hat{\boldsymbol{W}}_{11} = \text{diag}\ \{N_i \hat{P}_{1i}(1 - \hat{P}_{1i})\}$, $\hat{\boldsymbol{W}}_{22} = \text{diag}\ \{N_i \hat{P}_{2i}(1 - \hat{P}_{2i})\}$, $\hat{\boldsymbol{W}}_{12} = \hat{\boldsymbol{W}}_{21} = \text{diag}\ \{-N_i P_{1i} P_{2i}\}$. This leads to a three-allele counterpart of (17):

$$\hat{\boldsymbol{A}}^*_{(r+1)} = \hat{\boldsymbol{A}}^*_{(r)} + [\boldsymbol{Z}^{*\prime} \hat{\boldsymbol{W}}^*_{(r)} \boldsymbol{Z}^*]^{-1} [\boldsymbol{Z}^{*\prime} \boldsymbol{U}^* (\tilde{\boldsymbol{P}}^* - \hat{\boldsymbol{P}}^*_{(r)})], \quad r = 0, 1, 2, \ldots \tag{24}$$

which converges to the correct solution, provided $\hat{\boldsymbol{A}}^*$ is finite. Since (22) is strictly negative definite for all $\hat{\boldsymbol{A}}^*$, this solution is unique. The likelihood ratio test criteria developed by Smouse and Kojima (1972) are valid for the logistic model developed here as well. The reader is referred to that paper for details.

Two Loci

The two-locus, two-allele case may be treated as a one-locus, four-allele case. The probabilities of the four gametes may be defined as:

$$ln\ P_{11i} = \boldsymbol{B}_1' \boldsymbol{Z}_i;\ ln\ P_{12i} = \boldsymbol{B}_2' \boldsymbol{Z}_i;\ ln\ P_{21i} = \boldsymbol{B}_3' \boldsymbol{Z}_i;\ ln\ P_{22i} = \boldsymbol{B}_4' \boldsymbol{Z}_i \tag{25}$$

which may be alternatively written:

$$\left.\begin{aligned} ln\left(\frac{P_{i11}}{P_{i22}}\right) &= (\boldsymbol{B}_1 - \boldsymbol{B}_4)' \boldsymbol{Z}_i = \boldsymbol{A}_1' \boldsymbol{Z}_i \\ ln\left(\frac{P_{i12}}{P_{i22}}\right) &= (\boldsymbol{B}_2 - \boldsymbol{B}_4)' \boldsymbol{Z}_i = \boldsymbol{A}_2' \boldsymbol{Z}_i \\ ln\left(\frac{P_{i21}}{P_{i22}}\right) &= (\boldsymbol{B}_3 - \boldsymbol{B}_4)' \boldsymbol{Z}_i = \boldsymbol{A}_3' \boldsymbol{Z}_i \end{aligned}\right\} \tag{26}$$

or as the four-allele equivalent of (12). The estimates of $\boldsymbol{A}_1$, $\boldsymbol{A}_2$, and $\boldsymbol{A}_3$ must satisfy a four-allele equivalent of (15) and (18). An iterative scheme similar to (24) will yield the maximum likelihood estimates, which are unique.

The two-locus test criteria for the regression and lack of fit components of the geographic variation are, respectively:

$$\left.\begin{aligned} \Lambda_R(AB) &= -2 \sum_{i=1}^{I} \sum_{j=1}^{2} \sum_{k=1}^{2} X_{ijk} \times \\ &\quad \times [ln\ \bar{P}_{ijk} - ln\ \hat{P}_{ijk}] \sim \chi^2_{3K} \\ \Lambda_L(AB) &= -2 \sum_{i=1}^{I} \sum_{j=1}^{2} \sum_{k=1}^{2} X_{ijk} \times \\ &\quad \times [ln\ \hat{P}_{ijk} - ln\ \tilde{P}_{ijk}] \sim \chi^2_{3(I-K-1)} \end{aligned}\right\} \tag{27}$$

where

$$\left.\begin{aligned} \bar{P}_{ijk} &= X_{ijk} / \sum_{i=1}^{I} N_i, \\ \tilde{P}_{ijl} &= X_{ijk} / N_i, \end{aligned}\right\} \tag{28}$$

and $\hat{P}_{ijk}$ is the estimated value of P_{ijk} under (25). Various sub-hypotheses may be tested by partitioning $\Lambda_R(AB)$ along similar lines to those described by Smouse and Kojima (1972).

The hypothesis that each locus responds independently (in logistic fashion) to the environmental variables is equivalent to the assumption:

$$[\boldsymbol{B}_1 - \boldsymbol{B}_2 - \boldsymbol{B}_3 + \boldsymbol{B}_4] = [\boldsymbol{A}_1 - \boldsymbol{A}_2 - \boldsymbol{A}_3] = \boldsymbol{0}. \tag{29}$$

This is shown as follows. The logit of the marginal probability of the first locus ($P_{i1.}$) is written:

$$\begin{aligned} ln\left(\frac{P_{i1.}}{1 - P_{i1.}}\right) &= ln\left[\frac{P_{i11} + P_{i12}}{P_{i21} + P_{i22}}\right] \\ &= ln\left[\frac{\exp(\boldsymbol{A}_1' \boldsymbol{Z}_i) + \exp(\boldsymbol{A}_2' \boldsymbol{Z}_i)}{\exp(\boldsymbol{A}_3' \boldsymbol{Z}_i) + 1}\right] \end{aligned} \tag{30}$$

In view of (29), this may also be written:

$$\begin{aligned} ln\left(\frac{P_{i1.}}{1 - P_{i1.}}\right) &= ln\left[\exp(\boldsymbol{A}_2' \boldsymbol{Z}_i)\left\{\frac{1 + \exp(\boldsymbol{A}_3' \boldsymbol{Z}_i)}{1 + \exp(\boldsymbol{A}_3' \boldsymbol{Z}_i)}\right\}\right] \\ &= \boldsymbol{A}_2' \boldsymbol{Z}_i \end{aligned} \tag{31}$$

which is logistic. Similarly:

$$ln\left(\frac{P_{i.1}}{1 - P_{i.1}}\right) = \cdots = \boldsymbol{A}_3' \boldsymbol{Z}_i \tag{32}$$

and the two are independent. To show that independence implies (29) it is sufficient to note that the condition for independence is:

$$\frac{P_{i11} P_{i22}}{P_{i12} P_{i21}} = 1 \quad \text{for} \quad i = 1, \ldots, I \tag{33}$$

or alternatively:

$$\left.\begin{aligned} ln\left[\frac{P_{i11} P_{i22}}{P_{i12} P_{i21}}\right] &= (\boldsymbol{B}_1 - \boldsymbol{B}_2 - \boldsymbol{B}_3 + \boldsymbol{B}_4)' \boldsymbol{Z}_i \\ &= (\boldsymbol{A}_1 - \boldsymbol{A}_2 - \boldsymbol{A}_3)' \boldsymbol{Z}_i = \boldsymbol{0} \\ &\text{for} \quad i = 1, 2, \ldots, I, \end{aligned}\right\} \tag{34}$$

which implies (29).

A test of the validity of (29) is obtained by comparing $\Lambda_R(AB)$ of (27) with the sum of the two corresponding single-locus components:

$$\left.\begin{aligned} \Lambda_R(A) &= -2 \sum_{i=1}^{I} \sum_{j=1}^{2} X_{ij.}\ ln\ [\bar{P}_{ij.} - ln\ \hat{P}_{ij.}] \sim \chi^2_K \\ \Lambda_R(B) &= -2 \sum_{i=1}^{I} \sum_{k=1}^{2} X_{i.k}\ ln\ [\bar{P}_{i.k} - ln\ \hat{P}_{i.k}] \sim \chi^2_K \end{aligned}\right\} \tag{35}$$

where the analysis is done separately for each locus, as described above. The test of independence is thus seen to be:

$$[\Lambda_R(AB) - \Lambda_R(A) - \Lambda_R(B)] \sim \chi^2_K. \tag{36}$$

Similar treatment is possible for $\Lambda_T(AB)$, and the resulting analysis is depicted in Table 1. This treat-

Table 1. *Likelihood analysis of geographic variation in two-locus gametic frequencies*

Source	df		χ^2	
Regression	3K		$\Lambda_R(AB)$	
Locus A		K		$\Lambda_R(A)$
Locus B		K		$\Lambda_R(B)$
Interaction		K		$\Lambda_R(AB) - \Lambda_R(A) - \Lambda_R(B)$
Lack of fit	$3(I\text{-}K\text{-}1)$		$\Lambda_L(AB)$	
Among populations	$3(I\text{-}1)$		$\Lambda_T(AB)$	
Locus A		$(I\text{-}1)$		$\Lambda_T(A)$
Locus B		$(I\text{-}1)$		$\Lambda_T(B)$
Interaction		$(I\text{-}1)$		$\Lambda_T(AB) - \Lambda_T(A) - \Lambda_T(B)$

ment may be extended to multiple alleles and loci. The linear restrictions on the B-vectors which are required for independence among loci are more elaborate than (29), but the estimation and test criteria are entirely analogous to those above.

Illustrative examples

One Locus, Two Alleles

In a study of *Drosophila pavani* (Kojima, *et al.*, 1972) a linear regression model of P on latitude, elevation, and season required an environmental threshold for two loci (PGM and PGI). The observed frequencies of the PGM locus are plotted against the regression equation (used as an environmental index) in Fig. 2. For comparison, the data were fitted to a logistic regression model, and the plot of observed frequencies against the altered regression equation (as an alternative environmental index) is shown in Fig. 3. The fact that a whole set of populations (the January collections) have $\tilde{P} = 1$ suggests that the frequency is seasonally high and that finite sampling is responsible for the apparent fixation. The logistic model avoids the assertion that the PGM locus is seasonally "fixed".

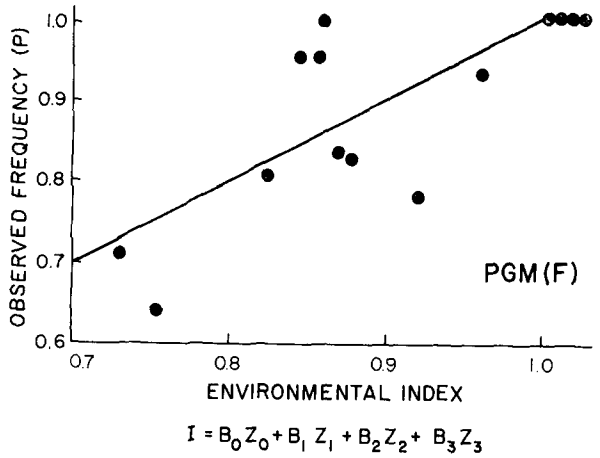

Fig. 2. Observed Frequencies of the Fast-Allele of Phosphoglucomutase (PGM) for Populations of *Drosophila pavani*, Plotted Against an Environmental Index (Estimated Linear Regression Equation) of Elevation (Z_1), Latitude (Z_2), and Season (Z_3)

One Locus, Three Alleles

Genotype-environment relationships have been carefully examined in the harvester ant *Pogonomyrmex barbatus* by means of principal components analysis (Johnson, *et al.*, 1969). Clear associations were found between allelic frequencies and environmental variables. I have examined the frequencies of the alleles of the Esh locus in the following fashion. The locus is a multiple allelic system, but all except three alleles are quite rare. I have pooled allele (6) and these rarer alleles into a single class, and have analyzed the locus as three-allele system. Only those populations with corresponding environmental measurements were used, and the reduced data set consists of twenty-five (25) collections, totalling 4806 alleles. The analysis of geographic variation in allelic frequencies is shown in Table 2. The order of fitting shown is not the only one possible, but is the order for which maximum variation is removed at each stage. The order clearly effects the values of the components.

It is worth noting that the allelic analysis shown above is correct for diploids if one may assume Hardy-Weinberg equilibrium within each population. This assumption is not always justified for this example (Johnson, *et al.*, 1969, Table II), but is not seriously violated except in a few cases. Both the parameter

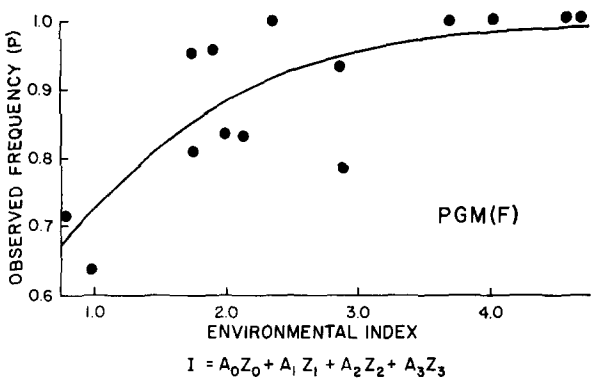

Fig. 3. Observed Frequencies of the Fast-Allele of Phospho glucomutase (PGM) for Populations of *Drosophila pavani*, Plotted Against an Environmental Index (Estimated Logistic Regression Equation) of Elevation (Z_1), Latitude (Z_2), and Season (Z_3)

Table 2. *Likelihood analysis of geographic variation in allelic frequencies at the Esh locus in the harvester ant (Pogonomyrmex barbatus)*

Source	df	χ^2		Accum. χ^2	%
Regression	10	1538.17			
Z_1	2		993.46	993.46	64.6%
$Z_2\|Z_1$	2		418.46	1411.92	91.8%
$Z_3\|Z_1, Z_2$	2		44.72	1456.64	94.7%
$Z_4\|Z_1, Z_2, Z_3$	2		74.52	1531.16	99.6%
$Z_5\|Z_1, Z_2, Z_3, Z_4$	2		7.01	1538.17	100.0%
Lack of fit	38	361.34			
Total	48	1899.51			

Z_1 = Annual precipitation Z_2 = Ave. Jan. temp. Z_3 = Ave. July temp.
Z_4 = Growing season Z_5 = Elevation

Estimated coefficients

	$\hat{A}_0$	$\hat{A}_1$	$\hat{A}_2$	$\hat{A}_3$	$\hat{A}_4$	$\hat{A}_5$
8-allele	+17.91647	+.13388	−.17256	−.12545	−.00648	−.00050
4-allele	−8.41258	+.06431	−.00983	+.12718	−.01637	−.00003

estimates and the test criteria must therefore be viewed as approximate. The sizes of the components in Table 2, however, are so large as to make significance testing pointless.

The regression model accounts for 81% of the total variation among populations. Elevation contributes very little to the description if the other variables are fitted first, and might be deleted with no real loss in information; a model involving annual precipitation, mean January Temperature, mean July Temperature, and Growing Season accounts for 99.6% of the total regression component.

The observed frequencies of allele (8) and allele (4) are plotted against their respective environmental indices (regression equations) in Fig. 4 and 5, respectively. Both of these indices may perhaps best be interpreted as measures of the transition from a subtropical coastal climate to an arid continental climate. The agreement between observed and expected frequencies is by no means perfect (the lack of fit variation is highly significant), but the pattern is nevertheless quite evident. The analysis shown in Table 2 constitutes an informative alternative to the principal components approach of Johnson, *et al.* (1969).

Two Loci, Two Alleles Each

There is a paucity of published two-locus gametic data for large numbers of geographically dispersed populations. While many studies are conducted in such a fashion that multiple-locus genotypes are recorded for each individual, preoccupation with single-locus patterns has precluded publishing multi-

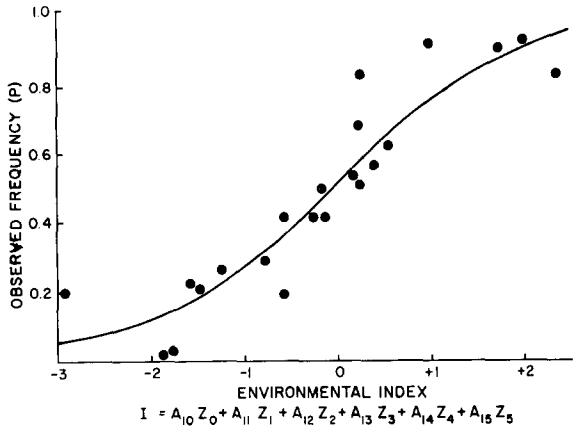

Fig. 4. Observed Frequencies of the (8)-Allele of Esterase-H (Esh) for Populations of *Pogonomyrmex barbatus*, Plotted Against a Logistic Environmental Index (See Table 2 for Z-Variables)

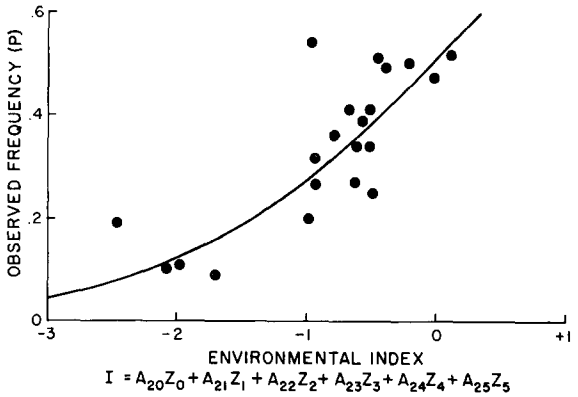

Fig. 5. Observed Frequencies of the (4)-Allele of Esterase-H (Esh) for Populations of *Pogonomyrmex barbatus*, Plotted Against a Logistic Environmental Index (See Table 2 for Z-Variables)

Table 3. *Environmental features and gametic frequencies of the PGM and G6PD loci for populations of Drosophila pavani*

Population location	Month of collection	Latitude	Elevation	Sample size N	Gametic frequencies			
					FF	FS	SF	SS
El Tabo	January	33°28′	25m	104	.000	1.000	.000	.000
Bellavista	January	33°34′	600m	40	.175	.825	.000	.000
San José de Maipo	January	33°39′	967m	98	.051	.949	.000	.000
Melipilla	January	33°41′	164m	102	.020	.980	.000	.000
Volcan	January	33°50′	1800m	92	.152	.783	.022	.043
La Serena	March	29°55′	60m	108	.278	.722	.000	.000
Vicuña	March	30°02′	650m	104	.328	.500	.086	.086
Leyda	March	33°37′	100m	106	.217	.566	.028	.189
San José de Maipo	March	33°39′	967m	104	.183	.654	.058	.107
Capiapó	April	27°34′	381m	108	.083	.630	.000	.287
Vallenar	April	28°44′	384m	80	.000	.638	.000	.362
Rancagua	April	34°10′	500m	108	.306	.648	.000	.046
San Fernando	April	34°35′	334m	108	.241	.713	.046	.000
Santa Cruz	April	34°38′	164m	108	.204	.630	.046	.120
				1370	.1635	.7255	.0219	.0891

FF = PGM(F), G6PD(F); FS = PGM(F), G6PD(S); SF = PGM(S), G6PD(F); SS = PGM(S), G6PD(S)

Table 4. *Likelihood analysis of geographic variation in two-locus gametic frequencies (PGM and G6PD) for Drosophila pavani*

Source	df	χ^2	
Regression	9	198.26***	
PGM	3		100.05***
G6PD	3		43.08***
PGM × G6PD	3		55.13***
Lack of fit	30	222.37***	
Among populations	39	420.63***	
PGM	13		200.89***
G6PD	13		185.15***
PGM × G6PD	13		34.59**

ple locus frequencies. In addition, most studies involve the assay of zygotic genotypes, and gametic frequencies are only indirectly estimable. It should be possible, by means of proper test-crossing, to obtain a gametic assay in many studies, but the desirability and/or utility of such a practice has yet to be established.

The study of *Drosophila pavani* (Kojima, *et al.*, 1972) already mentioned was conducted so that overlapping sets of loci were assayed on individuals. Although the two-locus genotypes were assayed as zygotes, it is possible to estimate two-locus gametic frequencies within a population by means of maximum likelihood procedures. The strategy is to partition the double heterozygotes into coupling and repulsion phases in such a manner that the total zygotic array is best predicted by the resulting two-locus gametic frequencies. These estimated frequencies may be used for the two-locus analysis outlined above, although the test criteria must be viewed as approximations. I have chosen to utilize the PGM and G6PD loci to illustrate the analysis, because the double heterozygotes represent no more than 10% of any given sample. The ambiguities arising from determining the gametic composition of this class are therefore minimal, and the analysis described above should constitute a good first approximation. The estimated gametic frequencies for these two loci are listed for all populations in Table 3, along with the three environmental measures of interest. The two-locus likelihood analysis is shown in Table 4.

The overall regression component accounts for only about 47% of the among populations variation, but is nevertheless highly significant. Considering the coarseness of the environmental variables, the large lack of fit term is not at all surprising. The interaction component of the among populations term is large, and indicates that the two loci do not vary independently over geography. The interaction component of the regression term is also quite large, indicating that the patterns of variation are not independent for the two loci. The sizes of the interaction terms suggest that perhaps more attention should be focused upon multiple-locus gametic patterns of geographic variation than has heretofore been the case, and that some effort toward careful gametic assay is warranted.

Discussion

I have described above the statistical utility of the logistic model. There remains the question of whether a sigmoidal response curve should be expected with allelic frequencies, or whether the linear model proposed by Smouse and Kojima (1972) and described by equation (1) is more appropriate. The type of geographic pattern to be expected depends entirely on the model employed. I shall only describe two

models below, both of the heterotic sort, but others are possible.

Consider first a single locus with two alleles (A and a) and zygotic fitness coefficients $(1 - f_{1i}) : 1 : (1 - f_{2i})$, where f_{1i} and f_{2i} are related to the environment by:

$$\left.\begin{aligned} f_{1i} &= B_{10} + B_1 Z_{1i} + \cdots + B_K Z_{Ki} \\ f_{2i} &= B_{20} - B_1 Z_{1i} - \cdots - B_K Z_{Ki} \,. \end{aligned}\right\} \quad (37)$$

The equilibrium allelic frequency of the i-th population is seen to be:

$$\hat{P}_i = \frac{f_{1i}}{f_{1i} + f_{2i}} = [B_{10} + B_{20})]^{-1} [B_{10} + B_1 Z_{1i} + \cdots + B_K Z_{Ki}] \,, \quad (38)$$

which is the equivalent of equation (1).

Alternatively, consider the selective model given by the zygotic selection coefficients $(1 - e^{-f_{1i}} : 1 : 1 - e^{-f_{2i}})$, where f_{1i} and f_{2i} are related to the environment by:

$$\left.\begin{aligned} f_{1i} &= B_{10} Z_{0i} + B_{11} Z_{1i} + \cdots + B_{1K} Z_{Ki} \\ f_{2i} &= B_{20} Z_{0i} + B_{21} Z_{1i} + \cdots + B_{2K} Z_{Ki} \end{aligned}\right\} \quad (39)$$

and the equilibrium condition for the i-th population is given by the relation:

$$ln\left(\frac{P_i}{Q_i}\right) = (f_{1i} - f_{2i}) = (\boldsymbol{B}_1 - \boldsymbol{B}_2)' \boldsymbol{Z}_i = \boldsymbol{A}' \boldsymbol{Z}_i \,, \quad (40)$$

the logistic model. The reader is referred to Endler (1973) for a discussion of further models.

I have assumed above a complete absence of migration among populations. If migration is added to the first model above, the pattern becomes more sigmoid. The greater the frequency of migration, the greater is the degree of curvature introduced. The second model becomes flatter with migration. Endler (1973) describes the effects of migration on several models. In general, the effect is to yield a sigmoidal pattern, and the logistic model described above should be quite generally useful in the analysis of geographic pattern in allelic frequencies.

The use of the logistic model is not restricted to the type of regression problem described above. The Z-variables may just as easily be chosen to represent the types of "contrast-variables" so familiar in analysis of variance. Using the same general approach described above, one may routinely deal with "analysis of variance" for multinomial situations [gametic frequencies], either at the univariate (one locus, two alleles) or the multivariate (one locus, multiple alleles; or multiple locus) levels.

Literature

Cox, D. R.: The Analysis of Binary Data. London: Methuen 1970.

Endler, J. A.: Gene flow and population differentiation. Science **179**, 243–250 (1973).

Finney, D. J.: Probit Analysis. New York: Cambridge University Press 1952.

Fisher, R. A.: The logic of inductive inference (with discussion). J. Roy. Statist. Soc. **98**, 39–54 (1935).

Johnson, F. M., Schaffer, H. E., Gillaspy, J. E., Rockwood, E. S.: Isozyme genotype-environment relationships in natural populations of the harvester ant, *Pogonomyrmex barbatus* from Texas. Biochem. Genet. **3**, 429–450 (1969).

Kojima, K., Smouse, P., Yang, S., Nair, P. S., Brncic, D.: Isozyme frequency patterns in *Drosophila pavani* associated with geographical and seasonal variables. Genetics **72**, 721–731 (1972).

Smouse, P. E., Kojima, K.: Maximum likelihood analysis of population differences in allelic frequencies. Genetics **72**, 709–719 (1972).

Received October 19, 1973

Communicated by R. W. Allard

Dr. Peter Smouse
Department of Human Genetics
Medical School
The University of Michigan
1137 E. Catherine Street
Ann Arbor, Michigan 48104 (USA)