Ciência Rural, Santa Maria, v.41, n.5, p.834-840, mai, 2011
ISSN 0103-8478

Comparação bayesiana de modelos com uma aplicação para o equilíbrio de Hardy-Weinberg usando o coeficiente de desequilíbrio

Bayesian comparison of models with an application to the Hardy-Weinberg equilibrium using the disequilibrium coefficient

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

RESUMO

O equilíbrio de Hardy-Weinberg é um dos principais assuntos estudados pela Genética de populações. Neste contexto, o presente trabalho aborda a análise e a comparação bayesiana de modelos utilizando o coeficiente de desequilíbrio (D_A). Para isso, realizou-se um estudo de simulação no qual as seguintes distribuições a priori *foram consideradas: Dirichlet (modelo 1); beta - função degrau uniforme (modelo 2); uniforme - função degrau uniforme (modelo 3); e as prioris independentes uniformes (modelo 4). Exemplos de aplicação a dados reais de grupos raciais também são apresentados e discutidos. As amostras das distribuições marginais a posteriori para os parâmetros de interesse foram obtidas mediante o algoritmo Metropolis-Hastings, o qual foi implementado no software livre R. A convergência das cadeias geradas por este algoritmo foi monitorada pelos critérios de Geweke e Gelman & Rubin, os quais estão implementados no pacote BOA do R. Quanto às comparações entre os modelos, efetuadas por meio do fator de Bayes, observa-se que, para os dados simulados, o modelo 4 é o mais indicado para os casos de D_A=0,146, D_A=0,02 e D_A=-0,02 com n=200; o modelo 2 é o mais indicado para D_A=-0,02 e n=50 e o modelo 3 é o mais indicado para D_A=-0,02 e n=1000. Para os dados reais, em cada caso analisado, nota-se uma grande diferenciação na escolha de modelos, em que apenas o modelo 1 não é recomendado.*

Palavras-chave: *fator de Bayes, Genética de populações, simulação de dados.*

ABSTRACT

One of the main subjects studied by population genetics is the Hardy-Weinberg equilibrium. In this context, this paper addresses the analysis and comparison of bayesian models used in its evaluation by the coefficient of disequilibrium. For this, it was carried out a simulation study in which the following prior distributions were considered: Dirichlet (model 1), beta - uniform step function (model 2), uniform - uniform step function (model 3) and independent uniform priors (model 4). Examples of application to real data for racial groups are presented and discussed. Samples from the marginal posterior distributions for parameters of interest were obtained by Metropolis-Hastings algorithm, which was implemented in the software R. The convergence of the chains generated by this algorithm was monitored by criteria of Geweke and Gelman & Rubin, which are implemented in the BOA package R. Regarding comparisons between models, performed using the Bayes factor, it was observed that model 4 is the most suitable for the cases of D_A=0.146, D_A=0.02 and D_A=-0.02 with n=200, the model 2 is the most suitable for D_A=-0.02 with n=50 and the model 3 is the most suitable for D_A=-0.02 and n=1000. For real data, in each case examined, there is a large difference in choice of models, where model 1 is the only one not recommended.

Key words: *Bayes factor, population genetics, simulation data.*

INTRODUÇÃO

Estudos referentes ao modo como os genes estão distribuídos nos indivíduos assumem grande importância para a obtenção de informações úteis ao estabelecimento de estratégias mais seguras na coleta e conservação da variabilidade genética. Em 1908, o matemático inglês Godfrey Harold Hardy e o médico

[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 26.05.10 Aprovado em 19.02.11 Devolvido pelo autor 29.04.11
CR-3632

alemão Wilhelm Weinberg chegaram independentemente e quase que simultaneamente à Lei do Equilíbrio de Hardy-Weinberg. Estes pesquisadores perceberam que, se não existissem fatores evolutivos atuando sobre uma população, as proporções alélicas permaneceriam inalteradas, e as proporções genotípicas atingiriam um equilíbrio estável, mostrando a mesma relação constante entre si ao longo do tempo. Portanto, considerando um gene com dois tipos de alelos, A e B, e definindo as proporções alélicas p_A e p_B=1-p_A, pode-se determinar as proporções genotípicas da população, como: $P_{AA} = p_A^2$ (proporção do genótipo homozigoto AA), P_{AB}=2$p_A p_B$ (proporção do genótipo heterozigoto AB) e $P_{BB} = p_B^2$ (proporção do genótipo homozigoto BB).

Um dos assuntos mais pesquisados na área da Genética de populações refere-se ao estudo das violações à Lei de Hardy-Weinberg. Neste caso, um dos parâmetros mais utilizados na avaliação desse desvio é o coeficiente de desequilíbrio D_A (HERNÁNDEZ & WEIR, 1989), em que este expressa a relação entre as proporções alélicas e o coeficiente de endogamia de uma população. Portanto, o coeficiente de desequilíbrio mede estas discrepâncias entre as proporções genotípicas sob cruzamentos aleatórios e estas sob cruzamentos endogâmicos na população (WEIR, 1996). A endogamia é definida como um sistema em que os acasalamentos se dão entre indivíduos aparentados, ou seja, relacionados pela ascendência, afetando diretamente a diversidade genética da população (MUNIZ et al., 2008; MUNIZ et al., 2010).

As proporções genotípicas homozigotas e heterozigotas para o caso de dois alelos sob a hipótese de violação do modelo de Hardy-Weinberg são definidas por (HERNÁNDEZ & WEIR, 1989):

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

em que os limites de D_A são dados por $\max[-p_A^2, -(1-p_A)^2] \le D_A \le p_A(1-p_A)$, que dependem das proporções alélicas.

Utilizando este modelo, AYRES & BALDING (1998) e ARMBORST (2005) aplicaram o método bayesiano na estimação dos parâmetros e concluíram que este possibilitou a incorporação dos efeitos de incerteza relativa aos parâmetros *nuisance* (parâmetros pelos quais não se tem interesse direto), isto é, as proporções alélicas. Assim, na comparação com os métodos frequentistas, o método bayesiano, utilizando distribuições *a priori* uniformes, apresentou os melhores resultados. SHOEMAKER et al. (1998) descreveram uma metodologia bayesiana para estudar o equilíbrio de Hardy-Weinberg considerando dois parâmetros, o coeficiente de desequilíbrio e o coeficiente de endogamia. Estes autores usaram três distribuições *a priori* para cada parâmetro (Dirichlet, beta - função degrau uniforme e uniforme - função degrau uniforme), mas não consideraram nenhuma forma específica de comparação entre estas.

O trabalho desenvolvido por REIS et al. (2008) utilizou a metodologia bayesiana na estimação do coeficiente de endogamia e da taxa de fecundação cruzada de uma população diploide por meio do modelo aleatório de COCKERHAM (1969) para frequências alélicas. Este trabalho propiciou resultados condizentes, validados pelo estudo de simulação de dados adotado. REIS et al. (2009) descreveram um método bayesiano para estudar o equilíbrio de Hardy-Weinberg através do coeficiente de endogamia. Neste trabalho, os autores analisaram vários modelos e concluíram que o melhor modelo é aquele que utiliza distribuições *a priori* Dirichlet.

Tendo em vista os aspectos apresentados, o objetivo do presente trabalho foi reescrever Hardy-Weinberg por meio do coeficiente de desequilíbrio. Objetivou-se também testar o método por meio de estudos de simulação de dados e aplicá-lo a um conjunto de dados reais.

MATERIAL E MÉTODOS

Atualmente, a inferência bayesiana é um método bastante utilizado em pesquisas genéticas para estimação do parâmetro D_A. Esta é definida através do Teorema de Bayes, que associa um modelo relacionado aos dados (função de verossimilhança) com a distribuição *a priori* dos parâmetros, que são considerados aleatórios e, a partir daí, resume essas informações através da distribuição condicional dos parâmetros sobre os dados observados, a distribuição *a posteriori* (GELMAN et al., 2000). Portanto, a estimação do parâmetro é realizada a partir da distribuição *a posteriori*, e esta informação pode ser resumida pela média, moda, mediana ou pelos intervalos de credibilidade (PAULINO et al., 2003).

Neste trabalho, as distribuições *a priori* utilizadas foram as Dirichlet, com hiperparâmetros inteiros γ_1, γ_2 e γ_3 definidas como $\frac{\Gamma(\gamma)}{\Gamma(\gamma_1)\Gamma(\gamma_2)\Gamma(\gamma_3)}(P_{AA})^{\gamma_1-1}(P_{AB})^{\gamma_2-1}(P_{BB})^{\gamma_3-1}$ em que $\gamma = \sum_{i=1}^{3}\gamma_i$.

De (1), $\pi(p_A, D_A) = \frac{\Gamma(\gamma)}{\Gamma(\gamma_1)\Gamma(\gamma_2)\Gamma(\gamma_3)}[p_A^2 + D_A]^{\gamma_1-1}$

$[2p_A(1-p_A) - 2D_A]^{\gamma_2-1}[(1-p_A)^2 + D_A]^{\gamma_3-1}$

A distribuição *a priori* conjunta beta-função degrau uniforme foi obtida por

$\pi(p_A, D_A) = \pi(p_A)\pi(D_A \mid p_A)$, em que a distribuição *a priori* para p_A, $\pi(p_A)$ foi condicionada por uma distribuição beta com hiperparâmetros α e β, e a distribuição condicional *a priori* para D_A, dado p_A, $\pi(D_A \mid p_A)$ foi determinada por uma função degrau uniforme sob cada um dos intervalos apresentados em SHOEMAKER et al. (1998). Dessa forma, a distribuição *a priori* conjunta é dada por:

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

A distribuição *a priori* conjunta uniforme - função degrau uniforme - é obtida por $\pi(p_A, D_A) = \pi(p_A)\pi(D_A \mid p_A)$ diferenciando da distribuição *a 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 distribuição *a priori* conjunta é dada por:

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

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 distribuição *a priori* conjunta é dada por $\pi(p_A, D_A) = \pi(p_A)\pi(D_A)$, em que $\pi(p_A) \sim U_{(0,1)}$ e $\pi(D_A) \sim U_{(\max[-p_A^2, -(1-p_A)^2], p_A(1-p_A))}$

Para a definição da função de verossimilhança, considere que n_1, n_2 e n_3 representem as quantidades observadas de genótipos *AA*, *AB* e *BB*, respectivamente, em uma amostra de tamanho $n=n_1+n_2+n_3$. Para esses dados, que apresentam uma distribuição multinomial com parâmetros n, p_A e D_A, a função de verossimilhança é dada por $\frac{n!}{n_1!n_2!n_3!}(P_{AA})^{n_1}(P_{AB})^{n_2}(P_{BB})^{n_3}$. De acordo com (1), tem-se: $L(p_A, D_A \mid n_1, n_2, n_3) = \frac{n!}{n_1!n_2!n_3!}[p_A^2 + D_A]^{n_1}$. $[2p_A(1-p_A) - 2D_A]^{n_2}[(1-p_A)^2 + D_A]^{n_3}$

A distribuição conjunta *a posteriori* encontrada deve ser integrada em relação a todos os outros parâmetros que a constituem, obtendo-se, assim, a distribuição marginal a *posteriori* de um parâmetro θ (PAULINO et al, 2003). Esta integração geralmente não é analítica, necessitando de algoritmos iterativos especializados denominados algoritmos MCMC (*Markov Chain Monte Carlo*), dentre os quais, destaca-se o algoritmo de Metropolis-Hastings. Este algoritmo gera um valor de uma distribuição auxiliar ou candidata e este valor é aceito com uma dada probabilidade (METROPOLIS et al., 1953; HASTINGS, 1970), caso contrário é rejeitado e um novo valor é amostrado. Para avaliar a convergência da cadeia de valores gerada pelo Metropolis-Hastings, são utilizados os critérios de GEWEKE (1992) e GELMAN & RUBIN (1992).

Muitas distribuições *a priori* são associadas aos parâmetros ligados ao desequilíbrio de Hardy-Weinberg, e a sua adequabilidade é avaliada por uma das grandes áreas da inferência bayesiana, a análise de sensitividade ou robustez, a qual se caracteriza pela comparação de distribuições *a priori* através de avaliadores de qualidade, como o fator de Baves (KASS & RAFTERY, 1995). Este é definido por: $FB_{(M_i, M_j)} = \frac{\pi(x \mid M_i)}{\pi(x \mid M_j)}$ em que $\pi(x \mid M_i)$ e $\pi(x \mid M_j)$ representam as verossimilhanças marginais de cada modelo. Em algumas situações, as quantidades $\pi(x \mid M_i)$ e $\pi(x \mid M_j)$ podem ser calculadas analiticamente, mas, em geral, os métodos MCMC são usados para obter soluções aproximadas. Uma interpretação para o fator de Bayes (FB) é dada em JEFFREYS (1961), em que valores de $FB_{(M_i, M_j)} < 1$ demonstram evidência a favor de M_j, $1 \le FB_{(M_i, M_j)} < 3,2$ demonstram evidência muito fraca a favor de M_i, $3,2 \le FB_{(M_i, M_j)} < 10$ demonstram evidência fraca a favor de M_i, $10 \le FB_{(M_i, M_j)} < 100$ demonstram evidência forte a favor de M_i e $FB_{(M_i, M_j)} \ge 100$ demonstram evidência muito forte a favor de M_i.

Um estudo de simulação foi realizado no intuito de avaliar o método empregado por cada uma das distribuições *a priori* implementadas. Assim, a partir de (1), vários cenários foram abordados e estes diferiram pelo tamanho da amostra (n=50; 200; 1000) e pela intensidade do parâmetro analisado, em que se considerou um valor próximo ao limite inferior do parâmetro (-0,02), um valor positivo próximo do EHW (0,02) e outro com alta endogamia (0,146) (ARMBORST, 2005), totalizando 9 cenários. Foram simuladas m=100 amostras para cada distribuição *a posteriori*, em que se estimaram os valores pontuais e por intervalo para cada um dos 9 cenários propostos. Utilizaram-se também os dados *FBI* e *Cellmark* retirados do trabalho de 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).

A implementação dos algoritmos computacionais foi realizada no *software* livre R (R DEVELOPMENT CORE TEAM, 2009). Vale ressaltar que: (a) utilizou-se como função candidata a distribuição uniforme no intervalo entre o limite inferior e superior de cada parâmetro; (b) em relação aos hiperparâmetros das distribuições beta e Dirichlet, foi utilizado o valor 2, pois, neste caso, as distribuições cobriam todo o espaço paramétrico das proporções alélicas e

genotípicas, respectivamente; (c) para se alcançar uma taxa de aceitação (número de vezes em que o parâmetro foi aceito ao longo das iterações) entre 20 a 50% (GILKS et al., 1996) foram testados valores adequados para o erro da proporção alélica (ε_{p_A}) e do coeficiente de desequilíbrio (ε_{D_A}) até atingirem valores satisfatórios para a convergência; (d) o procedimento proposto por NOGUEIRA et al. (2004) sugeriu 50000 iterações, em que se descartaram as 10000 iniciais (*burn-in*) e considerou-se um espaçamento de tamanho 40 (*thin*), obtendo uma amostra de tamanho 1000; (e) o critério de Geweke apresentou o valor *p* sempre maior que o nível de significância pré-fixado (5%), e o critério de Gelman & Rubin estimou valores de $\hat{R}$ próximos a 1.

RESULTADOS E DISCUSSÃO

A partir da função de verossimilhança e das distribuições *a priori* adotadas, encontram-se as distribuições conjuntas *a posteriori* para os modelos 1, 2, 3 e 4 dadas, respectivamente, por:

$$\pi(p_A, D_A \mid n_1, n_2, n_3) \propto [p_A^2 + D_A]^{n_1+\gamma_1-1} [2p_A(1-p_A) - 2D_A]^{n_2+\gamma_2-1} [(1-p_A)^2 + D_A]^{n_3+\gamma_3-1} \quad (2)$$

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

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

$$\pi(p_A, D_A \mid n_1, n_2, n_3) \propto [p_A^2 + D_A]^{n_1} [2p_A(1-p_A) - 2D_A]^{n_2} [(1-p_A)^2 + D_A]^{n_3} U_{(0,1)}(p_A) \; x \; U_{(\max[-p_A^2, -(1-p_A)^2], p_A(1-p_A))}(D_A) \quad (5)$$

As distribuições condicionais completas *a posteriori* para p_A, p_A, $\pi(p_A \mid f, n_1, n_2, n_3)$ e D_A, $\pi(D_A \mid p_A, n_1, n_2, n_3)$, apresentam a mesma expressão e correspondem à distribuição conjunta *a posteriori* dada em (2) para a distribuição *a priori* Dirichlet. Para as distribuições *a priori* 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. Para o caso do parâmetro D_A, estas apresentam a mesma forma e são dadas por:

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

As distribuições condicionais completas *a posteriori* para as distribuições *a priori* uniformes independentes são:

$$\pi(p_A \mid D_A, n_1, n_2, n_3) \propto [p_A^2 + D_A]^{n_1} [2p_A(1-p_A) - 2D_A]^{n_2} [(1-p_A)^2 + D_A]^{n_3} U_{(0,1)}(p_A)$$

para D_A, é

$$\pi(D_A \mid p_A, n_1, n_2, n_3) \propto [p_A^2 + D_A]^{n_1} [2p_A(1-p_A) - 2D_A]^{n_2} [(1-p_A)^2 + D_A]^{n_3} U_{(\max[-p_A^2, -(1-p_A)^2], p_A(1-p_A))}(D_A)$$

Os resultados referentes aos valores estimados (média, mediana, moda, desvio padrão e intervalo de credibilidade) para o parâmetro D_A, via simulação de dados, são apresentados na tabela 1. Observa-se que, exceto em alguns cenários para D_A =0,146, o intervalo de credibilidade contém o valor real adotado para o parâmetro. Nota-se também que os modelos apresentaram estimativas (média, mediana e moda) de D_A bastante próximas entre si, principalmente para os cenários de *n*=200 e *n*=1000, evidenciando distribuições do tipo simétricas. Vale ressaltar que estas estimativas estão bastante próximas do valor real, demonstrando que o método utilizado pode ser adaptado às diferentes situações encontradas.

Os resultados de comparação dos modelos, por meio do fator de Bayes, para os dados simulados e reais, são apresentados na tabela 2. Para os dados simulados, o valor real esteve presente em pelo menos 95% das vezes nos intervalos de credibilidade, demonstrando que o processo de simulação foi eficiente e propiciando uma melhor análise dos modelos, pois considerou vários cenários possíveis. Nestas análises observou-se que o modelo 4 é o mais indicado nos casos de D_A =0,146, D_A =0,02 e D_A =-0,02 com *n*=200, obtendo muitas vezes evidências fraca e muito fraca a seu favor, já o modelo 2 é o mais indicado para o caso de D_A=-0,02 com *n*=50, e o modelo 3 é o mais indicado considerando o parâmetro D_A =-0,02 com *n*=1000. Nota-se uma grande diferenciação na escolha de modelos para o parâmetro D_A, em que apenas o modelo 1 não é indicado, o que pode ser explicado, principalmente, pela distribuição *a priori* conjugada utilizada. Para o caso de D_A =-0,02, as diferentes escolhas dos modelos podem ser justificadas pela proximidade desse valor ao equilíbrio de Hardy-Weinberg, ou seja, D_A = 0.

Para os dados reais, observou-se que os modelos 2, 3 e 4, na maioria das vezes, apresentaram evidências muito forte e forte a seu favor quando comparados com o modelo 1 e, portanto, em nenhum dos casos, este modelo pode ser indicado, o que pode ser explicado pela utilização de uma distribuição *a priori* conjugada. Quando comparados entre si, os resultados foram os mais diversos possíveis, especificando um modelo para cada caso analisado, o

Tabela 1 - Média ($\hat{D}_{A_1}$), mediana ($\hat{D}_{A_2}$), moda ($\hat{D}_{A_3}$), desvio padrão (DP) e intervalo de credibilidade inferior (LI) e superior (LS), considerando o valor verdadeiro (D_A).

Modelo	*n*	D_A	$\hat{D}_{A_1}$	$\hat{D}_{A_2}$	$\hat{D}_{A_3}$	*DP*	*LI*	*LS*
	50		0,075	0,076	0,079	0,033	0,010	0,141
1	200	0,146	0,127	0,129	0,133	0,021	0,084	0,164
	1000		0,140	0,141	0,140	0,009	0,118	0,156
	50		0,047	0,043	0,039	0,030	-0,001	0,102
2	200	0,146	0,086	0,087	0,090	0,029	0,027	0,140
	1000		0,127	0,129	0,134	0,013	0,098	0,149
	50		0,042	0,038	0,029	0,028	-0,001	0,094
3	200	0,146	0,079	0,079	0,080	0,029	0,022	0,138
	1000		0,125	0,128	0,134	0,014	0,098	0,148
	50		0,064	0,062	0,056	0,031	0,001	0,118
4	200	0,146	0,122	0,125	0,131	0,023	0,076	0,165
	1000		0,138	0,140	0,141	0,010	0,115	0,157
	50		0,023	0,021	0,017	0,024	-0,020	0,071
1	200	0,02	0,016	0,016	0,016	0,012	-0,007	0,041
	1000		0,021	0,022	0,022	0,005	0,011	0,033
	50		0,016	0,013	0,002	0,024	-0,031	0,064
2	200	0,02	0,015	0,014	0,010	0,012	-0,011	0,037
	1000		0,021	0,021	0,021	0,005	0,010	0,032
	50		0,016	0,014	0,010	0,022	-0,022	0,063
3	200	0,02	0,014	0,014	0,014	0,012	-0,008	0,041
	1000		0,021	0,020	0,020	0,005	0,010	0,034
	50		0,016	0,013	0,008	0,023	-0,028	0,066
4	200	0,02	0,015	0,014	0,009	0,012	-0,008	0,041
	1000		0,021	0,020	0,019	0,005	0,009	0,032
	50		-0,006	-0,009	-0,011	0,022	-0,046	0,041
1	200	-0,02	-0,021	-0,022	-0,021	0,012	-0,046	-0,001
	1000		-0,024	-0,024	-0,024	0,004	-0,032	-0,016
	50		-0,005	-0,009	-0,010	0,025	-0,054	0,049
2	200	-0,02	-0,019	-0,021	-0,022	0,015	-0,047	0,014
	1000		-0,025	-0,024	-0,024	0,004	-0,033	-0,016
	50		-0,004	-0,007	-0,011	0,023	-0,047	0,049
3	200	-0,02	-0,019	-0,021	-0,024	0,014	-0,046	0,012
	1000		-0,025	-0,025	-0,025	0,004	-0,033	-0,017
	50		-0,011	-0,014	-0,016	0,022	-0,056	0,037
4	200	-0,02	-0,022	-0,024	-0,025	0,012	-0,048	0,002
	1000		-0,025	-0,025	-0,025	0,004	-0,033	-0,017

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

D_A	n	FB_{21}	FB_{31}	FB_{32}	FB_{41}	FB_{42}	FB_{43}
				Dados simulados			
	50	265,06	11,67	0,22	612,50	2,31	52,44
0,146	200	0,10	0,25	0,25	334,49	342,47	861,97
	1000	0,48	81,30	3967,42	1084,63	5291,9	13,33
	50	0,36	140,58	520,12	224,08	829,05	1,59
0,02	200	139,75	861,05	6,16	1238,79	8,86	1,43
	1000	20,34	135,47	6,65	439,75	21,60	3,24
	50	1078,19	12,79	0,84	786,33	0,13	614,91
-0,02	200	178,66	175,09	0,10	851,17	4,76	4,86
	1000	228,99	1395,38	6,09	1001,38	4,37	0,13
				Dados reais			
Grupo*	Loco	FB_{21}	FB_{31}	FB_{32}	FB_{41}	FB_{42}	FB_{43}
I		7,17	151,82	21,16	163,24	22,75	1,07
II		39,76	242,01	6,08	247,13	6,21	1,02
III		348,26	8133,29	23,35	344,97	1,00	0,23
IV	D7S8	169,06	54,27	0,31	145,82	0,11	2,68
V		1685,89	408,16	0,41	915,83	0,18	2,24
VI		310,91	96,34	0,32	181,92	0,17	1,88
VII		32,62	211,98	6,49	288,43	8,83	1,36
I		713,97	3533,01	4,94	94,31	0,75	0,37
II		220,87	54,21	0,40	96,18	0,22	1,77
III		63,25	468,25	7,40	502,57	7,94	1,07
IV	GYPA	22,11	205,81	9,30	277,89	12,56	1,35
V		861,39	4234,42	4,91	119,32	0,72	0,35
VI		320,11	106,14	0,30	159,51	0,20	1,50
VII		1,36	1034	756	13,88	10,14	0,74
I		126,87	7075,43	55,76	503,14	3,96	0,14
II		15,86	98,28	6,19	80,63	5,08	0,12
III		15,26	86,14	5,64	127,71	8,36	1,48
IV	LDLR	25,15	7092,91	281,94	433,87	17,24	0,16
V		20,13	13,27	0,15	0,18	0,37	0,25
VI		4,35	222,03	50,96	207,15	47,55	0,10
VII		8,29	69,05	8,32	140,03	16,87	2,02

* I – Afro-americanos (FBI), II – Caucasianos (FBI), III – Hispânicos sudeste (FBI), IV – Hispânicos sudoeste (FBI), V – Afro-americanos (Cellmark), VI – Caucasianos (Cellmark), VII – Hispânicos (Cellmark), n – tamanho da amostra.

que pode ser observado na tabela 2. Resultados semelhantes foram encontrados por SHOEMAKER et al. (1998), quando comparadas as *prioris* Dirichlet, beta - função degrau uniforme e uniforme - função degrau uniforme.

CONCLUSÃO

O modelo 1 (Dirichlet) é o único não indicado tanto para o processo de simulação como para cada grupo dos dados reais. A metodologia bayesiana mostrou-se eficiente no estudo do modelo de Hardy-Weinberg, sendo avaliada e comprovada pelo estudo de simulação, apresentando estimativas bem próximas ao valor real.

REFERÊNCIAS

ARMBORST, T. **Métodos para medir o desequilíbrio de Hardy-Weinberg através de medidas de endocruzamento**. 2005. 187f. Dissertação (Mestrado em Estatística) – Curso de Pós-graduação 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**, v.80, p.769-777, 1998. Disponível em: <http://www.nature.com/hdy/journal/v80/n6/full/6883600a.html>. Acesso em: 14 set. 2009. doi:10.1046/j.1365-2540.1998.00360.x.

COCKERHAM, C.C. Variance of gene frequencies. **Evolution**, v.23, p.72-84, 1969. Disponível em: <http://www.jstor.org/stable/2406485>. Acesso em: 12 jan. 2009.

GELMAN, A.; RUBIN, D.B. Inference from iterative simulation using multiple sequences. **Statistical Science**, v.7, p.457-511, 1992. Disponível em: <http://projecteuclid.org/DPubS?service=UI&version=1.0&verb=Display&handle=euclid.ss/1177011136>. Acesso em: 12 out. 2009. doi: 10.1214/ss/1177011136.

GELMAN, A. et al. **Bayesian data analysis**. USA: Chapman & Hall/CRC, 2000. 526p.

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

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

HASTINGS, W.K. Monte Carlo sampling methods using Markov chains and their applications. **Biometrika**, v.57, p.97-109, 1970. Disponível em: <http://biomet.oxfordjournals.org/content/57/1/97.abstract>. Acesso em: 12 ago. 2009. doi: 10.1093/biomet/57.1.97.

HERNÁNDEZ, J.L.; WEIR, B.S. A disequilibrium approach to Hardy-Weinberg testing. **Biometrics**, v.45, p.53-70, 1989. Disponível em: <http://www.jstor.org/stable/2532034>. Acesso em: 23 out. 2009.

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

KASS, R.E.; RAFTERY, A.E. Bayes factors and model uncertainty. **Journal of the American Statistical Association**, v.90, p.773-795, 1995. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.143.835>. Acesso em: 13 out. 2009. doi: 10.1.1.143.835.

METROPOLIS, N. et al. Equations of state calculations by fast computing machines. **Journal of Chemical Physics**, v.21, p.1087-1092, 1953. Disponível em: <http://m.jcp.aip.org/jcpsa6/v21/i6/p1087_s1>. Acesso em: 11 jun. 2009. doi: 10.1063/1.1699114.

MUNIZ, J.A. et al. Métodos de estimação do coeficiente de endogamia em uma população diplóide com alelos múltiplos. **Ciência e Agrotecnologia**, v.32, p.93-102, 2008. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-70542008000100014>. Acesso em: 23 jul. 2009. doi: 10.1590/S1413-70542008000100014.

MUNIZ, J.A. et al. Comparação entre métodos de estimação do coeficiente de endogamia com dados de frequências alélicas em uma população diplóide. **Ciência e Agrotecnologia**, v.34, p.43-54, 2010. Disponível em: <http://www.scielo.br/scielo.php?pid=S1413-70542010000100005&script=sci_arttext >. Acesso em: 15 out. 2009. doi: 10.1590/S1413-70542010000100005.

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

PAULINO, C.D. et al. **Estatística Bayesiana**. Lisboa: Fundação Calouste Gulbenkian, 2003. 446p.

R Development Core Team. **R: a language and environment for statistical computing**. Vienna. Disponível em: <http://www.R-project.org>. Acesso em: 12 mar. 2009.

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**, v.38, n.5, p.1258-1265, 2008. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-84782008000500009&lng=es&nrm=iso&tlng=es>. Acesso em: 18 out. 2009. doi: 10.1590/S0103-84782008000500009.

REIS, R.L. et al. Abordagem bayesiana da sensitividade de modelos para o coeficiente de endogamia. **Ciência Rural**, v.39, n.6, p.1752-1759, 2009. Disponível em: <http://www.scielo.br/scielo.php?pid=S0103-84782009000600018&script=sci_arttext>. Acesso em: 17 dez. 2009. doi: 10.1590/S0103-84782009000600018.

SHOEMAKER, J.S. et al. A Bayesian characterization of Hardy-Weinberg disequilibrium. **Genetics**, v.149, n.4, p.2079-2088, 1998. Disponível em: <http://www.genetics.org/cgi/content/full/149/4/2079>. Acesso em: 24 jun. 2009.

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