Ciência Rural, Santa Maria, v.38, n.5, p.1258-1265, ago, 2008
ISSN 0103-8478

Inferência Bayesiana na análise genética de populações diplóides: estimação do coeficiente de endogamia e da taxa de fecundação cruzada

Bayesian inference in genetic analysis of diploid populations: inbreeding coefficient and outcrossing rate estimation

Ricardo Luis dos Reis[I] Joel Augusto Muniz[I*] Fabyano Fonseca e Silva[II] Thelma Sáfadi[I] Luiz Henrique de Aquino[I]

RESUMO

Neste estudo, utilizou-se a metodologia Bayesiana para estimar o coeficiente de endogamia e a taxa de fecundação cruzada de uma população diplóide por meio do modelo aleatório de COCKERHAM para freqüências alélicas. Um sistema de simulação de dados foi estruturado para validar a metodologia utilizada. O algoritmo Gibbs Sampler foi implementado no software R para obter amostras das distribuições marginais a posteriori para o coeficiente de endogamia e para a taxa de fecundação. O método Bayesiano mostrou-se eficiente na estimação dos parâmetros, pois os valores paramétricos utilizados na simulação encontravam-se dentro do intervalo de credibilidade de 95% em todos os cenários considerados. A convergência do algoritmo Gibbs Sampler foi verificada, validando assim os resultados obtidos.

Palavras-chave: *parâmetros genéticos, Gibbs Sampler, simulação de dados.*

ABSTRACT

The Bayesian methodology was used to estimate the inbreeding coefficient and outcrossing rate in diploid populations by COCKERHAM random model to allelic frequency. The proposed methodology was evaluated by data simulation. The Gibbs Sampler algorithm was implemented in the R statistical software to obtain the random samples of the inbreeding coefficient and outcrossing rate posteriors marginal distributions. The Bayesian method showed good results, because the 95% credible intervals contained the true parameter values to all of the selected scenes. The Gibbs Sampler convergence was checked and this validated the estimation results.

Key words: *genetic parameters, Gibbs Sampler, simulated data.*

INTRODUÇÃO

Entre os diversos aspectos geralmente observados na caracterização genética de populações naturais, a avaliação do grau de estruturação da variabilidade genética dentro dos indivíduos assume grande importância. As relações com que os genes estão distribuídos nos indivíduos são importantes não só por fornecer subsídios para um melhor conhecimento acerca do sistema reprodutivo vigente em uma determinada espécie, mas também por resultar na obtenção de informações básicas que são úteis para o estabelecimento de estratégias mais seguras de coleta e conservação da variabilidade genética existente na população. Endogamia, freqüências alélicas, taxa de fecundação cruzada e tamanho efetivo de população são termos comuns no estudo de genética de populações (COCKERHAM, 1969), sendo que os conceitos e a maior parte da teoria são introduzidos pelos trabalhos clássicos de WRIGHT (1921) e FISHER (1949).

Endogamia é o acasalamento entre indivíduos que são relacionados por ascendência, apresentando como primeiro efeito uma mudança nas freqüências genotípicas de Hardy-Weinberg devido a um aumento na freqüência de genótipos homozigóticos em detrimento das freqüências de genótipos heterozigóticos, segundo a definição de FALCONER (1964). Para o autor, o coeficiente de endogamia *f* quantifica a probabilidade de dois genes, em qualquer loco de um indivíduo, serem originados da cópia de apenas um gene em uma geração anterior.

[I]Departamento de Ciências Exatas, Universidade Federal de Lavras (UFLA), CP 3037, 37200-000, Lavras, MG, Brasil. E-mail: joamuniz@ufla.br. *Autor para correspondência.
[II]Departamento de Informática, setor de Estatística, Universidade Federal de Viçosa (UFV), Campus Universitário, Centro, Viçosa, MG, Brasil.

Recebido para publicação 25.06.07 Aprovado em 19.12.07

Em termos biológicos, o coeficiente de endogamia *f* mede a redução fracionária na heterozigosidade, em relação a uma população de acasalamento aleatório com a mesma freqüência alélica (HARTL & CLARK, 1989). No caso de um loco com dois alelos, as freqüências genotípicas de AA, Aa e aa podem ser expressas em relação ao coeficiente de endogamia pelo modelo endogâmico:

$$p_{AA} = p_A^2 + p_A(1 - p_A) f$$
$$p_{Aa} = 2 p_A(1 - p_A)(1 - f)$$
$$p_{aa} = (1 - p_A)^2 + p_A(1 - p_A) f \qquad (1)$$

Para os autores, esta expressão facilita o entendimento e a comparação das freqüências genotípicas no princípio de Hardy-Weinberg. Quando $f = 0$, tem-se acasalamento aleatório, ou seja, não existe endogamia. Nesse caso, as freqüências genotípicas estarão de acordo com o equilíbrio de Hardy-Weinberg. Quando $f = 1$, tem-se endogamia total e a população só apresentará genótipos homozigóticos AA e aa, respectivamente, nas freqüências p e q.

Em relação à teoria estatística aplicada ao estudo de genética de populações, MUNIZ et al. (1996) estudaram as propriedades dos estimadores do coeficiente de endogamia e da taxa de fecundação cruzada obtidas pela análise de variância com dados de freqüências alélicas em populações diplóides. Os resultados da simulação de dados validaram a utilização destes estimadores.

Uma comparação de fórmulas para estimação da variância do estimador do coeficiente de endogamia obtido na análise de variância das freqüências alélicas em uma população diplóide foi feita por MUNIZ et al. (1997). Eles constataram, via simulação de dados, que as fórmulas propostas apresentaram valores semelhantes e satisfatórios quando a freqüência alélica da população estava entre 0,3 e 0,7, o coeficiente de endogamia da população era inferior a 0,5 e o número de indivíduos maior que 30.

Estudando a estimação do coeficiente de endogamia em uma população diplóide, MUNIZ et al. (1999) avaliaram a distribuição do quociente dos quadrados médios entre indivíduos e entre gene dentro de indivíduos, verificando que o teste F da análise de variância pode ser utilizado para testar a nulidade do coeficiente de endogamia quando a freqüência alélica estiver entre 0,3 e 0,7, trabalhando-se com 30 indivíduos, entre 0,25 e 0,75, com 50 indivíduos, e entre 0,20 e 0,80, com 100 indivíduos.

O método ANOVA foi utilizado com sucesso na estimação das freqüências alélicas, inclusive com resultados validados via simulação Monte Carlo. Porém, eles desconsideram a pressuposição de resíduos normais com média zero e variância σ^2, pois os dados são binários, sendo o valor um na presença do alelo e o valor zero na sua ausência, o que implica em uma possível falta de normalidade dos resíduos. Além disso, devido à complexidade estatística do estimador das freqüências alélicas, também não são apresentadas estimações intervalares, fato que pode ser relevante quando se tem interesse na elaboração de hipóteses a respeito destes parâmetros (MUNIZ et al., 1997; MUNIZ et al., 1999).

Outro parâmetro de grande importância é a taxa de fecundação cruzada, a qual, segundo (COELHO, 2002), é a união entre gametas de indivíduos diferentes, mas da mesma espécie. Populações de indivíduos que apresentam fecundação cruzada têm maiores possibilidades de aumentar a variabilidade genética sem adição de genes novos (por mutação, por exemplo) do que populações de indivíduos com autofecundação. Por meio da recombinação genética, que pode ser promovida pela autofecundação, uma população pode aumentar sua variabilidade genética sem adição de genes novos, produzindo por mutação ou por imigração de indivíduos de outras populações.

A metodologia que utiliza a informação de múltiplos locos para se estimar a taxa de fecundação cruzada e possibilita ainda a obtenção de uma série de outros parâmetros indicadores do sistema reprodutivo tem sido a mais utilizada atualmente (RITLAND & JAIN, 1981). Exemplos de aplicação dessa metodologia podem ser encontrados em MILLAR et al. (2000). Sob este enfoque, tem-se o seguinte para a taxa de fecundação cruzada (t): $t = \frac{1 - f}{1 + f}$

Uma discussão geral sobre os métodos de estimação de parâmetros genéticos com base em dados de freqüências alélicas é apresentado por WEIR (1996). Entre os diversos métodos, o autor destaca o método dos momentos, o método da máxima verossimilhança e a análise de variâncias das freqüências alélicas. O autor aborda ainda a possibilidade de usar técnicas Bayesianas na estimação dos parâmetros genéticos, uma vez que estas podem incorporar informações prévias ao procedimento de estimação, as quais são especificadas por meio da distribuição *a priori*.

Além desta possibilidade de incorporação de informações, outras vantagens relacionadas com a Inferência Bayesiana são caracterizadas pela ausência de pressuposições quanto aos modelos utilizados e pela facilidade da adoção da estimação por intervalo, neste caso denominado de intervalo de credibilidade, o qual é obtido diretamente pelos quantis da distribuição *a posteriori* (HOLSINGER, 2005). Dessa

forma, a utilização da metodologia Bayesiana pode suprir as carências relatadas anteriormente ao se utilizar o método ANOVA.

Para se realizar uma inferência Bayesiana, é necessário, além dos dados amostrais, o estabelecimento de uma informação *a priori* sobre o(s) parâmetro(s) e o cálculo da distribuição *a posteriori* do(s) parâmetro(s). A informação *a priori* é dada pela densidade de probabilidade $P(\theta)$, a qual expressa, de alguma forma, o conhecimento do pesquisador sobre o(s) parâmetro(s) a ser(em) estimado(s). Quando em determinado estudo o pesquisador tem pouca ou nenhuma informação para se incorporar, considera-se uma *priori* não-informativa, por exemplo, a *priori* de Jeffreys (JEFFREYS, 1961). Os dados $Y=\{y_1, y_2, ..., y_n\}$, representados por uma amostra aleatória de uma população com densidade *f* são utilizados na análise Bayesiana por meio da função de verossimilhança $L(y_1, ..., y_n \mid \theta)$, que é o valor da densidade conjunta da amostra aleatória, obtido para a amostra dada.

Portanto, a partir do momento que se opta por uma distribuição *a priori*, seja ela informativa ou não-informativa, e obtém-se a função de verossimilhança, é possível, por meio do Teorema de Bayes representado pela expressão

$$P(\theta \mid Y) = \frac{L(Y \mid \theta)P(\theta)}{\int L(Y \mid \theta)P(\theta)d\theta}$$

sendo $Y=\{y_1, y_2, ..., y_n\}$, obter a distribuição *a posteriori* de θ, de forma que qualquer conclusão a seu respeito é realizada a partir desta distribuição . O denominador é uma constante, chamada de constante de integração, pois só depende da amostra dada e não depende de θ. Portanto temos: $P(\theta|Y) \propto L(Y \mid \theta)P(\theta)$, ou seja, essa expressão pode ser entendida como: *Posteriori* $\propto$ *Verossimilhança* x *Priori*, em que $\propto$ é um símbolo que representa "proporcionalidade".

A distribuição *a posteriori* de um parâmetro θ contém toda a informação probabilística de interesse a respeito do mesmo. Assim, a inferência sobre o parâmetro é realizada por meio desta distribuição. Segundo ROSA (1998), para se inferir em relação a qualquer elemento (componente) de θ, a distribuição *a posteriori* conjunta dos parâmetros, $P(\theta \mid Y)$, deve ser integrada em relação a todos os outros elementos de θ. Assim, se $\theta = (\theta_1, \theta_2)$ e, se o interesse do pesquisador se concentra sobre os componentes de θ_1, tem-se a necessidade da obtenção da distribuição $p(\theta_1|Y)$, denominada de marginal, a qual é dada por:

$$P(\theta_1 \mid Y) = \int_{\boldsymbol{\theta} \neq \theta_1} P(\boldsymbol{\theta} \mid Y) d\boldsymbol{\theta}_{\boldsymbol{\theta} \neq \theta_1}$$

A integração da distribuição conjunta *a posteriori* para a obtenção das distribuições *a posteriori* marginais pode não ser fácil de realizar, necessitando de algoritmos iterativos especializados como, por exemplo, o Gibbs Sampler, o qual faz parte de uma classe de algoritmos denominada de MCMC (Markov Chain-Monte Carlo). Porém, para a utilização desse algoritmo, é necessário que se obtenha, a partir da distribuição *a posteriori*, um conjunto de distribuições chamadas de distribuições condicionais completas.

O algoritmo Gibbs Sampler é uma ferramenta extremamente útil na resolução de problemas que envolvem a estimação de mais de um parâmetro, como é o caso do modelo de COCHERHAM (1969), porém, para utilização desse algoritmo, exige-se que as distribuições condicionais *a posteriori* dos parâmetros tenham formas conhecidas, ou seja, que sejam dadas por distribuições de probabilidade conhecidas. Segundo GAMERMAN (1996), dada as distribuições condicionais completas dos parâmetros, $f(\theta_1|\theta_2, ..., \theta_n)$, $f(\theta_2|\theta_1, \theta_3..., \theta_n)$, ..., $f(\theta_n|\theta_1, \theta_2..., \theta_{n-1})$, o algoritmo Gibbs Sampler pode ser descrito de maneira prática e objetiva da seguinte forma: i) são dados valores iniciais $\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, ..., \theta_n^{(0)})$ para os parâmetros, ii) são gerados iterativamente *n* valores:

$$\begin{aligned}
&\theta_1^{(1)} \text{ de } f_1(\theta_1|\theta_2^{(0)}, \theta_3^{(0)}, ..., \theta_n^{(0)}),\\
&\theta_2^{(1)} \text{ de } f_2(\theta_2|\theta_1^{(1)}, \theta_3^{(0)}, ..., \theta_n^{(0)}),\\
&\qquad\qquad\downarrow\\
&\theta_n^{(1)} \text{ de } f_n(\theta_n|\theta_1^{(1)}, \theta_2^{(1)}, ..., \theta_{n-1}^{(1)}),
\end{aligned}$$

obtendo-se na primeira iteração $\theta^{(1)} = (\theta_1^{(1)}, \theta_2^{(1)}, ..., \theta_n^{(1)})$. Ao final de *N* iterações, é obtido então um conjunto de *N* valores para cada parâmetro, sendo que estes conjuntos representam as amostras das distribuições marginais *a posteriori* dos parâmetros. À medida que o número de iterações aumenta, o conjunto de valores gerados aproxima de sua condição de equilíbrio. Assim, assume-se que a convergência é atingida em uma iteração cuja distribuição esteja arbitrariamente próxima da distribuição de equilíbrio, ou seja, da distribuição marginal desejada (NOGUEIRA et al., 2004).

Uma vez que os algoritmos MCMC são processos iterativos, surge a seguinte questão: Quantas iterações são necessárias para que a convergência seja verificada? Existem vários critérios de convergência relatados na literatura. Entre estes, o critério de GELMAN & RUBIN (1992) é um dos mais utilizados (NOGUEIRA et al., 2004). Esse critério pressupõe que *k* seqüências sejam geradas pelos algoritmos MCMC, partindo de diferentes valores iniciais, num total de *m* iterações cada uma. Essas

seqüências fornecem *k* possíveis resultados inferenciais, relacionados com o valor de *m* assumido. Se esses resultados forem similares, tem-se então um indicativo de que a convergência foi alcançada, isto é, que o número *m* de iterações foi adequado. Estes autores propuseram um fator, dado por $\sqrt{R}$, que quantifica esta similaridade. Portanto, quando este apresentar valores bem próximos a um, verifica-se que a convergência foi alcançada.

Na área de Genética de Populações, trabalhos importantes têm sido desenvolvidos considerando técnicas Bayesianas. AYRES & BALDING (1998) aplicaram a metodologia Bayesiana no estudo do coeficiente de endogamia populacional e KARHU (2001) utilizou marcadores de microsatélites para estimar o coeficiente de endogamia em populações de Pinus por meio de técnicas Bayesianas usando MCMC (Cadeia de Markov e Simulação de Monte Carlo), sendo o Gibbs Sampler utilizado para obtenção de amostras da distribuição *a posteriori* dos parâmetros.

Mais recentemente, uma extensão da análise Bayesiana utilizada para estudar a estrutura genética de populações, sendo a distribuição Beta usada como aproximação da distribuição *a posteriori* do coeficiente de endogamia, foi realizada por HOLSINGER & WALLACE (2004). ARMBORST (2005) relata um caso de estimação multiparamétrica que pode ser tratado com o uso de técnicas Bayesianas e o método MCMC, que é a estimação das proporções alélicas e da medida de endocruzamento de forma simultânea.

O presente trabalho tem por objetivo utilizar a Inferência Bayesiana para estimar o coeficiente de endogamia e a taxa de fecundação cruzada de uma população diplóide, considerando o modelo aleatório de COCKERHAM (1969) para freqüências alélicas. Pretende-se também apresentar a implementação do algoritmo Gibbs Sampler no software livre R (R Development Core Team, 2006).

MATERIAL E MÉTODOS

Um sistema de simulação de dados foi estruturado visando, primeiramente, a avaliar os recursos computacionais empregados e também testar a metodologia. No teste do algoritmo, foram simulados vários cenários considerando uma população diplóide, em que se variou: número de indivíduos (n=10;50;100;200), proporção alélica (valores próximos ou iguais a 0,1-valor baixo; 0,5-valor médio; 0,9-valor alto) e coeficiente de endogamia por meio do modelo endogâmico mencionado anteriormente pela equação 1.

A descrição dos indivíduos em amostras extraídas de populações diplóides com dois alelos foi modelada pela função de COCKERHAM (COCKERHAM, 1969), cuja expressão é a seguinte: $y_{ij}=p+a_i+g_{(j)i}$ em que: y_{ij} é a freqüência do alelo *j* dentro do indivíduo *i*, que assume o valor um na presença do alelo e zero, caso contrário; p é a freqüência paramétrica do alelo A na população; a_i é o efeito do indivíduo *i*, com $i = 1, 2, ..., n$; $a_i \sim N(0,\sigma^2_a)$; $g_{(j)i}$ é o efeito do alelo *j* dentro do indivíduo *i*, com $j = 1, 2$, $g_{(j)i} \sim N(0,\sigma^2_g)$.

A função de verossimilhança foi obtida pela suposição de que os valores de y_{ij}, ou seja, a freqüência alélica de cada indivíduo, segue a distribuição Bernoulli, uma vez que se tem o valor um para a presença do alelo e o valor zero para a sua ausência. Portanto, de acordo com esta suposição, tem-se:

$$L(y|p,a,\sigma^2_a,\sigma^2_g) \propto p^T(1-p)^{2n-T}$$

em que: $T=\sum_{i=1}^{n}\sum_{j=1}^{2} y_{ij}$ e a=$\{a_1, ..., a_n\}$.

As distribuições *a priori* usadas para os parâmetros são representadas nas expressões de (2) a (5):

$$p \sim Beta(\alpha,\beta), \quad (2)$$

$$\mathbf{a} \sim N(0, I\sigma^2_a), \quad (3)$$

$$\sigma^2_a \sim IG(a_1,b_1), \quad (4)$$

$$\sigma^2_g \sim IG(a_2,b_2), \quad (5)$$

em que: a expressão (2) é uma distribuição Beta com hiperparâmetros α e β; a expressão (3) é uma distribuição normal multivariada com vetor de média zero e matriz de variâncias $I\sigma^2_a$; e as expressões (4) e (5) são representadas pelas distribuições gama inversa, com hiperparâmetros a_1, b_1, a_2 e b_2. Os valores utilizados para esses hiperparâmetros foi um, pois há o interesse em tomá-los de tal modo que as *prioris* são "praticamente" não-informativas.

De acordo com as definições apresentadas, obtém-se a seguinte distribuição conjunta *a posteriori:*

$$p(p,\sigma^2_a,\sigma^2_g,a|y) \propto L(y|p,\sigma^2_a,\sigma^2_g,a)p(p)p(\sigma^2_a)p(\sigma^2_g)\prod_{i=1}^{n}p(a_i) \quad (6)$$

E tem-se:

$$p(p,\sigma^2_a,\sigma^2_g,a|y) \propto p^T(1-p)^{2n-T}\,Beta(\alpha,\beta)IG(a_1,b_1)IG(a_2,b_2)N(0,\sigma^2_a)$$

As distribuições condicionais completas *a posteriori*, necessárias à implementação do algoritmo de Gibbs Sampler, foram obtidas da equação acima e são apresentadas como segue:

$$p(p \mid \sigma^2_a, \sigma^2_g, a, y) \sim Beta(\alpha, 2(n-T)+\beta) \quad (7)$$

$$p(a_i \mid \sigma^2_a, \sigma^2_g, a, y) \sim N(0, \sigma^2_a)$$

para cada $i=1, ..., n$, (8)

$$p(\sigma^2_a \mid p, a, \sigma^2_g, y) \sim IG(n/2 + a_1, \sum_{i=1}^{n} a_i^2/2 + b_1) \quad (9)$$

$$p(\sigma^2_g \mid \sigma^2_a, \sigma^2_g, y) \sim IG(-T+a_2, b_2) \quad (10)$$

O algoritmo Gibbs Sampler foi implementado utilizando-se a linguagem R. Considerou-se, em todas as análises efetuadas, um número fixo de 10.000 iterações, com espaçamento entre os pontos amostrados de 20 observações e no aquecimento (*burn-in*) desprezaram-se as 2.000 primeiras iterações. As inferências foram realizadas, portanto, considerando as 8.000 iterações restantes. O processo foi repetido 100 vezes para a validação do método.

Para avaliação da convergência do algoritmo Gibbs Sampler, optou-se pelo critério de GELMAN & RUBIN (1992), o qual encontra-se disponível no pacote BOA (*Bayesian Output Analisys*) do software livre R.

Quando se tem interesse na distribuição marginal de determinada função dos parâmetros, dada por $\phi = g(\theta) = g(\theta_1, \theta_2..., \theta_n)$, podem ser obtidas de forma indireta amostras desta distribuição marginal por meio dos valores gerados para $\theta_1, \theta_2..., \theta_n$ via algoritmo Gibbs Sampler, porém a convergência deste algoritmo deve ser fielmente constatada (ROSA, 1998).

Para a definição dos valores do coeficiente de endogamia, foram utilizados os valores gerados e calculados por meio da seguinte relação:

$$f = \frac{\sigma_a^2 - \sigma_g^2}{\sigma_a^2 + \sigma_g^2}.$$

Portanto, como há 8.000 valores para σ^2_a e para σ^2_g, automaticamente há também 8.000 valores para *f*, ou seja, dessa forma é possível obter indiretamente amostras da distribuição *a posteriori* de *f*. Este mesmo princípio foi usado para obter as amostras da distribuição *a posteriori* da taxa de fecundação cruzada: $t = \frac{1-f}{1+f}$.

As funções implementadas na linguagem R, para todos os cálculos abordados, são mostradas na tabela 1.

Tabela 1 - Funções implementadas na linguagem R.

Função Simulação de dados

```
IndAmostra = function (n,p) {
 nAlelo = 2;  y = matrix(0,n,2)
  for (i in 1:n)
  {
   for (j in 1:nAlelo)
   {
    y[i,j] = rbinom(1,1,p)
    }
   }
   pA = sum(y)/length(y); return(list(y=y,pA=pA))
}
Função Gibbs Sampler
Gibbs = function(Alfa,Beta,a1,a2,b1,b2,qmInd,qmGen,mTrat,F,Tfc,y,niter,nbur  n=n/2)
{
N    = length(y)/2
p    = matrix(0, nrow=niter); f    =  matrix(0, nrow=niter); t    = matrix(0, nrow=niter);
ai   = matrix(0, nrow=niter,ncol=N); vara = matrix(0, nrow=niter);
varg = matrix(0, nrow=niter)
somA = sum(y)
Beta = (2*(N-somA)) + Beta
a2   = - somA + a2; a1   = N/2 + a1;
p[1] = pA; ai[1,] = mTrat[1,]; vara[1] = qmInd; varg[1] = qmGen;  f[1] = F;  t[1] = Tfc;
  for (i in 2:niter)
  {
    p [i] = rbeta(1, Alfa, Beta )
    sumAi = 0
      for (l in 1:N) {
       ai [i,l] = rnorm(1,0, vara[i-1])
         sumAi = sumAi + (ai[i,l]^2)
    }
   b11 = sumAi/2 + b1
  vara[i]  =  rinvgamma(1, a1, b11)
varg [i] =  rinvgamma(1, a2, b2)
   f[i]    =  (vara[i]-varg[i])/ (vara[i]+varg[i]);    t[i]    =  (1-f[i])/(1+f[i]);
   }
return(list(p=p,ai=ai, vara=vara, varg=varg,f=f,t=t))
}
```

RESULTADOS E DISCUSSÃO

Por meio dos resultados apresentados na tabela 2 é possível avaliar o comportamento das estimativas dos parâmetros analisados. Nota-se que as estimativas obtidas pela metodologia Bayesiana foram muito próximas aos valores paramétricos da simulação, inclusive é possível verificar que todos os

Tabela 2 - Média *a posteriori*, desvio padrão, intervalo de credibilidade de 95% e fator de Gelman-Rubin para os parâmetros considerando n=10, 50, 100 e 200.

Freqüência alélica	Parâmetros	Valores paramétricos	Média	Desvio padrão	2,5%	97,5%	$\sqrt{R}$
	p	0,15	0,1499	0,0099	0,1312	0,1704	0,9998
Baixa (n=10)	*f*	0,00	0,0012	0,0041	0,0000	0,0118	1,0012
	t	1,00	0,9976	0,0080	0,9765	1,0000	1,0000
	p	0,40	0,4000	0,0099	0,3805	0,4199	0,9999
Média (n=10)	*f*	0,00	0,0009	0,0030	0,0000	0,0098	0,9994
	t	1,00	0,9980	0,0059	0,9804	1,0000	1,0012
	p	0,85	0,8499	0,0101	0,8296	0,8690	0,9999
Alta (n=10)	*f*	0,64	0,6435	0,0588	0,5132	0,7414	1,0000
	t	0,21	0,2184	0,0448	0,1484	0,3216	0,9999
	p	0,10	0,0999	0,0099	0,0814	0,1203	0,9999
Baixa (n=50)	*f*	0,00	0,0063	0,0185	0,0000	0,0688	1,0000
	t	1,00	0,9880	0,0339	0,8711	1,0000	1,0000
	p	0,55	0,5498	0,0099	0,5302	0,5692	1,0000
Média (n=50)	*f*	0,24	0,2456	0,0287	0,1884	0,3000	0,9999
	t	0,61	0,6064	0,0372	0,5384	0,6828	1,0002
	p	0,86	0,8599	0,0101	0,8393	0,8791	0,9998
Alta (n=50)	*f*	0,17	0,1883	0,0576	0,0710	0,2970	1,0000
	t	0,69	0,6871	0,0826	0,5419	0,8674	0,9999
	p	0,12	0,1249	0,0099	0,1062	0,1452	0,9999
Baixa (n=100)	*f*	0,00	0,0166	0,0295	0,0000	0,1036	1,0000
	t	1,00	0,9688	0,0537	0,8121	1,0000	1,0000
	p	0,46	0,4649	0,0099	0,4453	0,4846	0,9999
Média (n=100)	*f*	0,06	0,0697	0,0280	0,0139	0,1249	0,9999
	t	0,88	0,8708	0,0492	0,7778	0,9724	1,0000
	p	0,91	0,9152	0,0099	0,8950	0,9338	0,9999
Alta (n=100)	*f*	0,04	0,0867	0,0743	0,0000	0,2500	0,9999
	t	0,92	0,8487	0,1219	0,5999	1,0000	0,9999
	p	0,08	0,0849	0,0100	0,0664	0,1053	1,0000
Baixa (n=200)	*f*	0,16	0,2302	0,0889	0,0489	0,3966	1,0010
	t	0,71	0,6343	0,1207	0,4319	0,9065	1,0000
	p	0,48	0,4848	0,0099	0,4653	0,5044	0,9999
Média (n=200)	*f*	0,00	0,02622	0,0233	0,0000	0,0796	0,9999
	t	0,99	0,9498	0,0435	0,8524	1,0000	1,0000
	p	0,90	0,9050	0,0099	0,8846	0,9237	1,0000
Alta (n=200)	*f*	0,13	0,1911	0,0804	0,0235	0,3435	1,0000
	t	0,76	0,6868	0,1156	0,4886	0,9540	0,9999

intervalos de credibilidade contêm os verdadeiros valores dos parâmetros. Em relação à freqüência do alelo A, verificou-se que quanto maiores os valores de *p* e, conseqüentemente, maior número de homozigotos, a tendência é produzir coeficientes de endogamia altos.

De forma geral, os resultados obtidos foram mais consistentes que aqueles apresentados por MUNIZ et al. (2001), os quais utilizaram o mesmo sistema de simulação, mas com estimativas obtidas sob o ponto de vista freqüentista. De acordo com esses autores, para tamanho de amostra inferior a 50, ou nos casos em que as populações apresentaram freqüência alélica média fora do intervalo 0,30 a 0,70, os estimadores em questão não produzem valores concordantes com os valores teóricos da simulação. Portanto, esta limitação não foi observada na presente trabalho.

Também é possível observar que, nas configurações em que o coeficiente de endogamia (*f*) foi zero, este mesmo valor encontrava-se no intervalo de credibilidade de 95%, o que caracteriza um teste de hipótese para este parâmetro, uma vez que *f* igual a zero indica que a população encontra-se em equilíbrio de Hardy-Weinberg. Esta informação é muito importante, pois atualmente utiliza-se a metodologia dos Modelos Mistos (HENDERSON, 1984) como ferramenta no melhoramento genético, mas esta metodologia apenas apresenta propriedades ótimas se a população em questão apresentar este equilíbrio.

A cadeia de valores gerados pelo algoritmo Gibbs Sampler mostrou um comportamento estacionário, não sendo constatados picos e tendências. Esta estabilidade também é ratificada pelos valores de $\sqrt{R}$ na tabela 2. Por essas razões, é possível assumir que o Gibbs Sampler convergiu, validando assim as estimativas apresentadas.

As distribuições marginais *a posteriori* para o coeficiente de endogamia e para a taxa de fecundação cruzada caracterizam-se, respectivamente, como distribuições assimétricas à esquerda e à direita, sendo esta assimetria mais evidente quando o número de indivíduos é pequeno. Esta caracterização reforça a idéia da precaução estatística que se deve adotar em relação à suposição de normalidade para estes parâmetros de genética populacional.

CONCLUSÕES

Todo o processo de inferência Bayesiana fundamentado no modelo de COCKERHAM (1969) foi eficiente, pois propiciou resultados condizentes validados pelo estudo de simulação de dados adotados, destacando-se os intervalos de credibilidade para as estimativas dos parâmetros genéticos considerados.

A utilização do Software Livre R facilitou a execução deste trabalho, pois possibilitou que a realização do estudo de simulação de dados e a implementação do algoritmo Gibbs Sampler fossem realizadas de forma prática e objetiva.

REFERÊNCIAS

AYRES, K.L.; BALDING, D.J. Measuring departures from Hardy-Weinberg: a Markov chain Monte Carlo method for estimating the inbreeding coefficient. **Heredity**, v.80, p.769-777, 1998.

ARMBORST, T. **Métodos para medir o desequilíbrio de Hardy-Weinberg através de medidas de endocruzamento**. 2005. 205f. Dissertação (Mestrado em Estatística) - Universidade Federal de Minas Gerais, Belo Horizonte, MG.

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

COELHO, A.S.G. **Abordagem Bayesiana na análise genética de populações utilizando dados de marcadores moleculares**. 2002. 92f. Tese (Doutorado em Genética e Melhoramento de Plantas) - Universidade de São Paulo, Piracicaba, SP.

FALCONER, D.S. **Introduction to quantitative genetics**. New York: The Ronald, 1964. 365p.

FISHER, R.A. **The theory of inbreeding**. Edinburg: Oliver and Boyd, 1949. 120p.

GAMERMAN, D. Simulação estocástica via cadeias de Markov. In: SINAPE -ABE, 12., 1996, Caxambu, Minas Gerais, Brasil. **Anais...** Caxambu: Associação Brasileira de Estatística, 1996. 196p.

GELMAN, A.; RUBIN, D.B. Inference from iterative simulation using multiple sequences. **Statistical Science**, v.7, p.457-472, 1992.

HARTL, D.L.; CLARK, A.G. **Principles of population genetics**. Sunderland: Sinauer Associates, 1989. 468p.

HENDERSON, C.R. **Aplications of linear models in animal breeding.** Guelph: University of Guelph, 1984. 462p.

HOLSINGER, K.E.; WALLACE, L.E. Bayesian approaches for the analysis of populatioa genetic structure: an example from Platanthera leucophaea (Orchidaceae). **Molecular Ecology**, v.13, n.4, p.887-896, 2004.

HOLSINGER, K.E. **Bayesian population genetic data analysis**. Boston: Department of Ecology & Evolutionary Biology, University of Connecticut, 2005. 124p.

JEFFREYS, H. **Theory of probability**. Oxford: Claredon, 1961. 325p.

KARHU, A. **Estimation of inbreeding in radiata pine populations using microsatellites**. Finland: University of Oulu, 2001. Acesso em 20 de abr. 2006. Online. Disponível em: http://herkules.oulu.fi/isbn9514259246.

MILLAR, M.A. et al. Mating system studies in jarrah, ***Eucalyptus marginata*** (*Myrtaceae*). **Australian Journal of Botany**, v.48, n.4, p.475-479, 2000.

MUNIZ, J.A. et al. Properties of estimators of the inbreeding coefficient and the rate of cross fertilization obtained from gene frequency data in a diploid population. **Brazilian Journal of Genetics**, v.19, n.3, p.485-491, 1996.

MUNIZ, J.A. et al. A variância do estimador do coeficiente de endogamia obtido pelo método dos momentos em uma população diplóide. **Revista de Matemática e Estatística**, v.15, p.131-143, 1997.

MUNIZ, J.A. et al. Teste de hipótese sobre o coeficiente de endogamia de uma população diplóide. **Ciência e Agrotecnologia**, Lavras, v.23, n.2, p.410-420, 1999.

MUNIZ, J.A. et al. Teste de hipótese sobre o coeficiente de coancestria de populações haplóides. **Pesquisa Agropecuária Brasileira**, Brasília, v.36, n.1, p.15-25, 2001.

NOGUEIRA, D.A. et al. Avaliação de critérios de convergência para o método de Monte Carlo via Cadeias de Markov. **Revista Brasileira de Estatística**, IBGE, v.65, n.224, p.59-88, 2004.

ROSA, G.J.M. **Análise Bayesiana de modelos lineares mistos robustos via Amostrador de Gibbs**. 1998. 57f. Tese (Doutorado em Estatística e Experimentação Agronômica) - Universidade de São Paulo, Piracicaba, SP.

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

R Development Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, 2006. Acessado em 20 de mai. 2006. Online. Disponível em: http://www.R-project.org.

WEIR, B.S. **Genetic data analysis II**. Methods for discrete population genetic data. Sunderland: Sinauer Associates, 1996. 445p.

WRIGHT, S. System of mating. **Genetics**, v.6, p.111-178, 1921.