Ciência Rural, Santa Maria, v.39, n.6, p.1752-1759, set, 2009
ISSN 0103-8478

Abordagem bayesiana da sensitividade de modelos para o coeficiente de endogamia

Bayesian approach to sensitivity of models for inbreeding coefficient

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

RESUMO

Este trabalho tem como objetivo realizar uma análise bayesiana de modelos, por meio do fator de Bayes, para o desequilíbrio de Hardy-Weinberg. Pretende-se também testar a metodologia por meio da simulação de dados e aplicá-la a um conjunto de dados reais. Na definição dos modelos, utilizaram-se as prioris Dirichlet (modelo 1), Beta - função degrau Uniforme (modelo 2), Uniforme - função degrau Uniforme (modelo 3) e as prioris independentes Uniformes (modelo 4) relacionadas aos parâmetros coeficiente de endogamia e proporção alélica. Foi implementado um algoritmo no software livre R para realizar a amostragem pelo Metropolis-Hastings das distribuições condicionais a posteriori dos parâmetros dos modelos. A convergência das cadeias foram monitoradas por meio de procedimentos implementados no pacote BOA do software livre R. As comparações entre os modelos indicaram que o mais adequado, ou seja, o que melhor descreve o fenômeno em estudo, é o modelo 1, em comparação aos demais, tanto para os dados simulados, quanto para os dados reais. Em virtude dos resultados apresentados, pode-se atestar que a abordagem Bayesiana apresentou bons resultados, ou seja, por meio das distribuições a posteriori condicionais completas, foram verificadas a confiabilidade e a precisão da metodologia na comparação dos modelos.

Palavras-chave: *fator de Bayes, desequilíbrio de Hardy-Weinberg, simulação de dados.*

ABSTRACT

The aim of this research is to perform a Bayesian characterization of the Hardy-Weinberg disequilibrium through the Bayes factor. The methodology is tested by using both simulation study and actual data. It was used the following priors for the Bayesian models: Dirichlet (model 1), beta – step uniform function (model 2), uniform – step uniform function (model 3) and independent uniforms for the inbreeding coefficients and allele frequencies (model 4). Metropolis-Hasting algorithms were implemented using the software R to simulate multiple draws from the posterior distribution. Convergence of the Metropolis-Hasting algorithms was assessed by many methods available at R package BOA. Results showed that the model 1 presents the best performance for both simulation study and actual data. The results also showed that the Bayesian approach provides models that are useful for the analysis of the Hardy-Weinberg disequilibrium and inbreeding coefficient.

Key words: *Bayes factor, Hardy-Weinberg disequilibrium, simulation.*

INTRODUÇÃO

As informações obtidas pelos estudos sobre a caracterização da estrutura genética de diferentes espécies são utilizadas de modo decisivo no estabelecimento de estratégias mais adequadas para a conservação, o manejo e o melhoramento genético dessas espécies de interesse. Em 1908, Godfrey Harold Hardy e Wilhelm Weinberg demonstraram independentemente o princípio das proporções alélicas em uma população, ficando este conhecido como Lei de Hardy-Weinberg ou Lei do equilíbrio de Hardy-Weinberg. Esta afirma que, na ocorrência de cruzamentos ao acaso e na ausência de fatores evolutivos como seleção, mutação e migração, as proporções alélicas e genotípicas permanecem constantes de geração para geração. Considerando um

[I]Programa de Pós-graduação em Estatística e Experimentação Agropecuária, Universidade Federal de Lavras (UFLA), Campus Universitário, CP 3037, 37200-000, Lavras, MG, Brasil. E-mail: joamuniz@ufla.br. *Autor para correspondência.
[II]Departamento de Informática, Universidade Federal de Viçosa (UFV), Viçosa, MG, Brasil.

Recebido para publicação 17.03.08 Aprovado em 11.02.09

gene com dois alelos *A* e *B*, com proporções p_A e $p_B = 1 - p_A$, respectivamente, as proporções genotípicas na população seriam dadas pela seguinte relação: $P_{AA} = p^2_A$ Proporção do genótipo homozigoto *AA;* $P_{AB} = 2p_Ap_B$ Proporção do genótipo heterozigoto *AB*; $P_{BB} = p^2_B$ Proporção do genótipo homozigoto *BB*.

O estudo das violações à Lei de Hardy-Weinberg tem grande interesse em Genética de Populações. Um dos parâmetros mais utilizados para medir esse desequilíbrio é o coeficiente de endogamia *f* (WEIR, 1996), que avalia o quanto a endogamia (cruzamentos entre parentes) reduz o número de indivíduos heterozigotos, acarretando assim aumento no grau de parentesco entre indivíduos e na quantidade de alelos recessivos em sucessivas gerações. As proporções genotípicas homozigotas e heterozigotas para o caso de um gene com dois alelos, sob a violação do modelo de Hardy-Weinberg, segundo WEIR (1996), são dadas por:

$$\begin{cases} P_{AA} = p_A^2 + p_A(1-p_A)f \\ P_{AB} = 2p_A(1-p_A)(1-f) \\ P_{BB} = (1-p_A)^2 + p_A(1-p_A)f \end{cases}, \quad (1)$$

sendo os limites de f obtidos por: $\max[-p_A/(1-p_A), -(1-p_A)/p_A] \leq f \leq 1$ em que o limite inferior de *f* depende das proporções alélicas.

Atualmente, a abordagem bayesiana vem sendo utilizada com sucesso em várias áreas da ciência. Na inferência bayesiana, o parâmetro desconhecido é considerado uma variável aleatória, assumindo assim uma distribuição de probabilidade associada (*priori*), a qual é especificada antes da observação dos dados (BOX & TIAO, 1992). Esse conhecimento pode ser obtido considerando-se, por meio de análises anteriores, a experiência do pesquisador na área em questão ou as publicações sobre o assunto que se deseja pesquisar ou estudar.

Com base na definição do Teorema de Bayes, essa distribuição *a priori* é combinada com a informação contida nos dados amostrais (função de verossimilhança), induzindo uma distribuição *a posteriori*. Portanto, toda a inferência relativa a um determinado parâmetro é realizada utilizando-se a distribuição *a posteriori*, podendo esta ser resumida por meio da média, da moda, da mediana e do intervalo de credibilidade e/ou dos intervalos de máxima densidade a posteriori (PAULINO et al., 2003).

O Teorema de Bayes é fundamental na construção da inferência bayesiana e, basicamente, é o resultado de uma probabilidade condicional. Para o caso em que o parâmetro θ é contínuo, o teorema é dado por:

$$\pi(\theta|x) = \frac{L(x|\theta)\pi(\theta)}{\int L(x|\theta)\pi(\theta)d\theta}$$ e, no caso em que θ é discreto, tem-se: $$\pi(\theta|x) = \frac{L(x|\theta)\pi(\theta)}{\sum L(x|\theta)\pi(\theta)}$$, em que: $\pi(\theta|x)$ é a distribuição *a posteriori* do parâmetro θ, $L(x|\theta)$ é a função de verossimilhança e $\pi(\theta)$ é a distribuição *a priori* de θ. Como nessas expressões o denominador independe do parâmetro, este pode ser considerado como uma constante normalizadora e então essas expressões podem ser representadas, de forma proporcional, por: $\pi(\theta|x) \propto L(x|\theta)\,\pi(\theta)$.

Na aplicação do Teorema de Bayes, em estudos de modelos para o coeficiente de endogamia, a distribuição conjunta *a posteriori* dos parâmetros, deve ser integrada em relação a um parâmetro a fim de obter a distribuição marginal do outro parâmetro (PAULINO et al., 2003). A integração dessa distribuição geralmente não é analítica, necessitando de algoritmos especializados denominados de algoritmos MCMC (*Markov Chain Monte Carlo*).

O algoritmo de Metropolis-Hastings permite gerar uma amostra da distribuição conjunta *a posteriori* $\pi(\theta_1, \theta_2, \ldots, \theta_d|x)$, sendo *d* a dimensão do espaço de parâmetros, a partir das distribuições condicionais completas com formas desconhecidas. A ideia do algoritmo é que um valor é gerado de uma distribuição auxiliar ou candidata e este é aceito com uma dada probabilidade (METROPOLIS et al., 1953; HASTINGS, 1970). Se as distribuições condicionais completas possuem formas conhecidas, utiliza-se um caso especial do Metropolis-Hastings, o amostrador de Gibbs, no sentido que seja fácil amostrar de seus elementos. Nesse processo, as primeiras iterações são influenciadas pelo estado inicial e podem ser descartadas (*burn-in*. Para se obter uma amostra independente, as observações finais devem ser obtidas a cada *k* iterações *(thin)*.

Os algoritmos Metropolis-Hastings e o amostrador de Gibbs são processos iterativos, e as cadeias resultantes desses métodos necessitam ter sua convergência constatada. Os métodos comumente utilizados para diagnosticar a convergência são aqueles propostos por: GEWEKE (1992), que se baseia na igualdade de médias da primeira e última parte da cadeia; GELMAN & RUBIN (1992), baseado na comparação de duas ou mais cadeias paralelas; RAFTERY & LEWIS (1992b), baseado na acurácia da estimação do quantil, e HEIDELBERGER & WELCH (1993), que usa testes estatísticos para avaliar a hipótese nula de estacionariedade da amostra gerada. Estes estão implementados no pacote BOA (*Bayesian*

Output Analysis) do software livre R (R DEVELOPMENT CORE TEAM, 2007).

NOGUEIRA et al. (2004) sugerem a utilização desses critérios de forma combinada, ou seja, primeiro aplica-se RAFTERY & LEWIS (1992b) em uma amostra piloto para a determinação do tamanho ideal da sequência, depois determina-se o tamanho do *burn-in* pelo critério de HEIDELBERGER & WELCH (1993) e, por último, monitora-se a convergência por meio do critério de GEWEKE (1992) e GELMAN & RUBIN (1992).

Uma das formas de se comparar e selecionar modelos em estudo é por meio do fator de Bayes,

definido por: $FB_{(M_i,M_j)} = \frac{\pi(x \mid M_i)}{\pi(x \mid M_j)}$, em que $\pi(x|M_i)$ e $\pi(x|M_j)$ são as verossimilhanças marginais de cada modelo, dadas por: $\pi(x \mid M) = \int L(x \mid \theta, M)\pi(\theta \mid M)d\theta$, em que $L(x|\theta,M)$ é a função de verossimilhança para o modelo M, $(x|\theta,M)$ a distribuição *a priori* e θ o parâmetro do modelo M. Na prática, em algumas situações, as quantidades $\pi(x|M_i)$ e $\pi(x|M_j)$ podem ser calculadas analiticamente. Mas, em geral, essas integrais são de difícil solução, e os métodos MCMC são usados para obter soluções aproximadas (KASS & RAFTERY, 1995). Assim gera-se uma amostra de tamanho t ($\theta^1, \theta^2, \ldots, \theta^t$), a partir da qual pode ser calculada a verossimilhança marginal por meio da seguinte expressão:

$\hat{\pi}(x|\theta) = \frac{1}{T}\sum_{i=1}^{t} L(\theta^i | x, M)\pi(\theta^i | M)$, em que T é o tamanho da amostra final.

Uma interpretação para o fator de Bayes é dada em JEFFREYS (1961), em que: $FB_{(Mi, Mj)} < 1$ demonstra evidência a favor de M_j; $1 \leq FB_{(Mi, Mj)} < 3,2$ demonstra evidência muito fraca a favor de M_i; $3,2 \leq FB_{(Mi, Mj)} < 10$ demonstra evidência fraca a favor de M_i; $10 \leq FB_{(Mi, Mj)} < 100$ demonstra evidência forte a favor de M_i e $FB_{(Mi, Mj)} \geq 100$ demonstra evidência muito forte a favor de M_i.

Do ponto de vista biométrico, trabalhos recentes têm revelado uma grande aplicabilidade da abordagem bayesiana para estudos genéticos populacionais (COELHO, 2002). Os métodos frequentistas adotados por NEI & CHESSER (1983) e ROBERTSON & HILL (1984) levam em conta a presença de grupos na população. Esses grupos são estabelecidos de acordo com cruzamentos preferenciais em razão de alguma característica, ou seja, os indivíduos tendem a se cruzar com aqueles mais próximos ou com características em comum como: estatura, origem, entre outros.

Uma discussão geral sobre os métodos de estimação de parâmetros genéticos, com base em dados de proporções alélicas, é apresentada por WEIR (1996). Dentre os diversos métodos, o autor destaca o método dos momentos, o método da máxima verossimilhança e a análise de variâncias das proporções alélicas. No caso da técnica da análise de variância, são abordados os casos de organismos haplóides, bem como de populações diplóides com modelos hierárquicos de até quatro níveis. O autor enfatiza também a possibilidade de usar o enfoque bayesiano na estimação dos parâmetros genéticos, uma vez que este incorpora informações prévias ao procedimento de estimação, sendo úteis na descrição da estrutura genética de populações, principalmente, nas situações em que a estimação envolve a utilização de proporções alélicas. MUNIZ et al. (1999) estudaram as distribuições dos quadrados médios na análise de variância da proporção alélica de uma amostra de indivíduos de uma população diplóide, procurando-se avaliar o teste F proposto por COCKERHAM (1969). Os autores verificaram que este pode ser utilizado para testar a nulidade do coeficiente de endogamia quando a proporção 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. REIS et al. (2008) utilizaram a metodologia bayesiana para estimar o coeficiente de endogamia e a taxa de fecundação cruzada de uma população diplóide considerando o modelo aleatório de COCKERHAM (1969). Os resultados encontrados pelos autores foram validados por estudo de simulação e se mostraram condizentes com os resultados relatados na literatura.

Na avaliação de divergências do equilíbrio de Hardy-Weinberg, os métodos bayesianos permitem a incorporação da incerteza relativa aos parâmetros *nuisance* (parâmetros pelos quais não se tem interesse direto), isto é, as proporções alélicas, mesmo que o interesse maior esteja somente nas inferências sobre o coeficiente de endogamia *f*. AYRES & BALDING (1998) consideram a falta de restrição ao espaço paramétrico de *f* uma desvantagem da utilização do método frequentista, que apresentou piores resultados quando comparados ao método bayesiano, utilizando *priori* Uniforme. SHOEMAKER et al. (1998) descrevem uma metodologia bayesiana para estudar o equilíbrio de Hardy-Weinberg, considerando dois parâmetros, o coeficiente de desequilíbrio e o coeficiente de endogamia, avaliando a probabilidade dos parâmetros estarem em um determinado intervalo de equilíbrio. Os autores utilizaram três *prioris* para cada parâmetro (Dirichlet, Beta - função degrau Uniforme e Uniforme - função degrau Uniforme) e concluíram que as *prioris* que continham a função degrau apresentaram os melhores resultados.

Um caso de estimação multi-paramétrico que pode ser tratado com o uso de técnicas bayesianas e o método de MCMC é a estimação das proporções alélicas e da medida de endocruzamento ou endogamia, já que, em grande parte das situações, há vários alelos num mesmo loco na população e a estimação via máxima verossimilhança é complexa (ARMBORST, 2005). A autora relata que, entre os três métodos utilizados, sendo dois clássicos e um bayesiano, o último apresenta maior qualidade, pois respeita o espaço paramétrico em que o coeficiente de endogamia está definido.

A distribuição a *priori* desempenha um papel muito importante na inferência bayesiana, e o diferencial do presente estudo em relação ao de SHOEMAKER et al. (1998) está na comparação das *prioris* adotadas para incorporar a incerteza relativa ao parâmetro de interesse, o coeficiente de endogamia, e ao parâmetro *nuisance*, a proporção alélica. Estas são a Dirichlet, que é a conjugada natural da distribuição Multinomial, a *priori* não informativa Uniforme, a *priori* Beta e a função degrau, que é uma função de números reais escrita como uma combinação linear finita de funções indicadoras de certos intervalos. Além das *prioris* supracitadas foram consideradas também as *prioris* independentes Uniformes relatadas no trabalho de AYRES & BALDING (1998).

Este trabalho tem como objetivos utilizar a abordagem bayesiana na comparação de modelos para o coeficiente de endogamia, bem como testar a metodologia, por meio da simulação de dados, e aplicá-la a um conjunto de dados reais.

MATERIAL E MÉTODOS

Considere que n_1, n_2 *e* n_3 representem a quantidade observada de genótipos *AA, AB e BB*, respectivamente, em uma amostra de tamanho $n = n_1+n_2+n_3$. A função de verossimilhança é dada por uma distribuição multinomial com parâmetros p_A e f e pode ser escrita como:

$$L(p_A, f \mid n_1, n_2, n_3) = \frac{n!}{n_1!n_2!n_3!}[p_A^2 + p_A(1-p_A)f]^{n_1} \cdot [2p_A(1-p_A)(1-f)]^{n_2}[(1-p_A)^2 p_A(1-p_A)f]^{n_3}$$

As distribuições *a priori* utilizadas foram a Dirichlet com hiperparâmetros inteiros γ_1, γ_2 e γ_3, definida como: $\frac{\Gamma(\gamma)}{\Gamma(\gamma_1)\Gamma(\gamma_2)\Gamma(\gamma_3)}(P_{AA})^{\gamma_1-1}(P_{AB})^{\gamma_2-1}(P_{BB})^{\gamma_3-1}$, sendo $\gamma = \sum_{i=1}^{3}\gamma_i$. De (1) tem-se a *priori* conjunta dada por:

$$\pi(p_A, f) = \frac{\Gamma(\gamma)}{\Gamma(\gamma_1)\Gamma(\gamma_2)\Gamma(\gamma_3)}[p_A^2 + p_A(1-p_A)f]^{n_1} \cdot [2p_A(1-p_A)(1-f)]^{n_2-1}[(1-p_A)^2 p_A(1-p_A)f]^{n_3-1}.$$

A *priori* conjunta Beta - função degrau Uniforme é obtida por $\pi(p_A, f) = \pi(p_A)\,\pi(f|p_A)$

A distribuição *a priori* para pA, $\pi(p_A)$, foi condicionada por uma distribuição Beta com hiperparâmetros α e β, e a distribuição condicional *a priori* para f dado p_A, $\pi(f \mid p_A)$, foi determinada por uma função degrau Uniforme sob três intervalos (SHOEMAKER et al., 1998). Os intervalos são: limite inferior do parâmetro f, conforme a expressão (1), até -0,03, representando peso 0,25 (desequilíbrio); de -0,03 até 0,03, com peso 0,50 (equilíbrio) e de 0,03 até 1, com peso 0,25 (desequilíbrio). Dessa forma, a *priori* conjunta é dada por:

$$\pi(p_A, f) = p_A^{\alpha-1}(1-p_A)^{\beta-1}\sum_{i=0}^{2}\alpha_i.1_{A_i}(f).$$

A *priori* conjunta Uniforme - função degrau Uniforme é obtida por: $\pi(p_A, f) = \pi(p_A)\,\pi(f \mid p_A)$, diferenciando da *priori* conjunta Beta - função degrau Uniforme apenas pela distribuição *a priori* para p_A, $\pi(p_A)$, condicionada por uma distribuição Uniforme. Dessa forma, a *priori* conjunta é dada por:

$$\pi(p_A, f) = 1_{(0,1)}(p_A)\sum_{i=0}^{2}\alpha_i 1_{A_i}(f).$$

Considerando a independência entre os parâmetros e a falta de informação *a priori*, optou-se também pela utilização de uma distribuição Uniforme para cada um dos parâmetros. Portanto, a *priori* conjunta é dada por $\pi(p_A, f) = \pi(p_A)\,\pi(f)$, em que: $\pi(p_A) \sim 1_{(0,1)}(pA)$ e $\pi(f) \sim 1_{(max[-pA/(1-pA),\,-(1-pA)/pA],\,1)}(f)$.

As distribuições conjuntas *a posteriori* são dadas, respectivamente, por:

$$\pi(p_A, f \mid n_1, n_2, n_3) \propto [p_A^2 + p_A(1-p_A)f]^{n_1+n_3-1} \cdot [2p_A(1-p_A)(1-f)]^{n_2+n_3-1} \quad (2)$$

$$\pi(p_A, f \mid n_1, n_2, n_3) \propto [p_A^2 + p_A(1-p_A)f]^{n_1}[2p_A(1-p_A)(1-f)]^{n_2} \cdot [(1-p_A)^2 p_A(1-p_A)f]^{n_3} p_A^{\alpha-1}(1-p_A)^{\beta-1}\sum_{i=0}^{2}\alpha_i 1_{A_i}(f) \quad (3)$$

$$\pi(p_A, f \mid n_1, n_2, n_3) \propto [p_A^2 + p_A(1-p_A)f]^{n_1} \cdot [2p_A(1-p_A)(1-f)]^{n_2}[(1-p_A)^2 p_A(1-p_A)f]^{n_3} \cdot 1_{(0,1)}(p_A)\sum_{i=0}^{2}\alpha_i 1_{A_i}(f) \quad (4)$$

$$\pi(p_A, f \mid n_1, n_2, n_3) \propto [p_A^2 + p_A(1-p_A)f]^{n_1} \cdot [2p_A(1-p_A)(1-f)]^{n_2}[(1-p_A)^2 p_A(1-p_A)f]^{n_3} \times \times 1_{(0,1)}(p_A) \times 1_{(\max[-p_A/(1-p_A),\,-(1-p_A)/p_A],\,1)}(f) \quad (5)$$

As distribuições condicionais completas *a posteriori* para p_A, $\pi(p_A \mid f, n_1, n_2, n_3)$, e f, $\pi(f \mid pA, n_1, n_2, n_3)$, apresentam a mesma forma e correspondem à distribuição conjunta *a posteriori* dada em (2) para a *priori* Dirichlet. Para as *prioris* Beta – função degrau

Uniforme e Uniforme – função degrau Uniforme, as distribuições condicionais completas *a posteriori* para p_A são dadas pelas distribuições conjuntas *a posteriori* (3) e (4), respectivamente. Já para f, elas apresentam a mesma forma e são dadas por:

$$\pi(f \mid p_A, n_1, n_2, n_3) \propto [p_A^2 + p_A(1-p_A)f]^{n_1}[2p_A(1-p_A)(1-f)]^{n_2}[(1-p_A)^2 p_A(1-p_A)f]^{n_3} \sum_{i=0}^{2} \alpha_i . 1_{A_i}(f)$$

As distribuições condicionais completas *a posteriori* considerando-se as *prioris* uniformes independentes são dadas por:

$$\pi(p_A \mid f, n_1, n_2, n_3) \propto [p_A^2 + p_A(1-p_A)f]^{n_1} . [2p_A(1-p_A)(1-f)]^{n_2}[(1-p_A)^2 p_A(1-p_A)f]^{n_3} . \quad \text{e} \quad 1_{(0,1)}(p_A)$$

$$\pi(f \mid p_A, n_1, n_2, n_3) \propto [p_A^2 + p_A(1-p_A)f]^{n_1} . [2p_A(1-p_A)(1-f)]^{n_2}[(1-p_A)^2 p_A(1-p_A)f]^{n_3} \times \times 1_{(\max[-p_A/(1-p_A), -(1-p_A)/p_A], 1)}(f)$$

Na implementação do código, para obter uma taxa de aceitação (número de vezes em que o parâmetro foi aceito ao longo das iterações) entre 20 e 50%, sugerida por GILKS et al. (1996), foram considerados os erros das proporções alélicas (ε_{pA}) e do coeficiente de endogamia (εf) conforme procedimento descrito por ARMBORST (2005). Para serem obtidos valores amostrados dos parâmetros, utilizou-se, como função candidata, a distribuição Uniforme no intervalo entre o limite inferior e superior de cada parâmetro. Em relação aos hiperparâmetros das distribuições Beta e Dirichlet, foi utilizado o valor 2, pois, nesse caso, as distribuições cobriam todo o espaço paramétrico das proporções alélicas e genotípicas, respectivamente.

Um estudo de simulação foi realizado no intuito de avaliar a metodologia utilizada e comparar as características proporcionadas por cada uma das *prioris* testadas. A partir do modelo dado em (1) e com base nos estudos de ARMBORST (2005), nove cenários foram abordados, diferindo-se pelo tamanho da amostra (n=50; 200; 1000) e pela intensidade do parâmetro analisado, sendo considerado um valor próximo ao limite inferior do parâmetro (-0,217), um valor positivo próximo do EHW (0,02) e outro com alta endogamia (0,8). O número de alelos (k) foi fixo e igual a 2. Foram simuladas m=100 amostras para cada *posteriori*, sendo obtidas estimativas pontuais e por intervalo em cada um dos nove cenários. Utilizou-se também o conjunto de dados *FBI* e *Cellmark*, descritos em SHOEMAKER et al. (1998), que se referem às proporções genotípicas de três grupos raciais de imigrantes dos Estados Unidos (afro-americanos, caucasianos e hispânicos), localizados em três locos diferentes (D7S8, LDLR e GYPA).

RESULTADOS E DISCUSSÕES

Foi considerado, na análise dos dados, um número fixo de 50.000 iterações, segundo critério de RAFTERY & LEWIS (1992), sendo descartadas as 10.000 iniciais para o período de aquecimento da cadeia e para assegurar a independência da amostra, considerou-se também um espaçamento entre os pontos amostrados de tamanho 40, ou seja, obteve-se uma amostra final de tamanho 1.000. Em relação à convergência, observou-se, pelo critério de Geweke, que o p-valor estimado foi sempre maior que o nível de significância pré-fixado (5%) e que o critério de Gelman e Rubin, considerando duas cadeias com valores iniciais distintos, sempre apresentou valores de $\hat{R}$ próximos a um.

Com os dados simulados sob cada cenário, estimou-se o coeficiente de endogamia, obtendo-se, para cada uma das 100 amostras, a estimativa da média *a posteriori* e a média das 100 médias simuladas. O mesmo procedimento foi adotado para a mediana, a moda, o desvio padrão e o intervalo de credibilidade.

Com base nos resultados obtidos (Tabela 1) para os modelos 1 (*priori* Dirichlet), 2 (*priori* Beta - função degrau Uniforme), 3 (*priori* Uniforme - função degrau Uniforme) e 4 (*prioris* Uniformes), observa-se que: i) todos os modelos resultaram em estimativas da média, mediana e moda *a posteriori* próximas entre si, principalmente para os cenários de n=200 e n=1000, mostrando assim que a função de verossimilhança foi bem definida; ii) para o tamanho de amostra n=50 e coeficientes de endogamia, f = 0,8 e f= -0,217, nota-se que a moda *a posteriori* apresentou as melhores estimativas de$f$, demonstrando a assimetria da distribuição. Para o coeficiente de endogamia f= 0,02 e demais tamanhos de amostra, as estimativas da média, mediana e moda de f foram bem próximas entre si, e, para n=1.000, foram também bem próximas ao valor verdadeiro de f; iii) o processo de simulação propiciou uma melhor análise dos modelos, pois considerou vários cenários possíveis.

Na avaliação da metodologia, pode-se verificar que, em todas as repetições, o valor verdadeiro de f esteve presente nos intervalos de credibilidade. Resultados semelhantes são observados em ARMBORST (2005), em que o modelo com *prioris* Uniformes superestima o valor de f, quando este está muito próximo do limite inferior, possivelmente por considerar a restrição em seus limites.

Tabela 1 - Média ($\hat{f}_1$), mediana ($\hat{f}_2$), moda ($\hat{f}_3$), desvio padrão e HPD, considerando o valor verdadeiro (*f*).

Modelo	*n*	*f*	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	*DP*	*LI*	*LS*
	50		0,707	0,717	0,778	0,117	0,480	0,910
1	200	0,8	0,775	0,779	0,794	0,052	0,675	0,870
	1000		0,815	0,817	0,820	0,022	0,768	0,855
	50		0,749	0,762	0,831	0,115	0,525	0,941
2	200	0,8	0,783	0,787	0,802	0,053	0,673	0,877
	1000		0,818	0,820	0,826	0,023	0,771	0,859
	50		0,742	0,754	0,774	0,116	0,528	0,944
3	200	0,8	0,786	0,789	0,794	0,051	0,685	0,884
	1000		0,818	0,819	0,820	0,023	0,776	0,863
	50		0,742	0,759	0,812	0,121	0,522	0,956
4	200	0,8	0,784	0,788	0,795	0,052	0,680	0,880
	1000		0,818	0,819	0,819	0,022	0,771	0,857
	50		0,063	0,057	-0,005	0,139	-0,180	0,334
1	200	0,02	-0,011	-0,017	-0,029	0,069	-0,139	0,128
	1000		0,030	0,030	0,028	0,032	-0,028	0,093
	50		0,027	0,002	-0,053	0,155	-0,233	0,294
2	200	0,02	-0,021	-0,025	-0,045	0,070	-0,161	0,113
	1000		0,025	0,024	0,033	0,031	-0,027	0,093
	50		0,026	0,008	-0,052	0,141	-0,209	0,279
3	200	0,02	-0,025	-0,026	-0,020	0,066	-0,155	0,104
	1000		0,024	0,022	0,018	0,032	-0,037	0,088
	50		0,018	-0,001	-0,017	0,141	-0,206	0,302
4	200	0,02	-0,022	-0,024	-0,026	0,067	-0,142	0,108
	1000		0,025	0,025	0,028	0,032	-0,037	0,086
	50		-0,058	-0,085	-0,119	0,149	-0,297	0,215
1	200	-0,217	-0,180	-0,187	-0,185	0,053	-0,272	-0,069
	1000		-0,214	-0,216	-0,219	0,019	-0,250	-0,180
	50		-0,042	-0,091	-0,112	0,205	-0,327	0,451
2	200	-0,217	-0,163	-0,178	-0,193	0,068	-0,283	-0,006
	1000		-0,215	-0,217	-0,222	0,021	-0,248	-0,166
	50		-0,037	-0,088	-0,121	0,197	-0,295	0,431
3	200	-0,217	-0,164	-0,178	-0,192	0,066	-0,269	0,001
	1000		-0,213	-0,216	-0,218	0,023	-0,250	-0,171
	50		-0,075	-0,118	-0,147	0,172	-0,326	0,246
4	200	-0,217	-0,182	-0,193	-0,204	0,059	-0,275	-0,046
	1000		-0,217	-0,219	-0,224	0,020	-0,257	-0,180

As estimativas obtidas foram utilizadas na comparação dos modelos por meio do fator de Bayes, o qual se encontra na tabela 2. Percebe-se que o modelo 1 apresentou evidência muito forte em relação a todos os outros. O modelo 3 mostrou evidência fraca em relação ao modelo 2, já o modelo 4 demonstrou evidência fraca para o coeficiente de endogamia f =-0,217, muito fraca para f =0,02 e forte para f =0,8, quando comparado ao modelo 2, e evidência muito fraca, para f =0,8, em relação ao modelo 3. Portanto, o modelo 1 foi considerado mais adequado ao estudo, ou seja, aquele que representa a *priori* mais informativa, seguido pelo modelo 4 (*prioris* Uniformes), para um coeficiente de endogamia de 0,8, e pelo modelo 3 (Uniforme - função degrau Uniforme), para coeficientes de endogamia iguais a 0,02 e -0,217.

Pode-se verificar que, em relação aos outros, os resultados obtidos pelo fator de Bayes do modelo 1 apresentaram valores muito altos e contradizem os valores estimados do parâmetro (Tabela 1), principalmente para tamanhos de amostra n=200 e n=1000. Isso pode ser justificado pelo fato de as estimativas serem obtidas a partir das 100 repetições, enquanto que o fator de Bayes foi aplicado em uma repetição em específico. O mesmo comportamento foi observado para os dados reais. AYRES & BALDING (1998) e ARMBORST (2005) compararam, por meio de suas estimativas, dois métodos clássicos com o modelo

Tabela 2 - Fator de Bayes para dados simulados e para dados reais.

----------Dados simulados----------							
f	*n*	FB_{12}	FB_{13}	FB_{14}	FB_{32}	FB_{42}	FB_{43}
	50	2245	3101	1019	7,24	22,03	3,04
0,8	200	1196	2037	7024	5,87	17,02	2,89
	1000	4889	7690	2516	6,35	19,43	3,05
	50	5066	7614	2254	6,65	2,24	0,29
0,02	200	4112	6343	3030	6,48	1,35	0,47
	1000	6745	1072	6492	6,29	1,03	0,60
	50	1449	2065	3206	7,01	4,52	0,15
-0,217	200	3645	5871	6635	6,20	5,49	0,11
	1000	7695	1239	1639	6,20	4,69	0,13
----------Dados reais----------							
Grupo*	Loco	FB_{12}	FB_{13}	FB_{14}	FB_{32}	FB_{42}	FB_{43}
I		1550	366	295	4,22	5,24	1,24
II		5333	1218	3272	4,37	1,62	0,26
III		7354	1687	7530	4,35	0,10	0,44
IV	D7S8	2272	4860	2264	4,67	1,00	0,46
V		2768	7786	2081	3,55	1,33	0,26
VI		2992	7486	3930	3,99	0,13	0,52
VII		7217	1718	2394	4,20	3,01	0,13
I		1480	3762	2392	3,93	0,16	0,63
II		1105	2963	1278	3,73	0,11	0,43
III		6111	1495	1092	4,08	5,59	1,36
IV	GYPA	9447	2786	5330	3,38	1,77	0,19
V		2520	6478	3579	3,89	0,14	0,55
VI		6289	1520	7196	4,13	0,11	0,47
VII		1263	301	1635	4,19	0,12	0,54
I		2749	480	2161	5,71	1,27	0,44
II		2083	436	680	4,77	3,06	0,15
III		8571	1720	4459	4,98	1,92	0,25
IV	LDLR	3138	880	4060	3,56	0,12	0,46
V		8056	1594	1400	5,05	5,75	1,13
VI		7927	1789	4145	4,42	1,91	0,23
VII		1491	365	1986	4,08	0,13	0,54

I – Afro-americanos (FBI), II – Caucasianos (FBI), III – Hispânicos sudeste (FBI), IV – Hispânicos sudoeste (FBI), V – Afro-americanos (Cellmark), VI – Caucasianos (Cellmark) e VII – Hispânicos (Cellmark).

Bayesiano com *prioris* Uniformes e concluíram que esse modelo apresentou os melhores resultados.

Os resultados da comparação dos modelos por meio do fator de Bayes encontram-se na tabela 2. Em todos os casos, o modelo 1 foi o mais indicado, pois apresentou evidência muito forte a seu favor quando comparado aos demais modelos. Estes, por sua vez, apresentaram evidência muito fraca ou fraca entre si. SHOEMAKER et al. (1998), apesar de não terem utilizado nenhuma forma de comparação em específico das *prioris*, encontraram resultados bastante próximos no que se refere à probabilidade do parâmetro estar ou não em um intervalo de EHW, quando utilizaram as *prioris* Beta - função degrau Uniforme e Uniforme - função degrau Uniforme. Ainda segundo os autores, os resultados encontrados para essas *prioris* foram diferentes quando foi considerada a *priori* Dirichlet.

CONCLUSÕES

A abordagem bayesiana utilizada no estudo da Lei do equilíbrio de Hardy-Weinberg e aplicada aos dados reais teve sua eficiência comprovada pelo estudo de simulação.

O modelo que considera a *priori* Dirichlet, tanto para dados reais, quanto para os dados simulados, foi considerado como mais plausível, por meio do fator de Bayes, quando comparado aos demais.

AGRADECIMENTOS

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pelo financiamento do projeto, e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pela bolsa concedida.

REFERÊNCIAS

ARMBORST, T. **Métodos para medir o desequilíbrio de Hardy-Weinberg através de medidas de endocruzamento**. 2005. 187f. Dissertação (Mestrado em Estatística) - Universidade Federal de Minas Gerais, Belo Horizonte, MG.

AYRES, K.L.; BALDING, D.J. Measuring departures from Hardy-Weinberg: a Markov chain Monte Carlo method for estimating the inbreeding coefficient. **Heredity**, Oxford, v.80, n.6, p.769-777, 1998.

BOX, G.E.P.; TIAO, G.C. **Bayesian inference in statistical analysis**. New York, USA: John Wiley, 1992. 588p.

COCKERHAM, C.C. Variance of gene frequency. **Evolution**, Lawrence, n.1, 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.

GELMAN, A.; RUBIN, D.B. Inference from iterative simulation using multiple sequences. **Statistical Science**, Hayward, v.7, n.4, p.457-511, 1992.

GEWEKE, J. Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In: BERNARDO, J.M. et al. (Ed.). **Bayesian Statistics.** New York, USA: Oxford University, 1992. p.625-631.

GILKS, W.R.et al. **Markov chain Monte Carlo in practice**. London, UK: Chapman & Hall, 1996. 481p.

HASTINGS, W.K. Monte Carlo sampling methods using Markov chains and their applications. **Biometrika**, London, v.57, n.1, p.97-109, 1970.

HEIDELBERGER, P.; WELCH, P. Simulation run lenght control in the presence of an initial transient. **Operations Research**, Landing, v.31, n.6, p.1109-1144, 1983.

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

KASS, R.E.; RAFTERY, A.E. Bayes factors and model uncertainty. **Journal of the American Statistical Association**, Alexandria, v.90, n.430, p.773-795, 1995.

METROPOLIS, N. et al. Equations of state calculations by fast computing machines. **Journal of Chemical Physics**, Chicago, v.21, n.6, p.1087-1092, 1953.

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.

NEI, M.; CHESSER, R.K. Estimation of fixation indices and gene diversities. **Annals of Human Genetics**, London, v.47, n.3, p.253-259, 1983.

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

PAULINO, C.D. et al. **Estatística Bayesiana**. Lisboa, Portugal: Fundação Calouste Gulbenkian, 2003. 429p.

R DEVELOPMENT CORE TEAM. **R: a language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. Disponível em: <http://www.R-project.org>. On line. Acesso em: 2007.

RAFTERY, A.L.; LEWIS, S. How many iterations in the Gibbs sampler? In: BERNARDO, J.M. et al. (Ed.). **Bayesian statistics**. Oxford, USA: University, 1992B. p.763-774.

REIS, R. L. et al. Inferência bayesiana na análise genética de populações diplóides: estimação do coeficiente de endogamia e da taxa de fecundação cruzada. **Ciência Rural**, Santa Maria, v.38, n.5, p.1258-1265, 2008. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-84782008000500009&lng=pt&nrm=iso>. Acesso em: 12 fev. 2009. Doi: 10.1590/S0103-84782008000500009.

ROBERTSON, A.; HILL, W.G. Deviations from Hardy-Weinberg proportions: sampling variances and use in estimation of inbreeding coefficients. **Genetics**, Baltimore, v.107, p.703-718, 1984.

SHOEMAKER, J.S. et al. A Bayesian characterization of Hardy-Weinberg disequilibrium. **Genetics**, Bethesda, v.149, n.4, p.2079-2088, 1998.

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