# AMOSTRAGEM PARA ESTIMATIVA DE FREQÜÊNCIAS ALÉLICAS E ÍNDICES DE DIVERSIDADE GENÉTICA EM ESPÉCIES ARBÓREAS*

Alexandre Magno SEBBENN**


## RESUMO

O objetivo deste trabalho foi determinar o tamanho amostral adequado para a estimativa de freqüências alélicas e índices de diversidade genética dentro de populações em espécies arbóreas, quando dados de marcadores genéticos codominantes são avaliados. Foram consideradas as situações em que o objetivo é a estimativa das freqüências alélicas e índices de diversidade em uma ou várias populações de árvores adultas e progênies. Os tamanhos amostrais foram determinados com base em variâncias de freqüências gênicas, incorporando informações do sistema de reprodução, como taxa de autofecundação. O tamanho amostral ideal para a avaliação de árvores adultas em uma ou várias populações encontra-se entre o número de 60 a 100 árvores. Para amostrar progênies em uma população de uma espécie alógama é necessário avaliar pelo menos 10 sementes/progênie coletadas de 30 árvores matrizes. Caso a espécie se reproduza por autogamia, esse tamanho amostral deve ser de 50 progênies e 10 plantas/progênie. Finalmente, se o objetivo for avaliar progênies em várias populações será necessário amostrar pelo menos 10 sementes/progênie, coletadas em 30 a 10 árvores por população.


Palavras-chave: amostragem; freqüências alélicas; índices de diversidade genética; alelos raros; conservação genética; deriva genética.


## ABSTRACT

The aim of this work was to determine the ideal sample size for the estimation of allele frequencies and genetic diversity indexes in forest species, when data of codominant genetic markers are appraised. For so much, it was determined the sample size when objective to evaluate individuals in a simple population, individuals in several populations, populations of a species, families and progenies/families in a simple population and families and progenies/families in several populations. The sample size was estimated based on variances of gene frequencies, incorporating information of the mating system of species, as self-fertilization rate. The ideal sample size to evaluate adults individuals in one or several populations was determined to be among 60 to 100 individuals. In a population structured in families, when alogamus or mixed mating system, predominantly alogamus, species are used, the sampling of 30 families and 10 progenies/family are enough. To autogamus species, the sample size should be of 50 families and 10 progenies/families. Finally, to evaluate many populations structured in families it is necessary to adopt the sample of 30 to 10 families/population and 10 progenies/family.


Key words: sampling; allele frequencies; indexes of genetic diversity; rare alleles; genetic conservation; genetic drift.

## 1 INTRODUÇÃO

A diversidade genética em espécies arbóreas concentra-se principalmente dentro das populações (Hamrick *et al.*, 1979; Hamrick & Godt, 1989). Assim, caracterizar os níveis da diversidade genética dentro das populações é de importância primária em qualquer estudo evolutivo e trabalhos de conservação, melhoramento e manejo florestal. Esses níveis são caracterizados pelos índices de diversidade genética dentro de populações como: heterozigosidade observada, heterozigosidade esperada em Equilíbrio de Hardy-Weinberg, índice de fixação, porcentagem de locos polimórficos, número efetivo de alelos por loco e número de alelos por locos, estimados a partir de marcadores genéticos codominantes. Com exceção da heterozigosidade observada, os demais índices são dependentes da estimativa das freqüências alélicas, portanto, é fundamental estimar estas freqüências com precisão. Estes índices também permitem comparar os níveis de diversidade genética entre populações de uma espécie e entre diferentes espécies (Hamrick & Godt, 1989; Hamrick *et al.*, 1992).

(*) Aceito para publicação em janeiro de 2002.
(**) Instituto Florestal, Caixa Postal 1322, 01059-970, São Paulo, SP, Brasil.

Quando indivíduos de uma população ou diferentes espécies são comparados em termos genéticos, podem existir duas fontes de variação, a variância interlocos e a variância intralocos. A variância interlocos é causada pela amostragem de um número limitado de locos nos indivíduos. Aumentando-se o número de locos avaliados, reduz-se a variância interlocos (Nei, 1977). A grande maioria dos trabalhos realizados com espécies arbóreas, baseados em dados de marcadores genéticos, tem revelado de 10 a 30 locos (Berg & Hamrick, 1997). Estudos apontam que amostras de 20 locos já proporcionam estimativas confiáveis dos índices de diversidade e medidas de estrutura de populações, sendo poucas as alterações observadas nos parâmetros a partir deste número (Ayala & Kinger, 1984). Por outro lado, a variância intralocos é causada pela amostragem limitada de indivíduos em uma população, e é o objeto deste estudo. Aumentando-se o número de indivíduos amostrados, reduz-se a variância intralocos (Nei, 1977). El-Kassaby & Sziklai (1983) determinaram que amostras entre 42 a 60 árvores por população seriam suficientes para a obtenção de estimativas confiáveis das freqüências alélicas. No entanto, os autores não consideraram o efeito do tamanho amostral na precisão das estimativas dos índices de diversidade e as possíveis variações na taxa de cruzamento, sendo assumida alogamia total para as espécies. Nas espécies arbóreas, desvios de cruzamentos aleatórios são comuns, portanto, é necessário considerar variações na taxa de cruzamento na determinação do tamanho amostral adequado para estimar freqüências alélicas e índices de diversidade.

Este trabalho tem por objetivo determinar tamanhos amostrais adequados para estimativas de freqüências alélicas e índices de diversidade. Para isso, estudou-se o efeito do tamanho amostral e das variações no sistema de reprodução no erro das respectivas estimativas. Foram consideradas as seguintes situações de amostragem: a) amostragem de árvores adultas, jovens, plântulas ou do banco de sementes em uma ou várias populações, e b) amostragem de progênies e plantas/progênie em uma ou várias populações.

## 2 MATERIAL E MÉTODOS

Para estimar freqüências alélicas e índices de diversidade em uma simples população, é necessário definir o tamanho amostral em função do erro requerido. O erro na estimativa da freqüência de um alelo é máximo quando $p = 0,5$ (El-Kassaby & Sziklai, 1983), portanto, esta freqüência foi utilizada como referência para determinar o tamanho amostral. Os erros foram estimados em forma de intervalos de confiança, considerando as variações no sistema de reprodução das espécies. Foram abordadas duas situações de amostragem: primeiro, quando se pretende amostrar árvores adultas, jovens, plântulas ou do banco de sementes em uma população (amostragem individual) e, segundo, quando se pretende amostrar progênies em uma população (amostragem de progênies).

### 2.1 Amostragem de Indivíduos em uma População

Em uma população de indivíduos diplóides, em Equilíbrio de Endogamia de Wright (EEW), a variância na estimativa da freqüência de um alelo $p_i$ ($\hat{\sigma}^2_{(p_i)}$) pode ser obtida segundo Weir (1996) por:

$$\hat{\sigma}^2_{(p_i)} = \frac{p_i(1-p_i)(1+\hat{f})}{2n} \qquad (1)$$

em que, $\hat{f}$ representa o coeficiente de endogamia da população, estimado em EEW pela taxa de autofecundação ($\hat{s}$) por $\hat{f} = \hat{s}/(2-\hat{s})$ (2) e, $n$ é o número de indivíduos amostrados na população. Se o coeficiente de endogamia for zero (população em Equilíbrio de Hardy-Weinberg-EHW), a expressão 1 reduz-se a $\hat{\sigma}^2_{(p_i)} = p_i(1-p_i)/2n$ (3) (Weir, 1996).

O intervalo de confiança do erro da estimativa de $p_i$ ($IC_{(p_i)}$) para as equações 1 e 3 pode ser obtido por $IC_{(p_i)} = p_i \pm t\sqrt{\hat{\sigma}^2_{(p_i)}}$ (4) (El-Kassaby & Sziklai, 1983), sendo $t$ = t tabelado de Student para uma probabilidade prefixada.

Para avaliar a eficiência da amostragem determinada pela expressão 1, na estimativa das freqüências alélicas e índices de diversidade genética como heterozigosidade observada ( $\hat{H}_o$ ), heterozigosidade esperada segundo o EHW ( $\hat{H}_e$ ) e índice de fixação ( $\hat{f}$ ), criou-se uma população hipotética infinita, em EHW, constituída por 2.000 ($N$) indivíduos diplóides, onde três locos foram avaliados: Loco 1 com dois alelos ($p$ e $q$); Loco 2 com três alelos ($p$, $q$ e $r$), e Loco 3 com quatro alelos ($p$, $q$, $r$ e $s$). As freqüências genotípicas para estes locos foram obtidas pelos modelos de EHW apresentados na TABELA 1.

TABELA 1 - Modelo de Equilíbrio de Hardy-Weinberg para locos com 2 alelos (Loco 1), 3 alelos (Loco 2) e 4 alelos (Loco 3).

| Classes | Loco 1 | Loco 2 | Loco 3 |
|---|---|---|---|
| $A_1A_1$ | $p^2N$ | $p^2N$ | $p^2$ |
| $A_1A_2$ | 2pqN | 2pqN | 2pqN |
| $A_1A_3$ | ---- | 2prN | 2prN |
| $A_1A_4$ | ---- | ---- | 2psN |
| $A_2A_2$ | $q^2N$ | $q^2N$ | $q^2N$ |
| $A_2A_3$ | ---- | 2qrN | 2qrN |
| $A_2A_4$ | ---- | ---- | 2qsN |
| $A_3A_3$ | ---- | $r^2N$ | $r^2N$ |
| $A_3A_4$ | ---- | ---- | 2rsN |
| $A_4A_4$ | ---- | ---- | $s^2N$ |

Dessa população hipotética foram retiradas aleatoriamente 100 amostras de 10, 30, 60, 100, 150 e 200 indivíduos, com reposição, e estimadas as freqüências alélicas, e os índices $\hat{H}_o$, $\hat{H}_e$ e $\hat{f}$. Foram obtidas, em seguida, a média e o erro padrão das 100 estimativas. Os índices de diversidade foram estimados de acordo com Berg & Hamrick (1997) da seguinte forma: as freqüências alélicas foram obtidas por $\hat{p}_{ij} = n_{ij} / n_{.j}$ (5), em que: $\hat{p}_{ij}$ é a freqüência do alelo $i$, no loco $j$; $n_{ij}$ é o número de ocorrência do alelo $i$, no loco $j$; $n_{.j}$ é o número total de alelos amostrados, no loco; a heterozigosidade observada ( $\hat{H}_o$ ) foi estimada por: $\hat{H}_o = 1 - \sum P_{ii}$ (6), em que: $P_{ii}$ é a freqüência dos genótipos homozigotos, e a heterozigosidade esperada em EHW foi estimada por: $\hat{H}_e = 1 - \sum \hat{p}_{ij}^2$ (7) (Nei, 1977) e o índice de fixação não viesado foi estimado de acordo com Weir (1996):

$$\hat{f} = \frac{(\hat{H}_e - \hat{H}_o) + \frac{1}{2n}\hat{H}_o}{\hat{H}_e - \frac{1}{2n}\hat{H}_o} \qquad (8)$$

em que, $n$ é o número de indivíduos amostrados por população.

## 2.2 Amostragem de Progênies

A estimativa da variância amostral das freqüências gênicas ( $\hat{\sigma}^2_{(p_i)}$ ) em uma população estruturada em progênies foi obtida com base em Brown & Weir (1983) pela expressão:

$$\hat{\sigma}^2_{(p_i)} = \left[1 + \frac{(k-1)(1+\hat{s})^2}{4}\right]\left[\frac{p_i(1-p_i)}{n(2-\hat{s})}\right] \qquad (9)$$

em que, $k$ é o número de plantas por progênie; $\hat{s}$ é a taxa de autofecundação; $p_i$ é a freqüência do alelo $i$ na população, e $n$ é o tamanho amostral total (para amostragem do mesmo número de plantas por progênie, $n = m$ x $k$, sendo $m$ o número de progênies amostradas). O intervalo de confiança do erro da estimativa de $p_i$ para populações estruturadas em progênies também foi calculado pela expressão 4.

## 3 RESULTADOS E DISCUSSÕES

### 3.1 Amostragem de Indivíduos em uma População

Os intervalos de confiança do erro da freqüência alélica de $p_i$ = 0,5 a 95% de probabilidade ($\alpha$ = 0,05) são dados para diferentes tamanhos amostrais ($\hat{n}$) na TABELA 2. Para estas estimativas, admitiu-se ausência de parentesco ($\hat{\theta}$ = 0), indivíduos diplóides e população infinita em EEW.

TABELA 2 - Intervalo de confiança ($p_i$ ± IC) a 95% de probabilidade para diferentes tamanhos de amostras ($\hat{n}$), em função da taxa de autofecundação ($s$), em populações em equilíbrio de endogamia.

| | | Tamanho amostral ($\hat{n}$) | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $s$ | $\hat{f}$ | 10 | 20 | 30 | 50 | 60 | 100 | 150 | 200 |
| 0,000 | 0,000 | 0,253 | 0,165 | 0,132 | 0,101 | 0,091 | 0,071 | 0,057 | 0,049 |
| 0,050 | 0,026 | 0,256 | 0,168 | 0,133 | 0,102 | 0,092 | 0,072 | 0,057 | 0,050 |
| 0,100 | 0,053 | 0,259 | 0,170 | 0,135 | 0,103 | 0,094 | 0,073 | 0,058 | 0,050 |
| 0,200 | 0,111 | 0,267 | 0,174 | 0,139 | 0,106 | 0,096 | 0,075 | 0,060 | 0,052 |
| 0,300 | 0,176 | 0,274 | 0,179 | 0,143 | 0,109 | 0,099 | 0,077 | 0,061 | 0,053 |
| 0,500 | 0,333 | 0,292 | 0,191 | 0,152 | 0,116 | 0,105 | 0,082 | 0,065 | 0,057 |
| 1,000 | 1,000 | 0,358 | 0,234 | 0,186 | 0,142 | 0,129 | 0,100 | 0,080 | 0,069 |

Onde: $\hat{f}$ = coeficiente de endogamia em equilíbrio de endogamia.

Os resultados mostraram que quanto maior o tamanho amostral e menor a taxa de autofecundação, menor é o erro associado à estimativa de $\hat{p}_i$, sendo o inverso também verdadeiro. Assumiu-se estimativas das freqüências alélicas dentro do intervalo de 15% de $\hat{p}_i$ = 0,5 (0,075: $IC$ = 0,425 a 0,575) como boas e 20% (0,100: $IC$ = 0,400 a 0,500) como razoáveis. Assim, verifica-se que para se obter estimativas de $\hat{p}_i$ dentro do intervalo estabelecido como bom, até uma taxa de 20% de autofecundação a amostragem de 100 indivíduos garante com 95% de probabilidade que a verdadeira freqüência alélica da população estará entre o intervalo de 0,425 a 0,574. Para o intervalo considerado como razoável, a amostragem de 60 árvores garante com 95% de probabilidade, até uma taxa de 30% de autofecundação, que a freqüência de $p_i$ estará entre 0,401 a 0,599. Tais resultados revelam que para se obter estimativas razoáveis de freqüências alélicas é necessária a adoção de grandes tamanhos amostrais. Em estudos de populações naturais de espécies arbóreas por marcadores genéticos codominantes têm-se utilizado o tamanho amostral de 30 árvores/população. Porém, de acordo com os resultados da TABELA 1, a amostragem de 30 plantas gera estimativas com baixa precisão, sendo o intervalo obtido para $p_i$, a 95% de probabilidade e admitindo-se espécies arbóreas alógamas, de 0,368 a 0,632. El-Kassaby & Sziklai (1983) recomendaram para a estimativa de freqüências alélicas a amostragem de 42 a 60 árvores. Quando os autores estudaram os erros associados às estimavas de $p_i$, em uma população de *Pseudotsuga menziensii*, com taxa de autofecundação de 10%, observaram que estes tamanhos amostrais não eram suficientes para garantir precisão nas estimativas, em especial quando os alelos nos locos apresentavam freqüências homogêneas ($1/n_a$, sendo $n_a$ o número de alelos no loco). Para os tamanhos amostrais de 40 e 60 árvores, em populações praticando 10% de autofecundação, a freqüência estimada para o alelo $p_i$ estará entre o intervalo de 0,384 a 0,616, e 0,406 a 0,594, respectivamente, confirmando a baixa precisão obtida por El-Kassaby & Sziklai (1983). Estes resultados reforçam a conclusão de que pequenos tamanhos amostrais são insuficientes para se obter estimativas precisas nas freqüências alélicas. Assim, sugere-se a amostragem de 60 a 100 árvores como suficiente para estimar freqüências alélicas em uma população.

Em casos de populações pequenas, por exemplo, 15 árvores, recomenda-se a "amostragem" de todos os exemplares (censo) e o desprezo da estimativa de erro dos parâmetros em nível de locos individuais. O erro amostral obtido em nível de locos será fictício porque todas as plantas estão sendo representadas na amostra; logo, não existe erro intralocos e a estimativa obtida é o próprio parâmetro populacional. Porém, o erro ou intervalo de confiança em nível de média de locos permanece, porque estará medindo outro nível do processo amostral, a amostragem de locos para representar o genoma dos indivíduos.

O tamanho amostral de 60 a 100 árvores/população pode também ser adotado em estudos da distribuição da variabilidade genética entre e dentro de populações. Como anteriormente demonstrado, o tamanho amostral de 60 árvores/população garante, a 95% de probabilidade, que as freqüências alélicas dentro das populações não excedam a 20% da freqüência de $p_i$, até uma taxa de 30% de autofecundação nas populações. Apesar do intervalo amplo, como o principal objetivo é a medida de divergência entre populações ($G_{ST} \approx F_{ST} \approx \theta_P$), a amostragem de pelo menos quatro populações resultará em um tamanho amostral total de 240 árvores, gerando boas estimativas das freqüências alélicas médias entre populações, medidas estas fundamentais para uma boa precisão da divergência genética entre populações (Nei, 1973).

### 3.2 Precisão na Estimativa dos Índices de Diversidade

Na TABELA 3 são apresentadas as médias e os erros das estimativas das freqüências alélicas e na TABELA 4 a média e os erros das estimativas da heterozigosidade observada ($\hat{H}_e$), heterozigosidade esperada ($\hat{H}_e$) e índice de fixação ($\hat{f}$).

TABELA 3 - Freqüências alélicas médias ($\hat{p}_i$ e desvio padrão ($\hat{\sigma}$) de 100 amostras aleatórias de diferentes tamanhos ($n$), em uma população em Equilíbrio de Hardy-Weinberg, para três locos.

| | Média de 100 amostras aleatórias | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Loco 1 | | Loco 2 | | | | | |
| $n$ | $p = 0,500^a$ | $\hat{\sigma}$ | $p = 0,500$ | $\hat{\sigma}$ | $q = 0,450$ | $\hat{\sigma}$ | $r = 0,050$ | $\hat{\sigma}$ |
| 10 | 0,510 | 0,102 | 0,508 | 0,114 | 0,447 | 0,110 | 0,045 | 0,049 |
| 30 | 0,491 | 0,065 | 0,497 | 0,061 | 0,451 | 0,060 | 0,052 | 0,028 |
| 60 | 0,494 | 0,050 | 0,496 | 0,043 | 0,453 | 0,041 | 0,051 | 0,019 |
| 100 | 0,500 | 0,040 | 0,499 | 0,039 | 0,452 | 0,039 | 0,050 | 0,014 |
| 150 | 0,498 | 0,030 | 0,498 | 0,029 | 0,451 | 0,029 | 0,050 | 0,012 |
| | Loco 3 | | | | | | | |
| $n$ | $p = 0,500$ | $\hat{\sigma}$ | $q = 0,300$ | $\hat{\sigma}$ | $r = 0,150$ | $\hat{\sigma}$ | $s = 0,050$ | $\hat{\sigma}$ |
| 10 | 0,516 | 0,111 | 0,289 | 0,109 | 0,143 | 0,076 | 0,053 | 0,045 |
| 30 | 0,506 | 0,065 | 0,295 | 0,059 | 0,148 | 0,046 | 0,051 | 0,025 |
| 60 | 0,497 | 0,053 | 0,300 | 0,044 | 0,153 | 0,039 | 0,050 | 0,019 |
| 100 | 0,502 | 0,042 | 0,300 | 0,037 | 0,149 | 0,027 | 0,049 | 0,015 |
| 150 | 0,503 | 0,038 | 0,298 | 0,030 | 0,149 | 0,019 | 0,050 | 0,011 |

(a) Freqüência alélica esperada em EHW na população de referência ($N = 2.000$).

TABELA 4 - Média e desvio padrão ($\hat{\sigma}$) da heterozigosidade observada ($\hat{H}_o$), esperada em EHW ($\hat{H}_e$) e do índice de fixação ($\hat{f}$) estimada de 100 amostras aleatórias de diferentes tamanhos ($n$), de uma população em Equilíbrio de Hardy-Weinberg, para três locos.

| $n$ | $\hat{H}_e$ | $\hat{H}_o$ | $\hat{f}$ |
|---|---|---|---|
| Loco 1: $H_e = H_o = 0{,}5$; $f = 0{,}0$ | | | |
| 10 | 0,560 (0,029) | 0,504 (0,166) | 0,151 (0,279) |
| 30 | 0,517 (0,011) | 0,498 (0,092) | 0,053 (0,180) |
| 60 | 0,508 (0,007) | 0,499 (0,069) | 0,026 (0,134) |
| 100 | 0,504 (0,004) | 0,501 (0,053) | 0,011 (0,105) |
| 150 | 0,505 (0,003) | 0,501 (0,048) | 0,013 (0,096) |
| Loco 2: $H_e = H_o = 0{,}545$; $f = 0{,}0$ | | | |
| 10 | 0,598 (0,061) | 0,529 (0,161) | 0,167 (0,247) |
| 30 | 0,566 (0,027) | 0,535 (0,092) | 0,071 (0,153) |
| 60 | 0,557 (0,017) | 0,545 (0,064) | 0,030 (0,107) |
| 100 | 0,550 (0,013) | 0,544 (0,050) | 0,016 (0,087) |
| 150 | 0,551 (0,010) | 0,543 (0,042) | 0,019 (0,071) |
| Loco 3: $H_e = H_o = 0{,}635$; $f = 0{,}0$ | | | |
| 10 | 0,696 (0,092) | 0,606 (0,174) | 0,179 (0,211) |
| 30 | 0,563 (0,056) | 0,544 (0,092) | 0,051 (0,132) |
| 60 | 0,646 (0,033) | 0,639 (0,073) | 0,019 (0,093) |
| 100 | 0,639 (0,026) | 0,634 (0,056) | 0,014 (0,071) |
| 150 | 0,639 (0,021) | 0,633 (0,050) | 0,014 (0,067) |

( ) Desvio padrão da média.

Avaliando-se as médias das estimativas nas TABELAS 3 e 4 observa-se que quanto maior o tamanho amostral, mais próximas as estimativas das freqüências alélicas e índices de diversidade encontram-se do seu verdadeiro valor e menor é o erro padrão ($\hat{\sigma}$ ). Para o tamanho amostral de 60 plantas, o intervalo de confiança do alelo de maior freqüência ($p_i$), no loco com dois alelos variou de 0,444 a 0,544, no loco com três alelos de 0,453 a 0,539 e no loco com 4 alelos de 0,444 a 0,550. Para o tamanho amostral de 100 plantas o intervalo do alelo de maior freqüência ($p_i$) no loco com dois alelos variou de 0,460 a 0,540, no loco com três alelos de 0,461 a 0,539 e no loco com 4 alelos de 0,458 a 0,542. Comparando estes intervalos com os resultados dos intervalos esperados (TABELA 1) para espécies em EHW ($s = 0$) verifica-se que estes se encontram dentro do intervalo de confiança a 95% de probabilidade (60: $IC$ = 0,409 a 0,591; 100: $IC$ = 0,429 a 0,571), confirmando as predições teóricas da eficiência dos tamanhos amostrais determinados.

Avaliando-se os índices de diversidade estimados nos três locos (TABELA 4), verifica-se que quanto maior o número de alelos maior é o erro na estimativa da heterozigosidade esperada ($\hat{H}_e$), sugerindo a necessidade de maiores tamanhos amostrais para casos da utilização de locos hipervariáveis, como por exemplo locos de microssatélites. A heterozigosidade observada ($\hat{H}_o$), por sua vez, não apresenta um padrão definido, em função do número de alelos nos locos. Por outro lado, o índice de fixação ( $\hat{f}$ ) apresentou menores erros com o aumento do número de alelos nos locos, para todos os tamanhos amostrais. Entre as heterozigosidades, a $\hat{H}_o$ apresenta os maiores erros, indicando que é mais difícil amostrar freqüências genotípicas do que freqüências alélicas. A explicação é que cada indivíduo amostrado contribui com dois alelos e uma classe genotípica para a amostra, portanto, contribui duas vezes mais para a estimativa das freqüências alélicas do que para as freqüências genotípicas.

Entre os índices de diversidade, o índice de fixação apresentou os maiores erros, para qualquer tamanho amostral. O índice de fixação depende da precisão simultânea nas freqüências alélicas e genotípicas. A amostragem de freqüências alélicas é refletida na heterozigosidade esperada e a de freqüências genotípicas na heterozigosidade observada (equações 6 e 7), indicando que para se obter medidas precisas do índice de fixação são necessários grandes tamanhos amostrais (≥ 60 indivíduos). Isto pode ser melhor visualizado pela distribuição das classes do índice $\hat{f}$ para diferentes tamanhos amostrais, nas 100 amostras realizadas (FIGURAS 1 e 2). Observa-se que, mesmo com grandes tamanhos amostrais ou 60 e 150 indivíduos por populações, ainda existe uma grande probabilidade de se obter estimativas de $\hat{f}$ diferentes das esperadas. Por exemplo, das 100 amostras de 150 plantas, para o loco com dois alelos, 14 apresentaram valores inferiores a –0,1 e 19 superiores a 0,1, enquanto o valor esperado era zero, portanto 33% das amostras estavam fora da faixa considerada aceitável ($-0{,}1 \geq \hat{f} \leq 0{,}1$). Mas, com maior número de alelos por loco a precisão aumenta, sendo que somente 20% das estimativas no loco 2 (três alelos) e 13% no loco 3 (quatro alelos), estavam fora da faixa admitida como aceitável, confirmando a tendência de redução nas estimativas de $\hat{f}$ com o aumento do número de alelos nos locos. Portanto, quando locos hipervariáveis são utilizados para a avaliação genômica, um menor número de indivíduos é necessário para a estimativa do índice de fixação.

Nas FIGURAS 1 e 2 também é possível se verificar que com a amostragem de 100 plantas, as estimativas de $\hat{f}$ encontraram-se dentro do mesmo intervalo máximo apresentado para a amostra de 150 plantas, sendo que 39, 26 e 16% das estimativas nos locos 1, 2 e 3, respectivamente, excederam ao intervalo de -0,1 a 0,1, revelando que não existem grandes vantagens em adotar um tamanho amostral superior a 100 plantas, a não ser em casos especiais como estudos de simulações. Com a amostragem de 60 plantas, as estimativas de $\hat{f}$ podem variar de -0,3 a 0,4, porém mais de 55% das estimativas estarão próximas ao valor real do parâmetro, ou seja, neste caso serão zero.

De modo geral, os resultados reforçam a sugestão de que tamanhos amostrais de 60 a 100 plantas são suficientes para se obter uma boa precisão nas freqüências alélicas e índices de diversidade.

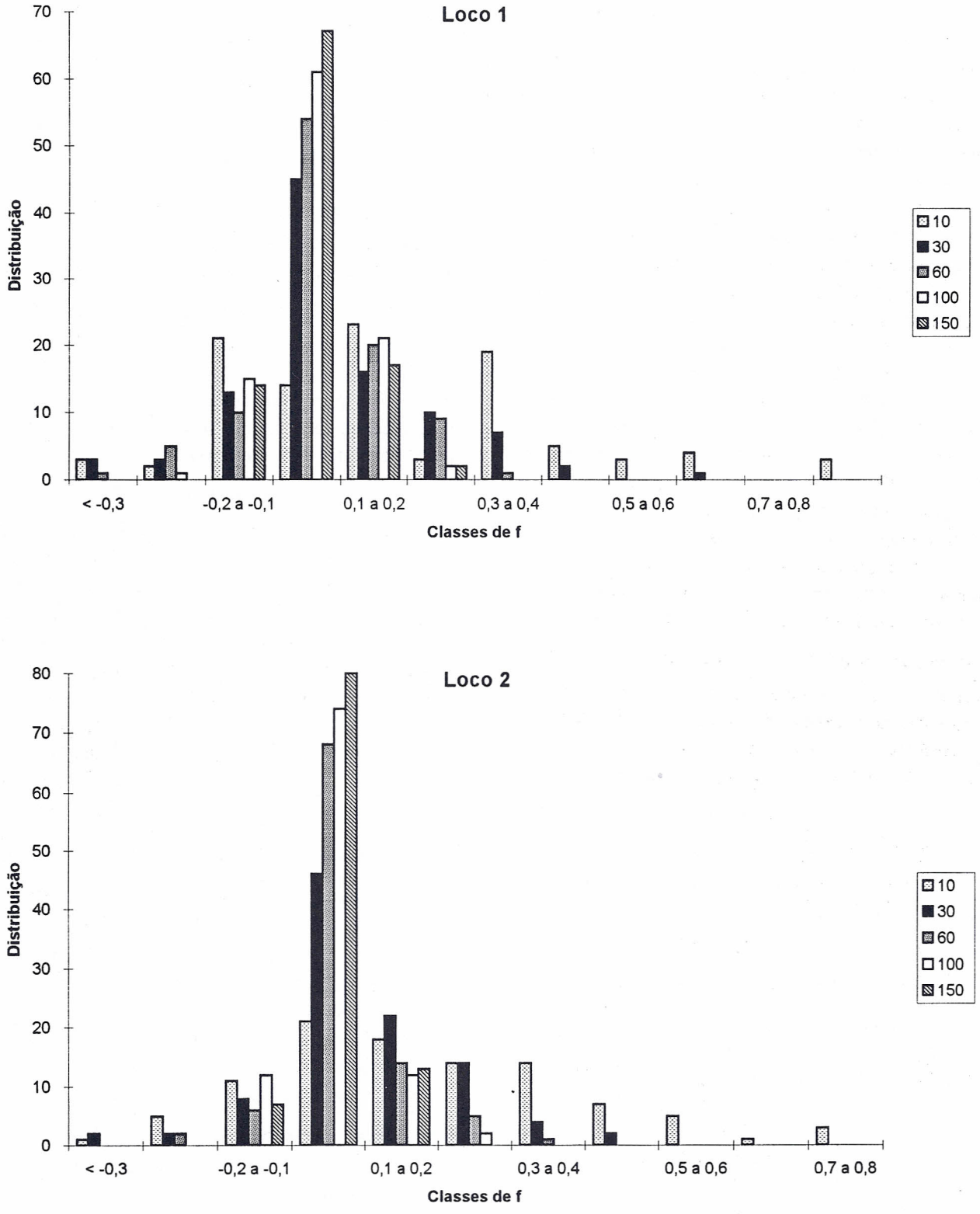


FIGURA 1 - Gráfico da distribuição do índice de fixação para 100 amostras de quatro diferentes tamanhos de uma população em Equilíbrio de Hardy-Weinberg, em um loco com dois (Loco 1) e três alelos (Loco 2).

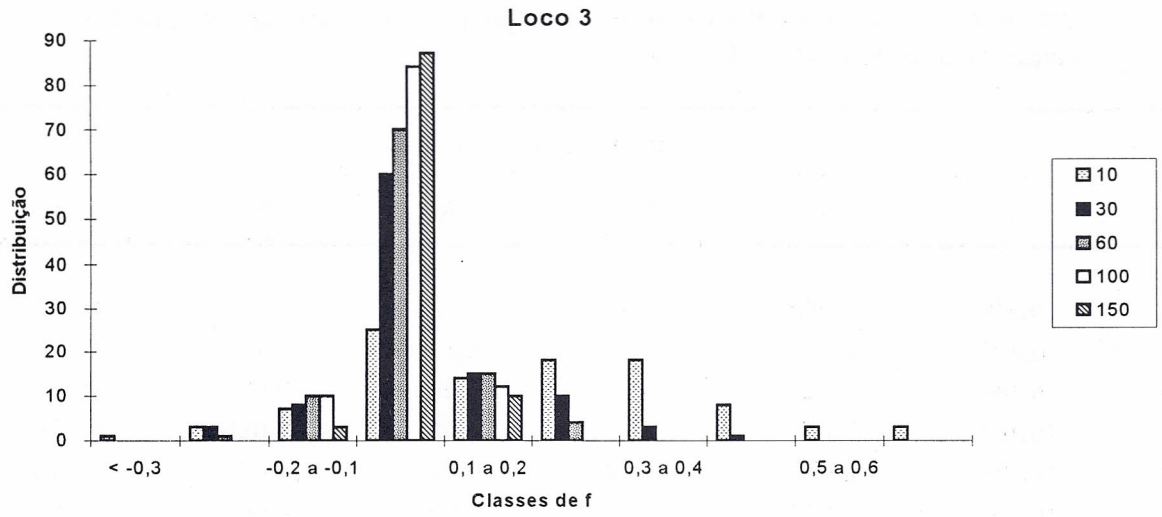


FIGURA 2 - Gráficos da distribuição do índice de fixação para 100 amostras de quatro diferentes tamanhos de uma população em Equilíbrio de Hardy-Weinberg, em um loco com quatro alelos.

### 3.3 Amostragem de Progênies

O tamanho amostral, em termos de número de árvores matrizes necessárias para a coleta de sementes ou número de progênies a amostrar ($m$) e do número de sementes a coletar por árvore ($k$) foi determinado pela amplitude do intervalo de confiança, associada à estimativa de um alelo de freqüência 0,5 (TABELA 5). Para estas estimativas foram também consideradas possíveis variações na taxa de autofecundação.

Pode-se concluir, primeiramente, que o erro nas estimativas das freqüências alélicas aumenta com o aumento da taxa de autofecundação. A segunda é que, independentemente da taxa de autofecundação, os erros amostrais das estimativas diminuem mais com o aumento do número de progênies ($m$) na amostra do que com o aumento no número de plantas por progênie ($k$), ou seja, para reduzir o erro de $p_i$ é mais eficiente aumentar o número de progênies do que o número de plantas por progênie na amostra. Outra vantagem em amostrar um maior número de progênies do que plantas/progênies, é que esta estratégia aumenta a chance de se detectar variações na taxa de cruzamento entre plantas individuais, cruzamentos entre aparentados, cruzamentos preferenciais e matrizes com alelos raros. A terceira conclusão é que em espécies autógamas ($s$ = 1,0) não existe vantagem em amostrar mais do que 10 plantas/progênie e em espécies alógamas, a partir de 20 plantas/progênie, muito pouco em termos de eficiência amostral é obtido na redução do erro de $p_i$. Em concordância, Cotterill & James (1984) determinaram que para se avaliar o efeito de progênies em testes de progênies de *Pinus radiata*, 10 a 20 plantas/progênies eram suficientes. O aumento do número de plantas por progênie só é interessante quando o número de progênies amostradas for baixo ($\leq$ 5), mas mesmo assim, não é necessário amostrar mais do que 30 plantas/progênie, se a espécie for de reprodução mista com predomínio de alogamia, ou mais de 20 plantas se a espécie for autógama. Quando o número de matrizes amostradas for igual ou superior a 30, não existem grandes vantagens em amostrar mais do que 10 plantas por progênie, independentemente do sistema de reprodução da espécie. Em tamanhos inferiores a 30 progênies, uma melhor eficiência amostral pode ser obtida coletando-se 20 sementes por progênie, se a espécie for de reprodução mista ou perfeitamente alógama ($t$ = 1,0). Para espécies autógamas continua valendo a recomendação de 10 plantas por progênie. Cotterill (1990) estudando o número de progênies necessárias para a seleção, determinou que 100 a 200 plantas estruturadas em 5 a 10 progênies seriam suficientes. Portanto, como o tamanho amostral aqui recomendado para estimar freqüências alélicas foi maior do que o determinado por Cotterill (1990), para programas de melhoramento, pode-se esperar boas perspectivas na sua utilização em programas de seleção.

TABELA 5 - Intervalo de confiança a 95% de probabilidade ($p_i \pm$ erro) para estimativa das freqüências alélicas ($p_i$ = 0,5) para diferentes números de progênies ($m$), tamanho de progênies ($k$) em função da taxas de autofecundação ($s$).

| | Plantas por progênie ($k$) | | | | | |
|---|---|---|---|---|---|---|
| $m$ | 1 | 10 | 20 | 30 | 50 | 100 |
| $s$ = 0 | | | | | | |
| 5 | 0,439 | 0,182 | 0,170 | 0,163 | 0,160 | 0,157 |
| 10 | 0,253 | 0,127 | 0,117 | 0,115 | 0,113 | 0,111 |
| 15 | 0,196 | 0,102 | 0,096 | 0,094 | 0,092 | 0,091 |
| 20 | 0,165 | 0,088 | 0,083 | 0,081 | 0,080 | 0,079 |
| 25 | 0,146 | 0,079 | 0,074 | 0,073 | 0,071 | 0,070 |
| 30 | 0,132 | 0,072 | 0,068 | 0,066 | 0,065 | 0,064 |
| 50 | 0,101 | 0,056 | 0,053 | 0,051 | 0,050 | 0,050 |
| 100 | 0,071 | 0,044 | 0,037 | 0,036 | 0,035 | 0,035 |
| $s$ = 0,1 | | | | | | |
| 5 | 0,450 | 0,200 | 0,188 | 0,181 | 0,179 | 0,177 |
| 10 | 0,259 | 0,140 | 0,131 | 0,128 | 0,126 | 0,125 |
| 15 | 0,201 | 0,112 | 0,107 | 0,105 | 0,103 | 0,102 |
| 20 | 0,170 | 0,097 | 0,092 | 0,091 | 0,089 | 0,088 |
| 25 | 0,150 | 0,087 | 0,076 | 0,081 | 0,080 | 0,079 |
| 30 | 0,135 | 0,079 | 0,075 | 0,074 | 0,073 | 0,072 |
| 50 | 0,104 | 0,061 | 0,058 | 0,057 | 0,057 | 0,056 |
| 100 | 0,073 | 0,048 | 0,041 | 0,041 | 0,040 | 0,040 |
| $s$ = 0,2 | | | | | | |
| 5 | 0,463 | 0,219 | 0,209 | 0,202 | 0,199 | 0,198 |
| 10 | 0,267 | 0,153 | 0,145 | 0,143 | 0,141 | 0,140 |
| 15 | 0,206 | 0,123 | 0,118 | 0,116 | 0,115 | 0,114 |
| 20 | 0,174 | 0,106 | 0,102 | 0,101 | 0,100 | 0,099 |
| 25 | 0,154 | 0,095 | 0,078 | 0,090 | 0,089 | 0,088 |
| 30 | 0,139 | 0,087 | 0,083 | 0,082 | 0,081 | 0,081 |
| 50 | 0,107 | 0,067 | 0,065 | 0,064 | 0,063 | 0,063 |
| 100 | 0,075 | 0,053 | 0,046 | 0,045 | 0,044 | 0,044 |
| $s$ = 1,0 | | | | | | |
| 5 | 0,621 | 0,452 | 0,447 | 0,438 | 0,438 | 0,438 |
| 10 | 0,358 | 0,316 | 0,310 | 0,310 | 0,310 | 0,310 |
| 15 | 0,277 | 0,253 | 0,253 | 0,253 | 0,253 | 0,253 |
| 20 | 0,234 | 0,219 | 0,219 | 0,219 | 0,219 | 0,219 |
| 25 | 0,206 | 0,196 | 0,196 | 0,196 | 0,196 | 0,196 |
| 30 | 0,187 | 0,179 | 0,179 | 0,079 | 0,079 | 0,079 |
| 50 | 0,143 | 0,139 | 0,139 | 0,139 | 0,139 | 0,139 |
| 100 | 0,100 | 0,098 | 0,098 | 0,098 | 0,098 | 0,098 |

Considerando que na maioria das situações o sistema de reprodução das espécies arbóreas não é conhecido *a priori*, mas o sistema misto e a alogamia predominam, recomenda-se o tamanho amostral de 30 progênies e 10 plantas/progênie, em uma população, para obtenção de estimativas confiáveis de freqüências alélicas. Desta forma, a verdadeira freqüência de $p_i$ estará, para uma espécie praticando 20% de autofecundação, entre 0,413 e 0,587. Para o caso de espécies autógamas, sugere-se a amostragem de 50 progênies e 10 plantas/progênie, como um tamanho suficiente para obtenção de precisão nas referidas estimativas. Neste caso, a verdadeira freqüência de $p_i$ estará entre o intervalo de 0,361 e 0,639. Estas recomendações agregam-se, em parte, às recomendações de Brown (1975). O autor, avaliando o tamanho amostral em quatro delineamentos de testes de progênies para a estimativa de parâmetros do sistema de reprodução, sugeriu para espécies alógamas a amostragem de mais de 25 progênies por população, com a avaliação de pelo menos 10 plantas por progênie.

Não existe justificativa para utilizar estruturas de progênies no estudo da distribuição da variabilidade genética entre e dentro de populações, por marcadores genéticos, devido ao grande tamanho amostral requerido para representar as populações individualmente. Por exemplo, considerando o tamanho amostral de 30 progênies e 10 plantas/progênie para representar uma população, se 10 populações forem amostradas, será necessário avaliar por marcadores 3.000 plantas (30 progênies x 10 plantas/progênies x 10 população), ou seja, um número excessivamente grande de plantas. Em situações onde a única alternativa é a avaliação de sementes, devido à distância da área de coleta, problemas de armazenamento das folhas, etc., é mais interessante trabalhar com amostragem de misturas de sementes. Para tanto, procede-se a coleta de sementes de um número grande de árvores ($\geq$ 30/população), mistura-se quantidades iguais de sementes de cada árvore matriz e retira-se aleatoriamente 60 sementes para representar a população. Repete-se o processo nas demais populações. O princípio de representatividade genética de cada população é o mesmo aplicado para o caso da amostragem individual (árvores adultas, jovens, plântulas e banco de sementes). Adotando-se esta estratégia amostral, para o exemplo anterior, o tamanho amostral total será de 600 plantas (60 plantas x 10 população) ou 80% menor, para uma mesma eficiência amostral nas estimativas das freqüências alélicas. A desvantagem da amostragem em misturas de sementes é que o sistema de reprodução não poderá ser avaliado de forma detalhada, pelo modelo de reprodução mista de Ritland & Jain (1981), o qual requer estruturas de progênies. A alternativa será estimar a taxa de cruzamento pelo método dos momentos ($t = (1-f)/(1+f)$; Weir, 1996). Em casos onde se deseja avaliar detalhadamente o sistema de reprodução de várias populações, pode-se optar pelas seguintes estratégias amostrais: primeiro, fixar o tamanho amostral por progênie, em 10 plantas; segundo, até um número de cinco populações, manter a amostragem de 30 progênies, resultando em 300 plantas por população e 1.500 plantas totais amostradas. Para amostragem de cinco a 10 populações, pode-se adotar o tamanho de 20 progênies/população, resultando no total de 200 plantas/população e 2.000 plantas no caso da avaliação de 10 populações. Para amostragem acima de 10 populações, o tamanho de 15 progênies/população pode ser suficiente, resultando em 150 plantas/população, e 3.000 plantas para o caso da análise de 20 populações. Finalmente, para a amostragem de mais de 20 populações, por exemplo, 30 populações, pode-se adotar o número de 10 progênies/população, totalizando a amostragem 100 plantas por população e 3.000 plantas para o conjunto das populações.

## 4 CONCLUSÕES E RECOMENDAÇÕES

1. Para se obter estimativas de freqüências alélicas e índices de diversidade genética precisas, em uma simples população, é necessário amostrar entre 60 a 100 indivíduos.
2. Com base no intervalo de confiança do erro das estimativas das freqüências alélicas, em uma população estruturada em progênies, se a espécie alvo de estudo for alógama ou de reprodução mista, com predomínio de alogamia, a amostragem de 30 progênies e 10 plantas por progênies é suficiente para a obtenção de estimativas confiáveis. Caso a espécie reproduza-se por autogamia, este tamanho amostral deve ser de 50 progênies e 10 plantas por progênies.

3. Para espécies alógamas, na amostragem de 5 a 10 populações, pode-se adotar o tamanho de 20 progênies/população, com 10 plantas/progênie.
4. Para espécies alógamas, na amostragem de 10 a 20 populações, pode-se adotar o tamanho de 15 progênies/população, com 10 plantas/progênie.
5. Para espécies alógamas, na amostragem de mais de 20 populações, pode-se adotar o tamanho de 10 progênies/população, com 10 plantas/progênie.

## REFERÊNCIAS BIBLIOGRÁFICAS

AYALA, F. J.; KIGER, J. A. **Modern genetics**. Menlo Park: The Benjamin/Cummings Co., 1984. 798 p.

BERG, E. E.; HAMRICK, J. L. Quantification of diversity at allozyme loci. **Canadian Journal Forest Research**, Edmonton, v. 27, p. 415-424, 1997.

BROWN, A. H. D. Efficient experimental designs for the estimation of genetic parameters in plant populations. **Biometrics**, Alexandria, v. 31, p. 145-160, 1975.

_____.; WEIR, S. B. Measuring genetic variability in plant populations. In: TANKSLEY, S. D.; ORTON, T. J. (Ed.). **Isozymes in plant genetics and breeding**. Amsterdam: Elsevier Science Publishers, 1983. part A, p. 219-239.

COTTERILL, P. P.; JAMES, J. W. Number of offspring and plot sizes required for progeny testing. **Silvae Genetica**, Frankfurt, v. 33, n. 6, p. 203-209, 1984.

COTTERILL, P. P. Short note: number of families and progeny required for provenance testing. **Silvae Genetica**, Frankfurt, v. 39, n. 2, p. 82-83, 1990.

EL-KASSABY, Y. A.; SZIKLAI, O. Effect of sample size on the precision of the estimate of allozyme frequencies in a natural stand of Douglas-Fir. **Egyptian Journal of Genetics and Cytology**, Ottawa, v. 24, p. 345-360, 1983.

HAMRICK, J. L.; LINHART, Y. B.; MITTON, J. B. Relationships between life history characteristic and eletrophoretically detectable genetic variation in plants. **Annual Review Ecology and Systematics**, Davis, v. 10, p. 173-200, 1979.

HAMRICK, J. L.; GODT, M. J. W. Allozyme diversity in plant species. In: BROWN, A. H. D. *et al.* (Ed.). **Plant population genetics, breeding and genetic resources**. Sunderland: Sinauer Press, 1989. p. 43-63.

_____.; _____.; SHERMAN-BROYLES, S. L. Factors influencing levels of genetic diversity in woody plant species. **New Forest**, Dordrecht, v. 5, p. 95-124, 1992.

NEI, M. Analysis of gene diversity in subdivided populations. **Proc. Nat. Acad. Sci.**, Washington, v. 70, n. 12, p. 3321-3323, 1973.

_____. *F*-statistics and analysis of gene diversity in subdivided populations. **Annals of Human Genetics**, London, v. 41, p. 225-233, 1977.

RITLAND, K.; JAIN, S. A model for the estimation of outcrossing rate and gene frequencies using independent loci. **Heredity**, Lund, v. 47, p. 35-52, 1981.

WEIR, B. S. **Genetic data analysis II. Methods for discrete population genetic data**. Sunderland: North Caroline State University, Sinauer Associates Inc. Pub., 1996. 445 p.