



Mating and offspring frequencies under partial outcrossing in a structured population

Alan E. Stark

Balgowlah NSW, Australia

Abstract

This paper gives a model of a structured population with respect to an autosomal locus with two alleles. The population reproduces in discrete and non-overlapping generations. The population is assumed to be in equilibrium in that exactly the same distribution of genotypic proportions is reproduced in each generation. The population is subdivided into 'localities' which are characterized by the local gene frequencies. Within each locality the genotypic proportions may depart from Hardy-Weinberg proportions and the same fixation index applies to all localities. The system departs from reality by assuming that the frequency of the first allele follows the beta distribution. However, this enables a convenient way to derive the mating frequencies of parents so that equilibrium is maintained. Wright's F -statistics are applied to characterize the population as a whole. The system is extended to permit an arbitrary level of outbreeding.

Key words: subdivided populations, F -statistics, mating frequencies.

Received: October 26, 2007; Accepted: January 22, 2008.

Introduction

This paper gives a mating system for a structured population with respect to an autosomal locus with two alleles. The population reproduces in discrete and non-overlapping generations and is assumed to be in equilibrium meaning that the population is exactly reproduced in each generation. The simplest model of an unstructured population is that of Weinberg (1908) and Hardy (1908). Stark (2006) and Li (1988) showed that certain forms of non-random mating produce and sustain Hardy-Weinberg proportions. While there were earlier attempts to add realism by going beyond the original model of Hardy and Weinberg it appears that Sewall Wright did most to develop models of structured populations. Some of his ideas are summarized in Wright (1965), which gives references to his earlier work. In that paper Wright wrote "A system was developed for describing the properties of hierarchically subdivided natural populations. Three parameters were proposed in the 1951 paper in terms of a total population (T), subdivisions (S), and individuals (I). F_{IT} is the correlation between gametes that unite to produce the individuals, relative to the gametes of the total population. F_{IS} is the average over all subdivisions of the correlation between uniting gametes relative to those of their own subdivision. F_{ST} is the correlation between random gametes within subdivisions, relative to gametes of the total population."

Send correspondence to Alan E. Stark. 3/20 Seaview Street, 2093 Balgowlah NSW, Australia. E-mail: alans@exemail.com.au.

The principal features contained in the quotation above are included here in that the total population is divided into sub-populations or localities which may be connected in the sense that the two members of a mating pair may come from different localities or may come from the same locality. Each locality is characterized by its gene frequencies. Nei (1987, p. 159) and Weir (1996, p.166) each have a section on fixation indices and so provide some background to the general theme of this paper.

The object here is to develop a compact model so that relations between various parameters can be explored. The most novel feature is that it is possible to give explicit expressions for mating frequencies, an aspect of the whole subject of subdivided populations to which Wright and others have given little attention.

The following sections give the notation used, mating frequencies at local and overall levels, distribution of genotypes from mating, a numerical example and some discussion.

Notation

The two alleles of the autosomal locus are denoted by A and B and the frequency of A within a locality by x and over the whole population by q . The corresponding frequencies of B are of course $1-x$ and $1-q$, respectively. The frequency of A varies over the whole population according to the beta distribution so that the probability of finding a

locality with gene frequency x in a small interval of width dx containing the value x is given by the expression

$$(x^\alpha - 1)(1-x)^{\beta-1}/B(\alpha, \beta)dx, \quad 0 \leq x \leq 1; \quad \alpha, \beta > 0, \quad (1)$$

where $B(\alpha, \beta)$ is the beta function with parameters α, β given by

$$B(\alpha, \beta) = \int u^{\alpha-1} (1-u)^{\beta-1} du,$$

the integral being taken over the interval $(0, 1)$.

The properties of (1) are given in Kendall and Stuart (1977, p. 35). In particular the first four moments about the origin of (1) are:

$$\mu_1 = \frac{\alpha}{(\alpha + \beta)}$$

$$\mu_2 = \frac{\alpha(\alpha + 1)}{(\alpha + \beta)(\alpha + \beta + 1)}$$

$$\mu_3 = \frac{\alpha(\alpha + 1)(\alpha + 2)}{(\alpha + \beta)(\alpha + \beta + 1)(\alpha + \beta + 2)}$$

$$\mu_4 = \frac{\alpha(\alpha + 1)(\alpha + 2)(\alpha + 3)}{(\alpha + \beta)(\alpha + \beta + 1)(\alpha + \beta + 2)(\alpha + \beta + 3)}$$

From these the mean frequency of A over the whole population is given by

$$q = u_1 = \frac{\alpha}{(\alpha + \beta)}$$

and the standard deviation of the distribution of frequency by

$$\sigma_x = \sqrt{\frac{\alpha\beta}{(\alpha + \beta)^2 (\alpha + \beta + 1)}}.$$

Stark (2007) gives a general mating system which can maintain a given departure from the Hardy-Weinberg form. Here a particular case of this is used (see Stark, 1976a, 1976b). If the frequencies of genotypes AA, AB and BB are respectively $f_0 = q^2 + Fpq, f_1 = 2pq - 2Fpq, f_2 = p^2 + Fpq$, ($p = 1 - q$), then the mating frequencies, denoted by f_{ij} , are given by

$$f_{ij} = f_i f_j (1 + \rho d_i d_j / V), \quad (2)$$

where $\rho = 2F/(1+F)$, $d_0 = -2p$, $d_1 = q - p$, $d_2 = 2q$ and $V = 2pq(1 + F)$.

The mating frequencies

It is assumed that the genotypic frequencies within a locality characterized by gene frequency x are:

$$\text{type } AA \quad x^2 + \omega x(1-x);$$

$$\text{type } AB \quad 2x(1-x) - 2\omega x(1-x);$$

$$\text{type } BB \quad (1-x)^2 + \omega x(1-x). \quad (3)$$

Thus ω is the within locality fixation index and corresponds to Wright's F_{IS} . The symbol ω is used here for convenience and to avoid complications is assumed to be non-negative. When a pair of mates is chosen from the same locality it is assumed that the frequency of mating pairs of the various types follow formula (2), with appropriate substitutions, such as ω for F , x for q , $x^2 + \omega x(1-x)$ for f_0 , etc.

On the assumption that the distribution of types within localities follows (3) and that the distribution of gene frequencies follows (1), matings within localities being as given in (2), the overall distribution of types in the whole population can be calculated, and in particular the overall population frequency of A , denoted by q . This will be demonstrated in the next section.

The complete model permits mating between individuals from different localities. In order to maintain the overall population structure the frequencies of pair types will again follow formula (2) but substituting whole population parameters. This will be shown in a later section. A final refinement allows for an arbitrary choice of the proportion of inter-locality matings denoted by χ .

Integrated mating frequencies consistent with a uniform within-locality fixation index

The probability of observing a locality with gene frequency close to x was given as $(x^{\alpha-1}(1-x)^{\beta-1}/B(\alpha, \beta))dx$ in formula (1). To maintain genotype proportions with fixation index ω , as given in (3), apply mating system (2) to parental frequencies (3). The resulting mating frequencies can be integrated using (1) as the weighting factor to obtain the overall mating frequencies in terms of the first four moments of the beta distribution, as follows:

$$AA \times AA \quad \frac{4\omega^3 \mu_1 + \omega^2 (1-\omega)(9-\omega)\mu_2 + 2\omega(1-\omega)^2 (3-\omega)\mu_3 + (1-\omega)^4 \mu_4}{(1+\omega)^2}$$

$$AA \times AB \quad 2(1-\omega) \frac{2\omega^2 \mu_1 + \omega(3-6\omega+\omega^2)\mu_2 + (1-6\omega+7\omega^2-2\omega^3)\mu_3 - (1-\omega)^3 \mu_4}{(1+\omega)^2}$$

$$AA \times BB \quad (1-\omega)^2 \frac{\omega \mu_1 + (1-3\omega+\omega^2)\mu_2 - 2(1-\omega)^2 \mu_3 + (1-\omega)^2 \mu_4}{(1+\omega)^2}$$

$$AB \times AB \quad 4(1-\omega)^2 \frac{\omega \mu_1 + (1-3\omega+\omega^2)\mu_2 - 2(1-\omega)^2 \mu_3 + (1-\omega)^2 \mu_4}{(1+\omega)^2}$$

$$AB \times BB \quad \frac{2(1-\omega) \left[(1+\omega^2)\mu_1 - (3-3\omega+3\omega^2-\omega^3)\mu_2 + (3-6\omega+5\omega^2-2\omega^3)\mu_3 - (1-\omega)^3\mu_4 \right]}{(1+\omega)^2}$$

$$BB \times BB \quad \frac{(1+\omega^2) - 2(2+\omega-\omega^3)\mu_1 + (6-6\omega+3\omega^2-4\omega^3+\omega^4)\mu_2 - 2(2-5\omega+5\omega^2-3\omega^3+\omega^4)\mu_3 + (1-\omega)^4\mu_4}{(1+\omega)^2}$$

In the above, where male and female are of different type, the reciprocal mating also shares the same frequency. Thus $AA \times BB$ and $BB \times AA$ have identical frequency. Note that the frequency of $AB \times AB$ matings is four times that of $AA \times BB$ matings. The resulting matrix of mating frequencies is denoted by L . A numerical example employing this formula in combination with mating frequencies of outbreeding pairs is given in Table 1. The method of calculating tables such as Table 1 is completed in the next section.

Overall genotypic distribution

The application of the mating frequencies given in the preceding section produces the following overall genotypic distribution among offspring:

$$AA \quad \frac{\alpha(\alpha+1+\omega\beta)}{(\alpha+\beta)(\alpha+\beta+1)} \quad (4)$$

$$AB \quad \frac{2(1-\omega)\alpha\beta}{(\alpha+\beta)(\alpha+\beta+1)} \quad (5)$$

$$BB \quad \frac{\beta(\beta+1+\omega\alpha)}{(\alpha+\beta)(\alpha+\beta+1)} \quad (6)$$

The frequency of gene A borne by this distribution is $q = \alpha/(\alpha + \beta)$, that of B is $\beta/(\alpha + \beta)$ and the fixation index is $F_{IT} = (1 + \omega(\alpha + \beta))/(1 + \alpha + \beta)$. Applying the relation $(1-F_{IT}) = (1 - \omega)(1-F_{ST})$ produces $F_{ST} = 1/(1 + \alpha + \beta)$. Thus F_{ST} is seen to be equal to the variance of the (beta) distribution of gene frequencies over localities $\alpha\beta/((\alpha + \beta)^2(\alpha + \beta + 1))$ divided by the product of the overall frequencies of genes A and B . An alternative way of calculating F_{ST} is given in the final section.

The distribution of frequencies given by (4)-(6) can be maintained by outbreeding, that is by forming couples in which the two partners come from different localities. The frequencies of the various pair types are taken from formula (2) using parental frequencies from (4)-(6), gene frequency $q = \alpha/(\alpha + \beta)$ and fixation index $F_{IT} = (1 + \omega(\alpha + \beta))/(1 + \alpha + \beta)$. Denote the matrix of outbreeding mating frequencies by M .

A further refinement is possible by permitting an arbitrary proportion of outbreeding denoted by χ . Combining the two forms of mating yields a mating matrix given by $C = (1 - \chi)L + \chi M$. Since both L and M produce offspring following frequencies (4)-(6), their combination C does the same.

Numerical example

Table 1 gives an example of the system: the parameters are $\alpha = 3$, $\beta = 5$, $\omega = 1/8$ and $\chi = 1/2$. The matrix of overall mating frequencies C is given in Table 1. The overall genotypic proportions are: type AA 37/192; type AB 70/192; type BB 85/192. The overall frequencies of genes A and B are respectively 3/8 and 5/8. Other properties of the population are: variance of the distribution of the A gene frequency is 5/192, $F_{IT} = 2/9$ and $F_{ST} = 1/9$.

Discussion

The quantity F_{ST} can be derived directly in a way which illustrates the following description of Wright given above: “ F_{ST} is the correlation between random gametes within subdivisions, relative to gametes of the total population”. Assign the gametic values 0 and 1 respectively to alleles A and B , as in Table 2. Calculate the probabilities of drawing a pair of genes independently within a locality characterized by allele frequency x , as in Table 2. Calculate the overall frequencies of the respective gene pairs by integration over distribution (1), using the moments of the beta distribution given above to yield the integrated values given in Table 3. Calculate the uncorrected sum of products of genic values from Table 3 to obtain $\beta(\beta+1)/((\alpha+\beta)(\alpha+\beta+1))$. Correct this by deducting the product of overall mean values, namely $\beta^2/(\alpha + \beta)^2$, to derive the corrected sum of products of genic values (covariance of genes in a pair), that is $\alpha\beta/((\alpha+\beta)^2(\alpha+\beta+1))$. Divide the covariance by the product of the standard deviations of the values of the genes in a pair over the whole population. Since the distributions of genic values are identical, this is equal to the variance of either, that is $\alpha\beta/((\alpha+\beta)^2)$. The result is the cor-

Table 1 - Mating scheme with $\alpha = 3$, $\beta = 5$, $\omega = 1/8$, $\chi = 1/2$.

Male x Female	AA	AB	BB
AA	0.0747	0.0823	0.0357
AB	0.0823	0.1427	0.1396
BB	0.0357	0.1396	0.2674

Table 2 - Probabilities of independently-drawn gene pairs in a locality.

Allele and value	A	0	B	1
A 0		x^2		$x(1-x)$
B 1		$(1-x)x$		$(1-x)^2$

Table 3 - Integrated probabilities of independently-drawn gene pairs over the whole population.

	A	0	B	1
0	$\alpha(\alpha+1)/((\alpha+\beta)(\alpha+\beta+1))$		$\alpha\beta/((\alpha+\beta)(\alpha+\beta+1))$	
1	$\alpha\beta/((\alpha+\beta)(\alpha+\beta+1))$		$\beta(\beta+1)/((\alpha+\beta)(\alpha+\beta+1))$	

relation between the values of the gene pair namely $F_{ST} = 1/(1 + \alpha + \beta)$, as noted earlier. Thus F_{ST} is the correlation between the values of a pair of genes drawn from the population that is due just to their sharing membership of the same locality.

A virtue of the model and analysis given here is that they show explicitly how the various quantities such as F_{ST} are expressed in terms of the basic parameters α , β , ω and χ . They show for example that F_{ST} depends only on the parameters of the distribution of gene frequency. They show also that a continuum of overall mating tables defined by χ can sustain the same population structure. Weir (1996, p. 166) discusses the use of an estimate of F_{ST} to examine “population differentiation” in a setting in which estimates of gene frequencies are available from samples of sub-populations. The analysis given here is a pointer as to what an estimate of F_{ST} should reflect. Weir (1996, p. 167) sounds a note of caution in interpreting a typical estimate of F_{ST} by a formula which he gives because it is difficult to assess the significance of sub-population divergence in this way. He recommends a test employing a contingency-table chi-squared statistic rather than F_{ST} . Nei (1987, p.163) mentions some approaches to empirical studies of population differentiation. In some populations there are no clear indications of sub-population boundaries. The formula given here for F_{IT}

shows that this quantity is determined by $\omega(F_{IS})$ as well as by α and β . Therefore, when a population is sampled as a single entity, the calculated fixation index reflects both sub-population differentiation and departure from Hardy-Weinberg proportions at a local level.

References

- Hardy GH (1908) Mendelian proportions in a mixed population. *Science* 28:49-50.
- Kendall MG and Stuart AS (1977) *The Advanced Theory of Statistics*, v 1: Distribution Theory. 4th ed. Charles Griffin & Company Limited, London, 472 pp.
- Li CC (1988) Pseudo-random mating populations. In celebration of the 80th anniversary of the Hardy-Weinberg law. *Genetics* 119:731-737.
- Nei M (1987) *Molecular Evolutionary Genetics*. Columbia University Press, New York, 512 pp.
- Stark AE (1976a) Generalisation of the Hardy-Weinberg law. *Nature* 259:44-44.
- Stark AE (1976b) Hardy-Weinberg law: Asymptotic approach to a generalized form. *Science* 193:1141-1142.
- Stark AE (2006) A clarification of the Hardy-Weinberg law. *Genetics* 174:1695-1697.
- Stark AE (2007) On extending the Hardy-Weinberg law. *Genet Mol Biol* 29:664-666.
- Weinberg W (1908) Über den Nachweis der Vererbung beim Menschen. *Jahresh Verein f vaterl Naturk Württem* 64:368-382. English version: On the demonstration of heredity in Man. In: Boyer SH (ed) *Papers on Human Genetics*. Prentice-Hall, Englewood Cliffs, pp 4-15.
- Weir BS (1996) *Genetic Data Analysis II: Methods for Discrete Population Genetic Data*. Sinauer Associates, Inc. Publishers, Sunderland, 445 pp.
- Wright S (1965) The interpretation of population structure by F-statistics with special regard to systems of mating. *Evolution* 19:395-420.

Associate Editor: Paulo A. Otto