RESEARCH NOTE

A note on the variance of the estimate of the fixation index F

PAULO A. OTTO* and RENAN B. LEMES*

Departamento de Genética e Biologia Evolutiva, Instituto de Biociências, Universidade de São Paulo, Caixa Postal (P.O. Box) 11.461, 05422-970 São Paulo, SP, Brazil

[Otto P. A. and Lemes R. B. 2015 A note on the variance of the estimate of the fixation index F. *J. Genet.* **94**, 759–763]

Introduction

In the two-allele case, the formulas for the estimated variances of allelic frequency $p = 1 - q$ and fixation index (average inbreeding coefficient) F are known in the specialized literature of statistical genetics. Besides presenting here an alternative manner to estimate the variance of both parameters, we also derive a very simple approximation for the estimate of the variance of F. The approximation, with adequate validity, can be applied not only to the two-allele case but also to the generalized case of any number of alleles segregating at an autosomal locus.

The variance of F has many practical applications in population genetics. For example, if geneticists are interested in a precise determination of its value, commonly the parameter is estimated from sets of data obtained from the genotypic analysis of several independent autosomal loci of the same population. If the estimates of F for loci $1, 2, \ldots, k$ are $\hat{F}_1$, $\hat{F}_2$, …, $\hat{F}_k$, the method of averaging these estimates is obtained usually by weighing them by the reciprocal of their corresponding variances:

$$\bar{F} = \frac{\sum \frac{\hat{F}_i}{var\left(\hat{F}_i\right)}}{\sum \frac{1}{var\left(\hat{F}_i\right)}} \quad .$$

Our paper deals with the population as specified by formulas (2.22) on page 65 of Weir's monograph (Weir 1996). The virtue of the resulting approximation for the estimate of *var* (*F*) we provide is a simple formula with adequate validity for multiple alleles, whereas Weir does leave his reader with details to be supplied.

Our results are presented below in three different sections: the first one deals with the case of two alleles, leading naturally to a second section on multiple alleles; a third section deals with simulation studies we performed to validate the approximations derived here.

The special case of two autosomal alleles

The generic population genotype frequencies in relation to an autosomal biallelic locus can be represented by equations

$$P\left(AA\right) = p^2 + pqF,$$

$$P\left(Aa\right) = 2pq\left(1 - F\right),$$

and

$$P\left(aa\right) = q^2 + pqF,$$

that represent a special case of Weir's population formulas referred to in the previous section, and where $p = P(A)$ is the frequency of allele A, $q = 1 - p = P(a)$ the frequency of its alternative allele a, and F the fixation index normally obtained from the formula,

$$F = 1 - \frac{h}{2pq},$$

where h is the heterozygous frequency $h = \frac{NAa}{N}$, NAa the observed number of heterozygous individuals, and N the total number of sampled subjects.

Since the expected values corresponding to observed numbers NAA, NAa and Naa of individuals AA, Aa and aa, respectively in a sample with size N and to a fixation index (average inbreeding coefficient) $F \neq 0$ are

$$N\left(p^2 + pqF\right),$$

$$2Npq\left(1 - F\right),$$

and

$$N\left(q^2 + pqF\right)$$

*For correspondence. E-mail: Paulo A. Otto, otto@usp.br; Renan B. Lemes, lemes.rb@usp.br.

Keywords. inbreeding; inbreeding coefficient; fixation index; variance estimation; variance of the inbreeding coefficient.

respectively, the likelihood function in logarithmic form is given by expression:

$$L = NAA \log\left[p^2 + p\,(1-p)\,F\right] + NAa \log\left[2p\,(1-p)\,(1-F)\right] + Naa \log\left[(1-p)^2 + p\,(1-p)\,F\right].$$

Maximum likelihood estimates of both p and F are obtained from the system $\{\frac{\partial \mathrm{L}}{\partial p} = 0,\ \frac{\partial \mathrm{L}}{\partial F} = 0\}$ and it is not difficult to determine that these solutions are identical to the estimates of p and F obtained through the application of intuitive direct counting methods: $\hat{p} = \hat{d} + \frac{\hat{h}}{2}$ and $\hat{F} = 1 - \frac{\hat{h}}{2\hat{p}(1-\hat{p})}$.

In the formulas above (and in many equations that follow) symbols like $\hat{p} = 1 - \hat{q}$ and $\hat{F}$ have carets because they are not unknown population (true) values but estimates of the corresponding parameters from the population, obtained from simple random sampling of a large population with genotype proportions occasionally different from Hardy–Weinberg ratios.

The determination of the values for the variances of p and F using iterative numerical procedures such as the usual generalized Newton–Raphson method is a complicated issue since it is practically impossible to get convergence to the estimation points p and F (Weir 1996), but values of $var(\hat{p})$ and $var(\hat{F})$, the variances of the estimated values of p and F can be taken directly from the variance–covariance matrix obtained by inverting the information matrix of second derivatives evaluated at estimation points $\{\hat{p}, \hat{F}\}$:

$$var\left(\hat{\mathrm{p}}\right) = \frac{a_{22}}{a_{11}.a_{22} - a_{12}.a_{21}}$$

and

$$var\left(\hat{\mathrm{F}}\right) = \frac{a_{11}}{a_{11}.a_{22} - a_{12}.a_{21}},$$

where $a_{11} = -\frac{\partial^2 \mathrm{L}}{\partial p^2}$, $a_{12} = -\frac{\partial^2 \mathrm{L}}{\partial p \partial F}$, $a_{21} = -\frac{\partial^2 \mathrm{L}}{\partial F \partial p}$, and $a_{22} = -\frac{\partial^2 \mathrm{L}}{\partial F^2}$, with all four second derivatives evaluated at estimation points

$$\hat{p} = \hat{d} + \frac{\hat{h}}{2}$$

and

$$\hat{F} = 1 - \frac{\hat{h}}{2\hat{p}\left(1 - \hat{p}\right)}.$$

In the case of the variance of the estimated value of p, we obtain $var(\hat{p}) = \frac{\hat{p}\hat{q}\left(1+\hat{F}\right)}{2N}$, as expected. This formula coincides with the expression obtained by Curie-Cohen (1982) and other authors (references of the many papers on the variances of p and F by Cockerham, Weir, and Cockerham and Weir, in Weir 1996) using different alternative methods.

Since

$$\frac{var\left(\hat{F}\right)}{var\left(\hat{p}\right)} = \frac{a_{11}}{a_{22}},$$

we get straightforwardly

$$var\left(\hat{F}\right) = \frac{\left(1 - \hat{F}\right)\left[2\hat{p}\hat{q} + 2\hat{F}\left(1 - 3\hat{p}\hat{q}\right) - \hat{F}^2\left(\hat{p} - \hat{q}\right)^2\right]}{2N\hat{p}\hat{q}}, \qquad (1)$$

a result that is algebraically equivalent to the formulas derived by Fyfe and Bailey (1951) and Curie-Cohen (1982) using alternative methods.

In the two-allele case, an approximate value of the variance of the estimate F can be obtained in a simple and straightforward way if we treat p, that can be directly calculated from the sample through $\hat{p} = \hat{d} + \frac{\hat{h}}{2}$, as an independently estimated parameter. Then the variance of $\hat{F}$ is obtained directly from $(a_{22})^{-1}$, taking form

$$var(\hat{F}) = \left\{\frac{NAA\left(1-\hat{p}\right)^2}{\left[\hat{p} + \left(1-\hat{p}\right)\hat{F}\right]^2} + \frac{NAa}{\left(1-\hat{F}\right)^2} + \frac{Naa\left(1-\hat{q}\right)^2}{\left[\hat{q} + \left(1-\hat{q}\right)\hat{F}\right]^2}\right\}^{-1}. \qquad (2)$$

This formula works as well as the one derived in this paper or other expressions from the literature.

The generalized case of any number of autosomal alleles

When the number of alleles (k) segregating at an autosomal locus is larger than two, estimates obtained through intuitive counting methods (and that correspond to maximum likelihood estimates under stringent conditions) are given by

$$\hat{p}_i = \frac{2N(a_i a_i) + \sum N(a_i a_j)}{2N},$$

$\hat{p}_j = \ldots, \ldots, \hat{p}_{k-1} = \ldots$, with i fixed and $j \neq i$ varying from 1 to k, that is $\sum N(a_i a_j)$ in the formula above represents the total number of heterozygous individuals as to the i allele, and

$$F = 1 - \frac{\sum\sum N(a_i a_j)}{2N\left(\sum\sum p_i p_j\right)},$$

with i varying from 1 to k and $j > i$, that is $\sum\sum N(a_i a_j)$ in the formula above represents the total number of heterozygous individuals as to alleles i and j.

In spite of being generally impossible to obtain convergence to the values shown above using numerical iterative procedures and to get the value of the variance of $\hat{F}$ by means of variations of Fisher's variance method (a rigorous argumentation on the subject is presented by Weir on pages 49–51 of his 1996 book), numerical values of $var(\hat{F})$ can be obtained either from large series of computer simulations or from the inspection of the main diagonal of the variance–covariance matrix evaluated at estimation points $\hat{p}_1, \ldots, \hat{p}_{k-1}, \hat{F}$. The variance of $\hat{p}_i$ in the multiallelic case can be determined independently through the formula (Curie-Cohen 1982; Weir 1996)

$$var\left(\hat{p}_i\right) = \frac{\hat{p}_i\left(1-\hat{p}_i\right)\left(1+\hat{F}\right)}{2N}.$$

Literal expressions for the variance of the estimated value of F when the number of alleles is larger than two can be obtained from the matrix method we used in the previous section (two-allele case), but they are however much more complicated; reliable, easily handled approximations should be preferred instead on practical grounds. Curie-Cohen (1982) and Robertson and Hill (1984) derived some of them under stringent statistical assumptions.

The real importance of the approximate formula derived for the two-allele case, however, stems from the fact that it is very easy to generalize it for the generic case of any number of alleles segregating at an autosomal locus. In fact, for the three-allele case, by treating the estimates $\hat{p}_1$, $\hat{p}_2$ and $\hat{p}_3 = 1-\left(\hat{p}_1+\hat{p}_2\right)$ as independently estimated parameters, each obtained by means of the intuitive formula

$$\hat{p}_i = \frac{2N\left(a_ia_i\right)+\sum N\left(a_ia_j\right)}{2N},$$

with i fixed and $j \neq i$ varying from 1 to k-1, that is $\sum N\left(a_ia_j\right)$ in the formula above represents the total number of heterozygous individuals as to the i allele, the corresponding formula for the variance of $\hat{F}$ is taken from

$$\left(-\frac{\partial^2 L}{\partial \hat{F}^2}\right) = \frac{1}{var\left(\hat{F}\right)} = \frac{N\left(a_1a_1\right)\left(1-\hat{p}_1\right)^2}{\left[\hat{p}_1+\left(1-\hat{p}_1\right)\hat{F}\right]^2} + \frac{N\left(a_2a_2\right)\left(1-\hat{p}_2\right)^2}{\left[\hat{p}_2+\left(1-\hat{p}_2\right)\hat{F}\right]^2} + \frac{N\left(a_3a_3\right)\left(1-\hat{p}_3\right)^2}{\left[\hat{p}_3+\left(1-\hat{p}_3\right)\hat{F}\right]^2} + \frac{N\left(a_1a_2\right)}{\left(1-\hat{F}\right)^2} + \frac{N\left(a_1a_3\right)}{\left(1-\hat{F}\right)^2} + \frac{N\left(a_2a_3\right)}{\left(1-\hat{F}\right)^2}$$

so that in the k-allele case we have

$$\left(-\frac{\partial^2 L}{\partial \hat{F}^2}\right)^{-1} = var\left(\hat{F}\right) = \left\{\sum\frac{N\left(a_ia_i\right)\left(1-\hat{p}_i\right)^2}{\left[\hat{p}_i+\left(1-\hat{p}_i\right)\hat{F}\right]^2} + \frac{\sum\sum N\left(a_ia_j\right)}{\left(1-\hat{F}\right)^2}\right\}^{-1}, \quad (3)$$

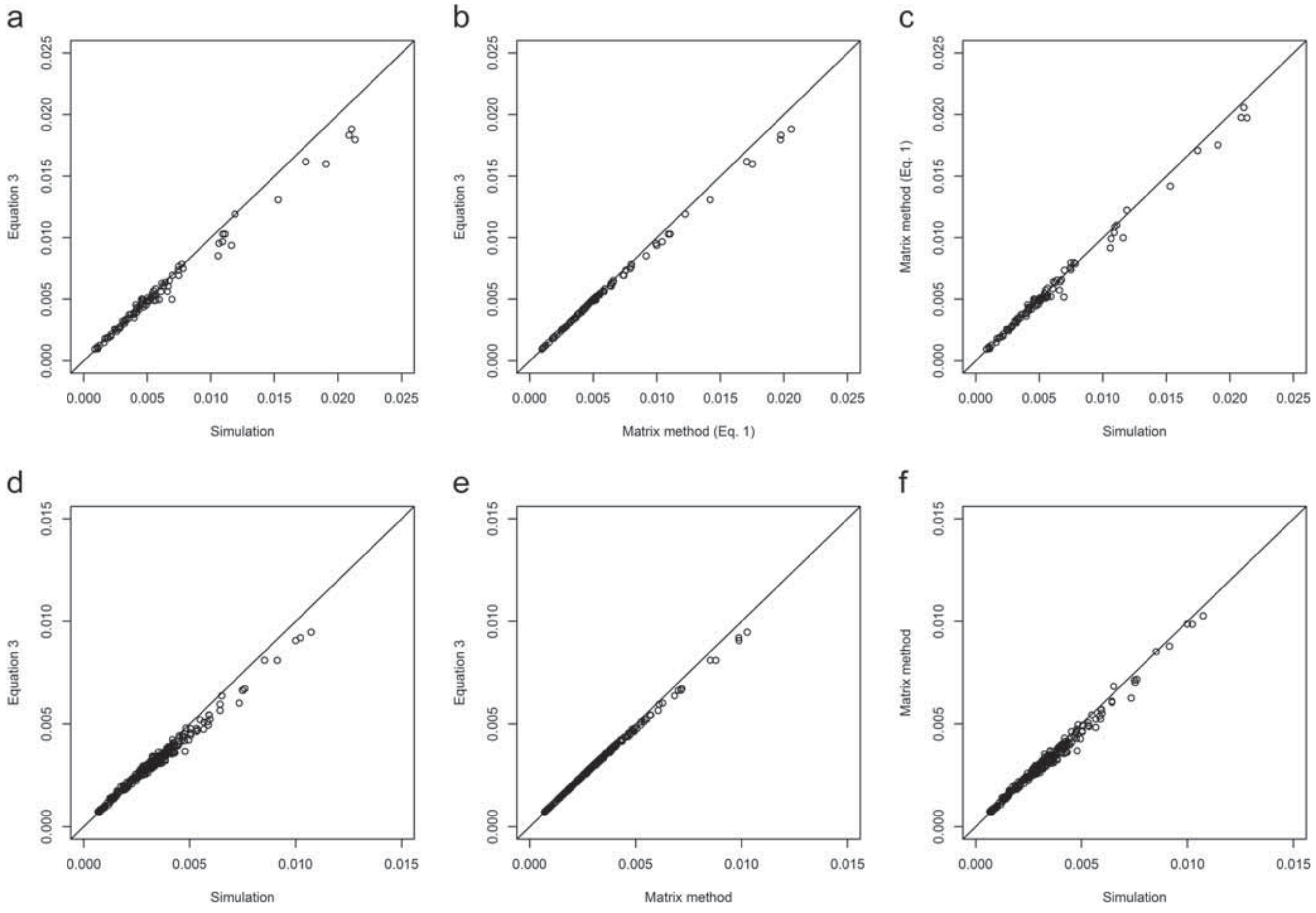

Figure 1. Comparison of values of *var* (F) corresponding to different combinations of values of p and F. In all cases p varied from 0.05 to 0.95 in intervals of 0.05, F varied from 0.1 to 0.9 in intervals of 0.1, and $N = 200$ in the cases of two alleles (graphs a,b,c) and three alleles (graphs d,e,f).

where $N(a_i a_i)$ indicates the observed number of homozygous individuals as to allele $\mathrm{a_i}$ and $N(a_i a_j)$ (with $j > i$) the observed number of heterozygous individuals as to both alleles a_i and a_j. This formula is valid for any value of $k \geq 2$, i.e. the case $k = 2$ (equation 2) is just a special case of equation 3.

Computer simulations

We also obtained values of *var* (*F*) using computer simulation methods, in which we proceeded as follows: from a relatively large number of sets of known values of F and allele frequencies $\{p_1, p_2, \ldots\}$, we determined the quantities $\{p_{11} = p_1 F + p_1^2 (1 - F),\ p_{12} = 2p_1 p_2 (1 - F),\ \ldots\}$, that were used to generate, through computer bootstrap simulations with replacement, for each combination of $\{p_1, p_2, \ldots, F\}$, 200 genotypes $\{a_1 a_1\ \ , a_1 a_2, \ldots\}$; from the genotype and allele frequencies estimated from each set of 200 genotypes so generated, we calculated the value of the fixation index *F*. The process was repeated 1000 times for each combination $\{p_1,\ p_2,\ \ldots,\ F\}$, and from the set of 1000 values of F so obtained we determined the value of *var* (*F*) after the usual formula $var\ (F) = \frac{\sum F_i^2}{1000} - \left(\frac{\sum F_i}{1000}\right)^2$. The values of *var* (*F*) obtained with different combinations of $\{p_1,\ p_2,\ \ldots,\ F\}$ could then be compared with the values calculated using the matrix method (detailed for the 2-allele case) or their corresponding approximations given by generalized equation 3.

The results we got when the values obtained (in the cases of two to six alleles) with either the simulation or the matrix method were compared to the values obtained with the approximation given by equation 3 were virtually the same beyond any reasonable doubt, as the graphs of figure 1 show for the cases of two or three alleles.

Taking into account the facts presented above, we studied, in the 2-allele case, the behaviour of the relative error, defined as $\frac{|v_1 - v_2|}{v_1}$, where v_1 and v_2 are respectively corresponding values of *var* (*F*) with same *p* and *F* obtained using equations 1 and 2. Extensive numerical analysis of the relative error showed that it is on average a bit large (its maximum value is around 11%) only when *F* has intermediate values (near 0.5) and the frequencies of the two alleles are very uneven. For other combinations of *p* and *F* the relative error is small, generally much less than 10%. For extreme *F* values (near 0 or 1) the relative error is very small (less than 2%) for any combination of allele frequencies and practically negligible when the allelic frequencies are approximately equal. The surface graph of figure 2, corresponding to the situation above discussed of two alleles and to a population size of $N = 200$, shows this in a straight forward manner. When the number of alleles was larger than two, the corresponding analyses were performed directly using the results shown by graphs as in figure 1 and the larger deviations from the diagonal line occurred exactly in the situations described for the case of two alleles, i.e. when *F* had intermediate values and allele frequencies were very uneven.

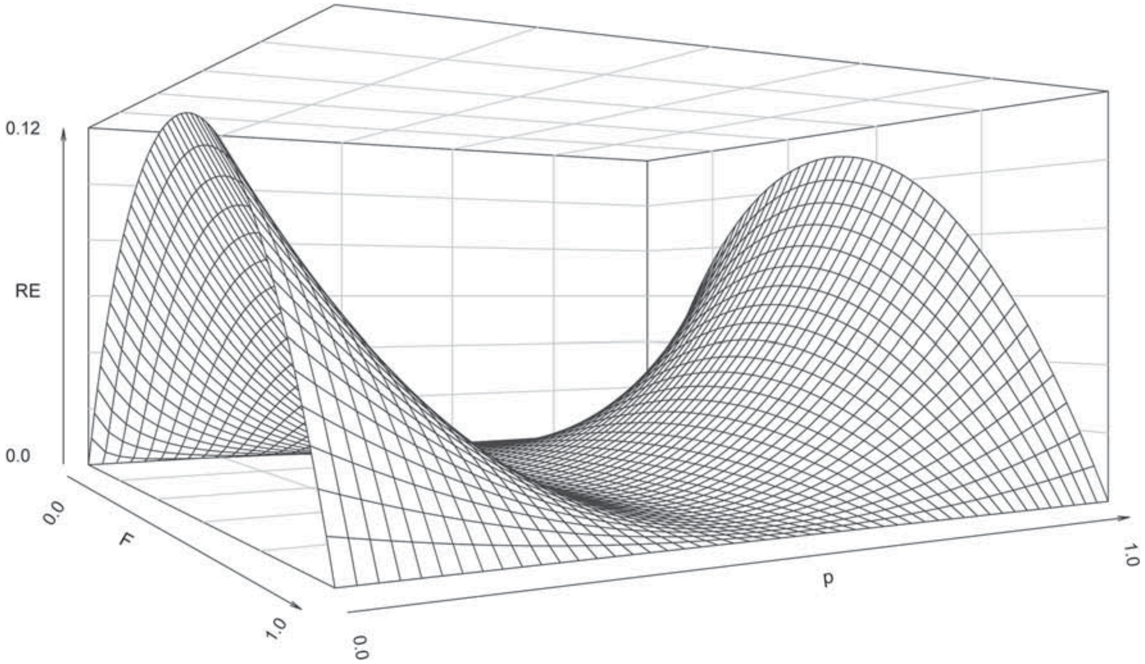

Figure 2. Relative error (RE) of *var* (*F*) values obtained using equations 1 and 2 in relation to all possible combinations of *p* and *F* for the case of two alleles. $RE = \frac{|v_1 - v_2|}{v_1}$, where $\mathrm{v_1}$ and $\mathrm{v_2}$ are corresponding values of *var* (*F*) obtained using equations 1 and 2, respectively.

Acknowledgements

We thank Drs Bruce S. Weir, Alan E. Stark and Richard M. Single for their valuable comments and suggestions on preliminary drafts of this manuscript. The very useful comments, suggestions, and important corrections made by two anonymous referees are also fully appreciated and acknowledged; most of them were incorporated (some of them literally) into the revised (final) version of the paper. This work was partly funded by grants from Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), all from Brazil.

References

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

Fyfe J. L. and Bailey N. T. J. 1951 Plant breeding studies in leguminous forage crops I. Natural cross-breeding in winter beans. *J. Agric. Sci.* **41**, 371–378.

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

Weir B. S. 1996 *Genetic data analysis II.* Sinauer Associates Inc, Sunderland, MA, USA.

Received 11 February 2015, in final revised form 11 May 2015; accepted 12 May 2015
Unedited version published online: 28 May 2014
Final version published online: 20 October 2015