Some Statistical Methods for the Simple Assortative Mating Problem

D. S. Robson *
Cornell University

BU-94-M
find year

Introduction

A breeding population composed of a fraction p of type A individuals and a fraction q=1-p of type B individuals, each type containing males and females in a 1:1 ratio, is said to be undergoing assortative mating if the expected frequencies of the matings AxA, AxB, BxA, and BxB differ from the frequencies p^2, pq, qp, and q^2 expected in a completely random mating population. Simple assortative mating may be characterized by three parameters p, α, and β, where α and β are defined by

Prob (a given type A male will mate with a type A female)=$p+\alpha p$

and

Prob (a given type B male will mate with a type B female)=$q+\beta q$.

The expected and observed frequencies of the four kinds of crosses may then be expressed as:

	Kind of mating				
	AxA	AxB	BxA	BxB	Total
Theoretic frequency	$p^2(1+\alpha)$	$pq-\alpha p^2$	$pq-\beta q^2$	$q^2(1+\beta)$	1
Observed number	n_{aa}	n_{ab}	n_{ba}	n_{bb}	n

The statistical problem considered here is to construct estimators and hypothesis testing procedures for p, α, and β, based upon the observed numbers n_{aa}, n_{ab}, and n_{bb} of the four kinds of crosses in a random sample of n crosses.

Estimation

The maximum likelihood estimators, obtained by maximizing the probability of the sample observations

*Paper ~~No. 361~~ Plant Breeding Department, and No. BU-47, Biometrics Unit, Plant Breeding Department, Cornell University.

$$\mathrm{Prob}(n_{aa},n_{ab},n_{ba},n_{bb})$$

$$= \frac{n!}{n_{aa}!n_{ab}!n_{ba}!n_{bb}!}[p^2(1+\alpha)]^{n_{aa}}[pq-\alpha p^2]^{n_{ab}}[pq-\beta q^2]^{n_{ba}}[q^2(1+\beta)]^{n_{bb}}$$

as a function of p, α, and β, are (see appendix)

$$\hat{p} = \frac{n_{aa}+n_{ab}}{n}$$

$$\hat{\alpha} = \frac{nn_{aa}}{(n_{aa}+n_{ab})^2} - 1$$

$$\hat{\beta} = \frac{nn_{bb}}{(n_{bb}+n_{ba})^2} - 1 \quad .$$

The estimator $\hat{p}$ is well defined, and $n\hat{p}$ is binomially distributed with mean np and variance npq. The estimators $\hat{\alpha}$ and $\hat{\beta}$, however, are undefined when $n_{aa}+n_{bb}=0$ or $n_{bb}+n_{ba}=0$, respectively, so their sampling distributions are likewise undefined in the small sample case. As the sample size n gets large, the probability of the event $n_{aa}+n_{ab}=0$ or $n_{bb}+n_{ba}=0$ approaches zero, and the asymptotic joint distribution of $\hat{p}$, $\hat{\alpha}$, and $\hat{\beta}$ is a multivariate normal distribution with variances and covariances given by (see appendix)

$$\mathrm{Var}(\hat{p}) = \frac{pq}{n}$$

$$\mathrm{Var}(\hat{\alpha}) = \frac{1+\alpha}{np^2}[1-p^2(1+\alpha)]$$

$$\mathrm{Var}(\hat{\beta}) = \frac{1+\beta}{nq^2}[1-q^2(1+\beta)]$$

$$\mathrm{Cov}(\hat{p},\hat{\alpha}) = -\frac{q(1+\alpha)}{n}$$

$$\mathrm{Cov}(\hat{p},\hat{\beta}) = \frac{p(1+\beta)}{n}$$

$$\mathrm{Cov}(\hat{\alpha},\hat{\beta}) = - \frac{(1+\alpha)(1+\beta)}{n}$$

Asymptotic normality of the estimators also provides a basis for constructing confidence intervals; the large sample 95% confidence interval estimators of p, α, and β are

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

$$\hat{\alpha} \pm 1.96\sqrt{\frac{1+\hat{\alpha}}{n\hat{p}^2}\left[1-\hat{p}^2(1+\hat{\alpha})\right]}$$

$$\hat{\beta} \pm 1.96\sqrt{\frac{1+\hat{\beta}}{n\hat{q}^2}\left[1-\hat{q}^2(1+\hat{\beta})\right]}$$

The large sample interval estimate of p may, of course, be replaced by the exact interval as tabulated from the binomial distribution.

Estimation when reciprocal crosses are not distinguished

A slightly modified version of this problem is encountered when the sexes are superficially indistinguishable or, for some other reason, the reciprocal crosses AxB and BxA are not separated in the sample. In this case the full model difined by p, α, and β is not identifiable; i.e., the information in the sample is not sufficient for estimating the parameters of the model. If the model is simplified by adding the restriction that α and β are equal, say, to some common value Θ, then identifiability is restored and maximum likelihood estimators of p and Θ are easily constructed (see appendix):

$$p^* = \frac{n_{aa} - \sqrt{n_{aa}n_{bb}}}{n_{aa} - n_{bb}}$$

$$\Theta^* = \frac{2p^*q^*n - m_{ab}}{(1-2p^*q^*)n}$$

where $m_{ab}(=n_{ab}+n_{ba})$ is the observed number of crosses in which one of the mates is type A and the other type B. The joint asymptotic distribution of p^* and Θ^* is the bivariate normal distribution with means p and Θ, respectively, and variances and covariance (see appendix):

$$\mathrm{Var}(p^*) = \frac{p^2+q^2}{4n(1+\Theta)}$$

$$\mathrm{Var}(\Theta^*) = \frac{1-\Theta^2}{n}$$

$$\mathrm{Cov}(p^*,\Theta^*) = \frac{q-p}{n}$$

Large sample confidence interval estimates may then be constructed as before.

Tests of hypotheses

Large sample test procedures for this problem may be obtained by the well known likelihood ratio method or, when this is not convenient, directly from the asymptotic normality of the maximum likelihood estimators. The hypothesis that $\alpha=\beta=0$, for example, may be conveniently tested by the likelihood ratio method; that is,

$$-2 \log \frac{\max\limits_{p} \left\{ [p^2]^{n_{aa}} [pq]^{n_{ab}+n_{ba}} [q^2]^{n_{bb}} \right\}}{\max\limits_{p,\alpha,\beta} \left\{ [p^2(1+\alpha)]^{n_{aa}} [pq-\alpha p^2]^{n_{ab}} [pq-\beta p^2]^{n_{ba}} [q^2(1+\beta)]^{n_{bb}} \right\}}$$

$$=-2 \log \frac{[\tilde{p}^2]^{n_{aa}} [\tilde{p}\tilde{c}]^{n_{ab}+n_{ba}} [\tilde{q}^2]^{n_{bb}}}{[\hat{p}^2(1+\hat{\alpha})]^{n_{aa}} [\hat{p}\hat{q}-\hat{\alpha}\hat{p}^2]^{n_{ab}} [\hat{p}\hat{q}-\hat{\beta}\hat{q}^2]^{n_{ba}} [\hat{q}^2(1+\hat{\beta})]^{n_{bb}}}$$

is approximately distributed as chi-square with 2 degrees of freedom when n is large, where

$$\tilde{p} = \frac{2n_{aa}+n_{ab}+n_{ba}}{2n} .$$

In the reduced problem where only three kinds of crosses are distinguished and $\alpha=\beta=\theta$ is an apriori assumption, the above hypothesis becomes, simply $\theta=0$. A test statistic for this hypothesis is

$$t = \theta^{*} \sqrt{n}$$

which is asymptotically distributed as the standard normal deviate when the hypothesis $\theta=0$ is true.

The hypothesis $\alpha=\beta$ for the original problem cannot be conveniently tested by the likelihood ratio method because the maximum likelihood estimator of the common value $\alpha=\beta=\theta$ can only be computed iteratively. Several alternative, asymptotically equivalent test statistics are available in this case, each of which has the same asymptotic distribution as

$$\frac{\hat{\alpha}-\hat{\beta}}{\sqrt{\frac{1+\theta}{n}\left(\frac{1}{p^2}+\frac{1}{q^2}\right)}} ,$$

The parameters in the denominator can be estimated in various ways; for example, two estimators $\hat{p}$ and p^* are already available for the parameter p. The choice from available estimators influences only the rate of approach to normality, and here intuition suggests that the best choice outside of the tedious iterative estimates is to sacrifice that information contained in n_{ab} and n_{ba} which is not also contained in $m_{ab}=n_{ab}+n_{ba}$, and use the estimators p^* and θ^*. Thus, the test statistic for the hypothesis $\alpha=\beta$ becomes

$$\frac{p^*q^*\sqrt{n}\;(\hat{\alpha}-\hat{\beta})}{\sqrt{(1+\theta^*)(1-2p^*q^*)}}$$

which is asympotically distributed as the standard normal deviate when $\alpha=\beta$.

Similar problems of choice arise when two samples are available, such as samples from two locations or two years, and hypotheses concerning homogeneity of the total population are to be tested. In two different years, for example, the parameters p_1 and p_2 may be expected to differ if assortative mating is taking place, but the hypothesis that $\alpha_1=\alpha_2$ and $\beta_1=\beta_2$ may be quite reasonable. A test of this hypothesis by the likelihood ratio method requires the iterative estimation of the common values of $\alpha_1=\alpha_2$ and $\beta_1=\beta_2$. An alternative is to test the hypotheses $\alpha_1=\alpha_2$ and $\beta_1=\beta_2$ separately by the asymptotically standard normal test statistics

$$\frac{\hat{\alpha}_1-\hat{\alpha}_2}{\sqrt{\frac{1+\hat{\alpha}_1}{n\hat{p}_1^2}\,[1-\hat{p}_1^2(1+\hat{\alpha}_1)]+\frac{1+\hat{\alpha}_2}{n\hat{p}_2^2}\,[1-\hat{p}_2^2(1+\hat{\alpha}_2)]}}$$

$$\frac{\hat{\beta}_1-\hat{\beta}_2}{\sqrt{\frac{1+\hat{\beta}_1}{n\hat{q}_1^2}[1-\hat{q}_1^2(1+\hat{\beta}_1)]+\frac{1+\hat{\beta}_2}{n\hat{q}_2^2}\,[1-\hat{q}_2^2(1+\hat{\beta}_2)]}}$$

When only three kinds of crosses are distinguished in the sample this hypothesis becomes $\theta_1=\theta_2$ and the corresponding test statistic would be

$$\frac{\theta_1^*-\theta_2^*}{\sqrt{\frac{1-\theta_1^{*2}}{n_1}+\frac{1-\theta_2^{*2}}{n_2}}}$$

In this case, however, an estimate of the common value $\bar{\theta}=\theta_1=\theta_2$ may be obtained explicitly as the weighted average of θ_1^* and θ_2^* ; i.e.,

$$\bar{\theta}^*=\frac{\frac{\theta_1^*}{\mathrm{Var}(\theta_1^*)}+\frac{\theta_2^*}{\mathrm{Var}(\theta_2^*)}}{\frac{1}{\mathrm{Var}(\theta_1^*)}+\frac{1}{\mathrm{Var}(\theta_2^*)}}=\frac{n_1\theta_1^*+n_2\theta_2^*}{n_1+n_2}$$

Consequently,

$$\frac{\theta_1^*-\theta_2^*}{\sqrt{(1-\bar{\theta}^{*2})(\frac{1}{n_1}+\frac{1}{n_2})}}$$

is a better large sample test statistic in the sense that its distribution approaches the standard normal distribution at a faster rate.

Appendix

The likelihood function of the observations n_{aa}, n_{ab}, and n_{ba}, and n_{bb} is, except for a factor depending only upon the observations,

$$L=[p^2(1+\alpha)]^{n_{aa}}[pq-\alpha p^2]^{n_{ab}}[pq-\beta q^2]^{n_{ba}}[q^2(1+\beta)]^{n_{bb}} \quad .$$

The maximizing values $\hat{p}$, $\hat{\alpha}$, $\hat{\beta}$ may be obtained by the usual device of equating to zero the partial derivatives of log L; thus,

$$\frac{\delta \log L}{\delta \alpha} = 0 = \frac{n_{aa}}{1+\hat{\alpha}} - \frac{\hat{p}^2 n_{ab}}{\hat{p}\hat{q}-\hat{\alpha}\hat{p}^2} \qquad (1)$$

$$\frac{\delta \log L}{\delta \beta} = 0 = \frac{n_{bb}}{1+\hat{\beta}} - \frac{\hat{q}^2 n_{ba}}{\hat{p}\hat{q}-\hat{\beta}\hat{q}^2} \qquad (2)$$

$$\frac{\delta \log L}{\delta p} = 0 = \frac{2n_{aa}}{\hat{p}} + \frac{(1-2\hat{p}-2\hat{\alpha}\hat{p})n_{ab}}{\hat{p}\hat{q}-\hat{\alpha}\hat{p}^2} - \frac{(1-2\hat{q}-2\hat{\beta}\hat{q})n_{ba}}{\hat{p}\hat{q}-\hat{\beta}\hat{q}^2} - \frac{2n_{bb}}{\hat{q}} \qquad (3)$$

Substitution of the relations (1) and (2) into (3) then gives

$$\frac{n_{aa}}{\hat{p}^2(1+\hat{\alpha})} = \frac{n_{bb}}{\hat{q}^2(1+\hat{\beta})}$$

so the three likelihood equations imply that in this case the maximum likelihood estimators are identical to those obtained by equating the observations to their expected values; i.e.,

$$\frac{n_{aa}}{\hat{p}^2(1+\hat{\alpha})} = \frac{n_{ab}}{\hat{p}\hat{q}-\hat{\alpha}\hat{p}^2} = \frac{n_{ba}}{\hat{p}\hat{q}-\hat{\beta}\hat{q}^2} = \frac{n_{bb}}{\hat{q}^2(1+\hat{\beta})} = n \quad .$$

The solution of this set of equations is easily seen to be the estimators given in the text.

The variance-covariance matrix V of the asymptotic multivariate distribution of these estimators is the reciprocal of the information matrix,

$$V^{-1} = \begin{pmatrix} -E\,\frac{\delta^2 \log L}{\delta p^2} & -E\,\frac{\delta^2 \log L}{\delta p \delta \alpha} & -E\,\frac{\delta^2 \log L}{\delta p \delta \beta} \\ -E\,\frac{\delta^2 \log L}{\delta \alpha \delta p} & -E\,\frac{\delta^2 \log L}{\delta \alpha^2} & -E\,\frac{\delta^2 \log L}{\delta \alpha \delta \beta} \\ -E\,\frac{\delta^2 \log L}{\delta \beta \delta p} & -E\,\frac{\delta^2 \log L}{\delta \beta \delta \alpha} & -E\,\frac{\delta^2 \log L}{\delta \beta^2} \end{pmatrix}$$

$$= \begin{pmatrix} n\left(\frac{1}{pq-\alpha p^2} + \frac{1}{pq-\beta q^2}\right) & \frac{np^2}{pq-\alpha p^2} & -\frac{nq^2}{pq-\beta q^2} \\ \frac{np^2}{pq-\alpha p^2} & \frac{np^3}{(1+\alpha)(pq-\alpha p^2)} & 0 \\ -\frac{nq^2}{pq-\beta q^2} & 0 & \frac{nq^3}{(1+\beta)(pq-\beta q^2)} \end{pmatrix}$$

The determinant of this information matrix is

$$\frac{n^3 p^2 q^2}{(1+\alpha)(1+\beta)(pq-\alpha p^2)(pq-\beta q^2)} ,$$

from which it follows that the elements of V are the variances and covariances given in the text.

For the reduced problem where $\alpha=\beta=\theta$ and $m_{ab}=n_{ab}+n_{ba}$ the likelihood function is proportional to

$$L = [p^2(1+\theta)]^{n_{aa}}[2pq-\theta(p^2+q^2)]^{m_{ab}}[q^2(1+\theta)]^{n_{bb}}$$

so that

$$\frac{\delta \log L}{\delta\Theta} = 0 = \frac{n_{aa}}{1+\Theta^*} - \frac{m_{ab}(1-2p^*q^*)}{2p^*q^*-\Theta^*(1-2p^*q^*)} \tag{4}$$

and

$$\frac{\delta \log L}{\delta p} = 0 = \frac{n_{aa}}{p^*} + \frac{m_{ab}(1-2p^*)(1+\Theta^*)}{2p^*q^*-\Theta^*(1-2p^*q^*)} - \frac{n_{bb}}{q^*} \tag{5}$$

Substitution of the relation (4) into (5) gives the quadratic equation in p *

$$p^{*2}(n_{aa}-n_{bb})-2n_{aa}p^*+n_{aa}=0$$

with the two roots

$$p^* = \frac{n_{aa}+\sqrt{n_{aa}n_{bb}}}{n_{aa}-n_{bb}}, \qquad p^* = \frac{n_{aa}-\sqrt{n_{aa}n_{bb}}}{n_{aa}-n_{bb}}$$

the first of which always exceeds unity. Equation (4) may then be solved directly for Θ* in terms of p*.

The variance covariance matrix of the limiting bivariate normal distribution of p* and Θ* is given by

$$V = \begin{pmatrix} -E\,\dfrac{\delta^2 \log L}{\delta p^2} & -E\,\dfrac{\delta^2 \log L}{\delta p\delta\Theta} \\ \\ -E\,\dfrac{\delta^2 \log L}{\delta\Theta\delta p} & -E\,\dfrac{\delta^2 \log L}{\delta\Theta^2} \end{pmatrix}^{-1}$$

$$= \begin{pmatrix} \dfrac{4n(1-\theta^2)}{2pq-\theta(p^2+q^2)} & -\dfrac{2n(q-p)}{2pq-\theta(p^2+q^2)} \\ \\ \dfrac{2n(q-p)}{2pq-\theta(p^2+q^2)} & \dfrac{n(p^2+q^2)}{(1+\theta)[2pq-\theta(p^2+q^2)]} \end{pmatrix}^{-1}$$

$$= \begin{pmatrix} \dfrac{p^2+q^2}{4n(1+\theta)} & \dfrac{q-p}{n} \\ \\ \dfrac{q-p}{n} & \dfrac{1-\theta^2}{n} \end{pmatrix}$$