Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística

Modelo com Erros de Classificação para a Proporção de Não-disjunção Cromossômica na Meiose I

Cristiane Silva Souto

Orientadora: Prof. Dra. Rosângela H. Loschi

Sumário

1 Introdução **5**

2 A Base do Modelo Probabilístico para Trissomias **7**

3 Modelos Estatísticos **9**
3.1 Modelo sem Erro de Classificação (Franco *et al.*, 2003) 9
3.2 Modelo com Erro de Classificação 10

4 Inferência Usando uma Simplificação do Modelo **11**

5 A Falta de Identificabilidade **12**

6 Comparação de Modelos **14**

7 Métodos Computacionais **15**

8 Estudo Monte Carlo Comparando as Estimativas Obtidas via ambos os Modelos **15**
8.1 Comparando as Estimativas de φ para Diferentes Probabilidades de Má Classificação 16
8.2 As Estimativas de β 23

9 Aplicação **26**

10 Conclusão **29**

11 Apêndice 1: Estudo da Convergência e Autocorrelação **30**

12 Apêndice 2: Programas **32**

Agradecimentos

Agradeço à minha orientadora Rosângela H. Loschi pela paciência, tempo empreendido e incentivo no desenvolvimento deste trabalho. Agradeço aos professores e colegas dos Departamentos de Matemática e Estatística do Icex-UFMG e aos amigos dos demais cursos. Em especial, agradeço à Profa. Glaura C. Franco pela idéia que deu origem a este trabalho e aos professores Sérgio D. J. Pena e Flávia C. Parra por fornecerem os dados. À João Vitor deixo minha gratidão pelo desenvolvimento da parte computacional sem a qual não seria possível concluir este trabalho. Agradeço aos professores Reinaldo e Roberto pelas sugestões que tanto enriqueceram este trabalho.

Agradeço ao meu noivo, Gil e aos meus familiares e amigos pela paciência, apoio e incentivo nos momentos mais difíceis.

Resumo

As anomalias cromossômicas numéricas ocorrem, geralmente, como eventos esporádicos de não-disjunção na meiose. Uma das técnicas mais usadas para se analisar as anomalias cromossômicas é a Reação em Cadeia de Polimerase (PCR) seguida por uma análise quantitativa via densitometria laser, na qual o indivíduo é classificado como tendo 1, 2 ou 3 picos em um loco de microssatélite polimórfico. Foi mostrado em trabalhos anteriores que o número de indivíduos com 1, 2 ou 3 picos em uma dada amostra, tem distribuição multinomial cujos parâmetros dependem da fração de não-disjunção φ. Neste trabalho, propomos um modelo que leva em conta o erro de classificação que pode ser cometido na coleta dos dados para estimar a proporção de não-disjunção φ na meiose I. Usamos métodos numéricos para extrair informações da distribuição *a posteriori* de φ. Os estimadores de Bayes (média e moda *a posteriori*) dos modelos com e sem erro de classificação são comparados através de um estudo Monte Carlo, onde analisamos a influência de diferentes tamanhos amostrais e diferentes probabilidades de má classificação nas estimativas. Nós aplicamos o modelo proposto para estimar φ em pacientes com trissomia no cromossomo 21 e fizemos uma análise de sensibilidade para o modelo. Neste caso usamos o *Deviance Information Criterium* (DIC) para comparar os modelos com e sem erro de classificação. Os resultados obtidos mostram que o modelo proposto não é o mais adequado para estimar φ, o que se justifica pela baixa proporção estimada de erro de classificação encontrada nos dados analisados.

Palavras Chave: Trissomia, distribuição Multinomial, erro de classificação, identificabilidade, estimador de Bayes.

Abstract

The main causes of numerical chromosomal anomalies, including trisomies, arise from an error in the chromosomal segregation during the meiotic process, named a non-disjuntion. One of the most used techniques to analyse chromosomal anomalies is the Polymerase Chain Reaction (PCR) followed by a quantitative analysis via laser densitometry, which counts the number of peaks or alleles in polymorphic microsatellite locus. It was shown in previous works that the number of peaks has a multinomial distribution whose parameters depend on the non-disjunction fraction φ. In this work, we propose a misclassification model for estimating the meiosis I non-disjunction fraction φ. We consider the Gauss Legendre method and de Simpson rule to extract information from the posterior distribution of φ. Bayes estimators are compared through Monte Carlo studies which focus in the influence of different sample sizes and differents probabilities of misclassification in the estimates. We apply the proposed method to estimate φ for patients with trisomy of chromosome 21 providing a sensitivity analysis for the method. In this case we use the Deviance Information Criterium (DIC) to compare the proposed model and the model proposed by Franco *et al.* (2003). The results obtained show that the proposed model is not the best. A possible reason for its low performance is the small proportion of misclassification.

Key Words: Trisomy, Multinomial distribution, misclassification, identifiability, Bayes estimator.

1 Introdução

As anomalias cromossômicas numéricas, chamadas aneuploidias, são causas comuns de retardamento mental, má formação congênita e aborto, ocorrendo, geralmente, como eventos esporádicos de não-disjunção na meiose. A meiose consiste da duplicação do DNA seguida de duas divisões celulares gerando, no final, células haplóides. Dessa forma, as aneuploidias podem ocorrer devido a uma falha na primeira fase do processo meiótico (meiose I), na qual a não-disjunção dos cromossomos homólogos pode ser observada, ou na segunda fase do processo meiótico (meiose II), na qual as aneupoidias são uma conseqüência da não-disjunção de duas cromátides ligadas ao mesmo centrômero. Para maiores detalhes, ver Parra (1999).

Pouco se sabe sobre a origem e as causas genéticas das aneuploidias. De acordo com Parra (1999), a determinação da proporção de vezes em que a não-disjunção na segregação cromossômica ocorre na meiose I (ou na meiose II) em cada cromossomo pode ser explicada por possíveis fatores tais como geografia, nutrição, idade, etc. Por exemplo, na trissomia do cromossomo 21 que origina a Síndrome de Down e é uma das aneuploidias mais conhecidas, há evidência de que a proporção de não-disjunção aumenta com a idade da mãe (ver Pena, 1998). No estudo populacional desenvolvido por Yoon *et al.* (1996) foi demonstrado que o risco de se obter um feto com trissomia 21 para mulheres com idades de 35-39 anos aumenta 3.7 vezes para erros na meiose I e cerca de 62.8 vezes para erros na meiose II. Para cromossomos sexuais, entretanto, a alta idade da mãe influencia somente a proporção de não-disjunção na meiose I.

Muitos métodos foram propostos para estimar φ (dado que a não-disjunção ocorreu, φ denota a probabilidade condicional de que a não-disjunção ocorra na meiose I) como, por exemplo, em Hassold e Hunt (2001), Hassold e Jacobs (1984), Yoon *et al.* (1996) e Zaragosa *et al.* (1994). Nestes trabalhos, além da informação da pessoa afetada, também se considera a informação vinda de seus pais. Mais recentemente, voltado para o estudo de trissomias, Franco *et al.* (2003) propõem um modelo de probabilidade para descrever o comportamento do número de indivíduos na amostra que apresentam 1, 2 e 3 alelos distintos no loco de interesse em função de φ e obtêm o seu estimador de máxima verossimilhança (EMV). O modelo proposto por Franco *et al.* (2003) trouxe um ganho na estimação de φ, pois leva em consideração apenas a informação extraída do paciente trissômico possibilitando assim o uso de bancos de dados pré-existentes. Loschi e Franco (2006) comparam estimadores de Bayes (média e moda *a posteriori*) para estimar φ com o EMV e concluem que a moda obtida numericamente é similar ao EMV, como esperado, se uma distribuição *a priori* uniforme é considerada e que os estimadores de Bayes, em geral, são melhores se distribuições *a priori* Beta justas são consideradas. Os estimadores de Bayes também apresentam melhores estimativas para pequenas amostras. Nos casos em que as distribuições *a priori* têm valores médios distantes do verdadeiro valor de φ, notou-se que o EMV é, em geral, melhor.

Embora o modelo proposto por Franco *et al.* (2003) traga alguns avanços na análise de trissomias, este modelo pode não ser o mais adequado, uma vez que na coleta dos dados, ou seja, ao classificarmos um indivíduo quanto ao número de alelos presentes no loco de interesse, um erro de classificação pode ser cometido. A informação é extraída do DNA da pessoa afetada através de técnica denominada Reação em Cadeia de Polimerase (PCR) seguida por uma análise quantitativa via densitometria laser. A saída desta análise pode ser vista na Figura 1 apresentada em Parra (1999).

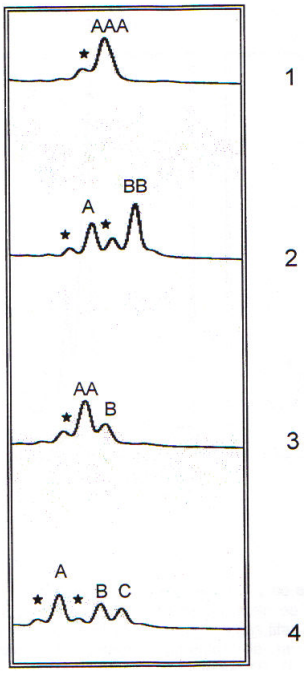

Figura 1: A canaleta 1 representa o padrão de 1 pico, as canaletas 2 e 3 representa o padrão de 2 picos e a canaleta 4 representa o padrão de 3 picos. Os picos principais representando os alelos estão indicados pelas letras A, B e C. Os picos de menor intensidade, localizados à esquerda dos picos principais e indicados pelas $\star$ são originados durante a PCR, devido à derrapagem das fitas de DNA durante a replicação da região repetitiva.

De acordo com Parra (1999) os picos de menor intensidade, indicados na Figura 1 por $\star$ são originados durante a PCR, devido à derrapagem das fitas de DNA durante a replicação da região repetitiva. A presença destes picos podem nos conduzir à um erro de classificação.

Este trabalho tem como objetivo principal propor um modelo para descrever o comportamento do número de indivíduos trissômicos com 1, 2 e 3 picos levando em conta a presença de erro de classificação. A abordagem bayesiana será considerada para a análise. Compararemos as estimativas para φ obtidas usando o modelo proposto e o modelo sugerido por Franco *et al.* (2003) através de um estudo Monte Carlo. Tem-se como meta no estudo de simulação, avaliar o efeito do tamanho da amostra e de diferentes probabilidades para o erro de classificação nas estimativas de Bayes (média e moda *a posteriori*). Consideraremos distribuições *a priori* uniforme para φ e para a probabilidade de se classificar erroneamente um indivíduo. Como ilustração utilizaremos o modelo proposto para analisar dados de pacientes Brasileiros com Síndrome de Down. Neste caso, apresentaremos uma análise de sensibilidade para o modelo considerando, distribuições *a priori* Beta (inclusive uniforme) para φ. Utilizamos as informações existentes na literatura para construir distribuições *a priori* mais realistas. Uma comparação com o modelo proposto por Franco *et al.* (2003) também é apresentada e, neste caso, o *DIC* (*Deviance Information Criterium*) é utilizado para identificar o melhor modelo. Compararemos as estimativas obtidas via WINBUGS e os Métodos Numéricos descritos na Seção 8.

Existe uma vasta literatura sobre modelos estatísticos que levam em conta o erro de classificação. Paulino *et al.* (2003) propõem um modelo de regressão Binomial no qual a variável resposta está sujeita à erro de classificação e o aplica no estudo da infecção do papilomavírus em mulheres. Em Stewart *et al.* (1998) o problema de erro de classificação em casos de câncer

entre dois grupos étnicos é considerado. Swartz *et al.* (2004) fazem uma análise Bayesiana de dados multinomiais na presença de erro de classificação. Viana (1994) propõe um modelo com erro de classificação para dados multinomiais extendendo os resultados propostos por Lew e Levy (1989) e Viana *et al.*(1993).

Ao construirmos modelos com erros de classificação, em geral, surgem problemas de falta de identificabilidade. Em particular, a falta de identificabilidade surge quando estamos fazendo a análise de dados multinomiais na presença de erro de classificação pois há, em geral, muitos parâmetros a serem estimados e o modelo pode não conseguir diferenciar entre eles. Segundo Swartz *et al.* (2004), em Estatística Bayesiana, a falta de identificabilidade entre os parâmetros pode afetar as estimativas de φ quando as distribuições *a posteriori* são obtidas através de métodos Monte Carlo via Cadeias de Markov (MCMC). Neste trabalho verificaremos que, para uma simplificação do modelo proposto, todos os parâmetros são identificáveis.

Este trabalho está organizado da seguinte forma. Na Seção 2, faremos uma breve descrição do problema genético a ser tratado neste trabalho (Pena, 1998). Na Seção 3, daremos uma breve descrição do modelo probabilístico proposto por Franco *et al.* (2003) que, seguindo Parra (1999) e Pena (1998), descreve o comportamento do número de indivíduos com 1, 2 ou 3 picos dado a proporção de vezes em que a não-disjunção cromossômica ocorre na meiose I. Além disso, introduziremos um modelo alternativo para este fim, o qual leva em conta a possibilidade de erro de classificação. Na Seção 4, abordaremos o problema de inferência sobre φ e sobre a probabilidade de se classificar erroneamente um indivíduo considerando uma simplicação do modelo proposto. Na Seção 5, apresentaremos duas definições de falta de identificabilidade e avaliaremos como se relacionam e verificaremos que para o modelo simplificado os parâmetros são identificáveis. Na seção 6 daremos uma breve descrição do cálculo do DIC que será usado na Seção 9 para comparar os modelos com e sem erro de classificação. Na Seção 7, descreveremos os métodos computacionais usados para calcular médias e modas *a posteriori* para os parâmetros de interesse nos modelos considerados. Na Seção 8, discutiremos alguns resultados obtidos nas simulações para estas estimativas considerando diferentes cenários. Na Seção 9 faremos uma aplicação do modelo proposto para analisar uma amostra de 34 indivíduos com trissomia no cromossomo 21 da população brasileira. Para finalizar, na Seção 10, concluímos o trabalho e apresentamos alguns pontos para trabalhos futuros e na Seção 11 (Apêndice 1) faremos um estudo da convergência e autocorrelação.

2 A Base do Modelo Probabilístico para Trissomias

Na construção do modelo Estatístico (ver Seção 3) consideraremos que o equilíbrio de Hardy-Weinberg é verificado para a população, o que significa dizer que as freqüências dos alelos não se alteram ao longo do tempo, ou seja, as freqüências alélicas tendem a manter-se constantes nas populações, ver Hardy (1908) e Weinberg (1908) para maiores detalhes. Além disso, as seguintes informações são levadas em conta.

De acordo com Parra (1999), pacientes trissômicos apresentam, no estudo de locos de microssatélite polimórficos, a presença de três picos de mesma intensidade, dois picos com uma dosagem relativa 2:1 ou apenas um pico (casos não-informativos). A proporção relativa de ocorrência desses três casos depende, além do índice de heterozigosidade do loco analisado, do momento do acidente meiótico.

Ignorando a recombinação, para um padrão com 3 picos ocorrer em um feto trissômico, é necessário que a não-disjunção tenha ocorrido na meiose I, a mãe seja heterozigota no loco de interesse e o pai transmita um alelo diferente daqueles transmitidos pela mãe. O padrão de 2

picos retrata uma situação em que as cromátides irmãs permanecem unidas e foram transmitidos dois alelos iguais e um diferente. Neste caso, a não-disjunção pode ter ocorrido tanto na meiose I quanto na meiose II. O padrão de 1 pico ocorre quando os pais são homozigotos, neste caso, são transmitidos três alelos iguais e a não-disjunção pode ter ocorrido tanto na meiose I quanto na meiose II.

Para entender os diferentes padrões que podem ocorrer devido a não-disjunção no processo meiótico, é necessário ver com mais detalhe, o processo como um todo. Sejam A e B dois alelos presentes em um particular loco de microsatélite de um dos pais e C o alelo do outro pai o qual se juntará aos dois primeiros alelos no processo de fertilização. Na Figura 2, apresentada em Franco *et. al* (2003), exibimos esquematicamente o evento da não-disjunção ocorrendo tanto na meiose I quanto na meiose II.

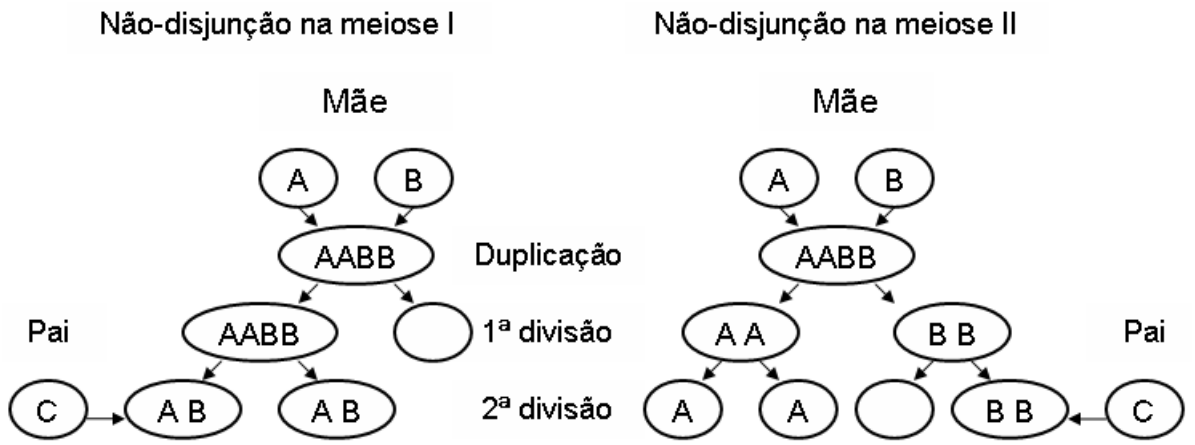

Figura 2: Não-disjunção na meiose I e II.

Uma vez que a não-disjunção ocorreu, o número de picos em um particular loco de um indivíduo trissômico será 1, 2 ou 3 se após o processo de fertilização observarmos, respectivamente as seguintes configurações:

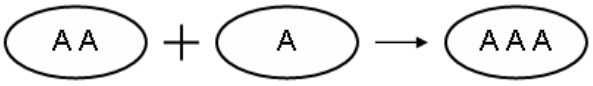

Figura 3: 1 pico (três alelos iguais).

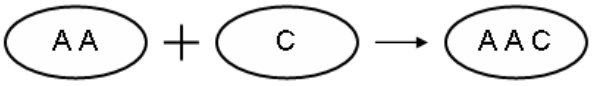

Figura 4: 2 picos (dois alelos iguais).

Figura 5: 3 picos (três alelos diferentes).

3 Modelos Estatísticos

Nesta seção apresentaremos brevemente o modelo proposto por Franco *et al.* (2003) que descreve o comportamento do número de indivíduos numa amostra com 1, 2 ou 3 picos, condicional no parâmetro φ, $\varphi \in [0,1]$. Ao longo dete trabalho φ denotará a probabilidade da não-disjunção cromossômica ter ocorrido na meiose I uma vez que a não-disjunção ocorreu. Também introduziremos um modelo alternativo para este fim o qual leva em conta a possibilidade de erro de classificação na coleta dos dados.

3.1 Modelo sem Erro de Classificação (Franco *et al.*, 2003)

Denotaremos por Y_l o número de indivíduos com l picos, $l = 1, 2, 3$, a ser observado em uma amostra de n indivíduos trissômicos e por $\mathbf{Y}$ o vetor aleatório (Y_1, Y_2, Y_3). Dado φ, Franco *et al.* (2003) propõem que o vetor aleatório $\mathbf{Y}$ tem distribuição multinomial com parâmetros $n, \theta_1(\varphi) > 0, \theta_2(\varphi) > 0, \theta_3(\varphi) > 0$, denotada por $\mathbf{Y} \sim Multinomial(n, \theta_1(\varphi), \theta_2(\varphi), \theta_3(\varphi))$, e cuja função de probabilidade é dada por:

$$P(\mathbf{Y} = \mathbf{y}|\varphi) = \frac{n!}{y_1!y_2!y_3!}[\theta_1(\varphi)]^{y_1}[\theta_2(\varphi)]^{y_2}[\theta_3(\varphi)]^{y_3} \quad , \tag{1}$$

onde $\mathbf{y} = (y_1, y_2, y_3)$, $y_l = 0, \cdots, n$, $l = 1, 2, 3$ e $\sum_{l=1}^{3} y_l = n$,

$$\begin{aligned}
\theta_1(\varphi) &= \varphi \textstyle\sum_{i=1}^{m} p_i^3 + (1-\varphi) \sum_{i=1}^{m} p_i^2; \\
\theta_2(\varphi) &= 3\varphi \textstyle\sum_{i=1}^{m} \sum_{j=1}^{m} p_i^2 p_j + (1-\varphi) \sum_{i=1}^{m} \sum_{j=1}^{m} p_i p_j, \ \ \forall\, i \neq j; \\
\theta_3(\varphi) &= \varphi \textstyle\sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k=1}^{m} p_i p_j p_k, \ \ \forall\, i \neq j \neq k;
\end{aligned} \tag{2}$$

onde m é o número de alelos na população e p_i, $i = 1, \ldots, m$ denota a freqüência relativa do alelo i em um dado loco. Essas freqüências alélicas são calculadas considerando a informação de bancos de dados pré-existentes e refletem a freqüência com que cada um dos m alelos é observado na população. Note que $\theta_l(\varphi)$ é a probabilidade de cada indivíduo apresentar l picos (alelos), $l = 1, 2, 3$. Podemos observar ainda que $\sum_{l=1}^{3} \theta_l(\varphi) = \varphi(\sum_{l=1}^{m} p_i)^3 + (1-\varphi)(\sum_{l=1}^{m} p_i)^2 = 1$.

Franco *et al.* (2003) também estabelecem que a função de verossimilhança em (1) possui os seguintes valores críticos:

$$R = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A}$$

onde

$$\begin{aligned}
A &= (a-b)[3c - (1-b)]n; \\
B &= (a-b)(1-b)(n-y_2) + b[3c - (1-b)](n-y_1); \\
C &= y_3(1-b)b,
\end{aligned}$$

os quais $a = \sum_i p_i^3$, $b = \sum_i p_i^2$ e $c = \sum_i \sum_j p_i^2 p_j$, $i, j = 1, \ldots, m$. O estimador de máxima verossimilhança são os valores de R tais que $0 \leq R \leq 1$.

Note que o modelo proposto por Franco *et al.* (2003) não leva em conta a possibilidade de erro de classificação que pode ser cometido na coleta dos dados como mencionamos na Seçao 1. Numa tentativa de descrever mais realísticamente o comportamento de $\mathbf{Y}$, dado φ, na próxima seção introduzimos um modelo que assume a possibilidade de haver erro de classificação na coleta dos dados.

3.2 Modelo com Erro de Classificação

Admitamos que a não-disjunção meiótica ocorreu. Segundo Paulino *et al.* (2003) e Swartz *et al.* (2004), para construirmos o modelo com erro de classificação, considera-se variáveis aleatórias auxiliares X e Z em que X denota o número real (não observado) de picos num indivíduo qualquer e Z denota o número observado de picos neste mesmo indivíduo. No caso em que o indivíduo é trissômico $X, Z \in \{1, 2, 3\}$.

Denote por $\theta_j(\varphi) = P(X = j)$ a probabilidade do número real de picos no indivíduo ser j e por $\pi_j(\varphi) = P(Z = j)$ a probabilidade do número observado de picos no indivíduo ser j; $j = 1, 2, 3$. As probabilidades $\theta_j(\varphi)$ são dadas em (2). Considere $\alpha_{j|k} = P(Z = j \mid X = k)$, a probabilidade do número observado de picos no indivíduo ser j dado que o número real de picos no indivíduo é k; $j, k = 1, 2, 3$. Note que se $j = k$, então $\alpha_{j|k}$ denota a probabilidade de classificarmos corretamente um indivíduo como tendo j picos e se $j \neq k$, então $\alpha_{j|k}$ denota a probabilidade de classificarmos erroneamente um indivíduo como tendo j picos quando, na realidade, tem k. Assumindo esta notação, temos que a probabilidade em função de φ do número observado de picos no indivíduo ser j é:

$$\begin{aligned} \pi_j(\varphi, \boldsymbol{\alpha}) \quad = P(Z = j) \quad &= \textstyle\sum_{k=1}^{3} P(Z = j, X = k) \\ &= \textstyle\sum_{k=1}^{3} P(Z = j \mid X = k) P(X = k) \\ &= \textstyle\sum_{k=1}^{3} \alpha_{j|k} \theta_k(\varphi), \end{aligned} \tag{3}$$

onde $\boldsymbol{\alpha} = (\alpha_{11}, \alpha_{12}, \alpha_{13}, \alpha_{21}, \alpha_{22}, \alpha_{23}, \alpha_{31}, \alpha_{32}, \alpha_{33})$. Note que usando resultados de cálculo de probabilidades, de (3) segue que:

$$\begin{aligned} \textstyle\sum_{j=1}^{3} \pi_j(\varphi, \boldsymbol{\alpha}) \quad &= \quad \textstyle\sum_{j=1}^{3} \sum_{k=1}^{3} \alpha_{j|k} \theta_k(\varphi) \\ &= \quad \textstyle\sum_{k=1}^{3} \theta_k(\varphi) \sum_{j=1}^{3} \alpha_{j|k} \\ &= \quad \textstyle\sum_{k=1}^{3} \theta_k(\varphi) \end{aligned}$$

uma vez que $\sum_{j=1}^{3} \alpha_{j|k} = \sum_{j=1}^{3} P(Z = j \mid X = k) = 1$, para cada $k = 1, 2, 3$. Como conseqüência, do que foi visto na Seção 3.1, tem-se que $\sum_{j=1}^{3} \pi_j(\varphi, \boldsymbol{\alpha}) = \theta_1(\varphi) + \theta_2(\varphi) + \theta_3(\varphi) = 1$.

Suponha que uma amostra de n indivíduos trissômicos seja observada e que cada indivíduo seja classificado como tendo k picos, independentemente um do outro, com probabilidade $\pi_j(\varphi, \boldsymbol{\alpha})$. Denote por Y_j, $j = 1, 2, 3$ o número de indivíduos na amostra que foram classificados como tendo j picos e faça $\mathbf{Y} = (Y_1, Y_2, Y_3)$. Então, por construção, tem-se que

$$\mathbf{Y} = (Y_1, Y_2, Y_3) \sim Multinomial(n, \pi_1(\varphi, \boldsymbol{\alpha}), \pi_2(\varphi, \boldsymbol{\alpha}), \pi_3(\varphi, \boldsymbol{\alpha}));$$

cuja função de probabilidade é dada por:

$$\begin{aligned} f(Y_1, Y_2, Y_3 \mid \boldsymbol{\alpha}, \varphi) \quad &= \frac{n!}{y_1! y_2! y_3!} \prod_{j=1}^{3} [\pi_j(\varphi, \boldsymbol{\alpha})]^{y_j} \\ &= \frac{n!}{y_1! y_2! y_3!} \prod_{j=1}^{3} \left[\sum_{k=1}^{3} \alpha_{j|k} \theta_k(\varphi) \right]^{y_j}, \end{aligned} \tag{4}$$

em que $\sum_{j=1}^{3} y_j = n$. Uma vez que $\sum_{j=1}^{3} \alpha_{j|k} = 1$, perceba da expressão (4) que há 7 parâmetros a serem estimados.

4 Inferência Usando uma Simplificação do Modelo

Neste trabalho, assumiremos uma simplificação do modelo apresentado em (3) e (4). Consideraremos que

$$\alpha_{j|k} = \begin{cases} \alpha, \text{ se } j = k \\ \beta, \text{ se } j \neq k \end{cases}, \tag{5}$$

onde α denota a probabilidade de classificarmos cada indivíduo corretamente quanto ao número de picos e β representa a probabilidade de classificarmos erroneamente cada indivíduo quanto ao número de picos. Como $\sum_{j=1}^{3} \alpha_{j|k} = 1$, então, temos que $\alpha = 1 - 2\beta$. Sob estas suposições, temos, da expressão (3) que:

$$\begin{aligned} \pi_j(\varphi, \beta) &= \textstyle\sum_{k=1}^{3} \alpha_{j|k}\theta_k(\varphi) \\ &= \alpha\ \theta_j(\varphi) + \beta \textstyle\sum_{k\neq j} \theta_k(\varphi) \\ &= \alpha\ \theta_j(\varphi) + \beta(1 - \theta_j(\varphi)) \\ &= (1 - 2\beta)\theta_j(\varphi) + \beta(1 - \theta_j(\varphi)). \end{aligned} \tag{6}$$

Consequentemente, de (4) e (6) tem-se que a função de verossimilhança adequada para descrever o comportamento de $\mathbf{Y}$ torna-se:

$$\begin{aligned} f(Y_1, Y_2, Y_3 \mid \beta, \varphi) &= \frac{n!}{y_1!y_2!y_3!} \prod_{j=1}^{3} [\pi_j(\varphi, \beta)]^{y_j} \\ &= \frac{n!}{y_1!y_2!y_3!} \prod_{j=1}^{3} [(1 - 2\beta)\theta_j(\varphi) + \beta(1 - \theta_j(\varphi))]^{y_j}. \end{aligned} \tag{7}$$

Perceba que o modelo proposto por Franco *et al.* (2003) é um caso particular do modelo em (7) se assumirmos $\beta = 0$.

Para fazermos inferências sobre β e φ construiremos a distribuição *a priori* de (β, φ) da seguinte forma:

(i) assumiremos que β e φ são independentes o que significa dizer que estamos assumindo que a incerteza sobre a probabilidade de má classificação β não se modifica ao sabermos a proporção φ da não-disjunção ocorrer na meiose I;

(ii) admitiremos que a probabilidade de classificarmos corretamente é maior do que a probabilidade de classificarmos erroneamente, isto é, assumiremos que $\alpha > 2\beta$, o que implica dizer que assumiremos $\beta < \frac{1}{4}$;

(iii) consideraremos pouca informação inicial sobre β e φ, ou seja, assumiremos que, *a priori*, $\beta \sim \mathcal{U}(0, \frac{1}{4})$ e $\varphi \sim \mathcal{U}(0, 1)$.

Desta forma, a distribuição *a priori* conjunta de (β, φ) é dada por:

$$\pi(\beta, \varphi) = \begin{cases} 4, & \text{se } \varphi \in (0, 1) \text{ e } \beta \in (0, \frac{1}{4}), \\ 0, & \text{cc.} \end{cases} \tag{8}$$

Como conseqüência de (7) e (8), temos que a distribuição *a posteriori* conjunta de (β, φ) é dada por:

$$\begin{aligned} f(\beta, \varphi \mid y) &= \frac{f(y \mid \beta, \varphi)\pi(\beta, \varphi)}{\int_0^1 \int_0^{\frac{1}{4}} f(y \mid \beta, \varphi)\pi(\beta, \varphi) \ d\beta d\varphi} \\ &= \frac{\frac{n!}{y_1!y_2!y_3!} \prod_{j=1}^{3} [(1-2\beta)\theta_j(\varphi) + \beta(1-\theta_j(\varphi))]^{y_j} \ 4}{\int_0^1 \int_0^{\frac{1}{4}} \frac{n!}{y_1!y_2!y_3!} \prod_{j=1}^{3} [(1-2\beta)\theta_j(\varphi) + \beta(1-\theta_j(\varphi))]^{y_j} \ 4 \ d\beta d\varphi} \\ &= \frac{\prod_{j=1}^{3} [(1-2\beta)\theta_j(\varphi) + \beta(1-\theta_j(\varphi))]^{y_j}}{\int_0^1 \int_0^{\frac{1}{4}} \prod_{j=1}^{3} [(1-2\beta)\theta_j(\varphi) + \beta(1-\theta_j(\varphi))]^{y_j} \ d\beta d\varphi}. \end{aligned}$$

Tem-se ainda que as distribuições marginais de φ e β são dadas, respectivamente, por:

$$\begin{aligned} f(\varphi \mid y) &= \int_0^{\frac{1}{4}} f(\beta, \varphi \mid y) \ d\beta \\ &= \frac{\int_0^{\frac{1}{4}} \prod_{j=1}^{3} [(1-2\beta)\theta_j(\varphi) + \beta(1-\theta_j(\varphi))]^{y_j} \ d\beta}{\int_0^1 \int_0^{\frac{1}{4}} \prod_{j=1}^{3} [(1-2\beta)\theta_j(\varphi) + \beta(1-\theta_j(\varphi))]^{y_j} \ d\beta d\varphi} \end{aligned} \tag{9}$$

e

$$\begin{aligned} f(\beta \mid y) &= \int_0^1 f(\beta, \varphi \mid y) \ d\varphi \\ &= \frac{\int_0^1 \prod_{j=1}^{3} [(1-2\beta)\theta_j(\varphi) + \beta(1-\theta_j(\varphi))]^{y_j} \ d\varphi}{\int_0^1 \int_0^{\frac{1}{4}} \prod_{j=1}^{3} [(1-2\beta)\theta_j(\varphi) + \beta(1-\theta_j(\varphi))]^{y_j} \ d\beta d\varphi}. \end{aligned} \tag{10}$$

5 A Falta de Identificabilidade

Segundo Swartz *et al.* (2004), uma dificuldade que ocorre freqüentemente no estudo de dados categorizados é a existência de erros de classificação, isto é, a diferença entre a classificação observada e a classificação verdadeira. Erros de classificação em dados modelados por distribuições multinomiais, em geral, conduzem a problemas de falta de identificabilidade do modelo isto por que, por serem modelos inflacionados de parâmetros, os dados podem não fornecer informação sobre alguns deles. Formalmente, Swartz *et al.* (2004) definem não identificabilidade como segue:

Definição 1 *(ver Swartz et al. (2004) por exemplo) Seja U uma variável aleatória com função de distribuição F_θ pertencente à família $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$ de distribuições indexadas pelo*

parâmetro θ. Aqui θ pode ser um escalar ou um vetor. Dizemos que θ é não identificável por U se existe ao menos um par (θ, θ'), onde $\theta \neq \theta'$ ambos pertencentes à Θ tal que $F_\theta(u) = F_{\theta'}(u)$. Caso contrário, diremos que θ é identificável.

Segundo Gelfand e Sahu (1999), no contexto bayesiano, a falta de identificabilidade no modelo pode gerar problemas na obtenção da distribuição *a posteriori* via métodos MCMC pois as distribuições condicionais completas não ficam atualizadas pelos dados. Na realidade, como, no contexto Bayesiano, tanto θ quanto X são objetos aleatórios, Dawid (1979) define falta de identificabilidade da seguinte forma:

Definição 2 *Considere um vetor de parâmetros $\boldsymbol{\theta} = (\theta_1, \theta_2)$. Suponha que o modelo possua função de verossimilhança dada por $f(y \mid \boldsymbol{\theta})$ e assuma que a distribuição a priori para $\boldsymbol{\theta}$ é $f(\boldsymbol{\theta})$. Diz-se que θ_2 é não identificável se*

$$f(\theta_2|\theta_1, y) = f(\theta_2|\theta_1). \tag{11}$$

Note de que como conseqüência de (11) tem-se que a função de verossimilhança não é função de θ_2, pois

$$\begin{aligned} f(\theta_2|\theta_1, y) = f(\theta_2|\theta_1) \quad &\Leftrightarrow \frac{f(y \mid \theta_1, \theta_2) f(\theta_2|\theta_1) f(\theta_1)}{f(y, \theta_1)} = f(\theta_2|\theta_1) \\ &\Leftrightarrow f(y \mid \theta_1, \theta_2) = \frac{f(y, \theta_1)}{f(\theta_1)} = f(y \mid \theta_1) \end{aligned} \tag{12}$$

Perceba ainda que θ_2 ser não identificável, não implica em que as distribuições *a priori* e *a posteriori* atribuídas a θ_2 sejam iguais, isto é, que $f(\theta_2|y) = f(\theta_2)$.

Pode-se mostrar que as definições 1 e 2 são equivalentes.

Teorema 1 *Definição* 1 $\Leftrightarrow$ *Definição* 2.

Prova Para isso, suponha que θ_2 é não identificável no sentido da Definição 2 e considere dois valores distintos de θ_2, isto é, tome os vetores (θ_1, θ_2') e (θ_1, θ_2''). Segue da Definição 2 que $f(\theta_2|\theta_1, y) = f(\theta_2|\theta_1)$

$$\Rightarrow \begin{cases} f(\theta_2'|\theta_1, y) = f(\theta_2'|\theta_1) \\ f(\theta_2''|\theta_1, y) = f(\theta_2''|\theta_1) \end{cases} . \tag{13}$$

Da expressão (13) segue que

$$\begin{cases} f(y \mid \theta_2', \theta_1) = \dfrac{f(y, \theta_1)}{f(\theta_1)} = f(y \mid \theta_1) \\ f(y \mid \theta_2'', \theta_1) = \dfrac{f(y, \theta_1)}{f(\theta_1)} = f(y \mid \theta_1) \end{cases} . \tag{14}$$

Logo, de (14), concluímos que $f(y \mid \theta_2', \theta_1) = f(y \mid \theta_2'', \theta_1)$ e a Definição 1 é satisfeita.

Reciprocamente suponha, por absurdo, que θ_2 é identificável segundo a Definição 2. Então tem-se que

$$f(\theta_2 \mid \theta_1, y) \neq f(\theta_2 \mid \theta_1), \ \ \forall \text{ valor assumido por } \theta_2 \ .$$

Daí, por (12) tem-se que $f(y \mid \theta_1, \theta_2) \neq f(y \mid \theta_1), \ \ \forall\, \theta_2$, ou seja, $f(y \mid \theta_1, \theta_2)$ não é constante com respeito a θ_2. Admita que $f(y \mid \theta_1, \theta_2) = g(\theta_2), \ \ \forall\, \theta_2$. Daí, para dois valores distintos de θ_2, digo, θ_2' e θ_2'', segue que $f(y \mid \theta_1, \theta_2') = g(\theta_2') \neq g(\theta_2'') = f(y \mid \theta_1, \theta_2'')$. Portanto, tem-se que $f(y \mid \theta_1, \theta_2') \neq f(y \mid \theta_1, \theta_2''), \ \ \forall\, \theta_2$ o que nos leva à uma contradição. Portanto Definição 1 $\Rightarrow$ Definição 2 e o teorema está provado.

Note que adotando a restrição em (5) e assumindo a distribuição *a priori* em (8) tem-se que β e φ são identificáveis, pois

$$\begin{aligned} f(\beta \mid \varphi, y) &= \frac{P(y \mid \varphi, \beta) P(\beta \mid \varphi) P(\varphi)}{\displaystyle\int_\theta P(y \mid \varphi, \beta) P(\beta \mid \varphi) P(\varphi) \ d\beta} \\ &= \frac{\displaystyle\prod_{j=1}^{3} [(1-2\beta)\theta_j(\varphi) + \beta(1-\theta_j(\varphi))]^{y_j}}{\displaystyle\int_0^{\frac{1}{4}} \prod_{j=1}^{3} [(1-2\beta)\theta_j(\varphi) + \beta(1-\theta_j(\varphi))]^{y_j} \ d\beta} \\ &\neq f(\beta|\varphi) = f(\beta). \end{aligned}$$

e

$$\begin{aligned} f(\varphi \mid \beta, y) &= \frac{P(y \mid \varphi, \beta) P(\varphi \mid \beta) P(\beta)}{\displaystyle\int_\Theta P(y \mid \varphi, \beta) P(\varphi \mid \beta) P(\beta) \ d\varphi} \\ &= \frac{\displaystyle\prod_{j=1}^{3} [(1-2\beta)\theta_j(\varphi) + \beta(1-\theta_j(\varphi))]^{y_j}}{\displaystyle\int_0^{1} \prod_{j=1}^{3} [(1-2\beta)\theta_j(\varphi) + \beta(1-\theta_j(\varphi))]^{y_j} \ d\varphi} \\ &\neq f(\varphi|\beta) = f(\varphi). \end{aligned}$$

Sendo assim, ao usarmos os métodos MCMC na obtenção das distribuições *a posteriori* de φ e β (ver Seção 9) estaremos gerando valores das distribuições condicionais completas *a posteriori* propriamente ditas.

6 Comparação de Modelos

Para comparar os modelos com e sem erro de classificação, na aplicação que será feita na Seção 9, utilizaremos o DIC (*Deviance Information Criterium*) que é dado por

$$\begin{aligned} DIC(\boldsymbol{\theta}_i, M_i) &= -2 \log p(\mathbf{y} \mid \overline{\boldsymbol{\theta}}_i, M_i) + 2p_D \\ &= \overline{D} + p_D \end{aligned}$$

em que $\overline{\boldsymbol{\theta}}_i = E(\boldsymbol{\theta}_i \mid \mathbf{y})$ determinada sob o modelo M_i, $\overline{D} = E(D(\boldsymbol{\theta}_i) \mid \mathbf{y})$, $p_D = \overline{D} - D(\overline{\boldsymbol{\theta}}_i)$ e $D(\theta) = -2 \ln p(y|\theta)$ é a função de *deviance* (ver Spiegelhalter *et al.* (2002) para maiores detalhes). O melhor modelo é aquele que apresenta menor valor para o DIC. Computacionalmente, o DIC é mais atrativo do que os diversos fatores de Bayes, pois seus termos podem ser facilmente incorporados dentro das rotinas MCMC os quais já estão implementados no programa WINBUGS.

7 Métodos Computacionais

Perceba de (9) e (10) que as distribuições *a posterirori* de φ e β não são exatamente conhecidas. Quando necessitamos encontrar soluções para problemas matemáticos que geralmente não têm solução exata ou cuja solução é difícil de ser obtida analiticamente, utilizamos ferramentas do Cálculo Numérico para resolver tais problemas. Uma possibilidade para calcularmos numericamente a média e a moda *a posteriori* de $\varphi \in (0,1)$ e $\beta \in (0, \frac{1}{4})$ é utilizando alguma técnica de quadratura. Consideramos um método chamado Método de Gauss-Legendre para aproximar as integrais duplas e a Regra de Simpson para calcular todas as integrais simples envolvidas no processo de estimação. Ver, por exemplo, Campos (2001) e Migon e Gamerman (1999) para maiores detalhes sobre estes métodos.

Denote por $\hat{I}$ o valor estimado para a integral no denominador das equações (9) e (10) obtido aplicando-se o algoritmo de Gauss-Legendre. Então, a média *a posteriori* de φ é estimada por:

$$E(\varphi \mid y) = \frac{\hat{I}_1}{\hat{I}}$$

onde $\hat{I}_1$ denota o valor obtido pelo método de Gauss-Legendre para a integral:

$$I_1 = \int_0^1 \int_0^{\frac{1}{4}} \varphi \prod_{j=1}^{3} [(1-2\beta)\theta_j(\varphi) + \beta(1-\theta_j(\varphi))]^{y_j} \; d\beta d\varphi.$$

Para calcular a moda *a posteriori* de φ, dividimos o intervalo $(0,1)$ em um número grande de subintervalos e avaliamos a função

$$f(\varphi \mid y) = \frac{\int_0^{\frac{1}{4}} \prod_{j=1}^{3} [(1-2\beta)\theta_j(\varphi) + \beta(1-\theta_j(\varphi))]^{y_j} \; d\beta}{\hat{I}}, \tag{15}$$

em cada ponto final destes subintervalos. A moda *a posteriori* é o ponto final que maximiza a expressão em (15). Devemos observar que quanto maior for o número de subintervalos, melhor será a aproximação da moda de φ. Analogamente, obtem-se a média e a moda *a posteriori* de β.

8 Estudo Monte Carlo Comparando as Estimativas Obtidas via ambos os Modelos

Para compararmos as estimativas *a posteriori* de φ obtidas utilizando os modelos com e sem erro de classificação e avaliarmos a qualidade das estimativas de β fizemos um estudo Monte Carlo. Uma das metas deste estudo é avaliar o efeito do tamanho da amostra e também de diferentes probabilidades de má classificação nas estimativas *a posteriori* de φ. Motivados pela aplicação real que será discutida posteriormente na Seção 9, consideramos que no loco de microssatélite de interesse há apenas 6 alelos distintos cujas freqüências alélicas são $p_1 = 0,12$, $p_2 = 0,45$, $p_3 = 0,09$, $p_4 = 0,31$, $p_5 = 0,01$ e $p_6 = 0,02$. Usamos diferentes valores de φ ($\varphi = 0,01,\ 0,10,\ 0,50,\ 0,90$ e $0,99$), diferentes valores de β ($\beta = 0,01,\ 0,05$ e $0,20$) e dois tamanhos amostrais ($n = 10$ e $n = 100$). Amostras de tamanhos pequenos são incluídas no estudo para avaliar o comportamento dos modelos em trissomias raras sobre as quais tem-se

pouca informação. Consideramos 500 réplicas Monte Carlo. Para gerar os dados, geramos amostras de indivíduos trissômicos usando o modelo definido em (7), o qual leva em conta o erro de classificação e aplicamos nos modelos com e sem erro de classificação para fazermos a análise. Para descrever a incerteza *a priori* de φ e β consideramos distribuições uniforme com suporte $(0, 1)$ e $(0, \frac{1}{4})$, respectivamente.

8.1 Comparando as Estimativas de φ para Diferentes Probabilidades de Má Classificação

Nesta seção, apresentamos uma análise dos resultados para as estimativas *a posteriori* de φ obtidas usando tanto o modelo proposto, o qual leva em conta a possibilidade de erro de classificação, quanto o modelo proposto por Franco *et. al* (2003).

Nas Tabelas 1-3 encontram-se os resultados obtidos considerando amostras de tamanho $n = 10$.

Tabela 1: Estatísticas descritivas para média e moda *a posteriori* de φ, caso $\beta = 0,01$, $n = 10$.

phi	Modelo	Estimativa	Média	EQM	Desvio Padrão	Q1	Mediana	Q3
0,01	c/ erro	media	0,3258	0,1019	0,0463	0,2931	0,3112	0,3353
		moda	0,0279	0,0041	0,0614	0,0100	0,0100	0,0100
	s/ erro	media	0,2822	0,0800	0,0773	0,2303	0,2598	0,2961
		moda	0,0408	0,0133	0,1115	0,0010	0,0010	0,0010
0,10	c/ erro	media	0,3536	0,0687	0,0662	0,2931	0,3353	0,3934
		moda	0,0764	0,0212	0,1439	0,0100	0,0100	0,0800
	s/ erro	media	0,3293	0,0648	0,1108	0,2303	0,2961	0,4065
		moda	0,1208	0,0384	0,1950	0,0010	0,0010	0,2770
0,50	c/ erro	media	0,4715	0,0105	0,0983	0,4051	0,4880	0,5325
		moda	0,4164	0,1258	0,3450	0,0800	0,3900	0,6800
	s/ erro	media	0,5213	0,0241	0,1538	0,4065	0,5433	0,6217
		moda	0,4889	0,1043	0,3231	0,2770	0,5120	0,7520
0,90	c/ erro	media	0,5656	0,1222	0,1020	0,4942	0,5699	0,6245
		moda	0,7269	0,1421	0,3352	0,4600	0,8600	1,0000
	s/ erro	media	0,6560	0,0801	0,1435	0,5788	0,6703	0,7644
		moda	0,7599	0,1057	0,2938	0,5690	0,8700	1,0000
0,99	c/ erro	media	0,5846	0,1738	0,0969	0,5290	0,5837	0,6621
		moda	0,7808	0,1371	0,3058	0,6200	0,9500	1,0000
	s/ erro	media	0,6813	0,1128	0,1324	0,6217	0,7065	0,7816
		moda	0,8079	0,1022	0,2631	0,6820	0,8830	1,0000

Tabela 2: Estatísticas descritivas para média e moda *a posteriori* de φ, caso $\beta = 0,05$, $n = 10$.

phi	Modelo	Estimativa	Média	EQM	Desvio Padrão	Q1	Mediana	Q3
0,01	c/ erro	media	0,3680	0,1343	0,0784	0,3112	0,3353	0,4184
		moda	0,1190	0,0477	0,1895	0,0100	0,0100	0,1500
	s/ erro	media	0,3542	0,1357	0,1313	0,2598	0,2961	0,4474
		moda	0,1789	0,0824	0,2323	0,0010	0,0010	0,3240
0,10	c/ erro	media	0,3843	0,0875	0,0817	0,3112	0,3740	0,4499
		moda	0,1508	0,0519	0,2223	0,0100	0,0200	0,2500
	s/ erro	media	0,3822	0,0983	0,1369	0,2598	0,3708	0,4933
		moda	0,2229	0,0787	0,2524	0,0010	0,2420	0,3940
0,50	c/ erro	media	0,4859	0,0104	0,1011	0,4183	0,4880	0,5699
		moda	0,4667	0,1289	0,3578	0,1500	0,3900	0,8600
	s/ erro	media	0,5448	0,0264	0,1562	0,4474	0,5433	0,6648
		moda	0,5405	0,1032	0,3190	0,3240	0,5120	0,8700
0,90	c/ erro	media	0,5656	0,1226	0,1040	0,4941	0,5837	0,6246
		moda	0,7355	0,1400	0,3364	0,4600	0,9500	1,0000
	s/ erro	media	0,6587	0,0802	0,1483	0,5789	0,7063	0,7644
		moda	0,7646	0,1040	0,2930	0,5690	0,8830	1,0000
0,99	c/ erro	media	0,5818	0,1772	0,1023	0,5290	0,5861	0,6621
		moda	0,7736	0,1492	0,3203	0,6200	0,9750	1,0000
	s/ erro	media	0,6798	0,1165	0,1421	0,6218	0,7066	0,7917
		moda	0,8004	0,1110	0,2742	0,6820	0,8855	1,0000

Tabela 3: Estatísticas descritivas para média e moda *a posteriori* de φ, caso $\beta = 0,20$, $n = 10$.

phi	Modelo	Estimativa	Média	EQM	Desvio Padrão	Q1	Mediana	Q3
0,01	c/ erro	media	0,4764	0,2279	0,1014	0,3934	0,4880	0,5482
		moda	0,4410	0,3173	0,3631	0,0800	0,3900	0,7600
	s/ erro	media	0,5383	0,3056	0,1627	0,4066	0,5618	0,6703
		moda	0,5279	0,3601	0,3035	0,2770	0,5320	0,7520
0,10	c/ erro	media	0,4783	0,1534	0,1017	0,3934	0,4940	0,5482
		moda	0,4500	0,2540	0,3631	0,0800	0,4600	0,7600
	s/ erro	media	0,5404	0,2203	0,1624	0,4066	0,5789	0,6703
		moda	0,5303	0,2751	0,3003	0,2770	0,5690	0,7520
0,50	c/ erro	media	0,5202	0,0104	0,1006	0,4300	0,5290	0,5887
		moda	0,5916	0,1410	0,3644	0,2500	0,6200	1,0000
	s/ erro	media	0,6064	0,0349	0,1537	0,4989	0,6334	0,7348
		moda	0,6502	0,1090	0,2941	0,4350	0,6820	0,8880
0,90	c/ erro	media	0,5480	0,1333	0,0978	0,4940	0,5482	0,6229
		moda	0,6884	0,1604	0,3405	0,4600	0,7600	1,0000
	s/ erro	media	0,6473	0,0842	0,1431	0,5789	0,6703	0,7410
		moda	0,7270	0,1019	0,2685	0,5690	0,7520	1,0000
0,99	c/ erro	media	0,5593	0,1944	0,0950	0,4941	0,5699	0,6229
		moda	0,7183	0,1808	0,3273	0,4600	0,9000	1,0000
	s/ erro	media	0,6632	0,1251	0,1357	0,5789	0,7038	0,7644
		moda	0,7542	0,1219	0,2577	0,5690	0,8700	1,0000

Podemos observar da Tabela 1 que, em média e para ambos os modelos, a moda *a posteriori*, em geral, fornece melhores estimativas para o parâmetro φ. Além disso, notamos que, em média, a média e a moda *a posteriori* computadas utilizando o modelo sem erro de classificação, estimam melhor o valor real de φ, exceto para $\varphi = 0,01$ onde a moda *a posteriori* média é $0,0279$ no modelo com erro enquanto que no modelo sem erro esse valor é igual a $0,0408$. Pode-se observar ainda que, em média, na maioria dos casos a moda *a posteriori* estima melhor do que a média *a posteriori*, exceto quando $\varphi = 0,50$, no modelo com erro onde a média *a posteriori*, em média, estima melhor do que a moda *a posteriori* (média *a posteriori* média = $0,4715$ e a moda *a posteriori* média = $0,4164$). Note ainda que conclusão similar pode ser tirada considerando o EQM salvo para $\varphi = 0,50$. Devemos observar também que os desvios padrões da moda *a posteriori* são maiores do que os desvios padrões da média *a posteriori* em todos os casos, o que significa que as estimativas da moda *a posteriori* estão mais dispersas em torno da média. Nota-se por exemplo que se $\varphi = 0,01$ e $\varphi = 0,10$ em pelo menos 75% dos casos, a média *a posteriori* superestima o valor de φ, para ambos os modelos. Já para $\varphi = 0,90$ e $\varphi = 0,99$, nota-se que a média *a posteriori* subestima φ em 75% dos casos. Para $\varphi = 0,50$, percebe-se que no modelo com erro a média *a posteriori* subestima φ em pelo menos 50% dos casos e, no modelo sem erro, ela superestima em 50% dos casos. Percebe-se, no entanto, que o modelo com erro tende a produzir médias *a posteriori* mais próximas do real uma vez que pelo menos 50% das estimativas estão entre $0,4051$ e $0,5325$.

Note que, neste caso, a probabilidade de má classificação é muito baixa (1%) e isto poderia explicar o mal desempenho do modelo proposto neste caso.

Da Tabela 2 observamos que, em média, a moda *a posteriori* fornece melhores estimativas para o valor real de φ do que a média *a posteriori*, exceto para $\varphi = 0,5$ no modelo com erro onde a média *a posteriori* média é $0,4859$ enquanto que a moda *a posteriori* média é $0,4667$. Note ainda que em relação a moda *a posteriori* temos que para $\varphi = 0,01$ pelo menos 50% das estimativas são menores ou iguais à $0,0100$ no modelo com erro enquanto que no modelo sem erro pelo menos 75% são maiores ou iguais a $0,0010$. Para $\varphi = 0,99$ pelo menos 50% das estimativas são maiores ou iguais à $0,9750$ no modelo com erro enquanto que no modelo sem erro pelo menos 50% das estimativas são menores ou iguais à $0,8855$. Considerando o EQM para $\varphi = 0,01$ e $0,10$, nota-se que as estimativas obtidas considerando o modelo com erro de classificação estão mais próximas do verdadeiro valor de φ embora, em média, a média *a posteriori* no modelo sem erro esteja mais próxima do valor real. Para $\varphi = 0,50$, a média *a posteriori* fornece estimativas mais próximas do valor real em ambos os modelos embora, em média, a média *a posteriori* no modelo sem erro esteja mais próxima do valor real. Para $\varphi = 0,90$ nota-se que a média *a posteriori* fornece estimativas mais próximas do valor real em ambos os modelos embora, em média, a moda *a posteriori* em ambos os modelos esteja mais próxima do valor real. Para $\varphi = 0,99$ nota-se que a moda *a posteriori* fornece estimativas mais próximas do valor real de φ. Uma possível explicação para esta divergência entre o EQM e a estimativa média é a presença de estimativas atípicas que puxam o valor da média para cima ou para baixo. Isso pode ser observado nos seguintes box-plots:

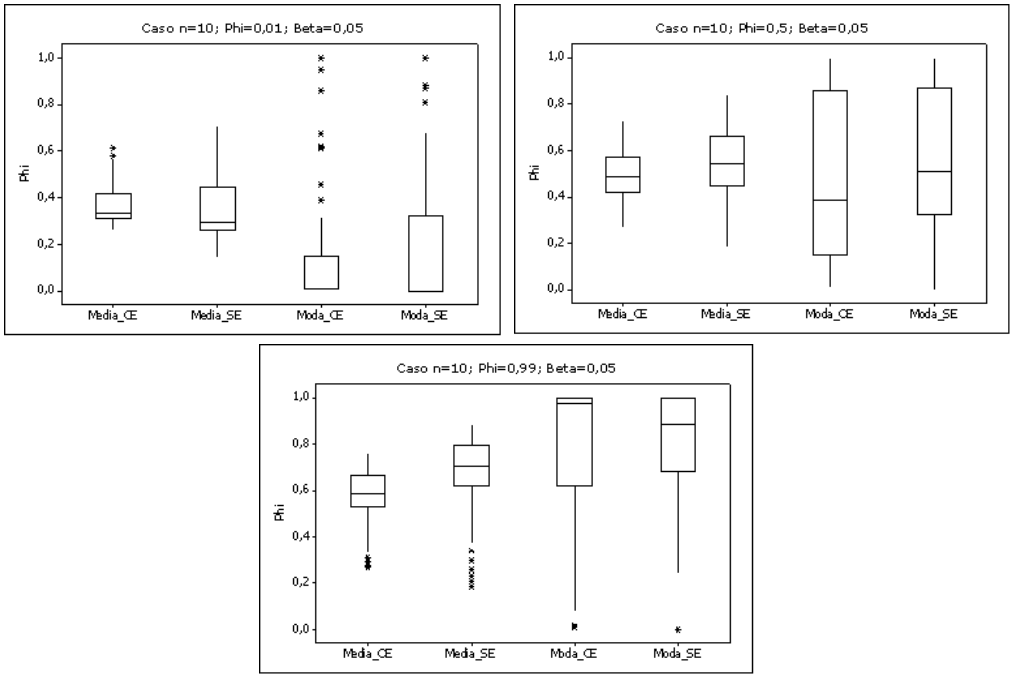

Figura 6: Box-Plot para média e moda *a posteriori* de φ, $\beta = 0,05$, $n = 10$.

Da Tabela 3 notamos que, em média, as estimativas (tanto a média quanto a moda *a posteriori*) obtidas utilizando o modelo com erro de classificação são melhores para $\varphi = 0,01, 0,10$ e $0,50$. Estas conclusões são corroboradas pelo EQM. Salvo para o caso em que $\varphi = 0,50$, a moda *a posteriori* é, em geral, melhor estimador que a média *a posteriori* em ambos os modelos.

Comparando as Tabelas 1 à 3 percebemos que para $\varphi = 0,01$ e $0,10$ a tendência é termos uma superestimação dos parâmetros por ambos os modelos sendo que esta superestimação é mais acentuada quando $\beta = 0,20$ (note que o EQM é bem maior neste caso). Para $\varphi = 0,90$ e $\varphi = 0,99$ tende a ocorrer uma subestimação de φ por ambos os modelos e, neste caso, as estimativas tendem a estar mais próximas do valor real para um valor de β menor ($\beta = 0,01$). No caso em que $\varphi = 0,50$, há uma tendência em ambos os modelos de superestimarmos φ, se $\beta = 0,20$ e de subestimarmos, se $\beta = 0,01$. Percebemos também que para valores maiores de β o modelo proposto tende a ter melhor desempenho.

As Tabelas 4-6 apresentam as estimativas de φ para o tamanho amostral $n = 100$.

Tabela 4: Estatísticas descritivas para média e moda *a posteriori* de φ, caso $\beta = 0,01$, $n = 100$.

phi	Modelo	Estimativa	Média	EQM	Desvio Padrão	Q1	Mediana	Q3
0,01	c/ erro	media	0,0603	0,0030	0,0230	0,0403	0,0550	0,0746
		moda	0,0170	0,0003	0,0170	0,0100	0,0100	0,0100
	s/ erro	media	0,0796	0,0063	0,0394	0,0388	0,0704	0,1043
		moda	0,0472	0,0030	0,0404	0,0010	0,0360	0,0720
0,10	c/ erro	media	0,1092	0,0020	0,0444	0,0798	0,1000	0,1333
		moda	0,0579	0,0050	0,0563	0,0100	0,0400	0,0900
	s/ erro	media	0,1623	0,0081	0,0657	0,1168	0,1583	0,2043
		moda	0,1326	0,0057	0,0680	0,0840	0,1310	0,1760
0,50	c/ erro	media	0,4410	0,0209	0,1320	0,3466	0,4340	0,5330
		moda	0,4506	0,0236	0,1453	0,3600	0,4500	0,5400
	s/ erro	media	0,5320	0,0129	0,1090	0,4612	0,5324	0,6040
		moda	0,5174	0,0136	0,1151	0,4440	0,5170	0,5910
0,90	c/ erro	media	0,8140	0,0150	0,0868	0,7738	0,8324	0,8794
		moda	0,8948	0,0137	0,1167	0,8200	0,9300	1,0000
	s/ erro	media	0,8380	0,0091	0,0732	0,7996	0,8534	0,8916
		moda	0,8838	0,0111	0,1046	0,8160	0,8990	0,9888
0,99	c/ erro	media	0,8598	0,0212	0,0652	0,8239	0,8786	0,9079
		moda	0,9500	0,0083	0,0824	0,9126	1,0000	1,0000
	s/ erro	media	0,8778	0,0158	0,0566	0,8482	0,8916	0,9192
		moda	0,9388	0,0091	0,0808	0,8980	0,9810	1,0000

Tabela 5: Estatísticas descritivas para média e moda *a posteriori* de φ, caso $\beta = 0,05$, $n = 100$.

phi	Modelo	Estimativa	Média	EQM	Desvio Padrão	Q1	Mediana	Q3
0,01	c/ erro	media	0,1258	0,0157	0,0478	0,0907	0,1192	0,1486
		moda	0,0643	0,0076	0,0683	0,0100	0,0300	0,1000
	s/ erro	media	0,2020	0,0416	0,0688	0,1533	0,1973	0,2440
		moda	0,1754	0,0324	0,0710	0,1270	0,1700	0,2210
0,10	c/ erro	media	0,1730	0,0097	0,0666	0,1283	0,1611	0,1993
		moda	0,1183	0,0102	0,0996	0,0100	0,1100	0,1800
	s/ erro	media	0,2734	0,0366	0,0808	0,2216	0,2656	0,3233
		moda	0,2492	0,0293	0,0837	0,1950	0,2410	0,3030
0,50	c/ erro	media	0,4655	0,0217	0,1431	0,3477	0,4594	0,5606
		moda	0,4814	0,0284	0,1676	0,3600	0,4900	0,5900
	s/ erro	media	0,5736	0,0176	0,1103	0,4936	0,5723	0,6495
		moda	0,5627	0,0178	0,1180	0,4800	0,5590	0,6420
0,90	c/ erro	media	0,7896	0,0219	0,0987	0,7413	0,8093	0,8623
		moda	0,8791	0,0167	0,1278	0,7925	0,9100	1,0000
	s/ erro	media	0,8269	0,0114	0,0781	0,7819	0,8383	0,8855
		moda	0,8645	0,0130	0,1084	0,7913	0,8740	0,9590
0,99	c/ erro	media	0,8336	0,0310	0,0807	0,7936	0,8542	0,8948
		moda	0,9325	0,0133	0,1002	0,8900	1,0000	1,0000
	s/ erro	media	0,8628	0,0205	0,0654	0,8304	0,8786	0,9128
		moda	0,9140	0,0144	0,0932	0,8575	0,9430	1,0000

Tabela 6: Estatísticas descritivas para média e moda *a posteriori* de φ, caso $\beta = 0,20$, $n = 100$.

phi	Modelo	Estimativa	Média	EQM	Desvio Padrão	Q1	Mediana	Q3
0,01	c/ erro	media	0,3547	0,1351	0,1281	0,2622	0,3309	0,4402
		moda	0,2694	0,1210	0,2320	0,0325	0,2300	0,4400
	s/ erro	media	0,5599	0,3122	0,0993	0,4927	0,5545	0,6277
		moda	0,5504	0,3027	0,1035	0,4800	0,5435	0,6200
0,10	c/ erro	media	0,3810	0,0966	0,1328	0,2733	0,3579	0,4674
		moda	0,3121	0,1040	0,2430	0,0525	0,3100	0,4900
	s/ erro	media	0,5808	0,2408	0,0982	0,5116	0,5792	0,6469
		moda	0,5722	0,2336	0,1029	0,5005	0,5690	0,6400
0,50	c/ erro	media	0,5377	0,0216	0,1419	0,4356	0,5365	0,6521
		moda	0,5875	0,0712	0,2525	0,4100	0,6000	0,7775
	s/ erro	media	0,6950	0,0469	0,0942	0,6358	0,7000	0,7638
		moda	0,6943	0,0486	0,1041	0,6290	0,6960	0,7650
0,90	c/ erro	media	0,6653	0,0713	0,1276	0,5777	0,6751	0,7656
		moda	0,8003	0,0500	0,2003	0,6700	0,8400	1,0000
	s/ erro	media	0,7819	0,0210	0,0844	0,7223	0,7888	0,8485
		moda	0,7962	0,0220	0,1057	0,7200	0,7940	0,8710
0,99	c/ erro	media	0,7009	0,0980	0,1202	0,6314	0,7112	0,7899
		moda	0,8547	0,0520	0,1839	0,7600	0,9200	1,0000
	s/ erro	media	0,8069	0,0400	0,0804	0,7579	0,8139	0,8679
		moda	0,8283	0,0370	0,1046	0,7580	0,8240	0,9000

Observando os dados da Tabela 4 que considera amostras com tamanho $n = 100$ notamos que, em média, os estimadores de Bayes, média e moda *a posteriori*, fornecem melhores estimativas para o valor real de φ no modelo com erro de classificação para $\varphi = 0,01$ e $\varphi = 0,50$. Para $\varphi = 0,10$ a média *a posteriori* fornece melhores estimativas considerando o modelo com erro enquanto que a moda *a posteriori* fornece melhores estimativas no modelo sem erro de classificação e para $\varphi = 0,90$ e $\varphi = 0,99$ nota-se que a média *a posteriori* fornece melhores estimativas no modelo sem erro enquanto que a moda *a posteriori* fornece melhores estimativas no modelo com erro. Nota-se ainda que, em média, a moda *a posteriori* fornece melhores estimativas para o verdadeiro valor de φ em ambos os modelos, exceto para $\varphi = 0,10$ no modelo com erro de classificação onde a média *a posteriori* é, em média, melhor. No entanto, se considerarmos o EQM, percebe-se que para $\varphi = 0,50$ e $\varphi = 0,90$ a média *a posteriori* fornece estimativas mais próximas do verdadeiro valor de φ em ambos os modelos. Nota-se também que o modelo com erro fornece estimativas mais próximas do valor real de φ para $\varphi = 0,01$ e $\varphi = 0,10$ enquanto que para $\varphi = 0,50$ e $\varphi = 0,90$ o modelo sem erro fornece estimativas mais próximas do verdadeiro valor do parâmetro. Algumas divergências entre o EQM e a estimativa média podem ser explicadas pela presença de estimativas atípicas. Isso pode ser observado nos box-plots que se seguem.

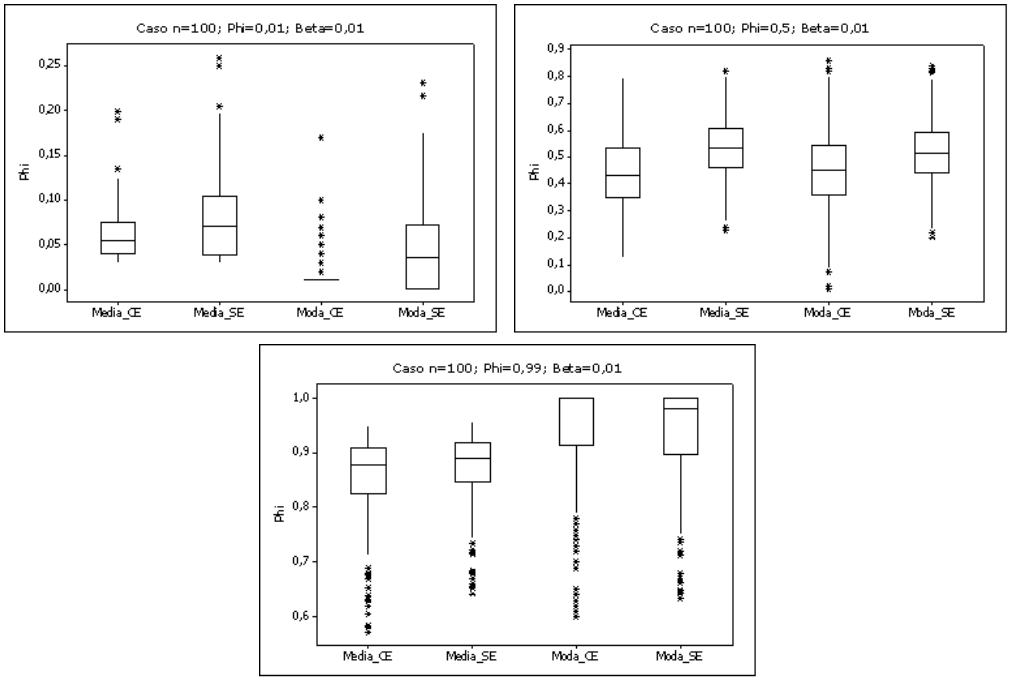

Figura 7: Box-Plot para média e moda *a posteriori* de φ, $\beta = 0,01$, $n = 100$.

Analisando os dados da Tabela 5 vemos que, em média, os estimadores de Bayes, média e moda *a posteriori* obtidos usando o modelo proposto, fornecem melhores estimativas para o valor real de φ do que se o modelo sem erro de classificação é considerado, exceto para $\varphi = 0,90$ e $\varphi = 0,99$ se a média *a posteriori* é escolhida como estimador. Estas conclusões são corroboradas pelo EQM exceto quando $\varphi = 0,50$ e $\varphi = 0,90$ em que o EQM indica o modelo sem erro como aquele que fornece melhores estimativas. Em média, em ambos os modelos, as estimativas obtidas considerando-se a moda *a posteriori* como estimador são melhores. Conclusões similares podem ser obtidas observado-se os seguintes box-plots:

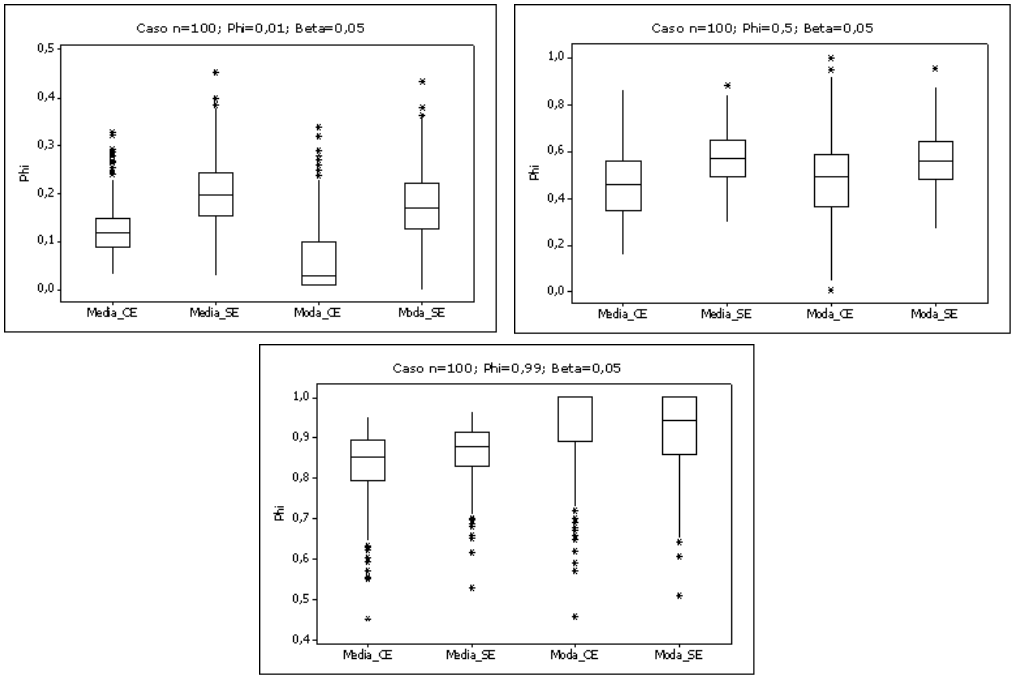

Figura 8: Box-Plot para média e moda *a posteriori* de φ, $\beta = 0,05$, $n = 100$.

Analisando os dados da Tabela 6 vemos que, em média, os estimadores de Bayes, média e moda *a posteriori*, fornecem melhores estimativas para o verdadeiro valor de φ no modelo com erro de classificação do que no modelo sem erro, exceto para $\varphi = 0,90$ e $\varphi = 0,99$ se a média *a posteriori* é considerada como estimador. Similar ao que ocorria no caso $n = 10$, ambos os modelos tendem a superestimar φ se $\varphi = 0,01, 0,10$ e $0,50$ e a subestimar, caso contrário.

Resumidamente, comparando as Tabelas 1 à 6 notamos que quando aumentamos o tamanho amostral, os estimadores tendem a fornecer, em média, melhores estimativas para o verdadeiro valor de φ em ambos os modelos. Considerando-se o EQM nota-se que as estimativas tendem a estar, em geral, mais próximas do valor real de φ quando aumentamos o tamanho da amostra. Também concluímos que o modelo proposto tende a produzir melhores estimativas para situações com maior probabilidade de ocorrer erros de classificação e para maiores tamanhos de amostra. A moda *a posteriori* tende a fornecer melhores estimativas em ambos os modelos e estas estimativas são melhores, como esperado, para amostras maiores.

8.2 As Estimativas de β

Nesta seção, apresentamos os resultados obtidos para as estimativas de β em todos os casos considerados na seção anterior. As Tabelas 7-12 apresentam algumas estatísticas descritivas para a média e moda *a posteriori* de β para $\beta = 0,01$, $\beta = 0,05$ e $\beta = 0,20$, respectivamente.

As Tabelas 7-9 apresentam as estimativas de β para o tamanho amostral $n = 10$.

Tabela 7: Estatísticas descritivas para média e moda *a posteriori* de β, caso $\beta = 0,01$, $n = 10$.

phi	Estimativa	Média	EQM	Desvio Padrão	Q1	Mediana	Q3
0,01	media	0,0865	0,0061	0,0176	0,0713	0,0800	0,0973
	moda	0,0121	0,0016	0,0395	0,0010	0,0010	0,0010
0,10	media	0,0910	0,0070	0,0212	0,0713	0,0894	0,0997
	moda	0,0261	0,0033	0,0552	0,0010	0,0010	0,0190
0,50	media	0,1081	0,0103	0,0268	0,0846	0,1104	0,1235
	moda	0,0740	0,0114	0,0855	0,0010	0,0580	0,1080
0,90	media	0,1164	0,0119	0,0248	0,0991	0,1179	0,1342
	moda	0,0904	0,0154	0,0946	0,0010	0,0800	0,1860
0,99	media	0,1173	0,0121	0,0239	0,1023	0,1179	0,1342
	moda	0,0920	0,0160	0,0962	0,0010	0,0800	0,1860

Tabela 8: Estatísticas descritivas para média e moda *a posteriori* de β, caso $\beta = 0,05$, $n = 10$.

phi	Estimativa	Média	EQM	Desvio Padrão	Q1	Mediana	Q3
0,01	media	0,0969	0,0027	0,0228	0,0800	0,0894	0,1104
	moda	0,0387	0,0046	0,0669	0,0010	0,0010	0,0580
0,10	media	0,1005	0,0032	0,0254	0,0800	0,0973	0,1220
	moda	0,0523	0,0057	0,0757	0,0010	0,0010	0,0940
0,50	media	0,1137	0,0048	0,0266	0,0973	0,1115	0,1342
	moda	0,0886	0,0095	0,0899	0,0010	0,0580	0,1630
0,90	media	0,1210	0,0057	0,0258	0,1023	0,1186	0,1361
	moda	0,1097	0,0133	0,0990	0,0010	0,0800	0,1900
0,99	media	0,1219	0,0058	0,0247	0,1023	0,1186	0,1361
	moda	0,1117	0,0136	0,0990	0,0010	0,0800	0,1900

Tabela 9: Estatísticas descritivas para média e moda *a posteriori* de β, caso $\beta = 0,20$, $n = 10$.

phi	Estimativa	Média	EQM	Desvio Padrão	Q1	Mediana	Q3
0,01	media	0,1327	0,0053	0,0286	0,1130	0,1342	0,1526
	moda	0,1494	0,0117	0,0955	0,0580	0,1736	0,2500
0,10	media	0,1308	0,0056	0,0280	0,1104	0,1342	0,1509
	moda	0,1420	0,0126	0,0961	0,0580	0,1630	0,2500
0,50	media	0,1370	0,0048	0,0280	0,1186	0,1361	0,1613
	moda	0,1633	0,0101	0,0941	0,0800	0,1860	0,2500
0,90	media	0,1394	0,0043	0,0261	0,1186	0,1362	0,1613
	moda	0,1722	0,0093	0,0929	0,0940	0,2410	0,2500
0,99	media	0,1399	0,0042	0,0260	0,1186	0,1362	0,1613
	moda	0,1730	0,0091	0,0920	0,0940	0,1900	0,2500

Analisando as Tabelas 7-9, notamos que, em média, a moda *a posteriori* fornece melhores estimativas para o verdadeiro valor de β em todos os casos. Entretanto, se considerarmos o EQM tem-se que a média *a posteriori* fornece estimativas mais próximas do valor real de β, exceto para $\beta = 0,01$ (quando $\varphi = 0,01$ ou $\varphi = 0,10$) se considerarmos a moda *a posteriori* como estimador.

Percebemos também que, *a posteriori*, para $\beta = 0,01$ e $0,05$, tanto a média quanto a moda tendem, em média, a superestimar o valor de β, sendo que o EQM tende a ser maior para valores maiores de φ. No caso em que $\beta = 0,20$ a tendência é haver uma subestimação de β. Neste caso, as melhores estimativas são observadas para maiores valores de φ.

As Tabelas 10-12 apresentam as estimativas de β para o tamanho amostral $n = 100$.

Tabela 10: Estatísticas descritivas para média e moda *a posteriori* de β, caso $\beta = 0,01$, $n = 100$.

phi	Estimativa	Média	EQM	Desvio Padrão	Q1	Mediana	Q3
0,01	media	0,0170	0,0000	0,0064	0,0118	0,0159	0,0204
	moda	0,0037	0,0000	0,0057	0,0010	0,0010	0,0030
0,10	media	0,0288	0,0004	0,0111	0,0217	0,0267	0,0334
	moda	0,0121	0,0002	0,0143	0,0010	0,0070	0,0190
0,50	media	0,0619	0,0033	0,0251	0,0431	0,0570	0,0789
	moda	0,0388	0,0023	0,0390	0,0010	0,0310	0,0700
0,90	media	0,0578	0,0030	0,0276	0,0380	0,0510	0,0717
	moda	0,0289	0,0019	0,0382	0,0010	0,0090	0,0500
0,99	media	0,0540	0,0027	0,0264	0,0354	0,0476	0,0634
	moda	0,0237	0,0014	0,0358	0,0010	0,0010	0,0350

Tabela 11: Estatísticas descritivas para média e moda *a posteriori* de β, caso $\beta = 0,05$, $n = 100$.

phi	Estimativa	Média	EQM	Desvio Padrão	Q1	Mediana	Q3
0,01	media	0,0406	0,0003	0,0157	0,0288	0,0380	0,0489
	moda	0,0268	0,0010	0,0217	0,0080	0,0240	0,0420
0,10	media	0,0517	0,0004	0,0203	0,0376	0,0476	0,0622
	moda	0,0383	0,0010	0,0290	0,0140	0,0360	0,0580
0,50	media	0,0824	0,0021	0,0325	0,0576	0,0778	0,1051
	moda	0,0679	0,0026	0,0480	0,0280	0,0660	0,1055
0,90	media	0,0801	0,0022	0,0362	0,0515	0,0718	0,1033
	moda	0,0595	0,0028	0,0525	0,0110	0,0500	0,0950
0,99	media	0,0768	0,0019	0,0348	0,0508	0,0706	0,0982
	moda	0,0545	0,0025	0,0503	0,0055	0,0460	0,0865

Tabela 12: Estatísticas descritivas para média e moda *a posteriori* de β, caso $\beta = 0,20$, $n = 100$.

phi	Estimativa	Média	EQM	Desvio Padrão	Q1	Mediana	Q3
0,01	media	0,1658	0,0022	0,0329	0,1434	0,1715	0,1910
	moda	0,1781	0,0023	0,0433	0,1470	0,1790	0,2095
0,10	media	0,1679	0,0020	0,0325	0,1482	0,1727	0,1913
	moda	0,1813	0,0021	0,0427	0,1540	0,1840	0,2108
0,50	media	0,1779	0,0014	0,0311	0,1576	0,1841	0,2009
	moda	0,1989	0,0020	0,0447	0,1680	0,2050	0,2390
0,90	media	0,1777	0,0012	0,0335	0,1594	0,1852	0,2039
	moda	0,2010	0,0024	0,0493	0,1710	0,2160	0,2495
0,99	media	0,1783	0,0014	0,0319	0,1627	0,1853	0,2019
	moda	0,2023	0,0022	0,0475	0,1753	0,2135	0,2500

Da Tabela 10 nota-se que, em média, a moda *a posteriori* fornece melhores estimativas para o verdadeiro valor de β exceto para $\varphi = 0,01$. Se considerarmos o EQM nota-se que a moda *a posteriori* fornece estimativas mais próximas do valor real de β em todos os casos. Da Tabela 11 conclusões similares podem ser tomadas em relação a tabela anterior exceto que para $\varphi = 0,10$ a média *a posteriori* fornece uma estimativa melhor para o valor real de β. Nota-se da Tabela 12 que, em média, a moda *a posteriori* fornece melhores estimativas para o valor real de β. Ao contrário, quando consideramos o EQM tem-se que a média *a posteriori* fornece estimativas mais próximas do verdadeiro valor do parâmetro. Como é esperado há uma melhora nas estimativas com o aumento do tamanho da amostra.

9 Aplicação

Nesta seção, analisaremos uma amostra de 34 pacientes com síndrome de Down considerada por Franco *et al.* (2003) considerando os dois modelos apresentados na Seção 3. Tem-se como objetivo fazer uma análise de sensibilidade para os modelos, avaliando o efeito de diferentes distribuições *a priori* nas estimativas de φ e β. Também objetiva-se comparar as estimativas obtidas via WINBUGS e os métodos computacionais usados na Seção 8 e fazer uma comparação entre os modelos com e sem erro de classificação via DIC. Na amostra encontrou-se 6, 22 e 6 pacientes com 1, 2 e 3 picos, respectivamente. Na população brasileira há seis alelos distintos cujas freqüências alélicas são $p_1 = 0,12$, $p_2 = 0,45$, $p_3 = 0,09$, $p_4 = 0,31$, $p_5 = 0,01$ e $p_6 = 0,02$. Calculando as estimativas de máxima verossimilhança de φ da população brasileira e sua variância assintótica, Franco *et al.* (2003) obteve $0,6552$ e $0,0481$, respectivamente.

Dentre as aneuploidias, uma das mais estudadas é a trissomia do cromossomo 21, que produz o fenótipo conhecido como síndrome de Down. Esta trissomia é a causa mais comum de retardamento mental de origem genética em humanos (Parra, 1999). Muitos estudos, considerando grandes grupos de pacientes com síndrome de Down e alguma informação dos pais, podem ser encontradas na literatura. Alguns deles encontram-se na Tabela 13, que mostra também as respectivas estimativas de φ e os tamanhos amostrais considerados nos respectivos estudos.

Tabela 13: Referências para algumas estimativas de φ.

Referência	Tamanho Amostral	$\hat{\varphi}$
Lorber *et al.* (1992)	52	0,5192
Pertensen *et al.* (1992)	60	0,6833
Zaragosa *et al.* (1994)	249	0,6867
Griffin (1996)	436	0,7133
Koehler *et al.* (1996)	776	0,7384
Yoon *et al.* (1996)	103	0,6893
Savage *et al.* (1998)	606	0,6930
Nicolaidis e Petersen (1998)	797	0,7189

Os estudos prévios cujos resultados são apresentados na Tabela 13 são utéis na construção de uma distribuição *a priori* mais informativa para φ. Considerando tais informações, temos que o valor esperado de φ é $0,6803$ com desvio-padrão $0,0678$. Desta forma, consideraremos distribuições *a priori* para φ centradas em torno de um valor próximo a este valor médio com diferentes variâncias. Assumiremos que, *a priori*, $\varphi \sim \mathcal{B}(a, b)$, onde a e b são os valores especificados na Tabela 14 (ver também Figura 9). A distribuição uniforme foi escolhida para avaliarmos o efeito da pouca informação nas estimativas.

Tabela 14: Sumário da distribuição *a priori* de φ para pacientes brasileiros com síndrome de Down.

Especificações *a priori*			
a	b	Média	Variância
1,0	1,0	0,500	0,080
2,0	1,0	0,667	0,060
4,0	2,0	0,667	0,030
20,0	10,0	0,667	0,007

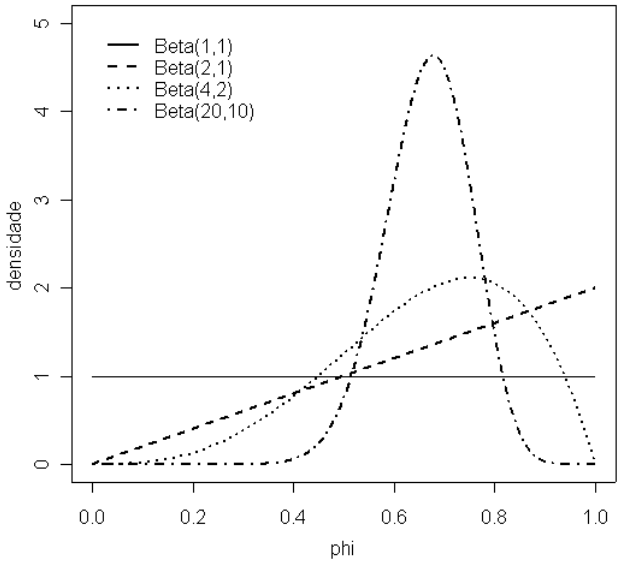

Figura 9: Distribuições *a priori* de φ.

Também assumimos que, *a priori*, a probabilidade de má classificação β tem distribuição $\mathcal{U}(0, \frac{1}{4})$. Para os métodos MCMC, implementados via WINBUGS, foram feitas 10.000 iterações com período de *burn-in* de 5.000 e *lag* 10 gerando, no final, amostras de tamanho 500. Para estas amostras verificou-se que a convergência é rápida e a amostra final é não autocorrelacionada. Na Seção 11 (Apêndice 1) encontram-se os gráficos de convergência e autocorrelação para a distribuição *a posteriori* de φ e para a distribuição *a posteriori* de β. A moda *a posteriori* das distribuições obtidas considerando os métodos MCMC foram obtidas por aproximação via método de Czuber.

Na Tabela 15 encontram-se alguns resumos da distribuição *a posteriori* de φ, entre eles média e moda *a posteriori*, obtidas via MCMC e via os Métodos Numéricos descritos na Seção 7.

Tabela 15: Sumário da distribuição *a posteriori* de φ para pacientes brasileiros com síndrome de Down.

Especificações *a priori*			MCMC			Métodos Numéricos		
a	b	Modelo	Média	Moda	Coef. Variação	Média	Moda	Coef. Variação
1,0	1,0	c/ erro	0,5558	0,6874	0,4257	0.5660	0,6100	0,4207
		s/ erro	0,6504	0,6031	0,2712	0,6549	0,6553	0,2667
2,0	1,0	c/ erro	0,6647	0,7910	0,2976	0.6662	0,7300	0,3025
		s/ erro	0,6903	0,7300	0,2425	0,7015	0,7244	0,2351
4,0	2,0	c/ erro	0,6655	0,8735	0,2359	0.6612	0,7000	0,2415
		s/ erro	0,6764	0,6592	0,1996	0,6814	0,7013	0,2049
20,0	10,0	c/ erro	0,6559	0,7164	0,1273	0.6631	0,6700	0,1234
		s/ erro	0,6665	0,6220	0,1209	0,6670	0,6753	0,1190

Perceba que as estimativas obtidas utilizando MCMC e os Métodos Numéricos são bem próximas em todos os casos. Também observamos coeficientes de variação bem próximos para tais distribuições, ou seja, apesar do erro Monte Carlo envolvido nos métodos MCMC as estimativas das distribuições *a posteriori* fornecidas por ele são muito próximas das obtidas via Métodos Numéricos, como esperado.

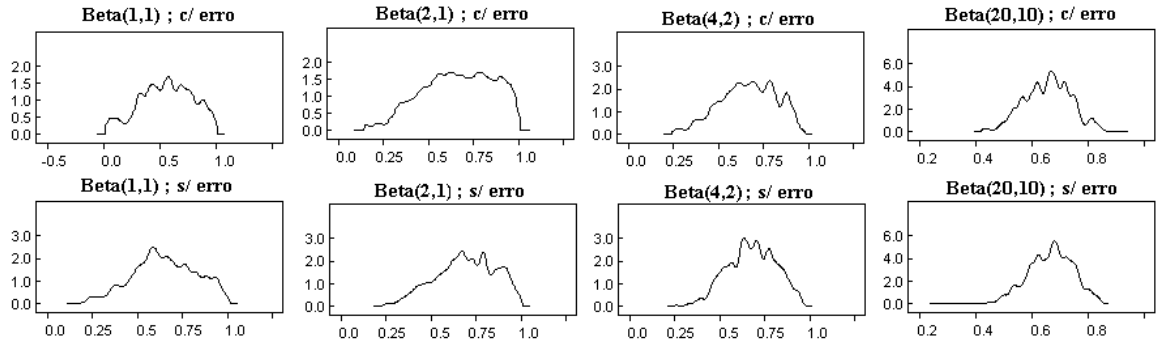

Figura 10: Distribuições *a posteriori* de φ para os modelos com e sem erro de classificação.

Perceba que, em geral, as estimativas *a posteriori* tendem a serem próximas das estimativas encontradas na literatura (ver Tabela 13). Note que as médias *a posteriori* obtidas tendem a ser menores que o valor médio das estimativas da literatura, exceto, para o modelo sem erro de classificação nos casos em que $\varphi \sim \mathcal{B}(2,1)$ usando ambos os métodos de obtenção da distribuição *a posteriori* e se $\varphi \sim \mathcal{B}(4,2)$ e usando os Métodos Numéricos. Note que se $\varphi \sim \mathcal{U}(0,1)$, as estimativas *a posteriori* no modelo sem erro de classificação são bem próximas do EMV encontrado por Franco *et al.* (2003) e, neste caso, as estimativas no modelo com erro de classificação são, em geral, bem inferiores aos valores apontados na literatura e a variância *a posteriori* é bem maior, indicando mais incerteza *a posteriori*. O modelo com erro de classificação fornece estimativas mais próximas do EMV apenas quando $\varphi \sim \mathcal{B}(20,10)$. Note que as modas *a posteriori*, em ambos os modelos, são bem próximas e tendem a ser maiores que a média *a posteriori*.

As estimativas obtidas via modelos com e sem erro de classificação tendem a tornar-se mais próximas quando a distribuição *a priori* apresenta menor variabilidade e a variabilidade *a posteriori* tende a ser menor se a distribuição *a priori* tem menor variabilidade, ou seja, maior certeza *a priori* implica em maior certeza *a posteriori*. Quando a variância *a priori* é menor, as distribuições *a posteriori* nos modelos com e sem erro de classificação tendem a ficar mais parecidas (ver Figura 10 a qual mostra também que as distribuições *a posteriori* tendem a ser unimodais e a apresentarem uma certa assimetria).

A Tabela 16 e a Figura 11 mostram, respectivamente, alguns resumos e os gráficos das distribuições *a posteriori* para a probabilidade de má classificação β.

Tabela 16: Sumário da distribuição *a posteriori* de β para pacientes brasileiros com síndrome de Down no modelo com erro de classificação.

Especificações *a priori*		MCMC			Métodos Numéricos		
a	b	Média	Moda	Coef. Variação	Média	Moda	Coef. Variação
1,0	1,0	0,0753	0,0447	0,7317	0.0802	0,0120	0,6942
2,0	1,0	0,0737	0,0266	0,7368	0.0741	0,0050	0,7267
4,0	2,0	0,0739	0,0240	0,7118	0.0718	0,0010	0,7370
20,0	10,0	0,0745	0,0099	0,7302	0.0690	0,0010	0,7531

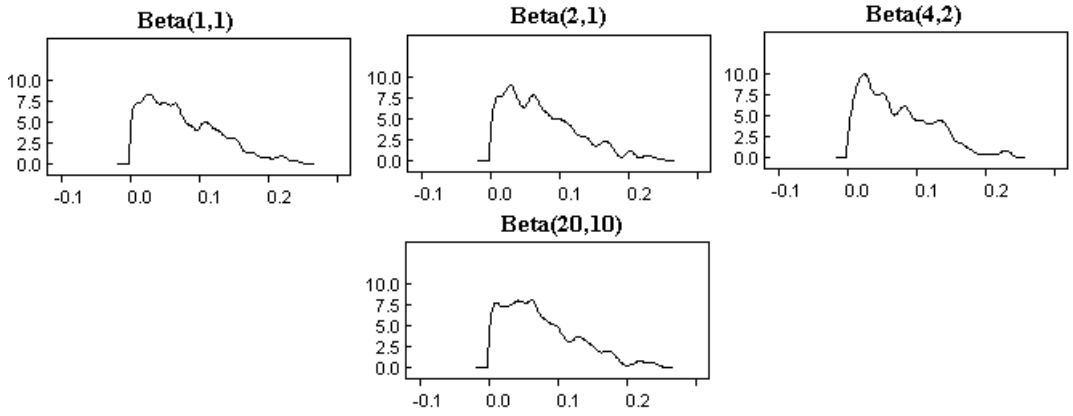

Figura 11: Distribuições *a posteriori* de β no modelo com erro de classificação.

Nota-se que os métodos MCMC e os Métodos Numéricos fornecem estimativas bem próximas em todos os casos. Quando $\varphi \sim \mathcal{B}(1,1)$ e $\varphi \sim \mathcal{B}(2,1)$, os métodos MCMC fornecem menores valores para a média *a posteriori* e a variabilidade *a posteriori* é maior indicando mais incerteza *a posteriori*. Quando $\varphi \sim \mathcal{B}(4,2)$ e $\varphi \sim \mathcal{B}(20,10)$, os métodos MCMC fornecem maiores valores para a média *a posteriori* e a variabilidade *a posteriori* é menor indicando menos incerteza *a posteriori*. A variabilidade *a posteriori* tende a ser menor se $\varphi \sim \mathcal{B}(1,1)$ para os Métodos Numéricos. Já para os métodos MCMC, a variabilidade *a posteriori* tende a ser menor se $\varphi \sim \mathcal{B}(4,2)$. As médias *a posteriori* tendem a ser menores do que as modas *a posteriori* obtidas via ambos os métodos de obtenção da distribuição *a posteriori*. As estimativas da moda *a posteriori* são bem inferiores quando consideramos os Métodos Numéricos. Da Figura 11 perceba que as distribuições *a posteriori* de β tendem a ser unimodais e assimétricas, colocando maior parte da massa em valores de β pequenos.

Tabela 17: Comparação dos modelos via *DIC*, pacientes brasileiros com síndrome de Down.

		Modelos	
a	b	c/ erro	s/ erro
1,0	1,0	9,6730	8,3610
2,0	1,0	9,3180	8,1540
4,0	2,0	9,0080	7,8070
20,0	10,0	8,5990	7,2410

Temos da Tabela 17 que os menores valores para o DIC ocorrem quando consideramos o modelo sem erro de classificação. Logo, concluímos que o modelo que leva em conta o erro de classificação não é o modelo mais adequado para tratar estes dados. Isto pode ser explicado pelo fato de que a probabilidade estimada $\widehat{\beta}$ de má classificação ser baixa (em torno de 7%, ver Tabela 16). Como foi visto no estudo de simulação apresentado na Seção 8, quando a probabilidade de má classificação é baixa o modelo com erro de classificação tende a ter pior desempenho.

10 Conclusão

Neste trabalho propomos um modelo para descrever o número de picos em um loco polimórfico em função da proporção φ da não-disjunção na meiose I levando em conta o erro de classificação que pode ser cometido na coleta dos dados. Usamos o método de Gauss-Legendre

e a Regra de Simpson para extrair informações *a posteriori* de φ bem como da probabilidade de má classificação β. Comparamos as estimativas fornecidas pelos estimadores de Bayes (média e moda *a posteriori*) nos modelos com e sem erro de classificação através de um estudo Monte Carlo considerando diferentes tamanhos amostrais, diferentes valores de φ e β e distribuições *a priori* uniforme para φ e β.

Neste estudo de simulação concluímos que quando aumentamos o tamanho amostral, os estimadores de Bayes (média e moda *a posteriori*) tendem a fornecer, em média, melhores estimativas para o verdadeiro valor de φ em ambos os modelos. Também concluímos que o modelo proposto tende a produzir melhores estimativas para situações com maior probabilidade de ocorrer erros de classificação e para maiores tamanhos de amostra. A moda *a posteriori* tende a fornecer melhores estimativas em ambos os modelos e estas estimativas são melhores para amostras maiores. Concluímos ainda que para valores maiores de β o modelo proposto tende a ter melhor desempenho. Um estudo Monte Carlo comparando estes dois modelos utilizando distribuições *a priori* Beta (mais e menos informativas) para φ é visto em Monteiro (2006). Neste trabalho concluiu-se que o modelo com erro de classificação, em geral, fornece melhores estimativas (considerando a moda *a posteriori* como estimador) que o modelo sem erro de classificação, principalmente se forem consideradas distribuições *a priori* menos informativas e/ou o tamanho amostral for grande.

Usamos o modelo proposto para analisar uma amostra de 34 pacientes brasileiros com trissomia no cromossomo 21 assumindo diferentes especificações *a priori*. Nós concluímos que as estimativas obtidas utilizando MCMC e os Métodos Numéricos são bem próximas em todos os casos e que, em geral, os Métodos Numéricos fornecem distribuições *a posteriori* com menor variabilidade. Concluímos também que as estimativas obtidas via modelos com e sem erro de classificação tendem a tornar-se mais próximas quando a distribuição *a priori* apresenta menor variabilidade e que a variabilidade *a posteriori* tende a ser menor se a distribuição *a priori* tem menor variabilidade. Notamos que quando a variância *a priori* é menor, as distribuições *a posteriori* nos modelos com e sem erro de classificação tendem a ficar mais parecidas, unimodais e apresentam uma certa assimetria. Notamos ainda que, em geral, as estimativas *a posteriori* de φ tendem a serem próximas das estimativas encontradas na literatura e que se $\varphi \sim \mathcal{U}(0,1)$, a média e a moda *a posteriori* no modelo sem erro de classificação são bem próximas do EMV encontrado por Franco *et al.* (2003) e, neste caso, as estimativas no modelo com erro de classificação são, em geral, bem inferiores aos valores apontados na literatura e a variância *a posteriori* é bem maior, indicando mais incerteza *a posteriori*. Comparamos os modelos com e sem erro de classificação via DIC e concluímos que o modelo com erro de classificação não é o mais adequado para tratar estes dados.

Tem-se como objetivo para trabalhos futuros realizar um estudo de simulação para o modelo com erro de classificação mais geral proposto na Seção 3.2. Podemos ainda fazer uma análise de sensibilidade para este modelo utilizando os dados de pacientes trissômicos usados por Franco *et al.* (2003) e compará-lo com o modelo sem erro de classificação e com a simplicação do modelo proposto (ver Seção 4) via DIC. Pretende-se também comparar os modelos com e sem erro de classificação (inclusive o modelo mais geral) utilizando outros métodos de comparação de modelos como, por exemplo, Fator de Bayes, AIC e BIC.

11 Apêndice 1: Estudo da Convergência e Autocorrelação

Nesta seção tem-se como objetivo fazer um estudo da convergência e autocorrelação das estimativas de φ nos modelos com e sem erro de classificação e da probabilidade de má classificação β obtidas via MCMC. Para gerar amostras das distribuições *a posteriori* de φ e β,

foram feitas 10.000 iterações com período de *burn-in* 5.000 e *lag* 10 gerando, ao final, amostras de tamanho 500. As Figuras 12-16 mostram os gráficos de convergência e autocorrelação para φ e β, respectivamente.

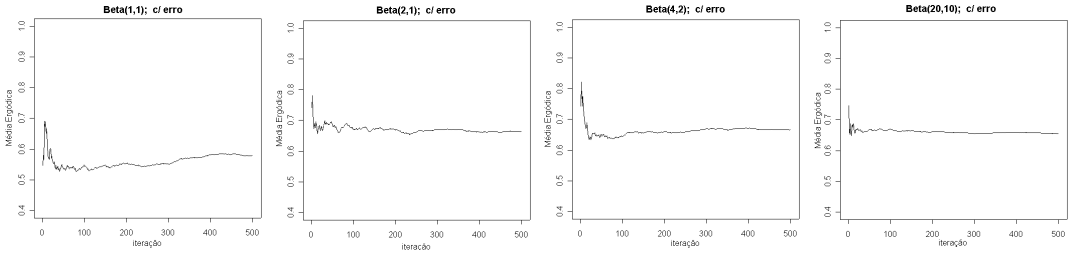

Figura 12: Gráficos de convergência para φ para o modelo com erro de classificação.

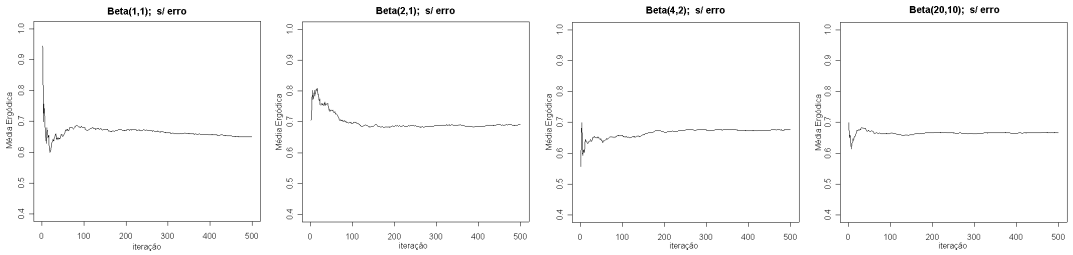

Figura 13: Gráficos de convergência para φ para o modelo sem erro de classificação.

Das Figuras 12 e 13 nota-se que a convergência dos métodos MCMC para as distribuições *a posteriori* de φ e β é rápida (por volta da iteração de número 200) em todos os casos. Perceba ainda que as estimativas tendem a convergir mais rapidamente se a variabilidade *a priori* é menor para ambos os modelos.

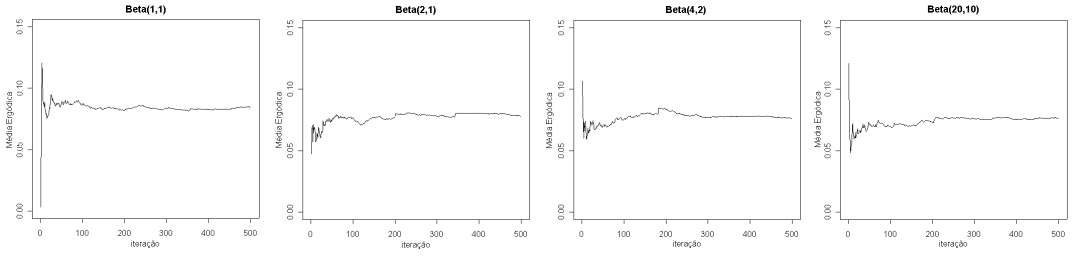

Figura 14: Gráficos de convergência para β.

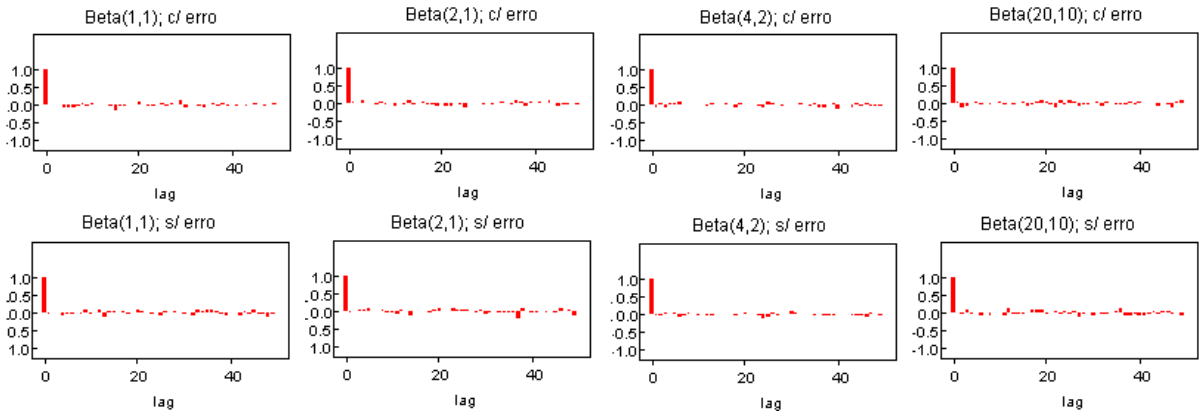

Figura 15: Funções de autocorrelação para φ para os modelos com e sem erro de classificação.

Perceba da Figura 15 que as estimativas de φ considerando-se os modelos com e sem erro de classificação são não autocorrelacionadas em todos os casos, exceto para o primeiro valor, o que é esperado já que a correlação é entre este valor e ele mesmo. Conclusões similares podem ser tomadas quando consideramos a probabilidade de má classificação β, ver Figura 16.

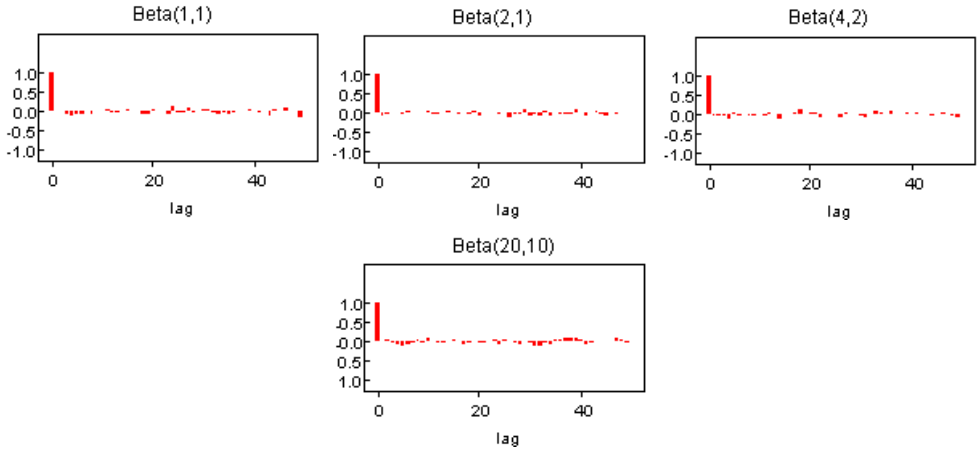

Figura 16: Funções de autocorrelação para β para o modelo com erro de classificação.

12 Apêndice 2: Programas

```
/* inclusion of standard libraries */

\#include <stdlib.h> \#include <stdio.h> \#include <math.h>

\#ifndef PI \#define PI 3.14159265358979323846 \#endif

//*****************************
//   random number generators
//*****************************

\#include "randx.c" \#define float double

//*************************************
// seeds for the random number generators
//*************************************

unsigned long runifSeed1 = 362436069, runifSeed2 = 521288629; int
rnormSeed = 13579; int rbetaSeed = 35791;

////////////////
//  FUNÇÕES   //
////////////////

//*************************************************
// FUNÇÕES PARA O MODELO COM ERRO DE CLASSIFICAÇÃO
//*************************************************

/* theta1(phi)*/

long double theta1(long double phi, float spi3, float spi2)
 {
 long double func;
 func = phi*spi3+(1-phi)*spi2;
 return (func);
 }

/* theta2(phi)*/

long double theta2(long double phi, float spi2j, float spij)
 {
 long double func;
 func = 3*phi*spi2j+(1-phi)*spij;
 return (func);
 }

/* theta3(phi)*/

long double theta3(long double phi, float spijk)
 {
 long double func;
 func = phi*spijk;
 return (func);
 }

/* Função de Verossimilhança do Modelo com erro de classificação
(modelo 1)*/

long double modelo1(long double phi, long double beta, float spi3,
float spi2, float spi2j, float spij, float spijk, int y1, int y2,
int y3)
 {
  long double func, teta1, teta2, teta3;
  teta1 = theta1 (phi, spi3, spi2);
  teta2 = theta2(phi, spi2j, spij);
```

```
teta3 = theta3(phi, spijk);
  func = pow(((1-2*beta)*teta1 + beta*(1-teta1)),y1) * pow(((1-2*beta)*teta2 +
beta*(1-teta2)),y2) * pow(((1-2*beta)*teta3 + beta*(1-teta3)),y3);
 return (func);
 }

////////////////////////////////////////////////////////////////////////
// ROTINAS DO MÉTODO DE INTEGRAÇÃO DUPLA VIA FÓRMULAS DE GAUSS LEGENDRE //
////////////////////////////////////////////////////////////////////////

/* Cálculo dos pesos para a fórmula de Gauss Legendre */

long double PesAbsGLA(int n, int k)
     {
     long double *A, z, p1, p2, p3, pp, z1 ;
     int m;
     m = ((n+1)*0.5);
     A = new long double[m+2];

     for (int i = 1; i <= m; i++)
        {
        z = cos(PI*(i-0.25)/(n+0.5));
        do
          {
          p1 = 1;
          p2 = 0;
          for (int j =1; j<=n; j++)
             {
             p3 = p2;
             p2 = p1;
             p1 = ((2*j-1)*z*p2 - (j-1)*p3)/j;
             }
          pp = n*(z*p1-p2)/(z*z - 1);
          z1 = z;
          z = z1 - p1/pp;
          } while (abs(z-z1) > pow(10, -15));

        A[m+1-i] = 2/((1-z*z)*pp*pp);
         }
     return(A[k]);
     }

/* Cálculo das abscissas para a fórmula de Gauss Legendre */

long double PesAbsGLt(int n, int k)
    {
    long double *t, z, p1, p2, p3, pp, z1 ;
    int m;
    m = ((n+1)*0.5);
    t = new long double[m+2];

    for (int i = 1; i <= m; i++)
       {
       z = cos(PI*(i-0.25)/(n+0.5));
       do
         {
         p1 = 1;
         p2 = 0;
         for (int j =1; j<=n; j++)
            {
            p3 = p2;
            p2 = p1;
            p1 = ((2*j-1)*z*p2 - (j-1)*p3)/j;
            }
```

```
pp = n*(z*p1-p2)/(z*z - 1);
         z1 = z;
         z = z1 - p1/pp;
         } while (abs(z-z1) > pow(10, -15));
       t[m+1-i] = z;
       }
    return(t[k]);
    }

/* função sinal */

int sinal (long double x)
  {
  if (x > 0) {return 1;}
  if (x < 0) {return -1;}
  if (x==0) { return 0;}
  }

////////////////////////////////////////////////////////////////////////////
/////////
// ROTINAS DO MÉTODO DE INTEGRAÇÃO SIMPLES VIA FÓRMULAS DE SIMPSON PARA CÁCULO
DA MODA //
////////////////////////////////////////////////////////////////////////////
/////////

/* Integral simples (com respeito a Beta) do numerador da posteriori
calculada pelo método de Newton para obter moda de Phi */

long double modafunc(float a, float b, long double phi, float spi3,
float spi2, float spi2j, float spij, float spijk, int n, int y1, int
y2, int y3)
   {
   long double som, par, som1, par1, fa, fb, integ1;
   som = 0;

   for (int k=1; k <= (n/2); k++)
     {
     par = a+(4*k+1)*b/(2*n);
     som = som + modelo1(phi, par,spi3,spi2,spi2j,spij,spijk,y1,y2,y3);
     }

     som1 = 0;
     for (int k=1; k <= (n/2); k++)
       {
       par1 = a+(4*k+3)*b/(2*n);
       som1 = som1 + modelo1(phi, par1,spi3,spi2,spi2j,spij,spijk,y1,y2,y3);
       }

       fa = modelo1(phi,a,spi3,spi2,spi2j,spij,spijk,y1,y2,y3);
       fb = modelo1(phi,b,spi3,spi2,spi2j,spij,spijk,y1,y2,y3);
       integ1 = ((b-a)/(3*n))*(fa + 4*som + 2*som1 + fb);

       return(integ1);
   }

/* Integral simples (com respeito a Phi) do numerador da a
posteriori calculada pelo método de Newton para obter moda de Beta
*/

long double modaBfunc(float a, float b, long double beta, float
spi3, float spi2, float spi2j, float spij, float spijk, int n, int
y1, int y2, int y3)
   {
   long double som, par, som1, par1, fa, fb, integ1;
```

```
som = 0;
   for (int k=1; k <= (n/2); k++)
     {
     par = a+(4*k+1)*b/(2*n);
     som = som + modelo1(par, beta,spi3,spi2,spi2j,spij,spijk,y1,y2,y3);
     }
     som1 = 0;
     for (int k=1; k <= (n/2); k++)
       {
       par1 = a+(4*k+3)*b/(2*n);
       som1 = som1 + modelo1(par1, beta,spi3,spi2,spi2j,spij,spijk,y1,y2,y3);
       }

       fa = modelo1(a,beta,spi3,spi2,spi2j,spij,spijk,y1,y2,y3);
       fb = modelo1(b,beta,spi3,spi2,spi2j,spij,spijk,y1,y2,y3);
       integ1 = ((b-a)/(3*n))*(fa + 4*som + 2*som1 + fb);

       return(integ1);
   }

//*************************************************************************
*****
// FUNÇÕES PARA O MODELO SEM ERRO DE CLASSIFICAÇÃO UTILIZANDO A PRIORI
UNIFORME(0,1)
//*************************************************************************
*****

/* Função da integral do denominador da f(Phi|X) */

 long double func1(long double phi, float spi3, float spi2, float spi2j, float
spij, float spijk, int y1, int y2, int y3)
   {
   long double vfunc1, teta1, teta2, teta3;
   teta1 = theta1 (phi, spi3, spi2);
   teta2 = theta2(phi, spi2j, spij);
   teta3 = theta3(phi, spijk);
   vfunc1 = (pow(teta1,y1)) * (pow(teta2,y2)) * (pow(teta3,y3));
   return (vfunc1);
   }

/* Função da integral do numerador utilizada para calcular a média
*/

 long double func2(long double phi, float spi3, float spi2, float spi2j, float
spij, float spijk, int y1, int y2, int y3)
 {
 long double vfunc2, teta1, teta2, teta3;
   teta1 = theta1 (phi, spi3, spi2);
   teta2 = theta2(phi, spi2j, spij);
   teta3 = theta3(phi, spijk);
   vfunc2 = phi*(pow(teta1,y1)) * (pow(teta2,y2)) * (pow(teta3,y3));
   return (vfunc2);
 }

long double integr1(float spi3, float spi2, float spi2j, float spij,
float spijk, int y1, int y2, int y3, int n)
    {
    long double som, par, som1, par1, fa, fb, integ1;

    som = 0;
    for (int k=1; k <= (n/2); k++)
      {
      par = 0.0000001+(4*k+1)*0.9999999/(2*n);
```

```
som = som + func1(par,spi3,spi2,spi2j,spij,spijk,y1,y2,y3);
      }

      som1 = 0;
      for (int k=1; k <= (n/2); k++)
        {
        par1 = 0.0000001+(4*k+3)*0.9999999/(2*n);
        som1 = som1 + func1(par1,spi3,spi2,spi2j,spij,spijk,y1,y2,y3);
        }

        fa = func1(0.0000001,spi3,spi2,spi2j,spij,spijk,y1,y2,y3);
        fb = func1(1,spi3,spi2,spi2j,spij,spijk,y1,y2,y3);

        integ1 = (0.9999999/(3*n))*(fa + 4*som + 2*som1 + fb);
      return(integ1);
    }

long double integr2(float spi3, float spi2, float spi2j, float spij,
float spijk, int y1, int y2, int y3, int n)
    {
    long double som, par, som1, par1, fa, fb, integ2;

    som = 0;
    for (int k=1; k <= (n/2); k++)
      {
      par = 0.0000001+(4*k+1)*0.9999999/(2*n);
      som = som + func2(par,spi3,spi2,spi2j,spij,spijk,y1,y2,y3);
      }

      som1 = 0;
      for (int k=1; k <= (n/2); k++)
        {
        par1 = 0.0000001+(4*k+3)*0.9999999/(2*n);
        som1 = som1 + func2(par1,spi3,spi2,spi2j,spij,spijk,y1,y2,y3);
        }

        fa = func2(0.0000001,spi3,spi2,spi2j,spij,spijk,y1,y2,y3);
        fb = func2(1,spi3,spi2,spi2j,spij,spijk,y1,y2,y3);
        integ2 = (0.9999999/(3*n))*(fa + 4*som + 2*som1 + fb);
      return(integ2);
    }

////////////////////////////////////////////////////////////////////////////
//                              INICIO DO PROGRAMA                        //
////////////////////////////////////////////////////////////////////////////

int main (void)

  {

  FILE *fileFreqAle;
  FILE *Estimativas;
  FILE *Estimativas2;
  FILE *Dados;
  FILE *Dadosreais;
  FILE *fatorbayes;

////////////////////////////////////////////////////////////////////
//    DECLARAÇÃO DAS VARIÁVEIS  MODELO COM ERRO DE CLASSIFICAÇÃO  //
////////////////////////////////////////////////////////////////////

int tamFA, y1, y2, y3, tamX, NRep; float *X, *Xr, *FreqAle, dummy,
un, un1; float spi3, spi2, spi2j, spij, spijk; long double beta,
phi, t1, t2, mediaphi, mediabeta;
```

```
/* Variáveis utilizadas no método Gauss Legendre para cálculo da
média */

int contador, contador1, nx, ny; float ax, bx, ay, by; long double
integral, integral1, integral2, *A, *B, *t, *u, ex1, ex2, ey1, ey2,
kx, tx, Ax, x, ky, ty, Ay, y; long double Som, Som1, Som2, Soma,
Soma1, Soma2; long double fxy, fxy1, fxy2;

/* Variáveis utilizadas para cálculo da moda */

long double vfunc, controle, modaphi, modabeta, fi , bet;

////////////////////////////////////////////////////////////////////
//    DECLARAÇÃO DAS VARIÁVEIS MODELO SEM ERRO DE CLASSIFICAÇÃO   //
////////////////////////////////////////////////////////////////////

/* Variáveis para o cálculo da Média de Phi */

long double mediafi, integ1, integ2;

/* Variáveis para o cálculo da Moda de Phi */

long double Gf, modafi;

////////////////////////////////////////////////////////////////////
//    DECLARAÇÃO DAS VARIÁVEIS - CÁLCULO DO FATOR DE BAYES         //
////////////////////////////////////////////////////////////////////

long double fatbayes, prop, contad;

////////////////////////////////////////////////////////////////////
// PARÂMETROS DE ENTRADA (INTERNOS) PARA O MÉTODO GAUSS-LEGENDRE  //
////////////////////////////////////////////////////////////////////

/*Limites de integração Phi e Beta respectivamente */

ax = 0.00001; bx = 1; ay = 0; by = 0.25;

/* Números de pontos usaddos na integração de Phi e Beta
respectivamente */

nx = 15; ny = 15;

////////////////////////////////////////////////////////////////////
//    PARÂMETROS DE ENTRADA (INTERNOS) PARA O FATOR DE BAYES    //
////////////////////////////////////////////////////////////////////

contad = 0;

////////////////////////////////////////////////////////
//  PARÂMETROS DE ENTRADA (FORNECIDOS PELO USUÁRIO)  //
////////////////////////////////////////////////////////

printf("\nForneca o numero de replicas: "); scanf("\%d",&NRep);
printf("\n\nGeracao de uma amostra da distribuicao a posteriori de F
\n\n"); printf("\nForneca o tamanho da amostra a ser gerada: ");
scanf("\%d",&tamX); printf("\nForneca o parametro Beta: ");
scanf("\%Lf",&beta); printf("\nForneca o parametro Phi: ");
scanf("\%Lf",&phi);

// Vetores //

/* Vetor dos picos reais */
```

```
Xr = new float [tamX];

/* Vetor dos picos sujeito a erro de classificação */

X = new float [tamX];

printf("\n\n"); printf("\nAguarde um instante");


////////////////////////////////////////////////////
// CÁLCULO DA SOMA DAS FREQUÊNCIAS ALÉLICAS (pi's) //
////////////////////////////////////////////////////

fileFreqAle = fopen("freqale.txt","r");

/* count frequence size */

tamFA = 0; while (fscanf(fileFreqAle,"\%lf\n", &dummy) == 1)
 {
 tamFA++;
 }

/* lendo arquivo de frequencias alelicas pi's */

FreqAle = new float[tamFA+1]; rewind(fileFreqAle); for (int k=1; k
<= tamFA; k++)
 {
 fscanf(fileFreqAle, "%lf\n", &FreqAle[k]);
 }

printf("\n\n");

spi2=0; spi3=0; spij=0; spijk=0; spi2j=0;

/* calculo da soma de pi^2 e pi^3 */

for (int i=1; i <= tamFA; i++)
 {
 spi2 += pow( FreqAle[i],2 );
 spi3 += pow( FreqAle[i],3 );
 }

/* calculo da soma de pi^2*pj e pi*pj */

for (int i=1; i <= tamFA; i++)
 {
 for(int j=1; j <=tamFA; j++ )
  {
  if(i!= j)
   {
   spij +=  FreqAle[i]*FreqAle[j];
   spi2j += pow(FreqAle[i],2) * FreqAle[j];
   }
  }
 }

/* calculo da soma de pi*pj*pk*/

for (int i=1; i <= tamFA; i++)
 {
 for(int j=1; j <= tamFA; j++ )
  {
  for(int k=1; k<= tamFA;k++)
```

```
{
   if(i!=j)
    {
     if (i!=k)
      {
       if (j!=k)
        {
        spijk +=  FreqAle[i]*FreqAle[j]*FreqAle[k];
        }
      }
    }
   }
  }
 }


////////////////////////////////////////////////////////////////////
// CÁLCULOS DOS PESOS PARA A GERAÇÃO DA DISTRIBUIÇÃO A POSTERIORI //
////////////////////////////////////////////////////////////////////

 t1 = theta1(phi, spi3, spi2);
 t2 = theta2(phi, spi2j,spij);


////////////////////////////////////////////////////////////////////
// CÁLCULO DOS PESOS E ABSCISSAS PARA A FÓRMULA DE GAUSS LEGENDRE //
////////////////////////////////////////////////////////////////////

A = new long double[((nx+1)/2)+1]; t = new long
double[((nx+1)/2)+1]; B = new long double[((ny+1)/2)+1]; u = new
long double[((ny+1)/2)+1];

contador = ((nx+1)*0.5)+ 1;

for (int i =1; i<= contador; i++)
  {
  A[i] = PesAbsGLA(nx, i);
  t[i] = PesAbsGLt(nx, i);
  }

if (ny == nx)
  {
  for (int j =1; j<= contador; j++)
    {
    B[j] = A[j];
    u[j] = t[j];
    }
  }
  else
    {
    contador1 = ((ny+1)*0.5)+1;
    for (int i =1; i<= contador1; i++)
      {
      B[i] = PesAbsGLA(ny, i);
      u[i] = PesAbsGLt(ny, i);
      }
    }

////////////////////////////////////////////////////////////////////
// CABEÇALHO DO ARQUIVO QUE VAI CONTER AS ESTIMATIVAS DE PHI E BETA //
////////////////////////////////////////////////////////////////////

 Estimativas = fopen("Estimativas.txt","w");
 fprintf(Estimativas, " ESTIMATIVAS          \n\n");
```

```
fprintf(Estimativas, "Media Phi\t      Media Beta\t     Moda Phi\t      Moda
Beta\n");


 Estimativas2 = fopen("Estimativas2.txt","w");
 fprintf(Estimativas2, " ESTIMATIVAS            \n\n");
 fprintf(Estimativas2, "Media Phi\t          Moda Phi\n");

 fatorbayes = fopen("fatorbayes.txt","w");
 fprintf(fatorbayes, " FATOR DE BAYES            \n\n");


 Dados = fopen("Data.txt","w");
 Dadosreais = fopen("Datareal.txt","w");

 for (int cont =1; cont<=NRep; cont++)
    {
    printf("\n\n Replica = %d", cont);

////////////////////////////////////////////////
//  GERAÇÃO DA AMOSTRA E CÁLCULODE y1,y2 e y3  //
////////////////////////////////////////////////

    for (int k=1;k<=tamX;k++)
       {
       un = rUnif2(&runifSeed1,&runifSeed2);
       un1 = rUnif(&rnormSeed);
       if ( (un>=0.0) & (un<=t1))
           {
           Xr[k]=1;
           if ((un1  >= 0.0) & (un1  <= beta))
              {X[k]=2;}
           if ((un1  > beta) & (un1  <= 2*beta))
              {X[k]=3;}
           if (un1 > 2*beta)
              {X[k] = 1;}
           }
       if ( (un>t1) & (un<=t1+t2))
           {
           Xr[k]=2;
           if ((un1  >= 0.0) & (un1  <= beta))
              {X[k]=1;}
           if ((un1  > beta) & (un1  <= 2*beta))
              {X[k]=3;}
           if (un1 > 2*beta)
              {X[k] = 2;}
           }
       if ( (un> (t1+t2)) & (un<=1.0))
           {
           Xr[k]=3;
           if ((un1  >= 0.0) & (un1  <= beta))
              {X[k]=1;}
           if ((un1  > beta) & (un1  <= 2*beta))
              {X[k]=2;}
           if (un1 > 2*beta)
              {X[k] = 3;}
           }

       fprintf(Dadosreais,"%1.0f\n", Xr[k]);
       fprintf(Dados,"%1.0f\n", X[k]);
       }

    fclose(Dados);
    fclose(Dadosreais);
```

```
printf("\n\n");

    y1=0;
    y2=0;
    y3=0;

    for(int k=1; k <= tamX; k++)
       {
       if(X[k]==1)
        {y1 += 1;}
        else
         {
         if(X[k]==2)
           {y2+= 1;}
         else
           { y3+=1;}
         }
       }

//////////////////////////////////////
// MODELO COM ERRO DE CLASSIFICAÇÃO  //
//////////////////////////////////////

////////////////////////////////////////////////
// Média de Phi e Beta - Método Gauss Legendre //
////////////////////////////////////////////////

   ex1 = (bx-ax)*0.5;
   ex2 = (ax+bx)*0.5;
   ey1 = (by-ay)*0.5;
   ey2 = (ay+by)*0.5;
   Soma = 0;
   Soma1 = 0;
   Soma2 = 0;

   for (int i =1; i<=nx; i++)
     {
     kx = (sinal(i - (nx + (nx+1)%2)*0.5))*(nx%2+((nx+1)%2)*0.5);
     kx = kx + i - (nx+1)*0.5;
     tx = (sinal(kx))*t[abs(kx)];
     Ax = A[abs(kx)];
     x = ex1*tx + ex2;
     Som = 0;
     Som1 = 0;
     Som2 = 0;

     for (int j =1; j<=ny; j++)
       {
       ky = (sinal (j - (ny + (ny+1)%2)*0.5))*(ny%2+((ny+1)%2)*0.5);
       ky = ky + j - (ny+1)*0.5;
       ty = (sinal(ky))*u[abs(ky)];
       Ay = B[abs(ky)];
       y = ey1*ty + ey2;
       fxy  = x*modelo1(x, y, spi3, spi2, spi2j, spij, spijk, y1, y2,y3);
       fxy1 = modelo1(x, y, spi3, spi2, spi2j, spij, spijk, y1, y2,y3);
       fxy2 = y*modelo1(x, y, spi3, spi2, spi2j, spij, spijk, y1, y2,y3);
       Som = Som +Ay*fxy;
       Som1 = Som1 +Ay*fxy1;
       Som2 = Som2 +Ay*fxy2;
       }
     Soma = Soma +Ax*Som;
     Soma1 = Soma1 +Ax*Som1;
     Soma2 = Soma2 +Ax*Som2;
     }
```

```
  integral = 0.25*(bx-ax)*(by-ay)*Soma;
    integral1 = 0.25*(bx-ax)*(by-ay)*Soma1;
    integral2 = 0.25*(bx-ax)*(by-ay)*Soma2;

    mediaphi = (integral/integral1);
    mediabeta = (integral2/integral1);

//////////////
// Moda Phi //
//////////////

    fi = 0;
    controle = 0;
    for (int i=1; i<=100; i++)
      {
      fi = fi + 0.01;
      vfunc =  modafunc(0, 0.25, fi, spi3, spi2, spi2j, spij, spijk, 10000, y1,
y2, y3);

      if (vfunc >= controle)
        {
        modaphi = fi;
        controle = vfunc;
        }
      }

///////////////
// Moda Beta //
///////////////

    bet = 0;
    controle = 0;
    for (int i=1; i<=250; i++)
      {
      bet = bet + 0.001;
      vfunc =  modaBfunc(0.00001, 1, bet, spi3, spi2, spi2j, spij, spijk, 10000,
y1, y2, y3);

      if (vfunc >= controle)
        {
        modabeta = bet;
        controle = vfunc;
        }
      }

/* Guardando as estimativas no Arquivo Estimativas */

    fprintf(Estimativas,"%Lf\t  %Lf\t  %Lf\t  %Lf\n", mediaphi, mediabeta,
modaphi, modabeta);

////////////////////////////////////
// MODELO SEM ERRO DE CLASSIFICAÇÃO //
////////////////////////////////////

/* Média Phi */

    integ1 =  integr1(spi3, spi2, spi2j, spij, spijk, y1, y2, y3, 10000);
    integ2 =  integr2(spi3, spi2, spi2j, spij, spijk, y1, y2, y3, 10000);
    mediafi = integ2/integ1;

/* Moda Phi */

    fi = 0;
```

```
controle = 0;

     for (int i=1; i<=1000; i++)
        {
      fi = fi + 0.001;
        Gf = func1(fi,spi3,spi2,spi2j,spij,spijk,y1,y2,y3)/integ1;

        if (Gf>controle)
            {
            modafi=fi;
            controle=Gf;
            }
        }

/* Guardando as estimativas no Arquivo Estimativas2 */

    fprintf(Estimativas2,"%Lf\t   %Lf\n", mediafi, modafi);

////////////////////////////////
// CÁLCULO DO FATOR DE BAYES //
////////////////////////////////

    fatbayes = integral1/integ1;
    fprintf(fatorbayes,"%Lf\n", fatbayes);

    if (fatbayes > 1)
     {
     contad = contad + 1;
     }

     }

     prop = (contad/NRep)*100;

     fprintf(fatorbayes,"%Lf\n", prop);
     fprintf(fatorbayes,"%\n\n");

   printf("\n\n");
   system("pause");

   return 0;
   }
```

```
model
 {X[1:3] ~ dmulti (pi[], N)
 teta[1] <- F*spi3 + (1-F)*spi2
 teta[2] <-3*F*spi2j + (1-F)*spij
 teta[3] <- F*spijk
 pi[1] <- (1-2*b)*teta[1] + b*(1-teta[1])
 pi[2] <- (1-2*b)*teta[2] + b*(1-teta[2])
 pi[3] <- (1-2*b)*teta[3] + b*(1-teta[3])
 F ~ dbeta (20,10)
 b ~ dunif(0,0.25)
 }
 list(N = 34, spi3 = 0.123382, spi2 = 0.3216, spi2j = 0.198218, spij = 0.6784,
spijk = 0.281964, X = c(6, 22, 6))
 list(F = 0.5, b = 0.1)
```

```
model
 {X[1:3] ~ dmulti (teta[], N)
 teta[1] <- F*spi3 + (1-F)*spi2
 teta[2] <-3*F*spi2j + (1-F)*spij
 teta[3] <- F*spijk
 F ~ dbeta (20,10)
 }
 list( N = 34, spi3 = 0.123382, spi2 = 0.3216, spi2j = 0.198218, spij = 0.6784,
spijk = 0.281964, X = c(6, 22, 6))
 list(F = 0.5)
```

Referências

[1] Campos, F. F. (2001). *Algoritmos Numéricos*. LTC, Rio de Janeiro.

[2] Dawid, A. P. (1979). Conditional Independence in Statistical Theory (with discussion). *Journal of the Royal Statistical Society*, Ser. B, 41, 1-31.

[3] Franco, G. C., Lucio, P. S., Parra, F. C. and Pena, D. J. (2003). A probability model for the meioses I non-disjunction fraction in numerical chromosomal anomalies. *Statistics in Medicine*, 22, 2015-2024.

[4] Gelfand, A. E. and Sahu, S. K. (1999). Identifiability, Improper Priors, and Gibbs Sampling for Generalized Linear Models. *Journal of the American Statistical Association*, 94(445), 247-253.

[5] Hardy, G. H. (1908). Mendelian proportions in a mixed population. *Science*, 78, 49-50.

[6] Hassold, T. J. and Hunt, P. (2001). To er (meiotically) is human: the genesis of human aneuploidy. *Nature Reviews in Genetics*, 2, 280-291.

[7] Hassold, T. J. and Jacobs, P. A. (1984). Trissomy in man. *Annual Reviews in Genetics*, 18, 69-67.

[8] Lew, R. A. and Levy, P. (1989). Estimation of Prevalence on the basis of screening tests. *Statistics in Medicine*, 8, 1225-1230.

[9] Loschi, R. H. and Franco, G. L. (2006). Bayesian Analysis for the meiosis I Non-disjunction Fraction in Numerical Chromosomal Anomalies. *Biometrical Journal*, 48(2), 220-232.

[10] Migon, H. S. and Gamerman, D. (1999). *Statistical Inference: An integrated approach*. New York: Arnold.

[11] Monteiro, J. V. D. (2006). *Avaliando o efeito da informação a priori nas estimativas da fração de não-disjunção meiótica em modelos com e sem erro de classificação*. Relatório de Iniciaço Científica-programa PIBIC/CNPq. Departamento de Estatística. Universidade Federal de Minas Gerais. Belo Horizonte.

[12] Parra, F. C. (1999). *Estudo da origem de não-disjunção meiótica em indivíduos brasileiros com trissomia do cromossomo 21 utilizando um estimador de máxima verossimilhança*. Tese de Mestrado, Departamento de Bioquímica e Imunologia, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais.

[13] Paulino, C. D., Soares, P. and Neuhaus, J. (2003). Binomial Regression with Misclassification. *Biometrics*, 59, 670-675.

[14] Pena, S. D. J. (1998). Molecular Cytogenetics I: PCR-based diagnosis of trisomies using computer-assisted laser desitometry. *Genetic Molecules Biology*, 3, 371-322.

[15] Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and Linde, A. v. d. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of The Royal Statistical Society*, 64, 583-639.

[16] Stewart, S. L., Swallen, K. C., Glaser, S. L., Horn-Ross, P. L. and West, D. W. (1998). Adjustment of Cancer Incidence Rates for Ethnic Misclassification. *Biometrics*, 54, 774-781.

[17] Swartz, T., Haitovsky, Y., Vexler, A. and Yang, T. (2004). Bayesian identifiability and misclassification in multinomial data. *The Canadian Journal of Statistics*, 32(3), 1-18.

[18] Viana, M. A. G., Ramakrishnan, V. and Levy, P. (1993). Bayesian analysis of prevalence from the results of small screening samples. *Communications in Statistics-Theory and Methods*, 22, 575-585.

[19] Viana, M. A. G. (1994). Bayesian Small-Sample Estimation of Misclassified Multinomial Data. *Biometrics*, 50, 237-243.

[20] Weinberg, W. (1908). Über den Nashweis der Vererbung beim Menschen. *Jahreshefte Verein f. Vaterl. Naturk. in Wurtemberg*, 64, 368-382.

[21] Yoon, P. W., Sherman, S. L., Taft, L. F., Gu, Y., Pettay, D., Flanders, W. D., Khoury, M. J. and Hassold, T. J. (1996). Advanced maternal age and risk of down syndrome characterized by the meiotic stage of the chromosomal error: a population based study. *American Journal of Human Genetics*, 58, 628-633.

[22] Zaragosa, M. V., Millie, E., Redline, R. W. and Hassold, T. J. (1994). Studies of nondisjunction in trissomies 2, 7, 15 e 22: does the parental origin of trissomy influence placental morphology? *Journal of Medical Genetics*, 35, 924-931.