Mathematical Population Studies, 13:83–103, 2006

ISSN: 0889-8480 print/1543-5253 online
DOI: 10.1080/08898480600620027

Routledge
Taylor & Francis Group

Estimating Allele Frequencies and Inbreeding Coefficients in *K*-Allele Models

Prakash Gorroochurn
Division of Statistical Genetics, Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, New York, USA

Susan E. Hodge
Division of Statistical Genetics, Department of Biostatistics, Mailman School of Public Health, Columbia University, and Clinical-Genetic Epidemiology Unit, New York State Psychiatric Institute, New York, New York, USA

Some of the methods of estimation of allele frequencies and inbreeding coefficients in a K-allele model are examined. A result that has long been assumed to be true is proved. That is, in the presence of inbreeding, the maximum likelihood estimators of the allele frequencies and of the inbreeding coefficient f do not in general equal their observed (or sample) values (except when $K = 2$). A least-squares way of looking at the estimation problem is presented, and simulations are used to compare the three types of estimators (sample, maximum likelihood, and least-squares) in a 3-allele model. Probability generating functions are used to derive exact expressions for the bias of the sample estimator of f in a 2-allele model for any sample size, and those biases are calculated for a number of situations. Finally, an approximately unbiased estimator of the inbreeding coefficient when an allele is rare or common is proposed, and its bias is compared with that of the sample estimator and with that of an estimator proposed by Weir (1996).

Keywords: Hardy-Weinberg equilibrium; maximum likelihood; least-squares; probability generating function

INTRODUCTION

Inbreeding coefficients have played a prominent role in the population genetics literature. These coefficients were first put forward by Wright

Address correspondence to P Gorroochurn, Columbia University, Department of Biostatistics, Rm 620, 722 W 168th Street, New York, NY 10032. E-mail: pg2113@columbia.edu

(1922) as measures of "relatedness" between individuals in a reference population. Usually, inbreeding arises from nonrandom mating between closely related individuals and results in an excess of homozygotes over heterozygotes, relative to proportions under Hardy-Weinberg equilibrium (HWE); however, allele frequencies are not changed. Moreover, mating between close relatives is not the only cause for inbreeding. In fact, any small population has some degree of inbreeding associated with it, due to the sharing of common ancestors.

Wright (1922) originally defined the inbreeding coefficient with respect to a single locus as the correlation between two genes in a uniting gamete by assigning numerical values to the genes. An alternative definition was given by Haldane and Moshinsky (1939), Cotterman (1940), and Malécot (1948), who used the concept of *identity by descent* (IBD). Two homologous alleles at a locus are said to be IBD if they are both derived from the same allele copy in a common ancestor (assuming that the common ancestor is so recent in the past that there are no intervening mutations). The inbreeding coefficient f is then defined as the probability that two homologous alleles are IBD. We shall define f as the proportionate reduction in heterozygozity relative to HWE:

$$f = 1 - \frac{\sum_{(i,j):i<j} P_{ij}}{1 - \sum_i p_i^2}, \tag{1}$$

where p_i is the proportion of allele i, and P_{ij} is the frequency of genotype A_iA_j, such that $\sum_{(i,j):i\leq j} P_{ij} = \sum_i p_i = 1$. For two alleles, A_1 and A_2 having respective frequencies p and $q(=1-p)$, Eq. (1) is equivalent to:

$$\begin{aligned} P_{11} &= p^2 + pqf, \\ P_{12} &= 2pq - 2pqf, \\ P_{22} &= q^2 + pqf, \end{aligned} \tag{2}$$

where $P_{11} + P_{12} + P_{22} = 1$. Eq. (2) suggests that we can also view f as a standardized measure of the deviation from HWE proportions (Nagylaki, 1998). For a general K-allele with inbreeding coefficient f, we thus have

$$P_{ij} = \begin{cases} p_i^2 + p_i(1-p_i)f, & i = j, \\ 2p_ip_j - 2p_ip_jf, & i \neq j, \end{cases} \tag{3}$$

From Eq. (3), we see that $-p_i/(1-p_i) \leq f \leq 1$ for $i = 1, 2, \ldots, K$. Another approach is to view f as genotype-specific (i.e., there are as many f_{ij}'s as there are genotypes), but it is often reasonable to use (as we do here) a single inbreeding coefficient as the loci are unlikely to be under the influence of selection (Weir, 2001). Both here and in what follows, we assume that the alleles are co-dominant.

In general, any deviation from HWE (whether due to inbreeding or not) is measured by the fixation index (F), a term first introduced by Wright (1951, 1965). Even more generally, Wright's three fixation indices (F_{IS}, F_{IT} and F_{ST}), or F- statistics, have been used to quantify deviations from HWE when populations are hierarchically structured (e.g., Excoffier (2001)). In such a framework, F_{IS} corresponds to the inbreeding coefficient f considered here.

Early authors have done much work involving estimation, including Li and Horvitz (1953), Yasuda (1968), and Robertson and Hill (1984), and Rousset and Raymond (1995). For example, Li and Horvitz propose five methods for estimating f. They show that, in a 2-allele model with inbreeding, the maximum likelihood estimates (MLEs) of p and f are equal to their sample (or observed) values; however, they do not provide explicit expressions for the p_i's and f when $K \geq 3$. They also conjecture that the MLEs of the p_i's do not equal their sample values. The same conjecture has been made by other Curie-Cohen (1982), Robertson and Hill (1984), and others, but has not to our knowledge been proven.

We prove that the MLEs of the p_i's do not equal their sample values in the section Maximum Likelihood Estimation. We also examine some of the methods of estimation of the p_i's and f. In the section, A Least-Squares Approach to the Estimation Problem, we evaluate a least-squares method of estimation of the p_i's and f, and we use simulations in the next section to compare the three estimators (sample, MLE and LSE) in a 3-allele model. In the section on exact bias, we use probability generating functions to derive exact expressions for the bias of the sample estimator of f in a two-allele model for any sample size. In the final section, we derive an approximately unbiased estimator of f when p is very small (or very large) and compare its performance with other estimators.

STATISTICAL MODEL AND ESTIMATION

In general, for a K-allele model, suppose the genotype A_iA_j has observed count x_{ij} and true proportion P_{ij} in the population, where $1 \leq i \leq j \leq K$. Then the joint distribution of the vector $\mathbf{x} = \{x_{ij}, 1 \leq i \leq j \leq K\}$ is given by the multinomial distribution:

$$p(\mathbf{x}) = \frac{n!}{\prod_{(i,j):i\leq j} x_{ij}!} \prod_{(i,j):i\leq j} P_{ij}^{x_{ij}}, \tag{4}$$

where $n = \sum_{(i,j):i\leq j} x_{ij}$ is the sample size (or total number of genotypes), the P_{ij}'s are given in Eq. (3) and $\sum_{(i,j):i\leq j} P_{ij} = 1$. The distribution in

Eq. (4) is strictly valid only for an infinitely large population. From standard statistical theory, marginally each x_{ij} has a binomial distribution:

$$x_{ij} \sim \text{bino}\left(n, P_{ij}\right). \tag{5}$$

Moreover, the population proportion of allele i is

$$p_i = P_{ii} + \frac{1}{2}\sum_{j:j\neq i} P_{ij},$$

where we define $\sum_{j:j\neq i}(\cdot)_{ij} \equiv \sum_{j:j<i}(\cdot)_{ji} + \sum_{j:j>i}(\cdot)_{ij}$. The estimation of the parameters p_i and f has been a central and natural problem in the literature. The p_i's are usually estimated by using the sample (i.e., observed) proportions:

$$p_i^{(s)} = \frac{2x_{ii} + \sum_{j:j\neq i} x_{ij}}{2n}. \tag{6}$$

The estimator $p_i^{(s)}$ is actually a moment estimator and is unbiased:

$$\text{E}p_i^{(s)} = p_i.$$

This comes from Eq. (3), (5) and (6). The estimation of f is more difficult since it is a function of a *ratio* of parameters. A simple estimator of f could be obtained by replacing sample values in Eq. (1), resulting in what we denote as the *sample* estimator:

$$f^{(s)} = 1 - \left\{\frac{\sum_{(i,j):i<j} x_{ij}/n}{1 - \sum_i \left(p_i^{(s)}\right)^2}\right\}. \tag{7}$$

The latter estimator is biased. An alternative approach is to obtain unbiased estimators for both the numerator and denominator of f (Weir, 1996), resulting in the estimator:

$$f^{(weir)} = \frac{\sum_i \left\{\frac{x_{ii}}{n} - \left(p_i^{(s)}\right)^2\right\} + \frac{1}{2n}\left(1 - \sum_i \frac{x_{ii}}{n}\right)}{\left\{1 - \sum_i \left(p_i^{(s)}\right)^2\right\} - \frac{1}{2n}\left(1 - \sum_i \frac{x_{ii}}{n}\right)} \tag{8}$$

Both estimators in Eqs. (7) and (8) are consistent, meaning that both their bias and variance decrease with increasing sample size.

MAXIMUM LIKELIHOOD ESTIMATION

The method of maximum likelihood has been the most popular method of estimation in statistical theory. Maximum likelihood estimators (MLEs)

enjoy many optimal properties, such as consistency and large-sample efficiency (Silvey, 1971: 73). To obtain the MLEs of the p_i's and f in the multinomial model in Eq. (4), we start by writing the log-likelihood as:

$$\begin{aligned} l(\mathbf{p},f) &\equiv \sum_{(i,j):i\le j} x_{ij}\ln P_{ij} + C_1 \\ &= \sum_{i=1}^{K} x_{ii}\ln\{p_i^2 + p_i(1-p_i)f\} + \sum_{(i,j):i<j} x_{ij}\ln\{2p_ip_j(1-f)\} + C_1, \end{aligned} \tag{9a}$$

where $C_1 = \ln(n!/\prod_{(i,j):i\le j} x_{ij}!)$ is a constant, the vector $\mathbf{p} = \{p_i, i = 1,\ldots,K-1\}$ and $p_K \equiv 1 - \sum_{i=1}^{K-1} p_i$. Eq. (9a) can be simplified to:

$$\begin{aligned} l(\mathbf{p},f) = &\sum_{i=1}^{K} x_i \ln p_i + \sum_{i=1}^{K} x_{ii}\ln\{p_i + (1-p_i)f\} \\ &+ \sum_{(i,j):i<j} x_{ij}\ln(1-f) + C_2, \end{aligned} \tag{9b}$$

where $x_i \equiv x_{ii} + \sum_{j:j\neq i} x_{ij}$ and C_2 is a constant. The next step is to set the first partial derivatives of Eq. (9b) with respect to $p_1,\ldots,p_{K-1}$ and f (the so-called *score functions*) to zero and solve for the K parameters.

Considerable simplification occurs in one special case, when the number of independent parameters equals the number of independent counts (or the number of degrees of the freedom (d.f.) in the model). This happens when $K = 2$ and yields a *saturated* model in which the MLEs of the parameters can be obtained by simply equating the observed cell counts with the expected ones; that is, the MLEs are then equal to the sample estimators, Bailey's Rule (Bailey, 1951). Although the condition of equality in the number of d.f. and the number of parameters is sufficient, it is not necessary. To illustrate this, we take the 2-allele model with inbreeding (see Eq. (2)). This has 2 d.f. and 2 independent parameters (p and f) so that Bailey's Rule gives the MLEs:

$$\begin{aligned} p^{(mle)} &= p^{(s)} = \frac{2x_{11} + x_{12}}{2n}, \\ f^{(mle)} &= f^{(s)} = 1 - \frac{x_{12}/n}{2p^{(s)}q^{(s)}}, \end{aligned}$$

where $q^{(s)} \equiv 1 - p^{(s)}$. However, for any K-allele model under HWE ($f \equiv 0$ in Eq. (3)), the MLEs are still equal to the sample estimators, even though the number of d.f. exceeds the number of independent parameters:

$$p_i^{(mle)} = p_i^{(s)} = \frac{2x_{ii} + \sum_{j:j\neq i} x_{ij}}{2n}, \quad i = 1,2,\ldots,K. \tag{10}$$

There exist two sufficient conditions for the sample estimators to equal the MLEs. One condition is that the population must be in HWE ($f \equiv 0$), and the other is that there should be only two alleles ($K = 2$). However, when neither condition holds, the sample estimators do not in general maximize the likelihood (except in the rare case of a "perfect fit"; see Conclusion), as we now show:

Theorem 1: For the K-allele model with inbreeding coefficient f (where $f \neq 0$), having multinomial distribution given in Eq. (4) and cell probabilities P_{ij}'s defined in Eq. (3), the MLEs of the parameters $p_i (i = 1, \ldots, K)$ and f do not in general equal the sample estimators $p_i^{(s)}$ and $f^{(s)}$ (given in Eqs. (6) and (7)) for $K \geq 3$.

Proof. Differentiating the log-likelihood in Eq. (9b) with respect to $p_i (i = 1, \ldots, K-1)$ and f, and setting to zero, we obtain

$$\frac{\partial l}{\partial p_i} = \frac{x_i}{p_i} + \frac{x_{ii}(1-f)}{p_i + (1-p_i)f} - \frac{x_K}{p_K} - \frac{x_{KK}(1-f)}{p_K + (1-p_K)f} = 0, \tag{11}$$

$$\frac{\partial l}{\partial f} = \sum_{i=1}^{K} \left\{ \frac{x_{ii}(1-p_i)}{p_i + (1-p_i)f} \right\} - \frac{\sum_{(i,j):i<j} x_{ij}}{1-f} = 0. \tag{12}$$

For Eq. (11) to be valid, we need, for $i = 1, \ldots, K-1$,

$$\frac{x_i}{p_i} + \frac{x_{ii}(1-f)}{p_i + (1-p_i)f} = \frac{x_K}{p_K} + \frac{x_{KK}(1-f)}{p_K + (1-p_K)f}.$$

It follows from the above that, for all $i = 1, \ldots, K$,

$$\frac{x_i}{p_i} + \frac{x_{ii}(1-f)}{p_i + (1-p_i)f} = C_3, \tag{13}$$

where C_3 is a constant. We now replace p_i and f by their sample values in Eqs. (6) and (7), and show that Eq. (13) cannot in general be satisfied when $K \geq 3$. Let $p_i = p_i^{(s)} = \left(2x_{ii} + \sum_{j:j\neq i} x_{ij}\right)/(2n)$ and $f = f^{(s)}$ so that Eq. (13) becomes, for $i = 1, \ldots, K$,

$$\frac{x_{ii}}{\left(p_i^{(s)}\right)^2 + p_i^{(s)}\left(1 - p_i^{(s)}\right) f^{(s)}} = C_4, \tag{14}$$

where C_4 is a constant. If the sample estimators did equal the MLEs, then Eq. (14) would be valid for all possible x_{ij}'s and the corresponding $p_i^{(s)}$'s and $f^{(s)}$. However, this is not the case, in general. We construct a

simple counterexample: For $K \geq 3$, let $x_{11}/n = 1/K$; let $x_{ii}/n = 1/(2K)$ for $i = 2, \ldots, K$; let $x_{1j} = 0$ for $j = 2, \ldots, K$; and let $x_{ij}/n = 1/\{K(K-2)\}$ for $2 \leq i < j \leq K$. Then from Eq. (6), $p_i^{(s)} = 1/K$ for all $i = 1, \ldots, K$. This example shows that it is possible to have all $p_i^{(s)}$'s the same even though not all the x_{ii}'s are the same. Therefore, no one value of $f^{(s)}$ can possibly satisfy Eq. (14) for all i. Thus the sample estimators do not in general equal the MLEs when $K \geq 3$ and $f \neq 0$. On the other hand, when $K = 2$, $p_1^{(s)} = p_2^{(s)}$ implies $x_{11} = x_{22}$ and vice-versa, so that Eq. (14) is always valid. This completes the proof.

Remarks:

(i) For a model assuming HWE (i.e. $f \equiv 0$), Eq. (13) is valid for any K-allele model when p_i is replaced by $p_i^{(s)}$, showing that the sample estimators are indeed the MLEs, as previously stated (Li and Horvitz, 1953).

(ii) If the value of f is fixed, then Eq. (11) can be solved for the MLEs of the p_i's. Similarly, the values of p_i's could be fixed at their sample values $p_i^{(s)}$, and Eq. (12) could be solved for the MLE of f (Curie-Cohen, 1982). However, the MLEs would still not equal the sample values of the respective parameters, in general.

For a $K \geq 3$ allele model with inbreeding, we must obtain the MLEs by a numerical technique such as the method of scoring, or Newton's method with numerical or analytical derivatives (Monahan, 2001) (see section 5 for some numerical experiments).

In the next section, we evaluate a method of estimating the parameters p_i and f that has been popular in regression problems and goes back to the times of Legendre (1752–1833) and Gauss (1777–1855) (for example, Stigler (1986)).

A LEAST-SQUARES APPROACH TO THE ESTIMATION PROBLEM

The method of least-squares has the advantage over the method of maximum likelihood in that it does not require any distributional (or statistical) assumptions for the estimation itself and, in many cases, is more computationally tractable (especially with the use of matrix algebra) than the method of maximum likelihood. However, in general, MLEs have more attractive properties than least-squares estimators (LSEs) and are more widely used. For large samples, MLEs are nearly unbiased and have variances nearly equal to the Cramer-Rao lower bound (Silvey, 1971:77).

To obtain the LSEs of the p_i's and f for a general K-allele model, we need only the observed genotype counts $\mathbf{x} = \{x_{ij}, 1 \leq i \leq j \leq K\}$ and the expected genotype counts nP_{ij}, where P_{ij} is given in Eq. (3). The function that needs to be minimized here is called the *error sum of squares* and is given by:

$$SSE \equiv \sum_{(i,j):i\leq j} \left(x_{ij} - nP_{ij}\right)^2 = \sum_{i=1}^{K} \left\{x_{ii} - np_i^2 - np_i(1-p_i)f\right\}^2 + \sum_{(i,j):i<j} \left\{x_{ij} - 2np_ip_j(1-f)\right\}^2, \quad (15)$$

where $p_K \equiv 1 - \sum_{i=1}^{K-1} p_i$, as before. Eq. (15) is non-linear and the corresponding LSEs do not, in general, have the attractive properties of the linear case, namely unbiasedness and minimum variance (of all linear unbiased estimators). However, the LSEs of Eq. (15) are asymptotically unbiased and have minimum variance (Myers, 1990: 426).

To obtain the LSEs in Eq. (15), the usual approach is to set the first partial derivatives of SSE to zero, as in the MLE case. If the p_i's are fixed (say to their sample values), then f can be obtained analytically since the score equations for Eq. (15) are linear in f. Again, there is a simplification in the 2-allele case: The LSEs are simply the sample estimators, since the number of independent parameters is equal to the number of d.f. (i.e., there is no d.f. for the error and the model is saturated). So for $K = 2$, we have:

$$p^{(lse)} = p^{(mle)} = p^{(s)} = \frac{2x_{11} + x_{12}}{2n},$$
$$f^{(lse)} = f^{(mle)} = f^{(s)} = 1 - \frac{x_{12}/n}{2p^{(s)}q^{(s)}},$$

Substituting these LSEs into Eq. (15) yields a sum of zero terms, thus minimizing SSE.

For a $K \geq 3$ allele model, the LSEs are biased. They should be obtained using numerical techniques such as the Gauss-Newton procedure (Monahan, 2001).

NUMERICAL RESULTS

Using the delta method (Casella and Berger, 2001), Curie-Cohen (1982) shows that the expected value of $f^{(s)}$ is

$$\mathrm{E}f^{(s)} = f + O\left(\frac{1}{n}\right),$$

and its variance is

$$\operatorname{var} f^{(s)} = \frac{(1-f)\left\{1-2pq(1-f)-(p-q)^2(1-f)^2\right\}}{2pqn} + O\left(\frac{1}{n^2}\right).$$

Moments of several other related estimators have also been given in Robertson and Hill (1984). We use simulations to numerically compare the three types of estimators (sample, MLE and LSE) for a 3-allele model with $p_1 = .2$, $p_2 = .5$, $f = .05$ for various sample sizes n, and we investigate their biases and standard deviations (see Table 1). We observe that the biases of both the MLEs and LSEs of f are quite small in magnitude, decreasing as n gets larger. Theoretically, the biases of the sample estimators of the p_i's are zero, but since our simulations are based on a *finite* number of simulations the biases are not exactly zero. Moreover, in most cases, the MLEs outperform the LSEs both in terms of bias and standard deviation. Finally, as n gets larger, the standard deviations of the MLEs become slightly smaller than those of the sample estimators. In many cases, the sample estimators are better than both the MLEs and LSEs both in terms of bias and standard deviation. Table 2 shows more results with different allele- and f-combinations, with similar results to those of Table 1.

THE EXACT BIAS OF $f^{(s)}$ IN A 2-ALLELE MODEL

In this section, we investigate the bias of $f^{(s)}$, as defined in Eq. (7), for the 2-allele model and an arbitrary sample size n. We start by obtaining an algebraic expression for the expectation $\mathcal{E} \equiv \mathrm{E}\{(x_{12}/n)/(2p^{(s)}q^{(s)})\}$.

Theorem 2: For the 2-allele model with inbreeding coefficient f, and x_{11}, x_{12}, x_{22} as the number of genotypes A_1A_1, A_1A_2, A_2A_2, respectively, such that $x_{11} + x_{12} + x_{22} = n$, we have

$$\begin{aligned}\mathcal{E} = 2 - 2n\int_{z=0}^{1} \Big[&p^*z\left\{p^*z^2 + (1-p^*-q^*)z + q^*\right\}^{n-1} \\ &+ q^*z\left\{q^*z^2 + (1-p^*-q^*)z + p^*\right\}^{n-1}\Big]dz, \end{aligned} \tag{16}$$

where $p^* = p^2(1-f) + pf$, $q^* = q^2(1-f) + qf$, $p^{(s)} = (2x_{11} + x_{12})/(2n)$ and $q^{(s)} = 1 - p^{(s)}$.

Proof: Substituting $p^{(s)} = (2x_{11} + x_{12})/(2n)$ and $x_{12} = n - x_{11} - x_{22}$, we have

$$\frac{x_{12}/n}{2p^{(s)}q^{(s)}} = 2\left(1 - \frac{x_{11}}{n + x_{11} - x_{22}} - \frac{x_{22}}{n - x_{11} + x_{22}}\right), \tag{17}$$

TABLE 1 Bias and Standard Deviation (SD) of Sample Estimates (Sam), Maximum Likelihood Estimates (MLE), and Least-Squares Estimates (LSE) of p_1, p_2 and f for Various Sample Sizes in a 3-Allele Model with $p_1 = .2$

	p_1		p_2		f	
	bias	SD	bias	SD	bias	SD
$n = 10$	sam = –.00030	sam = .09212	sam = –.00105	sam = .11412	sam = –.05627	sam = .24461
	mle = .00296	mle = .14338	mle = –.00512	mle = .18701	mle = –.07264	mle = .30792
	lse = –.00436	lse = .10089	lse = .00629	lse = .12031	lse = –.08023	lse = .27441
$n = 20$	sam = .00042	sam = .06550	sam = –.00015	sam = .08204	sam = –.02458	sam = .17020
	mle = .00005	mle = .08629	mle = .00116	mle = .09206	mle = –.03355	mle = .19470
	lse = –.00305	lse = .07168	lse = .00518	lse = .08448	lse = –.03793	lse = .18922
$n = 30$	sam = .00102	sam = .05319	sam = –.00061	sam = .06579	sam = –.02069	sam = .13639
	mle = .00015	mle = .06870	mle = .00015	mle = .06902	mle = –.02601	mle = .14861
	lse = –.00157	lse = .05852	lse = .00335	lse = .06742	lse = –.03015	lse = .15053
$n = 40$	sam = .00016	sam = .04619	sam = –.00028	sam = .05716	sam = –.01359	sam = .11901
	mle = –.00009	mle = .04928	mle = –.00001	mle = .05790	mle = –.01673	mle = .12316
	lse = –.00179	lse = .05077	lse = .00288	lse = .05820	lse = –.02083	lse = .13127
$n = 100$	sam = .00023	sam = .02886	sam = –.00031	sam = .03622	sam = –.00015	sam = .07476
	mle = .00019	mle = .02894	mle = –.00031	mle = .03623	mle = –.00555	mle = .07362
	lse = –.00065	lse = .03228	lse = .00115	lse = .03668	lse = –.00821	lse = .08201
$n = 200$	sam = –.00015	sam = .02047	sam = –.00009	sam = .02569	sam = –.00188	sam = .05290
	mle = –.00014	mle = .02047	mle = –.00009	mle = .02569	mle = –.00220	mle = .05193
	lse = –.00052	lse = .02305	lse = –.00067	lse = .02604	lse = –.00317	lse = .05787

$p_2 = .5$ and $f = .05$, from computer simulations. For each sample size, with the given values of p_1, p_2 and f, multinomial observations were generated according to Eq. (4), using Maple V Release 4.00a. The sample estimates were calculated using Eqs. (6) and (7). The MLEs and LSEs were obtained by using the method of scoring and the Gauss-Newton procedure, respectively. All three estimators of f were assigned the value zero when the samples were monomorphic. Biases and standard deviations were calculated over 10,000 simulations for each sample size.

TABLE 2 Bias and Standard Deviation (SD) of Sample Estimates (sam), Maximum Likelihood Estimates (MLE), and Least-Squares Estimates (LSE) of p_1, p_2, and f for Various Sample Sizes in a 3-Allele Model from Computer Simulations

	p_1		p_2		f	
	bias	SD	bias	SD	bias	SD
(a) $p_1 = .2$, $p_2 = .5$ and $f = .01$						
$n = 20$	sam = –.00068	sam = .06420	sam = –.00069	sam = .07907	sam = –.04270	sam = .18613
	mle = .00044	mle = .10224	mle = .00119	mle = .09830	mle = –.03777	mle = .20778
	lse = –.00312	lse = .07105	lse = .00439	lse = .08174	lse = –.04264	lse = .18611
$n = 40$	sam = .00085	sam = .04506	sam = –.00047	sam = .05619	sam = –.01992	sam = .13034
	mle = .00032	mle = .04606	mle = –.00019	mle = .05631	mle = –.01597	mle = .12066
	lse = –.00142	lse = .05042	lse = .00280	lse = .05731	lse = –.01989	lse = .13031
(b) $p_1 = .2$, $p_2 = .5$ and $f = .05$ (from the Table 1)						
$n = 20$	sam = .00042	sam = .06550	sam = –.00015	sam = .08204	sam = –.02458	sam = .17020
	mle = .00005	mle = .08629	mle = .00116	mle = .09206	mle = –.03355	mle = .19470
	lse = –.00305	lse = .07168	lse = .00518	lse = .08448	lse = –.03793	lse = .18922
$n = 40$	sam = .00016	sam = .04619	sam = –.00028	sam = .05716	sam = –.01359	sam = .11901
	mle = –.00009	mle = .04928	mle = –.00001	mle = .05790	mle = –.01673	mle = .12316
	lse = –.00179	lse = .05077	lse = .00288	lse = .05820	lse = –.02083	lse = .13127
(c) $p_1 = .4$, $p_2 = .5$ and $f = .01$						
$n = 20$	sam = –.00003	sam = .07774	sam = .00022	sam = .07862	sam = –.02157	sam = .18318
	mle = .00211	mle = .11449	mle = .00255	mle = .10831	mle = –.04372	mle = .23114
	lse = –.00010	lse = .08086	lse = .00155	lse = .08088	lse = –.02445	lse = .20809
$n = 40$	sam = .00084	sam = .05534	sam = –.00033	sam = .05626	sam = –.01205	sam = .12932
	mle = .00173	mle = .07707	mle = .00053	mle = .07379	mle = –.02944	mle = .18129
	lse = .00073	lse = .05684	lse = .00035	lse = .05782	lse = –.01308	lse = .14863

(*Continued*)

TABLE 2 Continued

	p_1		p_2		f	
	bias	SD	bias	SD	bias	SD
(d) $p_1 = .4$, $p_2 = .5$ and $f = .05$						
$n = 20$	sam = –.00078	sam = .08025	sam = .00101	sam = .08114	sam = –.02288	sam = .20858
	mle = –.00069	mle = .11540	mle = .00083	mle = .11654	mle = –.04210	mle = .22889
	lse = –.00040	lse = .08308	lse = .00229	lse = .08313	lse = –.02278	lse = .20873
$n = 40$	sam = .00093	sam = .05669	sam = –.00082	sam = .05725	sam = –.01103	sam = .14848
	mle = .00163	mle = .07302	mle = –.00176	mle = .08933	mle = –.02207	mle = .15633
	lse = .00100	lse = .05834	lse = .00002	lse = .05883	lse = –.01103	lse = .14848

by partial fractions. Now (x_{11}, x_{22}) has a multinomial distribution:

$$(x_{11}, x_{22}) \sim \text{multi}(n, p^*, q^*) \Rightarrow \mathrm{E}\left(z_1^{x_{11}} z_2^{x_{22}}\right) = (p^* z_1 + q^* z_2 + 1 - p^* - q^*)^n.$$

$\mathrm{E}\left(z_1^{x_{11}} z_2^{x_{22}}\right)$ is the joint *probability generating function* of (x_{11}, x_{22}). Differentiating it partially with respect to z_1,

$$\mathrm{E}\left(x_{11} z_1^{x_{11}-1} z_2^{x_{22}}\right) = np^*(p^* z_1 + q^* z_2 + 1 - p^* - q^*)^{n-1},$$

so that

$$\mathrm{E}\left(x_{11} z_1^{n+x_{11}-1} z_2^{x_{22}}\right) = np^* z_1^n (p^* z_1 + q^* z_2 + 1 - p^* - q^*)^{n-1}.$$

Setting $z_2 = 1/z_1$ and integrating with respect to z_1,

$$\mathrm{E}\left(\frac{x_{11} z_1^{n+x_{11}-x_{22}}}{n + x_{11} - x_{22}}\right) = np^* \int z_1 \left\{p^* z_1^2 + q^* + (1 - p^* - q^*) z_1\right\}^{n-1} dz_1.$$

Therefore,

$$\mathrm{E}\left(\frac{x_{11}}{n + x_{11} - x_{22}}\right) = np^* \int_{z_1=0}^{1} z_1 \left\{p^* z_1^2 + q^* + (1 - p^* - q^*) z_1\right\}^{n-1} dz_1. \tag{18}$$

Similarly,

$$\mathrm{E}\left(\frac{x_{22}}{n - x_{11} + x_{22}}\right) = nq^* \int_{z_2=0}^{1} z_2 \left\{p^* + q^* z_2^2 + (1 - p^* - q^*) z_2\right\}^{n-1} dz_2. \tag{19}$$

Using Eq. (17), (18) and (19), we get the required expression in Eq. (16). This completes the proof.

From Eq. (16), we have

$$\mathrm{E} f^{(s)} = 1 - \mathcal{E}.$$

Thus, the bias of $f^{(s)}$ is given by

$$\text{bias}\left\{f^{(s)}\right\} = \mathrm{E} f^{(s)} - f = 1 - \mathcal{E} - f. \tag{20}$$

Remembering that, for the 2-allele case, $f^{(s)} = f^{(mle)} = f^{(lse)}$, Eq. (20) also gives the bias for both the MLE and LSE of f.

In Table 3, we compute the bias of $f^{(s)}$ for various sample sizes n in the 2-allele case with $p = .1$, $.25$ and $.5$, $f = -.30, -.10, 0, .10, .30$ and $.75$, and $n = 20$ and 40. In all cases, $f^{(s)}$ underestimates the inbreeding coefficient f.

TABLE 3 Exact Bias of $\mathbf{f}^{(s)}$ for Sample Sizes $n = 20, 40$ in a 2-Allele Model with $p = .1, .25, .5,$ and $f = -.30, -.10, 0, .10, .30, .75$

		f					
p	n	$-.30$	$-.10$	0	.10	.30	.75
.1	20	n/a	$-.00513$	$-.02526$	$-.04386$	$-.07607$	$-.11940$
	40	n/a	$-.00258$	$-.01265$	$-.02135$	$-.03421$	$-.03585$
.25	20	$-.00977$	$-.02125$	$-.02564$	$-.02904$	$-.03271$	$-.02271$
	40	$-.00488$	$-.01058$	$-.01265$	$-.01418$	$-.01554$	$-.00952$
.50	20	$-.02278$	$-.02518$	$-.02564$	$-.02559$	$-.02392$	$-.01198$
	40	$-.01138$	$-.01248$	$-.01265$	$-.01258$	$-.01165$	$-.00570$

Note that the bias of $\mathbf{f}^{(s)}$ at p is the same as the bias at $1 - p$. Two cell combinations were impossible (n/a) because $-\mathbf{p}/(\mathbf{1} - \mathbf{p}) \leq \mathbf{f} \leq \mathbf{1}$. for each sample size, with the given values of p and f, the bias was calculated using Eq. (20).

AN UNBIASED ESTIMATOR OF f FOR p SMALL OR p LARGE

Let us consider the case when p is very small such that $p \gg p^2 \approx 0$. Then $p^* \approx pf$ and $q^* \approx 1 - 2p + pf$. Then, by using Taylor's series expansions, Eq. (18) can be simplified to

$$\mathrm{E}\frac{x_{12}/n}{2p^{(s)}q^{(s)}} \approx \frac{2n - 1 + 2np}{2n - 1} - \frac{np(2n + 1)f}{2n - 1}.$$

An approximately unbiased estimator of f for small p can thus be obtained:

$$f^{(unb)} \approx \frac{2n - 1 + 2np}{np(2n + 1)} - \frac{2n - 1}{np(2n + 1)}\left\{\frac{x_{12}/n}{2p^{(s)}q^{(s)}}\right\}. \tag{21}$$

Eq. (21) involves p, which can be estimated from the sample by $p^{(s)}$. The corresponding approximate estimator in Eq. (21) then becomes

$$\widetilde{f}^{(unb)} \approx \frac{2n - 1 + 2np^{(s)}}{np^{(s)}(2n + 1)} - \frac{2n - 1}{np^{(s)}(2n + 1)}\left\{\frac{x_{12}/n}{2p^{(s)}q^{(s)}}\right\}. \tag{22}$$

Analogous estimators when p is large are be obtained by interchanging p and q in Eq. (22).

Table 4 compares the three estimators $f^{(s)}, f^{(weir)}$ and $\widetilde{f}^{(unb)}$ for $p = .001, f = .05$, and for various sample sizes n. From $n = 20$ through $n = 50$, the bias of $\widetilde{f}^{(unb)}$ is almost the same as that of $f^{(weir)}$, and are slightly less than that $f^{(s)}$. All three standard deviations are very close. Table 5 again compares the same three estimators for $n = 25$ and 75, $f = .05$, and $p = .001, .005, \ldots, .025$. For $n = 25, \widetilde{f}^{(unb)}$ outperforms $f^{(s)}$

TABLE 4 Bias and Standard Deviation (SD) of $\mathbf{f}^{(s)}$ (sam), $\mathbf{f}^{(weir)}$ (weir) and $\widetilde{\mathbf{f}}^{(unb)}$ (unb) in a 2-Allele Model with n = 20, 25, 30, 40, 50, 75, 100, 125, 150, $f = .05$, and $p = .001$ from Computer Simulations

$n = 20$		$n = 25$		$n = 30$	
Bias	SD	bias	SD	bias	SD
sam = −.04974	sam = .03565	sam = −.04952	sam = .03758	sam = −.05001	sam = .03099
weir = −.04876	weir = .03524	weir = −.04863	weir = .03740	weir = −.04866	weir = .03071
unb = −.04876	unb = .03492	unb = −.04860	unb = .03739	unb = −.04866	unb = .03032
$n = 35$		$n = 40$		$n = 50$	
Bias	SD	bias	SD	bias	SD
sam = −.04964	sam = .03680	sam = −.04890	sam = .04593	sam = −.04855	sam = .04830
weir = −.04866	weir = .03664	weir = −.04795	weir = .04578	weir = −.04766	weir = .04823
unb = −.04866	unb = .03631	unb = −.04790	unb = .04578	unb = −.04767	unb = .04749
$n = 75$		$n = 100$		$n = 125$	
Bias	SD	bias	SD	bias	SD
sam = −.04789	sam = .05514	sam = −.04625	sam = .06721	sam = −.04557	sam = .07037
weir = −.04700	weir = .05509	weir = −.04537	weir = .06717	weir = −.04472	weir = .07035
unb = −.04696	unb = .05487	unb = −.04547	unb = .06577	unb = −.04502	unb = .06723

For each sample size, with the given values of p_1, p_2 and f, multinomial observations were generated according to Eq. (4), using Maple V Release 4.00a. The three estimates were calculated using Eqs. (7), (8) and (22), respectively. All three estimators of f were assigned the value zero when the samples were monomorphic. Biases and standard deviations were calculated over 10,000 simulations for each sample size.

TABLE 5 Bias and Standard Deviation (SD) of $\mathbf{f}^{(s)}$ (sam), $\mathbf{f}^{(weir)}$ (weir) and $\widetilde{\mathbf{f}}^{(unb)}$ (unb) in a 2-Allele Model with p = .001, .005, .009, .013, .017, .021, .025, $f = .05$, and $n = 25, 75$ from Computer Simulations

	$n = 25$		$n = 75$	
	Bias	SD	Bias	SD
$p = .001$	sam = −.04924	sam = .04121	sam = −.04787	sam = .05502
	weir = −.04831	weir = .04104	weir = −.04698	weir = .05498
	unb = −.04831	unb = .04048	unb = −.04697	unb = .05432
$p = .005$	sam = −.04948	sam = .07159	sam = −.04041	sam = .10937
	weir = −.04521	weir = .07112	weir = −.03700	weir = .10919
	unb = −.04493	unb = .06928	unb = −.03789	unb = .09868
$p = .009$	sam = −.04759	sam = .10173	sam = −.03124	sam = .14257
	weir = −.04044	weir = .10093	weir = −.02645	weir = .14216
	unb = −.03972	unb = .09640	unb = −.02929	unb = .11977
$p = .013$	sam = −.04540	*sam* = .12313	sam = −.02854	sam = .14770
	weir = −.03606	weir = .12218	weir = −.02309	weir = .14731
	unb = −.03490	unb = .11386	unb = −.02763	unb = .11328
$p = .017$	sam = −.04326	sam = .13681	sam = −.02434	sam = .15324
	weir = −.03226	weir = .13570	weir = −.01862	weir = .15301
	unb = −.03108	unb = .12258	unb = −.02526	unb = .11067
$p = .021$	sam = −.04245	sam = .15081	sam = −.02133	sam = .15435
	weir = −.02940	weir = .14950	weir = −.01537	weir = .15445
	unb = −.02737	unb = .13168	unb = −.02491	unb = .09985
$p = .025$	sam = −.04299	sam = .15335	sam = −.01915	sam = .15614
	weir = −.02877	weir = .15189	weir = −.01285	*weir* = .15653
	unb = −.02621	unb = .12924	unb = −.02413	unb = .09593

For each sample size, with the given values of p_1, p_2 and f, multinomial observations were generated according to Eq. (4), using Maple V Release 4.00a. The three estimates were calculated using Eqs. (7), (8) and (22), respectively. All three estimators of f were assigned the value zero when the samples were monomorphic. Biases and standard deviations were calculated over 10,000 simulations.

and $f^{(weir)}$ both in terms of bias and standard deviation. However, for $n = 75$, $f^{(weir)}$ is the best in terms with respect to bias, while $\widetilde{f}^{(unb)}$ still remains the best with respect to standard deviation.

CONCLUSION

We have investigated some of the methods of estimation of allele frequencies and inbreeding coefficients in a K-allele model. We have proved that in a $K \geq 3$ allele model with inbreeding, the sample values of the allele frequencies and inbreeding coefficient differ from their

maximum likelihood estimates. We have introduced and evaluated a least-squares way of looking at the estimation problem. For the $K = 3$ allele case, we have numerically compared the sample, maximum likelihood, and least-squares estimates of the parameters in our model. We have shown that, while the biases are relatively small for all three estimators, the standard deviations are slightly smaller for the MLEs than for the sample estimators with increasing sample size. We have used probability generating functions to calculate the exact bias of the sample estimator of the inbreeding coefficient in a $K = 2$ allele model for any sample size n. Finally, we have derived an approximately unbiased estimator of f when p is very small (or very large) and have shown that it outperforms $f^{(s)}$ and $f^{(weir)}$ both in terms of bias and standard deviation when p is very small ($\sim .001$) and n is moderate (~ 25).

Is it surprising that the sample estimators of the parameters do not equal their MLEs, when $K \geq 3$? This fact is perhaps less surprising for the inbreeding coefficient f, but it seems counterintuitive for the allele frequencies. After all, the actual count of A_i alleles in the sample is indeed given by Eq. (10), the sample estimator. On the other hand, one can realize that in the presence of inbreeding, homozygotes with alleles IBD yield only half the information of homozygotes whose alleles are not autozygous. In any case, the sample estimators do not maximize the likelihood for any of the parameters when $K \geq 3$ and $f \neq 0$. The only exception occurs when there is a "perfect fit," that is, if and only if the $K(K+1)/2$ equations $nP_{ij} = x_{ij}$ (where P_{ij} is defined in Eq. (3)) yield a *unique* solution for $\mathbf{p}$ and f. Li and Horvitz (1953) did state this result, which they regarded as reasonable; however, they did not prove it. What we provide here is a rigorous mathematical proof.

Additionally, three observations can be made from the proof of Theorem 1. First, no *one* value of $f^{(s)}$ satisfies Eqs. (14) for all $i = 1, 2, \ldots, K$; thus, no matter what the value of f, $\mathbf{p} = \mathbf{p}^{(s)}$ does not maximize the likelihood in general. Second, the proof does not involve the actual value of $f^{(s)}$; thus even if the true value of f were known, so that f did not need to be estimated, it would still follow from Eqs. (14) that the sample estimators of the p_i's would not maximize the likelihood. Third, even if the true value of the p_i's were known, so that they did not need to be estimated, it would still follow from Eqs. (14) that the sample estimates of f would not maximize the likelihood.

The least-squares approach shows an interesting way of looking at the estimation problem, that of fitting the $K(K+1)/2$ genotype counts, or data points, x_{ij} to their expected values. The least-squares solutions thus obtained represent the estimates that best fit the system of equations $nP_{ij} = x_{ij}$. When $K \geq 3$ and $f \neq 0$, these estimates

are also different from both the sample values and the MLEs (except in the case of a perfect fit). Since least-squares are not as widely used as maximum likelihood, we did not dwell too much on the theoretical properties of the former.

The numerical calculations in the section, Numerical Results, show that, at least for the cases we have considered, the biases of both the MLEs and LSEs are quite small in magnitude, decreasing as n gets larger. The MLEs are slightly better than the LSEs both in terms of bias and standard deviation. For larger n, the standard deviations of the MLEs become slightly smaller than those of the sample estimates.

In the subsequent section, we give an exact algebraic expression for $\mathrm{E}\{(x_{12}/n)/(2p^{(s)}q^{(s)})\}$ in integral form for any sample size n in a $K = 2$ allele model with inbreeding. This enables not only the exact bias of $f^{(s)}$ to be calculated, but also the exact bias of any estimator of f which is of the form $1 - \kappa(x_{12}/n)/(2p^{(s)}q^{(s)})$, where κ is a constant.

We use the results from that section to derive an approximately unbiased estimator of f when p is very small (or very large), and we compare its performance with other estimators. We show that this estimator outperforms $f^{(s)}$ and $f^{(weir)}$ both in terms of bias and standard deviation when p is very small ($\sim .001$) and n is moderate (~ 25).

Our primary aim in this paper was to *investigate* the performances of various estimators (SAM, MLE, LSE). Therefore, we have refrained from giving clear-cut guidelines on the "best" estimator based on bias and standard deviation, but rather have left it to the reader to make that choice. Moreover, we would like to point out that, although most of the biases seem to be small, they are on average only one order of magnitude smaller than the corresponding standard deviations, and therefore unlikely to be insignificant.

A final interesting point arises serendipitously from the expression for $\mathrm{E}f^{(s)}$ derived in Theorem 2. Consider the extreme situation in which either $x_{11} = n$ or $x_{22} = n$ (i.e., all observed genotypes are homozygous for one particular allele). What value should one assign to $f^{(s)}$ in that case? One cannot simply use the usual expression for $f^{(s)}$ (see Eq. 17) because that expression, viewed as a function of x_{11} and x_{22}, $g(x_{11}, x_{22})$, has a singularity at the points $(n, 0)$ and $(0, n)$. Moreover, applying L'Hôpital's rule does not resolve the indeterminacy because the value of $g(x_{11}, x_{22})$ near each singularity differs, depending on how one approaches the singularity.

The question of what value to assign to $f^{(s)}$ when $x_{11} = n$ or $x_{22} = n$ becomes relevant if, for example, one wishes to calculate the exact $\mathrm{E}f^{(s)}$ in some finite sample. Hartl and Clark (1997:112), imply that $f^{(s)}$ equals zero in this situation, based on the reasoning that, "the genotype frequencies, though extreme, still satisfy the Hardy-Weinberg

principle." We can derive the same value (0), based on a mathematical argument, namely that the PGF *implicitly* assigns that value to $f^{(s)}$ in that situation. See the Appendix for details.

APPENDIX

Proof that the probability generating function, as formulated in Theorem 2, implicitly assigns the value 0 to $f^{(s)}$ when $(x_{11}, x_{22}) = (0, n)$ or when $(x_{11}, x_{22}) = (n, 0)$.

We write Eq. (17) as

$$1 - f^{(s)} = \frac{x_{12}/n}{2p^{(s)}q^{(s)}} = 2 - 2\Phi_1(x_{11}, x_{22}) - 2\Phi_2(x_{11}, x_{22}),$$

where

$$\Phi_1(x_{11}, x_{22}) \equiv \frac{x_{11}}{n + x_{11} - x_{22}}, \quad \Phi_2(x_{11}, x_{22}) \equiv \frac{x_{22}}{n - x_{11} + x_{22}}.$$

Let us consider the case when $(x_{11}, x_{22}) = (0, n)$. Then $\Phi_2(0, n) = 1/2$, so

$$1 - f^{(s)} = 1 - 2\Phi_1(0, n) \Rightarrow f^{(s)} = 2\Phi_1(0, n). \tag{23}$$

Now $\Phi_1(0, n)$ is indeterminate and cannot be evaluated even by applying L'Hôpital's rule. For example, if we approach $\Phi_1(0, n)$ along the line $x_{11} + x_{22} = n$, then $\lim_{x_{22} \to n} \Phi_1(0, n) = 1/2$; if along the line $2x_{11} + x_{22} = n$, then $\lim_{x_{22} \to n} \Phi_1(0, n) = 1/3$; if along the line $x_{11} = 0$, then $\lim_{x_{22} \to n} \Phi_1(0, n) = 0$; etc. Thus, the value of $\Phi_1(x_{11}, x_{22})$ at $(0, n)$ represents a true singularity. However, we can write

$$\begin{aligned}
\mathrm{E}\Phi_1(x_{11}, x_{22}) &= \sum_{x_{11}=0}^{n} \sum_{x_{22}=0}^{n-x_{11}} \Phi_1(x_{11}, x_{22}) \left(\frac{n!}{x_{11}!\, x_{22}!\, (n - x_{11} - x_{22})!} \right) \\
&\quad \times (p^*)^{x_{11}} (q^*)^{x_{22}} (1 - p^* - q^*)^{n - x_{11} - x_{22}} \\
&= \Phi_1(0, n) + \sum_{x_{11}=1}^{n} \sum_{x_{22}=0}^{n-x_{11}} \left(\frac{x_{11}}{n + x_{11} - x_{22}} \right) \\
&\quad \times \left(\frac{n!}{x_{11}!\, x_{22}!\, (n - x_{11} - x_{22})!} \right) (p^*)^{x_{11}} (q^*)^{x_{22}} \\
&\quad \times (1 - p^* - q^*)^{n - x_{11} - x_{22}},
\end{aligned} \tag{24}$$

since $(x_{11}, x_{22}) \sim \text{multi}(n, p^*, q^*)$. But also, using the PGF to find expected values, as in Eq. (18), we know

$$\begin{aligned}
\mathrm{E}\Phi_1(x_{11}, x_{22}) &= np^* \int_{z=0}^{1} z\left\{p^* z^2 + q^* + (1-p^*-q^*)z\right\}^{n-1} dz \\
&= np^* \int_{z=0}^{1} z \sum_{i=0}^{n-1} \sum_{j=0}^{n-1-i} \frac{(n-1)!}{i!\,j!(n-1-i-j)!} \left(p^* z^2\right)^i (q^*)^j \\
&\quad \times \{(1-p^*-q^*)z\}^{n-1-i-j} dz \\
&= np^* \int_{z=0}^{1} z \sum_{i=1}^{n} \sum_{j=0}^{n-i} \frac{(n-1)!}{(i-1)!j!(n-i-j)!} \left(p^* z^2\right)^{i-1} (q^*)^j \\
&\quad \times \{(1-p^*-q^*)z\}^{n-i-j} dz \\
&= np^* \sum_{i=1}^{n} \sum_{j=0}^{n-i} \frac{(n-1)!}{(i-1)!j!(n-i-j)!} (p^*)^{i-1} (q^*)^j \\
&\quad \times (1-p^*-q^*)^{n-i-j} \int_{z=0}^{1} z^{n-1+i-j} dz \\
&= \sum_{i=1}^{n} \sum_{j=0}^{n-i} \frac{n!i}{i!j!(n-i-j)!} (p^*)^i (q^*)^j (1-p^*-q^*)^{n-i-j} \int_{z=0}^{1} z^{n-1+i-j} dz \\
&= \sum_{i=1}^{n} \sum_{j=0}^{n-i} \left(\frac{i}{n+i-j}\right) \left(\frac{n!}{i!j!(n-i-j)!}\right) (p^*)^i (q^*)^j (1-p^*-q^*)^{n-i-j}. \qquad (25)
\end{aligned}$$

Clearly Eq. (24) is equal to Eq. (25) if and only if

$$\Phi_1(0, n) \equiv 0.$$

That is, if we accept that the PGF does yield the expected value of $\Phi_1(x_{11}, x_{22})$, then perforce $\Phi_1(0, n)$ must be set to zero. Eq. (23) means that when $(x_{11}, x_{22}) = (0, n)$, then $f^{(s)} \equiv 0$. Similarly, when $(x_{11}, x_{22}) = (\dot{n}, 0)$, then $f^{(s)} \equiv 0$.

ACKNOWLEDGEMENTS

We acknowledge helpful comments from two reviewers that led to a much-improved manuscript. This work was supported in part by grants DK31813, AA13654, MH48858, DK31775, NS27941.

REFERENCES

Bailey, N.T.J. (1951). Testing the solubility of maximum likelihood equations in the routine application of scoring methods. *Biometrics 7*(3): 268–274.

Casella, G. and Berger, R.L. (2001). *Statistical Inference* (2nd Edition). Brooks: Cole.

Cotterman, C.W. (1940). *A Calculus for Statistico-Genetics*. PhD Dissertation Ohio University, Columbus, Ohio.

Curie-Cohen, M. (1982). Estimates of inbreeding in a natural population: a comparison of sampling properties. *Genetics 100*(2): 339–358.

Excoffier, L. (2001). Analysis of population subdivision. In D.J. Balding, M. Bishop, and C. Cannings (Eds.), *Handbook of Statistical Genetics*. Chichester: Wiley.

Haldane, J.B.S. and Moshinsky, P. (1939). Inbreeding in Mendelian populations with special reference to human cousin marriage. *Annals of Eugenics 9*: 321–340.

Li, C.C. and Horvitz, D.G. (1953). Some methods of estimating the inbreeding coefficient. *American Journal of Human Genetics 5*(2): 107–117.

Malécot, G. (1948). *Les Mathématiques de l'Hérédité*. Paris: Masson et Cie.

Monahan, J.F. (2001). *Numerical Methods of Statistics*. Cambridge: Cambridge University Press.

Myers, R.M. (1990). *Classical and Modern Regression with Applications* (2nd Edition). Kent: PWS.

Nagylaki, T. (1998). Fixation indices in subdivided populations. *Genetics 148*(3): 1325–1332.

Robertson, A. and Hill, W.G. (1984). Deviations from Hardy-Weinberg proportions: sampling variances and use in estimation of inbreeding coefficients. *Genetics 107*(4): 703–718.

Rousset, F. and Raymond, M. (1995). Testing heterozygosity excess and deficiency. *Genetics 140*: 1413–1419.

Silvey, S.D. (1971). *Statistical Inference*. Duxbury Press.

Stigler, S.M. (1986). *The History of Statistics: The Measurement of Uncertainty Before 1900*. Cambridge: Harvard University Press.

Weir, B.S. (1996). *Genetic Data Analysis 2*. Sinauer Associates, Sunderland, MA.

Weir, B.S. (2001). *Forensics*. In D.J. Balding, M. Bishop, and C. Cannings (Eds.), *Handbook of Statistical Genetics*. Chichester: Wiley.

Wright, S. (1922). Coefficients of inbreeding and relationship. *The American Naturalist 56*: 330–338.

Wright, S. (1951). The genetical structure of populations. *Annals of Eugenics 15*: 323–354.

Wright, S. (1965). The interpretation of population structure by F-statistics with special regard to systems of mating. *Evolution 19*(3): 395–420.

Yasuda, N. (1968). Estimation of the inbreeding coefficient from phenotype frequencies by a method of maximum likelihood scoring. *Biometrics 24*(4): 915–935.