

UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação
ISSN 0103-2569

**Variabilidade Extra-Binomial: Uso de Métodos
Bayesianos**

**Juliano José Guimarães Junqueira
Jorge Alberto Achcar**

Nº 141

RELATÓRIOS TÉCNICOS



São Carlos – SP
Mai./2001

SYSNO	124487
DATA	/ /
ICMC - SBAB	

Summary

In many applications of the binomial distribution, we can have an observed variability larger or smaller than the expected variability. This variability is defined as extra-binomial variability. Some models are introduced in the literature to fit extra-binomial variability: beta-binomial models, correlated binomial models and mixtures of binomial models. In this paper, we analyse these models using the Bayesian approach and using Markov Chain Monte Carlo (MCMC) methods. We also consider models in the presence of covariates. Two examples with real data are introduced.

Variabilidade Extra-Binomial: Uso de Métodos Bayesianos

Juliano José Guimarães Junqueira

Jorge Alberto Achcar

SCE, ICMC / USP

Caixa Postal 668

13560-970, São Carlos - SP

Resumo

Em muitas aplicações da distribuição binomial, podemos ter uma variabilidade observada maior ou menor do que a variabilidade esperada. Essa variabilidade é chamada variabilidade extra-binomial. Alguns modelos são introduzidos na literatura para ajustar a variabilidade extra-binomial: modelos beta-binomial, binomial correlacionado e modelos de misturas de distribuições binomiais. Neste artigo, analisamos esses modelos sob o enfoque Bayesiano utilizando métodos de Monte Carlo em cadeias de Markov (MCMC). Também consideramos modelos na presença de covariáveis. Dois exemplos com dados reais são introduzidos.

palavras chave: variabilidade extra-binomial, análise Bayesiana, métodos MCMC, covariáveis.

1. Introdução

A distribuição Binomial é comumente usada para dados de contagens de y_i sucessos num total de n_i ensaios independentes, onde cada ensaio admite duas respostas possíveis, sucessos ou fracassos. Contudo, em muitas aplicações podemos ter uma variabilidade observada dos dados maior do que a variabilidade esperada quando consideramos a suposição ordinária de uma distribuição binomial $b(n_i, p)$ para Y_i (variável aleatória que representa o número de sucessos em n_i ensaios), $i = 1, \dots, N$. Essa variabilidade é chamada variabilidade extra-Binomial (ver por exemplo, Skellam, 1948 ; Altham, 1978 ou Rudolfer, 1990).

Alguns modelos tem sido propostos na literatura para ajustar a variabilidade extra-Binomial. Entre eles se destacam os modelos Beta-Binomial, Binomial Correlacionado e misturas de distribuições Binomiais.

Com o modelo Beta-Binomial (ver Skellam, 1948), a variabilidade extra-Binomial é originada pela distribuição de probabilidade na probabilidade de sucesso p da distribuição Binomial. A distribuição Beta com parâmetros α e β é a distribuição escolhida para p e a função densidade de probabilidade para a variável Y_i , dado n_i , α e β , é dada por,

$$f(y_i / n_i, \alpha, \beta) = \binom{n_i}{y_i} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha + y_i)\Gamma(\beta + n_i - y_i)}{\Gamma(\alpha + \beta + n_i)} \quad (1)$$

(densidade Beta-Binomial)

O modelo Binomial Correlacionado (ver Altham, 1978 ou Kupper & Hasemam, 1978) assume que a fonte da variabilidade extra-Binomial é dada pela correlação δ entre as variáveis binárias $X_{i1}, X_{i2}, \dots, X_{in_i}$ de Y_i . Neste caso a função densidade de probabilidade de Y_i , dado n_i , δ e p , é dada por,

$$f(y_i / n_i, \delta, p) = \binom{n_i}{y_i} p^{y_i} (1-p)^{n_i - y_i} \left\{ 1 + \frac{\delta}{2p(1-p)} [(y_i - n_i p)^2 + y_i(2p-1) - n_i p^2] \right\} \quad (2)$$

Uma terceira possibilidade de modelagem para ajustar a variabilidade extra-Binomial é considerar que a contagem de sucessos y_i é supostamente gerada por uma mistura finita de distribuições Binomiais (ver Hsiao, 1994). Ou seja, a observação y_i origina-se de uma entre J categorias mas a fonte exata é desconhecida. Considerando o caso particular $J = 2$ (mistura de duas Binomiais), temos,

$$f(y_i / n_i, \theta_1, \theta_2, \pi) = \pi \binom{n_i}{y_i} \theta_1^{y_i} (1 - \theta_1)^{n_i - y_i} + (1 - \pi) \binom{n_i}{y_i} \theta_2^{y_i} (1 - \theta_2)^{n_i - y_i} \quad (3)$$

Neste artigo, esses modelos são analisados sob o enfoque Bayesiano, através dos métodos de Monte Carlo em Cadeia de Markov (MCMC). Em particular, utilizamos os algoritmos Gibbs Sampling (ver por exemplo, Gelfand & Smith, 1990) e Metropolis-Hastings (ver por exemplo, Smith & Roberts, 1993) para obter estimadores de Monte Carlo das quantidades a posteriori de interesse dos parâmetros. Utilizamos o critério de Gelman & Rubin (ver Gelman & Rubin, 1992) para o diagnóstico de convergência das amostras geradas por esses algoritmos.

2. Análise Bayesiana do modelo Beta-Binomial

Assumindo o modelo Beta-Binomial (1), consideramos as seguintes distribuições a priori para os parâmetros:

$$\begin{aligned} \text{(i)} \quad \alpha &\sim \Gamma(a_1, b_1); \quad a_1, b_1 \text{ conhecidos} \\ \text{(ii)} \quad \beta &\sim \Gamma(a_2, b_2); \quad a_2, b_2 \text{ conhecidos} \end{aligned} \quad (4)$$

onde $\Gamma(a, b)$ denota uma distribuição gama com média a/b e variância a/b^2 .

Considerando que as distribuições a priori para os parâmetros são independentes, a densidade a posteriori conjunta para α e β é dada por,

$$\pi(\alpha, \beta / \underset{\sim}{y}, \underset{\sim}{n}) \propto \alpha^{a_1-1} e^{-b_1 \alpha} \beta^{a_2-1} e^{-b_2 \beta} \left\{ \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \right\}^N \prod_{i=1}^N \left[\frac{\Gamma(\alpha + y_i) \Gamma(\beta + n_i - y_i)}{\Gamma(\alpha + \beta + n_i)} \right] \quad (5)$$

onde: $\underset{\sim}{y} = (y_1, \dots, y_N)$ e $\underset{\sim}{n} = (n_1, \dots, n_N)$.

A partir de (5), encontramos as distribuições condicionais a posteriori de α e β , que são necessárias para o uso dos métodos MCMC. Essas distribuições são dadas por,

$$\begin{aligned} \text{(i)} \quad \pi(\alpha / \beta, \underset{\sim}{y}, \underset{\sim}{n}) &\propto \alpha^{a_1-1} e^{-b_1 \alpha} \left\{ \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)} \right\}^N \prod_{i=1}^N \left[\frac{\Gamma(\alpha + y_i)}{\Gamma(\alpha + \beta + n_i)} \right] \\ &\propto \alpha^{a_1-1} e^{-b_1 \alpha} \exp \left\{ N \log[\Gamma(\alpha + \beta)] - N \log[\Gamma(\alpha)] + \sum_{i=1}^N \log[\Gamma(\alpha + y_i)] - \sum_{i=1}^N \log[\Gamma(\alpha + \beta + n_i)] \right\} \end{aligned}$$

$$\text{ou } \pi(\alpha / \beta, \underset{\sim}{y}, \underset{\sim}{n}) \propto \Gamma(a_1, b_1) \psi_1(\alpha, \beta)$$

onde:

$$\psi_1(\alpha, \beta) = \exp \left\{ N \log[\Gamma(\alpha + \beta)] - N \log[\Gamma(\alpha)] + \sum_{i=1}^N \log[\Gamma(\alpha + y_i)] - \sum_{i=1}^N \log[\Gamma(\alpha + \beta + n_i)] \right\}$$

$$\begin{aligned} \text{(ii)} \quad \pi(\beta / \alpha, \underset{\sim}{y}, \underset{\sim}{n}) &\propto \beta^{a_2-1} e^{-b_2 \beta} \left\{ \frac{\Gamma(\alpha + \beta)}{\Gamma(\beta)} \right\}^N \prod_{i=1}^N \left[\frac{\Gamma(\beta + n_i - y_i)}{\Gamma(\alpha + \beta + n_i)} \right] \\ &\propto \beta^{a_2-1} e^{-b_2 \beta} \exp \left\{ N \log[\Gamma(\alpha + \beta)] - N \log[\Gamma(\beta)] + \sum_{i=1}^N \log[\Gamma(\beta + n_i - y_i)] - \sum_{i=1}^N \log[\Gamma(\alpha + \beta + n_i)] \right\} \end{aligned}$$

$$\text{ou } \pi(\beta / \alpha, \underset{\sim}{y}, \underset{\sim}{n}) \propto \Gamma(a_2, b_2) \psi_2(\alpha, \beta)$$

onde:

$$\psi_2(\alpha, \beta) = \exp \left\{ N \log[\Gamma(\alpha + \beta)] - N \log[\Gamma(\beta)] + \sum_{i=1}^N \log[\Gamma(\beta + n_i - y_i)] - \sum_{i=1}^N \log[\Gamma(\alpha + \beta + n_i)] \right\} \quad (6)$$

Como as distribuições condicionais para α e β não apresentam formas de distribuições conhecidas, utilizamos o algoritmo Metropolis-Hastings para realizar a simulação das amostras dos parâmetros e fazer as inferências desejadas.

3. Análise Bayesiana do modelo Binomial Correlacionado

Assumindo o modelo Binomial Correlacionado (2), consideramos as seguintes distribuições a priori para os parâmetros:

- (i) $\delta \sim U(a, b)$; a, b conhecidos
 - (ii) $p \sim \text{Beta}(c, d)$; c, d conhecidos
- (7)

onde $U(a, b)$ denota uma distribuição uniforme no intervalo (a, b) e $\text{Beta}(c, d)$ denota uma distribuição beta com média $\frac{c}{c+d}$ e variância $\frac{cd}{(c+d)^2(c+d+1)}$.

Assumindo independência a priori para os parâmetros, encontramos a densidade a posteriori conjunta para δ e p dada por,

$$\pi(\delta, p / \underline{y}, \underline{n}) \propto p^{c + \sum_{i=1}^N y_i - 1} (1-p)^{d + \sum_{i=1}^N (n_i - y_i) - 1} \prod_{i=1}^N \left\{ 1 + \frac{\delta}{2p(1-p)} [(y_i - n_i p)^2 + y_i(2p-1) - n_i p^2] \right\} \quad (8)$$

onde: $\underline{y} = (y_1, \dots, y_N)$ e $\underline{n} = (n_1, \dots, n_N)$.

Portanto, as distribuições condicionais a posteriori para os parâmetros δ e p são dadas por,

(i)

$$\pi(\delta / p, y, n) \propto \prod_{i=1}^N \left\{ 1 + \frac{\delta}{2p(1-p)} \left[(y_i - n_i p)^2 + y_i(2p-1) - n_i p^2 \right] \right\}$$

$$\text{ou } \pi(\delta / p, y, n) \propto U(a, b) \psi(\delta, p)$$

(ii)

$$\pi(p / \delta, y, n) \propto p^{c + \sum_{i=1}^N y_i - 1} (1-p)^{d + \sum_{i=1}^N (n_i - y_i) - 1} \prod_{i=1}^N \left\{ 1 + \frac{\delta}{2p(1-p)} \left[(y_i - n_i p)^2 + y_i(2p-1) - n_i p^2 \right] \right\}$$

$$\text{ou } \pi(p / \delta, y, n) \propto \text{Beta} \left(c + \sum_{i=1}^N y_i, d + \sum_{i=1}^N (n_i - y_i) \right) \psi(\delta, p)$$

$$\text{onde: } \psi(\delta, p) \propto \prod_{i=1}^N \left\{ 1 + \frac{\delta}{2p(1-p)} \left[(y_i - n_i p)^2 + y_i(2p-1) - n_i p^2 \right] \right\}$$

(9)

Como as distribuições condicionais para δ e p não apresentam formas de distribuições conhecidas, utilizamos o algoritmo Metropolis-Hastings para realizar a simulação das amostras dos parâmetros e fazer as inferências desejadas.

4. Análise Bayesiana do modelo de mistura de duas distribuições Binomiais

Para simplificar a análise Bayesiana do modelo de mistura de duas distribuições Binomiais (3), introduzimos no modelo variáveis não observáveis Z_i , ($i=1, \dots, N$), também chamadas de variáveis latentes.

A variável latente Z_i , ($i=1, \dots, N$), é definida como:

$$Z_i = \begin{cases} 1 & \text{se a } i\text{-ésima observação foi gerada da primeira distribuição Binomial} \\ 0 & \text{se a } i\text{-ésima observação foi gerada da segunda distribuição Binomial} \end{cases}$$

ou seja, Z_i tem uma distribuição de Bernoulli com probabilidade de sucesso dada por,

$$h_i = P(Z_i = 1 / y_i, n_i, \theta_1, \theta_2, \pi) = \frac{\pi \binom{n_i}{y_i} \theta_1^{y_i} (1 - \theta_1)^{n_i - y_i}}{\pi \binom{n_i}{y_i} \theta_1^{y_i} (1 - \theta_1)^{n_i - y_i} + (1 - \pi) \binom{n_i}{y_i} \theta_2^{y_i} (1 - \theta_2)^{n_i - y_i}} \quad (10)$$

A função de verossimilhança simplificada com a introdução das variáveis latentes $(Z_i, i = 1, \dots, N)$ é dada por,

$$L\left(\underset{\sim}{n}, \theta_1, \theta_2, \pi / \underset{\sim}{y}, \underset{\sim}{z}\right) = \pi^{\sum_{i=1}^N z_i} (1 - \pi)^{N - \sum_{i=1}^N z_i} \prod_{i=1}^N \left[\left(\binom{n_i}{y_i} \theta_1^{y_i} (1 - \theta_1)^{n_i - y_i} \right)^{z_i} \left(\binom{n_i}{y_i} \theta_2^{y_i} (1 - \theta_2)^{n_i - y_i} \right)^{1 - z_i} \right] \quad (11)$$

A partir daí, desenvolvemos uma análise Bayesiana para o modelo de mistura de duas distribuições Binomiais, assumindo as seguintes distribuições a priori para os parâmetros:

- (i) $\theta_1 \sim \text{Beta}(a_1, b_1)$
 - (ii) $\theta_2 \sim \text{Beta}(a_2, b_2)$
 - (iii) $\pi \sim \text{Beta}(c, d)$
- (12)

Assumindo que as distribuições a priori para os parâmetros θ_1 , θ_2 e π são independentes, encontramos a densidade a posteriori conjunta para θ_1 , θ_2 e π , dada por,

$$\begin{aligned} g\left(\theta_1, \theta_2, \pi / \underset{\sim}{n}, \underset{\sim}{y}, \underset{\sim}{z}\right) &\propto \theta_1^{a_1 - 1} (1 - \theta_1)^{b_1 - 1} \theta_2^{a_2 - 1} (1 - \theta_2)^{b_2 - 1} \pi^{c - 1} (1 - \pi)^{d - 1} \pi^{\sum_{i=1}^N z_i} (1 - \pi)^{N - \sum_{i=1}^N z_i} \\ &\prod_{i=1}^N \left[\left(\binom{n_i}{y_i} \theta_1^{y_i} (1 - \theta_1)^{n_i - y_i} \right)^{z_i} \left(\binom{n_i}{y_i} \theta_2^{y_i} (1 - \theta_2)^{n_i - y_i} \right)^{1 - z_i} \right] \\ &\propto \theta_1^{a_1 + \sum_{i=1}^N y_i z_i - 1} (1 - \theta_1)^{b_1 + \sum_{i=1}^N z_i (n_i - y_i) - 1} \theta_2^{a_2 + \sum_{i=1}^N (1 - z_i) y_i - 1} (1 - \theta_2)^{b_2 + \sum_{i=1}^N (1 - z_i) (n_i - y_i) - 1} \pi^{c + \sum_{i=1}^N z_i - 1} (1 - \pi)^{d + N - \sum_{i=1}^N z_i - 1} \end{aligned} \quad (13)$$

Portanto, as distribuições condicionais necessárias para o uso dos métodos MCMC são dadas por,

$$\begin{aligned}
(i) \quad g\left(\theta_1 / \theta_2, \pi, n, z, y\right) &\propto \theta_1^{a_1 + \sum_{i=1}^N y_i z_i - 1} (1 - \theta_1)^{b_1 + \sum_{i=1}^N z_i (n_i - y_i) - 1} \\
(ii) \quad g\left(\theta_2 / \theta_1, \pi, n, z, y\right) &\propto \theta_2^{a_2 + \sum_{i=1}^N (1 - z_i) y_i - 1} (1 - \theta_2)^{b_2 + \sum_{i=1}^N (1 - z_i) (n_i - y_i) - 1} \\
(iii) \quad g\left(\pi / \theta_1, \theta_2, n, z, y\right) &\propto \pi^{c + \sum_{i=1}^N z_i - 1} (1 - \pi)^{d + N - \sum_{i=1}^N z_i - 1}
\end{aligned} \tag{14}$$

Como as distribuições condicionais para os parâmetros θ_1 , θ_2 e π apresentam formas de distribuições Beta, utilizamos o algoritmo Gibbs Sampling para realizar a simulação das amostras dos parâmetros e fazer as inferências desejadas.

O algoritmo computacional para gerar amostras dos parâmetros θ_1 , θ_2 e π baseado na introdução de variáveis latentes é dado pelas seguintes etapas:

(i) Dados os valores iniciais $\theta_1^{(0)}$, $\theta_2^{(0)}$ e $\pi^{(0)}$, gerar uma amostra $(z_1, z_2, \dots, z_N)$ a partir da distribuição de Bernoulli p/ z_i com probabilidade de sucesso,

$$h_i = \frac{\pi^{(0)} \binom{n_i}{y_i} (\theta_1^{(0)})^{y_i} (1 - \theta_1^{(0)})^{n_i - y_i}}{\pi^{(0)} \binom{n_i}{y_i} (\theta_1^{(0)})^{y_i} (1 - \theta_1^{(0)})^{n_i - y_i} + (1 - \pi^{(0)}) \binom{n_i}{y_i} (\theta_2^{(0)})^{y_i} (1 - \theta_2^{(0)})^{n_i - y_i}}, \quad i = 1, \dots, N.$$

(ii) Gerar $\theta_1^{(1)}$ a partir da distribuição Beta $\left(a_1 + \sum_{i=1}^N y_i z_i, b_1 + \sum_{i=1}^N z_i (n_i - y_i)\right)$

(iii) Gerar $\theta_2^{(1)}$ a partir da distribuição Beta $\left(a_2 + \sum_{i=1}^N y_i (1 - z_i), b_2 + \sum_{i=1}^N (1 - z_i) (n_i - y_i)\right)$

(iv) Gerar $\pi^{(1)}$ a partir da distribuição Beta $\left(c + \sum_{i=1}^N z_i, d + N - \sum_{i=1}^N z_i\right)$

Repetir (i)-(iv) até gerar uma amostra de tamanho S.

Resultados similares são obtidos para misturas finitas de mais de duas distribuições binomiais.

5. Discriminação dos modelos

A verossimilhança marginal é uma ferramenta de grande importância numa análise Bayesiana quando estamos interessados na construção de testes de hipóteses e na discriminação dos modelos. Para um modelo representado por M , a verossimilhança marginal é definida na forma,

$$p\left(\underline{y}/M\right)=\int L\left(\underline{\theta}_M,M\right)\pi\left(\underline{\theta}_M/M\right)p\left(\underline{\theta}_M\right) d\underline{\theta}_M \quad (15)$$

onde: $L\left(\underline{\theta}_M,M\right)$ é a função de verossimilhança para o modelo M ,

$\pi\left(\underline{\theta}_M/M\right)$ é a distribuição a priori conjunta para os parâmetros do modelo M ,

e $\underline{\theta}_M$ é o vetor de parâmetros do modelo M .

Quando nosso objetivo é a comparação de dois modelos M_1 e M_2 , utilizamos o fator de Bayes B_{12} , que é definido como a razão das verossimilhanças marginais desses modelos, ou seja:

$$B_{12}=\frac{p\left(\underline{y}/M_1\right)}{p\left(\underline{y}/M_2\right)} \quad (16)$$

Desta forma, o fator de Bayes é um valor numérico baseado nos dados, que está a favor de um determinado modelo estatístico e contra um outro modelo. Uma interpretação do fator de Bayes sugerido por Kass e Raftery (1995), é dado pela tabela a seguir.

Tabela 1: Interpretação do fator de Bayes.

B_{12}	$2\log B_{12}$	Evidência a favor de M_1
< 1	< 0	Negativa (apoia M_2)
$1 \mapsto 3$	$0 \mapsto 2$	Vale ser mencionada
$3 \mapsto 20$	$2 \mapsto 6$	Positiva
$20 \mapsto 150$	$6 \mapsto 10$	Forte
> 150	> 10	muito forte

Pela tabela 1, concluímos que quando a verossimilhança marginal do modelo M_1 for maior que a verossimilhança marginal do modelo M_2 , ou seja quando $B_{12} > 1$, o modelo M_1 apresenta um melhor ajuste dos dados.

Para o caso de r modelos, $M_1, M_2, \dots, M_r$, escolhemos aquele modelo que apresentar a maior verossimilhança marginal.

6. Variabilidade extra-Binomial na presença de covariáveis

Se variáveis explanatórias estão presentes, usualmente consideramos regressão logística para modelagem. A variação residual, contudo, pode ser maior que a variação esperada, o que caracteriza a existência da variabilidade extra-Binomial (ver por exemplo, Williams, 1982).

Na presença de um vetor $\tilde{x} = (x_1, \dots, x_k)$ de covariáveis, o modelo de regressão logística é dado por,

$$y_i / n_i, y_i, x_i \sim b(n_i, p_i) \quad (17)$$

$$\text{onde } p_i = \frac{e^{\beta' x_i}}{1 + e^{\beta' x_i}}, \quad \beta' x_i = \beta_0 + \sum_{l=1}^k \beta_l x_{li}, \quad i = 1, \dots, N$$

Vamos assumir os mesmos valores para as covariáveis $\tilde{x}_i = (x_{1i}, \dots, x_{ki})$ para as n_i observações na distribuição $b(n_i, p)$, $i = 1, \dots, N$.

A variação residual contudo pode ser maior do que a variação esperada assumindo o modelo Binomial, quando todas as covariáveis do modelo são ajustadas.

Para modelar a variação extra-Binomial na presença de covariáveis, assumimos o modelo Binomial Correlacionado (2), onde:

$$P(Y_i = y_i / n_i, \tilde{x}_i, \beta, \delta) = \binom{n_i}{y_i} p_i^{y_i} (1 - p_i)^{n_i - y_i} A_i(n_i, y_i, \tilde{x}_i, \beta, \delta) \quad (18)$$

$$\text{onde } A_i(n_i, y_i, \tilde{x}_i, \beta, \delta) = 1 + \frac{\delta}{2p_i(1-p_i)} [(y_i - n_i p_i)^2 + y_i(2p_i - 1) - n_i p_i^2],$$

$$p_i = \frac{e^{\beta' \tilde{x}_i}}{1 + e^{\beta' \tilde{x}_i}}, \quad \beta' = (\beta_0, \beta_1, \dots, \beta_k), \quad i = 1, \dots, N.$$

Observar que no modelo (18), assumimos que a correlação intra-classe entre as variáveis Binárias $U_{i1}, \dots, U_{in_i}$ de $Y_i = \sum_{l=1}^{n_i} U_{il}$ é homogênea e independente do valor das covariáveis $\tilde{x}_i$, $i = 1, \dots, N$.

Assumindo independência a priori entre os parâmetros, consideramos as seguintes distribuições a priori para os parâmetros:

- (i) $\beta_0 \sim N(\mu_0, \sigma_0^2)$; μ_0, σ_0^2 conhecidos;
- (ii) $\beta_l \sim N(\mu_l, \sigma_l^2)$; μ_l, σ_l^2 conhecidos;
- (iii) $\delta \sim U(a, b)$; a, b conhecidos;

onde $l = 1, 2, \dots, k$, e $N(u, \sigma^2)$ denota uma distribuição normal com média u e variância σ^2 .

Assumindo as distribuições a priori dadas em (19), utilizamos os métodos MCMC para realizar a simulação dos parâmetros e obter as inferências desejadas.

7. Alguns Exemplos

7.1 – Um exemplo com dados genéticos

Nesta aplicação, consideramos os dados genéticos introduzidos por Skellam (1948), onde o objetivo é a associação secundária de cromossomos numa espécie de couve flor e repolho, chamada Brassica.

As unidades observadas são 337 núcleos, sendo que em cada núcleo existem 3 pares de cromossomos durante a meiose. Cada par pode ou não mostrar uma associação entre esses cromossomos, ou seja em cada núcleo podemos ter 0, 1, 2 ou 3 pares de cromossomos associados. As frequências observadas com 0, 1, 2 e 3 pares associados nesses núcleos são dadas por 32, 103, 122 e 80, respectivamente.

Observamos que neste exemplo, a variável Y_i representa o número de pares de cromossomos associados no i -ésimo núcleo, $i = 1, \dots, 337$. Assumimos para Y_i os modelos Binomial, Beta-Binomial (1), Binomial Correlacionado (2) e mistura de duas distribuições Binomiais (3), utilizando os métodos Clássicos e Bayesianos.

Na tabela 2 apresentamos os resultados das análises Clássicas dos modelos, considerando os dados genéticos.

Tabela 2 – Resultados das análises Clássicas dos modelos para os dados genéticos.

Modelo	Parâmetro	EMV	Int. Conf. (95%)
Binomial	p	0,5806	(0,5502 ; 0,6110)
Beta	α	6,1198	(0,9563 ; 11,2833)
Binomial	β	4,4186	(0,6945 ; 8,1427)
Binomial	p	0,5809	(0,5480 ; 0,6139)
Correlacionado	δ	0,0883	(0,0214 ; 0,1553)
Mistura de 2	θ_1	0,5138	(0,3738 ; 0,6538)
Distribuições	θ_2	0,8941	(0,3644 ; 1)
Binomiais	π	0,8242	(0,2911 ; 1)

Na análise Bayesiana consideramos os seguintes valores para os hiperparâmetros: modelo Beta-Binomial ($\alpha_1 = 90$, $b_1 = 15$, $\alpha_2 = 70$ e $b_2 = 16$), modelo Binomial Correlacionado ($a = 30$, $b = 20$, $c = 0$ e $d = 0,15$) e modelo de mistura de duas distribuições Binomiais ($\alpha_1 = 30$, $b_1 = 30$, $\alpha_2 = 20$, $b_2 = 2$, $c = 18$ e $d = 4$). Esses valores para os hiperparâmetros foram escolhidos após uma análise preliminar dos dados.

Utilizamos os métodos MCMC para gerar amostras das distribuições a posteriori marginais dos parâmetros.

Em todos os modelos geramos 5 cadeias de 2000 valores para cada parâmetro, utilizando as distribuições condicionais a posteriori. Para cada cadeia descartamos as primeiras 1000 iterações e consideramos a 10.^a, 20.^a,....., iterações. A convergência do algoritmo foi verificada graficamente e usando o critério de Gelman e Rubin (1992).

As estimativas Bayesianas de todos os modelos considerando os dados genéticos encontram-se na tabela 3. Também são dados na tabela 3 os valores do critério de Gelman e Rubin (GR) para todos os parâmetros.

Tabela 3 - Sumários a posteriori para os dados genéticos

Modelo	Parâmetro	Média	DP	Int. Cred. (95%)	GR
Binomial	p	0,5805	0,0155	(0,5501 ; 0,6108)	
Beta	α	6,0839	0,5028	(5,0783 ; 7,0830)	1,0016
Binomial	β	4,3901	0,3759	(3,7198 ; 5,2018)	1,0014
Binomial	p	0,0883	0,0294	(0,0309 ; 0,1429)	1,0002
Correlacionado	δ	0,5806	0,0166	(0,5487 ; 0,6110)	1,0026
Mistura de 2	θ_1	0,5159	0,0241	(0,4691 ; 0,5624)	1,0010
Distribuições	θ_2	0,9067	0,0529	(0,7801 ; 0,9845)	1,0020
Binomiais	π	0,8316	0,0574	(0,6860 ; 0,9239)	1,0033

A tabela 4 apresenta o valor da verossimilhança marginal exata do modelo Binomial e as estimativas de Monte Carlo dos outros modelos obtidas para os dados genéticos.

Tabela 4 – Discriminação dos modelos para os dados genéticos

Modelo	Estimativa da Ver. Marginal
Binomial	$0,2149 \times 10^{-192}$
Beta-Binomial	$1,4615 \times 10^{-190}$
Binomial Correlacionado	$1,1294 \times 10^{-190}$
Mistura de 2 distribuições Binomiais	$1,4958 \times 10^{-190}$

Observamos que os modelos Beta-Binomial, Binomial Correlacionado e Mistura de 2 distribuições Binomiais tiveram um comportamento semelhante no ajuste dos dados genéticos, ao passo que ambos superaram o ajuste do modelo Binomial, indicando a presença da variabilidade extra-Binomial nesses dados.

7.2 – Um exemplo com covariáveis

Consideramos o experimento analisado por Crowder (1978), onde uma quantidade de sementes é colocada numa placa coberta com um extrato numa dada diluição. Os números de sementes que germinaram e não germinaram são anotados. Foram considerados dois tipos de sementes (*O. Aegyptiaca* 75 e *O. Aegyptiaca* 73), dois tipos de extratos (feijão e vagem) e algumas réplicas para as combinações.

Apresentamos os dados desse experimento na tabela 5. Para cada réplica, temos o número total de sementes (n_i), o número de sementes que germinaram (y_i) e a proporção de sementes que germinaram (y_i/n_i).

Tabela 5 – Dados de Crowder (1978)

(I) <i>O. Aegyptiaca</i> 75						(II) <i>O. Aegyptiaca</i> 73					
Feijão			Vagem			Feijão			Vagem		
n_i	y_i	y_i/n_i	n_i	y_i	y_i/n_i	n_i	y_i	y_i/n_i	n_i	y_i	y_i/n_i
39	10	0,26	6	5	0,83	16	8	0,50	12	3	0,25
62	23	0,37	74	53	0,72	30	10	0,33	41	22	0,54
81	23	0,28	72	55	0,76	28	8	0,29	30	15	0,50
51	26	0,51	51	32	0,63	45	23	0,51	51	32	0,63
39	17	0,44	79	46	0,58	4	0	0	7	3	0,43
			13	10	0,77						

Numa inspeção dos dados da tabela 5, observamos uma significativa heterogeneidade entre as proporções das réplicas e consideramos na análise dos dados a presença de duas covariáveis x_{1i} e x_{2i} : x_{1i} assume os valores -1 para o tipo O. Aegyptiaca 75 e 1 para o tipo O. Aegyptiaca 73 e x_{2i} assume os valores -1 para o extrato de feijão e 1 para o extrato de vagem.

Assumimos para a variável Y_i (número de sementes que germinaram em n_i sementes), os modelos de regressão Logística (17) sem assumir a presença de variabilidade extra-binomial e o modelo Binomial Correlacionado (18), considerando as covariáveis x_{1i} , x_{2i} e a interação $x_{1i} x_{2i}$:

$$p_i = \frac{e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i}}}{1 + e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i}}}, \text{ para } i = 1, \dots, 21 \quad (20)$$

Na tabela 6 apresentamos os resultados das análises Clássicas dos modelos, considerados com covariáveis.

Tabela 6 – Resultados das análises Clássicas dos modelos para os dados da tabela 5.

Modelo	Parâmetro	Média	Int. Conf (95%)
Regressão Logística	β_0	-0,0206	(-0,1708 ; 0,1296)
	β_1	-0,1216	(-0,2718 ; 0,0286)
	β_2	0,4646	(0,3144 ; 0,6148)
	β_3	-0,1945	(-0,3447 ; -0,0443)
Binomial Correlacionado	δ	0,02085	(0,0033 ; 0,0384)
	β_0	-0,0562	(-0,2410 ; 0,1286)
	β_1	-0,1557	(-0,3422 ; 0,0307)
	β_2	0,4214	(0,2279 ; 0,6150)
	β_3	-0,2003	(-0,3843 ; -0,0163)

Na análise Bayesiana consideramos após uma análise preliminar dos dados, os seguintes valores para os hiperparâmetros: modelo de regressão logística ($\mu_0 = 0,05$, $\sigma_0^2 = 0,2$, $\mu_1 = -0,2$, $\sigma_1^2 = 0,2$, $\mu_2 = 0,3$, $\sigma_2^2 = 0,2$, $\mu_3 = -0,3$ e $\sigma_3^2 = 0,2$) e Binomial Correlacionado ($\mu_0 = -0,05$, $\sigma_0^2 = 0,1$, $\mu_1 = -0,15$, $\sigma_1^2 = 0,1$, $\mu_2 = 0,42$, $\sigma_2^2 = 0,1$, $\mu_3 = -0,2$, $\sigma_3^2 = 0,1$, $a = 0,015$ e $b = 0,025$).

Utilizamos os métodos MCMC para gerar amostras das distribuições a posteriori marginais dos parâmetros.

Nos dois modelos geramos 5 cadeias de 2000 valores para cada parâmetro, utilizando as distribuições condicionais a posteriori. Para cada cadeia descartamos as primeiras 1000 iterações e consideramos a 10.^a, 20.^a,....., iterações. Usando o critério de Gelman e Rubin (1992), foi observado o convergência do algoritmo Gibbs Sampling (valores do critério Gelman e Rubin, GR são dados na tabela 7).

As estimativas Bayesianas dos modelos considerando a presença de covariáveis são dadas na tabela 7.

Tabela 7 – Sumários a posteriori para o exemplo com covariáveis

Modelo	Parâmetro	Média	DP	IC (95%)	GR
Regressão Logística	β_0	-0,0144	0,0691	(-0,1501 ; 0,1265)	1,0002
	β_1	-0,1268	0,0711	(-0,2638 ; 0,0144)	1,0005
	β_2	0,4398	0,0648	(0,3163 ; 0,5666)	1,0006
	β_3	-0,2155	0,0704	(-0,3604 ; -0,0865)	1,0010
Binomial Correlacionado	δ	0,0201	0,0028	(0,0152 ; 0,0248)	0,9999
	β_0	-0,0503	0,0658	(-0,1878 ; 0,0702)	0,9998
	β_1	-0,16	0,0635	(-0,2857 ; -0,0447)	1,0001
	β_2	0,423	0,0626	(0,3109 ; 0,5532)	1,0007
	β_3	-0,1998	0,0615	(-0,3256 ; -0,0704)	1,0004

A tabela 8 apresenta as estimativas de Monte Carlo da verossimilhança marginal obtidas para este exemplo com covariáveis.

Tabela 8 – Discriminação dos modelos para o exemplo com covariáveis

Modelo	Estimativa da Ver. Marginal
Regressão Logística	$4,1349 \times 10^{-25}$
Binomial Correlacionado	$6,6823 \times 10^{-24}$

Observamos pelo resultados da tabela 8 que o modelo Binomial Correlacionado na presença de covariáveis, se ajustou melhor ao dados introduzidos por Crowder (1978) do que o

modelo de Regressão Logística. Também observamos que considerando o modelo Binomial Correlacionado, as covariáveis x_{1i} , x_{2i} e a interação $x_{1i} x_{2i}$ tem efeitos significativos.

8. Conclusões

O uso dos modelos para ajustar a variabilidade extra-Binomial é necessário em muitas aplicações, quando a variação observada é maior do que a variação esperada a partir da suposição Binomial. Também, em muitas aplicações, podemos ter a presença de covariáveis.

O uso dos métodos MCMC é apropriado para obter sumários a posteriori de interesse desses modelos.

Referências

- Altham, P.M.E (1978). Two Generalizations of the Binomial Distribution. *Applied Statistics*, n.27, p.162-167.
- Crowder, M.J. (1978). Beta-Binomial Anova for Proportions. *Applied Statistics*, n.27, p.34-37.
- Gelfand, A.E.; Smith, A.F.M. (1990). Sampling Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, n.85, p.398 – 409.
- Gelman, A.E.; Rubin, D. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Sciences*, n.7, p.457-472.
- Hsiao, C.K. (1994). Bayesian Tests of Extra-Binomial Variability with Emphasis on the Boundary Case, *PhD thesis*, Carnegie-Mellon University, U.S.A.
- Kass, R.E.; Raftery, A.E. Bayes Factors and Model Uncertainty. *Journal of the American Statistical Association*, n.90, p.773-775, 1995.
- Kupper, L.L.; Haseman, J.K. (1978). The Use of a Correlated Binomial Model for the Analysis of Certain Toxicological Experiments. *Biometrics*, n.34, p.69–76.
- Rudolfer, S.,M. (1990). A Markov Chain Model of Extra-Binomial Variation. *Biometrika*, n.77, p.255-264.
- Skellam, J.G. (1948). A Probability Distribution Derived from the Binomial Distribution by Regarding the Probability of Success as Variable between the Sets of Trials. *Journal of the Royal Statistical Society*, B, p.257-261.
- Smith, A.F.M.; Roberts, G.O. (1993). Bayesian Methods via the Gibbs Sampler and related Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society*, B, n.55, p.3-23.

Williams, D.A. (1982). Extra-Binomial Variation in Logistic Linear Models. *Applied Statistics*, p.31, n.144-148.