MÉTODO BOOTSTRAP APLICADOS EM NÍVEIS DE REAMOSTRAGEM NA ESTIMAÇÃO DE PARÂMETROS GENÉTICOS POPULACIONAIS

Luciana Aparecida Carlini-Garcia[1]*; Roland Vencovsky[2]; Alexandre Siqueira Guedes Coelho[3]

[1]*Pós-Graduanda em Genética e Melhoramento de Plantas - USP/ESALQ.*
[2]*Depto. de Genética - USP/ESALQ, C.P. 83 - CEP: 13418-970 - Piracicaba, SP.*
[3]*Depto. de Biologia Geral - ICB/UFG, C.P. 131 - CEP: 74001-970 - Goiânia, GO.*
**Autor correspondente <lacarlin@carpa.ciagri.usp.br>*

RESUMO: Marcadores isoenzimáticos ou moleculares têm sido empregados em estudos da estrutura genética e do sistema reprodutivo de populações naturais. Parâmetros genéticos populacionais de interesse são estimados, porém, há pouca informação sobre o erro associado a essas estimativas em função dos diferentes níveis de amostragem. Neste trabalho, o método de reamostragem bootstrap foi aplicado sobre locos, indivíduos, populações e indivíduos e populações concomitantemente, em dados de populações naturais. Para os parâmetros índice de fixação total (F), índice de fixação intrapopulacional (f), diversidade entre populações (θ) e taxa aparente de cruzamento (t_a) foram obtidos, em função das fontes de reamostragem, os erros associados às estimativas desses parâmetros, a distribuição de empírica dessas estimativas e intervalos de confiança para tais parâmetros. Geralmente, os menores erros estão associados a $\hat{\theta}$, e verificou-se que apenas as distribuições empíricas de $\hat{F}$ e $\hat{f}$ tendem à normalidade quando indivíduos estão envolvidos na reamostragem. Foram feitas reamostragens com tamanho variável de amostras bootstrap, visando obter o número necessário de locos, indivíduos e populações para atingir um dado nível de precisão na estimação de F, f e θ. Em geral, os tamanhos amostrais utilizados nas pesquisas com populações naturais foram suficientes apenas para estimar θ, com a precisão estabelecida. A fonte de variação de locos foi responsável pelos maiores erros associados a $\hat{F}$ e $\hat{f}$, sendo recomendável aumentar o número de locos em pesquisas dessa natureza. A amostragem de populações também deverá merecer atenção no planejamento de pesquisas futuras.
Palavras-chave: estrutura populacional, genética de populações, estimação

BOOTSTRAP METHOD APPLIED OVER RESAMPLING LEVELS IN THE ESTIMATION OF GENETIC PARAMETERS OF POPULATIONS

ABSTRACT: Genetic structure and reproductive system of natural populations have often been studied using molecular markers and isozymes. Specific genetic parameters are therefore being estimated for this purpose. There is little information about errors associated with these estimates due to different possible sampling levels. Some real datasets were used in this study and bootstrapping was performed over loci, individuals, populations and both individuals and populations simultaneously. These different resampling strategies produced associated errors of the estimates and their empiric distributions and confidence intervals for the following parameters: total fixation index (F), fixation index within populations (f), diversity between populations (θ) and apparent outcrossing rate (t_a). Generally, smaller errors were associated with $\hat{\theta}$. The only empirical distribution of estimates approaching normality were for $\hat{F}$ and $\hat{f}$ when individuals were envolved in the resampling process. Bootstrap samples of variable sizes were used to obtain the number of loci, individuals and populations necessary to achieve a given precision of $\hat{F}$, $\hat{f}$ and $\hat{\theta}$. Generally, the magnitude of errors of the estimates in most datasets indicated that sample sizes currently used in research involving natural populations were suitable only for the estimation of θ. The largest errors of $\hat{F}$ and $\hat{f}$, were associated to sampling errors due to loci. Therefore, future studies should use larger numbers of loci. Sampling errors due to population was also important and should not be neglected.
Keywords: population structure, population genetics, estimation

INTRODUÇÃO

A Genética de Populações é fundamental para o desenvolvimento e entendimento dos processos evolutivos e do melhoramento genético. Com os marcadores bioquímicos e moleculares, houve um salto qualitativo e quantitativo em estudos da estrutura populacional e do sistema reprodutivo de diversas espécies. Nestes estudos, podem ser estimados vários parâmetros populacionais como grau de endogamia, sistema reprodutivo predominante, entre outros, que são muito importantes na determinação de estratégias de conservação da variabilidade genética na natureza.

Porém, no processo de estimação, nem sempre os erros das estimativas são fornecidos e há pouca informação sobre a natureza da distribuição empírica das

mesmas. Além disso, obter expressões explícitas das estimativas das variâncias de $\hat{F}$, $\hat{f}$ e $\hat{\theta}$ é tarefa bastante difícil, pois esses estimadores são razões entre variáveis aleatórias com distribuição desconhecida. Expressões explícitas das variâncias de parâmetros como esses, geralmente decorrem de aproximações advindas da expansão da série de Taylor. Muniz (1994) utilizou tal aproximação e apresentou estimadores de F, f, θ e t_a e de suas respectivas variâncias. Além disso, ao trabalhar com esperanças de quadrados médios, obtidas da análise de variância, supôs normalidade no processo de obtenção das variâncias dos estimadores, o que foi outra aproximação. Os estimadores propostos por esse autor apresentaram desempenho satisfatório somente quando as freqüências gênicas foram intermediárias e os tamanhos amostrais elevados. Weir & Cockerham (1984) também apresentaram estimadores para F, f, θ e suas variâncias, porém, considerando apenas um loco.

Uma alternativa é o uso de técnicas de reamostragem para mensurar a acurácia das estimativas dos parâmetros de interesse e os desvios-padrão a elas associados. Dentre os vários procedimentos de reamostragem, o método bootstrap vem se destacando, pois, além de fornecer estimativas de parâmetros e seus desvios-padrão, permite obter intervalos de confiança para os parâmetros analisados, bem como a distribuição empírica de suas estimativas (Efron & Tibishirani, 1993; Manly, 1997).

Nesse contexto, Weir (1996) ressalta o uso desse procedimento para testar se as freqüências alélicas entre populações de uma espécie diferem entre si. Para tanto, populações devem ser reamostradas. Vencovsky et al. (1997) reamostraram indivíduos, visando obter as distribuições empíricas das estimativas do índice de fixação intrapopulacional ($\hat{f}$) e da taxa de cruzamento ($\hat{t}$), a variância desses estimadores e os intervalos de confiança para *f* e *t*.

Uma questão importante em Genética de Populações ainda não resolvida na literatura é o estabelecimento de qual nível hierárquico deve ser reamostrado. Van Dongen (1995) estudou o uso de bootstrap para estimar a distribuição de estatísticas calculadas a partir de dados alozimáticos. Concluiu que a reamostragem de genótipos individuais deve ser preferida em relação à reamostragem de locos. Petit & Pons (1998) avaliaram qual a unidade a ser reamostrada para estudos de diversidade. Três formas de reamostragem foram consideradas: populações, indivíduos dentro de populações e populações e indivíduos conjuntamente. Verificaram que os estimadores de variância mais apropriados foram obtidos com bootstrap sobre populações somente.

Em vista do exposto, o objetivo geral deste trabalho foi fornecer subsíduos para uma aplicação mais eficiente do método de reamostragem bootstrap em estudos de populações naturais via marcadores genéticos. Especificamente, os objetivos foram: a) demonstrar a necessidade de realizar reamostragens para obtenção dos erros de estimação associados às estimativas dos parâmetros F (índice de fixação total), f (índice de fixação intrapopulacional), θ (divergência entre populações) e t_a (taxa de cruzamento aparente), em função de diversas fontes de variação estudadas, a saber, locos, indivíduos dentro de populações, populações e indivíduos e populaçoes simultaneamente; b) comparar a grandeza dos erros provenientes dos vários níveis de reamostragem associados às estimativas desses parâmetros; c) testar se as distribuições empíricas de $\hat{F}$, $\hat{f}$, $\hat{\theta}$ e $\hat{t}_a$ tendem à normalidade; d) comparar intervalos de confiança provenientes das diferentes fontes de variação na estimação desses parâmetros; e) verificar a quais parâmetros estão associados os maiores erros de estimação; f) estabelecer estratégias de amostragem para obter uma dada precisão na estimação de F, f e θ.

MATERIAL E MÉTODOS

Os dados utilizados nesta pesquisa foram produzidos por Reis (1996), Seoane (1998), Telles & Coelho (1998), Ciampi (1999), Auler (2000) e Sebbenn[1] et al. (s.d.). Em todos estes trabalhos, os autores estudaram a estrutura genética populacional e/ou o sistema reprodutivo das espécies, por meio de marcadores isoenzimáticos ou, no caso de Ciampi (1999), por marcadores microssatélites. Reis (1996) estudou oito populações de palmiteiro (*Euterpe edulis*), considerando sete locos isoenzimáticos. Coletou, em média, 25 indivíduos por população. Seoane (1998) coletou cerca de 22 indivíduos em cada uma das quatro populações de *Esenbeckia leiocarpa* analisadas. Utilizou onze locos de isoenzimas, cinco dos quais foram polimórficos. Telles & Coelho (1998) estudaram a magnitude e a distribuição da variabilidade genética de seis populações de Araticum (*Anonna crassiflora*) provenientes de duas regiões do Estado de Goiás. Foram considerados quatro marcadores isoenzimáticos e coletados 30 indivíduos por população. Ciampi (1999) coletou 24 indivíduos em cada uma das quatro populações de copaíba (*Copaifera langsdorffii*) amostradas no Cerrado. Analisou oito locos de microssatélites. Auler (2000) caracterizou a estrutura genética de nove populações de *Araucaria angustifolia* em Santa Catarina. Coletou cerca de 36 indivíduos por população. Foram analisados 15 locos isoenzimáticos, dos quais doze eram polimórficos. Sebbenn[2] et al.(s.d.) estudaram a estrutura genética, o sistema reprodutivo, a distribuição genética espacial, o fluxo de genes e o

[1]SEBBENN, A.M.; SEOANE, C.E.S.; KAGEYAMA, P.Y.; LACERDA, C.B. Estrutura genética de populações de *Tabebuia cassinoides*: implicações para o manejo e a conservação genética. (enviado para publicação).

tamanho efetivo populacional de duas populações de *Tabebuia cassinoides*, uma não manejada, situada na localidade de Juréia e outra manejada localizada em Iguape. Utilizaram doze locos de isoenzimas e coletaram, em média, 46 indivíduos por população.

A análise da estrutura populacional foi feita de acordo com a análise de variância de freqüências gênicas (Cockerham, 1969; Cockerham, 1973; Weir & Cockerham, 1984; Weir, 1996). Esta análise foi realizada para cada conjunto de dados, visando a obtenção de $\hat{F}$, $\hat{f}$ e $\hat{\theta}$. Em todos os conjuntos de dados, a estrutura hierárquica considerada incluiu as fontes de variação de populações (P), de indivíduos dentro de populações (I) e de genes dentro de indivíduos (G).

Em princípio, o método de reamostragem bootstrap foi aplicado com a finalidade de obter estimativas bootstrap de F, f e θ e de seus respectivos desvios-padrão, em função das fontes de variação de locos, indivíduos, populações e indivíduos e populações simultaneamente. As reamostragens foram feitas para verificar quais as principais fontes de erro das estimativas desses parâmetros, obter intervalos de confiança percentil para os mesmos e as distribuições empíricas de suas estimativas.

Primeiramente, as amostras bootstrap foram tomadas com o mesmo tamanho da amostra original. Para cada nível de reamostragem, foram obtidas 100.000 amostras bootstrap e, para cada uma delas, calcularam-se, pelo método da análise de variância de freqüências gênicas, estimativas bootstrap de F, f e θ. bem como de seus desvios-padrão. A média dessas estimativas, por parâmetro, forneceu a estimativa bootstrap do parâmetro e a variância dessas estimativas resultou na estimativa da variância da estimativa bootstrap do parâmetro. Tais análises foram realizadas com uso do programa computacional EG (Coelho, 2000a).

Apenas os locos polimórficos foram usados na reamostragem de locos. Além disso, como a reamostragem é feita com reposição, algumas configurações de amostras bootstrap podem conduzir a estimativas que são quocientes, cujo denominador é zero. Nesses casos, a amostra bootstrap foi descartada automaticamente pelo programa computacional usado. Isso se aplica a qualquer nível reamostrado.

A partir de $\hat{f}$ em cada amostra bootstrap, obtiveram-se estimativas de t_a acordo com a expressão

$$\hat{t}_a = (1-\hat{f})/(1+\hat{f})$$

Em cada conjunto de dados, foram obtidos intervalos de confiança percentil para os quatro parâmetros estudados. Foram feitos testes de normalidade de Kolmogorov-Smirnov (Sokal & Rohlf, 1995) para cada uma das distribuições das estimativas obtidas.

Como estimativas das variâncias de $\hat{F}$, $\hat{f}$ e $\hat{\theta}$ possuem vários componentes e dada a dificuldade de obter expressões explícitas dessas variâncias, adotou-se metodologia semelhante àquela utilizada por Tivang et al. (1994) e por Hállden et al. (1994), visando determinar quais os números mínimos necessários de locos, indivíduos e populações para obter uma dada magnitude de desvio-padrão dessas estimativas, para cada conjunto de dados avaliados. O número mínimo de populações e indivíduos considerados na reamostragem variável foi dois, uma vez que deve-se ter pelo menos dois indivíduos e duas populações para poder estimar F, f e θ.

Diferindo dos trabalhos de Tivang et al. (1994) e de Hállden et al. (1994), ao invés de calcular o coeficiente de variação associado às estimativas dos parâmetros para cada tamanho de amostra, foram obtidos os desvios-padrão. O número de amostras bootstrap para cada tamanho amostral foi 1.000.

Para cada estimativa de parâmetro, foi admitido erro máximo de 0,02 a ela associado. Este valor foi determinado em função das magnitudes de $\hat{F}$, $\hat{f}$ e $\hat{\theta}$ freqüentemente encontradas na literatura. O número de locos, indivíduos e populações pôde ser estimado a partir da equação de regressão do tipo potência obtida em cada caso, substituindo a variável resposta pelo valor 0,02. Dessa forma, com base na metodologia proposta, foi possível determinar estratégias de amostragem adequadas para cada um dos conjuntos de dados avaliados. As reamostragens bootstrap com tamanhos de amostra variáveis e cálculos das estimativas dos parâmetros e de seus desvios-padrão foram realizadas pelo programa computacional EGBV (Coelho, 2000b).

RESULTADOS E DISCUSSÃO

As estimativas $\hat{F}$ obtidas com a reamostraram de indivíduos foram as que mais se aproximaram da estimativa original, obtida pelo método da análise de variância das freqüências gênicas, porém sem a realização de bootstrap (TABELA 1). Para f, isso ocorreu quando foram feitas reamostragem de locos com os dados de Sebbenn[3] et al. (s.d.) e de populações para os demais conjuntos de dados. No caso de θ, estimativas mais parecidas com a estimativa original sempre ocorreram quando foram reamostrados locos.

O fato de ser observada tendência a um padrão de comportamento das estimativas oriundas de diferentes tipos de reamostragem, em relação às estimativas originais dos parâmetros, pode sugerir a presença de algum viés nas estimativas bootstrap. No entanto, com o aumento do tamanho amostral, espera-se que o viés tenda a diminuir. Posto isso, sugere-se que pesquisas futuras sejam conduzidas para verificar se esses padrões de comportamento observados são de

[2,3]SEBBENN, op. citi. p.2.

TABELA 1 - Estimativas dos parâmetros [a] F, [b] f, [c] θ, e [d] t_a, de seus desvios-padrão ($\hat{\sigma}$) e intervalos de confiança oriundos da reamostragem de locos (L), indivíduos (I), populações (P) e indivíduos e populações concomitantemente (I e P). Dados de diversos autores.

Reis (1996)							
		$\hat{F}$			$\hat{f}$		
Nível	nbd+	Estimativa	$\hat{\sigma}$	I.C.$_{95\%}$	Estimativa	$\hat{\sigma}$	I.C.$_{95\%}$
O*		-0,028			-0,046		
L	0	-0,031	0,041	[-0,110; 0,050]	-0,048	0,039	[-0,120, 0,029]
I	0	-0,028	0,023	[-0,073; 0,018]	-0,067	0,024	[-0,115; -0,019]
P	0	-0,033	0,040	[-0,110; 0,047]	-0,046	0,037	[-0,116; 0,029]
I e P	0	-0,032	0,047	[-0,122; 0,060]	-0,066	0,044	[-0,152; 0,023]
		$\hat{\theta}$			$\hat{t}_a$		
Nível	nbd+	Estimativa	$\hat{\sigma}$	I.C.$_{95\%}$	Estimativa	$\hat{\sigma}$	I.C.$_{95\%}$
O*		0,017			1,097		
L	0	0,017	0,007	[0,005; 0,031]	1,105	0,085	[0,944; 1,273]
I	0	0,036	0,007	[0,023; 0,051]	1,145	0,056	[1,039; 1,259]
P	0	0,013	0,006	[-0,001; 0,022]	1,099	0,081	[0,944; 1,261]
I e P	0	0,032	0,009	[0,015; 0,051]	1,147	0,102	[0,956; 1,359]
Seoane (1998)							
		$\hat{F}$			$\hat{f}$		
Nível	nbd+	Estimativa	$\hat{\sigma}$	I.C.$_{95\%}$	Estimativa	$\hat{\sigma}$	I.C.$_{95\%}$
O*		0,085			-0,018		
L	0	0,090	0,082	[-0,067; 0,236]	-0,011	0,095	[-0,149; 0,205]
I	11	0,084	0,046	[-0,006; 0,174]	-0,045	0,054	[-0,152; 0,059]
P	0	0,055	0,060	[-0,082; 0,162]	-0,017	0,055	[-0,128; 0,087]
I e P	14	0,055	0,076	[-0,101; 0,197]	-0,043	0,077	[-0,194; 0,107]
		$\hat{\theta}$			$\hat{t}_a$		
Nível	nbd+	Estimativa	$\hat{\sigma}$	I.C.$_{95\%}$	Estimativa	$\hat{\sigma}$	I.C.$_{95\%}$
O*		0,101			1,036		
L	0	0,099	0,047	[0,017; 0,194]	1,041	0,185	[0,660; 1,350]
I	11	0,123	0,024	[0,078; 0,174]	1,100	0,121	[0,889; 1,359]
P	0	0,071	0,031	[-0,012; 0,102]	1,041	0,116	[0,840; 1,293]
I e P	14	0,094	0,038	[0,006; 0,161]	1,104	0,173	[0,806; 1,482]
Telles & Coelho (1998)							
		$\hat{F}$			$\hat{f}$		
Nível	nbd+	Estimativa	$\hat{\sigma}$	I.C.$_{95\%}$	Estimativa	$\hat{\sigma}$	I.C.$_{95\%}$
O*		0,231			0,017		
L	0	0,229	0,078	[0,083; 0,372]	0,020	0,058	[-0,081; 0,141]
I	0	0,231	0,035	[0,162; 0,300]	-0,001	0,043	[-0,084; 0,083]
P	0	0,198	0,059	[0,079; 0,310]	0,017	0,041	[-0,062; 0,094]
I e P	0	0,199	0,070	[0,058; 0,330]	0,001	0,059	[-0,116; 0,117]
		$\hat{\theta}$			$\hat{t}_a$		
Nível	nbd+	Estimativa	$\hat{\sigma}$	I.C.$_{95\%}$	Estimativa	$\hat{\sigma}$	I.C.$_{95\%}$
O*		0,218			0,967		
L	0	0,215	0,039	[0,148; 0,294]	0,967	0,111	[0,755; 1,176]
I	0	0,231	0,018	[0,197; 0,268]	1,004	0,087	[0,847; 1,185]
P	0	0,185	0,044	[0,096; 0,268]	0,969	0,079	[0,828; 1,133]
I e P	0	0,198	0,048	[0,101; 0,284]	1,006	0,120	[0,790; 1,261]

Ciampi (1999)

		$\hat{F}$			$\hat{f}$		
Nível	nbd+	Estimativa	$\hat{\sigma}$	I.C.$_{95\%}$	Estimativa	$\hat{\sigma}$	I.C.$_{95\%}$
O*		0,035			-0,016		
L	0	0,035	0,030	[-0,026; 0,091]	-0,016	0,032	[-0,081; 0,044]
I	0	0,035	0,012	[0,012; 0,060]	-0,037	0,014	[-0,064; -0,011]
P	0	0,018	0,021	[-0,032; 0,044]	-0,016	0,019	[-0,058; 0,020]
I e P	0	0,018	0,024	[-0,039; 0,057]	-0,037	0,023	[-0,037; 0,008]
		$\hat{\theta}$			$\hat{t}_a$		
Nível	nbd+	Estimativa	$\hat{\sigma}$	I.C.$_{95\%}$	Estimativa	$\hat{\sigma}$	I.C.$_{95\%}$
O*		0,050			1,032		
L	0	0,050	0,004	[0,043; 0,058]	1,035	0,067	[0,915; 1,177]
I	0	0,070	0,005	[0,060; 0,081]	1,077	0,029	[1,022; 1,136]
P	0	0,033	0,013	[0,009; 0,050]	1,033	0,039	[0,960; 1,122]
I e P	0	0,053	0,014	[0,026; 0,074]	1,078	0,050	[0,984; 1,179]

Auler (2000)

		$\hat{F}$			$\hat{f}$		
Nível	nbd+	Estimativa	$\hat{\sigma}$	I.C.$_{95\%}$	Estimativa	$\hat{\sigma}$	I.C.$_{95\%}$
O*		0,175			0,148		
L	0	0,182	0,059	[0,091; 0,327]	0,155	0,060	[0,061; 0,303]
I	0	0,174	0,029	[0,117; 0,231]	0,134	0,029	[0,077; 0,189]
P	0	0,173	0,051	[0,073; 0,272]	0,151	0,052	[0,052; 0,251]
I e P	0	0,173	0,059	[0,060; 0,289]	0,137	0,059	[0,024; 0,253]
		$\hat{\theta}$			$\hat{t}_a$		
Nível	nbd+	Estimativa	$\hat{\sigma}$	I.C.$_{95\%}$	Estimativa	$\hat{\sigma}$	I.C.$_{95\%}$
O*		0,031			0,742		
L	0	0,032	0,003	[0,026; 0,039]	0,736	0,085	[0,535; 0,885]
I	0	0,047	0,009	[0,031; 0,066]	0,765	0,045	[0,682; 0,857]
P	0	0,026	0,008	[0,012; 0,042]	0,741	0,079	[0,599; 0,902]
I e P	0	0,042	0,012	[0,022; 0,068]	0,764	0,092	[0,596; 0,953]

Sebbenn (s.d.)

		$\hat{F}$			$\hat{f}$		
Nível	nbd+	Estimativa	$\hat{\sigma}$	I.C.$_{95\%}$	Estimativa	$\hat{\sigma}$	I.C.$_{95\%}$
O*		0,107			0,037		
L	0	0,108	0,057	[-0,008; 0,215]	0,037	0,048	[-0,060; 0,126]
I	0	0,107	0,030	[0,046; 0,165]	0,025	0,026	[-0,028; 0,076]
P	0	0,068	0,040	[0,016; 0,107]	0,038	0,008	[0,028; 0,051]
I e P	0	0,067	0,050	[-0,019; 0,156]	0,026	0,028	[-0,028; 0,083]
		$\hat{\theta}$			$\hat{t}_a$		
Nível	nbd+	Estimativa	$\hat{\sigma}$	I.C.$_{95\%}$	Estimativa	$\hat{\sigma}$	I.C.$_{95\%}$
O*		0,073			0,929		
L	0	0,074	0,025	[0,028; 0,125]	0,933	0,090	[0,777; 1,127]
I	0	0,084	0,018	[0,052; 0,124]	0,953	0,050	[0,859; 1,057]
P	0	0,030	0,043	[-0,013; 0,073]	0,926	0,015	[0,902; 0,945]
I e P	0	0,042	0,044	[-0,008; 0,116]	0,950	0,053	[0,847; 1,057]

+n.d.b. = número de bootstraps descartados; *estimativa original dos parâmetros; [a]índice de fixação total; [b]índice de fixação dentro das populações; [c]grau de divergência entre populações; [d]taxa de cruzamento aparente.

fato caracterizados como viéses e, em caso afirmativo, estudar a importância desses viéses na estimação de parâmetros populacionais e a partir de que tamanhos amostrais eles são desprezíveis.

Desconsiderando o erro conjunto devido a indivíduos e populações, para $\hat{F}$ e $\hat{f}$, os maiores desvios-padrão foram obtidos quando foram reamostrados locos. A segunda maior fonte de erro geralmente esteve associada a populações. No caso de $\hat{\theta}$, os desvios-padrão máximos obtidos em função dos níveis de reamostragem variaram entre os diferentes conjuntos de dados, não havendo um padrão aparente para tal variação. Dessas observações, sugere-se que um maior número de populações e principalmente de locos seja considerado em pesquisas futuras com populações naturais. Deve-se lembrar que, nos dados do presente estudo, o número de locos variou de quatro a quinze e o de populações de duas a nove.

No entanto, muitas vezes não é possível aumentar o número de populações amostradas. O mesmo pode ocorrer com o número de locos isoenzimáticos, o que talvez possa ser contornado pelo o emprego de marcadores microssatélites, por exemplo. Porém, uma outra observação a ser feita é que, ao aumentar o número de indivíduos, mesmo sem alterar o número de populações e locos, deve haver redução nos erros associados a essas duas últimas fontes de variação.

Considerando apenas os parâmetros F, f e θ, observou-se que, em geral, os maiores erros estiveram associados às estimativas de F e f. A magnitude dos erros de $\hat{\theta}$ foram relativamente menores, mesmo quando as estimativas pontuais de θ foram maiores, como ocorre nos conjuntos de Seoane (1998) e Ciampi (1999), por exemplo. Portanto, considerando uma mesma magnitude de erro para tais estimativas, tamanhos amostrais maiores seriam necessários para estimar F e f em comparação a θ.

As estimativas $\hat{t}_a$ que mais se aproximaram da original ocorreram com a reamostragem de locos para os dados Telles & Coelho (1998), de locos e de populações no caso de Seoane (1998) e de populações nos demais conjuntos de dados (TABELA 1).

Para o conjunto de Reis (1996), $\hat{t}_a$ não diferiu de 1 exceto com a reamostragem de indivíduos, quando foi superior a este valor. Embora taxa aparente de cruzamento não possa ser superior a 1, isso ocorreu porque, para esse nível de reamostragem, $\hat{f}$ foi negativo, como pode ser visto pelo seu intervalo de confiança (TABELA 1), e com o estimador de t_a usado, obteve-se estimativa superior a 1 para tal parâmetro. Comportamento idêntico foi encontrado com os dados de Ciampi (1999).

Embora a espécie estudada por Auler (2000) seja dióica e, portanto, fosse esperado que $\hat{t}_a$ não diferisse de 1, observou-se que, para todos os níveis reamostrados $\hat{t}_a$, foi menor que 1. Isso ocorreu porque $\hat{f}$ foi maior que zero em todos os níveis de reamostragem, indicando, como o autor sugeriu, a presença de cruzamentos entre indivíduos aparentados. Para os dados de Sebbenn[4] et al. (s.d.), $\hat{t}_a$ foi menor do que 1 somente quando foram feitas reamostragens de populações. Com os dados de Seoane (1998) e Telles & Coelho (1998), observou-se que, para todos os níveis de reamostragem considerados $\hat{t}_a$ não diferiu de 1, o que era esperado uma vez que $\hat{f}$ não diferiu de zero em nenhum dos intervalos de confiança calculados para esse parâmetro.

Com os resultados obtidos, ficou evidente que se deve tomar cuidado com as estimativas de t_a pois o estimador usado não é muito eficiente e conclusões incorretas podem ser tomadas se os intervalos de confiança não forem considerados.

Deve-se atentar para a importância da reamostragem bootstrap na estimação dos erros das estimativas dos parâmetros. É relevante conhecer a magnitude desses erros, pois se ela for elevada, a precisão das estimativas dos parâmetros é pequena, o que pode ser prejudicial no que se refere, por exemplo, a planos de conservação da variabilidade genética na natureza. A preservação de populações em condições naturais, muitas vezes baseia-se nas medidas de diversidade intra e interpopulacionais, quantificadas pelas estimativas dos parâmetros populacionais. Se eles forem estimados com baixa precisão, decisões errôneas podem ser tomadas quanto à manutenção ou não de populações na natureza, o que pode levar a perda da variabilidade genética, ou ter outras conseqüências.

Seria interessante verificar futuramente se existem fontes de variação mais apropriadas a serem reamostradas em função dos diversos parâmetros de interesse. Van Dongen (1995) afirma que seria melhor reamostrar indivíduos ao invés de locos quando se trabalha com isoenzimas, pois, locos isoenzimáticos podem não ser independentes e nem identicamente distribuídos. Mencionou também que um número pequeno de locos, pode comprometer a reamostragem. Quanto à independência, poderiam ser considerados na reamostragem somente locos em equilíbrio de ligação entre si, o que, em muitos casos, levaria a um número ainda menor de locos analisados. Nesse ponto, uma alternativa seria usar marcadores como microssatélites e RFLP's, que são mais numerosos. Persistiria ainda o problema dos locos poderem não ser identicamente distribuídos. Por outro lado, Petit & Pons (1998) afirmam que reamostrar populações seria mais adequado para estimar a variância dos parâmetros diversidade média intrapopulacional, diversidade total e diferenciação entre populações. Consideraram vários alelos segregando em um loco haplóide e, com base no trabalho de Van

[4]SEBBENN, op. citi. p.2.

Dongen (1995), não reamostraram locos. Assim, as conclusões foram obtidas para três parâmetros específicos, considerando-se apenas um único loco haplóide, sendo, portanto, necessário expandir tal estudo para situações com vários locos em espécies diplóides, incluindo outros níveis de reamostragem e outros parâmetros também.

Estabelecer regras gerais sobre qual fonte de variação é mais importante na reamostragem é uma atitude questionável, pois isso depende de como a diversidade está distribuída entre populações e indivíduos e da homogeneidade entre as estimativas provenientes de diferentes locos. Portanto, no momento, o mais aconselhável para obter conclusões com maior segurança seria considerar os resultados provenientes das reamostragens feitas para cada fonte de variação, em função dos parâmetros de interesse.

Como fontes de variação podem levar a intervalos de confiança conflitantes, é conveniente uma consideração mais geral que possa servir como orientação em pesquisas como as abordadas neste trabalho. Em tais situações, querendo-se fazer inferências a respeito de um dado parâmetro, um critério aqui sugerido é o seguinte: a) o parâmetro é considerado maior do que zero se todos os intervalos forem positivos; b) o parâmetro é considerado menor do que zero se todos os intervalos forem negativos; c) o parâmetro é admitido como nulo se pelo menos um intervalo contiver o zero. Dessa forma, as inferências terão considerável margem de segurança. Esse mesmo raciocínio pode ser aplicado para t_a, sendo que o valor de referência passa a ser o 1 e não o zero, quando se quer investigar se a espécie é alógama.

Pelo teste de normalidade de Kolmogorov-Smirnov, as estimativas de *F* apresentaram distribuição normal somente quando foram reamostrados indivíduos nos dados de Seoane (1998) e Auler (2000), em que os *p*-valores obtidos foram $p>0,15$ e $p=0,13$, respectivamente. Quanto a $\hat{f}$, o teste de normalidade foi não significativo apenas na reamostragem de indivíduos com os dados de Telles & Coelho (1998) ($p>0,15$) e com a reamostragem concomitante de indivíduos e populações nos dados de Ciampi (1999) ($p>0,15$). Para os demais parâmetros, nos diversos níveis de reamostragem, rejeitou-se a hipótese da normalidade, considerando nível mínimo de significância de 0,05. Como conseqüência, muitos procedimentos estatísticos clássicos que tenham como pressuposto a normalidade não podem ser aplicados. Exemplos seriam o uso de teste *z* ou χ^2 para testar a hipótese da nulidade (H_0 : f = *0*) (Weir, 1996) e a construção de intervalos de confiança baseados na distribuição normal. Dessa forma, fica evidente a vantagem e a necessidade do emprego de técnicas de reamostragem para contornar esse problema.

As distribuições empíricas de $\hat{F}$ e $\hat{f}$ tendem à normalidade quando o tamanho amostral é grande. Isso deveu-se ao número de indivíduos amostrados nas pesquisas, que foi maior do que o de locos e de populações. Assim, quando indivíduos estão incluídos na reamostragem, parece que tais estimativas tendem à normalidade. Vencovsky et al. (1997) obtiveram resultados semelhantes. Verificaram que, para oito dos nove locos analisados, $\hat{f}$ teve distribuição normal quando indivíduos foram reamostrados. Por outro lado, observaram que $\hat{t}_a$ não tem distribuição normal.

No que se refere à reamostragem com tamanhos variáveis de amostra é importante esclarecer que, em alguns conjuntos, houve elevado número de bootstraps descartados quando poucos indivíduos foram reamostrados. Nesses casos, estabeleceu-se que o número mínimo de indivíduos a ser considerado nas análises de regressão seria aquele a partir do qual houvesse menos de 10% de descartes. Assim, com os dados de Seoane (1998), Auler (2000) e Sebbenn[5] et al. (s.d.), as regressões foram realizadas considerando número mínimo de seis, cinco e três indivíduos, respectivamente, quando se reamostraram indivíduos. Para os dados de Sebbenn[6] et al. (s.d.), a reamostragem com número variável de populações não foi realizada, uma vez que somente duas populações foram amostradas.

Os valores apresentados na TABELA 2 foram obtidos considerando que locos, indivíduos e populações foram amostrados de conjuntos infinitos de locos, indivíduos e populações na natureza. Tais resultados devem ser considerados com cautela, por se tratarem, geralmente, de extrapolações.

Em geral, observa-se que para obter erros com magnitude de 0,02, as estimativas de *F* e *f*, requereriam maiores números de locos, de indivíduos e de populações do que os geralmente usados em pesquisas com populações naturais (TABELA 2). Isso já era esperado, pois, como discutido anteriormente, tais estimativas apresentaram maiores erros do as estimativas de θ.

Para estimar θ com desvio-padrão máximo de 0,02, o número de locos foi suficiente para os dados de Reis (1996), Ciampi (1999) e Auler (2000). O número de indivíduos foi suficiente nesses mesmos conjuntos e também nos dados de Telles & Coelho (1998) e Sebbenn[7] et al. (s.d.) e o número de populações foi suficiente em todos os casos, exceto na pesquisa de Seoane (1998).

Algumas considerações gerais podem ser feitas ao observar a TABELA 2. Os conjuntos de Reis (1996) e Ciampi (1999) apresentam várias semelhanças entre si, pois tratam de espécies alógamas ($\hat{f}$ não diferiu de zero), com baixa divergência entre populações ($\hat{\theta}$ não diferiu de zero). O número de populações de Reis (1996)

[5,6,7]SEBBENN, op. citi. p.2.

TABELA 2 - Número de locos (L), indivíduos por população (I) e populações (P) necessários para obter erros-padrão das estimativas de [a] F, [b] f, e [c] θ iguais ou menores a 0,02. Dados de diversos autores.

	Reis (1996)			Seoane (1998)			Telles & Coelho (1998)		
Nível	$\hat{F}$	$\hat{f}$	$\hat{\theta}$	$\hat{F}$	$\hat{f}$	$\hat{\theta}$	$\hat{F}$	$\hat{f}$	$\hat{\theta}$
L	51	44	1	19	23	61	160	43	20
I	34	34	7	112	138	38	2	4	2
P	34	33	2	24	51	6	2	1	2
	Ciampi (1999)			**Auler (2000)**			**Sebbenn et al. (s.d.)**		
Nível	$\hat{F}$	$\hat{f}$	$\hat{\theta}$	$\hat{F}$	$\hat{f}$	$\hat{\theta}$	$\hat{F}$	$\hat{f}$	$\hat{\theta}$
L	17	20	1	51	52	1	121	94	18
I	9	13	6	69	69	12	105	72	38
P	4	4	3	56	53	3	-	-	-

[a]índice de fixação total; [b]índice de fixação dentro das populações; [c]grau de divergência entre populações.

foi o dobro do de Ciampi (1999), porém o número de locos e o número médio de indivíduos por população foram praticamente iguais nos dois conjuntos. No entanto, o número de indivíduos e de populações usados por Ciampi (1999) foram suficientes para estimar todos os parâmetros com a precisão desejada, o que não ocorreu com os dados de Reis (1996). O número de locos em ambos os casos foi suficiente apenas para estimar θ com a precisão determinada, mas, observa-se que, para a pesquisa de Ciampi (1999) o número de locos necessário visando estimativas de F e f foi menor do que no caso de Reis (1996). Este fato deve-se provavelmente, ao tipo marcador empregado. Enquanto Ciampi (1999) usou microssatélites, em que o número de alelos por loco variou de 19 a 31, Reis (1996) empregou isoenzimas, em que esse número foi de no máximo quatro. Assim, não só o número de locos, mas também o número total de alelos parece determinante na estimação dos parâmetros F, f e θ com boa precisão.

A espécie estudada por Telles & Coelho (1998) é alógama, como as duas anteriores, mas apresenta divergência entre populações relativamente alta. Também neste caso, o número de locos foi o fator crítico, o que era esperado, uma vez que apenas quatro locos foram usados. O conjunto de Seoane (1998) é referente a uma espécie alógama com baixa divergência entre populações. Pelo que se observa da TABELA 2, para conjuntos desse tipo, são necessários elevados números de populações, de locos e principalmente de indivíduos.

A espécie estudada por Auler (2000) é dióica e apresenta $\hat{F}$, $\hat{f}$ e $\hat{\theta}$ significativos. Numa situação como essa, também são necessários grandes tamanhos amostrais. Nesse conjunto, assim como no de Seoane (1998) e de Sebbenn[8] et al. (s.d.) o polimorfismo não foi elevado, o que pode ter contribuído para que tamanhos amostrais elevados fossem requeridos.

Nos casos estudados não é simples chegar a uma recomendação geral, pois, embora existam algumas semelhanças entre alguns casos, os resultados são particulares de cada conjunto. Assim, seria importante fazer análises prévias do material a ser estudado, como as realizadas neste trabalho, e então, se necessário, voltar ao campo e coletar mais material e/ou aumentar o número de locos usados na avaliação. Considerando o conjunto de Telles & Coelho (1998), por exemplo, observou-se que o número de populações e indivíduos foi adequado, mas o de locos foi insuficiente. Assim, seria necessário apenas aumentar o número de locos analisados.

De uma visão geral dos resultados, é possível afirmar que o número de locos isoenzimáticos que vem sendo utilizado nas pesquisas com populações naturais não é suficiente. Amostrar maior número de populações, igualmente, é estratégia que deve ser adotada. É conveniente enfatizar que, ao aumentar o número de locos, haverá aumento concomitante do número total de alelos. Também, com incremento do número de populações, deverá crescer o número total de indivíduos amostrados. Esses detalhes são importantes e devem ser levados em conta para atingir desvios-padrão das estimativas dos parâmetros, que sejam compatíveis e adequados.

AGRADECIMENTOS

Aos autores que prontamente forneceram os dados analisados neste trabalho e ao CNPq, pelo auxílio financeiro.

REFERÊNCIAS BIBLIOGRÁFICAS

AULER, N.M.F. Caracterização da estrutura genética de populações naturais de *Araucaria angustifolia* (Bert) O. Ktze. no Estado de Santa Catarina. Florianópolis, 2000. 93p. Dissertação (Mestrado) - Universidade Federal de Santa Catarina.

CIAMPI, A.Y. Desenvolvimento e utilização de marcadores microsatélites, AFLP e seqüênciamento de cpDNA, no estudo da estrutura genética e parentesco em populações de copaíba (*Copaifera langsdorffii*) em matas de galeria no cerrado. Botucatu, 1999. 204p. Tese (Doutorado) - Universidade Estadual Paulista "Júlio de Mesquita Filho".

COCKERHAM C.C. Variance of gene frequencies. **Evolution**, v.23, p.72-84, 1969.

COCKERHAM, C.C. Analysis of gene frequencies. **Genetics**, v.74, p.679-700, 1973.

COELHO, A.S.G. **Programa EG:** Análise de estrutura genética de populações pelo método da análise de variância (software). Goiânia: UFG, Inst. de Ciências Biológicas, Depto. de Biologia Geral, 2000a.

COELHO, A.S.G. **Programa EGBV**: Análise de estrutura genética de populações pelo método da análise de variância, com a utilização de bootstraps com amostras de tamanho variável (software). Goiânia: UFG, Inst. de Ciências Biológicas, Depto. de Biologia Geral, 2000b.

EFRON, B.; TIBSHIRANI, R.J. **An introduction to the bootstrap.** New York: Chapman & Hall, 1993. 436p.

HÁLLDEN, C.; NILSSON, N.O.; RADING, I.M.; SÄLL, T. Evaluation of RFLP and RAPD markers in a comparison of *Brassica napus* breeding lines. **Theoretical and Applied Genetics**, v.88, p.123-128, 1994.

MANLY, B.F.J. **Randomization, bootstrap and Monte Carlo methods in biology.** 2. ed. New York: Chapman & Hall, 1997. 399p.

MUNIZ, J.A. Inferência sobre parâmetros relativos à estrutura genética de populações com dados de freqüências gênicas. Piracicaba, 1994. 224p. Tese (Doutorado) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo.

PETIT, R.J.; PONS, O. Bootstrap variance of diversity and differentiation estimators in a subdiveded population. **Heredity**, v.80, p.56-61, 1998.

REIS, M.S. Distribuição e dinâmica da variabilidade genética em populações naturais de palmiteiro (*Euterpe edulis* Martius). Piracicaba, 1996. 210p. Tese (Doutorado) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo.

SEOANE, C.E.S. Efeitos da fragmentação florestal sobre a estrutura genética de populações de *Esenbeckia leiocarpa* Engl. - guarantã - um exemplo de espécie arbória tropical climácica de distribuição agregada. Campinas, 1998. 80p. Dissertação (Mestrado) - Universidade Estadual de Campinas.

SOKAL, R.R.; ROHLF, F.J. **Biometry**: the principles and practice of statistics in biological research. 3.ed. New York: W.H. Freeman & Company, 1995. 887p.

TELLES, M.P.C.; COELHO, A.S.G. Caracterização genética de populações naturais de araticum (*Anonna crassiflora*). **Genetic and Molecular Biology**, v.21, p.199, 1998. Supplement. /Apresentado ao 44. Congresso Nacional de Genética, Águas de Lindóia, 1998 - Resumo/

TIVANG, J.G.; NIENHUIS, J.; SMITH, O.S. Estimation of samplimg variance o molecular marker data using the bootstrap procedure. **Theoretical and Applied Genetics**, v.89, p.259-264, 1994.

VAN DONGEN, S. How should we bootstrap allozyme data? **Heredity**, v.74, p.445-447, 1995.

VENCOVSKY, R.; DIAS, C.T.S.; DEMÉTRIO, C.G.B.; LEANDRO, R.A.; PIEDADE, S.M.S. Reamostragem por "bootstrap" na estimação de parâmetros baseados em marcadores genéticos. In: ENCONTRO SOBRE TEMAS DE GENÉTICA E MELHORAMENTO, 14., Piracicaba, 1997. **Anais**. Piracicaba: ESALQ, Depto. de Genética , 1997. p.59-72.

WEIR, B. S. **Genetic data analysis II**. 2.ed. Sunderland: Sinauer Associates, 1996. 445p.

WEIR, B.S.; COCKERHAM, C.C. Estimating *F*-statistics for the analysis of population structure. **Evolution**, v.38, p.1358-1370, 1984.

Recebido em 08.02.01