

**COMPARAÇÃO BAYESIANA DE MODELOS  
PARA O DESEQUILÍBRIO DE HARDY-  
WEINBERG**

**RICARDO LUIS DOS REIS**

**2008**

**RICARDO LUIS DOS REIS**

**COMPARAÇÃO BAYESIANA DE MODELOS PARA O  
DESEQUILÍBRIO DE HARDY-WEINBERG**

Dissertação apresentada à Universidade Federal de Lavras como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, para a obtenção do título de "Mestre".

Orientador

Prof. Dr. Joel Augusto Muniz

LAVRAS  
MINAS GERAIS - BRASIL  
2008

**Ficha Catalográfica Preparada pela Divisão de Processos Técnicos da  
Biblioteca Central da UFLA**

Reis, Ricardo Luis dos.

Comparação Bayesiana de modelos para o desequilíbrio de Hardy-Weinberg /  
Ricardo Luis dos Reis. – Lavras : UFLA, 2008.

86 p. : il.

Dissertação (Mestrado) – Universidade Federal de Lavras, 2008.

Orientador: Joel Augusto Muniz.

Bibliografia.

1. Fator de Bayes. 2. Genética de populações. 3. Inferência Bayesiana.  
4. Coeficiente de endogamia. 5. Coeficiente de desequilíbrio. I. Universidade  
Federal de Lavras. II. Título.

CDD-519.542

**RICARDO LUIS DOS REIS**

**COMPARAÇÃO BAYESIANA DE MODELOS PARA O  
DESEQUILÍBRIO DE HARDY-WEINBERG**

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, para obtenção do título de "Mestre".

APROVADA em 1º de fevereiro de 2008.

Profa. Dra. Thelma Sáfadi UFLA

Prof. Dr. Fabyano Fonseca e Silva UFV

Prof. Dr. Renato Ribeiro de Lima UFLA

Prof. Dr. Joel Augusto Muniz

UFLA

(Orientador)

LAVRAS

MINAS GERAIS - BRASIL

2008

*A Deus e a Santo Expedito.  
Aos meus pais, Luzia e Pedro.  
A minha irmã, Dinha.*

*“Eu me lembro  
o dia em que aqui cheguei  
foi tudo como eu sonhei  
meu sonho eu realizei...”*

*Trecho da música Lembranças, de Elias dos Santos.*

## AGRADECIMENTOS

Agradeço a DEUS e a Santo Expedito, por sempre guiarem meus passos aonde quer que eu vá. A meus pais, Luzia e Pedro, e a minha irmã, Dinha, pelo carinho, amor, alegria e incentivo em todos os momentos da minha vida. Que Deus abençoe vocês!

Ao professor Joel, pela orientação, confiança, amizade, ensinamentos e, sobretudo, pelo exemplo.

À professora Thelma, pela co-orientação, seriedade, ética, paciência, incentivo e preciosíssimas sugestões, que enriqueceram significativamente a qualidade deste trabalho, além do grande convívio. Seus ensinamentos foram de extrema importância. Obrigado pelo exemplo, pela amizade e pelo compromisso com a pesquisa.

Ao professor Fabyano Fonseca e Silva, pela amizade, ensinamentos, exemplo. Obrigado por tudo! Ao professor Luiz Henrique de Aquino (Caveira), por todos os ensinamentos e, principalmente, pela grande amizade. As reuniões, em sua casa, são inesquecíveis. Obrigado por tudo!

Gostaria de agradecer também aos professores Delly e Marcelo Cirillo, por me ajudarem no início de minhas atividades na Estatística. Aos colegas e amigos do mestrado e do doutorado, do DEX e de outros departamentos, pela amizade, pelas festas e os grandes momentos de risadas; em especial, a Natascha e a Luciana Lanchote e aos eternos amigos da graduação.

Aos professores e funcionários do Departamento de Ciências Exatas da UFLA e ao Programa de Pós-Graduação em Estatística e Experimentação Agropecuária.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pelo apoio com bolsas de estudo. A todos que, de forma direta ou indireta, me auxiliaram neste trabalho.

## SUMÁRIO

	Página
LISTA DE TABELAS.....	i
LISTA DE QUADROS.....	ii
LISTA DE FIGURAS.....	iii
RESUMO.....	iv
ABSTRACT.....	v
1 INTRODUÇÃO .....	1
2 REFERENCIAL TEÓRICO.....	4
2.1 Inferência Bayesiana.....	4
2.1.1 Distribuições <i>a priori</i> .....	5
2.1.1.1 <i>Prioris</i> informativas.....	6
2.1.1.2 <i>Prioris</i> não informativas.....	7
2.1.2 Função de Verossimilhança.....	7
2.1.3 Distribuições <i>a posteriori</i> .....	8
2.1.3.1 Intervalos de credibilidade e HPD.....	8
2.1.3.2 Teste de hipótese Bayesiano.....	9
2.1.4 Métodos computacionais.....	9
2.1.4.1 Métodos de Monte Carlo via cadeias de Markov.....	9
2.1.4.2 Algoritmo de Metropolis-Hastings.....	10
2.1.5 Método de seleção de modelos (fator de Bayes).....	11
2.1.6 Testes para avaliação da convergência.....	12
2.1.6.1 Teste de Geweke .....	13
2.1.6.2 Teste de Gelman & Rubin .....	13
2.1.6.3 Teste de Raftery & Lewis .....	14
2.1.6.4 Teste de Heidelberger & Welch .....	14
2.1.6.5 Autocorrelação.....	15
2.1.6.6 Verificação da convergência.....	15
2.2 Conceitos genéticos.....	16
2.2.1 Conceitos básicos.....	16
2.2.2 Fatores evolutivos.....	17
2.2.3 Equilíbrio de Hardy-Weinberg.....	17
2.2.4 Violações ao modelo de Hardy-Weinberg.....	18
2.2.4.1 Coeficiente de endogamia.....	19
2.2.4.2 Coeficiente de desequilíbrio.....	21
2.3 Abordagem clássica em genética de populações.....	22
2.4 Abordagem Bayesiana em genética de populações.....	24
3 MATERIAL E MÉTODOS .....	27
3.1 Função de verossimilhança .....	27
3.2 Distribuições <i>a priori</i> .....	28
3.2.1 <i>Priori</i> Dirichlet (Modelo 1) .....	28

3.2.2 <i>Priori</i> Beta - função degrau Uniforme (Modelo 2) e <i>priori</i> Uniforme - função degrau Uniforme (Modelo 3).....	28
3.2.3 <i>Prioris</i> Uniformes (Modelo 4).....	31
3.3 Distribuições conjuntas <i>a posteriori</i> .....	31
3.3.1 Distribuição conjunta <i>a posteriori</i> (Modelo 1) .....	31
3.3.2 Distribuição conjunta <i>a posteriori</i> (Modelo 2).....	32
3.3.3 Distribuição conjunta <i>a posteriori</i> (Modelo 3).....	32
3.3.4 Distribuição conjunta <i>a posteriori</i> (Modelo 4).....	33
3.4 Distribuições condicionais completas <i>a posteriori</i> .....	33
3.4.1 Distribuições condicionais completas <i>a posteriori</i> (Modelo 1) .....	33
3.4.2 Distribuições condicionais completas <i>a posteriori</i> (Modelo 2).....	34
3.4.3 Distribuições condicionais completas <i>a posteriori</i> (Modelo 3).....	35
3.4.4 Distribuições condicionais completas <i>a posteriori</i> (Modelo 4).....	35
3.5 Implementação do código.....	36
3.6 Análise dos dados simulados .....	38
3.6.1 Avaliação da acurácia.....	39
3.7 Análise dos dados reais.....	40
4 RESULTADOS E DISCUSSÃO .....	42
4.1 Dados simulados .....	42
4.1.1 Considerações gerais.....	42
4.1.2 Parâmetro $f$ .....	43
4.1.2.1 Taxas de aceitação.....	47
4.1.2.2 Fator de Bayes.....	47
4.1.2.3 Avaliação da acurácia.....	48
4.1.3 Parâmetro $D_A$ .....	51
4.1.3.1 Taxas de aceitação.....	55
4.1.3.2 Fator de Bayes.....	56
4.1.3.3 Avaliação da acurácia.....	57
4.2 Dados reais.....	60
4.2.1 Considerações gerais.....	60
4.2.2 Parâmetro $f$ .....	61
4.2.2.1 Taxas de aceitação.....	67
4.2.2.2 Fator de Bayes.....	68
4.2.3 Parâmetro $D_A$ .....	69
4.2.3.1 Taxas de aceitação.....	75
4.2.3.2 Fator de Bayes.....	76
4.3 Considerações finais.....	77
5 CONCLUSÕES .....	78
6 REFERÊNCIAS BIBLIOGRÁFICAS.....	79
ANEXO .....	85



## LISTA DE TABELAS

<b>TABELA 1:</b> Interpretação do fator de Bayes, segundo Jeffreys (1961).....	12
<b>TABELA 2:</b> Intervalos sugeridos pela <i>National Research Council</i> (1996).....	29
<b>TABELA 3:</b> Proporções genotípicas para o loco D7S8.....	41
<b>TABELA 4:</b> Proporções genotípicas para o loco GYPA.....	41
<b>TABELA 5:</b> Proporções genotípicas para o loco LDLR.....	41
<b>TABELA 6:</b> Média ( $\hat{f}_1$ ), mediana ( $\hat{f}_2$ ), moda ( $\hat{f}_3$ ), desvio padrão e HPD, considerando o valor verdadeiro $f = 0,8$ .....	43
<b>TABELA 7:</b> Média ( $\hat{f}_1$ ), mediana ( $\hat{f}_2$ ), moda ( $\hat{f}_3$ ), desvio padrão e HPD, considerando o valor verdadeiro $f = 0,02$ .....	43
<b>TABELA 8:</b> Média ( $\hat{f}_1$ ), mediana ( $\hat{f}_2$ ), moda ( $\hat{f}_3$ ), desvio padrão e HPD, considerando o valor verdadeiro $f = -0,217$ .....	44
<b>TABELA 9:</b> Erros das proporções alélicas e do coeficiente de endogamia.....	47
<b>TABELA 10:</b> Fator de Bayes para o parâmetro $f$ .....	48
<b>TABELA 11:</b> Avaliação da acurácia, sendo $\hat{f}$ a média das médias.....	49
<b>TABELA 12:</b> Média ( $\hat{D}_A$ ), mediana ( $\hat{D}_B$ ), moda ( $\hat{D}_C$ ), desvio padrão e HPD, considerando o valor verdadeiro $D_A = 0,146$ .....	52
<b>TABELA 13:</b> Média ( $\hat{D}_A$ ), mediana ( $\hat{D}_B$ ), moda ( $\hat{D}_C$ ), desvio padrão e HPD, considerando o valor verdadeiro $D_A = 0,02$ .....	52
<b>TABELA 14:</b> Média ( $\hat{D}_A$ ), mediana ( $\hat{D}_B$ ), moda ( $\hat{D}_C$ ), desvio padrão e HPD, considerando o valor verdadeiro $D_A = -0,02$ .....	53
<b>TABELA 15:</b> Erros das proporções alélicas e do coeficiente de desequilíbrio.....	56
<b>TABELA 16:</b> Fator de Bayes para o parâmetro $D_A$ .....	57
<b>TABELA 17:</b> Avaliação da acurácia com a média das médias de $\hat{D}_A$ .....	58
<b>TABELA 18:</b> Média ( $\hat{f}_1$ ), mediana ( $\hat{f}_2$ ), moda ( $\hat{f}_3$ ), desvio padrão e HPD, considerando o loco D7S8.....	62
<b>TABELA 19:</b> Média ( $\hat{f}_1$ ), mediana ( $\hat{f}_2$ ), moda ( $\hat{f}_3$ ), desvio padrão e HPD, considerando o loco GYPA.....	63
<b>TABELA 20:</b> Média ( $\hat{f}_1$ ), mediana ( $\hat{f}_2$ ), moda ( $\hat{f}_3$ ), desvio padrão e HPD, considerando o loco LDLR.....	64
<b>TABELA 21:</b> Erros das proporções alélicas e do coeficiente de endogamia.....	67

<b>TABELA 22:</b> Fator de Bayes para o parâmetro $f$ .....	68
<b>TABELA 23:</b> Média ( $\hat{D}_{A_1}$ ), mediana ( $\hat{D}_{A_2}$ ), moda ( $\hat{D}_{A_3}$ ), desvio padrão e HPD, considerando o loco D7S8.....	70
<b>TABELA 24:</b> Média ( $\hat{D}_{A_1}$ ), mediana ( $\hat{D}_{A_2}$ ), moda ( $\hat{D}_{A_3}$ ), desvio padrão e HPD, considerando o loco GYP A.....	71
<b>TABELA 25:</b> Média ( $\hat{D}_{A_1}$ ), mediana ( $\hat{D}_{A_2}$ ), moda ( $\hat{D}_{A_3}$ ), desvio padrão e HPD, considerando o loco LDLR.....	72
<b>TABELA 26:</b> Erros das proporções alélicas e do coeficiente de desequilíbrio.	75
<b>TABELA 27:</b> Fator de Bayes para o parâmetro $D_A$ .....	76

## LISTA DE QUADROS

<b>QUADRO 1:</b> Algoritmo de Metropolis-Hastings.....	10
<b>QUADRO 2:</b> Simulação de dados.....	38

## LISTA DE FIGURAS

<b>FIGURA 1:</b> Função degrau com 5 intervalos ( $A_0, A_1, A_2, A_3, A_4$ ) .....	6
<b>FIGURA 2:</b> Gráfico da distribuição marginal <i>a posteriori</i> de $f$ ( $n=50$ ).....	45
<b>FIGURA 3:</b> Gráfico da distribuição marginal <i>a posteriori</i> de $f$ ( $n=200$ ).....	46
<b>FIGURA 4:</b> Gráfico da distribuição marginal <i>a posteriori</i> de $f$ ( $n=1000$ ).....	46
<b>FIGURA 5:</b> Gráfico de dispersão para o caso de $f = 0,8$ .....	50
<b>FIGURA 6:</b> Gráfico de dispersão para o caso de $f = 0,02$ .....	51
<b>FIGURA 7:</b> Gráfico de dispersão para o caso de $f = -0,217$ .....	51
<b>FIGURA 8:</b> Gráfico da distribuição marginal <i>a posteriori</i> de $D_A$ ( $n=50$ ).....	54
<b>FIGURA 9:</b> Gráfico da distribuição marginal <i>a posteriori</i> de $D_A$ ( $n=200$ ).....	54
<b>FIGURA 10:</b> Gráfico da distribuição marginal <i>a posteriori</i> de $D_A$ ( $n=1000$ )...55	
<b>FIGURA 11:</b> Gráfico de dispersão para o caso de $D_A = 0,146$ .....	59
<b>FIGURA 12:</b> Gráfico de dispersão para o caso de $D_A = 0,02$ .....	60
<b>FIGURA 13:</b> Gráfico de dispersão para o caso de $D_A = -0,02$ .....	60
<b>FIGURA 14:</b> Gráfico da distribuição marginal <i>a posteriori</i> de $f$ para o grupo afro-americano do loco D7S8 ( <i>FBI</i> ).....	66
<b>FIGURA 15:</b> Gráfico da distribuição marginal <i>a posteriori</i> de $f$ para o grupo hispânico do sudeste no loco GYPA ( <i>FBI</i> ).....	66
<b>FIGURA 16:</b> Gráfico da distribuição marginal <i>a posteriori</i> de $f$ para o grupo afro-americano no loco LDLR ( <i>Cellmark</i> ).....	67
<b>FIGURA 17:</b> Gráfico da distribuição marginal <i>a posteriori</i> de $D_A$ para o grupo afro-americano no loco D7S8 ( <i>FBI</i> ).....	74
<b>FIGURA 18:</b> Gráfico da distribuição marginal <i>a posteriori</i> de $D_A$ para o grupo hispânico do sudeste no loco GYPA ( <i>FBI</i> ).....	74
<b>FIGURA 19:</b> Gráfico da distribuição marginal <i>a posteriori</i> de $D_A$ para o grupo afro-americano no loco LDLR ( <i>Cellmark</i> ).....	75

## RESUMO

REIS, Ricardo Luis dos. **Comparação Bayesiana de modelos para o desequilíbrio de Hardy-Weinberg**. 2008. 86 p. Dissertação (Mestrado em Estatística e Experimentação Agropecuária) - Universidade Federal de Lavras, Lavras, MG.\*

Dentre os diversos aspectos geralmente observados na caracterização genética de populações naturais, a avaliação do grau de estruturação da variabilidade genética assume grande importância. Neste trabalho, considerou-se uma análise Bayesiana de modelos para o desequilíbrio de Hardy-Weinberg, usando a comparação de *prioris* e os parâmetros coeficiente de endogamia e coeficiente de desequilíbrio. A metodologia utilizada foi testada mediante um estudo de simulação usando as *prioris* 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 são discutidos. Implementou-se 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 análise gráfica e pelos critérios de Geweke e Gelman & Rubin, que estão implementados no pacote BOA do software livre R. As comparações entre os modelos, realizadas por meio do fator de Bayes, indicaram superioridade do modelo 1 em relação aos demais para o parâmetro  $f$  e inferioridade do modelo 1 para o parâmetro  $D_A$ . Em virtude dos resultados apresentados, pode-se atestar que a abordagem Bayesiana apresentou bons resultados na comparação dos modelos, demonstrando a aplicabilidade das metodologias propostas e o elevado potencial de aplicação da estatística Bayesiana em estudos de genética de populações.

**Palavras Chaves:** inferência Bayesiana, fator de Bayes, genética de populações, coeficiente de endogamia e desequilíbrio.

---

\* Comitê Orientador: Prof. Joel Augusto Muniz - UFLA (orientador), Profa. Thelma Sáfydi - UFLA e Prof. Luiz Henrique de Aquino - UFLA.

## ABSTRACT

REIS, Ricardo Luis dos. **Bayesian comparison of models for the Hardy-Weinberg disequilibrium.** 2008. 86 p. Dissertation (Master Program in Statistics and Agricultural Experimentation) - Federal University of Lavras, Lavras, Minas Gerais, Brazil.\*

Among the various aspects generally considered in the genetic characterization of natural populations, the evaluation of the degree of genetic structure are of great importance. In these work, considered a Bayesian analysis of models for the Hardy-Weinberg disequilibrium using the comparison of prioris and the disequilibrium and inbreeding coefficient parameters. The methodology was tested by a simulation study using the prioris Dirichlet (model 1), Beta - step function Uniform (model 2), Uniform - step function Uniform (model 3) and the independents prioris Uniforms (model 4). Examples of application using real data of breed groups are presented. A Metropolis-Hastings algorithm was implemented in the free software R to get sampling of the posterior conditionals distributions of the models parameters. The convergence of the chains was monitored through graphic analysis and for the criteria of Geweke and Gelman & Rubin, implemented in the BOA package of the free software R. These comparisons were realized by Bayes factor showing the model 1 superiority for the  $f$  parameter and a model 1 inferiority for the  $D_A$  parameter. Because of the presented results, it can be attested that the Bayesian approach presented good results in the comparison of the models, illustrating the applicability of the proposed methods and reveal the great potential of use of Bayesian statistics in population genetic studies.

**Keywords:** Bayesian inference, Bayes factor, population genetic, disequilibrium and inbreeding coefficient.

---

\*Guidance Committee: Prof. Joel Augusto Muniz - UFLA (adviser), Profa. Thelma Sáfadi - UFLA and Prof. Luiz Henrique de Aquino - UFLA.

## 1 INTRODUÇÃO

As informações obtidas pelos estudos sobre a caracterização da estrutura genética existente nas mais 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. Esta é uma das principais áreas de estudo da genética de populações. Esse campo da genética estuda as proporções alélicas e genotípicas e os fatores capazes de alterá-las, destacando-se dentre estes aspectos evolutivos, como a seleção, a migração e a mutação (Falconer, 1989). Neste contexto, população é definida como um conjunto de indivíduos da mesma espécie, que ocupam o mesmo local, apresentam continuidade no tempo e possuem a capacidade de se intercasalar ao acaso, portanto, trocar alelos entre si (Gardner, 1977).

Em 1908, Godfrey Harold Hardy e Wilhelm Weinberg demonstraram, independentemente, o princípio das proporções alélicas em uma população, que ficou conhecido como Lei de Hardy-Weinberg ou Lei do equilíbrio de Hardy-Weinberg. Esta lei relata que, na ocorrência de cruzamentos ao acaso e na ausência de fatores evolutivos, as proporções alélicas e genotípicas permanecem constantes, de geração para geração.

O estudo das violações à Lei de Hardy-Weinberg é um dos principais assuntos pesquisados. 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. Outro parâmetro relacionado é o coeficiente de desequilíbrio  $D_A$  (Hernández & Weir, 1989), que expressa a relação entre as proporções alélicas e o coeficiente de endogamia de uma população. Esses parâmetros têm a vantagem de serem de

fácil interpretação biológica, porém, em alguns casos, podem apresentar tratamento matemático complexo. Uma das vantagens relacionadas a esse desequilíbrio é a capacidade de adaptação do indivíduo a um determinado meio, mas, por outro lado, esta característica pode envolver algum tipo de doença.

Alguns métodos usados na estimação desses parâmetros são encontrados nos trabalhos clássicos de Emigh (1980), Hernández & Weir (1989) e Weir (1996) e nos trabalhos Bayesianos de Pereira & Rogatko (1984), Lindley (1988), Chow & Fong (1992), Ayres & Balding (1998) e Shoemaker et al. (1998).

Atualmente, a abordagem Bayesiana vem sendo utilizada com sucesso, em várias áreas da ciência. A inferência Bayesiana é o processo de encontrar um modelo de probabilidade para um conjunto de dados e resumir o resultado por uma distribuição de probabilidade sobre os parâmetros do modelo (Gelman et al., 2000). Ou seja, ela 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 por meio da distribuição condicional dos parâmetros sobre os dados observados, a distribuição *a posteriori*. Em estudos de genética de populações é atraente a idéia de se associar a um determinado parâmetro uma distribuição de probabilidade e não um único valor fixo, pois os elementos sob caracterização, as populações, estão sujeitos a uma série de fatores estocásticos (Coelho, 2002). Muitas distribuições *a priori* são associadas aos parâmetros ligados ao desequilíbrio de Hardy-Weinberg, mas quais destas seriam mais adequadas para o problema em questão? Essa pergunta pode ser respondida por uma das grandes áreas da inferência Bayesiana, a análise de sensibilidade ou robustez, a qual se caracteriza pela comparação de *prioris* por meio de avaliadores de qualidade.

Este trabalho foi realizado com o objetivo de utilizar a abordagem Bayesiana para o problema do desequilíbrio de Hardy-Weinberg, considerando a comparação de *prioris* como foco principal. Pretendem-se também testar a

metodologia por meio da simulação de dados, tendo sido estudados nove cenários que diferiram pelo tamanho da amostra e pela intensidade do parâmetro, sendo, posteriormente, aplicada a um conjunto de dados reais de grupos raciais de imigrantes.

A proposta deste trabalho foi realizar uma extensão dos trabalhos de Ayres & Balding (1998) e Armbrorst (2005), utilizando apenas modelos Bayesianos e também realizando as análises para o parâmetro coeficiente de desequilíbrio e do trabalho de Shoemaker et al. (1998), realizando um processo de simulação antes da aplicação nos dados reais e fazendo a comparação entre os modelos por métodos mais específicos.



## 2 REFERENCIAL TEÓRICO

### 2.1 Inferência Bayesiana

Fazer inferências é uma das principais finalidades da estatística. Na abordagem clássica, os parâmetros desconhecidos são considerados fixos e toda a análise se restringe àquelas informações contidas na amostra dos dados. Segundo Paulino et al. (2003), esta abordagem foi adotada, de forma quase unânime, pelos estatísticos durante a primeira metade do século XX. No entanto, uma abordagem alternativa ressurgiu devido aos avanços computacionais, a inferência Bayesiana, que considera úteis todas as informações disponíveis para reduzir as incertezas na análise.

Portanto, a inferência Bayesiana trata o parâmetro desconhecido como quantidade aleatória e, conseqüentemente, permite incorporar algum conhecimento sobre esse, assumindo, assim, distribuições de probabilidade (*prioris*), antes que os dados tenham sido coletados (Box & Tiao, 1992). Este conhecimento pode ser obtido por meio de análises anteriores, experiência do pesquisador na área em questão ou publicações sobre o assunto que se deseja pesquisar ou estudar.

De maneira bem geral, a distribuição *a priori* é o único fator que diferencia a estatística clássica da Bayesiana. Portanto, é alvo de crítica para muitos dos estatísticos clássicos, que alegam que a situação *a priori* é um processo subjetivo, já que dois pesquisadores podem ter diferentes graus de incertezas sobre uma quantidade desconhecida (Paulino et al., 2003).

Por meio do Teorema de Bayes, essa distribuição *a priori* é combinada com a informação contida nos dados amostrais (função de verossimilhança), induzindo a uma distribuição *a posteriori*. Portanto, toda inferência é realizada usando-se a distribuição *a posteriori* e esta informação pode ser resumida pela

média, moda, mediana ou pelos intervalos de credibilidade (Paulino et al., 2003). Dessa forma, a inferência Bayesiana oferece resultados mais completos, pois, além da verossimilhança, utiliza também a informação *a priori* relativa aos parâmetros.

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  $\theta$  é contínuo, o Teorema é dado por:

$$\pi(\theta|x) = \frac{L(\theta|x)\pi(\theta)}{\int L(\theta|x)\pi(\theta)d\theta} \quad (1)$$

e, no caso em que  $\theta$  é discreto, tem-se:

$$\pi(\theta|x) = \frac{L(\theta|x)\pi(\theta)}{\sum L(\theta|x)\pi(\theta)}. \quad (2)$$

Como nas expressões (1) e (2), o denominador independe do parâmetro  $\theta$ , este pode ser considerado como uma constante e, então, estas expressões podem ser representadas por:

$$\pi(\theta|x) \propto L(\theta|x)\pi(\theta), \quad (3)$$

ou seja, a expressão (3) pode ser entendida como:

distribuição *a posteriori*  $\propto$  verossimilhança x distribuição *a priori*.

### 2.1.1 Distribuições *a priori*

A distribuição *a priori* representa, probabilisticamente, o conhecimento a respeito do parâmetro  $\theta$ , antes da obtenção dos dados (Leandro, 2001). Esta é especificada por meio de experiências ou crenças do pesquisador, sendo, muitas vezes, baseada simplesmente na expectativa subjetiva e, portanto, diferindo de

pesquisador para pesquisador. O problema é como convertê-la na forma de uma distribuição de probabilidade (Smith, 1991).

### 2.1.1.1 *Prioris informativas*

A partir do conhecimento prévio sobre o parâmetro  $\theta$  pode-se descrever, aproximadamente, a distribuição de probabilidade que melhor o representa. Uma forma de expressar esse conhecimento é utilizando informações fornecidas por pesquisas realizadas sobre um determinado assunto. Por meio delas, Shoemaker et al. (1998) representaram seus resultados na forma de uma função degrau, considerada como *priori* nesse trabalho envolvendo o desequilíbrio de Hardy-Weinberg. Uma função de números reais é chamada de função degrau se pode ser escrita como uma combinação linear finita de funções indicadoras de certos intervalos (Shoemaker et al., 1998). Uma função degrau com 5 intervalos é ilustrada na Figura 1.

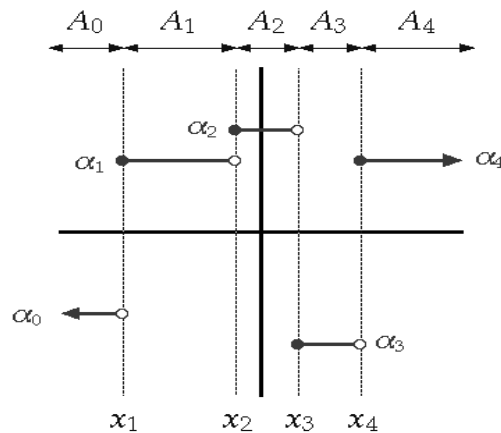


FIGURA 1: Função degrau com 5 intervalos ( $A_0, A_1, A_2, A_3, A_4$ ).

Na Figura 1, a seqüência de coeficientes  $\alpha_0, \dots, \alpha_4$  representa os pesos associados a cada intervalo, os valores  $x_1, \dots, x_4$  correspondem aos limites dos intervalos  $(A_0, A_1, A_2, A_3, A_4)$  e os intervalos estão definidos por  $A_0 = (-\infty, x_1)$ ,  $A_1 = [x_1, x_2)$ ,  $A_2 = [x_2, x_3)$ ,  $A_3 = [x_3, x_4)$  e  $A_4 = [x_4, \infty)$ . Portanto, uma função degrau com  $n$  intervalos pode ser escrita como:

$$f(x) = \sum_{i=0}^{n-1} \alpha_i \cdot 1_{A_i}(x), \text{ onde } 1_A(x) = \begin{cases} 1, & \text{se } x \in A \\ 0, & \text{se } x \notin A \end{cases}.$$

*Prioris* conjugadas ocorrem quando as distribuições *a priori* e *a posteriori* pertencem à mesma classe de distribuições e, portanto, a atualização do conhecimento que se tem sobre  $\theta$  envolve apenas uma mudança nos hiperparâmetros (Ehlers, 2007).

### 2.1.1.2 *Prioris* não informativas

A utilização de uma distribuição *a priori* não informativa implica que a informação contida nos dados é dominante, no sentido de que o conhecimento *a priori* seja vago (Gelman et al., 2000). Nestes casos, pode-se considerar uma distribuição *a priori* Uniforme ( $\pi(\theta) \propto k$ ), pois os possíveis valores de  $\theta$  são igualmente prováveis ou *a priori* de Jeffreys (Jeffreys, 1961).

### 2.1.2 Função de verossimilhança

A função de verossimilhança pode ser considerada como a representação da informação de  $\theta$  obtida dos dados (Box & Tiao, 1992). É por meio dela que os dados podem modificar o conhecimento que se tem *a priori* sobre  $\theta$ . Assim, o princípio da verossimilhança postula que, para fazer inferência sobre uma

quantidade de interesse  $\theta$ , só importa aquilo que for realmente observado e não aquilo que poderia ter ocorrido, mas efetivamente não ocorreu (Ehlers, 2007).

### 2.1.3 Distribuições *a posteriori*

Segundo Paulino et al. (2003), por meio do Teorema de Bayes, a distribuição *a posteriori* é obtida. Suas informações podem ser resumidas por meio de alguns valores numéricos de interesse como a média, mediana, moda, intervalos de credibilidade e HPD (*highest posterior density interval*).

#### 2.1.3.1 Intervalos de credibilidade e HPD

Um resumo da distribuição *a posteriori* mais informativo do que qualquer estimativa pontual é obtido de uma região do espaço paramétrico  $\Theta$  que contenha uma parte substancial desta distribuição (Paulino et al., 2003). Assim, um intervalo  $(a, b)$  é chamado de intervalo de credibilidade  $100(1-\alpha)\%$  para  $\theta$ , se:  $\int_a^b \pi(\theta | x) d\theta = 1 - \alpha$ ,  $(0 \leq \alpha \leq 1)$ .

Dada a infinidade de intervalos de credibilidade com o mesmo nível de credibilidade  $100(1-\alpha)\%$ , interessa selecionar aquele com o menor comprimento possível (Bolfarine & Sandoval, 2001). Os intervalos de comprimento mínimo são obtidos tomando-se os valores de  $\theta$  com maior densidade *a posteriori*. Essa região é denominada de HPD ou intervalo de máxima densidade *a posteriori*. Assim, quanto menor for o tamanho do intervalo, mais concentrada é a distribuição do parâmetro e a amplitude do intervalo informa sobre a dispersão de  $\theta$ .

### 2.1.3.2 Teste de hipótese Bayesiano

Os intervalos de credibilidade são úteis na análise de um teste de hipótese Bayesiano. Suponha que o interesse seja testar  $H_0 : \theta = \theta_0$  versus  $H_1 : \theta \neq \theta_0$ . Para isso, primeiramente, constrói-se o intervalo de credibilidade para  $\theta$  e, se  $\theta_0$  está contido no intervalo, aceita-se  $H_0$ . Caso contrário, rejeita-se  $H_0$  (Bolfarine & Sandoval, 2001).

### 2.1.4 Métodos computacionais

Segundo Paulino et al. (2003), para se inferir em relação a qualquer elemento de  $\theta$ , a distribuição conjunta *a posteriori* dos parâmetros deve ser integrada em relação a todos os outros parâmetros que a constituem, ou seja, procura-se obter a distribuição marginal de cada um dos parâmetros. A integração dessa distribuição, geralmente, não é analítica, necessitando de algoritmos iterativos especializados denominados de algoritmos MCMC (*Markov Chain Monte Carlo*).

#### 2.1.4.1 Métodos de Monte Carlo via cadeias de Markov

Uma cadeia de Markov é um processo estocástico no qual o próximo estado da cadeia,  $\phi_{t+1}$ , depende somente do estado atual,  $\phi_t$  e dos dados e não da história passada da cadeia (Gelman, 1997). Segundo este autor, as primeiras iterações são influenciadas pelo estado inicial,  $\phi_1$  e são descartadas. Este período é conhecido como aquecimento da cadeia ou *burn-in*. Também se considera uma dependência entre as observações subsequentes da cadeia e, para se obter uma amostra independente, as observações finais devem ser obtidas a cada  $k$

iterações, sendo este valor conhecido como salto, *thin* ou intervalo de amostragem.

A idéia dos métodos MCMC é obter uma amostra das distribuições marginais *a posteriori* dos parâmetros de interesse, por meio de um processo iterativo, utilizando as distribuições condicionais completas de cada parâmetro. Por sua vez, esses valores gerados são considerados amostras aleatórias de uma determinada distribuição de probabilidade, caracterizando, assim, o método de simulação Monte Carlo. Dessa forma, tem-se uma ação conjunta dos métodos, que resulta no processo MCMC, cujos principais algoritmos são o Metropolis-Hastings e o amostrador de Gibbs.

#### 2.1.4.2 Algoritmo de Metropolis-Hastings

O algoritmo de Metropolis-Hastings está estruturado no Quadro 1.

**I** => Inicialize o contador de iterações  $t = 0$  e especifique valores iniciais  $\theta^{(0)} = (\theta_1^{(0)}, \dots, \theta_d^{(0)})$ ;

**II** => Gere um novo valor  $\theta^c$  da distribuição proposta  $q(\cdot | \theta_1)$ ;

**III** => Calcule a probabilidade de aceitação  $\alpha(\theta_1, \theta^c)$  e gere  $u \sim U(0,1)$ ;

$$\alpha(\theta_1, \theta^c) = \min\left(1, \frac{\pi(\theta^c | \theta_2, \dots, \theta_d)q(\theta_1 | \theta^c)}{\pi(\theta_1 | \theta_2, \dots, \theta_d)q(\theta^c | \theta_1)}\right)$$

**IV** => Se  $u \leq \alpha$ , então aceite o novo valor e faça  $\theta_1^{(t+1)} = \theta^c$ . Caso contrário, rejeite e faça  $\theta_1^{(t+1)} = \theta^t$ ;

**V** => Incremente o contador de  $t$  para  $t + 1$  e volte ao passo **II** até atingir a convergência.

QUADRO 1: Algoritmo de Metropolis-Hastings.

O algoritmo de Metropolis-Hastings permite gerar uma amostra da distribuição conjunta *a posteriori*  $\pi(\theta_1, \theta_2, \dots, \theta_d | x)$ , sendo  $d$  parâmetros, a partir das distribuições condicionais completas com formas desconhecidas. Eles usam a idéia de 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 essas distribuições possuem formas conhecidas, utiliza-se um caso especial do Metropolis-Hastings, o amostrador de Gibbs, no sentido de que seja fácil amostrar de seus elementos.

### 2.1.5 Método de seleção de modelos (fator de Bayes)

O fator de Bayes é um critério Bayesiano para a comparação e a seleção de modelos (Kass & Raftery, 1995). Considere a razão entre dois modelos,  $M_i$  e

$$M_j: FB_{(M_i, M_j)} = \frac{\pi(M_i | x) / \pi(M_j | x)}{\pi(M_i) / \pi(M_j)}, \text{ em que } \pi(M_i) \text{ e } \pi(M_j) \text{ são as}$$

probabilidades *a priori* e  $\pi(M_i | x)$  e  $\pi(M_j | x)$ , as probabilidades *a posteriori*.

Sendo assim, o fator de Bayes é igual à razão *a posteriori* dividida pela razão *a priori* de dois modelos, que pode ser simplificada para:

$$FB_{(M_i, M_j)} = \frac{\pi(x | M_i)}{\pi(x | M_j)}, \text{ em que } \pi(x | M_i) \text{ e } \pi(x | M_j) \text{ são as verossimilhanças}$$

marginais de cada modelo e definidas por:

$$\pi(x | M) = \int L(x | \theta, M) \pi(\theta | M) d\theta, \quad (4)$$

em que  $L(x | \theta, M)$  é a função de verossimilhança para o modelo  $M$ ,  $\pi(\theta | M)$  a distribuição *a priori* e  $\theta$  o parâmetro do modelo  $M$ .

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  $t(\theta^1, \theta^2, \dots, \theta^t)$ , na qual a verossimilhança marginal dada em (4) pode ser calculada por:

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

Uma interpretação para o fator de Bayes é dada por Jeffreys (1961) e está apresentada na Tabela 1, a qual contém categorias de evidências em relação aos modelos  $M_i$  e  $M_j$  e deve ser utilizada na escolha do melhor modelo.

TABELA 1: Interpretação do fator de Bayes, segundo Jeffreys (1961).

Valores de $FB_{(M_i, M_j)}$	Conclusão
$FB_{(M_i, M_j)} < 1$	Evidência a favor de $M_j$
$1 \leq FB_{(M_i, M_j)} < 3,2$	Evidência muito fraca a favor de $M_i$
$3,2 \leq FB_{(M_i, M_j)} < 10$	Evidência fraca a favor de $M_i$
$10 \leq FB_{(M_i, M_j)} < 100$	Evidência forte a favor de $M_i$
$FB_{(M_i, M_j)} \geq 100$	Evidência muito forte a favor de $M_i$

O fator de Bayes é eficiente quando não se utilizam *prioris* impróprias, como a *priori* de Jeffreys (Jeffreys, 1961). Para este caso, uma alternativa seria o uso do pseudofator de Bayes (Gelfand, 1996). Este é de aplicação mais prática, pois é calculado por meio de distribuições preditivas, as quais podem ser obtidas mediante valores gerados a cada iteração dos métodos MCMC (Silva et al., 2008).

### 2.1.6 Testes para avaliação da convergência

As técnicas MCMC constituem uma ótima ferramenta para a resolução de muitos problemas práticos na análise Bayesiana, porém, apresentam algumas

limitações. Entre elas incluem-se as influências dos valores iniciais da cadeia, a correlação entre os dados gerados e a falta de convergência da cadeia.

Para avaliar estas características, alguns testes implementados por Smith (2007) no pacote BOA (*Bayesian Output Analysis*) são apresentados. Este pacote pode ser instalado no software livre R (R Development Core Team, 2007) e contém um conjunto de funções para a análise de convergência dos dados gerados por meio dos métodos MCMC.

#### **2.1.6.1 Teste de Geweke**

O critério de Geweke (1992) propõe uma avaliação de convergência que se baseia no teste de igualdade de médias da primeira e da última parte da cadeia, geralmente, dos primeiros 10% e dos últimos 50%. Segundo Geweke (1992), a cadeia converge quando a maioria dos dados está entre os limites de uma distribuição normal padronizada.

#### **2.1.6.2 Teste de Gelman & Rubin**

O critério de Gelman & Rubin (1992) é um teste baseado na comparação entre duas ou mais cadeias paralelas. Ele analisa as variâncias dentro e entre as cadeias e utiliza esta informação para estimar o fator pelo qual o parâmetro escalar da distribuição marginal *a posteriori* deveria ser reduzido se a cadeia fosse repetida infinitas vezes. Este fator é expresso pelo valor  $\hat{R}$  (fator de redução de escala potencial ou fator de diagnóstico da convergência) e sugere que valores de  $\hat{R}$  próximos a 1 indicam que a convergência foi atingida para  $n$  iterações.

### **2.1.6.3 Teste de Raftery & Lewis**

O critério de Raftery & Lewis (1992b) é baseado na acurácia de estimação do quantil. Assim, determina-se o número de iterações necessárias para se estimar o quantil  $q$  dentro de uma acurácia  $\pm r$ , com probabilidade  $p$  e uma tolerância de convergência  $\delta$ . Portanto, além do número de iterações, este teste também apresenta como resultado o número de iterações iniciais que devem ser descartadas (*burn-in*) e a distância mínima ( $k$ ) de uma iteração a outra (*thin*) para se obter uma amostra independente. Outra saída importante é o fator de dependência, que é responsável pelo acréscimo multiplicativo ao número de iterações necessárias para se alcançar a convergência, sendo esta atingida quando esse valor é menor que 5 (Nogueira et al., 2004).

### **2.1.6.4 Teste de Heidelberger & Welch**

O critério de Heidelberger & Welch (1993) utiliza testes estatísticos para avaliar a hipótese nula de estacionariedade da amostra gerada. Se a hipótese nula for rejeitada para um dado valor, o teste é repetido depois de descartados os 10% valores iniciais da amostra. Se a hipótese é novamente rejeitada, mais 10% dos valores iniciais são descartados, e assim sucessivamente, até serem descartados os 50% valores iniciais. Se a hipótese for novamente rejeitada, isso indica que é necessário um número maior de iterações. Caso contrário, o número de iterações descartadas é indicado como o tamanho do *burn-in* (Nogueira et al., 2004).

### 2.1.6.5 Autocorrelação

Este teste calcula a função de autocorrelação da cadeia com determinados valores de defasagens ou *lags*. Se a cadeia apresenta alta correlação, isto é um forte indicativo de falta da convergência.

### 2.1.6.6 Verificação da convergência

Um procedimento que também é útil na avaliação da convergência é a utilização de gráficos. Segundo Gelman et al. (2000), os gráficos mais frequentes nesta análise são o gráfico de  $\theta$  ao longo das iterações e um gráfico da estimativa da distribuição *a posteriori* de  $\theta$ , por exemplo, um histograma ou uma densidade *kernel*.

Após um estudo de avaliação dos critérios de convergência para os métodos MCMC, Nogueira et al. (2004) concluíram que o seguinte procedimento deveria ser utilizado para que se obtivesse uma avaliação mais precisa da convergência:

1. aplicar o critério de Raftery & Lewis em uma amostra piloto e determinar o tamanho ideal da seqüência;
2. monitorar a convergência das seqüências nas proximidades do tamanho ideal, indicado pelo critério de Raftery & Lewis, por meio dos critérios de Gelman & Rubin e Geweke;
3. determinar o tamanho de *burn-in*, pelo critério de Heidelberger & Welch.

## **2.2 Conceitos genéticos**

A genética é a ciência que estuda a hereditariedade e as formas de evolução desses organismos. Uma de suas áreas é a genética de populações, que estuda as distribuições e as mudanças nas proporções alélicas e genótípicas sob influência, principalmente, de fatores evolutivos e endogamia. Esta também pode explicar alguns fenômenos relativos à adaptação ao meio (Falconer, 1989).

### **2.2.1 Conceitos básicos**

O gene é a unidade funcional básica da hereditariedade e está presente em locais dos cromossomos denominados loco (Gardner, 1977). Quando estes genes se localizam em um mesmo loco, ou seja, nos cromossomos homólogos, eles assumem formas alternativas responsáveis pelas diferentes manifestações fenotípicas e são denominados alelos. Uma população pode conter um gene que apresente, por exemplo, dois alelos: *A* e *B*. Segundo Griffiths et al. (1996), o genótipo é a constituição genética de um organismo, ou seja, cada organismo possui dois conjuntos gênicos, um herdado do pai e outro da mãe. Quando um indivíduo apresenta uma combinação de duas cópias *A*, então, seu genótipo é *AA* e ele é homocigoto para o alelo *A*. Um indivíduo com genótipo *AB* é heterocigoto e um indivíduo com genótipo *BB* é homocigoto para o alelo *B*. Se cada um dos três genótipos (*AA*, *AB* e *BB*) determinarem três fenótipos distintos, então, os alelos são ditos co-dominantes. Entretanto, se os indivíduos com genótipos *AA* e *AB* expressarem o mesmo fenótipo, então, *A* pode ser definido como alelo dominante e *B*, como alelo recessivo.

### **2.2.2 Fatores evolutivos**

Segundo Falconer (1989), os fatores evolutivos podem alterar as proporções alélicas e genotípicas de uma população e estes são:

- a) seleção natural: quando os genótipos que possuem maior chance de sobreviverem e produzirem descendentes, ou seja, que estão mais bem adaptados a uma determinada condição ecológica, são selecionados;
- b) migração: quando há entrada ou saída de indivíduos, assim permitindo mistura de populações;
- c) mutação: quando ocorre alteração permanente no DNA. Somente considera-se como uma força de mudança se o material genético mutante for herdado pelos filhos.

### **2.2.3 Equilíbrio de Hardy-Weinberg**

Segundo Gardner (1977), em 1908, o matemático inglês Godfrey Harold Hardy e o médico alemão Wilhelm Weinberg chegaram, independentemente e quase que simultaneamente, às mesmas conclusões do que é considerado o fundamento da genética de populações. As conclusões a que chegaram passaram a ser conhecidas como a Lei do Equilíbrio de Hardy-Weinberg ou, mais simplesmente, a Lei de Hardy-Weinberg.

Hardy e Weinberg 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. Para demonstrar este princípio, eles estudaram o efeito do cruzamento, em gerações sucessivas, nas

proporções alélicas numa população. Esta não estava sujeita a fatores evolutivos e sem presença de endogamia.

Considerando, então, os alelos  $A$  e  $B$  com proporções  $p_A$  e  $p_B = 1 - p_A$ , respectivamente, as proporções genótípicas na população seriam dadas pela seguinte relação:

- $P_{AA} = p_A^2$ : proporção do genótipo homozigoto  $AA$ ;
- $P_{AB} = 2p_A p_B$ : proporção do genótipo heterozigoto  $AB$ ;
- $P_{BB} = p_B^2$ : proporção do genótipo homozigoto  $BB$ .

Tal conjunto de proporções genótípicas é conhecido como modelo de Hardy-Weinberg. O equilíbrio de Hardy-Weinberg (EHW) ocorre quando as proporções alélicas e genótípicas permanecem constantes, ou seja, seguem o modelo de Hardy-Weinberg.

De acordo com Falconer (1989), as condições para que uma população esteja em equilíbrio de Hardy-Weinberg são:

- ausência de fatores evolutivos, como seleção natural, migração e mutação;
- os indivíduos devem ser diplóides e se reproduzir sexuadamente;
- acasalamento ao acaso e população grande.

#### **2.2.4 Violações ao modelo de Hardy-Weinberg**

De acordo com Gardner (1977), na natureza e nas populações controladas pelo homem, encontram-se dois sistemas de acasalamento: o endogâmico e o exogâmico. O cruzamento endogâmico é aquele que ocorre entre indivíduos aparentados, enquanto no cruzamento exogâmico este cruzamento é sem qualquer parentesco.

Portanto, endogamia é um sistema no qual os acasalamentos se dão entre indivíduos aparentados, ou seja, relacionados pela ascendência (Falconer, 1989). Esse cruzamento entre parentes possui algumas vantagens e desvantagens. Como desvantagem, tem-se uma maior possibilidade de se observar doenças de caráter recessivo. Por outro lado, uma vantagem refere-se à questão de adaptação do indivíduo a um determinado meio.

Em populações humanas, os casamentos consanguíneos podem ocorrer não porque sejam preferenciais, mas o tamanho reduzido da população provoca um aumento na probabilidade de parentesco consanguíneo próximo entre os cônjuges. Isso tem sido observado em aldeias localizadas em ilhas, em tribos indígenas e em pequenas comunidades religiosas e de imigrantes (Beiguelman, 2005).

Há várias maneiras de descrever violações ao modelo de Hardy-Weinberg. Entre estas, destacam-se o coeficiente de endogamia e o coeficiente de desequilíbrio.

#### **2.2.4.1 Coeficiente de endogamia**

Estudos referentes ao modo como os genes estão distribuídos nos indivíduos assumem grande importância por resultar na obtenção de informações básicas, que são úteis para o estabelecimento de estratégias mais seguras de coleta e conservação da variabilidade genética (Coelho, 2002). Um parâmetro de grande interesse neste caso é o coeficiente de endogamia.

O coeficiente de endogamia,  $f$ , mede a diminuição no número de heterozigotos em detrimento do aumento de homozigotos em uma população resultante de cruzamentos endogâmicos (Falconer, 1989). Assim, quanto menor a quantidade de genótipos heterozigotos na população, maior o grau de



parentesco entre os indivíduos em gerações sucessivas. Por causa desta diminuição na heterozigosidade, os alelos recessivos se expressam mais vezes.

É importante ressaltar que o coeficiente de endogamia pode ser negativo, embora alguns trabalhos na literatura sejam contrários a essa possibilidade. Isto porque, nestes trabalhos, o coeficiente de endogamia mede somente o efeito da endogamia, sem considerar a possibilidade de exogamia (Armborst, 2005).

As proporções genóticas homozigotas e heterozigotas, para o caso de dois alelos sob a violação do modelo de Hardy-Weinberg (modelo endogâmico), são (Weir, 1996):

$$\begin{cases} P_{AA} = p_A^2 + p_A(1-p_A)f \\ P_{AB} = 2p_A(1-p_A)(1-f) \\ P_{BB} = (1-p_A)^2 + p_A(1-p_A)f \end{cases} . \quad (5)$$

A partir deste modelo, pode-se definir que, quando  $f = 0$ , as proporções genóticas seguem a Lei de Hardy-Weinberg e, se  $f = 1$ , as proporções genóticas dos heterozigotos são nulas. Outra observação é que valores negativos de  $f$  correspondem a uma diminuição de homozigotos, enquanto os valores positivos de  $f$  indicam um aumento de homozigotos (Shoemaker et al., 1998). Segundo Weir (1996), os limites de  $f$  a partir do modelo endogâmico são dados por:

$$\max[-p_A/(1-p_A), -(1-p_A)/p_A] \leq f \leq 1, \quad (6)$$

em que o limite inferior de  $f$  depende das proporções alélicas.

### 2.2.4.2 Coeficiente de desequilíbrio

O coeficiente de desequilíbrio,  $D_A$ , mede as discrepâncias entre as proporções genótípicas sob cruzamentos aleatórios e as mesmas sob cruzamentos endogâmicos na população (Weir, 1996). Tanto o coeficiente de desequilíbrio como o coeficiente de endogamia podem ser usados para descrever violações no modelo de Hardy-Weinberg, devido a qualquer causa e não somente relativo à endogamia.

As proporções genótípicas homozigotas e heterozigotas, para o caso de dois alelos sob a violação do modelo de Hardy-Weinberg (modelo endogâmico), são (Hernández & Weir, 1989):

$$\begin{cases} P_{AA} = p_A^2 + D_A \\ P_{AB} = 2p_A(1-p_A) - 2D_A, \\ P_{BB} = (1-p_A)^2 + D_A \end{cases} \quad (7)$$

em que  $D_A = p_A(1-p_A)f$ .

A partir deste modelo, pode-se definir que, quando  $D_A = 0$ , as proporções genótípicas seguem a Lei de Hardy-Weinberg. Valores negativos de  $D_A$  correspondem a uma diminuição do número de homozigotos, enquanto os valores positivos de indicam um aumento do número de homozigotos (Shoemaker et al., 1998). Segundo Hernández & Weir (1989), os limites de  $D_A$  a partir do modelo endogâmico são dados por:

$$\max[-p_A^2, -(1-p_A)^2] \leq D_A \leq p_A(1-p_A), \quad (8)$$

em que os limites inferior e superior de  $D_A$  dependem das proporções alélicas.

A utilização desses dois parâmetros deve-se ao fato de serem os mais usados para o desequilíbrio de Hardy-Weinberg, deixando a cargo dos pesquisadores interessados nessa área a escolha do mais adequado.

### **2.3 Trabalhos clássicos em genética de populações**

De acordo com Cockerham (1969), endogamia, proporções alélicas e tamanho efetivo de população, entre outros, são termos comuns no estudo de genética de populações, sendo os conceitos e a maior parte da teoria introduzidos pelos trabalhos clássicos de Wright (1921) e Fisher (1949).

Cockerham & Weir (1983) apontaram que o coeficiente de endogamia é um parâmetro importante em genética quantitativa e de populações, sendo útil para informar sobre homozigosidade, deriva, endogamia e variação quantitativa.

Os métodos freqüentistas de Nei & Chesser (1983) e Robertson & Hill (1984) foram construídos levando-se em conta a presença de grupos na população. Esses referidos grupos são estabelecidos de acordo com cruzamentos preferenciais devido a alguma característica, ou seja, os indivíduos tendem a se cruzar com aqueles mais próximos ou com características em comum, como estatura, origem, etc. Hernández & Weir (1989) compararam vários testes para o equilíbrio de Hardy-Weinberg usando o coeficiente de desequilíbrio com ênfase em múltiplos alelos e verificaram que estes obtiveram bons resultados.

Guo & Thompson (1992) propuseram dois algoritmos para estimar o nível de significância no teste de equilíbrio de Hardy-Weinberg, considerando múltiplos alelos. Weir (1996) apresentou uma discussão geral sobre os métodos de estimação de parâmetros genéticos com base em dados de proporções alélicas. Entre 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, o autor aborda o caso de organismos haplóides, bem como populações diplóides, com modelos hierárquicos de até quatro níveis. O autor aborda, ainda, a possibilidade de usar técnicas Bayesianas na estimação dos parâmetros genéticos, uma vez que estas incorporam 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. (1996) estudaram as propriedades dos estimadores do coeficiente de endogamia e da taxa de fecundação cruzada obtidos pela análise de variância com dados de proporções alélicas em populações diplóides. Muniz et al. (1997) compararam fórmulas para a estimação da variância do estimador do coeficiente de endogamia obtido na análise de variância das proporções alélicas em uma população diplóide. Resultados de simulação mostraram que as três fórmulas propostas apresentam valores semelhantes e satisfatórios quando a proporção alélica da população estiver entre 0,3 e 0,7 e o coeficiente de endogamia da população for inferior a 0,5, trabalhando com, pelo menos, 30 indivíduos.

Chen & Thomson (1999) testaram a variância em genótipos heterozigotos individuais utilizando o parâmetro coeficiente de desequilíbrio. Muniz et al. (1999), estudando a estimação do coeficiente de endogamia em uma população diplóide, avaliaram a distribuição do quociente dos quadrados médios entre indivíduos e entre genes dentro de indivíduos. Estes autores verificaram que o teste F da análise de variância 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. Estudo de simulação validou os resultados teóricos.

Leutenegger et al. (2003) avaliaram a estimação do coeficiente de endogamia pelo método da máxima verossimilhança, levando em consideração a dependência do marcador molecular, que envolveu desenvolvimento de cadeias de Markov. O método foi avaliado com dados de manifestação autossômica recessiva obtida de estudo envolvendo a doença dos dentes *Charcot-Marie*.

## 2.4 Trabalhos Bayesianos na genética de populações

Na área de genética de populações, trabalhos importantes têm sido desenvolvidos, considerando técnicas Bayesianas. Além dos dois parâmetros já relatados,  $f$  e  $D_A$ , que se referem ao desequilíbrio de Hardy-Weinberg, um outro parâmetro é apresentado por Pereira & Rogatko (1984) e Lindley (1988),

dado por:  $\theta = \frac{P_{AB}^2}{P_{AA}P_{BB}}$ . Pereira & Rogatko (1984) utilizaram a abordagem

Bayesiana fazendo uma analogia à inferência clássica, ou seja, testes de hipótese e intervalos de confiança. Eles utilizaram a distribuição Dirichlet como *priori* para as proporções genotípicas, e assim, estimaram  $\theta$ . Lindley (1988) considerou várias *prioris*, Uniformes e não Uniformes, e dois novos parâmetros dados por:  $\alpha = \frac{1}{2}\log(4\theta)$  e  $\beta = \frac{1}{2}\log(P_{AA}/P_{BB})$ .

Chow & Fong (1992) estudaram o problema da estimação das proporções alélicas sob a perspectiva Bayesiana e estavam interessados na comparação entre as metodologias clássica e Bayesiana. Balding & Nichols (1997) utilizaram a metodologia Bayesiana na estimação do coeficiente de endogamia de uma população. Wilson & Balding (1998) aplicaram técnicas Bayesianas MCMC em estudo de amostragem de árvores para locus microsatélites. Soria et al. (1998) desenvolveram técnicas Bayesianas para fazer inferência sobre parâmetros genéticos na cultura de Eucalyptus. Os autores não encontraram diferenças nos resultados obtidos para as estimativas dos valores genéticos pelo método BLUP (*best linear unbiased prediction*) e pela técnica Bayesiana.

Segundo Ayres & Balding (1998), na avaliação de divergências do equilíbrio de Hardy-Weinberg (EHW), os métodos freqüentistas não possibilitam que se incorporem os efeitos de incerteza relativa aos parâmetros

*nuisance* (parâmetros pelos quais não se tem interesse direto), isto é, as proporções alélicas, pois o interesse maior está nas inferências sobre o coeficiente de endogamia  $f$ . Além disso, tais métodos não permitem que se imponham restrições ao espaço paramétrico de  $f$ . Sendo assim, na comparação com os métodos freqüentistas, o método Bayesiano, utilizando *prioris* Uniformes, apresentou os melhores resultados. 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 de esses parâmetros estarem em um determinado intervalo de equilíbrio. Eles usaram três *prioris* para cada parâmetro (Dirichlet, Beta - função degrau Uniforme e Uniforme - função degrau Uniforme), não considerando nenhuma forma específica de comparação entre estas.

Um guia para iniciantes, retratando a abordagem Bayesiana em estudos de genética de populações, é apresentado por Shoemaker et al. (1999). Mcguire et al. (2001) implementaram o MCMC em amostragem de árvores filogenéticas usando modelos de substituição de nucleotídeos. Ayres & Balding (2001) descreveram uma abordagem Bayesiana para análises de genótipos em multilocos para definição do equilíbrio gamético. Montoya-Delgado et al. (2001) introduziram um teste exato para descrever o equilíbrio de Hardy-Weinberg em um caso bialélico baseado no fator de Bayes.

Coelho (2002) propôs um modelo hierárquico para estimação do coeficiente de endogamia no qual há a incorporação de informações obtidas de múltiplos locos. Devido à condicionalidade do processo de estimação em relação ao polimorfismo, o que torna o processo altamente complexo, foi utilizada uma abordagem Bayesiana, a qual facilitou a implementação das hierarquias consideradas e providenciou estimativas precisas, relacionadas com a baixa amplitude dos intervalos de credibilidade obtidos. Rogatko et al. (2002)

propuseram dois diagnósticos para a avaliação do equilíbrio de Hardy-Weinberg usando o intervalo de credibilidade Bayesiano, obtendo bons resultados.

Wilson & Rannala (2003) apresentaram uma abordagem Bayesiana para estimar a taxa de migração dentro de populações, no caso de genótipos multilocos. O método exige pressuposições de que os estimadores do fluxo gênico de longo prazo podem ser aplicados em populações não estacionárias que não estejam em equilíbrio genético. Os parâmetros foram estimados usando o MCMC.

Beaumont & Rannala (2004) apresentaram uma revisão mostrando toda a revolução na área da genética com a introdução da metodologia Bayesiana. Holsinger & Wallace (2004) utilizaram uma extensão da análise Bayesiana para estudar a estrutura genética de populações. A distribuição Beta foi utilizada como aproximação da distribuição *a posteriori* do coeficiente de endogamia. Para ilustrar o método, foram utilizados dados obtidos de amostragem de DNA de orquídeas.

Amborst (2005) relatou um caso de estimação multiparamétrico que pode ser tratado com o uso de técnicas Bayesianas e o método de MCMC, que 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. A autora relata que, dentre os três métodos utilizados, sendo dois clássicos e um Bayesiano, o último apresenta maior qualidade, pois respeita o espaço paramétrico no qual o coeficiente de endogamia está definido.

### 3 MATERIAL E MÉTODOS

O objetivo principal em toda análise Bayesiana é a obtenção dos resumos *a posteriori*. Como visto, a inferência Bayesiana combina a verossimilhança dos dados com as distribuições *a priori* dos parâmetros, resultando na distribuição *a posteriori* de quantidades desconhecidas. Todas essas etapas são relacionadas na seqüência, ressaltando que, neste trabalho, utilizam-se dois parâmetros referentes ao desequilíbrio de Hardy-Weinberg, um condicionado ao coeficiente de endogamia  $f$  e o outro por meio do coeficiente de desequilíbrio  $D_A$ , sendo estes tratados separadamente.

#### 3.1 Função de verossimilhança

Considere que  $n_1$ ,  $n_2$  e  $n_3$  representam a quantidade observada de genótipos  $AA$ ,  $AB$  e  $BB$ , respectivamente, em uma amostra de tamanho  $n = n_1 + n_2 + n_3$ . Utilizou-se, neste trabalho, a definição  $p_B = 1 - p_A$ , para a proporção do alelo  $B$ . Esses dados apresentam uma distribuição multinomial e, portanto, 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 (5) e (7) têm-se, respectivamente:

$$L(p_A, f | n_1, n_2, n_3) = \frac{n!}{n_1!n_2!n_3!} [p_A^2 + p_A(1-p_A)f]^{n_1} [2p_A(1-p_A)(1-f)]^{n_2} [(1-p_A)^2 + p_A(1-p_A)f]^{n_3}. \quad (9)$$

$$L(p_A, D_A | n_1, n_2, n_3) = \frac{n!}{n_1!n_2!n_3!} [p_A^2 + D_A]^{n_1} [2p_A(1-p_A) - 2D_A]^{n_2} [(1-p_A)^2 + D_A]^{n_3}. \quad (10)$$



### 3.2 Distribuições *a priori*

Foram adotadas quatro *prioris* para cada caso estudado e estas estão baseadas na conveniência matemática, por informações obtidas em estudos, como aqueles da pesquisa da *National Research Council* (1996) e pela ausência de informação. Portanto, seguindo a seqüência dessas *prioris*, esse conhecimento relativo ao parâmetro tende a ser menos informativo.

#### 3.2.1 *Priori* Dirichlet (Modelo 1)

A distribuição Dirichlet com hiperparâmetros inteiros  $\gamma_1$ ,  $\gamma_2$  e  $\gamma_3$  é a conjugada natural da distribuição multinomial e 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 (5) e (7) têm-se,

respectivamente, a *priori* conjunta dada por:

$$\pi(p_A, f) = \frac{\Gamma(\gamma)}{\Gamma(\gamma_1)\Gamma(\gamma_2)\Gamma(\gamma_3)} [p_A^2 + p_A(1-p_A)f]^{\gamma_1-1} [2p_A(1-p_A)(1-f)]^{\gamma_2-1} [(1-p_A)^2 + p_A(1-p_A)f]^{\gamma_3-1}. \quad (11)$$

$$\pi(p_A, D_A) = \frac{\Gamma(\gamma)}{\Gamma(\gamma_1)\Gamma(\gamma_2)\Gamma(\gamma_3)} [p_A^2 + D_A]^{\gamma_1-1} [2p_A(1-p_A) - 2D_A]^{\gamma_2-1} [(1-p_A)^2 + D_A]^{\gamma_3-1}. \quad (12)$$

#### 3.2.2 *Priori* Beta - função degrau Uniforme (Modelo 2) e *priori* Uniforme - função degrau Uniforme (Modelo 3)

A pesquisa da *National Research Council* (1996) sugeriu que, para populações humanas, valores de  $f$  ou  $D_A$  dentro de certos intervalos indicariam

que esta população estaria em desequilíbrio de Hardy-Weinberg. Vale ressaltar que essa pesquisa é válida apenas em populações humanas, não sendo os mesmos resultados utilizados em análises com animais e vegetais. Segundo esta pesquisa, o intervalo para o parâmetro  $f$  é :

$$\max[-0,03; -p_A/(1-p_A), -(1-p_A)/p_A] \leq f \leq 0,03 \quad (13)$$

e para o parâmetro  $D_A$  é :

$$\max[-0,03p_A(1-p_A); -p_A^2; -(1-p_A)^2] \leq D_A \leq 0,03p_A(1-p_A). \quad (14)$$

Na Tabela 2 são apresentados os intervalos deste relatório, assim como a probabilidade de o parâmetro  $f$  ou  $D_A$  estar no intervalo e a situação desta população (EHW => equilíbrio ou DHW => desequilíbrio). Os limites dos intervalos são demonstrados no Anexo.

TABELA 2: Intervalos sugeridos pela *National Research Council* (1996)\*.

Situação	Prob. ( $\alpha$ )	Limite inferior	Limite superior
$f$			
DHW	0,25	I	II
EHW	0,50	II	III
DHW	0,25	III	IV
$D_A$			
DHW	0,25	V	VI
EHW	0,50	VI	VII
DHW	0,25	VII	VIII

\* I =>  $\max[-p_A/(1-p_A), -(1-p_A)/p_A]$  ; II =>  $\max[-0,03; -p_A/(1-p_A), -(1-p_A)/p_A]$  ; III => 0,03; IV => 1;

V =>  $\max[-p_A^2, -(1-p_A)^2]$  ; VI =>  $\max[-0,03p_A(1-p_A); -p_A^2; -(1-p_A)^2]$  ; VII =>  $0,03p_A(1-p_A)$  ; VIII =>  $p_A(1-p_A)$ .

Portanto, a *priori* conjunta Beta - função degrau Uniforme é obtida por  $\pi(p_A, f) = \pi(p_A)\pi(f | p_A)$  ou  $\pi(p_A, D_A) = \pi(p_A)\pi(D_A | p_A)$ . A distribuição *a priori* para  $p_A$ ,  $\pi(p_A)$ , foi condicionada por uma distribuição Beta com hiperparâmetros  $\alpha$  e  $\beta$  e a distribuição condicional *a priori* para  $f$  dado  $p_A$ ,  $\pi(f | p_A)$  ou  $D_A$  dado  $p_A$ ,  $\pi(D_A | p_A)$ , foi determinada por uma função degrau Uniforme sob cada um dos três intervalos dados pela Tabela 2. Dessa forma, a *priori* conjunta para cada um dos parâmetros,  $f$  e  $D_A$ , é dada por:

$$\begin{aligned}\pi(p_A, f) &= p_A^{\alpha-1}(1-p_A)^{\beta-1} \sum_{i=0}^2 \alpha_i \cdot 1_{A_i}(f) \\ \pi(p_A, D_A) &= p_A^{\alpha-1}(1-p_A)^{\beta-1} \sum_{i=0}^2 \alpha_i \cdot 1_{A_i}(D_A)\end{aligned}\tag{15}$$

sendo o intervalo  $A_i$  Uniforme e determinado pela Tabela 2, de acordo com os parâmetros  $f$  ou  $D_A$ , respectivamente.

A *priori* conjunta Uniforme - função degrau Uniforme é obtida por  $\pi(p_A, f) = \pi(p_A)\pi(f | p_A)$  ou  $\pi(p_A, D_A) = \pi(p_A)\pi(D_A | p_A)$ , diferenciando-se 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 para cada um dos parâmetros,  $f$  e  $D_A$ , é dada por:

$$\begin{aligned}\pi(p_A, f) &= U_{(0,1)}(p_A) \sum_{i=0}^2 \alpha_i \cdot 1_{A_i}(f) \\ \pi(p_A, D_A) &= U_{(0,1)}(p_A) \sum_{i=0}^2 \alpha_i \cdot 1_{A_i}(D_A)\end{aligned}\tag{16}$$

sendo o intervalo  $A_i$  Uniforme e determinado pela Tabela 2 de acordo com os parâmetros  $f$  ou  $D_A$ , respectivamente.

### 3.2.3 Prioris Uniformes (Modelo 4)

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  $f$  ou  $D_A$ . Portanto, a *priori* conjunta é dada por  $\pi(p_A, f) = \pi(p_A)\pi(f)$  ou  $\pi(p_A, D_A) = \pi(p_A)\pi(D_A)$  em que:

$$\begin{aligned}\pi(p_A) &\sim U_{(0,1)} \\ \pi(f) &\sim U_{(\max[-p_A/(1-p_A), -(1-p_A)/p_A], 1)} \\ \pi(D_A) &\sim U_{(\max[-p_A^2, -(1-p_A)^2], p_A(1-p_A))}\end{aligned}\quad (17)$$

### 3.3 Distribuições conjuntas *a posteriori*

Toda a abordagem apresentada nas seções 3.1 e 3.2 resulta na distribuição conjunta *a posteriori* para  $f$  e  $D_A$  e estas são representadas, respectivamente, por  $\pi(p_A, f | n_1, n_2, n_3) \propto L(p_A, f | n_1, n_2, n_3)\pi(p_A, f)$  e  $\pi(p_A, D_A | n_1, n_2, n_3) \propto L(p_A, D_A | n_1, n_2, n_3)\pi(p_A, D_A)$ .

#### 3.3.1 Distribuição conjunta *a posteriori* (Modelo 1)

Nesse caso, tem-se, devido à propriedade de conjugação entre as distribuições Dirichlet e Multinomial, uma reparametrização da distribuição *a priori* Dirichlet, resultando em uma distribuição conjunta *a posteriori* Dirichlet reparametrizada. Para  $f$ , esta é de (9) e (11), dada por:

$$\begin{aligned}\pi(p_A, f | n_1, n_2, n_3) &\propto [p_A^2 + p_A(1-p_A)f]^{n_1+\gamma_1-1} \\ &\quad [2p_A(1-p_A)(1-f)]^{n_2+\gamma_2-1} [(1-p_A)^2 + p_A(1-p_A)f]^{n_3+\gamma_3-1}\end{aligned}\quad (18)$$

e para  $D_A$ , de (10) e (12) por:

$$\begin{aligned} \pi(p_A, D_A | n_1, n_2, n_3) &\propto \\ & [p_A^2 + D_A]^{n_1 + \gamma_1 - 1} [2p_A(1 - p_A) - 2D_A]^{n_2 + \gamma_2 - 1} [(1 - p_A)^2 + D_A]^{n_3 + \gamma_3 - 1}. \end{aligned} \quad (19)$$

### 3.3.2 Distribuição conjunta *a posteriori* (Modelo 2)

Para  $f$ , de (9) e (15) tem-se:

$$\begin{aligned} \pi(p_A, f | n_1, n_2, n_3) &\propto [p_A^2 + p_A(1 - p_A)f]^{n_1} \\ & [2p_A(1 - p_A)(1 - f)]^{n_2} [(1 - p_A)^2 + p_A(1 - p_A)f]^{n_3} p_A^{\alpha - 1} (1 - p_A)^{\beta - 1} \sum_{i=0}^2 \alpha_i \cdot 1_{A_i}(f). \end{aligned} \quad (20)$$

Para  $D_A$ , de (10) e (15) tem-se:

$$\begin{aligned} \pi(p_A, D_A | n_1, n_2, n_3) &\propto [p_A^2 + D_A]^{n_1} \\ & [2p_A(1 - p_A) - 2D_A]^{n_2} [(1 - p_A)^2 + D_A]^{n_3} p_A^{\alpha - 1} (1 - p_A)^{\beta - 1} \sum_{i=0}^2 \alpha_i \cdot 1_{A_i}(D_A). \end{aligned} \quad (21)$$

### 3.3.3 Distribuição conjunta *a posteriori* (Modelo 3)

Para  $f$ , de (9) e (16), tem-se:

$$\begin{aligned} \pi(p_A, f | n_1, n_2, n_3) &\propto [p_A^2 + p_A(1 - p_A)f]^{n_1} \\ & [2p_A(1 - p_A)(1 - f)]^{n_2} [(1 - p_A)^2 + p_A(1 - p_A)f]^{n_3} U_{(0,1)}(p_A) \sum_{i=0}^2 \alpha_i \cdot 1_{A_i}(f). \end{aligned} \quad (22)$$

Para  $D_A$ , de (10) e (16), tem-se:

$$\begin{aligned} \pi(p_A, D_A | n_1, n_2, n_3) &\propto [p_A^2 + D_A]^{n_1} \\ & [2p_A(1 - p_A) - 2D_A]^{n_2} [(1 - p_A)^2 + D_A]^{n_3} U_{(0,1)}(p_A) \sum_{i=0}^2 \alpha_i \cdot 1_{A_i}(D_A). \end{aligned} \quad (23)$$

### 3.3.4 Distribuição conjunta *a posteriori* (Modelo 4)

Para  $f$ , de (9) e (17), tem-se:

$$\begin{aligned} \pi(p_A, f | n_1, n_2, n_3) &\propto [p_A^2 + p_A(1-p_A)f]^{n_1} [2p_A(1-p_A)(1-f)]^{n_2} \\ &[(1-p_A)^2 + p_A(1-p_A)f]^{n_3} U_{(0,1)}(p_A) U_{(\max[-p_A/(1-p_A), -(1-p_A)/p_A], 1)}(f). \end{aligned} \quad (24)$$

Para  $D_A$ , de (10) e (17), tem-se:

$$\begin{aligned} \pi(p_A, D_A | n_1, n_2, n_3) &\propto [p_A^2 + D_A]^{n_1} [2p_A(1-p_A) - 2D_A]^{n_2} \\ &[(1-p_A)^2 + D_A]^{n_3} U_{(0,1)}(p_A) U_{(\max[-p_A^2, -(1-p_A)^2], p_A(1-p_A))}(D_A). \end{aligned} \quad (25)$$

### 3.4 Distribuições condicionais completas *a posteriori*

A inferência será baseada em amostras obtidas por meio das distribuições condicionais completas *a posteriori*, usando algoritmos MCMC, pois não é possível obter as distribuições marginais *a posteriori* a partir da distribuição conjunta *a posteriori*, ou seja, as integrais desta não possuem solução analítica. Portanto, apresentam-se as distribuições condicionais completas *a posteriori* necessárias à implementação do algoritmo Metropolis-Hastings, verificando-se que estas possuem forma desconhecida.

#### 3.4.1 Distribuições condicionais completas *a posteriori* (Modelo 1)

As distribuições condicionais completas *a posteriori*, para  $p_A$  e  $f$ , são dadas, respectivamente, por  $\pi(p_A | f, n_1, n_2, n_3)$  e  $\pi(f | p_A, n_1, n_2, n_3)$ , que apresentam a mesma forma e correspondem à distribuição conjunta *a posteriori* dada em (18).

As distribuições condicionais completas *a posteriori* para  $p_A$  e  $D_A$  são dadas, respectivamente, por  $\pi(p_A | D_A, n_1, n_2, n_3)$  e  $\pi(D_A | p_A, n_1, n_2, n_3)$ . Elas apresentam a mesma forma e correspondem à distribuição conjunta *a posteriori* dada em (19).

### 3.4.2 Distribuições condicionais completas *a posteriori* (Modelo 2)

A distribuição condicional completa *a posteriori* para  $p_A$ ,  $\pi(p_A | f, n_1, n_2, n_3)$  corresponde à distribuição conjunta *a posteriori* dada em (20). Já a distribuição condicional completa *a posteriori* de  $f$ ,  $\pi(f | p_A, n_1, n_2, n_3)$ , a partir da expressão (20), é dada por:

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

A distribuição condicional completa *a posteriori* para  $p_A$ ,  $\pi(p_A | D_A, n_1, n_2, n_3)$  corresponde à distribuição conjunta *a posteriori* dada em (21). Já a distribuição condicional completa *a posteriori* de  $D_A$ ,  $\pi(D_A | p_A, n_1, n_2, n_3)$ , a partir da expressão (21), é dada por:

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

### 3.4.3 Distribuições condicionais completas *a posteriori* (Modelo 3)

A distribuição condicional completa *a posteriori* para  $p_A$ ,  $\pi(p_A | f, n_1, n_2, n_3)$ , corresponde à distribuição conjunta *a posteriori* dada em (22). Já a distribuição condicional completa *a posteriori* de  $f$ ,  $\pi(f | p_A, n_1, n_2, n_3)$ , a partir da expressão (22), é dada por:

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

A distribuição condicional completa *a posteriori* para  $p_A$ ,  $\pi(p_A | D_A, n_1, n_2, n_3)$ , corresponde à distribuição conjunta *a posteriori* dada em (23). Já a distribuição condicional completa *a posteriori* de  $D_A$ ,  $\pi(D_A | p_A, n_1, n_2, n_3)$ , a partir da expressão (23), é dada por:

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

### 3.4.4 Distribuições condicionais completas *a posteriori* (Modelo 4)

As distribuições condicionais completas *a posteriori* para  $p_A$ ,  $\pi(p_A | f, n_1, n_2, n_3)$  e  $f$ ,  $\pi(f | p_A, n_1, n_2, n_3)$ , são dadas, respectivamente, por:

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



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

As distribuições condicionais completas *a posteriori* para  $p_A$ ,  $\pi(p_A | D_A, n_1, n_2, n_3)$  e  $D_A$ ,  $\pi(D_A | p_A, n_1, n_2, n_3)$ , são dadas, respectivamente, por:

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

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

### 3.5 Implementação do código

O código para a obtenção de uma amostra da distribuição conjunta *a posteriori* e, conseqüentemente, das distribuições marginais, análise da convergência e fator de Bayes, foi implementado na linguagem R (R Development Core Team, 2007) e consta, basicamente, dos seguintes passos:

Passo 1: definição dos valores iniciais para os parâmetros.

Passo 2: definição dos hiperparâmetros.

Passo 3: definição dos valores do erro das proporções alélicas ( $\varepsilon_{p_A}$ ), erro do coeficiente de endogamia ( $\varepsilon_f$ ) e erro do coeficiente de desequilíbrio ( $\varepsilon_{D_A}$ ). Estes serão usados nos limites da distribuição Uniforme para o controle da taxa de aceitação.

Passo 4: repetir.

- a) Amostrar  $p_A$  a partir de sua distribuição condicional completa;

b) Amostrar  $f$  ou  $D_A$ , a partir de sua distribuição condicional completa.

O algoritmo Metropolis-Hastings é implementado, sendo esta etapa realizada para cada *priori* abordada. Para facilitar os cálculos, foi utilizada a função *log* em cada membro da probabilidade de aceitação.

Passo 5: fazer o passo 4  $n$  vezes (número de iterações) e considerar o período de aquecimento da cadeia (*burn-in*) e o espaçamento entre pontos amostrados (*thin*).

Passo 6: calcular a taxa de aceitação de cada parâmetro.

Para se obter 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) devem-se escolher valores adequados dos erros das proporções alélicas, do coeficiente de endogamia e do coeficiente de desequilíbrio. O valor de  $\varepsilon_{p_A}$  não poderá ser muito alto, pois a cadeia não irá se mover. Já o valor de  $\varepsilon_f$  deve ser

maior que  $\frac{k^2 \varepsilon_{p_A}}{(k-1)(k-1-k\varepsilon_{p_A})}$ , sendo  $k$  o número de alelos (Armborst, 2005).

$p_A$

Foi verificado também que o  $\varepsilon_{D_A}$  não obedece à mesma expressão do  $\varepsilon_f$ , pois deve assumir valores bem menores. Sendo assim, estes foram testados até atingirem valores satisfatórios para a convergência.

Passo 7: realizar a análise de convergência da cadeia.

Passo 8: calcular as estimativas dos parâmetros, resumindo-as por meio da média, mediana, moda, desvio-padrão e HPD.

Passo 9: calcular o fator de Bayes.

Para obter valores amostrados dos parâmetros  $p_A$ ,  $f$  e  $D_A$ , utilizou-se como função candidata a distribuição Uniforme no intervalo entre o limite inferior e superior de cada parâmetro, sendo a mesma utilizada nos trabalhos de Ayres & Balding (1998) e Armborst (2005). 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. Quando as *prioris* baseadas na função degrau são utilizadas, um processo de escolha de intervalo é realizado, ou seja, estimado o valor de  $p_A$ , os três intervalos dados pela Tabela 2 são definidos e, então, escolhe-se aquele que contém o valor candidato do parâmetro e, depois, o valor atual do parâmetro, sendo este executado durante o processo de aceitação-rejeição no Metropolis-Hastings. E, assim, é definido o intervalo que será usado pela distribuição Uniforme na função degrau. Os códigos implementados para todas as análises se encontram no site [www.dex.ufla.br/~muniz](http://www.dex.ufla.br/~muniz).

### 3.6 Análise dos dados simulados

Um estudo de simulação foi realizado no intuito de avaliar a metodologia empregada e comparar as características proporcionadas por todas as *prioris* testadas. O algoritmo para a definição do número de genótipos de cada tipo está resumido no Quadro 2:

**I** => Gerar um valor para  $p_A$  de um  $U \sim (0,1)$ .

**II** => Fixar um valor para  $f$  ou  $D_A$  como se esse fosse o verdadeiro valor da população.

**III** => Calcular as proporções genotípicas de acordo com cada modelo endogâmico dado em (5) ou (7).

**IV** => Gerar valores de  $n_1, n_2$  e  $n_3$  por meio de uma multinomial, sendo  $n$  o número de indivíduos amostrados.

QUADRO 2: Simulação de dados.

Assim, a partir do modelo endogâmico, vários cenários foram abordados e estes diferiram pelo tamanho da amostra ( $n=50;200;1000$ ) e pela intensidade do parâmetro analisado, sendo considerado um valor próximo ao limite inferior do parâmetro ( $f = 0,8$  e  $D_A = 0,146$ ), um valor positivo próximo do EHW ( $f = 0,02$  e  $D_A = 0,02$ ) e outro que apresente alta endogamia ( $f = -0,217$  e  $D_A = -0,02$ ). Esta mesma regra é apresentada em Armbrorst (2005). Perfaz-se, então, um total de 9 cenários, sendo que, neste trabalho, o número de alelos foi 2 e a proporção alélica usada foi 0,21.

### 3.6.1 Avaliação da acurácia

Foram simuladas  $m=100$  amostras para cada *posteriori*, sendo estimados os valores pontuais e por intervalo para cada um dos 9 cenários descritos. Para avaliação da metodologia, utilizou-se o vício, o erro quadrático médio (EQM) e a probabilidade estimada de cobertura dos intervalos de credibilidade dos parâmetros. Tais valores estão descritos para  $f$ , sendo as mesmas regras utilizadas para o parâmetro  $D_A$ :

1. vício: mede a diferença entre a média dos  $m$  valores estimados de  $f$  e o seu valor verdadeiro,  $f^*$ , ou seja, Vício = ;
2. erro quadrático médio: mede a média dos desvios quadráticos de  $\hat{f}$  com respeito a  $f^*$ , ou seja, EQM = ;
3. probabilidade estimada de cobertura: descreve a proporção dos  $m$  intervalos que contêm o valor verdadeiro  $f^*$ .

Armbrorst (2005) utilizou este mesmo procedimento na avaliação de métodos clássicos e Bayesianos.

### 3.8 Análise dos dados reais

Os dados *FBI* e *Cellmark* analisados foram retirados do trabalho de Shoemaker et al. (1998) e referem-se 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 expressão raças humanas refere-se ao conceito antropológico que classifica grupos populacionais com base em vários conjuntos de características somáticas e crenças sobre ancestralidade comum. As categorias mais amplamente usadas neste sentido restrito baseiam-se em traços visíveis, tais como cor da pele, conformação do crânio e do rosto e tipo de cabelo, bem como a auto-identificação (Long & Kittles, 2003).

Afro-americanos é uma das designações oficiais para os cidadãos dos Estados Unidos descendentes de africanos. O termo caucasianos foi criado para classificar o grupo humano que é mais conhecido como "raça branca", pelo seu tom de pele geralmente claro. O termo surgiu de antigos estudos de antropologia que acreditavam que tal grupo havia se originado no Cáucaso. Os hispânicos são um grupo racial cuja ascendência remonta à Espanha ou, no seu uso mais freqüente, são cidadãos residentes nos Estados Unidos que sejam originários de países de língua espanhola da América Latina (Olsen, 2003). As Tabelas 3, 4 e 5 apresentam as proporções genotípicas e o número de indivíduos amostrados ( $n$ ) de cada grupo, sendo também atribuídos algarismos romanos a cada um, para uma melhor visualização posteriormente.

TABELA 3: Proporções genotípicas para o loco D7S8.

Grupo racial	<i>n</i>	AA	AB	BB
<i>FBI</i>				
Afro-americanos (I)	145	0,338	0,552	0,110
Caucasianos (II)	148	0,358	0,514	0,128
Hispânicos (sudeste) (III)	93	0,344	0,495	0,161
Hispânicos (sudoeste) (IV)	97	0,454	0,443	0,103
<i>Cellmark</i>				
Afro-americanos (V)	100	0,450	0,420	0,130
Caucasianos (VI)	103	0,301	0,485	0,214
Hispânicos (VII)	200	0,410	0,425	0,165

Fonte: Shoemaker et al. (1998).

TABELA 4: Proporções genotípicas para o loco GYPA.

Grupo racial	<i>n</i>	AA	AB	BB
<i>FBI</i>				
Afro-americanos (I)	145	0,228	0,503	0,269
Caucasianos (II)	148	0,351	0,466	0,182
Hispânicos (sudeste) (III)	93	0,333	0,409	0,258
Hispânicos (sudoeste) (IV)	97	0,443	0,412	0,144
<i>Cellmark</i>				
Afro-americanos (V)	100	0,220	0,500	0,280
Caucasianos (VI)	103	0,301	0,476	0,223
Hispânicos (VII)	200	0,370	0,490	0,140

Fonte: Shoemaker et al. (1998).

TABELA 5: Proporções genotípicas para o loco LDLR.

Grupo racial	<i>n</i>	AA	AB	BB
<i>FBI</i>				
Afro-americanos (I)	145	0,048	0,352	0,600
Caucasianos (II)	148	0,176	0,554	0,270
Hispânicos (sudeste) (III)	93	0,194	0,441	0,365
Hispânicos (sudoeste) (IV)	97	0,309	0,505	0,186
<i>Cellmark</i>				
Afro-americanos (V)	100	0,020	0,500	0,480
Caucasianos (VI)	103	0,165	0,544	0,291
Hispânicos (VII)	200	0,245	0,480	0,275

Fonte: Shoemaker et al. (1998).

## 4 RESULTADOS E DISCUSSÃO

### 4.1 Dados simulados

#### 4.1.1 Considerações gerais

O procedimento proposto por Nogueira et al. (2004) foi utilizado na definição do número de iterações necessário à convergência, assim como na determinação dos valores de *burn-in* e *thin*. O teste de Raftery & Lewis sugeriu um processo com 50.000 iterações, sendo, de acordo com o teste de Heidelberger & Welch, descartadas as 10.000 iniciais, para o período de aquecimento da cadeia (*burn-in*). Para assegurar a independência da amostra, pela análise da autocorrelação, considerou-se um espaçamento, entre os pontos amostrados, de tamanho 40 (*thin*), ou seja, obteve-se uma amostra final, para cada parâmetro, de tamanho 1.000.

A convergência das cadeias de todos os parâmetros do modelo foi monitorada por meio da visualização gráfica do traço e dos critérios disponíveis no pacote BOA do software livre R (R Development Core Team, 2007), não existindo evidências contra a convergência. Sendo assim, pelo critério de Geweke, o p-valor estimado foi sempre maior que o nível de significância pré-fixado (5%) e, em relação ao critério de Gelman & Rubin, este sempre apresentou valores de  $\hat{R}$  próximos a 1. Foram consideradas nesta análise, duas cadeias com diferentes valores iniciais.

#### 4.1.2 Parâmetro $f$

Os resultados observados em relação ao parâmetro  $f$ , referentes a uma amostra, ou seja,  $m=1$ , são apresentados nas Tabelas 6 a 8.

TABELA 6: Média ( $\hat{f}_1$ ), mediana ( $\hat{f}_2$ ), moda ( $\hat{f}_3$ ), desvio padrão e HPD, considerando o valor verdadeiro  $f = 0,8$ .

Modelo	$n$	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$DP$	$LI$	$LS$
1	50	0,707	0,717	0,778	0,117	0,480	0,910
	200	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
2	50	0,749	0,762	0,831	0,115	0,525	0,941
	200	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
3	50	0,742	0,754	0,774	0,116	0,528	0,944
	200	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
4	50	0,742	0,759	0,812	0,121	0,522	0,956
	200	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

TABELA 7: Média ( $\hat{f}_1$ ), mediana ( $\hat{f}_2$ ), moda ( $\hat{f}_3$ ), desvio padrão e HPD, considerando o valor verdadeiro  $f = 0,02$ .

Modelo	$n$	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$DP$	$LI$	$LS$
1	50	0,063	0,057	-0,005	0,139	-0,180	0,334
	200	-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
2	50	0,027	0,002	-0,053	0,155	-0,233	0,294
	200	-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
3	50	0,026	0,008	-0,052	0,141	-0,209	0,279
	200	-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
4	50	0,018	-0,001	-0,017	0,141	-0,206	0,302
	200	-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



TABELA 8: Média ( $\hat{f}_1$ ), mediana ( $\hat{f}_2$ ), moda ( $\hat{f}_3$ ), desvio padrão e HPD, considerando o valor verdadeiro  $f = -0,217$ .

Modelo	$n$	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$DP$	$LI$	$LS$
1	50	-0,058	-0,085	-0,119	0,149	-0,297	0,215
	200	-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
2	50	-0,042	-0,091	-0,112	0,205	-0,327	0,451
	200	-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
3	50	-0,037	-0,088	-0,121	0,197	-0,295	0,431
	200	-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
4	50	-0,075	-0,118	-0,147	0,172	-0,326	0,246
	200	-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

Tendo em vista os resultados obtidos, observa-se que:

- em todos os cenários, o HPD continha o valor verdadeiro do parâmetro  $f$  ;
- todos os modelos apresentaram estimativas (média, mediana e moda) de  $f$  semelhantes, principalmente para os cenários de  $n=200$  e  $n=1000$ ;
- nota-se que, de modo geral, a moda apresentou os melhores valores para a estimativa de  $f$  , nos cenários de  $n=50$  e  $f = 0,8$  ou  $f = -0,217$ , mostrando a assimetria da distribuição e que a média apresentou os melhores valores para o cenário de  $n=50$  e  $f = 0,02$ . Para os outros cenários, as estimativas da média, mediana e moda de  $f$  foram bem próximas entre si, sendo que, para  $n=1000$ , foram também bem próximas ao valor verdadeiro de  $f$  ;
- o processo de simulação propiciou uma melhor análise dos modelos, pois considerou vários cenários possíveis.

Resultados semelhantes são apresentados por Ayres & Balding (1998) e Armbrorst (2005), para o modelo com *prioris* Uniformes, destacando-se que, além de respeitarem o espaço paramétrico, estes propiciam também a estimação das proporções alélicas. Shoemaker et al. (1998) não utilizaram, em seu trabalho, nenhum processo de simulação para as outras três *prioris* abordadas.

Nas Figuras de 2 a 4 está representada a distribuição marginal *a posteriori* para o parâmetro  $f$ , para cada modelo utilizado, considerando os três cenários de  $f = 0,8$ . Observa-se que, à medida que  $n$  aumenta, a amplitude de  $f$  diminui, a densidade aumenta e a distribuição vai se tornando mais simétrica, com uma concentração de valores de  $f$  em torno de seu valor verdadeiro. As demais figuras foram omitidas, por serem semelhantes aos três cenários apresentados.

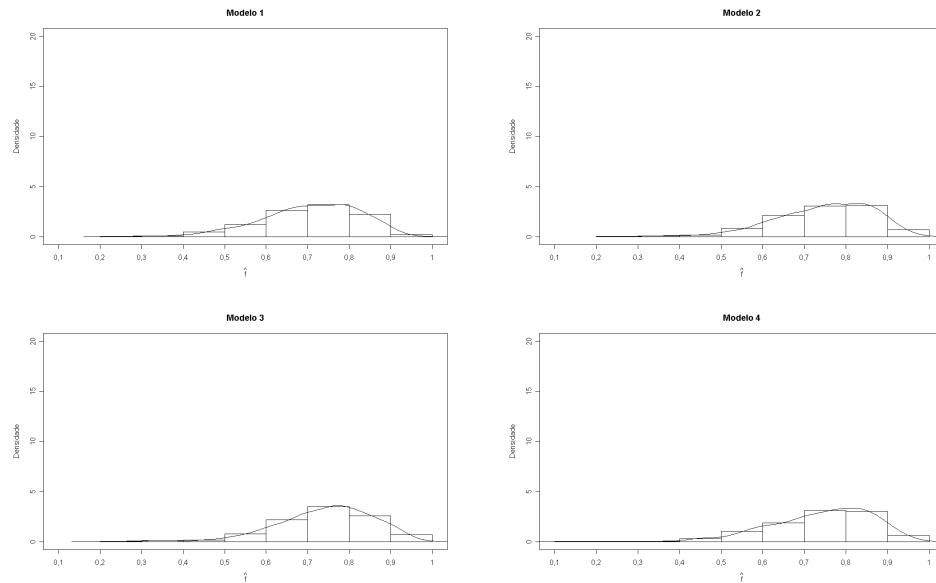


FIGURA 2: Gráfico da distribuição marginal *a posteriori* de  $f$  ( $n=50$ ).

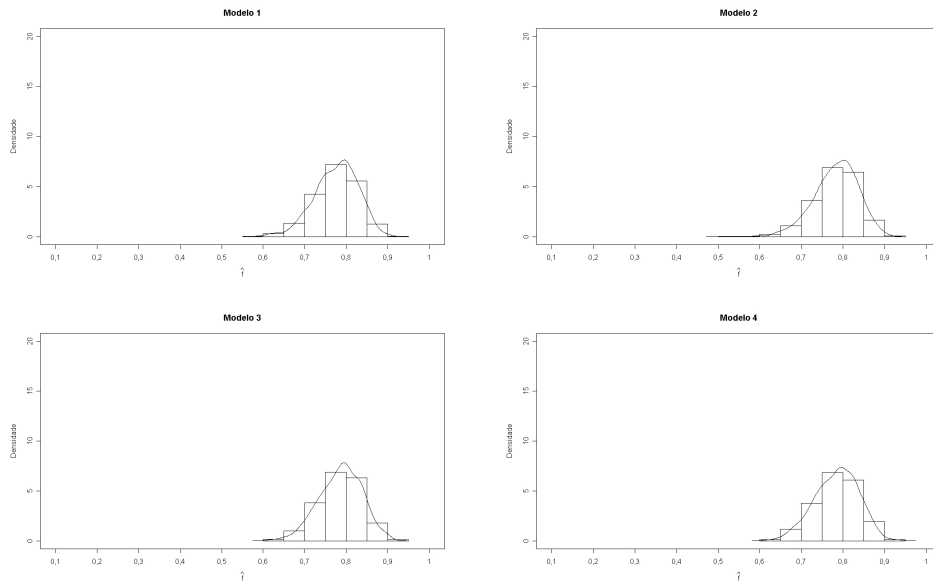


FIGURA 3: Gráfico da distribuição marginal *a posteriori* de  $f$  ( $n=200$ ).

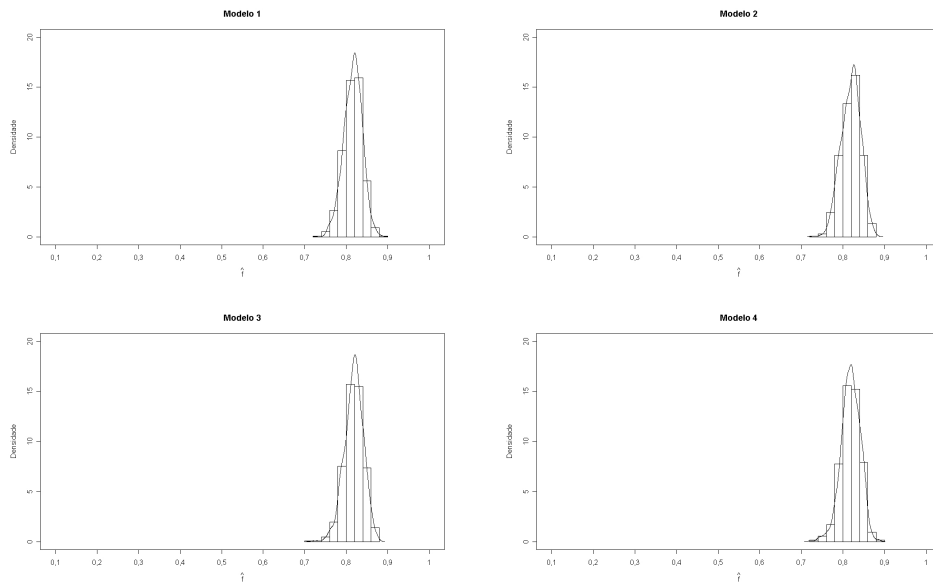


FIGURA 4: Gráfico da distribuição marginal *a posteriori* de  $f$  ( $n=1000$ ).

#### 4.1.2.1 Taxas de aceitação

Houve grande dificuldade em se alcançar taxas de aceitação, para  $p_A$  e  $f$ , entre 20% a 50%. Uma série de tentativas foi realizada com o objetivo de determinar valores de  $\varepsilon_p$  e  $\varepsilon_f$ , para cada um dos cenários. Na Tabela 9 são apresentados esses valores.

TABELA 9: Erros das proporções alélicas e do coeficiente de endogamia.

$f$	$n$	$\varepsilon_p$	$\varepsilon_f$
0,8	50	0,2	1,333
	200	0,1	0,5
	1000	0,038	0,164
0,02	50	0,2	1,333
	200	0,1	0,5
	1000	0,038	0,164
-0,217	50	0,3	3
	200	0,07	0,325
	1000	0,038	0,164

Esses valores são sugeridos para os modelos apresentados em conjunto e foram diferentes daqueles estabelecidos por Armbrorst (2005), pois, em seu trabalho, apenas o modelo com *prioris* Uniformes foi analisado. A autora ainda relata que, em alguns casos, devem-se evitar valores de  $\varepsilon_f$  menores que 0,1, o que conduziria a intervalos de credibilidade muito estreitos, podendo não conter o valor verdadeiro de  $f$ .

#### 4.1.2.2 Fator de Bayes

Os resultados da comparação dos modelos por meio do fator de Bayes encontram-se na Tabela 10.

TABELA 10: Fator de Bayes para o parâmetro  $f$ .

$f$	$n$	FB <sub>12</sub>	FB <sub>13</sub>	FB <sub>14</sub>	FB <sub>32</sub>	FB <sub>42</sub>	FB <sub>43</sub>
0,8	50	2245	3101	1019	7,24	22,03	3,04
	200	1196	2037	7024	5,87	17,02	2,89
	1000	4889	7690	2516	6,35	19,43	3,05
0,02	50	5066	7614	2254	6,65	2,24	0,29
	200	4112	6343	3030	6,48	1,35	0,47
	1000	6745	1072	6492	6,29	1,03	0,60
-0,217	50	1449	2065	3206	7,01	4,52	0,15
	200	3645	5871	6635	6,20	5,49	0,11
	1000	7695	1239	1639	6,20	4,69	0,13

Observa-se que o modelo 1 apresentou evidência muito forte em relação a todos os outros. O modelo 3 apresenta evidência fraca em relação ao modelo 2, já o modelo 4 mostra evidência fraca para os cenários com  $f = -0,217$ , muito fraca para os cenários com  $f = 0,02$  e forte para os cenários com  $f = 0,8$  em relação ao modelo 2 e evidência muito fraca para os cenários com  $f = 0,8$  em relação ao modelo 3, não sendo indicado para os outros cenários. Portanto, o modelo considerando mais adequado foi o 1, ou seja, aquele que representa a *priori* mais informativa, seguido pelo modelo 4 (*prioris* Uniformes), para  $f = 0,8$  e pelo modelo 3 (Uniforme - função degrau Uniforme), para  $f = 0,02$  e  $f = -0,217$ . Ayres & Balding (1998) e Armbrorst (2005) compararam, por meio de suas estimativas, dois métodos clássicos com o modelo Bayesiano com *prioris* Uniformes e concluíram que este modelo apresentou os melhores resultados.

#### 4.1.2.3 Avaliação da acurácia

Os resultados relativos à verificação da acurácia dos modelos Bayesianos por meio do vício, erro quadrático médio (EQM) e probabilidade de cobertura para cada cenário são apresentados na Tabela 11.

TABELA 11: Avaliação da acurácia, sendo  $\hat{f}$  a média das médias.

$f$	$n$	Modelo	$\hat{f}$	Vício	EQM	Prob. Cober.	
0,8	50	1	0,715	-0,085	0,071	1	
		2	0,745	-0,055	0,029	1	
		3	0,741	-0,059	0,034	1	
		4	0,740	-0,060	0,035	1	
	200	1	0,777	-0,023	0,005	1	
		2	0,785	-0,015	0,001	1	
		3	0,785	-0,015	0,002	1	
		4	0,784	-0,016	0,002	1	
	1000	1	0,816	0,016	0,002	1	
		2	0,818	0,018	0,003	1	
		3	0,818	0,018	0,003	1	
		4	0,818	0,018	0,003	1	
	0,02	50	1	0,062	0,042	0,018	1
			2	0,023	0,003	0,003	1
			3	0,026	0,006	0,005	1
			4	0,017	-0,003	0,002	1
200		1	0,007	-0,013	0,027	1	
		2	0,022	0,002	0,018	1	
		3	0,022	0,002	0,017	1	
		4	0,022	0,002	0,018	1	
1000		1	0,029	0,009	0,089	1	
		2	0,026	0,006	0,042	1	
		3	0,026	0,006	0,041	1	
		4	0,026	0,006	0,046	1	
-0,217		50	1	-0,062	0,155	0,238	1
			2	-0,049	0,168	0,282	1
			3	-0,045	0,172	0,295	1
			4	-0,074	0,143	0,203	1
	200	1	-0,177	0,040	0,015	1	
		2	-0,162	0,055	0,029	1	
		3	-0,160	0,057	0,031	1	
		4	-0,180	0,037	0,013	1	
	1000	1	-0,214	0,003	0,068	1	
		2	-0,214	0,003	0,078	1	
		3	-0,214	0,003	0,104	1	
		4	-0,214	0,003	0,067	1	

Assim, o modelo 1 apresentou as piores estimativas, ou seja, os maiores vícios e EQMs para os cenários de  $f = 0,8$  e  $f = 0,02$  e, juntamente com o modelo 4, obteve as melhores estimativas para  $f = -0,217$ , ou seja, os menores vícios e EQMs. Nota-se também que, à medida que  $n$  aumentava, as estimativas de  $f$  se tornavam mais próximas do seu valor verdadeiro. Em todos os casos, a probabilidade estimada de cobertura dos intervalos de credibilidade foi 1, significando que o valor verdadeiro de  $f$  esteve presente nestes intervalos, validando, assim, a metodologia

Resultados semelhantes são observados em Armbrorst (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 dos limites de  $f$ . Já em relação à probabilidade de cobertura, os resultados foram diferentes dos apresentados pela autora.

As Figuras de 5 a 7 representam as estimativas de  $f$  para cada modelo, considerando os três tamanhos amostrais ( $n=50;200;1000$ ), sendo que a linha pontilhada representa o valor verdadeiro de  $f$ . O que se observa é que, em todos os casos, as estimativas vão se aproximando do valor real, à medida que  $n$  aumenta.

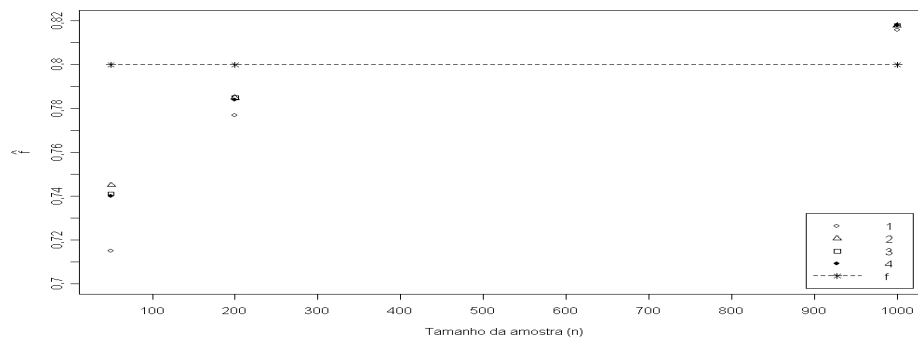


FIGURA 5: Gráfico de dispersão para o caso de  $f = 0,8$ .

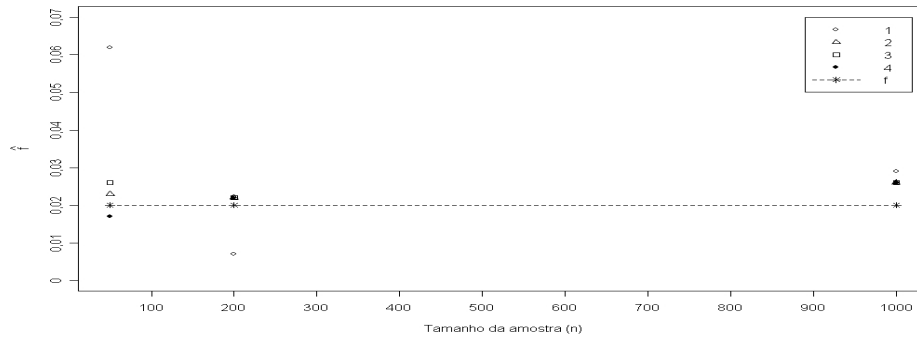


FIGURA 6: Gráfico de dispersão para o caso de  $f = 0,02$ .

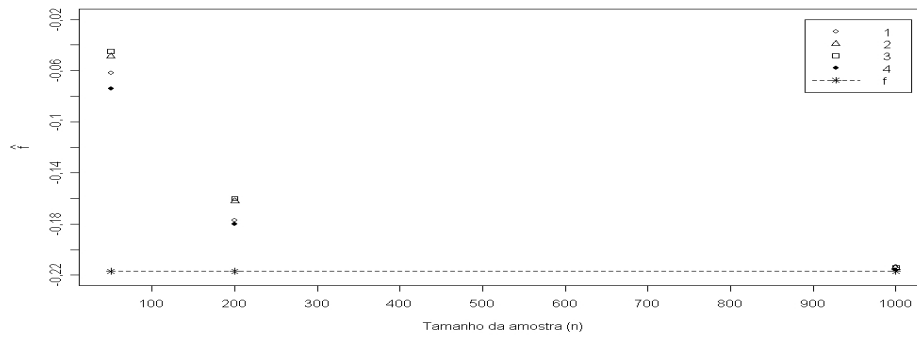


FIGURA 7: Gráfico de dispersão para o caso de  $f = -0,217$ .

#### 4.1.3 Parâmetro $D_A$

Os resultados referentes ao parâmetro  $D_A$  são apresentados nas Tabelas de 12 a 14. Esse parâmetro, por ser influenciado pelas estimativas da proporção alélica, apresentou resultados bem discrepantes em relação ao parâmetro  $f$ .



TABELA 12: Média ( $\hat{D}_{A_1}$ ), mediana ( $\hat{D}_{A_2}$ ), moda ( $\hat{D}_{A_3}$ ), desvio padrão e HPD, considerando o valor verdadeiro  $D_A = 0,146$ .

Modelo	$n$	$\hat{D}_{A_1}$	$\hat{D}_{A_2}$	$\hat{D}_{A_3}$	$DP$	$LI$	$LS$
1	50	0,075	0,076	0,079	0,033	0,010	0,141
	200	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
2	50	0,047	0,043	0,039	0,030	-0,001	0,102
	200	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
3	50	0,042	0,038	0,029	0,028	-0,001	0,094
	200	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
4	50	0,064	0,062	0,056	0,031	0,001	0,118
	200	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

TABELA 13: Média ( $\hat{D}_{A_1}$ ), mediana ( $\hat{D}_{A_2}$ ), moda ( $\hat{D}_{A_3}$ ), desvio padrão e HPD, considerando o valor verdadeiro  $D_A = 0,02$ .

Modelo	$n$	$\hat{D}_{A_1}$	$\hat{D}_{A_2}$	$\hat{D}_{A_3}$	$DP$	$LI$	$LS$
1	50	0,023	0,021	0,017	0,024	-0,020	0,071
	200	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
2	50	0,016	0,013	0,002	0,024	-0,031	0,064
	200	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
3	50	0,016	0,014	0,010	0,022	-0,022	0,063
	200	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
4	50	0,016	0,013	0,008	0,023	-0,028	0,066
	200	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

TABELA 14: Média ( $\hat{D}_{A_1}$ ), mediana ( $\hat{D}_{A_2}$ ), moda ( $\hat{D}_{A_3}$ ), desvio padrão e HPD, considerando o valor verdadeiro  $D_A = -0,02$ .

Modelo	$n$	$\hat{D}_{A_1}$	$\hat{D}_{A_2}$	$\hat{D}_{A_3}$	$DP$	$LI$	$LS$
1	50	-0,006	-0,009	-0,011	0,022	-0,046	0,041
	200	-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
2	50	-0,005	-0,009	-0,010	0,025	-0,054	0,049
	200	-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
3	50	-0,004	-0,007	-0,011	0,023	-0,047	0,049
	200	-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
4	50	-0,011	-0,014	-0,016	0,022	-0,056	0,037
	200	-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

Tendo em vista os resultados obtidos, observa-se que:

- o HPD não contém o valor verdadeiro do parâmetro, para alguns cenários, em  $D_A = 0,146$ ;
- todos os modelos apresentaram estimativas (média, mediana e moda) de  $D_A$  semelhantes, principalmente para os cenários de  $n=200$  e  $n=1000$ ;
- nota-se que a média, a mediana e a moda apresentaram valores muito próximos entre si. Para  $n=1000$ , eles foram também bem próximos ao valor verdadeiro de  $D_A$ , demonstrando o caráter simétrico da distribuição empírica;
- o processo de simulação propiciou uma melhor análise dos modelos, pois considerou vários cenários possíveis.

Os gráficos das Figuras de 8 a 10 representam a distribuição marginal *a posteriori* para o parâmetro  $D_A = 0,146$ , em cada modelo utilizado.

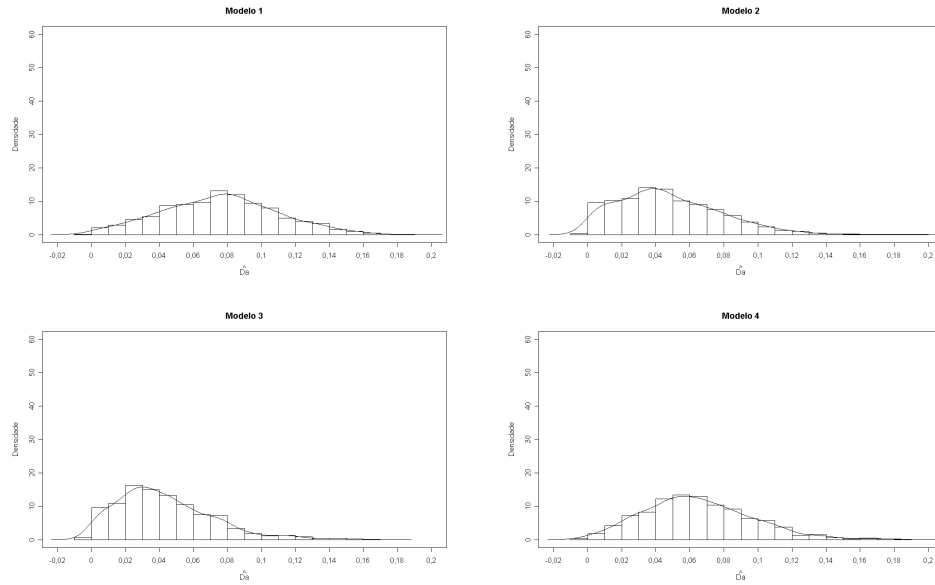


FIGURA 8: Gráfico da distribuição marginal *a posteriori* de  $D_A$  ( $n=50$ ).

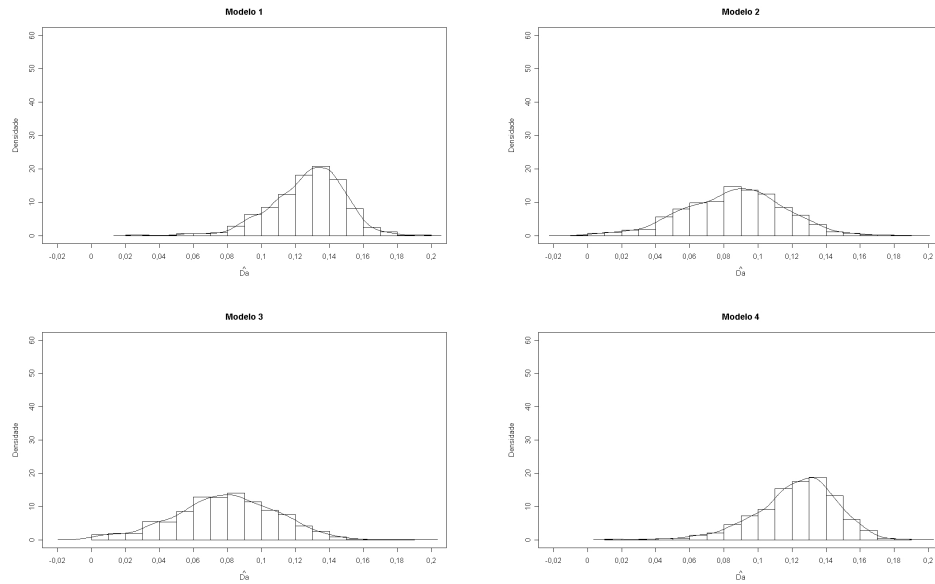


FIGURA 9: Gráfico da distribuição marginal *a posteriori* de  $D_A$  ( $n=200$ ).

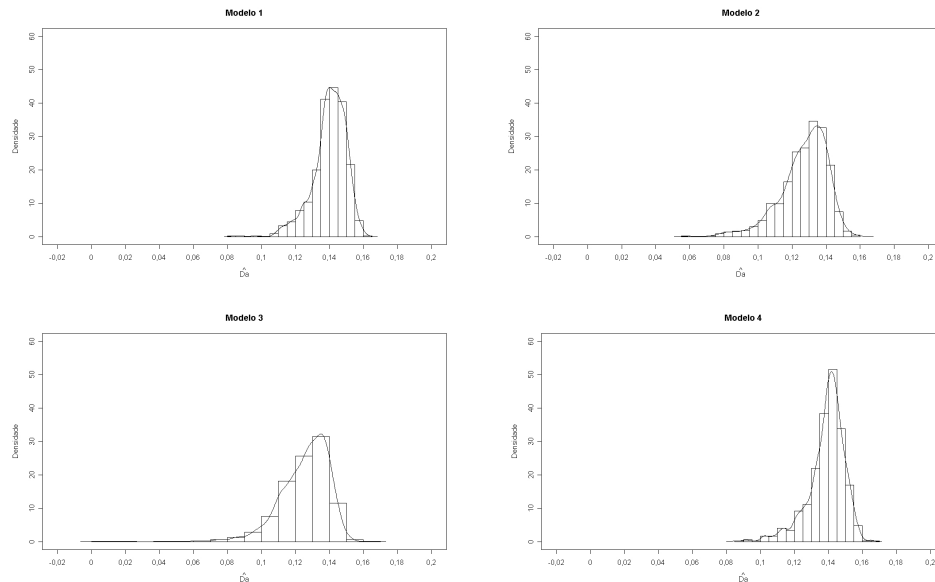


FIGURA 10: Gráfico da distribuição marginal *a posteriori* de  $D_A$  ( $n=1000$ ).

Para  $n=50$  ou  $n=200$ , as amplitudes de  $D_A$  são maiores, sendo que a concentração de valores  $D_A$  não fica em torno do valor verdadeiro, observado, principalmente, para o caso de  $n=50$ . Para  $n=1000$ , a amplitude de  $D_A$  diminui e a distribuição vai se tornando mais simétrica, com uma concentração de valores em torno de  $D_A$ . As demais figuras foram omitidas por apresentarem comportamento semelhante.

#### 4.1.3.1 Taxas de aceitação

Os dados da Tabela 15 propõem valores dos erros da proporção alélica e do coeficiente de desequilíbrio que possam ser utilizados nos 4 modelos aqui estudados em conjunto.

TABELA 15: Erros das proporções alélicas e do coeficiente de desequilíbrio.

$D_A$	$n$	$\varepsilon_p$	$\varepsilon_{D_A}$
0,146	50	0,3	0,08
	200	0,1	0,07
	1000	0,06	0,05
0,02	50	0,3	0,1
	200	0,1	0,07
	1000	0,05	0,04
-0,02	50	0,3	0,1
	200	0,1	0,08
	1000	0,05	0,04

Houve grande dificuldade em se atingir taxas de aceitação para  $p_A$  e  $D_A$  entre 20% a 50%. Realizou-se uma série de tentativas com o objetivo de determinar valores de  $\varepsilon_p$  e  $\varepsilon_{D_A}$  para cada um dos cenários. Vale ressaltar que, apenas para o cenário  $D_A=0,146$  e  $n=1000$ , as taxas de aceitação de  $p_A$  e  $D_A$  estiveram fora do intervalo de 20% a 50%, sendo estas de, aproximadamente, 15%.

#### 4.1.3.2 Fator de Bayes

Os resultados da comparação dos modelos por meio do fator de Bayes encontram-se na Tabela 16. Observa-se uma grande diferenciação de escolha de modelos para o parâmetro  $D_A$ , mostrando que este apresenta resultados bem discrepantes em relação ao parâmetro  $f$ .

TABELA 16: Fator de Bayes para o parâmetro  $D_A$ .

$D_A$	$n$	FB <sub>21</sub>	FB <sub>31</sub>	FB <sub>32</sub>	FB <sub>41</sub>	FB <sub>42</sub>	FB <sub>43</sub>
0,146	50	265,06	11,67	0,22	612,50	2,31	52,44
	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
0,02	50	0,36	140,58	520,12	224,08	829,05	1,59
	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
-0,02	50	1078,19	12,79	0,84	786,33	0,13	614,91
	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

Segundo esses resultados, o modelo 4 é o mais indicado para os cenários de  $D_A=0,146$  e  $D_A=0,02$ , obtendo, muitas vezes, evidência fraca e muito fraca a seu favor. Já para o cenário de  $D_A=-0,02$  e  $n=50$ , indica-se o modelo 2, para  $D_A=-0,02$  e  $n=200$  o modelo 4 e, para  $D_A=-0,02$  e  $n=1000$ , o modelo 3. Uma justificativa para esses resultados é que o coeficiente de desequilíbrio é um parâmetro relacionado às proporções alélicas e ao coeficiente de endogamia, portanto, totalmente dependente.

#### 4.1.3.3 Avaliação da acurácia

Os resultados relativos à verificação da acurácia dos modelos Bayesianos por meio do vício, do erro quadrático médio (EQM) e da probabilidade de cobertura para cada cenário são apresentados na Tabela 17.

TABELA 17: Avaliação da acurácia com a média das médias de  $\hat{D}_A$ .

$D_A$	$n$	Modelo	$\hat{D}_A$	Vício	EQM	Prob. Cober.
0,146	50	1	0,076	-0,070	0,049	0,9
		2	0,049	-0,097	0,094	0,9
		3	0,042	-0,104	0,107	0,9
		4	0,063	-0,083	0,068	0,9
	200	1	0,126	-0,020	0,003	1
		2	0,086	-0,060	0,036	0,9
		3	0,078	-0,068	0,046	0,9
		4	0,122	-0,024	0,005	1
	1000	1	0,142	-0,004	0,001	1
		2	0,130	-0,016	0,002	1
		3	0,129	-0,017	0,003	1
		4	0,141	-0,005	0,002	1
0,02	50	1	0,023	0,003	0,014	1
		2	0,016	-0,004	0,013	1
		3	0,015	-0,005	0,017	1
		4	0,016	-0,004	0,015	1
	200	1	0,017	-0,003	0,009	1
		2	0,014	-0,006	0,025	1
		3	0,014	-0,006	0,027	1
		4	0,014	-0,006	0,027	1
	1000	1	0,021	0,001	0,033	1
		2	0,021	0,001	0,018	1
		3	0,021	0,001	0,019	1
		4	0,021	0,001	0,020	1
-0,02	50	1	-0,007	0,013	0,016	1
		2	-0,001	0,019	0,019	1
		3	-0,005	0,015	0,022	1
		4	-0,010	0,010	0,008	1
	200	1	-0,021	-0,001	0,026	1
		2	-0,019	0,001	0,009	1
		3	-0,018	0,002	0,018	1
		4	-0,022	-0,002	0,066	1
	1000	1	-0,024	-0,004	0,020	1
		2	-0,025	-0,005	0,025	1
		3	-0,024	-0,004	0,024	1
		4	-0,025	-0,005	0,026	1

Verifica-se que, em todos os cenários, os modelos 1 e 4 tiveram as melhores estimativas, ou seja, os menores vícios e EQMs, tendo, para  $n=1000$ , sido muito próximas ao valor verdadeiro. Nota-se também que à medida que  $n$  aumentava, as estimativas de  $D_A$  se tornavam mais próximas do seu valor verdadeiro. A probabilidade estimada de cobertura, na maioria dos casos, foi 1, significando que o valor verdadeiro esteve presente nos intervalos de credibilidade e foi de 0,9 para os outros casos, confirmando a validação da metodologia para o parâmetro  $D_A$ .

Gráficos de dispersão apresentados nas Figuras 11 a 13 foram construídos com as estimativas de  $D_A$  para cada modelo, considerando os três tamanhos amostrais utilizados ( $n=50;200;1000$ ). A linha pontilhada reta representa o valor verdadeiro de  $D_A$ . O que se observa é que, para  $D_A=0,146$ , as estimativas para cada modelo são discrepantes se aproximando quando  $n=1000$ . Para  $D_A=0,02$ , essas estimativas sempre estiveram próximas ao valor real. Já para  $D_A=-0,02$ , nota-se que, quando  $n=200$ , as estimativas se localizam bem próximas ao valor real.

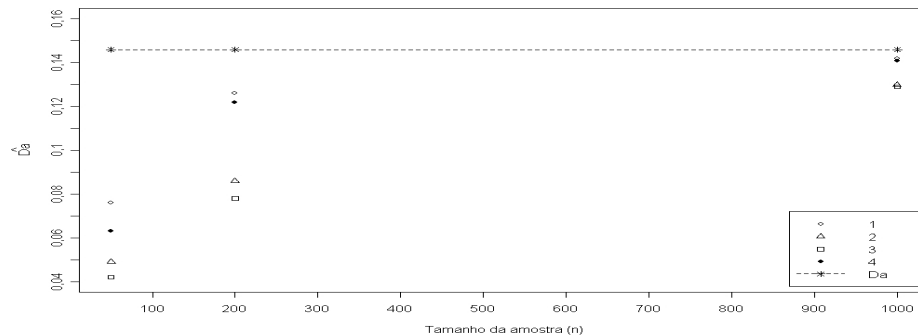


FIGURA 11: Gráfico de dispersão para o caso de  $D_A=0,146$ .



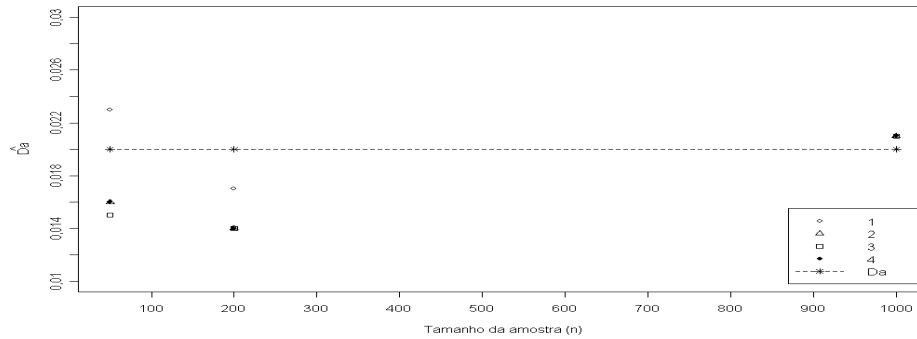


FIGURA 12: Gráfico de dispersão para o caso de  $D_A = 0,02$ .

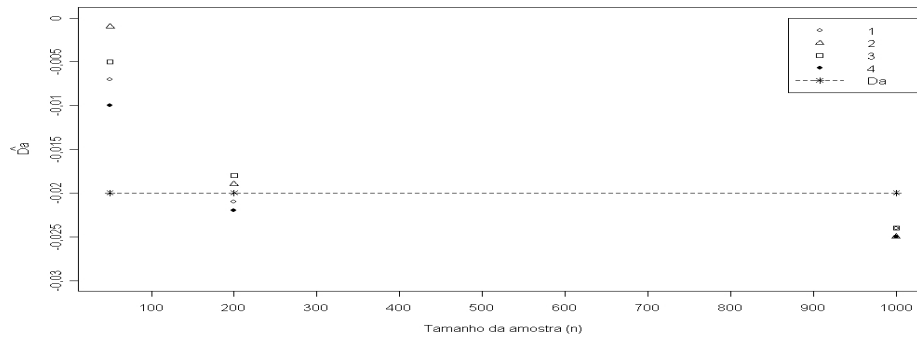


FIGURA 13: Gráfico de dispersão para o caso de  $D_A = -0,02$ .

## 4.2 Dados reais

### 4.2.1 Considerações gerais

Também foi utilizado, para os dados reais, um processo com 50.000 iterações, sendo descartadas as 10.000 iniciais, para o período de aquecimento da cadeia (*burn-in*). Para assegurar a independência da amostra, considerou-se um espaçamento entre os pontos amostrados de tamanho 40 (*thin*), ou seja, obteve-se uma amostra final de tamanho 1.000 para cada parâmetro.

A convergência das cadeias de todos os parâmetros do modelo foi monitorada por meio da visualização gráfica do traço e dos critérios disponíveis no pacote BOA do software livre R (R Development Core Team, 2007), não existindo evidências contra a convergência. Assim, pelo critério de Geweke, o p-valor estimado foi sempre maior que o nível de significância pré-fixado (5%) e, em relação ao critério de Gelman & Rubin, este sempre apresentou valores de  $\hat{R}$  próximos a 1. Foram consideradas nesta análise duas cadeias com diferentes valores iniciais. Sendo assim, foi possível estimar o coeficiente de endogamia e o coeficiente de desequilíbrio para os 21 conjuntos de dados reais relacionados, anteriormente, nas Tabelas 3 a 5.

#### 4.2.2 Parâmetro $f$

Os resultados referentes às inferências sobre  $f$  são apresentados nas Tabelas 18 a 20. Para uma melhor visualização, utilizaram-se as denominações dadas nas Tabelas 3 a 5:

Destaca-se que, para um grupo ser considerado em equilíbrio de Hardy-Weinberg, é necessário que suas estimativas estejam dentro do intervalo  $[-0,03;0,03]$ , como visto em (13). Ou seja, vários valores são definidos para a proporção alélica no intuito de se encontrar o limite inferior para o coeficiente de endogamia. Um teste de hipótese Bayesiano é avaliado como  $H_0 : f = [-0,03;0,03]$  (rejeita-se a hipótese de desequilíbrio) *versus*  $H_1 : f \neq [-0,03;0,03]$  (aceita-se a hipótese de desequilíbrio).

TABELA 18: Média ( $\hat{f}_1$ ), mediana ( $\hat{f}_2$ ), moda ( $\hat{f}_3$ ), desvio padrão e HPD, considerando o loco D7S8.

Grupo	Modelo	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$DP$	$LI$	$LS$
I	1	-0,143	-0,144	-0,141	0,079	-0,287	-0,016
	2	-0,155	-0,160	-0,168	0,078	-0,308	-0,005
	3	-0,155	-0,155	-0,154	0,076	-0,303	-0,011
	4	-0,156	-0,153	-0,138	0,076	-0,309	-0,017
II	1	-0,068	-0,068	-0,070	0,080	-0,221	0,084
	2	-0,075	-0,077	-0,072	0,080	-0,231	0,078
	3	-0,076	-0,079	-0,086	0,079	-0,239	0,072
	4	-0,081	-0,084	-0,101	0,076	-0,219	0,071
III	1	0,001	0,001	0,002	0,097	-0,196	0,172
	2	-0,020	-0,022	-0,045	0,101	-0,205	0,181
	3	-0,019	-0,021	-0,020	0,104	-0,227	0,174
	4	-0,012	-0,010	-0,001	0,099	-0,220	0,171
IV	1	0,012	0,008	-0,004	0,096	-0,174	0,190
	2	-0,004	-0,008	-0,013	0,100	-0,178	0,210
	3	-0,001	-0,001	-0,001	0,100	-0,203	0,188
	4	0,001	-0,001	0,006	0,100	-0,194	0,184
V	1	0,080	0,078	0,074	0,097	-0,110	0,267
	2	0,071	0,069	0,068	0,096	-0,103	0,271
	3	0,067	0,067	0,065	0,102	-0,127	0,277
	4	0,072	0,074	0,081	0,098	-0,097	0,280
VI	1	0,039	0,039	0,041	0,096	-0,134	0,237
	2	0,027	0,025	0,024	0,095	-0,154	0,214
	3	0,023	0,026	0,040	0,097	-0,171	0,209
	4	0,025	0,025	0,011	0,092	-0,153	0,201
VII	1	0,101	0,103	0,112	0,067	-0,019	0,234
	2	0,098	0,098	0,088	0,069	-0,036	0,237
	3	0,097	0,098	0,100	0,070	-0,053	0,225
	4	0,098	0,097	0,091	0,067	-0,036	0,227

TABELA 19: Média ( $\hat{f}_1$ ), mediana ( $\hat{f}_2$ ), moda ( $\hat{f}_3$ ), desvio padrão e HPD, considerando o loco GYPA.

Grupo	Modelo	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$DP$	$LI$	$LS$
I	1	0,003	0,001	-0,010	0,081	-0,147	0,158
	2	-0,003	-0,006	-0,014	0,084	-0,180	0,154
	3	-0,005	-0,005	-0,014	0,079	-0,156	0,154
	4	-0,002	-0,001	-0,002	0,080	-0,147	0,162
II	1	0,052	0,052	0,052	0,081	-0,105	0,211
	2	0,045	0,048	0,060	0,080	-0,099	0,209
	3	0,043	0,040	0,038	0,081	-0,103	0,202
	4	0,039	0,042	0,059	0,077	-0,100	0,188
III	1	0,187	0,190	0,172	0,098	-0,019	0,357
	2	0,176	0,171	0,168	0,097	-0,000	0,381
	3	0,178	0,183	0,208	0,098	-0,019	0,364
	4	0,178	0,178	0,182	0,099	-0,012	0,367
IV	1	0,109	0,108	0,104	0,099	-0,077	0,299
	2	0,099	0,098	0,102	0,100	-0,092	0,298
	3	0,099	0,097	0,085	0,103	-0,102	0,290
	4	0,098	0,100	0,131	0,100	-0,101	0,296
V	1	0,013	0,014	-0,013	0,093	-0,161	0,198
	2	0,002	0,001	0,004	0,098	-0,185	0,194
	3	0,004	0,004	0,012	0,097	-0,196	0,192
	4	0,002	0,003	0,005	0,099	-0,197	0,189
VI	1	0,055	0,058	0,067	0,093	-0,112	0,234
	2	0,048	0,043	0,035	0,096	-0,139	0,228
	3	0,044	0,043	0,023	0,096	-0,171	0,216
	4	0,045	0,047	0,061	0,092	-0,139	0,221
VII	1	-0,024	-0,023	-0,022	0,068	-0,154	0,108
	2	-0,030	-0,030	-0,035	0,069	-0,163	0,105
	3	-0,029	-0,031	-0,034	0,069	-0,157	0,118
	4	-0,030	-0,029	-0,023	0,066	-0,158	0,099

TABELA 20: Média ( $\hat{f}_1$ ), mediana ( $\hat{f}_2$ ), moda ( $\hat{f}_3$ ), desvio padrão e HPD, considerando o loco LDLR.

Grupo	Modelo	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$DP$	$LI$	$LS$
I	1	0,020	0,015	-0,015	0,081	-0,143	0,166
	2	0,006	0,003	-0,010	0,081	-0,139	0,166
	3	0,003	-0,001	0,005	0,081	-0,158	0,152
	4	0,007	0,003	0,006	0,082	-0,156	0,154
II	1	-0,098	-0,098	-0,089	0,078	-0,236	0,064
	2	-0,112	-0,111	-0,113	0,079	-0,281	0,031
	3	-0,113	-0,114	-0,142	0,080	-0,267	0,043
	4	-0,111	-0,114	-0,137	0,077	-0,264	0,038
III	1	0,100	0,101	0,110	0,098	-0,101	0,279
	2	0,096	0,097	0,080	0,099	-0,078	0,310
	3	0,096	0,096	0,105	0,102	-0,124	0,277
	4	0,094	0,096	0,093	0,097	-0,095	0,276
IV	1	-0,005	-0,004	-0,018	0,099	-0,192	0,198
	2	-0,020	-0,021	-0,007	0,096	-0,206	0,167
	3	-0,019	-0,023	-0,038	0,100	-0,216	0,163
	4	-0,014	-0,015	-0,019	0,094	-0,198	0,163
V	1	-0,201	-0,210	-0,227	0,087	-0,353	-0,03
	2	-0,182	-0,209	-0,231	0,126	-0,397	0,107
	3	-0,177	-0,204	-0,221	0,124	-0,400	0,087
	4	-0,210	-0,230	-0,245	0,112	-0,376	0,028
VI	1	-0,083	-0,084	-0,100	0,092	-0,262	0,096
	2	-0,095	-0,096	-0,093	0,094	-0,277	0,087
	3	-0,095	-0,096	-0,086	0,094	-0,263	0,099
	4	-0,094	-0,095	-0,096	0,094	-0,290	0,074
VII	1	0,048	0,049	0,048	0,069	-0,082	0,181
	2	0,043	0,042	0,033	0,070	-0,098	0,175
	3	0,039	0,040	0,045	0,069	-0,109	0,174
	4	0,043	0,043	0,039	0,065	-0,089	0,160

Observa-se que:

- na maioria das vezes, os quatro modelos apresentaram estimativas muito próximas entre si para a média, a mediana e a moda;
- de acordo com os dados da Tabela 18, todos os grupos no loco D7S8 estão em EHW, destacando-se que, para os grupos III e IV, esse

- equilíbrio é visualizado tanto pelo HPD quanto pelas suas estimativas, muito próximas a zero. Outro detalhe é que o HPD dos grupos I e VII não contém totalmente o intervalo de equilíbrio;
- c) de acordo com a Tabela 19, todos os grupos no loco GYPA estão em EHW, destacando-se que, para os grupos I e V esse equilíbrio é visualizado tanto pelo HPD quanto pelas suas estimativas, muito próximas a zero. Já no grupo III, o HPD de todos os modelos não contém o intervalo de equilíbrio;
  - d) na Tabela 20, todos os grupos no loco LDLR estão em EHW, destacando-se que, para os grupos I e IV, esse equilíbrio é visualizado tanto pelo HPD quanto pelas suas estimativas, muito próximas a zero. Já no grupo V, o HPD do modelo 4 não contém o intervalo de equilíbrio.

Armborst (2005) utilizou *prioris* Uniformes e estimou estes mesmos parâmetros para dados de parasitologia e tribos indígenas, verificando que os modelos Bayesianos respeitam o espaço paramétrico. Trabalhando com dados de nativos da ilha de Salomão, Ayres & Balding (1998) chegaram às mesmas conclusões. Quanto às outras *prioris*, resultados semelhantes podem ser observados em Shoemaker et al. (1998), utilizando a probabilidade de as estimativas dos parâmetros se localizarem em intervalos de equilíbrio.

Os gráficos das Figuras de 14 a 16 representam a distribuição marginal *a posteriori* para o parâmetro  $f$  em cada modelo. Consideraram-se, dentre os 21 conjuntos de dados reais, os três que apresentavam características mais próximas de desequilíbrio de Hardy-Weinberg.

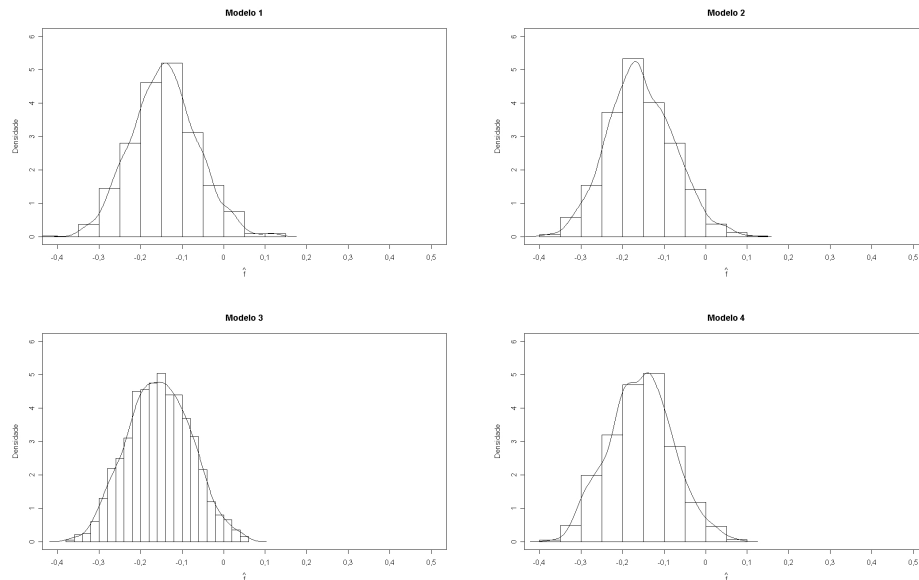


FIGURA 14: Gráfico da distribuição marginal *a posteriori* de  $f$ , para o grupo afro-americano do loco D7S8 (FBI).

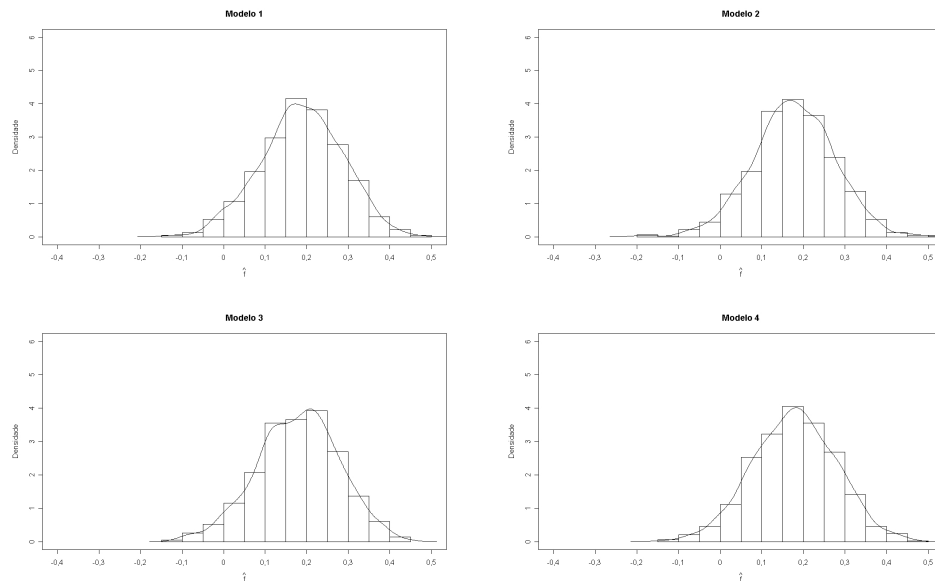


FIGURA 15: Gráfico da distribuição marginal *a posteriori* de  $f$ , para o grupo hispânico do sudeste no loco GYP A (FBI).

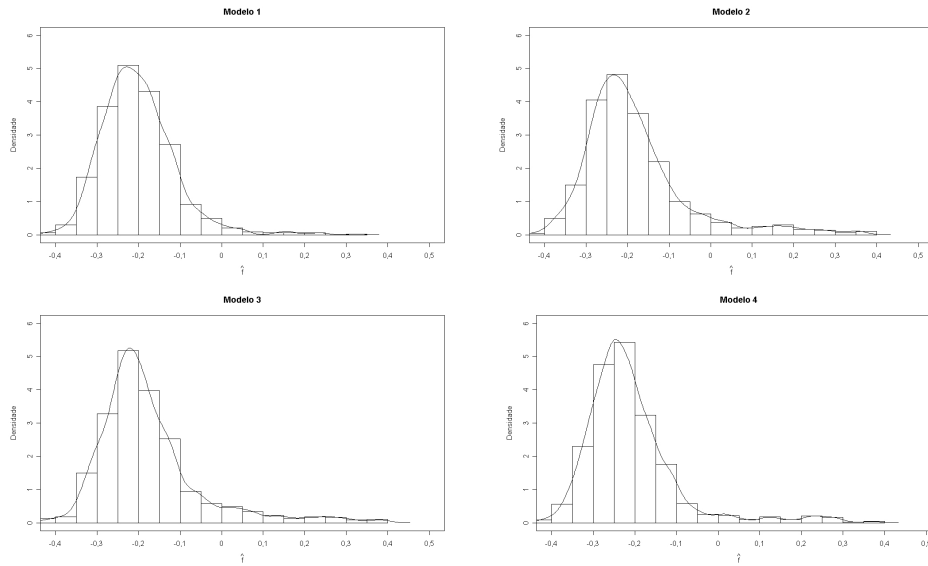


FIGURA 16: Gráfico da distribuição marginal *a posteriori* de  $f$ , para o grupo afro-americano no loco LDLR (*Cellmark*).

Esses gráficos apresentam aspecto de uma distribuição simétrica com valores concentrados afastando-se do intervalo de equilíbrio. Já na Figura 16 é apresentado o aspecto de distribuição assimétrica, sendo que, para este caso, os valores se afastam ainda mais do intervalo de equilíbrio.

#### 4.2.2.1 Taxas de aceitação

Na Tabela 21 são apresentados os seguintes  $\varepsilon_p$  e  $\varepsilon_f$  para cada um dos conjuntos de dados reais, os quais não diferiram de loco para loco.

TABELA 21: Erros das proporções alélicas e do coeficiente de endogamia.

n	93	97	100	103	145	148	200
$\varepsilon_p$	0,13	0,13	0,13	0,12	0,10	0,10	0,09
$\varepsilon_f$	0,70	0,70	0,70	0,63	0,50	0,50	0,43



O que se observa é que, à medida que  $n$  aumenta, os valores de  $\varepsilon_p$  e  $\varepsilon_f$  tendem a diminuir.

#### 4.2.2.2 Fator de Bayes

Os resultados da comparação dos modelos por meio do fator de Bayes encontram-se na Tabela 22.

TABELA 22: Fator de Bayes para o parâmetro  $f$ .

Grupo	Loco	FB <sub>12</sub>	FB <sub>13</sub>	FB <sub>14</sub>	FB <sub>32</sub>	FB <sub>42</sub>	FB <sub>43</sub>
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

Para todos os conjuntos de dados, o modelo mais indicado é o 1, pois apresentou evidência muito forte a seu favor em relação a todos os outros. Estes,

por sua vez, apresentaram evidência muito fraca ou fraca entre si para  $f$ . Estes resultados concordam com o processo de simulação.

Resultados diferentes foram encontrados por Shoemaker et al. (1998), quando comparadas as *prioris* Dirichlet, Beta - função degrau Uniforme e Uniforme - função degrau Uniforme. Segundo esses autores, utilizando resultados referentes à probabilidade de o parâmetro estar ou não em um intervalo de EHW, as duas últimas *prioris* apresentaram valores próximos e bem diferentes da primeira *priori*, não sendo utilizada nenhuma forma de comparação em específico. Ayres & Balding (1998) e Armbrorst (2005) concluíram que o modelo Bayesiano com *prioris* Uniformes apresentou resultados mais satisfatórios em comparação com os métodos clássicos.

#### 4.2.3 Parâmetro $D_A$

Os resultados iniciais relativos às inferências sobre  $D_A$  são apresentados nas Tabelas 23 a 25. Para uma melhor visualização, utilizam-se as mesmas denominações sugeridas nas Tabelas 3 a 5. Destaca-se que, para que um grupo seja considerado em equilíbrio de Hardy-Weinberg, é necessário que suas estimativas estejam dentro do intervalo  $[-0,007;0,007]$ , como visto em (14), ou seja, vários valores são definidos para a proporção alélica no intuito de se encontrar o limite inferior e o superior para o coeficiente de desequilíbrio. Um teste de hipótese Bayesiano é avaliado como  $H_0 : D_A = [-0,007;0,007]$  (rejeita-se a hipótese de desequilíbrio) *versus*  $H_1 : D_A \neq [-0,007;0,007]$  (aceita-se a hipótese de desequilíbrio).

TABELA 23: Média ( $\hat{D}_{A_1}$ ), mediana ( $\hat{D}_{A_2}$ ), moda ( $\hat{D}_{A_3}$ ), desvio padrão e HPD, considerando o loco D7S8.

Grupo	Modelo	$\hat{D}_{A_1}$	$\hat{D}_{A_2}$	$\hat{D}_{A_3}$	$DP$	$LI$	$LS$
I	1	-0,034	-0,034	-0,028	0,019	-0,068	0,004
	2	-0,036	-0,037	-0,039	0,018	-0,072	-0,001
	3	-0,036	-0,037	-0,040	0,019	-0,073	0,001
	4	-0,037	-0,037	-0,035	0,018	-0,070	0,001
II	1	-0,016	-0,015	-0,014	0,019	-0,054	0,017
	2	-0,017	-0,017	-0,012	0,018	-0,055	0,018
	3	-0,018	-0,019	-0,019	0,018	-0,054	0,019
	4	-0,018	-0,018	-0,016	0,017	-0,050	0,018
III	1	-0,001	-0,001	0,006	0,024	-0,045	0,045
	2	-0,004	-0,004	-0,007	0,024	-0,052	0,041
	3	-0,003	-0,003	-0,009	0,024	-0,047	0,051
	4	-0,004	-0,004	-0,002	0,023	-0,051	0,039
IV	1	0,003	0,002	-0,005	0,021	-0,034	0,051
	2	-0,001	-0,001	-0,001	0,022	-0,045	0,042
	3	0,001	0,001	0,001	0,022	-0,042	0,041
	4	0,001	-0,001	0,001	0,021	-0,039	0,039
V	1	0,018	0,018	0,020	0,022	-0,025	0,064
	2	0,015	0,015	0,016	0,022	-0,025	0,059
	3	0,015	0,015	0,015	0,021	-0,025	0,057
	4	0,015	0,014	0,010	0,020	-0,026	0,056
VI	1	0,008	0,008	0,016	0,023	-0,038	0,048
	2	0,005	0,005	0,003	0,023	-0,041	0,049
	3	0,007	0,005	0,002	0,023	-0,040	0,051
	4	0,006	0,005	0,007	0,022	-0,036	0,052
VII	1	0,024	0,023	0,025	0,016	-0,006	0,056
	2	0,022	0,022	0,027	0,016	-0,010	0,054
	3	0,022	0,022	0,019	0,017	-0,011	0,054
	4	0,021	0,021	0,019	0,015	-0,005	0,051

TABELA 24: Média ( $\hat{D}_{A_1}$ ), mediana ( $\hat{D}_{A_2}$ ), moda ( $\hat{D}_{A_3}$ ), desvio padrão e HPD, considerando o loco GYPA.

Grupo	Modelo	$\hat{D}_{A_1}$	$\hat{D}_{A_2}$	$\hat{D}_{A_3}$	$DP$	$LI$	$LS$
I	1	0,001	0,001	-0,003	0,020	-0,038	0,036
	2	-0,001	-0,001	-0,001	0,020	-0,038	0,043
	3	-0,001	-0,001	-0,002	0,020	-0,044	0,039
	4	-0,001	-0,001	-0,001	0,019	-0,039	0,036
II	1	0,012	0,012	0,009	0,019	-0,023	0,047
	2	0,010	0,011	0,005	0,019	-0,028	0,047
	3	0,010	0,010	0,011	0,020	-0,030	0,052
	4	0,009	0,010	0,012	0,019	-0,025	0,049
III	1	0,044	0,044	0,051	0,024	-0,002	0,088
	2	0,044	0,045	0,052	0,024	-0,004	0,088
	3	0,044	0,043	0,039	0,025	-0,008	0,088
	4	0,043	0,043	0,048	0,0243	-0,004	0,090
IV	1	0,026	0,026	0,028	0,023	-0,018	0,069
	2	0,022	0,0227	0,022	0,023	-0,021	0,066
	3	0,023	0,023	0,024	0,023	-0,022	0,068
	4	0,021	0,021	0,0121	0,021	-0,023	0,061
V	1	0,002	0,002	0,003	0,024	-0,042	0,047
	2	0,001	0,001	0,005	0,023	-0,048	0,044
	3	0,001	0,001	-0,003	0,024	-0,054	0,041
	4	0,001	0,001	0,004	0,023	-0,039	0,051
VI	1	0,013	0,013	0,018	0,024	-0,029	0,062
	2	0,012	0,012	0,006	0,023	-0,037	0,055
	3	0,012	0,011	0,011	0,024	-0,036	0,059
	4	0,011	0,010	0,013	0,023	-0,031	0,057
VII	1	0,125	0,127	0,132	0,023	0,078	0,169
	2	0,085	0,087	0,094	0,029	0,028	0,143
	3	0,077	0,078	0,075	0,031	0,010	0,132
	4	0,120	0,124	0,132	0,024	0,068	0,163

TABELA 25: Média ( $\hat{D}_{A_1}$ ), mediana ( $\hat{D}_{A_2}$ ), moda ( $\hat{D}_{A_3}$ ), desvio padrão e HPD, considerando o loco LDLR.

Grupo	Modelo	$\hat{D}_{A_1}$	$\hat{D}_{A_2}$	$\hat{D}_{A_3}$	$DP$	$LI$	$LS$
I	1	0,003	0,003	0,005	0,014	-0,025	0,033
	2	0,001	0,001	0,001	0,014	-0,027	0,028
	3	0,001	0,001	0,001	0,014	-0,028	0,029
	4	0,001	0,001	0,001	0,014	-0,025	0,030
II	1	-0,025	-0,025	-0,023	0,019	-0,061	0,010
	2	-0,027	-0,027	-0,028	0,019	-0,067	0,010
	3	-0,027	-0,027	-0,033	0,020	-0,071	0,009
	4	-0,029	-0,028	-0,026	0,019	-0,060	0,011
III	1	0,0253	0,026	0,030	0,024	-0,018	0,073
	2	0,0237	0,024	0,024	0,023	-0,026	0,068
	3	0,0228	0,021	0,015	0,025	-0,032	0,068
	4	0,0219	0,020	0,017	0,023	-0,018	0,073
IV	1	-0,002	-0,003	-0,004	0,024	-0,046	0,042
	2	-0,004	-0,004	-0,007	0,023	-0,053	0,039
	3	-0,004	-0,005	-0,009	0,024	-0,053	0,044
	4	-0,005	-0,004	-0,001	0,023	-0,047	0,042
V	1	-0,037	-0,039	-0,040	0,020	-0,080	0,021
	2	-0,030	-0,033	-0,035	0,025	-0,076	0,028
	3	-0,029	-0,032	-0,033	0,025	-0,078	0,025
	4	-0,040	-0,042	-0,048	0,021	-0,076	0,014
VI	1	-0,020	-0,020	-0,014	0,022	-0,064	0,021
	2	-0,023	-0,023	-0,024	0,022	-0,066	0,019
	3	-0,023	-0,024	-0,021	0,023	-0,069	0,023
	4	-0,024	-0,023	-0,020	0,022	-0,063	0,022
VII	1	0,011	0,011	0,015	0,017	-0,024	0,042
	2	0,010	0,010	0,011	0,017	-0,020	0,048
	3	0,010	0,009	0,006	0,017	-0,023	0,047
	4	0,009	0,009	0,010	0,016	-0,018	0,044

Observa-se, então, que:

- todos os quatro modelos apresentaram estimativas muito próximas, tanto para a média, como para a mediana e a moda;
- na Tabela 23, todos os grupos no loco D7S8 estão em EHW, destacando-se que, para os grupos III, IV e VI, esse equilíbrio é

visualizado tanto pelo HPD quanto pelas suas estimativas, muito próximas a zero. Para o grupo VII, o HPD dos modelos 1 e 4 não contém o intervalo do equilíbrio. Já no grupo I, o HPD de todos os modelos não contém o intervalo de equilíbrio;

- c) na Tabela 24, todos os grupos no loco GYPA estão em EHW, destacando-se que, para os grupos I e V, esse equilíbrio é visualizado tanto pelo HPD quanto pelas suas estimativas, muito próximas a zero. Já no grupo III, o HPD de todos os modelos não contém o intervalo de equilíbrio;
- d) Na Tabela 25, todos os grupos no loco LDLR estão em EHW, destacando-se que, para os grupos I e IV, esse equilíbrio é visualizado tanto pelo HPD quanto pelas suas estimativas, muito próximas a zero.

Shoemaker et al. (1998) utilizaram as *prioris* Dirichlet, Beta - função de grau Uniforme e Uniforme - função de grau Uniforme para estas análises e os resultados obtidos foram semelhantes.

Os gráficos das Figuras 17 a 19 representam a distribuição marginal *a posteriori* para o parâmetro  $D_A$  em cada modelo. Foram considerados aqui três conjuntos de dados que apresentavam as maiores características de desequilíbrio de Hardy-Weinberg. Essas apresentam aspecto de uma distribuição simétrica com valores concentrados afastando-se do intervalo de equilíbrio. Já na Figura 19, apresenta-se o aspecto de distribuição assimétrica, sendo que para este caso, os valores se afastam ainda mais do intervalo de equilíbrio.

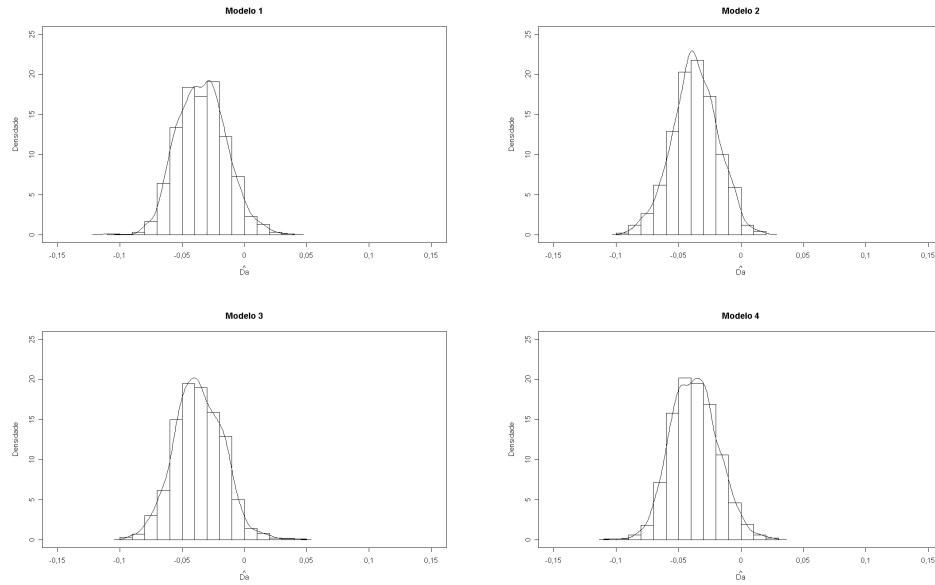


FIGURA 17: Gráfico da distribuição marginal *a posteriori* de  $D_A$  para o grupo afro-americano no loco D7S8 (FBI).

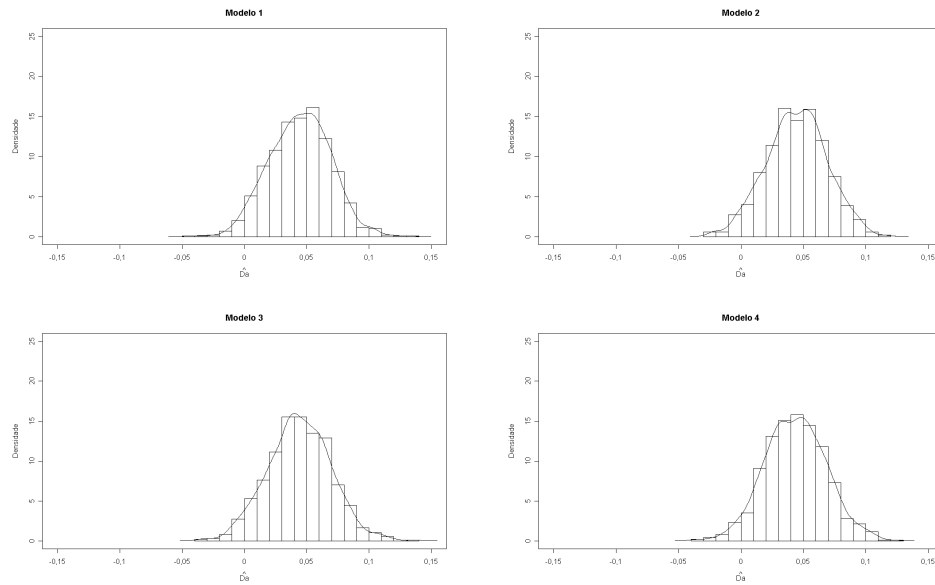


FIGURA 18: Gráfico da distribuição marginal *a posteriori* de  $D_A$  para o grupo hispânico do sudeste no loco GYPA (FBI).

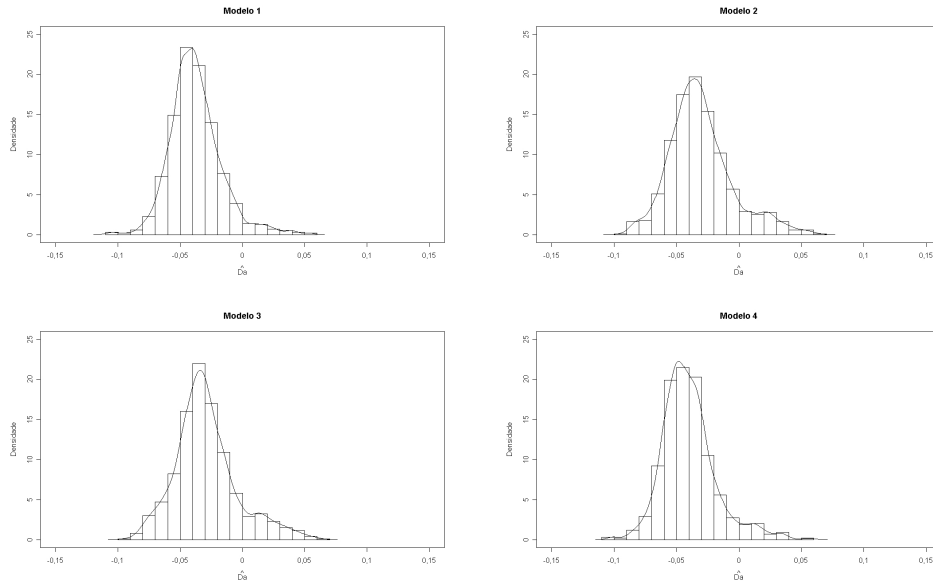


FIGURA 19: Gráfico da distribuição marginal *a posteriori* de  $D_A$  para o grupo afro-americano no loco LDLR (*Cellmark*).

#### 4.2.3.1 Taxas de aceitação

Foram utilizados os seguintes valores de  $\varepsilon_p$  e  $\varepsilon_{D_A}$ , apresentados na Tabela 26, para cada um dos conjuntos de dados reais, os quais não diferiram de loco para loco. À medida que  $n$  aumenta, as medidas de  $\varepsilon_p$  e  $\varepsilon_{D_A}$  tendem a diminuir.

TABELA 26: Erros das proporções alélicas e do coeficiente de desequilíbrio

n	93	97	100	103	145	148	200
$\varepsilon_p$	0,15	0,15	0,15	0,15	0,13	0,13	0,1
$\varepsilon_{D_A}$	0,11	0,11	0,11	0,11	0,09	0,09	0,08



#### 4.2.3.2 Fator de Bayes

Os resultados da comparação dos modelos por meio do fator de Bayes encontram-se na Tabela 27.

TABELA 27: Fator de Bayes para o parâmetro  $D_A$ .

Grupo	Loco	FB <sub>21</sub>	FB <sub>31</sub>	FB <sub>32</sub>	FB <sub>41</sub>	FB <sub>42</sub>	FB <sub>43</sub>
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

Observa-se que os modelos 2, 3 e 4, na maioria das vezes, apresentaram evidências forte, muito forte, fraca ou muito fraca a seu favor, quando comparados com o modelo 1. Quando comparados entre si, os resultados foram os mais diversos possíveis. Para cada grupo racial em um determinado loco pode ser indicado um desses modelos e apenas o modelo 1 não pode ser indicado para nenhum dos conjuntos de dados reais.

Resultados semelhantes foram encontrados por Shoemaker et al. (1998), quando comparadas as *prioris* Dirichlet, Beta - função degrau Uniforme e Uniforme - função degrau Uniforme. Segundo esses autores, utilizando resultados referentes à probabilidade do parâmetro estar ou não em um intervalo de EHW, as duas últimas *prioris* apresentaram valores próximos e bem diferentes da primeira *priori*, não sendo utilizada nenhuma forma de comparação em específico.

### 4.3 Considerações finais

Os parâmetros  $f$  e  $D_A$  propiciaram resultados diferentes em relação à escolha do melhor modelo, justificando, assim, o uso desses dois parâmetros, cabendo ao pesquisador a utilização de um ou outro em suas análises. Ressalta-se também que os modelos indicados, em cada cenário e em cada conjunto de dados reais, foram os mesmos tanto no processo de simulação como para os dados reais. Outro detalhe é a importância da avaliação da metodologia por um processo de simulação e a utilização de métodos mais específicos, como o fator de Bayes, para a escolha do melhor modelo, não sendo os mesmos usados no trabalho de Shoemaker et al. (1998).

## 5 CONCLUSÕES

Diante dos resultados obtidos, pode-se concluir que:

- a) Para o parâmetro  $f$ , o modelo mais indicado aos dados simulados e também aos dados reais foi o modelo 1 (*priori* Dirichlet) e, para o parâmetro  $D_A$ , o único modelo não indicado foi o 1. Para cada caso do processo de simulação e para cada grupo dos dados reais, um modelo, em específico, pode ser indicado, sendo os modelos 2 (*priori* Beta - função degrau Uniforme), 3 (*priori* Uniforme - função degrau Uniforme) e 4 (*prioris* Uniformes).
- b) A metodologia Bayesiana mostrou-se eficiente, sendo esta avaliada e comprovada pelo processo de simulação, que apresentou estimativas sempre bem próximas ao valor verdadeiro do parâmetro.
- c) A análise dos dados reais foi condizente com os resultados já obtidos pela literatura, podendo a população em estudo ser caracterizada como em equilíbrio de Hardy-Weinberg.

Em estudos futuros, recomenda-se trabalhar com um número maior de alelos ( $k > 2$ ), analisar outros parâmetros referentes ao desequilíbrio de Hardy-Weinberg, como aqueles citados em Pereira & Rogatko (1984) e Lindley (1988) e verificar neles a influência do parâmetro proporção alélica ( $p_A$ ). No tocante à inferência Bayesiana, novas distribuições *a priori* informativas ou não poderiam ser avaliadas e uma fórmula geral para o erro do coeficiente de desequilíbrio ( $\epsilon_{D_A}$ ) poderia ser definida, evitando-se assim, grandes esforços na obtenção de taxas de aceitação plausíveis.

## 6 REFERÊNCIAS BIBLIOGRÁFICAS

ARMBORST, T. **Métodos para medir o desequilíbrio de Hardy-Weinberg através de medidas de endocruzamento**. 2005. 187 p. 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, June 1998.

AYRES, K. L.; BALDING, D. J. Measuring gametic disequilibrium from multi-locus data. **Genetics**, Bethesda, v.157, n. 1, p. 413-423, Jan. 2001.

BALDING, D. J.; NICHOLS, R. A. Significant genetic correlations among Caucasian at forensic DNA loci. **Heredity**, Oxford, v. 78, n. 6, p. 583-589, June 1997.

BEAUMONT, M. A.; RANNALA, B. The Bayesian revolution in Genetics. **Genetics**, Bethesda, v. 5, n. 4, p. 251-261, Apr. 2004.

BEIGUELMAN, B. **Genética de populações humanas**. 2005. 230 p. Disponível em: <[http://lineu.icb.usp.br/bbeiguel/Genetica Populacoes](http://lineu.icb.usp.br/bbeiguel/Genetica%20Populacoes)>. Acesso em: 02 maio 2005.

BOLFARINE, H.; SANDOVAL, M. C. **Introdução à inferência estatística**. Rio de Janeiro: Sociedade Brasileira de Matemática, 2001. 125 p.

BOX, G. E. P.; TIAO, G. C. **Bayesian inference in statistical analysis**. New York: J. Wiley, 1992. 603 p.

CHEN, J. J.; THOMSON, G. The variance for the disequilibrium coefficient in the individual Hardy-Weinberg test. **Biometrics**, Arlington, v. 55, n. 4, p. 1269-1272, Dec. 1999.

CHOW, M.; FONG, D. K. H. Simultaneous estimation of the Hardy-Weinberg proportions. **The Canadian Journal of Statistics**, Ottawa, v. 20, n. 3, p. 291-296, Sept. 1992.

COCKERHAM, C. C. Variance of gene frequencies. **Evolution**, Lancaster, v. 23, n. 1, p. 72-84, Mar. 1969.

COCKERHAM, C. C.; WEIR, B. S. Variance of actual inbreeding. **Theoretical Population Biology**, San Diego, v. 23, n. 1, p. 85-109, Feb. 1983.

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

EHLERS, R. S. **Introdução à inferência Bayesiana**. UFPR. Departamento de Estatística. Disponível em: <<http://leg.ufpr.br/~ehlers/notas/bayes.pdf>>. Acesso em: 21 nov. 2007.

EMIGH, T. H. A comparison of tests for Hardy-Weinberg equilibrium. **Biometrics**, Arlington, v. 36, n. 4, p. 627-642, Dec. 1980.

FALCONER, D. S. **Introduction to quantitative genetics**. 3<sup>rd</sup>ed. New York: Logman Scientific & Technical, 1989. 438 p.

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

GAMERMAN, D. **Markov Chain Monte Carlo - stochastic simulation for bayesian inference**. London: Chapman & Hall, 1997. 245 p.

GARDNER, E. J. **Genética**. 5<sup>nd</sup>ed. Utah: Interamericana/Utah State University, 1977. 515 p.

GELFAND, A. E. Model determination using sampling based methods. In: GILKS, W. R.; RICHARDSON, S.; SPIEGELHALTER, D. J. (Ed.). **Markov chain monte carlo in practice**. London: Chapman & Hall, 1996. p. 145-162.

GELMAN, A.; CARLIN, J. B.; STER, H. S.; RUBIN, D. B. **Bayesian data analysis**. Boca Raton: Chapman & Hall/CRC, 2000. 526 p.

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

GEWEKE, J. Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In: BERNARDO, J. M.; BERGER, J. O.; DAWID, A.P.; SMITH, A.F.M. (Ed.). **Bayesian statistics**. New York: Oxford University, 1992. p. 625-631.

GILKS, W. R.; RICHARDSON, S.; SPIGELHATER, D. J. . **Markov chain monte carlo in practice**. London: Chapman & Hall, 1996. 481 p.

GRIFFITHS, A. J. F.; MILLER, J. H.; SUZUKI, D. T.; LEWONTIN, R. C.; GELBART, W. M. **An Introduction to genetic analysis**. 6<sup>nd</sup>ed. New York: W.H. Freeman & Company, 1996. 915 p.

GUO, S. W.; THOMPSON, E. A. Performing the exact test of Hardy-Weinberg proportion for multiple aleles. **Biometrics**, Arlington, v. 48, n. 6, p. 361-372, June 1992.

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

HEIDELBERG, P.; WELCH, P. Simulation run lenght control in the presence of an initial transient. **Operations Research**, Landing, v. 31, n. 6, p. 1109-1144, Nov./Dec. 1993.

HERNÁNDEZ, J. L.; WEIR, B. S. A disequilibrium approach to Hardy-Weinberg testing. **Biometrics**, Arlington, v. 45, n. 1, p. 53-70, Mar. 1989.

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

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

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, June 1995.

LEANDRO, R.A. **Introdução à estatística Bayesiana**. Piracicaba, SP: Universidade de São Paulo. Escola Superior de Agricultura “Luiz de Queiroz”. Departamento de Ciências Exatas, 2001. 51 p.

LEUTENEGGER, A. L.; PRUM, B.; GENIN, E.; VERNY, C.; LEMAINQUE, A.; CLERGET-DARPOUX, F.; THOMPSON, E. A. Estimation of inbreeding coefficient trough use of genomic data. **The American Journal Human Genetics**, Chicago, v. 73, n. 3, p. 516-523, Sept. 2003.

LINDLEY, D. Statistical inference concerning Hardy-Weinberg equilibrium. **Bayesian Statistics**, Oxford, v. 3, n. 1, p. 307-326, Jan. 1988.

LONG J. C.; KITTLES R. A. Human genetic diversity and the nonexistence of biological races. **Human Biology**, Wayne, v. 75, n. 4, p. 449-471, Aug. 2003.

McGUIRE, G.; DENHAM, M. C.; BALDING, D. J. Models of evolution for DNA sequences including gaps. **Molecular Biology and Evolution**, Oxford, v. 18, n. 4, p. 481-490, Apr. 2001.

METROPOLIS, N.; ROSENBLUTH, A. W.; ROSENBLUTH, M. N.; TELLER, A. H.; TELLER, E. Equations of state calculations by fast computing machines. **Journal of Chemical Physics**, Chicago, v. 21, n. 6, p. 1087-1092, June 1953.

MONTOYA-DELGADO, L. E.; IRONY, T. Z.; PEREIRA, C. A. B; WHITTLE, M. R. An unconditional exact test for the Hardy-Weinberg law: sample-space ordering using the Bayes factor. **Genetics**, Bethesda, v. 158, n. 2, p. 875-883, June 2001.

MUNIZ, J. A.; BARBIN, D.; VENCOVSKY, R. 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**, Ribeirão Preto, v. 19, n. 3, p. 485-491, July 1996.

MUNIZ, J. A.; BARBIN, D.; VENCOVSKY, R. A variância do estimador do coeficiente de endogamia obtido pelo método dos momentos em uma população diplóide. **Revista de Matemática e Estatística**, São Paulo, v. 15, p. 131-143, 1997.

MUNIZ, J. A.; BARBIN, D.; VENCOVSKY, R.; VEIGA, R. D. Teste de hipótese sobre o coeficiente de endogamia de uma população diplóide. **Ciência e Agrotecnologia**, Lavras, v. 23, n. 2, p. 410-420, abr./jun. 1999.

NATIONAL RESEARCH COUNCIL. **The Evaluation of Forensic DNA Evidence**. Washington: National Academy, 1996. 328 p.

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

NOGUEIRA, D. A.; SÁFADI, T. ; FERREIRA, D. F. 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.

OLSEN, S. **Mapping human history: genes, race, and our common origins.** New York: Houghton Mifflin, 2003. 308 p.

PAULINO, C. D.; TURKMAN, M. A. A.; MURTEIRA, B. **Estatística bayesiana.** Lisboa: Fundação Calouste Gulbenkian, 2003. 429 p.

PEREIRA, C.; ROGATKO, A. The hardy-weinberg equilibrium under a bayesian perspective. **Brazilian Journal of Genetics**, Ribeirão Preto, v. 7, n. 4, p. 689-707, Oct. 1984.

R DEVELOPMENT CORE TEAM . **R: a language and environment for statistical computing.** Vienna, Austria. R Foundation for Statistical Computing. Disponível em: <<http://www.R-project.org>>. Acesso em: 10 nov. 2007.

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

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.

ROGATKO, A.; SLIFKER, M. J.; BABB, J. S. Hardy-weinberg equilibrium diagnostics. **Theoretical Population Biology**, New York, v. 62, n. 3, p. 251-257, Nov. 2002.

SHOEMAKER, J. S.; PAINTER, I. S.; WEIR, W. S. A Bayesian characterization of Hardy-Weinberg disequilibrium. **Genetics**, Bethesda, v. 149, n. 4, p. 2079-2088, Aug. 1998.

SHOEMAKER, J. S.; PAINTER, I. S.; WEIR, W. S. Bayesian statistics in genetics: a guide for uninitiated. **Trends in Genetics**, London, v. 15, n. 9, p.354-358, Sept. 1999.

SILVA, F. F.; SÁFADI, T.; MUNIZ, J.A.; AQUINO, L.H.; MOURÃO, G.B. Comparação Bayesiana de modelos de previsão para diferenças esperadas nas progênes no melhoramento genético da gado Nelore. **Pesquisa Agropecuária Brasileira**, Piracicaba, v. 43, 2008. No prelo.

SMITH, A. F. M. Bayesian Methods in Reliability. In: SANDER, P.; BADEUX, R. (Ed.). **Topics in safety reliability and quality.** London: Kluwer Academic, 1991. p. 34-79.



SMITH, B. J. **Bayesian output analysis program (BOA) for MCMC**. R version 1.1.6-1. Disponível em: <<http://www.public-health.uiowa.edu/boa>>. Acesso em: 15 nov. 2007.

SORIA, F.; BASURCO, F.; TOVAL, G.; SILIÓ, L.; RODRIGUEZ, M. C.; TORO, M. An application of Bayesian techniques to the genetic evaluation of growth traits in *Eucalyptus globulus*. **National Research Council of Canada**, Ottawa, v. 28, n. 9, p. 1286-1294, Sept. 1998.

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

WILSON, G. A.; RANNALA, B. Bayesian inference of recent migration rates using a multilocus genotypes. **Genetics**, Baltimore, v. 163, p. 1177-1191, 2003.

WILSON, I. J.; BALDING, D. J. Genealogical inference from microsatellite data. **Genetics**, Baltimore, v. 150, p. 499-510, 1998.

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

## ANEXO

### LIMITES DOS PARÂMETROS DO DESEQUILÍBRIO DE HARDY-WEINBERG

Os limites do coeficiente de endogamia e do coeficiente de desequilíbrio,  $f$  e  $D_A$ , são determinadas a partir da restrição das proporções genotípicas do modelo de Hardy-Weinberg. Segundo Weir (1996), os limites das proporções genotípicas homozigotas são dadas por  $0 \leq p_{AA} \leq p_A$  e  $0 \leq p_{BB} \leq p_B$ . Adota-se que  $p_A = 1 - p_B$ .

Usando (5) para o cálculo dessas proporções homozigotas, obtém-se o seguinte resultado para  $f$ , considerando  $p_{AA}$ :

$$\begin{aligned} 0 &\leq p_A(f + (1-f)p_A) \leq p_A \\ 0 &\leq f + (1-f)p_A \leq 1 \\ 0 &\leq f + p_A - p_A f \leq 1 \\ -p_A &\leq f - p_A f \leq 1 - p_A \\ -p_A &\leq f(1 - p_A) \leq 1 - p_A \\ \frac{-p_A}{(1 - p_A)} &\leq f \leq \frac{1 - p_A}{1 - p_A} \\ \frac{-p_A}{(1 - p_A)} &\leq f \leq 1, \end{aligned} \tag{26}$$

e para para  $f$ , considerando  $p_{BB}$ :

$$\begin{aligned} 0 &\leq p_B(f + (1-f)p_B) \leq p_B \\ 0 &\leq f + (1-f)p_B \leq 1 \\ 0 &\leq f + p_B - p_B f \leq 1 \\ -p_B &\leq f - p_B f \leq 1 - p_B \end{aligned}$$

$$\begin{aligned}
-p_B &\leq f(1-p_B) \leq 1-p_B \\
\frac{-p_B}{(1-p_B)} &\leq f \leq \frac{1-p_B}{1-p_B} \\
\frac{-p_B}{(1-p_B)} &\leq f \leq 1.
\end{aligned} \tag{27}$$

Então, os limites inferiores e superiores de  $f$ , a partir de (26) e (27), são dados por (6).

Usando (7) para o cálculo das proporções homozigotas, obtém-se o seguinte resultado para  $D_A$ , considerando  $p_{AA}$ :

$$\begin{aligned}
0 &\leq p_A^2 + D_A \leq p_A \\
-p_A^2 &\leq D_A \leq p_A - p_A^2 \\
-p_A^2 &\leq D_A \leq p_A(1-p_A),
\end{aligned} \tag{28}$$

e para para  $D_A$ , considerando  $p_{BB}$ :

$$\begin{aligned}
0 &\leq p_B^2 + D_A \leq p_B \\
-p_B^2 &\leq D_A \leq p_B - p_B^2 \\
-p_B^2 &\leq D_A \leq p_B(1-p_B).
\end{aligned} \tag{29}$$

Então, os limites inferiores e superiores de  $D_A$ , a partir de (28) e (29), são dados por (8).