Estimation of fixation indices and gene diversities

BY M. NEI* AND R. K. CHESSER†

* Center for Demographic and Population Genetics, The University of Texas at Houston, Houston, Texas 77025
† Department of Biological Sciences, Texas Tech University, Lubbock, Texas 79409

SUMMARY

Considering the multinomial sampling of genotypes, unbiased estimators of various gene diversity measures in subdivided populations are presented. Using these quantities, formulae for estimating Wright’s fixation indices \(F_{IS}, F_{IT}, \text{ and } F_{ST}\) from a finite sample are developed.

INTRODUCTION

Wright’s (1943, 1951) fixation indices \(F_{IS}, F_{IT}, \text{ and } F_{ST}\) are useful tools for studying the genetic differentiation of populations. These indices were originally defined in terms of the correlations of two uniting gametes. In practice, however, there are a number of difficulties in relating the correlations of uniting gametes or the probabilities of gene identities in subdivided populations to real-world evolutionary processes (Smith, 1974). In view of this situation Nei (1977) redefined these indices in terms of observed and expected heterozygosities and simplified the computation substantially. His definition is applicable to any diploid population whether there is selection or not or whether there are multiple alleles or not. However, his formulae are based on population allele frequencies, and he has not shown how to estimate these quantities from sample allele frequencies. Some authors (e.g. Eanes & Koehn, 1978; Avise & Felley, 1979) have attempted to make corrections for the effect of small sample size by using Levene’s (1949) method, but their corrections are not justified, because Levene’s method is based on the assumption of random mating. Such methods may result in erroneous estimations of genetic differentiation of populations. Earlier Cockerham (1973) developed statistical methods for estimating fixation indices (allelic correlations) for the case of two alleles under the assumption of no selection and no mutation, but his methods are quite complicated. Furthermore, he defines allelic correlations in reference to an imaginary population, as in the case of Wright’s definition of fixation indices. In the following we shall study the sampling properties of observed and expected heterozygosities in subdivided populations and present a simple statistical method for estimating the fixation indices defined by Nei from small samples. We shall also discuss the effect of sample size in Nei’s (1973) gene diversity analysis. A small-sample theory will be important in the future, because genetic variation is now studied at the DNA level and in this case sample size is usually very small.

FIXATION INDICES

Consider a diploid population which is subdivided into \(s\) subpopulations, and assume that there are \(r\) alleles \((A_1, A_2, \ldots, A_r)\) segregating in the population. Let \(p_{ik}\) be the frequency of allele
$A_k$ in the $i$th subpopulation, and $P_{ki}$ be the frequency of genotype $A_k A_i$ in this subpopulation. Nei (1977) has defined fixation indices $F_{IS}$, $F_{IT}$, and $F_{ST}$ in the following way:

$$F_{IS} = 1 - \frac{H_o}{H_S},$$

$$F_{IT} = 1 - \frac{H_o}{H_T},$$

$$F_{ST} = 1 - \frac{H_S}{H_T},$$

where

$$H_o = 1 - \sum_{k=1}^{r} P_{kk}, \quad H_S = 1 - \sum_{k=1}^{r} \overline{p_k}, \quad \text{and} \quad H_T = 1 - \sum_{k=1}^{r} p_k^2.$$

Here $P_{kk} = \sum_{l} w_i P_{ikl}$, $\overline{p_k} = \sum_{l} w_i \overline{p_{ikl}}$, and $\overline{p_k} = \sum_{l} w_i P_{ikl}$, in which $w_i$ is the relative size of the $i$th subpopulation with $\sum_{i} w_i = 1$. In most instances $w_i$ is not known, but $w_i = 1/s$ may be assumed, because population size is quite transitory and geneticists are interested in gene frequency differences disregarding the effect of population size (Nei, 1977). (Note that $w_i$ is not the relative sample size. The effect of the differences in relative sample size will be considered below.) $H_o$ denotes the frequency of all heterozygotes, whereas $H_S$ and $H_T$ represent the expected heterozygosities under Hardy–Weinberg equilibrium or gene diversities within subpopulations and in the total population, respectively (Nei, 1973). We note that the fixation indices defined by (1)–(3) satisfy the well-known identity

$$1 - F_{IT} = (1 - F_{IS})(1 - F_{ST}).$$

Since $p_k$ and $p_{ki}$ are not observed, our problem is how to estimate $H_o$, $H_S$, and $H_T$ from sample frequencies. We assume that $n_i$ individuals are randomly chosen from the $i$th subpopulation and the alleles under investigation are all codominant. Let $x_{ik}$ and $X_{ik}$ be the frequencies of allele $A_k$ and genotype $A_k A_i$ in the sample from the $i$th subpopulation, respectively. Estimation of $H_o$ is simple, because $x_{ik}$ is an unbiased estimate of $P_{kk}$ under our assumption. An unbiased estimate of $H_o$ in the $i$th subpopulation ($H_{oi} = 1 - \sum_{k} P_{ik}$) is $1 - \sum_{k} x_{ik}$, and thus $H_o$ may be estimated by

$$\hat{H}_o = 1 - \sum_{k} x_{ik} / s.$$

Estimation of $H_S$ is somewhat more complicated, since $x_{ik}^2$ is not an unbiased estimate of $\overline{p_k}$.

This can be seen in the following way. Obviously,

$$x_k = \left( \frac{X_{kk} + \sum_{l \neq k} X_{kl}}{2} \right),$$

where subscript $i$ is dropped for brevity. Therefore, under the multinomial sampling of genotypes, we have

$$E(x_k^2) = E(X_{kk} + X_{kk}(\sum_{l \neq k} X_{kl}) + (\sum_{l \neq k} X_{kl}^2)/2)$$

$$= P_{kk}^2 + P_{kk}(1 - P_{kk})/n + P_{kk}(\sum_{l \neq k} P_{kl})$$

$$- P_{kk}(\sum_{l \neq k} P_{kl})/n + (\sum_{l \neq k} P_{kl}^2)/4 + (\sum_{l \neq k} P_{kl}^2)/(4n)$$

$$- (\sum_{l \neq k} P_{kl}^2)/(4n)$$

$$= p_k^2 + P_{kk}/n + \sum_{l \neq k} P_{kl}/(4n) - p_k^2/n,$$
since \( p_k = P_{kk} + \Sigma_{i+k} P_{kl}/2 \). Thus, the expectation of sample gene diversity \((1 - \Sigma x^2_k)\) in the \( i \)th subpopulation is

\[
1 - E(\Sigma x^2_k) = H_{Si}(1 - 1/n_i) + H_{Si}/(2n_i),
\]

(6)

where \( H_{Si} = \Sigma p_{ik}^2 \), and an unbiased estimate of gene diversity \((H_{Si})\) is given by

\[
\bar{H}_{Si} = \frac{n_i}{n_i - 1} \left[ 1 - \Sigma x^2_k - \frac{H_{Si}}{2n_i} \right].
\]

(7)

In the case of \( r = 2 \) this is equivalent to Cockerham’s (1973) formula for \( D_{bk} \) (p. 687).

The expectation of the average of sample gene diversity over all subpopulations is

\[
1 - E(\Sigma x^2_k) = \frac{1}{s} [\Sigma_i H_{Si}(1 - 1/n_i) + H_{Si}/(2n_i)]
\]

\[
= H_{S}(1 - 1/\bar{n}) + H_{s}/(2\bar{n}),
\]

(8)

approximately, where \( x^2_k = \Sigma x^2_k/s \) and \( \bar{n} \) is the harmonic mean of \( n_i \). The above formula is approximate because \( \Sigma_i H_{Si}/n_i \) and \( \Sigma_i H_{Si}/(2n_i) \) are equated to \( H_s(\Sigma n^{-1}_i)/s \) and \( H_s(\Sigma n^{-1}_i)/2s \), respectively, under the assumption of no correlation of \( n_i \) with \( H_{Si} \) and \( H_{Si} \). When \( n_i \) is the same for all subpopulations, it is exact. Expression (8) is different from Cockerham’s equivalent quantity even for \( r = 2 \) because he has made different assumptions in his formulation. In the estimation of \( H_S \), \( E(\Sigma x^2_k) \) may be replaced by \( \Sigma_k x^2_k \) and \( H_0 \) by (5). Therefore, an unbiased estimate of \( H_S \) is given by

\[
\bar{H}_S = \frac{\bar{n}}{\bar{n} - 1} \left[ 1 - \Sigma_k x^2_k - \frac{\bar{H}_S}{2\bar{n}} \right].
\]

(9)

To get an unbiased estimate of \( H_T \), we must know the expectation of \( x^2_1 \equiv \Sigma x_{i1}/s^2 \). For simplicity, consider a locus with two codominant alleles. The mean sample frequency of allele \( A_1 \) for the entire population is given by

\[
\bar{x}_1 = \left( \Sigma_{i=1}^s x_{i1} \right)/s
\]

\[
= \left( \Sigma_{i=1}^s x_{i11} + \Sigma_{i=1}^s x_{i12}/2 \right)/s.
\]

Therefore, the expectation of \( \bar{x}_1^2 \) is

\[
E(\bar{x}_1^2) = E(\Sigma x_{i1} + \Sigma x_{i12}/2)/s^2
\]

\[
= (\Sigma x_{i1}^2 + \Sigma x_{i12}^2)/s^2
\]

\[
+ (\Sigma x_{i12}/2)^2 + \Sigma x_{i11}(1 - P_{i11})/n_i
\]

\[
- \Sigma x_{i11}/n_i + \Sigma x_{i15}(1 - P_{i15})/(4n_i)/s^2
\]

\[
= \bar{x}_1^2 + (\Sigma x_{i11} - \Sigma x_{i12}^2)/(n_i s^2) + \Sigma x_{i12}/(4n_i s^2).
\]

A similar expression can be obtained for \( E(\bar{x}_2^2) \). Thus, noting

\[
H_T = 1 - \bar{x}_1^2 - \bar{x}_2^2, \quad H_S = \Sigma_i (1 - p_{i1}^2 - p_{i2}^2)/s,
\]

and

\[
H_0 = \Sigma_i P_{i12}/s = \Sigma_i (1 - P_{i11} - P_{i22})/s,
\]

we have

\[
1 - E(\bar{x}_1^2 + \bar{x}_2^2) = H_T - H_S/(\bar{n}s) + H_0/(2\bar{n}s),
\]
Table 1. Genotype frequencies for the acid phosphatase locus in the Jewish populations who moved to Israel from Iran, Iraq and Yemen

Means were calculated assuming \( w_i = 1/s \), i.e. all populations are given equal weight. Data from Mourant et al. (1976).

<table>
<thead>
<tr>
<th>Population</th>
<th>Sample size (n)</th>
<th>AA</th>
<th>AB</th>
<th>AC</th>
<th>BB</th>
<th>BC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iran</td>
<td>49</td>
<td>0.2041</td>
<td>0.3265</td>
<td>—</td>
<td>0.4286</td>
<td>0.0408</td>
</tr>
<tr>
<td>Iraq</td>
<td>82</td>
<td>0.1342</td>
<td>0.4024</td>
<td>0.0122</td>
<td>0.4268</td>
<td>0.0244</td>
</tr>
<tr>
<td>Yemen</td>
<td>37</td>
<td>0.0270</td>
<td>0.2433</td>
<td>—</td>
<td>0.7027</td>
<td>0.0270</td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td>0.1218</td>
<td>0.3241</td>
<td>0.0040</td>
<td>0.5194</td>
<td>0.0307</td>
</tr>
</tbody>
</table>

approximately. This can easily be extended to the case of \( r \) alleles, and it becomes

\[
1 - E(\Sigma_k x_k^2) = H_T - H_S/(\bar{\bar{n}}s) + H_0/(2\bar{\bar{n}}s),
\]

where \( H_T = 1 - \Sigma_k p_{kk}^2 \), \( H_S = \Sigma_k(1 - \Sigma_k p_{kk}^2)/s \), and \( H_0 = \Sigma_k(1 - \Sigma_k P_{kk})/s \). Thus, an unbiased estimate of \( H_T \) is given by

\[
\hat{H}_T = 1 - \Sigma_k \hat{x}_k^2 + \hat{H}_S/(\bar{\bar{n}}s) - \hat{H}_0/(2\bar{\bar{n}}s),
\]

where \( \hat{H}_S \) and \( \hat{H}_0 \) are given by (9) and (5), respectively.

Using \( \hat{H}_0 \), \( \hat{H}_S \), and \( \hat{H}_T \), we can now estimate the fixation indices by

\[
\hat{F}_{IS} = 1 - \hat{H}_S/\hat{H}_S,
\]

\[
\hat{F}_{IT} = 1 - \hat{H}_0/\hat{H}_T,
\]

\[
\hat{F}_{ST} = 1 - \hat{H}_S/\hat{H}_T.
\]

These estimates again satisfy Wright's equation \((1 - \hat{F}_{IT}) = (1 - \hat{F}_{IS})(1 - \hat{F}_{ST})\). It should be noted, however, that \( \hat{F}_{IS} \), \( \hat{F}_{IT} \), and \( \hat{F}_{ST} \) are not unbiased estimates, though \( \hat{H}_0 \), \( \hat{H}_S \), and \( \hat{H}_T \) are all unbiased estimates. This is because the expectations of \( \hat{H}_0/\hat{H}_S \), \( \hat{H}_0/\hat{H}_T \), and \( \hat{H}_S/\hat{H}_T \) are not equal to \( E(\hat{H}_0)/E(\hat{H}_S) \), \( E(\hat{H}_0)/E(\hat{H}_T) \), and \( E(\hat{H}_S)/E(\hat{H}_T) \), respectively. Nevertheless, the latter are known to be generally very close to the former (e.g. Nei & Chakravarti, 1977), so that these estimates are better quantities than those obtained by (1), (2), and (3) when sample size is small.

The sampling variances of \( \hat{F}_{IS} \), \( \hat{F}_{IT} \), and \( \hat{F}_{ST} \) can be obtained by Nei & Chakravarti's (1977) method. However, their computation is quite tedious. For testing the statistical significances of these quantities we had better use the \( \chi^2 \) test illustrated in the following example.

**Numerical example**

Mourant, Kopéc & Domaniewska-Sobczak (1976) list the genotype frequencies for the acid phosphatase locus in Jewish populations from Iran, Iraq, and Yemen. The results obtained are given in Table 1. From this Table, we obtain \( \hat{H}_S = 1 - (0.1218 + 0.5194) = 0.3588 \). To estimate \( H_S \), we must compute the mean \( (1 - \Sigma x_k^2) \) of \( 1 - \Sigma x_k^2 \) for the three populations. This becomes 0.4009 from Table 2. In the present case the harmonic mean of sample size (\( \bar{\bar{n}} \)) is 50.3. Therefore, using (9) we have

\[
\hat{H}_S = \frac{50.3 - 1}{50.3} \times \frac{0.4009 - 0.3588}{100.6}
\]

\[= 0.4054.\]
Table 2. Gene frequencies and some other quantities obtained from the values of Table 1

<table>
<thead>
<tr>
<th>Population</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>1 - (\Sigma x^2_i)</th>
<th>(\hat{F}_{IS})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iran</td>
<td>0.3674</td>
<td>0.6122</td>
<td>0.0204</td>
<td>0.04898</td>
<td>0.2598</td>
</tr>
<tr>
<td>Iraq</td>
<td>0.3415</td>
<td>0.6402</td>
<td>0.0183</td>
<td>0.04372</td>
<td>0.0020</td>
</tr>
<tr>
<td>Yemen</td>
<td>0.1487</td>
<td>0.8513</td>
<td>0.0135</td>
<td>0.02573</td>
<td>0.0316</td>
</tr>
<tr>
<td>Mean</td>
<td>0.2859</td>
<td>0.6967</td>
<td>0.0174</td>
<td>0.04009</td>
<td>0.1149</td>
</tr>
</tbody>
</table>

The value of \(1 - \Sigma x^2_i\) can be computed from the mean gene frequencies in Table 2. It becomes 0.4326. Therefore, from (11), \(\hat{H}_T = 0.4326 + 0.4054/150.9 - 0.3588/301.8 = 0.4341\). We can now obtain the estimates of \(\hat{F}_{IS}\), \(\hat{F}_{IT}\), and \(\hat{F}_{ST}\). They are \(\hat{F}_{IS} = 0.1149\), \(\hat{F}_{IT} = 0.1735\) and \(\hat{F}_{ST} = 0.0661\).

The values indicate that there is a homozygote excess in these populations and a large part of the excess occurs within populations, \(\hat{F}_{IS}\) being 0.1149. It is possible to compute \(\hat{F}_{IS}\) for each subpopulation, applying (5) and (9) with \(s = 1\). The values obtained are given in Table 2. It is clear that a large deviation from Hardy–Weinberg proportions exists in Iranian Jews. The null hypothesis of \(\hat{F}_{IS} = 0\) can be tested by the \(\chi^2\) test of the differences between the observed and expected genotype frequencies. In the present case, however, this test is not appropriate because there are only four genotypes observed and the number of observed individuals for genotype BC is only 2 in Iranian Jews. A more appropriate test for this case is to compute \(\hat{F}_{IS}\) for each allele and then use \(\chi^2_{(1)} = n\hat{F}_{IS}\) (Li & Horvitz, 1953). For example, the \(\hat{F}_{IS}\) for allele A can be computed by treating all other alleles as a single non-A allele (N) and assuming that there is only one pair of alleles, A and N. Application of this method to Iranian Jews gives \(\hat{F}_{IS} = 0.3070\). Therefore, \(\chi^2_{(1)} = 4.6\) is significant at the 5% level. This suggests that in this population there is inbreeding or some other factor that increases the frequency of homozygotes. However, the \(\hat{F}_{IS}\) for allele B is 0.2362, and \(\chi^2\) is not significant. Computation of \(\hat{F}_{IS}\) for allele C is meaningless in the present case, because this allele is very rare (Table 2).

Our estimate of \(\hat{F}_{ST}\) is quite large compared with other estimates (e.g. Japanese subpopulations, Nei & Imaizumi, 1966). This is mainly due to the deviation of the gene frequencies in Yemen Jews from those of the other two populations. At the present time there seems to be no authentic test of the difference between two estimates of \(\hat{F}_{ST}\). The statistical significance of \(\hat{F}_{ST}\) from 0 can be tested by the usual \(\chi^2\) test of heterogeneity of gene frequencies (Workman & Niswander, 1970). In the present case \(\chi^2_{(0)} = 11.73\), which is significant at the 5% level.

Let us now compare our estimates of fixation indices with those obtainable from (1), (2), and (3) by equating sample allele frequencies to population frequencies. The latter values are \(\hat{F}_{IS} = 0.1050\), \(\hat{F}_{IT} = 0.1706\), and \(\hat{F}_{ST} = 0.0733\). Therefore, if we use (1), (2), and (3), \(\hat{F}_{IS}\) and \(\hat{F}_{IT}\) are underestimated, whereas \(\hat{F}_{ST}\) is overestimated.

**Gene Diversity Analysis**

In gene diversity analysis only the expected heterozygosities under Hardy–Weinberg equilibrium are considered, irrespective of the distribution of genotype frequencies (Nei, 1973). Nei’s previous theory of gene diversity analysis is based on the assumption of Hardy–Weinberg
equilibrium. Thus, \( H_0 = H_S \) is assumed, and the gene diversity within subpopulations is estimated by

\[
\hat{H}_S = \frac{2\hat{n}}{2\hat{n} - 1} (1 - \Sigma_k \bar{x}_k^2),
\]

(15)

which is a special case of (9) (Nei, 1978). Similarly, \( \hat{H}_T \) is estimated by

\[
\hat{H}_T = 1 - \Sigma_k x_k^2 + \hat{H}_S/(2\bar{n}_s),
\]

(16)

under the same condition. However, if there is any reason to suspect that the population is not in Hardy–Weinberg equilibrium, better estimates of \( H_S \) and \( H_T \) are given by (9) and (11), respectively. Of course, gene diversity analysis is intended to be applied for many loci simultaneously, so that the averages (\( \bar{H}_S \) and \( \bar{H}_T \)) of \( H_S \) and \( H_T \) over all loci should be used.

Nei’s (1973) \( D_{ST} \) and \( G_{ST} \) statistics are then estimated by \( \bar{D}_{ST} = \bar{H}_T - \bar{H}_S \) and \( \bar{G}_{ST} = \bar{D}_{ST}/\bar{H}_T \), respectively.

It is interesting to note that in selfing populations \( H_0 \) is generally 0, so that we have

\[
\hat{H}_S = \frac{\hat{n}}{\hat{n} - 1} (1 - \Sigma_k \bar{x}_k^2),
\]

(17)

\[
\hat{H}_T = 1 - \Sigma_k x_k^2 + \hat{H}_S/(\bar{n}_s).
\]

(18)

Therefore, the sample size for selfing populations must be two times larger than that for the case of random mating populations if one wants to get estimates of \( H_S \) and \( H_T \) with the same accuracy.

**DISCUSSION**

Since the invention of fixation indices, various methods have been used for estimating fixation indices from a sample of finite size. Wright (1943) and Nei & Imaizumi (1966) corrected for sampling errors for the estimates of \( F_{ST} \) but not for those of \( F_{IS} \) and \( F_{IT} \). They considered a single allele and estimated \( F_{ST} \) by \( (\sigma^2 - \sigma^2_1)/(\bar{x}(1 - \bar{x})) \), where \( \sigma^2 \) is the variance of allele frequency \( x \) among subpopulations and \( \sigma^2_1 \) is the sampling variance and given by \( \Sigma_x (x - \bar{x})/(2n) \), where sample size is assumed to be the same for all subpopulations. This method is equivalent to our method if genotype frequencies are in Hardy–Weinberg equilibrium and \( s \) is large. In this case \( H_S = H_0 \), so that

\[
1 - E(\Sigma_k x_k^2) - (1 - E(\Sigma_k \bar{x}_k^2)) = \Sigma_k \sigma_k^2
\]

\[
= H_T - H_S + H_S/(2n),
\]

(19)

where \( \sigma_k^2 \) is the variance of the frequency of the \( k \)th allele. When there are only two alleles, \( \sigma_1^2 = \sigma_2^2 = \sigma^2 \), and \( H_S/(2n) = 2x_1(1 - x_1)/(2n) = 2\sigma^2 \), Furthermore, \( H_T = 2\bar{x}(1 - \bar{x}) \). Therefore, \( F_{ST} = (H_T - H_S)/H_T = (\sigma^2 - \sigma_1^2)/(\bar{x}(1 - \bar{x})) \), which is identical with the formula used by Wright (1943). However, the method described in this paper is more general than Wright’s formula and applicable irrespective of the number of alleles involved and the distribution of genotype frequencies. Furthermore, in this method sample size correction is made for all of \( \bar{F}_{IS} \), \( \bar{F}_{IT} \), and \( \bar{F}_{ST} \).

Some authors (e.g. Eanes & Koehn, 1978; Avise & Felley, 1979) used Levene’s (1949) method for estimating the homozygote or heterozygote frequencies in the population and then estimated \( F_{IS} \). When the population is in Hardy–Weinberg equilibrium, this method is equivalent to our
formula for $F_{IS}$. However, this correction is not valid, since we are estimating $F_{IS}$ under the assumption that the population is not in Hardy–Weinberg equilibrium.

When sample size is small, the estimate of $F_{ST}$ may be improved by assuming $F_{IS} = 0$ if $\hat{F}_{IS}$ does not significantly deviate from 0 and there is good reason to believe that the population is in Hardy–Weinberg equilibrium. In this case $H_q = H_S$, so that we need to estimate only $H_S$ and $H_T$. Therefore, the variance of the estimate of $F_{ST}$ (and $F_{IT}$) would be reduced. In many cases the most important parameter of population differentiation is $F_{ST}$.

Finally, it should be noted that, although fixation indices are estimable from a small sample, a large number of individuals should be sampled if possible. As is clear from the example considered earlier, estimates of fixation indices are subject to a large sampling error when sample size is small. Therefore, meaningful conclusions about the genetic structure of populations cannot be obtained unless sample size is sufficiently large.

This study was supported by research grants from the National Science Foundation and the National Institutes of Health.

REFERENCES


