UNIVERSIDADE FEDERAL DE LAVRAS

VANESSA SIQUEIRA PERES DA SILVA

MODELO COM DISTRIBUIÇÃO POISSON INFLACIONADA DE ZEROS: ESTUDO DO TESTE VANDENBROEK E APLICAÇÃO USANDO ARGUMENTO BAYESIANO

LAVRAS - MG

2011

VANESSA SIQUEIRA PERES DA SILVA

MODELO COM DISTRIBUIÇÃO POISSON INFLACIONADA DE ZEROS: ESTUDO DO TESTE VANDENBROEK E APLICAÇÃO USANDO ARGUMENTO BAYESIANO

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

Orientador

Dr. Marcelo Ângelo Cirillo

Coorientadora

ra. Juliana Garcia Cespedes

LAVRAS - MG

2011

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

Silva, Vanessa Siqueira Peres da.
Modelo com distribuição Poisson Inflacionada de Zeros : estudo do teste Vandenbroek e aplicação usando argumento bayesiano / Vanessa Siqueira Peres da Silva. – Lavras : UFLA, 2011.
90 p. : il.

Dissertação (mestrado) – Universidade Federal de Lavras, 2011.
Orientador: Marcelo Ângelo Cirillo.
Bibliografia.

1. Poder do teste. 2. Teste escore. 3. Erro tipo I. 4. Distribuição ZIP. I. Universidade Federal de Lavras. II. Título.

CDD – 519.24

VANESSA SIQUEIRA PERES DA SILVA

MODELO COM DISTRIBUIÇÃO POISSON INFLACIONADA DE ZEROS: ESTUDO DO TESTE VANDENBROEK E APLICAÇÃO USANDO ARGUMENTO BAYESIANO

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

APROVADA em 10 de fevereiro de 2011.

Dra. Juliana Garcia Cespedes	UNIFESP
Dr. Mário Javier Ferrua Vivanco	UFLA
Dr. Fortunato Silva de Menezes	UFLA

Dr. Marcelo Ângelo Cirillo
Orientador

LAVRAS – MG

2011

Ao meu marido, **MICHAEL**, *minha outra metade, razão da minha vida.*

Ofereço e Dedico.

AGRADECIMENTOS

A Deus, por ter me dado forças para terminar este trabalho.

À Universidade Federal de Lavras (UFLA) e ao Departamento de Ciências Exatas (DEX), pela oportunidade concedida para a realização do mestrado.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes/REUNI), pela concessão de bolsa de estudos.

Aos professores do Departamento de Ciências Exatas da UFLA, pelos ensinamentos transmitidos e harmoniosa convivência, em especial ao professor Dr. Augusto Ramalho de Morais, pelos conhecimentos repassados na disciplina de Modelos Lineares.

Ao meu orientador, professor Dr. Marcelo Ângelo Cirillo, pela orientação, paciência, amizade, dedicação e pelos ensinamentos que foram de grande relevância para a realização deste trabalho e meu crescimento profissional.

A minha coorientadora, professora Dra. Juliana Garcia Céspedes, pela orientação, dedicação, acolhimento em sua casa em Itajubá, atenção, paciência e ensinamentos que foram de grande valia para a realização deste trabalho.

Aos professores Mário Vivanco, Juliana Cespedes e Fortunato Menezes, que aceitaram o convite para participar da banca examinadora, pelos comentários e observações que certamente contribuirão para o aprimoramento desta pesquisa.

A minha avó Abadia (*in memoriam*), que me ensinou a orar a Deus, pelo exemplo de vida.

A minha tia Cida, segunda mãe, pelo carinho, amor, dedicação, apoio, ajuda financeira nos momentos em que mais precisei e pelas orações constantes.

Aos meus pais, Paulo e Maria Terezinha, pela educação, amor, cuidado e carinho.

Ao meu irmão, Tiago, pelo apoio.

A todos meus familiares que torceram por mim e a todos aqueles que contribuíram de alguma forma para o meu crescimento e sucesso. DEUS os abençoe e proteja.

Aos amigos Hernani, Augusto, Leandro, Ana Lúcia, Felipe, Caroline (Carol), Izabela, Luzia e Andressa, pelos momentos alegres que passamos juntos e por tudo que fizeram por mim durante esta caminhada. As minhas amigas da República: Adriana (Dri), Thalita (Thá) e Thaís (Thatá), por terem me aceitado com eu sou, pela companhia, pelo carinho e por todos os ensinamentos de vida.

A todos os colegas do mestrado e doutorado, professores e funcionários do DEX, pela convivência e aprendizagem. Ao funcionário da biblioteca Carlos Rogério Coelho, por ter sido tão prestativo na pesquisa de artigos e teses para a realização deste trabalho.

Sou eternamente grata a todos que me ajudaram a trilhar esse caminho.

Vanessa Siqueira Peres da Silva.

RESUMO

A presente dissertação está organizada em dois artigos. No primeiro artigo o teste de escore, proposto por Vandenbroek, é avaliado em função das probabilidades do controle do erro tipo I e poder. O objetivo foi verificar qual o tamanho de amostra necessário para utilizar o teste de Vandenbroek na discriminação entre os modelos com distribuição Poisson e ZIP. Com este propósito, consideraram-se diferentes tamanhos de amostras (n) e valores da taxa média da distribuição Poisson (θ). Concluiu-se que o teste foi mais poderoso com a redução da taxa média da Poisson. No segundo artigo, tratou-se de uma aplicação da inferência bayesiana, estimando os parâmetros da distribuição Poisson inflacionada de zeros. Com este propósito utilizou-se um conjunto de dados referentes ao número de notificações de mortes por dia, em mulheres com 80 anos de idade ou mais, durante três anos consecutivos, em Londres. Esse conjunto de dados foi utilizado para ilustrar a praticidade do método MCMC proposto e a facilidade em implementar o programa utilizando o software WinBUGS.

Palavras-chave: Modelo com distribuição ZIP. Teste escore. Erro Tipo I. Poder do teste. Estimação bayesiana.

ABSTRACT

This dissertation is organized in two articles. In the first article test score, proposed by Vandenbroek is assessed in terms of probabilities of control of type I error and power. The objective is to verify the sample size necessary to use the test Vandenbroek in discriminating between models with Poisson and ZIP. For this purpose were considered different sample sizes (n) and values of the average rate of a Poisson (θ). It was conclude that the test was more powerful with reducing the average rate of Poisson. In the second article, this was an application of Bayesian inference by estimating the parameters of Zero-Inflated Poisson distribution. For this purpose was used a set of data on the number of reported deaths per day in women 80 years of age or older, for three consecutive years in London. This data set was used to illustrate the practicality of the proposed MCMC method and ease in implementing the program using the software WinBUGS.

Keywords: ZIP model with distribution. Test score. Type I error. Power of the test. Bayesian estimation.

LISTA DE GRÁFICOS

LISTA DE TABELAS

LISTA DE SIGLAS

ZIP	Poisson inflacionada de zeros
MMV	Método da máxima verossimilhança
MCMC	Monte Carlo via cadeias de Markov
GS	Gibbs sampling
MH	Metropolis-Hastings
MNMC	Métodos não Monte Carlo
MIMC	Métodos iterativos Monte Carlo
MNIMC	Métodos não iterativos Monte Carlo

LISTA DE SÍMBOLOS

n	Tamanho amostral
n_0	Número de observações iguais a zero
N	Número de simulações Monte Carlo
θ	Taxa média da distribuição Poisson
ϕ	Parâmetro de dispersão
p	Probabilidade de uma variável aleatória assumir o valor zero (porcentagem de zeros)
S_1	Estatística do teste escore
$L(p,\theta \mid \underset{\sim}{y})$	Função de verossimilhança do modelo com distribuição ZIP
$l(p,\theta)$	Logaritmo neperiano da função de verossimilhança
$\hat{\theta}$	Estimativa do parâmetro da Poisson sob a hipótese nula
$\hat{p}_{MLE}$	Estimador do parâmetro p pelo MMV
$\hat{\theta}_{MLE}$	Estimador do parâmetro θ pelo MMV
$\pi(p,\theta \mid \underset{\sim}{y})$	Distribuição *a posteriori* conjunta de p e θ
$\pi(p)$	Distribuição *a priori* de p
$\pi(\theta)$	Distribuição *a priori* de θ
$\pi(p \mid \theta, \underset{\sim}{y})$	Distribuição condicional completa *a posteriori* de p
$\pi(\theta \mid p, \underset{\sim}{y})$	Distribuição condicional completa *a posteriori* de θ
$\hat{p}_{post.}$	Média *a posteriori* de p obtido na estimação bayesiana
$\hat{\theta}_{post.}$	Média *a posteriori* de θ obtido na estimação bayesiana

SUMÁRIO

PRIMEIRA PARTE

1 INTRODUÇÃO

Em dados de contagem é razoável supor que a distribuição Poisson possa ser ajustada. Entretanto, se considerarmos uma amostra com determinadas quantidades de zeros, o uso dessa distribuição pode ser questionada, dada a ausência do conhecimento do quanto este modelo suporta observações nulas.

Em se tratando de amostras com excesso de zeros, Nagamine, Candolo e Moura (2008) mencionam duas classificações: zeros estruturais que naturalmente ocorrem, independentemente da distribuição discreta de probabilidade e os zeros amostrais que estão relacionados com a ocorrência de zeros segundo o modelo probabilístico adotado. Para exemplificar as diferenças entre essas duas classificações, em um estudo sobre a quantificação de lesões em plantas, Ridout, Demetrio e Hinde (1998) asseguram que uma planta pode não apresentar uma lesão porque é resistente àquela doença ("zero estrutural") ou simplesmente porque o agente causador da doença não recaiu sobre aquela planta ("zero amostral").

Em particular o modelo com distribuição Poisson inflacionada de zeros (ZIP) sem covariáveis, é discutido por vários autores, como Johnson, Kotz e Kemp (1992).

Mediante o exposto, assumindo que os dados representem eventos de contagem com excesso de zeros, torna-se conveniente verificar até que ponto o modelo com distribuição Poisson clássico suporta uma determinada quantidade de zeros, sem necessariamente estimar os parâmetros da distribuição ZIP. Nesse sentido, a estruturação desta dissertação é dada no formato de dois artigos, dos quais o primeiro artigo apresenta um estudo do teste de escore proposto por Vandenbroek (1995). Tal teste permite avaliar se uma amostra com

determinadas quantidades de valores nulos pode ser ajustada por um modelo com distribuição Poisson ou ZIP.

Decorrente do fato de que o ajuste do modelo com distribuição ZIP é feito pelo método de máxima verossimilhança, como uma alternativa, no segundo artigo se propõe em fazer uma inferência bayesiana, no que tange a estimação dos parâmetros. Com este propósito, utilizou-se um conjunto de dados referentes ao número de notificações de mortes por dia, em mulheres com 80 anos de idade ou mais, observado durante três anos consecutivos, em Londres. Na contagem desses dados geralmente aparece um número relativamente grande de zeros, ou seja, a inexistência de mortes no dia. O modelo comumente utilizado na modelagem desses dados é o modelo com distribuição Poisson inflacionada de zeros (ZIP). Neste artigo utiliza-se a inferência bayesiana para se obter um resumo dos parâmetros do modelo com distribuição ZIP.

2 REFERENCIAL TEÓRICO

As estimativas dos parâmetros do modelo com distribuição Poisson para dados com excesso de zeros (ZIP) são obtidas por meio da inferência clássica, pelo método da máxima verossimilhança (MMV) e inferência bayesiana, descritos respectivamente nas seções 2.1 e 2.2.

2.1 Estimação de máxima verossimilhança dos parâmetros do modelo com distribuição Poisson inflacionada de zeros

Um fato comum para dados de contagem dar-se-á na ocorrência de um número excessivo de zeros, o que, supostamente, entende-se como uma das causas da superdispersão. A superdispersão é definida como uma perturbação na amostra em que a variabilidade amostral é superior à variabilidade esperada pelo modelo. Diante de tal situação, uma alternativa é realizar um ajuste que contemple o excesso de zeros.

Assumindo que a variável aleatória Y tem distribuição Poisson com excesso de zeros, a generalização do modelo proposto por (JOHNSON; KOTZ; KEMP, 1992) é dada por:

$$f(y;p,\theta)=\begin{cases} p+(1-p)\exp(-\theta), & se \quad y=0 \\ (1-p)\exp(-\theta)\dfrac{\theta^{y}}{y!}, & se\ y>0 \end{cases}, \qquad (1)$$

em que p é um indicador de proporção de zeros e θ a taxa média da distribuição Poisson padrão. Observe-se que, quando $p=0$, a distribuição ZIP se reduz ao modelo com distribuição Poisson.

A esperança de Y do modelo com distribuição Poisson inflacionada de zeros é dada como:

$$E(Y)=\sum_{y=0}^{\infty} yP[Y=y]=\sum_{y=1}^{\infty} yP[Y=y]=\sum_{y=1}^{\infty} y(1-p)\exp(-\theta)\frac{\theta^y}{y!}$$

$$=(1-p)\theta\exp(-\theta)\sum_{(y-1)=0}^{\infty}\frac{\theta^{(y-1)}}{(y-1)!}=(1-p)\theta.$$

Para se obter a variância de Y, é preciso calcular $E(Y^2)$:

$$E(Y^2)=\sum_{y=0}^{\infty} y^2P[Y=y]=\sum_{y=1}^{\infty} y^2P[Y=y]=\sum_{y=1}^{\infty} y^2(1-p)\exp(-\theta)\frac{\theta^y}{y!}$$

$$=(1-p)(\theta^2+\theta).$$

De tal forma que

$$Var(Y)=E(Y^2)-[E(Y)]^2=(1-p)(\theta^2+\theta)-[(1-p)\theta]^2$$

$$=(1-p)\theta+\frac{p}{1-p}[(1-p)\theta]^2.$$

Note que a variância da mistura é maior que a média da distribuição $Var(Y)=(1-p)\theta+\frac{p}{1-p}[(1-p)\theta]^2>E(Y)=(1-p)\theta$, ou seja, o modelo contempla a superdispersão gerada pelo excesso de zeros. Quanto maior a probabilidade de excesso de zeros maior a variância da variável. À medida que p se aproxima de zero a variância se aproxima de θ, ou seja, volta-se a lidar somente com uma distribuição Poisson padrão.

2.1.1 Obtenção dos estimadores de máxima verossimilhança de p e θ

A obtenção dos estimadores de p e θ, para o modelo com distribuição Poisson inflacionada de zeros, pode ser feita por meio do método da máxima verossimilhança (MMV).

2.1.2 Função de verossimilhança

Considera-se um conjunto de observações $\{y_1, y_2, ..., y_n\}$, com tamanho amostral n, sendo n_0 observações iguais a zero e n_y observações diferentes de zero para $y = 1,2,...$, tal que n_1 são as observações iguais a 1, n_2 são as observações iguais a 2 e assim por diante, de forma que $n = n_0 + \sum_{y=1}^{\infty} n_y$, tem-se a função de verossimilhança do modelo com distribuição ZIP definida por:

$$L(p,\theta \mid \underset{\sim}{y}) = \prod_{i=1}^{n} \left\{ p + (1-p)\exp(-\theta) I_{\{y_i=0\}} + (1-p)\exp(-\theta)\frac{\theta^{y_i}}{y_i!} I_{\{y_i>0\}} \right\}$$

$$\Rightarrow L(p,\theta \mid \underset{\sim}{y}) = \prod_{\substack{\{y_i=0\} \\ i=1,...,n}} [p_i + (1-p_i)\exp(-\theta_i)] \prod_{\substack{\{y_i>0\} \\ i=1,...,n}} (1-p_i)\exp(-\theta_i)\frac{\theta^{y_i}}{y_i!}$$

$$\Rightarrow L(p,\theta \mid \underset{\sim}{y}) = [p + (1-p)\exp(-\theta)]^{n_0} [(1-p)\exp(-\theta)\prod_{i=1}^{n}\frac{\theta^{y_i}}{y_i!}]^{(n-n_0)} \quad (2)$$

O logaritmo neperiano da função de verossimilhança é dado por

$$l(p,\theta) = n_0 \log_e \{p + (1-p)\exp(-\theta)\} + (n-n_0)\log_e \left\{ (1-p)\exp(-\theta)\prod_{i=1}^{n}\frac{\theta^{y_i}}{y_i!} \right\} \quad (3)$$

Derivando (3) em relação a p tem-se

$$\frac{\partial l(p,\theta)}{\partial p} = \frac{n_0(1-\exp(-\theta))}{p + (1-p)\exp(-\theta)} - \frac{(n-n_0)}{(1-p)} \quad (4)$$

Derivando (3) em relação a θ tem-se

$$\frac{\partial(p,\theta)}{\partial\theta} = -\frac{n_0(1-p)\exp(-\theta)}{p+(1-p)\exp(-\theta)} + (n-n_0)\left(\frac{y-\theta}{\theta}\right) \tag{5}$$

Igualando-se as derivadas (4) e (5) a zero obtêm-se os estimadores de máxima verossimilhança (MMV) para os parâmetros p e θ, dados, respectivamente, por (6) e (7):

$$\hat{p}_{MLE} = \frac{n_0 - n\exp(-\hat{\theta}_{MLE})}{n[1-\exp(-\hat{\theta}_{MLE})]} \tag{6}$$

$$\hat{\theta}_{MLE} = \frac{\sum_{i=1}^{n}\frac{y_i}{n}}{1-\hat{p}_{MLE}} = \frac{\overline{y}}{1-\hat{p}_{MLE}} \tag{7}$$

As equações (6) e (7) formam um sistema não linear, que não possui solução analítica. Dessa forma, soluções numéricas obtidas por processos iterativos do tipo Newton-Raphson são necessárias (XIE; HE; GOH, 2001).

2.2 Estimação pelo método bayesiano do modelo com distribuição Poisson inflacionada de zeros

Para estimar os parâmetros do modelo com distribuição ZIP pelo método bayesiano devem-se definir as distribuições *a priori* para cada um desses parâmetros. A partir dessas distribuições *a priori* e da função de verossimilhança definida em (2) define-se a distribuição *a posteriori* conjunta de p e θ.

2.2.1 Distribuição *a priori* dos parâmetros p e θ

Na estimação do parâmetro p será considerada a seguinte restrição: $0 < p < 1$ e $\theta > 0$. Com base nessa informação, as distribuições *a priori* não informativas para p e θ são dadas por (8) e (9):

$$p \sim Uniforme(a,b) \Rightarrow \pi(p) = \frac{1}{(b-a)} \tag{8}$$

$$\theta \sim Gama(\alpha,\beta) \Rightarrow \pi(\theta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}\theta^{\alpha-1}\exp(-\beta\theta) \tag{9}$$

2.2.2 Distribuição *a posteriori* conjunta dos parâmetros p e θ

A distribuição *a posteriori* conjunta será definida como

$$\pi(p,\theta \mid \underset{\sim}{y}) \propto L(p,\theta \mid \underset{\sim}{y})\pi(p)\pi(\theta)$$

$$\Rightarrow \pi(p,\theta \mid \underset{\sim}{y}) \propto [p+(1-p)\exp(-\theta)]^{n_0}[(1-p)\exp(-\theta)\prod_{i=1}^{n}\frac{\theta^{y_i}}{y_i!}]^{(n-n_0)}$$

$$\times \frac{1}{(b-a)}\frac{\beta^{\alpha}}{\Gamma(\alpha)}\theta^{\alpha-1}\exp(-\beta\theta) \tag{10}$$

2.2.3 Distribuição condicional completa *a posteriori* de p

A distribuição condicional completa *a posteriori* de p é dada por:

$$\pi(p \mid \theta, \underset{\sim}{y}) \propto L(p,\theta \mid \underset{\sim}{y})\pi(p)$$

$$\Rightarrow \pi(p \mid \theta, \underset{\sim}{y}) \propto \underbrace{[p+(1-p)\exp(-\theta)]^{n_0}[(1-p)\exp(-\theta)\prod_{i=1}^{n}\frac{\theta^{y_i}}{y_i!}]^{(n-n_0)}}_{\text{núcleo de uma } ZIP(p,\theta)} \times \underbrace{\frac{1}{(b-a)}}_{constante} \tag{11}$$

Dessa forma, $p \mid \theta, \underset{\sim}{y} \sim ZIP(p,\theta)$.

2.2.4 Distribuição condicional completa *a posteriori* de θ

A distribuição condicional completa *a posteriori* de θ é dada por

$$\pi(p \mid \theta, \underset{\sim}{y}) \propto L(p,\theta \mid \underset{\sim}{y})\pi(\theta)$$

$$\Rightarrow \pi(\theta \mid p, \underset{\sim}{y}) \propto \underbrace{\theta^{(\alpha^*-1)}\exp[(-\theta)^{\beta^*}]}_{\text{núcleo de uma } Gama(\alpha^*,\beta^*)}. \tag{12}$$

Dessa forma,

$$\theta \mid p, \underset{\sim}{y} \sim Gama(\alpha^*, \beta^*); \quad \begin{cases} \alpha^* = \sum_{i=1}^{n} y_i(n-n_0)+\alpha; \\ \beta^* = \beta+(n-n_0)-1 \end{cases}$$

3 CONSIDERAÇÕES FINAIS

Em síntese, os resultados obtidos neste trabalho, dividido no formato de artigo, são de suma importância no sentido de que, em muitas situações experimentais, é comum a obtenção de dados de contagem com valores nulos, dos quais inúmeras interpretações podem ser dadas. Como exemplos podem-se citar número de sementes que não germinaram, amostras não contaminadas por determinados fungos, etc. Assim sendo, tendo o conhecimento de que observações nulas podem afetar substancialmente as estimativas do modelo Poisson, sugere-se o uso ou não do modelo ZIP, por meio do teste Vandenbroek, o que certamente agregará informações mais precisas nas estimativas das probabilidades. Neste aspecto, convém propor a abordagem da inferência bayesiana na estimação dos parâmetros do modelo, uma vez que a informação a priori do pesquisador possivelmente pode ser incorporada no processo de estimação.

REFERÊNCIAS

JOHNSON, N.; KOTZ, S.; KEMP, A.W. **Univariate discrete distributions**. 2nd ed. New York: J. Wiley, 1992. 565 p.

LAMBERT, D. Zero - inflated Poisson regression with application to defects in manufacturing. **Technometrics,** Alexandria, v. 34, p. 1-14, 1992.

MULLAHAY, J. Heterogeneity, excess zeros, and the structure of count data models. **Journal of Applied Econometrics**, Kingston, v. 12, p. 337-350, 1997.

NAGAMINE, C. M. L.; CANDOLO, C.; MOURA, M. S. A. An application of models for count data zero-inflated for modeling the number of eggs for the mosquito Aedes aegypti. **Revista de Matemática e Estatística**, Sao Paulo, v. 26, n. 10, p. 97-114, 2008.

RIDOUT, M.; DEMETRIO, C. G. B.; HINDE, J. Models for count data with many zeros. In: INTERNATIONAL BIOMETRIC CONFERENCE, 19., 1998, Cape Town. **Anais...** Cape Town: [s. n.], 1998. p. 179-190.

VANDENBROEK, J. A score test for zero inflation in a Poisson-distribution. **Biometrics,** Washington, v. 51, n. 2, p. 738-743, 1995.

XIE, M.; HE, B.; GOH, T.N. Zero-inflated Poisson model in statistical process control. **Computational Statistics & Data Analysis,** Amsterdam, v. 38, p. 191-201, 2001.

SEGUNDA PARTE – ARTIGOS

ARTIGO 1 Avaliação do poder e taxas de erro tipo I do teste escore de Vandenbroek por meio do método de Monte Carlo

RESUMO

Neste artigo estuda-se o tamanho de amostra necessário para utilizar o teste escore de Vandenbroek para verificar se o modelo com distribuição Poisson se ajusta adequadamente em uma amostra com excesso de zeros, no intuito de verificar se a rejeição de H_0 propicia a aplicação da distribuição Poisson inflacionada de zeros. Este trabalho foi realizado com o objetivo de verificar a adequabilidade do modelo com distribuição Poisson inflacionada de zeros, por meio do teste Vandenbroek. Para isso utilizou-se um estudo de simulação com as seguintes configurações: tamanhos de amostras (n) e valores da taxa média da distribuição Poisson (θ), computando-se as taxas empíricas do controle do erro tipo I e poder do teste. Conclui-se que, reduzindo a proporção de valores nulos e aumentando o tamanho amostral, os valores de poder têm diferenças muito pequenas. O teste de Vandenbroek tende a ser conservativo à medida que a taxa média da Poisson (θ) é aumentada.

Palavras-chave: Modelo com distribuição ZIP. Teste escore. Erro Tipo I. Poder do teste. Simulação Monte Carlo.

ARTICLE 1 Evaluation of power and type I error rates of test score Vandenbroek through the Monte Carlo method

ABSTRACT

This paper studies the sample size necessary to use the test score Vandenbroek to verify the model with Poisson fit neatly in a sample with excess zeros, in order to check whether the rejection of H_0 provides the application of the Zero-Inflated Poisson distribution. This work was carried out with the objective to verify the adequacy of the model with Zero-Inflated Poisson through the test Vandenbroek. For this was used a simulation study with the following settings: sample sizes (n) and values of the average rate of Poisson (θ), computing rates of empirical control of type I error and power of the test. It follows that by reducing the proportion of null values and increasing the sample size, the values of power has very small differences. The test of Vandenbroek tends to be conservative measure that the average of the Poisson (θ) is increased.

Keywords: ZIP model with distribution. Test score. Type I error. Power of the test. Monte Carlo simulation.

1 INTRODUÇÃO

É comum observar, na análise de dados de contagem, uma quantidade de zeros superior àquela esperada sob uma determinada distribuição. Em muitas situações, a utilização de um simples modelo com distribuição Poisson pode ser viável, desde que se tenha o conhecimento do quanto este modelo suporta observações nulas, no sentido de ter uma variação expressiva em relação à variação esperada por esse modelo. Em outras palavras, deseja-se verificar se a distribuição Poisson pode ser utilizada para representar esses dados. Com esse propósito, Vandenbroek deduziu um teste escore para fazer essa verificação, considerando como hipótese nula a distribuição Poisson e como hipótese alternativa a distribuição Poisson Inflacionada de Zeros (ZIP).

Com a utilização deste teste verifica-se a existência de alguma evidência para apoiar a suposição de que a distribuição Poisson inflacionada de zeros é realmente mais apropriada do que a distribuição Poisson para representar um conjunto de dados em que a porcentagem de observações nulas é conhecida. Para essa questão, a resposta é dada pela aplicação desse teste construído de forma a confrontar a amostra gerada por meio de uma variável aleatória com distribuição ZIP, considerando que a hipótese nula (p=0) represente a distribuição Poisson. Entretanto, por ser um teste assintótico, utiliza-se como premissa que a estatística possui, assintoticamente, distribuição normal. Contudo, dadas amostras pequenas, é sabido que as propriedades do controle do erro tipo I e poder são seriamente afetadas.

Mediante o exposto, a motivação deste trabalho foi justamente verificar o tamanho amostral desejável de forma que o controle do erro tipo I e, consequentemente, o poder possam apresentar valores condizentes com o nível nominal especificado. Neste sentido, outras configurações paramétricas são dignas de serem avaliadas.

2 REFERENCIAL TEÓRICO

2.1 Modelo com distribuição ZIP

Em geral, quando há muitos zeros na amostra, a distribuição de Poisson inflacionada de zeros pode ser utilizada para representar esses dados.

A variável resposta Y tem distribuição Poisson com excesso de zeros, assume o valor zero com probabilidade p e com probabilidade $(1-p)$ assume o valor de uma variável aleatória com distribuição Poisson de média θ.

Johnson, Kotz e Kemp (1992) discutiram uma maneira simples de modificar uma distribuição discreta para acomodar excesso de zeros. Uma parte com excesso de zeros, p, é adicionada à proporção de zeros a partir da distribuição original discreta, $f(0)$, enquanto a proporção restante diminui de forma adequada:

$$\begin{cases} P(Y_i = 0) = p + (1-p)f(0), \\ P(Y_i = y_i) = (1-p)f(y_i) \quad , \quad (y_i > 0) \end{cases} \tag{1}$$

em que $i = 1,...,n$; p é um indicador de proporção de zeros e θ a taxa média da distribuição Poisson padrão. Observe-se que quando $p = 0$, a distribuição ZIP se reduz ao modelo com distribuição Poisson.

Johnson, Kotz e Kemp (1992) afirmam que é possível considerar p inferior a zero, respeitando-se a condição: $p \geq -\dfrac{f(0)}{[1-f(0)]}$.

Uma aplicação desta distribuição na prática foi feita por Lambert (1992). Em um processo de fabricação, o número de defeitos dos produtos possui, frequentemente, distribuição Poisson. Assim, em uma amostra de n peças espera-se encontrar $n\exp(-\theta)$ peças sem defeito. Porém, pode-se encontrar um número ainda maior de peças não defeituosas, pois se espera, diante de um

"estado perfeito", que existam itens extremamente resistentes. Dessa forma, os dados podem possuir distribuição ZIP. Porém, se o processo de fabricação é bem estruturado, muitos produtos que não têm defeitos causam um grande número de zeros nos dados. Mullahay (1997) demonstrou que as observações não homogêneas são comumente assumidas como superdispersão em modelos de dados de contagem e têm implicações tais como a superdispersão para a estrutura de probabilidade de modelos de misturas.

Este exemplo assume que a distribuição Poisson inflacionada de zeros é adequada ou, em outras palavras, que a população considerada é constituída por duas subpopulações: a proporção p, constituída de peças que não têm defeitos e outra constituída por peças com defeitos. Isso nem sempre é óbvio. Deseja-se verificar se há alguma evidência, a partir dos dados observados, para apoiar essa hipótese. Para alcançar este objetivo, para a distribuição Poisson inflacionada de zeros um teste escore de Vandenbroek é proposto na próxima seção.

2.2 Teste Escore Vandenbroek (1995)

Usando (1), a função de probabilidade da ZIP é definida como:

$$P(Y_i = 0) = p + (1-p)\exp(-\theta)$$

$$P(Y_i = y_i) = (1-p)\frac{\exp(-\theta)\theta^{y_i}}{y_i!} \qquad (y_i > 0), \quad i = 1,...,n$$

em que n representa o tamanho amostral; p a porcentagem de zeros e θ a taxa média da distribuição Poisson.

Considere o caso em que há n observações, entre elas n_0 zeros e não existem covariáveis, $\bar{y}$ é a média das observações e $p_0 = \exp(-\hat{\theta})$, em que $\hat{\theta}$ é a estimativa do parâmetro da Poisson sob a hipótese nula. A estatística do teste

escore de Vandenbroek, considerando a hipótese nula $(H_0 : Y \sim Poisson(\theta))$, é dada por

$$S_1 = \frac{(n_0 - np_0)^2}{np_0(1-p_0) - n\bar{y}p_0^2} \quad (2)$$

Vandenbroek (1995) realizou um estudo de simulação para verificar se a aproximação qui-quadrado é conveniente. A partir de uma distribuição de Poisson com média θ=0,5 foram realizadas 5.000 simulações com amostras de tamanho 100 e 200. O mesmo foi feito para média θ =1. Para cada amostra, a estatística de escore foi calculada e, posteriormente, percentis foram obtidos, os quais foram comparados com os percentis de uma distribuição qui-quadrado com um grau de liberdade (Tabela 1).

Tabela 1 Percentis da estatística S_1, com base em 5.000 amostras de tamanho n de uma distribuição de Poisson com média θ e os mesmos pontos de distribuição $\chi^2(1)$.

	Percentis de uma $\chi^2(1)$				
	$P_{0,7} = 1,07$	$P_{0,8} = 1,64$	$P_{0,9} = 2,71$	$P_{0,95} = 3,84$	$P_{0,99} = 6,63$
$n = 100$ e $\theta = 0,5$	1,13	1,66	2,73	3,86	6,77
$n = 100$ e $\theta = 1$	1,11	1,68	2,75	3,97	6,87
$n = 200$ e $\theta = 0,5$	1,13	1,37	2,67	3,73	6,52
$n = 200$ e $\theta = 1$	1,02	1,58	2,56	3,68	6,61

Por meio dos resultados descritos na Tabela 1 nota-se que de fato existe uma concordância em considerar a distribuição $\chi^2(1)$ como uma regra de decisão do teste escore proposto por Vandenbroek, associado ao nível de significância (α), tomando em particular o valor de $P_{0,95}$ = 3,84, sendo este obtido da distribuição de S_1 sob hipótese nula e interpretado como o valor que proporciona

95% das observações abaixo de 3,84. Além do mais, pode-se perceber que em percentil é aproximado pela média de uma distribuição $\chi^2(1)$ acumulada. Assim sendo, sugere-se, empiricamente, assumir que a rejeição de H_0 deverá ocorrer quando $S_1 > \chi^2(1)$.

Assim, a conclusão do teste é: se a hipótese nula é rejeitada, então o modelo com distribuição Poisson não é viável de ser aplicado, no sentido de que o modelo com distribuição ZIP pode ser utilizado no ajuste dos dados.

2.2.1 Um exemplo proposto por Vandenbroek (1995)

Um estudo foi realizado no Departamento de Medicina Interna do Hospital da Universidade de Utrecht, com 98 homens infectados pelo vírus da imunodeficiência humana (HIV). Nesse estudo foi registrado o número de vezes que os pacientes tiveram uma infecção do trato urinário (número de episódios) (HOEPELMAN et al., 1992). Além disso, o estado imunológico de cada paciente foi determinado medindo-se a contagem de células CD4 +. Pelos dados da Tabela 2 observa-se que muitos pacientes não tiveram uma infecção urinária.

Tabela 2 Frequências do número de episódios.

Número de episódios	0	1	2	3
Frequências	81	9	7	1

O teste escore é dado por $S_1 = \frac{(n_0 - np_0)^2}{np_0(1-p_0) - n\bar{y}p_0^2} = 15{,}35$, em que $n_0 = 81$ é o número de pacientes que não tiveram uma infecção urinária; $n = 98$ é o número de pacientes infectados pelo HIV; $\bar{y} = 0{,}27$ é a média do número de vezes que os pacientes tiveram uma infecção do trato urinário e $p_0 = \exp(-\bar{y}) = \exp(-0{,}27) = 0{,}77$ é a estimativa da probabilidade de os pacientes não terem uma infecção urinária. Comparando-se esse resultado com $\chi^2_{(1)}$, que foi 3,84, considerando nível de significância de 5%, rejeita-se a hipótese nula, concluindo-se que o modelo Poisson para esta quantidade de zeros não é plausível de ser utilizado, sugerindo o modelo ZIP. Entretanto, convém salientar que a população pode ser constituída por duas partes: a proporção p constituída de pacientes que não estão em risco de desenvolver uma infecção do trato urinário e outra constituída por pacientes que correm o risco de desenvolver uma infecção urinária.

Segundo El-Shaarawi (1985), a rejeição da hipótese nula, de que a distribuição Poisson representa bem os seus dados, não implica que a distribuição Poisson inflacionada é a distribuição mais adequada, pois pode haver outras distribuições a serem utilizadas para representar os dados. A distribuição binomial negativa pode ser uma boa candidata.

3 METODOLOGIA

Em consonância com o objetivo proposto na introdução, a metodologia utilizada foi realizada conforme as seções seguintes.

3.1 Simulação das amostras dos modelos com distribuição Poisson e ZIP

Na condução deste trabalho, as amostras inflacionadas de zeros foram simuladas a partir de uma variável aleatória com distribuição ZIP.

Conforme a função de probabilidade definida na seção 2.1, especificamente, ao modelo Poisson as amostras obtidas foram dadas fixando-se $p = 0$. Considerando-se ambos os modelos, os valores paramétricos assumidos foram especificados nas seguintes situações:

a) parâmetro de dispersão (ϕ), fixado em 1;

b) Taxa média da distribuição Poisson (θ), fixada em 5 e 10;

c) diferentes tamanhos amostrais, determinados pelos valores: 25, 30, 40, 50, 100, 150, 200, 500, 1.000, 5.000 e 10.000.

3.2 Taxas empíricas da ocorrência do erro tipo I e poder do teste escore de Vandenbroek

Mantendo-se as especificações paramétricas definidas em 3.1, utilizando o método de Monte Carlo, as taxas empíricas da ocorrência do erro tipo I foram determinadas assumindo N=100.000 simulações, de tal forma que, considerando a hipótese $H_0 : p = 0$ (modelo Poisson), a proporção Z/N, em que Z corresponde à contagem do número de vezes que a estatística do teste S_1 (seção 2.2) foi significativo, fixado o nível de significância em 5%. Nesta situação, Z/N foi interpretada como a taxa empírica da ocorrência do erro tipo I.

Seguindo o mesmo procedimento, porém, realizando as simulações sob $H_1 : p \neq 0$ (modelo ZIP), o valor obtido Z/N referiu-se como uma aproximação do poder, proporcionado pelo teste escore de Vandenbroek.

Por fim, para a obtenção dos resultados via método Monte Carlo, implementou-se um programa no software R Development Core Team (2010), utilizando o pacote ZIGP (APÊNDICE A).

4 RESULTADOS E DISCUSSÃO

4.1 Taxa de erro tipo I

A discussão dos resultados referentes ao controle do erro tipo I foi feita de forma comparativa, em relação aos diferentes valores da taxa média (θ), conforme especificado na metodologia. Neste contexto, observou-se que, mantendo-se o valor do parâmetro de dispersão fixado em 1 e aumentando-se a taxa média da Poisson de θ=5 para θ=10, conforme descrito na Tabela 3, notou-se uma sensibilidade no controle do erro tipo I, no que tange à especificação da média da Poisson. Para uma melhor compreensão na ocorrência deste resultado, pode-se observar que, na situação inicial (θ=5), para $n \geq 100$ (Tabela 3), o teste Vandenbroek apresentou taxas empíricas da ocorrência do erro tipo I aproximadas ao nível nominal fixado. Contudo, comparando-se com os resultados obtidos para a especificação (θ=10), observou-se que, em todas as situações amostrais, o teste mostrou-se conservativo, com exceção para grandes amostras consideradas em n = 10.000.

A fim de confirmar esse efeito pronunciado na taxa média θ, em uma situação mais extrema (considerando θ=30 e θ=50)[1], as taxas da ocorrência do erro tipo I foram nulas, sugerindo fazer a afirmação de que o teste é conservativo, porém, deve-se considerar que elevados valores da taxa média, bem como elevados valores do parâmetro de dispersão, conduzem a uma aproximação da distribuição normal. Assim sendo, para essas configurações, o teste escore de Vandenbroek não é recomendável para discriminar entre o modelo com distribuição Poisson e ZIP quanto ao seu uso, perante uma amostra com quantidade de zeros expressiva.

---

[1] Resultados não apresentados na discussão. Para confirmação, o pesquisador poderá utilizar os programas em anexo.

Tabela 3 Taxas de erro tipo I, para o teste escore, em função do tamanho amostral (n), taxa média da Poisson (θ), parâmetro de dispersão (ϕ) fixado em 1, nível nominal de significância α=5% e N=100.000 simulações.

n	$\theta = 5$	$\theta = 10$
25	0,0890	0,0011
30	0,0750	0,0015
40	0,0480	0,0016
50	0,0470	0,0023
100	0,0508	0,0044
150	0,0460	0,0068
200	0,0475	0,0094
500	0,0499	0,0280
1.000	0,0553	0,0250
5.000	0,0578	0,0215
10.000	0,0572	0,0570

4.2 Poder do teste escore

Mantendo-se as mesmas especificações paramétricas referentes à taxa média da Poisson (θ=5 e 10) e diferentes tamanhos amostrais, o poder do teste Vandenbroek foi avaliado sob a hipótese alternativa, considerando as seguintes proporções de valores zeros hipotetizadas por H_1 em 1%, 3%, 5%, 10% e 30%. Nestas situações, os valores obtidos para o estudo do poder do teste estão descritos na Tabela 4. Entretanto, convém ressaltar que os resultados referentes ao controle do erro tipo I foram condizentes com o nível nominal de significância especificado em 5%, em ambas as situações de θ=5 e θ=10 (Tabela 3). Portanto, a discussão dos resultados se deteve para amostras superiores a n=100. Neste contexto, pode-se observar que, de modo geral, para θ=10, os

valores de poder foram inferiores à situação em que se considerou θ=5. Tal fato é corroborado justamente pelas baixas taxas empíricas observadas em θ=10 (Tabela 3). Porém, notoriamente, foi possível observar uma relação entre o tamanho amostral com a proporção de valores nulos (p), de tal forma que os valores de poder foram bem aproximados para θ=5 e θ=10. Tal relação é verificada na medida em que ocorre o aumento do tamanho amostral e, simultaneamente à redução do valor de p, os valores de poder tendem a igualar. Tendo por base esses resultados, nas situações avaliadas, principalmente devido ao fato de que todo o estudo realizado neste trabalho se deu mantendo o parâmetro de dispersão fixado em 1.

O teste não controlou o erro tipo I (quando fixado a taxa média da Poisson em θ=5 e 10) para os tamanhos amostrais fixados em n=25, 30. Por esse motivo, optou-se por não avaliar o poder para esses tamanhos amostrais, uma vez que o uso do teste escore de Vandenbroek é recomendável mediante a coerência entre o controle da ocorrência do erro tipo I e poder.

Em função dos resultados referentes ao erro tipo I e poder, descritos, respectivamente, nas Tabelas 3 e 4, tornou-se possível verificar a relação entre os valores de p e n, resumida na Tabela 5.

Tabela 4 Poder do teste escore S_1 (%) para diversos valores de θ e nível nominal de significância α=5%, em função da porcentagem de zeros (p) e do tamanho amostral (n).

	$\theta = 5$	$\theta = 10$
n=40 e p=1%	33,8	18,4
n=40 e p=3%	69,8	47,6
n=40 e p=5%	87,1	69,9
n=40 e p=10%	98,7	94,1
n=40 e p=30%	100,0	100,0
n=50 e p=1%	40,2	20,7
n=50 e p=3%	78,4	53,7
n=50 e p=5%	92,6	76,8
n=50 e p=10%	99,4	97,3
n=50 e p=30%	100,0	100,0
n=100 e p=1%	63,2	27,9
n=100 e p=3%	95,2	73,9
n=100 e p=5%	99,5	93,3
n=100 e p=10%	100,0	99,9
n=100 e p=30%	100,0	100,0
n=150 e p=1%	77,5	37,1
n=150 e p=3%	98,9	86,2
n=150 e p=5%	99,9	98,1
n=150 e p=10%	100,0	99,9
n=150 e p=30%	100,0	100,0
n=200 e p=1%	87,1	40,9
n=200 e p=3%	99,7	92,5
n=200 e p=5%	100,0	99,5
n=200 e p=10%	100,0	100,0
n=200 e p=30%	100,0	100,0
n=500 e p=1%	99,5	67,8
n=500 e p=3%	100,0	99,5
n=500 e p=5%	100,0	100,0
n=500 e p=10%	100,0	100,0
n=500 e p=30%	100,0	100,0
n=1.000 e p=1%	100,0	89,2
n=1.000 e p=3%	100,0	100,0

Tabela 4, continuação

	$\theta = 5$	$\theta = 10$
n=1.000 e p=5%	100,0	100,0
n=1.000 e p=10%	100,0	100,0
n=1.000 e p=30%	100,0	100,0
n=5.000 e p=1%	100,0	100,0
n=5.000 e p=3%	100,0	100,0
n=5.000 e p=5%	100,0	100,0
n=5.000 e p=10%	100,0	100,0
n=5.000 e p=30%	100,0	100,0
n=10.000 e p=1%	100,0	100,0
n=10.000 e p=3%	100,0	100,0
n=10.000 e p=5%	100,0	100,0
n=10.000 e p=10%	100,0	100,0
n=10.000 e p=30%	100,0	100,0

Tabela 5 Menor tamanho amostral em função da proporção de zeros que forneça valores de poder similares para a taxa média θ=5 e θ=10, para o teste de Vandenbroek utilizado para verificar a adequabilidade do modelo Poisson, ou ZIP, em uma amostra com excesso de zeros, considerando diferentes valores de p.

p	Tamanho amostral (n)
30%	$n \geq 40$
10%	$n \geq 100$
5%	$n \geq 150$
3%	$n \geq 200$
1%	$n \geq 500$

O tamanho da amostra adequado para verificar a adequabilidade do modelo Poisson ou ZIP numa amostra com 30% de valores nulos deve ser de, no mínimo, $n \geq 40$. Se essa proporção for de 10%, o tamanho da amostra aumenta para $n \geq 100$. Para uma proporção de 5%, o tamanho da amostra adequado deve ser de, no mínimo, $n \geq 150$. Se essa proporção diminuir para 3%, novamente tem-

se um aumento no tamanho amostral, que passa a ser $n \geq 200$. Finalmente, se houver uma diminuição na proporção de zeros para 1%, o tamanho da amostra aumenta para $n \geq 500$.

5 CONCLUSÃO

Reduzindo a proporção de valores nulos e aumentando o tamanho amostral, de forma simultânea, os valores de poder têm diferenças muito pequenas. O Teste de Vandenbroek tende a ser conservativo à medida que a taxa média (θ) é aumentada, dadas as configurações estudadas neste trabalho.

Considerando os valores de θ iguais a 5 e 10, o tamanho amostral desejável, de forma que o controle do erro tipo I e, consequentemente, o poder possam apresentar valores condizentes com o nível nominal especificado, é definido de acordo com a proporção de zeros observada na amostra. À medida que a proporção de zeros diminui, o tamanho amostral aumenta.

6 ESTUDOS FUTUROS

Da mesma forma como foi feito um estudo do teste escore Vandenbroek fixando o erro tipo I em α=0,05, pode-se estudar o erro tipo II (β) fixando um determinado valor para o poder, por exemplo, ($1 - \beta = 0,95$).

Deseja-se fazer um estudo comparativo do teste escore de Vandenbroek com outras distribuições discretas e inflacionadas de zeros.

REFERÊNCIAS

EL-SHAARAWI, A. H. Some goodness-of-fit methods for the Poisson plus added zeros distribution. **Applied and Environmental Microbiology,** Washington, v. 49, p. 1304-1306, 1985.

HOEPELMAN, A. I. M. et al. Bacteriuria in men infected with HIV-1 is related to their immune status (CD4+ cell count). **AIDS**, London, v. 6, p. 179-184, 1992.

JOHNSON, N.; KOTZ, S.; KEMP, A. W. **Univariate discrete distributions**, 2nd ed. New York: J. Wiley, 1992. p. 312-318.

LAMBERT, D. Zero - inflated Poisson regression with application to defects in manufacturing. **Technometrics,** Alexandria, v. 34, p. 1-14, 1992.

MULLAHAY, J. Heterogeneity, excess zeros, and the structure of count data models. **Journal of Applied Econometrics**, Kingston, v. 12, p. 337-350, 1997.

R DEVELOPMENT CORE TEAM. **R**: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2010. Disponível em: <http://www.R-project.org/>. Acesso em: 22 jan. 2010.

VANDENBROEK, J. A score test for zero inflation in a Poisson-distribution. **Biometrics,** Washington, v. 51, n. 2, p. 738-743, 1995.

APÊNDICE

APÊNDICE A Programa construído no software R utilizado no método Monte Carlo para a obtenção das taxas empíricas da ocorrência do erro tipo I e poder do teste escore de Vandenbroek

```
library(ZIGP)  ## pacote requerido ##

## Valores paramétricos de simulação ##

n=25       ### Tamanho da amostra ###
theta=8      #### Média da Poisson ###
thetaini=7.8  ####  Chute inicial do Newton Raphson ###
phi=1       ### Parâmetro de dispersão ###
p=0  # Porcentagem de zeros # H0: p=0 (ERRO TIPO I) MODELO POISSON #
#####  Definindo p≠ 0  ##### H1: p≠0 (PODER) MODELO ZIP #

### Variáveis auxiliares ###

nsim=100000        # Total sim.: Número de amostras de tamanho n geradas #
tolerancia=0.1  # Critério de tolerância do newton Raphson #
totint=1000      # Número total de iterações do newton Raphson #
n0=0             # Número de observações iguais a zero #
contaprop=0

##### Newton Raphson ####

nr=function(n0,thetaini,amos,n,tolerancia,totint)
{
```

```
        it=0
        status="T"
while (status=='T')
{
        it=it+1
        pest=(n0-n*exp(-thetaini))/(n*(1-exp(-thetaini)))
        thetaf=(sum(amos)/n)/(1-pest)
        crit=abs(thetaini-thetaf)
        if (crit > tolerancia || it < totint)  thetaini=thetaf
        if (crit < tolerancia || it > totint)
{
        estheta=thetaini
        estp=pest
        status="F"
        rstatus=status
}
}
        return (list(rpest=estp,rtheta=estheta,rcrit=crit,sit=it))
}
for (i in 1:nsim)
{
        n0=0
        amoszip=rzigp(n,theta,phi,p)

### Computar n0 ###

        for (j in 1:length(amoszip))
        if (amoszip[j]==0) n0=n0+1
```

```
        ### Newton Raphson: Estimativa de p e theta ###
        chama_nr=nr(n0,thetaini,amoszip,n,tolerancia,totint)
        while (chama_nr$rpest>1.0) nr(n0,thetaini,amoszip,n,tolerancia,totint)

#### Estatística do teste ###

        yb=sum(amoszip)/n
        p0=exp(-chama_nr$rtheta)
        s1=((n0-n*p0)^2)/((n*p0*(1-p0))-(n*yb*(p0^2)))
        qt=qchisq(c(0.05), df=1, lower.tail=FALSE)
        if (s1>qt) contaprop=contaprop+1
}
        prob=contaprop/nsim
        prob  # Proporção Z/N #
```

ARTIGO 2 Estimação dos parâmetros do modelo com distribuição ZIP via inferência bayesiana

RESUMO

Quando se trabalha com dados de contagem, não raras vezes pode-se encontrar a presença de excesso de zeros. Em geral, essas situações apresentam uma variabilidade maior do que a esperada pelo modelo com distribuição Poisson, conhecido como superdispersão. Neste trabalho utiliza-se o modelo com distribuição inflacionada de zeros para acomodar tal situação. A título de ilustração da aplicação da inferência bayesiana na estimação dos parâmetros do modelo ZIP, utilizaram-se dados referentes ao número de notificações de mortes por dia, em mulheres com 80 anos de idade ou mais, observadas durantes três anos consecutivos, em Londres. Na contagem desses eventos, geralmente aparece um número relativamente grande de zeros, ou seja, a inexistência de mortes de mulheres, por vários dias dos anos analisados. A abordagem utilizada no trabalho é a inferência bayesiana. O conjunto de dados apresentado acima é usado para ilustrar a praticidade do método proposto, facilmente implementado, utilizando-se o software WinBUGS. Dessa forma, o presente trabalho, em síntese, foi realizado com o objetivo de estimar os parâmetros de interesse do modelo com distribuição Poisson inflacionada de zeros, utilizando a inferência bayesiana, com o auxílio do programa computacional WinBUGS.

Palavras-chave: Inferência Bayesiana. MCMC. Gibbs Sampling. Modelo ZIP. Teste escore.

ARTICLE 2 Estimation of model parameters with ZIP distribution via Bayesian inference

ABSTRACT

When working with count data, sometimes not rare, you can find the presence of excess zeros. In general, these situations present a greater variability than expected by the model with Poisson distribution, known as overdispersion. In this paper was used the distribution model with Zero-Inflated to accommodate such a situation. For illustrate the application of Bayesian methods to estimate the parameters of the ZIP models, were used data on the number of reported deaths per day in women 80 years of age or older, observed during three consecutive years in London. In counting of these events usually appears a relatively large number of zeros, i.e., the absence of deaths in women for several days of the years analyzed. The approach used in this work is the Bayesian inference. The data set presented above is used to illustrate the practicality of the proposed method, easily implemented, using WinBUGS software. Thus, the present work in summary, was realized with the objective to estimate the parameters of interest to the model with distribution Zero-Inflated Poisson, using Bayesian inference, with the aid of the computer program WinBUGS.

Keywords: Bayesian Inference. MCMC. Gibbs Sampling. ZIP models. Test score.

1 INTRODUÇÃO

Dados de contagens que contêm muitos zeros são comuns em diversas áreas de estudos. Por exemplo, Comulada et al. (2007) utilizaram modelos com excesso de zeros para detectar a redução do uso abusivo de drogas por jovens portadores do vírus HIV e realçaram a importância desta técnica na análise de dados de estudos de comportamento. Karazsia e Van Dulmen (2008) apresentaram uma análise de dados longitudinais de contagens de atendimentos a crianças vítimas de lesões, em que muitos zeros estão presentes. Os autores enfatizaram neste estudo a utilidade de um modelo baseado na distribuição de Poisson, que considera o excesso de zeros na análise de seus dados. Ramis-Prieto et al. (2007) utilizaram um modelo bayesiano com excesso de zeros para descrever o padrão geográfico da mortalidade por tumores hematológicos na Espanha (Mazin et al., 2008). Outros exemplos da distribuição Poisson com zeros inflacionados (distribuição ZIP - *zero inflated Poisson*) podem ser encontrados em Cohen (1960), Goraski (1977), Kemp (1986) e Martin e Katti (1965)

No presente trabalho, foi utilizado um conjunto de dados do jornal britânico "London Times", retirado do artigo Schilling (1947). Esses dados representam o número de notificações de mortes por dia, de mulheres com 80 anos de idade ou mais, por três anos consecutivos, em Londres. Uma descrição detalhada dos dados é apresentada na Tabela 6. Na contagem desses dados, geralmente aparece um número relativamente grande de zeros, ou seja, a inexistência de mortes de mulheres por vários dias dos anos analisados. O modelo comumente utilizado na modelagem desses dados é o modelo com distribuição ZIP. Todavia, torna-se conveniente verificar se esse conjunto de dados pode ser bem representado pelo modelo com distribuição Poisson.

Essa questão pode ser respondida por um teste assintótico, chamado teste escore (S_1), proposto por Vandenbroek (1995), estudado no artigo 1 desta dissertação. Este teste é utilizado para verificar se situações em que o número de zeros é muito grande podem ser representadas por uma distribuição Poisson padrão. O objetivo do teste escore ZIP é verificar se existe alguma evidência para apoiar a suposição de que a distribuição Poisson inflacionada de zeros é realmente mais apropriada do que a distribuição Poisson.

A estatística de contagem do teste escore tem distribuição assintótica qui-quadrado χ^2_{α} com 1 grau de liberdade sob a hipótese nula e pode ser escrita como $S_1 = \frac{(n_0 - np_0)^2}{np_0(1-p_0) - n\bar{y}p_0^2}$, em que $p_0 = \exp(-\hat{\theta})$; n é o número total de observações; n_0 é o número de zeros observados; $\bar{y}$ é a média das observações e $\hat{\theta}$ é a estimativa do parâmetro da Poisson sob a hipótese nula H_0: $p = 0$ (Modelo Poisson).

Neste artigo, apresenta-se a estimação bayesiana dos parâmetros do modelo com distribuição ZIP. O conjunto de dados, apresentado anteriormente, é utilizado para ilustrar a praticidade do método proposto, facilmente implementado, usando o software WinBUGS.

Dessa forma, o presente trabalho, em síntese, foi realizado com o objetivo de estimar, utilizando inferência bayesiana, os parâmetros do modelo com distribuição ZIP.

2 REFERENCIAL TEÓRICO

2.1 Inferência bayesiana

Segundo Gelman et al. (2003), a inferência bayesiana é o processo de encontrar um modelo de probabilidade para um conjunto de dados e resumir o resultado por uma distribuição de probabilidade sobre os parâmetros do modelo e sobre quantidades não observadas, tais como predição para novas observações. A inferência bayesiana está fundamentada no Teorema de Bayes que, para muitos, é um dos poucos resultados que se propõem a caracterizar a aprendizagem com a experiência, isto é, modificar a atitude inicial em relação aos antecedentes, causas, hipóteses ou estados depois de ter a informação adicional de que certo acontecimento se realizou (PAULINO; TURKMAN; MURTEIRA, 2003).

2.1.1 Teorema de Bayes

Suponha que $\underset{\sim}{y}' = (y_1, y_2, \ldots, y_n)$ seja um vetor de n observações cuja distribuição de probabilidade $L(\theta, p \mid \underset{\sim}{y})$ depende dos valores de p e θ. Para fazer a inferência sobre p e θ, dado $\underset{\sim}{y}$, é necessário um modelo vindo de uma distribuição de probabilidade conjunta para p, θ e $\underset{\sim}{y}$ (GELMAN et al., 2003).

A condição sobre os valores conhecidos dos dados $\underset{\sim}{y}$, usando a propriedade condicional conhecida como Teorema de Bayes, produz a densidade *a posteriori*:

$$\pi(p, \theta \mid \underset{\sim}{y}) = \frac{\pi(p, \theta, \underset{\sim}{y})}{f(\underset{\sim}{y})} = \frac{L(p, \theta \mid \underset{\sim}{y})\pi(p)\pi(\theta)}{f(\underset{\sim}{y})} \qquad (1)$$

Uma forma equivalente de (1) omite o valor de $f(\underset{\sim}{y})$ que não depende de p e θ, com $\underset{\sim}{y}$ fixo, então, pode ser considerado constante, produzindo, assim, uma densidade *a posteriori* não normalizada:

$$\pi(p,\theta \mid \underset{\sim}{y}) \propto L(p,\theta \mid \underset{\sim}{y})\pi(p)\pi(\theta) \qquad (2)$$

Os métodos bayesianos passam, em certo sentido, por uma extensão do modelo clássico. No modelo clássico, os parâmetros p e θ são escalares ou vetores desconhecidos, porém, fixos, no modelo bayesiano p e θ são escalares ou vetores aleatórios, não observáveis. A filosofia bayesiana assume, então, que o que é desconhecido (parâmetros p e θ) é incerto e toda a incerteza deve ser quantificada em termos de probabilidade (PAULINO; TURKMAN; MURTEIRA, 2003).

A informação inicial ou *a priori* do que é incerto pode traduzir-se formalmente por uma distribuição de probabilidade chamada de distribuição *a priori.*

2.1.2 Distribuição *a priori*

Segundo Box e Tiao (1992), uma distribuição *a priori* tem um importante papel na análise bayesiana e é usada para representar o que é conhecido sobre a incerteza que deve ser quantificada antes de se avaliar os dados. Essa distribuição é utilizada para representar o conhecimento que se tem sobre p e θ, antes da realização do experimento ou, simplesmente, demonstrar a ignorância relativa.

Existem muitas formas de se especificar a distribuição *a priori* de forma a representar o conhecimento (ou ignorância) a respeito das incertezas. As mais

comuns são as distribuições *a priori* conjugadas, não informativas e hierárquicas.

2.1.2.1 Distribuições *a priori* hierárquicas

A distribuição *a priori* de p depende do valor do hiperparâmetro θ e pode-se escrever $\pi(p, \underset{\sim}{y})$, ao invés de $\pi(p)$. Além disso, em vez de fixar valores para o hiperparâmetro, pode-se especificar uma distribuição *a priori* $\pi(\theta)$, completando, assim, o segundo estágio na hierarquia. Assim, a distribuição *a priori* conjunta é simplesmente $\pi(p,\theta) = \pi(p)\pi(\theta)$ e a distribuição marginal de p pode ser, então, obtida por integração, como

$$\pi(p) = \int \pi(p,\theta)d\theta = \int \pi(p)\pi(\theta)d\theta \qquad (3)$$

A distribuição *a posteriori* conjunta fica definida como

$$\pi(p,\theta \mid \underset{\sim}{y}) \propto L(\underset{\sim}{y} \mid p,\theta)\pi(p)\pi(\theta) \qquad (4)$$

2.1.2.2 Distribuições não informativas

Quando a distribuição *a priori* adotada não interfere na distribuição *a posteriori*, sendo esta última caracterizada somente pelas informações provenientes dos dados, então, temos a distribuição não informativa. A primeira ideia de "não informação" *a priori* que se pode ter é pensar em todos os possíveis valores de θ como igualmente prováveis, isto é, com uma distribuição *a priori* uniforme (PAULINO; TURKMAN; MURTEIRA, 2003).

2.1.3 Distribuição *a posteriori*

Dados as distribuições *a priori* $\pi(p)$ e $\pi(\theta)$, o modelo de probabilidade $L(p,\theta \mid \underset{\sim}{y})$ e os dados $\underset{\sim}{y}$, é possível calcular a distribuição de probabilidade $\pi(p,\theta,\underset{\sim}{y})$, a qual é chamada de distribuição *a posteriori* conjunta de p e θ. Utilizando-se o princípio da inferência bayesiana, esse cálculo resume-se à equação em (4):

$$\pi(p,\theta \mid \underset{\sim}{y}) \propto L(p,\theta \mid \underset{\sim}{y})\pi(p)\pi(\theta) \qquad (4)$$

Essa distribuição $\pi(p,\theta \mid \underset{\sim}{y})$ contém toda a informação sobre os parâmetros p e θ. É importante apresentar alguns valores numéricos que possam resumir a informação contida na distribuição, tais como média, mediana, moda e intervalo de credibilidade dos parâmetros de interesse.

2.1.4 Intervalo de credibilidade

Intervalo de credibilidade (Gráfico 1), ou HPD, é o intervalo em que a densidade para todo ponto pertencente ao intervalo é maior do que para todo ponto não pertencente a ele, considerando o menor possível (BOX; TIAO, 1992). Para distribuições *a posteriori* simétricas, como a normal, tal intervalo é obtido de forma que seus extremos tenham densidades iguais. Esses intervalos são geralmente associados com 90%, 95% ou 99% da probabilidade total, sendo 95% o mais utilizado (CESPEDES, 2003).

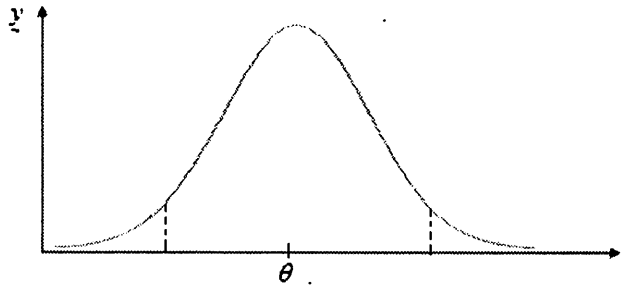

Gráfico 1 Intervalo de Credibilidade

2.2 Uma breve abordagem sobre os métodos numéricos utilizados na inferência clássica e bayesiana

Para a obtenção de uma amostra da distribuição *a posteriori* são utilizados alguns algoritmos. Tanner (1996) classificou esses algoritmos como métodos não Monte Carlo (MNMC), métodos iterativos Monte Carlo (MIMC) e métodos não iterativos Monte Carlo (MNIMC).

Os métodos MNMC caracterizam-se por não exigirem a entrada de números pseudoaleatórios. A aplicação desses métodos, em geral, se dá na estimação de parâmetros de locação, como, por exemplo, a moda de uma função de verossimilhança ou a moda de uma distribuição *a posteriori*. Dentre estes métodos podem-se citar o Newton-Raphson, EM (Esperança e Maximização) e aproximação de Laplace, sendo este último utilizado para calcular aproximações de alta ordem de precisão para a função de interesse.

A implementação dos métodos MIMC envolve um processo iterativo, portanto, requer a entrada de números pseudoaleatórios no contexto bayesiano. Os exemplos clássicos são Gibbs Sampling e Metropolis-Hastings, apresentados por Gelman et al. (2003) e Slice Sampling, propostos por Neal (2003). A utilização desses métodos é feita em função da forma como a distribuição *a posteriori* é apresentada. Se a distribuição *a posteriori* é escrita em função de

distribuições condicionais fechadas e conhecidas, deve-se utilizar o algoritmo Gibbs Sampling. Já diante de distribuições condicionais incompletas ou desconhecidas, utiliza-se o método de Metropolis-Hastings ou Slice Sampling.

Vale ressaltar que esses métodos têm o mesmo objetivo de aproximar a distribuição *a posteriori* da distribuição exata do parâmetro a ser estimado, representada pela distribuição marginal. Além disso, devido ao processo de iteração existente nesses métodos, a precisão de seus resultados é verificada por meio de vários critérios de convergência, podendo propiciar conclusões diferenciadas.

Os métodos MNIMC também requerem a entrada de números aleatórios, os quais não constituem um processo iterativo. O fato é que esses números representam amostras da função de interesse, seja a verossimilhança ou *a posteriori.* Os principais métodos representativos dessa classe são o de amostragem por importância conhecido como *importance sampling* (TANNER, 1996) e o método de aceitação/rejeição, sendo que este último utiliza uma densidade auxiliar para gerar amostras aleatórias para as densidades de interesse que são de difícil tratamento analítico (CIRILLO, 2006).

2.2.1 Monte Carlo via cadeias de Markov (MCMC)

Esse método consiste na simulação de amostras aleatórias de uma distribuição de interesse. Um exemplo é a simulação de valores de uma distribuição marginal *a posteriori* de um parâmetro de interesse. As amostras aleatórias geradas são dependentes. Para fazer inferência é necessário que as amostras aleatórias sejam independentes. Para isso é necessário tomar valores a cada t iterações ("thin").

2.2.1.1 Cadeias de Markov (Markov-Chain)

Considere um sistema envolvendo variáveis aleatórias discretas:

$$\left.\begin{array}{l} X_1 = \{x_1, x_2, ..., x_n\} \\ X_2 = \{x_1, x_2, ..., x_n\} \\ \vdots \\ X_J = \{x_1, x_2, ..., x_n\} \end{array}\right\}, X_j = x_i; \quad j = 1, ..., J; \quad i = 1, ..., n,$$

em que j representa diferentes instantes (tempo) ou iterações e i representa os possíveis valores da variável X_j.

O processo estocástico que define uma cadeia de Markov é dado por

$$P(X_j = x_i \mid X_{j-1} = x_k, X_{j-2} = x_l, ..., X_1 = x_m) = P(X_j = x_i \mid X_{j-1} = x_k).$$

Assume-se que a probabilidade da variável aleatória assumir um valor em um determinado tempo (iteração) depende apenas do valor assumido por ela em um instante anterior.

2.2.1.2 Algoritmo Gibbs Sampling (GS)

A distribuição de Gibbs (multivariada) é uma importante ferramenta na área de física.

Geman e Geman (1984) utilizaram a teoria de cadeias de Markov para gerar dados desta distribuição.

Gelfand e Smith (1990) aplicaram o algoritmo desenvolvido por Geman e Geman (1984) para obter amostras (gerar dados) de distribuições marginais *a posteriori* em modelos de regressão linear.

O Algoritmo Gibbs Sampling exige que as distribuições condicionais completas *a posteriori* sejam distribuições de probabilidade conhecidas.

O mecanismo gera valores das distribuições marginais *a posteriori* via distribuições condicionais completas *a posteriori*. O processo é feito da seguinte forma:

(1°) Inicialize a cadeia assumindo valores iniciais:

$\underset{\sim}{\theta}^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, ..., \theta_p^{(0)}) \rightarrow$ iteração zero

(2°) Obtenha $\underset{\sim}{\theta}^{(1)} = (\theta_1^{(1)}, \theta_2^{(1)}, ..., \theta_p^{(1)})$

$\theta_1^{(1)} \sim \pi(\theta_1 \mid \theta_2^{(0)}, ..., \theta_p^{(0)}, \underset{\sim}{y}) \rightarrow$ entra com valores iniciais $\theta_2^{(0)}, ..., \theta_p^{(0)}$

para gerar o valor $\theta_1^{(1)}$.

$\theta_2^{(1)} \sim \pi(\theta_2 \mid \theta_1^{(1)}, ..., \theta_p^{(0)}, \underset{\sim}{y})$

$\vdots$

$\theta_p^{(1)} \sim \pi(\theta_p \mid \theta_1^{(1)}, ..., \theta_{p-1}^{(1)}, \underset{\sim}{y})$.

O objetivo é gerar os valores $\theta_1^{(1)}, \theta_2^{(1)}, ..., \theta_p^{(1)}$ para substituir em $\underset{\sim}{\theta}^{(1)}$.

(3°) Obtenha $\underset{\sim}{\theta}^{(2)} = (\theta_1^{(2)}, \theta_2^{(2)}, ..., \theta_p^{(2)})$

$\theta_1^{(2)} \sim \pi(\theta_1 \mid \theta_2^{(1)}, ..., \theta_p^{(1)}, \underset{\sim}{y}) \rightarrow$ entra com valores iniciais $\theta_2^{(1)}, ..., \theta_p^{(1)}$

para gerar o valor $\theta_1^{(2)}$.

$\theta_2^{(2)} \sim \pi(\theta_2 \mid \theta_1^{(2)}, ..., \theta_p^{(1)}, \underset{\sim}{y})$

$\vdots$

$\theta_p^{(2)} \sim \pi(\theta_p \mid \theta_1^{(2)}, ..., \theta_{p-1}^{(2)}, \underset{\sim}{y})$.

(4°) Após $j \rightarrow \infty$ iterações $\pi(\theta_i \mid \underset{\sim}{y}) = \pi(\theta_i \mid \theta_{i-1}, \underset{\sim}{y})$, ou seja, ao gerar valores da distribuição condicional completa *a posteriori*, considera-se que tais valores são amostras das distribuições marginais *a posteriori*.

Um exemplo para representar o gráfico da distribuição marginal *a posteriori* do parâmetro θ_1 pode ser visto no Gráfico 2.

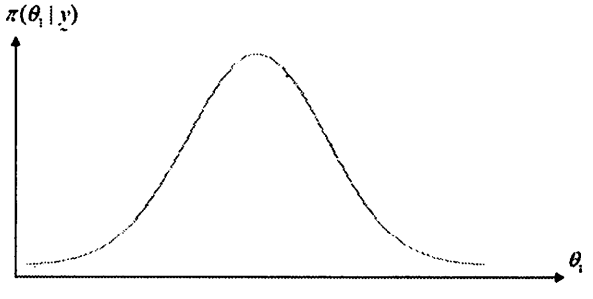

Gráfico 2 Distribuição marginal *a posteriori* de θ_1 proveniente do conjunto de valores gerados $\theta_1 = (\theta_1^{(0)}, \theta_1^{(1)}, ..., \theta_1^{(j)})$

2.2.1.3 Algoritmo Metropolis-Hastings (MH)

Metropolis et al. (1953) desenvolveram um método de simulação de valores de distribuição que não se caracterizaram como distribuições de probabilidade conhecidas (área de físico-química).

Hastings (1970) generalizou o método para situações envolvendo cadeias de Markov.

Chib e Greenberg (1995) apresentaram uma revisão detalhada do método, incluindo suas propriedades assintóticas e várias aplicações em inferência bayesiana.

Diferentemente do Gibbs Sampling, o Metropolis-Hastings não exige que as distribuições condicionais completas *a posteriori* sejam distribuições de probabilidade conhecidas.

2.2.1.3.1 Inferência uniparamétrica

Seja $\phi \mid \underset{\sim}{y}$ uma distribuição *a posteriori* cuja função densidade probabilidade (fdp) é dada por $\pi(\phi \mid \underset{\sim}{y})$ e tal distribuição não se caracteriza como uma distribuição de probabilidade conhecida.

Seja $q(\phi)$ uma distribuição candidata definida de acordo com o espaço paramétrico de ϕ, sendo esta distribuição conhecida.

A ideia básica do Metropolis-Hastings é gerar valores de $\phi \mid \underset{\sim}{y}$ indiretamente por meio de $q(\phi)$.

(1°) Especifica-se um valor inicial $\phi^{(0)}$.

(2°) Gerar $\phi^{(1)}$;

$\phi^{(1)} \sim q(\phi)$ (valor gerado da distribuição candidata).

$\alpha^{(1)} = \min\left(1, \dfrac{\pi(\phi^{(1)} \mid \underset{\sim}{y})}{\pi(\phi^{(0)} \mid \underset{\sim}{y})}\right)$; em que $\pi(\phi^{(1)} \mid \underset{\sim}{y})$ é o valor da *posteriori* ao substituir o valor gerado $\phi^{(1)}$ e $\pi(\phi^{(0)} \mid \underset{\sim}{y})$ é o valor da *posteriori* ao substituir o valor inicial.

Se $\alpha^{(1)} > U^{(1)}$, sendo $U^{(1)} \sim U(0,1)$, aceita-se $\phi^{(1)}$ como um valor gerado da distribuição de interesse (dita desconhecida).

Se $\alpha^{(1)} \leq U^{(1)}$, rejeita-se $\phi^{(1)}$ e assume-se que $\phi^{(1)} = \phi^{(0)}$.

(3°) Seja $\alpha^{(1)} > U^{(1)}$

$\phi^{(2)} \sim q(\phi)$ (valor gerado da distribuição candidata).

$\alpha^{(2)} = \min\left(1, \dfrac{\pi(\phi^{(2)} \mid \underset{\sim}{y})}{\pi(\phi^{(1)} \mid \underset{\sim}{y})}\right)$;

Aplica-se novamente a regra de decisão; se $\alpha^{(2)} > U^{(2)}$, aceita-se $\phi^{(2)}$. Caso contrário $\phi^{(2)} = \phi^{(1)}$.

Seja $\alpha^{(1)} \leq U^{(1)}$

$\phi^{(2)} \sim q(\phi)$

$$\alpha^{(2)} = \min\left(1, \frac{\pi(\phi^{(2)} \mid \underset{\sim}{y})}{\pi(\phi^{(1)} \mid \underset{\sim}{y})}\right) = \min\left(1, \frac{\pi(\phi^{(2)} \mid \underset{\sim}{y})}{\pi(\phi^{(0)} \mid \underset{\sim}{y})}\right);$$

(4°) Repetir os passos anteriores até a convergência.

A cadeia gerada pode apresentar valores repetidos quando $\phi^{(j)} = \phi^{(j-1)}$.

2.2.1.3.2 Escolha da distribuição candidata

Para escolher a distribuição candidata existem várias propostas, mas, geralmente, utiliza-se uma normal $N(\mu_\phi, \sigma_\phi^2 k)$ com valores de referência (literatura) para μ_ϕ e σ_ϕ^2 e uma constante k (envelope) para controlar a dispersão, como se pode observar no Gráfico 3.

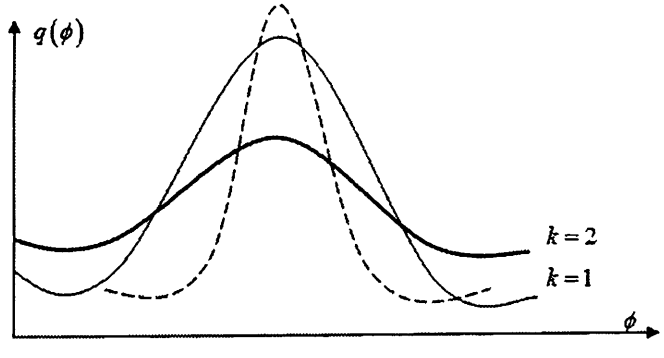

Gráfico 3 Distribuição proposta com diferentes valores para k

A constante k, também chamada de *envelope*, é a distribuição proposta. A curva para $k = 1$ "cobre" mais a curva pontilhada do que a curva para $k = 2$.

Existem divergências na literatura sobre a **taxa de aceitação** que uma distribuição candidata deve apresentar.

O cálculo é feito da seguinte forma:

$$\left.\begin{array}{l} N^{\circ}\ iterações \rightarrow 100\% \\ N^{\circ}\ iter.\ aceitas \rightarrow t_a \end{array}\right\} \Rightarrow \boxed{t_a = \frac{100\% N^{\circ}\ iter.\ aceitas}{N^{\circ}\ iterações}}.$$

Neal e Roberts (2008) e Roberts, Gelman e Gilks (1997) sugeriram que a taxa de aceitação deve variar entre 23% a 45%, sendo o ideal 25%. Já segundo Spiegelhalter et al. (2003), essa taxa de aceitação deve variar entre 20% a 40% e, para Roberts e Rosenthal (2001), essa variação deve ser entre 10% a 40%.

Teoricamente, qualquer candidata pode ser usada se $N^{\circ}\ iter. \rightarrow \infty$, mas candidatas que geram t_a entre $20\% \leq t_a \leq 40\%$ exigem um número menor de iterações até a convergência.

2.2.1.3.3 Inferência multiparamétrica

É necessário obter as distribuições condicionais completas *a posteriori*, da mesma forma que foi feito com o algoritmo Gibbs Sampling, porém, não é necessário encontrar tais distribuições com formas (núcleos das distribuições) conhecidas.

Sejam $\pi(\phi_1 \mid \phi_2, \underset{\sim}{y})$ e $\pi(\phi_2 \mid \phi_1, \underset{\sim}{y})$ as distribuições condicionais completas *a posteriori*.

$$\phi_1^{(1)} \sim \phi_1 \mid \phi_2^{(0)}, \underset{\sim}{y} \Rightarrow \begin{cases} \phi_1^{(1)} \sim q(\phi_1) \\ \alpha = \min\left(1, \dfrac{\pi(\phi_1^{(1)} \mid \phi_2^{(0)}, \underset{\sim}{y})}{\pi(\phi_1^{(0)} \mid \phi_2^{(0)}, \underset{\sim}{y})}\right) \\ se\ \alpha > U(0,1)\ aceita - se\ \phi_1^{(1)} \\ se\ \alpha \leq U(0,1)\quad \phi_1^{(1)} = \phi_1^{(0)} \end{cases}$$

(1°)

$$\phi_2^{(1)} \sim \phi_2 \mid \phi_1^{(1)}, \underset{\sim}{y} \Rightarrow \begin{cases} \phi_2^{(1)} \sim q(\phi_2) \\ \alpha = \min\left(1, \dfrac{\pi(\phi_2^{(1)} \mid \phi_1^{(1)}, \underset{\sim}{y})}{\pi(\phi_2^{(0)} \mid \phi_1^{(1)}, \underset{\sim}{y})}\right) \\ se\ \alpha > U(0,1)\ aceita - se\ \phi_2^{(1)} \\ se\ \alpha \leq U(0,1)\quad \phi_2^{(1)} = \phi_2^{(0)} \end{cases}$$

(2°) Repete-se o processo até obter convergência.

Neste processo assume-se que ambas as distribuições condicionais completas são desconhecidas (ϕ_1 e ϕ_2).

Pode ocorrer que a distribuição completa de ϕ_1 seja conhecida e a de ϕ_2 não, ou vice-versa, sendo, neste caso, necessário usar (GS) para um parâmetro e o (MH) para o outro.

2.2.1.4 Verificação de convergência

Sendo métodos MCMC algoritmos iterativos, a condição de convergência deve ser verificada (se realmente $j \to \infty$).

Os valores gerados das distribuições condicionais completas serão considerados valores da distribuição marginal apenas se $j \to \infty$.

Ao imprimir os valores que estão no Gráfico 4, na faixa de equilíbrio (região B), obtém-se a distribuição *a posteriori* do parâmetro em questão. Na

situação de equilíbrio as amostras das distribuições condicionais completas *a posteriori* são consideradas amostras das distribuições marginais *a posteriori*.

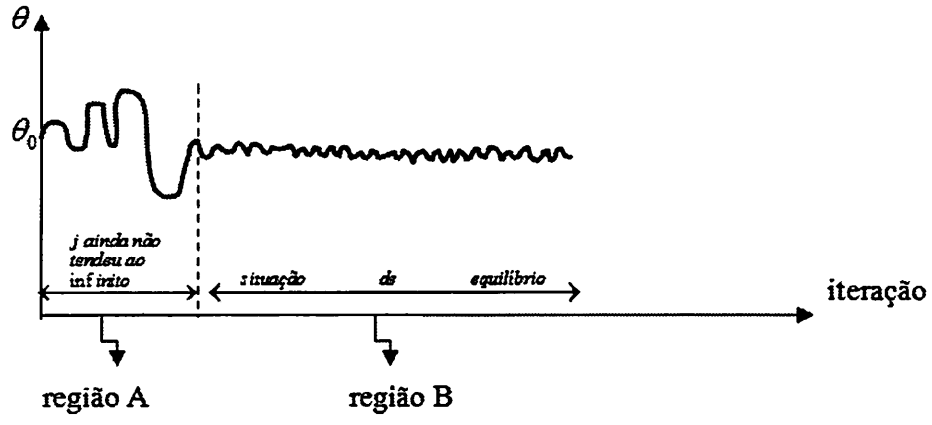

Gráfico 4 Valores gerados das distribuições condicionais completas a posteriori

As amostras observadas no Gráfico 4, na região A, devem ser retiradas (*burn-in*), ou seja, deve-se eliminar as amostras iniciais. Esses valores que foram gerados são dependentes (cadeia de Markov). Para fazer inferência, é preciso "quebrar" tal dependência. Para isso, é necessário tomar valores a cada t iterações (*thin*), garantindo, assim, que as amostras aleatórias da distribuição marginal sejam independentes. No Gráfico 5 está ilustrado o que foi exposto.

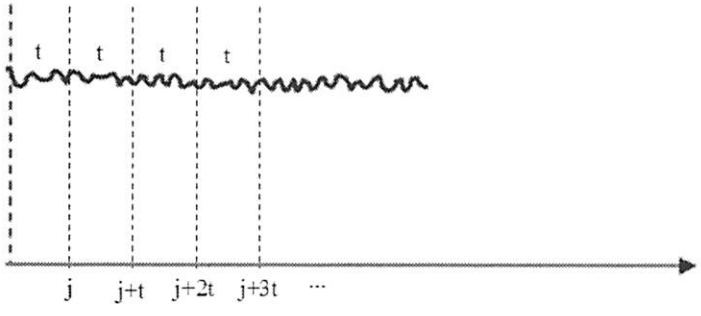

Gráfico 5 Valores tomados a cada t iterações ("thin"), garantindo, assim, que as amostras aleatórias da distribuição marginal sejam independentes

Um exemplo é dado para melhor compreensão do processo.

Para um conjunto de dados qualquer efetuou-se um processo de 60.000 iterações, *burn-in* sendo descartadas as 10.000 iniciais, para o período de aquecimento da cadeia. E, para assegurar a independência da amostra, considerou-se um espaçamento entre os pontos amostrados de tamanho 10 (*thin*). Assim, deve-se observar uma amostra de tamanho 5.000 para o parâmetro de interesse. A distribuição marginal a posteriori de θ proveniente das 5.000 iterações restantes pode ser vista no Gráfico 6.

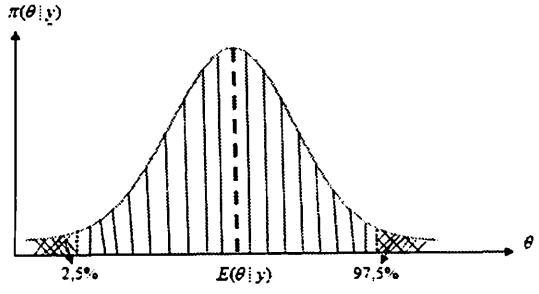

Gráfico 6 Distribuição marginal *a posteriori* de θ proveniente das 5.000 iterações restantes

Existem inúmeros métodos de verificação "não subjetiva" (não visual da convergência) e estes estão disponíveis no pacote Bayesian Output Analysis (BOA) do software R Development Core Team (2010).

2.2.1.5 Introdução aos métodos de Monte Carlo via cadeia de Markov no software WinBUGS

WinBUGS é um software estatístico que utiliza na análise bayesiana o método Monte Carlo via cadeias de Markov (MCMC).

É com base na *Bayesian inference Using Gibbs Sampling* (BUGS) que se iniciou o projeto. O WinBUGS foi desenvolvido pelo projeto BUGS, por uma equipe de pesquisadores do Reino Unido pertencentes à unidade de Bioestatística do MRC Biostatistics Unit (LUNN et al., 2000), em Cambridge, e da Escola Imperial da Faculdade de Medicina (Imperial College School of Medicine), em Londres. A última versão do WinBUGS é a versão 1.4.3, lançada por Spiegelhalter et al. (2007). O desenvolvimento agora está no foco do OpenBUGS (projeto da Universidade de Helsinki), uma versão de software livre do pacote. OpenBUGS é a plataforma principal de desenvolvimento e atualmente é experimental, mas acabará se tornando a versão padrão.

O WinBUGS 1.4.3 continua disponível como uma versão estável para uso rotineiro, mas não está mais sendo desenvolvido.

2.3 Distribuição zero - modificada

Segundo Dietz e Böhning (2000), a ocorrência de muitos ou poucos zeros pode advir das seguintes situações:

a) nem todos os membros da população de estudo são afetados pelo processo, de modo que a inflação de zero ocorre devido à resposta zero de membros não afetados;

b) certos problemas inevitáveis ao processo de amostragem levam a um crescimento ou decrescimento da chance de indivíduos da população com contagem zero serem amostrados;

c) não existe chance de ocorrer uma observação zero dentro da amostra (distribuições positivas, truncadas em zero);

d) uma combinação das situações (a) e (c), em que parte da população vem de um processo positivo e a outra parte, a qual não é afetada, fornece as observações de contagem zero.

Considere Y uma variável aleatória definida sobre os inteiros {0, 1, 2,...} com função de probabilidade $\pi_Y(\underset{\sim}{y} \mid \theta)$, tal que $\theta = E(Y)$. Considera-se para Y as funções de probabilidade de Poisson com parâmetro θ e binomial com parâmetros m e θ.

Teorema 1 A distribuição zero - modificada (ZMD) de Y é dada por:

$$\pi_{ZM}(\underset{\sim}{y} \mid p, \theta) = p\pi(\underset{\sim}{y}, 0) + (1-p)\pi_Y(\underset{\sim}{y}, \theta) \qquad (5)$$

em que p é o parâmetro que modela a ocorrência de zeros, com $0 \leq p \leq \frac{1}{1-\pi_Y(0|\theta)}$ e $\pi(\underset{\sim}{y},0)$ é uma distribuição degenerada com $\theta = 0$,

$$\pi(\underset{\sim}{y},0) = \begin{cases} 1, & se \quad y = 0 \\ 0, & se \quad y > 0 \end{cases} \tag{6}$$

Corolário 1 Seja Y uma variável aleatória com distribuição dada em (5). A média e a variância de Y são, respectivamente, $\theta_{ZM} = (1-p)\theta$ e $\sigma^2_{ZM} = (1-p)\theta + \frac{p}{(1-p)}[(1-p)\theta^2]$; em que θ é a média da distribuição usual de Y.

Proposição 1 Se $p = 1$, a equação em (5) é uma distribuição zero - modificada com massa no ponto zero.

Proposição 2 Se $0 < p < 1$, a equação em (5) é a distribuição inflacionada de zeros (ZID), a qual tem uma proporção de zero adicional (situações (a), (b) e (d)).

Proposição 3 Se $p = 0$, a equação em (5) é a distribuição usual (Poisson e binomial).

Proposição 4 Se $[1-\exp(\theta)]^{-1} \leq p < 0$, a equação em (5) é a distribuição deflacionada de zeros (ZDD) (situações (b) e (d)).

Proposição 5 Se $p = \frac{1}{1-\pi_Y(0|\theta)}$, a equação em (5) é uma distribuição zero truncada (ZTD).

De acordo com a proposição (2), apresentada por Dietz e Böhning (2000), pode-se afirmar que, para a distribuição inflacionada de zeros, $0 < p < 1$.

Se a distribuição *a posteriori* conjunta não for tratável algebricamente, resumos *a posteriori* de interesse são obtidos de amostras desta distribuição utilizando-se algoritmos MCMC, tais como Gibbs Sampling e Metropolis-Hastings. Para utilizar esses algoritmos é necessário obter as distribuições condicionais completas *a posteriori*.

3 MATERIAL E MÉTODOS

3.1 Material

O material utilizado neste trabalho foi um conjunto de dados do jornal britânico London Times, retirado do artigo de Schilling (1947). Esses dados correspondem ao número de notificações de mortes, por dia, de mulheres de 80 anos de idade ou mais, durante três anos consecutivos, em Londres. Esses dados podem ser vistos como um exemplo no qual a taxa de mortalidade durante os meses de inverno é maior do que nos meses de verão. Com isso, observa-se uma superdispersão nos dados que agora é devido à ocorrência de mais valores nulos na amostra do que seria esperado para dados que seguem a distribuição Poisson. Assim, esse conjunto de dados foi analisado por um modelo com distribuição ZIP. A descrição dos dados está apresentada na Tabela 6.

Tabela 6 Conjunto de dados do número de notificações de mortes por dia de mulheres com 80 anos de idade ou mais, durante três anos consecutivos, em Londres.

Observação do número de mortes (n)	**Frequência observada de n (F_{obs})**
0	162
1	267
2	271
3	185
4	111
5	61
6	27
7	8
8	3
9	1

3.2 Métodos

3.2.1 Modelo com distribuição ZIP

Assumindo que variável aleatória Y possui distribuição Poisson inflacionada de zeros, a generalização do modelo com distribuição Poisson (JOHNSON; KOTZ; KEMP, 1992) foi definida da seguinte forma:

$$f(y;p,\theta)=\begin{cases} p+(1-p)\exp(-\theta), & se \quad y=0 \\ (1-p)\exp(-\theta)\dfrac{\theta^{y}}{y!}, & se\ y>0 \end{cases}, \tag{7}$$

em que p é um indicador de proporção de zeros e θ a taxa média da distribuição Poisson padrão. Essa distribuição sob a hipótese nula $H_0: p=0$ se reduz ao modelo com distribuição Poisson padrão.

A função de verossimilhança do modelo com distribuição ZIP é definida em (8) por:

$$\Rightarrow L(p,\theta \mid \underset{\sim}{y})=[p+(1-p)\exp(-\theta)]^{n_0}[(1-p)\exp(-\theta)\prod_{i=1}^{n}\frac{\theta^{y_i}}{y_i!}]^{(n-n_0)} \tag{8}$$

e o logaritmo neperiano da função de verossimilhança é definido em (9) por:

$$\begin{aligned} l(p,\theta) &= n_0 \log_e\{p+(1-p)\exp(-\theta)\} \\ &+(n-n_0)\log_e\left\{(1-p)\exp(-\theta)\prod_{i=1}^{n}\frac{\theta^{y_i}}{y_i!}\right\} \end{aligned} \tag{9}$$

3.2.2 Estimação bayesiana dos parâmetros p e θ do modelo com distribuição Poisson para dados inflacionados de zeros

Ghosh, Mukhopadhy e Lu (2006) apresentam um modelo de regressão bayesiano que permitiu a modelagem de dados com muitos zeros, com base em

outras distribuições discretas. Outros trabalhos relevantes que utilizam a inferência bayesiana para o modelo com distribuição Poisson para dados inflacionados de zero são apresentados por Mazin et al. (2008) e Rodrigues (2003, 2006).

Observando-se o modelo proposto em (7) e o logaritmo neperiano da função de verossimilhança definida em (9), constata-se que as quantidades de incerteza presentes no modelo são p e θ. Para utilizar a inferência bayesiana é necessário definir a distribuição *a priori* para essas incertezas.

3.2.3 Distribuição *a priori* dos parâmetros p e θ

Na estimação dos parâmetros p e θ será considerada a seguinte restrição: $0 < p < 1$ e $\theta > 0$. Com base nessa informação, as distribuições *a priori* para p e θ são dadas por (10) e (11):

$$\pi(p) \sim \textit{Uniforme}(a,b) \Rightarrow \pi(p) = \frac{1}{(b-a)} \tag{10}$$

$$\pi(\theta) \sim \textit{Gama}(\alpha,\beta) \Rightarrow \pi(\theta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}\theta^{\alpha-1}\exp(-\beta\theta) \tag{11}$$

3.2.4 Distribuição conjunta *a posteriori*

A distribuição *a posteriori* conjunta $\pi(p,\theta \mid \underset{\sim}{y})$ pode ser obtida pelo produto das distribuições a *priori* $\pi(p)$, $\pi(\theta)$ e da distribuição amostral, função de verossimilhança $L(p,\theta \mid \underset{\sim}{y})$.

Assim, a distribuição *a posteriori* conjunta será definida como:

$$\pi(p,\theta \mid \underset{\sim}{y}) \propto L(p,\theta \mid \underset{\sim}{y})\pi(p)\pi(\theta)$$

Considerando as equações (8), (10) e (11), a distribuição *a posteriori* conjunta é definida por:

$$\pi(p,\theta \mid \underset{\sim}{y}) \propto [p+(1-p)\exp(-\theta)]^{n_0}[(1-p)\exp(-\theta)\prod_{i=1}^{n}\frac{\theta^{y_i}}{y_i!}]^{(n-n_0)} \times \frac{1}{(b-a)}\frac{\beta^{\alpha}}{\Gamma(\alpha)}\theta^{\alpha-1}\exp(-\beta\theta) \tag{12}$$

A distribuição *a posteriori* conjunta não é tratável analiticamente, então, é necessário utilizar algoritmos para obter amostras desta distribuição. Dessa forma, é necessário obter as distribuições condicionais completas *a posteriori* de p e θ.

3.2.5 Distribuição condicional completa *a posteriori* de p

Para obter a distribuição condicional completa *a posteriori* de p, devem-se manter constantes os termos que não dependem do parâmetro p. Portanto, essa distribuição é definida como:

$$\pi(p,\theta \mid \underset{\sim}{y}) \propto L(p,\theta \mid \underset{\sim}{y})\pi(p)$$

$$\Rightarrow \pi(p \mid \theta,\underset{\sim}{y}) \propto [p+(1-p)\exp(-\theta)]^{n_0}[(1-p)\exp(-\theta)\prod_{i=1}^{n}\frac{\theta^{y_i}}{y_i!}]^{(n-n_0)} \times \underbrace{\frac{1}{(b-a)}}_{\text{constante, não depende de } p}$$

$$\Rightarrow \pi(p \mid \theta,\underset{\sim}{y}) \propto \underbrace{[p+(1-p)\exp(-\theta)]^{n_0}[(1-p)\exp(-\theta)\prod_{i=1}^{n}\frac{\theta^{y_i}}{y_i!}]^{(n-n_0)}}_{\text{núcleo de uma ZIP}(p,\theta)} \tag{13}$$

Dessa forma, $p \mid \theta, \underset{\sim}{y} \sim ZIP(p,\theta)$.

A distribuição condicional completa *a posteriori* do parâmetro p é uma distribuição de probabilidade conhecida. Logo, utiliza-se o algoritmo Gibbs Sampling para se obter uma amostra da distribuição marginal *a posteriori* de p.

3.2.6 Distribuição condicional completa *a posteriori* de θ

Para obter a distribuição condicional completa *a posteriori* de θ, deve-se manter constantes os termos que não dependem do parâmetro θ. Portanto, essa distribuição é definida como

$$\pi(p \mid \theta, \underset{\sim}{y}) \propto L(p,\theta \mid \underset{\sim}{y})\pi(\theta)$$

$$\Rightarrow \pi(\theta \mid p, \underset{\sim}{y}) \propto [p+(1-p)\exp(-\theta)]^{n_0}[(1-p)\exp(-\theta)\prod_{i=1}^{n}\frac{\theta^{y_i}}{y_i!}]^{(n-n_0)}$$

$$\times \underbrace{\frac{\beta^{\alpha}}{\Gamma(\alpha)}}_{constante,\ não\ depende\ de\ \theta} \theta^{\alpha-1}\exp(-\beta\theta)$$

$$\Rightarrow \pi(\theta \mid p, \underset{\sim}{y}) \propto \sum_{j=0}^{n_0} p^j[(1-p)\exp(\theta)]^{n_0-j}(1-p)^{n-n_0}$$

$$\times \exp[-\theta(n-n_0)]\prod_{i=1}^{n}\frac{\theta^{y_i(n-n_0)}}{(y_i!)^{(n-n_0)}}\theta^{\alpha-1}\exp(-\beta\theta)$$

$$\Rightarrow \pi(\theta \mid p, \underset{\sim}{y}) \propto \sum_{j=0}^{n_0} p^j[(1-p)\exp(\theta)]^{n_0-j}$$

$$\times \exp\left[(-\theta)^{n-n_0+\beta}\right]\prod_{i=1}^{n}\theta^{y_i(n-n_0)+\alpha-1}$$

$$\Rightarrow \pi(\theta \mid p, \underset{\sim}{y}) \propto \sum_{j=0}^{n_0}[\exp(\theta)]^{j}\exp\left[(-\theta)^{n-n_0+\beta}\right]\prod_{i=1}^{n}\theta^{y_i(n-n_0)+\alpha-1}$$

$$\Rightarrow \pi(\theta \mid p, \underset{\sim}{y}) \propto \underbrace{\sum_{j=0}^{n_0} \exp(j)}_{constante} \exp\left[(-\theta)^{n-n_0+\beta-1}\right] \theta^{\sum_{i=1}^{n} y_i (n-n_0)+\alpha-1}$$

$$\Rightarrow \pi(\theta \mid p, \underset{\sim}{y}) \propto \theta^{\overbrace{\sum_{i=1}^{n} y_i (n-n_0)+\alpha-1}^{\alpha^*}} \exp\left[(-\theta)^{\overbrace{\beta-1+n-n_0}^{\beta^*}}\right]$$

$$\Rightarrow \pi(\theta \mid p, \underset{\sim}{y}) \propto \underbrace{\theta^{(\alpha^*-1)} \exp[(-\theta)^{\beta^*}]}_{núcleo\ de\ uma\ Gama(\alpha^*,\beta^*)}. \qquad (14)$$

Dessa forma,

$$\theta \mid p, \underset{\sim}{y} \sim Gama(\alpha^*, \beta^*); \quad \begin{cases} \alpha^* = \sum_{i=1}^{n} y_i (n-n_0) + \alpha; \\ \beta^* = \beta + (n-n_0) - 1 \end{cases}$$

A distribuição condicional completa *a posteriori* do parâmetro θ é uma distribuição de probabilidade conhecida. Logo, utiliza-se o algoritmo Gibbs Sampling para se obter uma amostra da distribuição marginal *a posteriori* de θ.

3.2.7 Alguns comentários sobre a implementação do programa para simulação do conjunto de dados

O programa (APÊNDICE B) para a obtenção de uma amostra da distribuição conjunta *a posteriori* (12) e, consequentemente, para as distribuições marginais de p e θ é implementado no programa estatístico WinBUGS (SPIEGELHALTER et al., 2003) e consta, basicamente, dos seguintes passos:

Passo 1: definir o conjunto de dados.

Passo 2: definir as *prioris* para os parâmetros p e θ;

Passo 3: definir os hiperparâmetros dessas prioris:

i) a, b para p;

ii) α, β para θ;

Passo 4: atribuir valores iniciais para os parâmetros p e θ;

Passo 5: efetuar um processo de 60.000 iterações, sendo descartadas as 10.000 iniciais, para o período de aquecimento da cadeia (*burn-in*);

Passo 6: para assegurar a independência da amostra, considerar um espaçamento entre os pontos amostrados de tamanho 10 (*thin*);

Passo 7: após todos esses procedimentos, obtém-se uma amostra de tamanho 5.000 para cada parâmetro analisado.

4 RESULTADOS E DISCUSSÃO

4.1 Resultados da inferência bayesiana aplicada aos dados reais

Para a análise bayesiana dos dados da Tabela 6 é necessário definir os hiperparâmetros das distribuições *a priori* de p e θ. Considerando independência entre as quantidades de incerteza e distribuições *a priori* não informativas, optou-se por $a=0$, $b=1$, $\alpha=0{,}1$ e $\beta=0{,}01$, sendo a e b hiperparâmetros da distribuição *a priori* de p, e α e β hiperparâmetros da distribuição *a priori* de θ, ou seja, $U(0,1)$ e $Gama(0{,}1;0{,}01)$.

Para o conjunto de dados da Tabela 6 foi feito um processo de 60.000 iterações, sendo descartadas as 10.000 iniciais, para o período de aquecimento da cadeia ("burn-in) e, para assegurar a independência da amostra, foi considerado um espaçamento entre os pontos amostrados de tamanho 10 ("thin"). Assim, uma amostra de tamanho 5.000 foi observada para cada parâmetro.

Na Tabela 7 é apresentado um resumo *a posteriori* para cada quantidade de incerteza do modelo com distribuição ZIP.

Tabela 7 Resumo *a posteriori* dos parâmetros p e θ.

Par.	Média	d.p.	2,5%	50%(Mediana)	97,5%
p	0,949	0,013	0,923	0,949	0,975
θ	2,269	0,054	2,161	2,270	2,374

Apesar de não conhecer a distribuição marginal *a posteriori* dos parâmetros p e θ, obtêm-se as esperanças para os parâmetros de interesse. Assim, sabe-se que p e θ têm, em média, os valores de 0,949 e 2,269, respectivamente.

A convergência das cadeias de todos os parâmetros do modelo foi monitorada por meio da visualização gráfica do traço e da densidade, não existindo evidências contra a convergência.

No Gráfico 7 apresentam-se o traço e a densidade da distribuição marginal *a posteriori* dos parâmetros p e θ.

Os traços de p e θ são gráficos entre os valores das variáveis e o número de iterações. Esses gráficos são aleatórios, pois os valores de p e θ não ficaram concentrados em nenhuma área ao longo das iterações.

Os gráficos das densidades da marginal *a posteriori* dos parâmetros p e θ retorna a densidade (aproximada) de cada uma das variáveis analisadas.

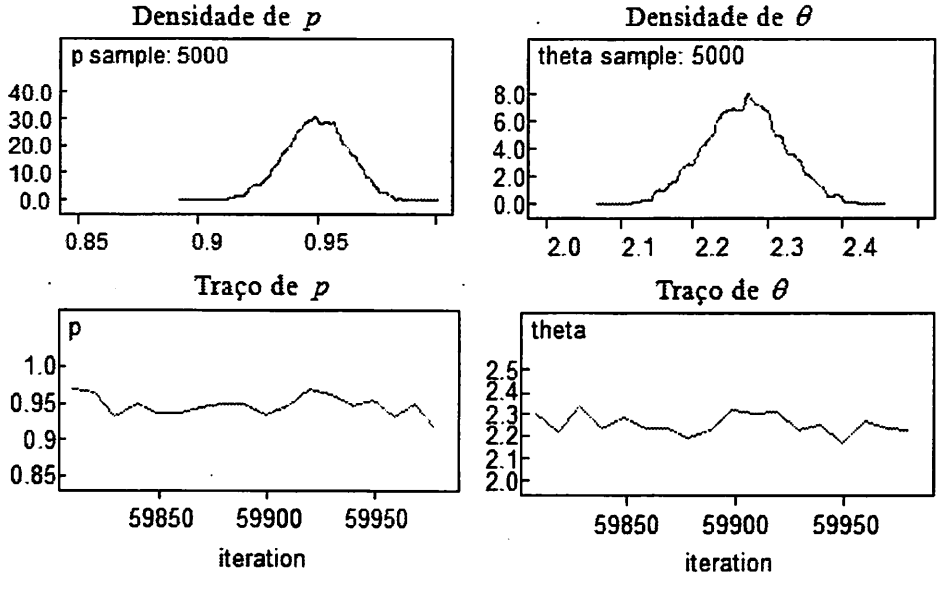

Gráfico 7 Traço e densidade da marginal *a posteriori* dos parâmetros p e θ

A autocorrelação da amostra gerada para os parâmetros p e θ é apresentada no Gráfico 8, no qual se observa um decaimento imediato nas barras do gráfico. Assim, a correlação existente entre os valores é a mais baixa possível, sendo possível afirmar que as estimativas desses parâmetros são confiáveis.

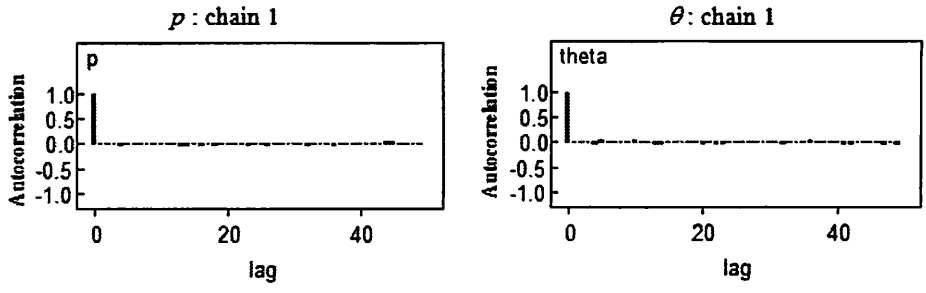

Gráfico 8 Autocorrelação da amostra gerada para os parâmetros p e θ

O Gráfico 9 representa as variáveis juntamente com seus respectivos intervalos de credibilidade com percentis 2,5% e 97,5%. Observa-se que estes intervalos não são muito amplos.

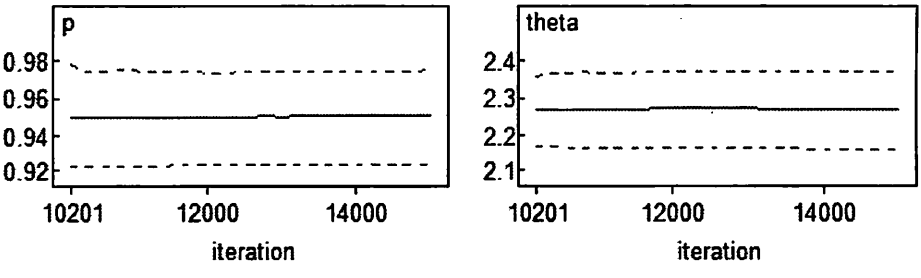

Gráfico 9 Intervalos de credibilidade de 95%

4.2 Teste escore de Vandenbroek (1995) com argumento bayesiano

O teste de escore, proposto por Vandenbroek (1995), foi utilizado para verificar se o conjunto de dados utilizado nesse artigo deve ser modelado pela distribuição Poisson ou pela distribuição Poisson com excesso de zeros (ZIP).

Foi utilizado como estimativa do parâmetro θ o valor da média *a posteriori* obtido na estimação bayesiana ($\hat{\theta}_{post.}$).

Assim, ao substituir o valor de $\hat{\theta}_{post.} = 2,269$ no teste escore de Vandenbroek $\left(S_1 = \frac{(n_0 - np_0)^2}{np_0(1-p_0) - n\bar{y}p_0^2} \right)$ para testar a hipótese nula $(H_0 : Y \sim Poisson(\theta))$, obtiveram-se os resultados descritos na Tabela 8.

Tabela 8 Resumo da estatística do teste escore (α=5%).

Estatística do teste (S_1)	**Quantil da χ_1^2**	**Região crítica**	**Aceit. / Rej. H**
31,01	3,84	$S_1 > 3,84$	Rejeita H_0

Observando-se os resultados descritos na Tabela 8, tem-se que $S_1 > \chi_1^2$. Assim, a hipótese nula deve ser rejeitada a 5% de significância. Conclui-se que o modelo com distribuição ZIP deve ser utilizado para representar os dados.

5 CONSIDERAÇÕES FINAIS

Em diversas áreas de estudo é comum encontrar dados de contagem e não são raras as situações em que são encontradas amostras com grande quantidade de zeros. O uso de métodos bayesianos é uma alternativa promissora para análise de dados desse tipo, conforme se observa nos resultados obtidos. O uso de métodos de simulação de amostras para a distribuição *a posteriori* de interesse via amostrador de Gibbs não requer grande custo computacional e pode facilmente ser implementado utilizando-se programas computacionais disponíveis. Segundo Mazin et al. (2008) uma limitação encontrada no uso do programa WinBUGS é que ele não permite a determinação do valor do *Deviance Information Criterion*, ou DIC (SPIEGELHALTER et al., 2002) em análises baseadas em modelos com misturas de distribuições. Isto pode ser um obstáculo em situações nas quais se pretende comparar diferentes modelos, o que pode ser contornado com alternativas como o fator de Bayes.

REFERÊNCIAS

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

CESPEDES, J. G. **Eficiência de produção**: um enfoque Bayesiano. 2003. 66 p. Dissertação (Mestrado em Agronomia) – Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2003.

CHIB, S.; GREENBERG, E. Understanding the metropolis-hastings algorithm. **American Statistician**, Alexandria, 49, p. 327-335, 1995.

CIRILLO, M. A. **Propostas de testes multivariados para comparar matrizes de covariâncias de populações normais dependentes**. 2006. 111 p. Tese (Doutorado em Estatística e Experimentação Agropecuária) – Universidade Federal de Lavras, Lavras, 2006.

COHEN, A. C. Estimating the parameteres of modified Poisson distribution. **Journal of the American Statistical Association**, Alexandria, v. 55, p.139-143, 1960.

COMULADA, W. S. et al. Reductions in drug use among young people living with HIV. **American Journal of Drug and Alcohol Abuse**, New York, v. 33, n. 3, p. 493-501, 2007.

DIETZ, E.; BÖHNING, D. On estimation of the Poisson parameter in zero-modified Poisson models. **Computational Statistics & Data Analysis**, Amsterdam, v. 34, p. 441-459, 2000.

GELFAND, A. E.; SMITH, A. F. M. Sampling-based approaches to calculating marginal densities. **Journal of the American Statistical Association**, Alexandria, v. 85, p. 398-409, 1990.

GELMAN, A. et al. **Bayesian data analysis**. London: Chapman and Hall, 2003. 668 p.

GEMAN, S.; GEMAN, D. Stochastic relaxation, Gibbs Distribution, and the Bayesian Restoration of images. **IEEE Transaction on Pattern Analysis and Machine Intelligence**, Washington, v. 6, p. 721-741, 1984.

GHOSH, S. K.; MUKHOPADHYAY , P.; LU, J. C. Bayesian analysis of zero-inflated regression models. **Journal of Statistical Planning and Inference**, Amsterdam, v.136, n. 4, p. 1360-1375, 2006.

GORASKI, A. **Distribution Z-Poisson**. Paris: Institut de Statistique de l'Université de Paris,1977. p. 43-45. (Publication, 12).

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

JOHNSON, N.; KOTZ, S.; KEMP, A. W. **Univariate discrete distributions**. 2nd ed. New York: J. Wiley, 1992. p. 312-318.

KARAZSIA, B. T.; VAN DULMEN, M. H. Regression models for count data: illustrations using longitudinal predictors of childhood injury. Journal of **Pediatric Psychology**,Washington, v. 33, n.10, p.1076-1084, 2008.

KEMP, A. W. Weighted discrepancies and maximum likelihood estimation for discrete distribution. **Communications in Statistics**, New York, v.15, p. 783-803, 1986.

LAMBERT, D. Zero - inflated Poisson regression with application to defects in manufacturing. **Technometrics**, Alexandria, v. 34, p. 1-4, 1992.

LUNN, D.J. et al. WinBUGS: a bayesian modelling framework: concepts, structure, and extensibility. **Statistics and Computing**, London, v. 10, p. 325-337, 2000.

MARTIN, D. C.; KATTI, S.K. Fitting of some contagious distributions to some available data by the maximum likelihood method. **Biometrics**, Washington, v. 21, p. 34-48, 1965.

MAZIN, S. C. et al. Uso de um modelo bayesiano de Poisson com excesso de zeros na análise de dados de lesões miocárdicas em recém-nascidos com cardiopatias congênitas complexas. **Revista Brasileira de Biometria**, São Paulo, v. 26 , n. 4 , p. 113 -125, 2008.

METROPOLIS, N. et al. Equation of state calculations by fast computing machines. **Journal of Chemical Physics,** Prince George's, v. 21, p. 1087-1092, 1953.

NEAL, P. J.; ROBERTS, G. O. Optimal scaling for random walk Metropolis on spherically constrained target densities. **Methodology and Computing in Applied Probability,** Hingham, v. 10, p. 277-297, 2008.

NEAL, R. M. Slice sampling. **The Annals of Statistics**, Philadelphia, v. 31, n. 3, p. 705-767, 2003.

PAULINO, C. D.; TURKMAN, M. A. A.; MURTEIRA, B. **Estatística bayesiana**. Lisboa: Fundação Calouste Gulbenkian, 2003. 446 p.

R DEVELOPMENT CORE TEAM. **R**: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2010.

RAMIS-PRIETO, R. et al. Modelling of municipal mortality due to haematological neoplasias in Spain. **Journal of Epidemiology and Community Health**, London, v. 61, n. 2, p.165-171, 2007.

ROBERTS, G. O.; GELMAN, A.; GILKS, W. R. Weak convergence and optimal scaling of random walk metropolis algorithms. **Annals of Applied Probability**, New York, v. 7, p. 110-120, 1997.

ROBERTS, G. O.; ROSENTHAL, J. S. Optimal scaling for various Metropolis-Hastings algorithms. **Statistical Science**, New York, v. 16, p. 351-67, 2001.

RODRIGUES, J. Bayesian analysis of zero-inflated distributions. **Communication in Statistics** , New York, v. 32, n. 2, p. 281-289, 2003.

RODRIGUES, J. Full Bayesian significance test for zero-inflated distributions. **Communication in Statistics**, New York, v. 35, n. 2, p. 299-307, 2006.

SCHILLING, W. A frequency distribution represented as the sum of two Poisson distributions. **Journal of the American Statistical Association**, Alexandria, v. 42, p. 407-424, 1947.

SPIEGELHALTER, D. et al. **Bayesian inference using gibbs sampling for windows (WinBUGS)**: version 1.4.3. Cambridge: MRC, 2007. Disponível em: <http://www.mrc-bsu.cam.ac.uk/bugs/>. Acesso em: 13 jan. 2011.

SPIEGELHALTER, D. J. et al. Bayesian measures of model complexity and fit (with discussion). **Journal of the Royal Statistical Society**, Oxford, v. 64, p. 583-640, 2002.

SPIEGELHALTER, D. J. et al. **WinBUGS**: version 1.4.3. Cambridge: MRC, 2003. Disponível em: <http://www.mrc-bsu.cam.ac.uk/ bugs/>. Acesso em: 13 jan. 2011.

TANNER, M. A. **Tools for statistical inference**: methods for the exploration of posterior distributions and likelihood functions. 3rd ed. New York: [s. n.]1996. 207 p.

VANDENBROEK, J. A score test for zero inflation in a Poisson-distribution. **Biometrics,** Washington, v. 51, n. 2, p. 738-743, 1995.

XIE, M.; HE, B.; GOH, T.N. Zero- inflated Poisson model in statistical process control. **Computing and Statistical Data Analysis**, Amsterdam, v. 38, p. 191-201, 2001.

APÊNDICE

APÊNDICE B: Programa WinBUGS

```
model {

  K<-10000

    for (i in 1:n) {

    # Definição das variáveis

    d[i]<-step(y[i]-1)                                  # d=I(y>0) para y>0

    # Verossimilhança

          ll[i]<-(1-d[i])*log(1-p*(1-exp(-theta))) + d[i]*(log(p*(1-exp(-
theta)))+y[i]*log(theta) -
                      theta - loggam(y[i]+1)-log(1-exp(-theta)))
       zeros[i]<-0
    zeros[i]~dpois(phi[i])
    phi[i]<-  - ll[i]+K
                    }


 #p0<- exp(-theta)


 # Priori

p ~ dunif(0,1)
theta ~ dgamma(0.1, 0.01)
          }

#Valor verdadeiro p=0.1478102 e theta=2.157
```

DADOS

```
list(y=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
1,1,1,1,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,
2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,
2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,
2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,
2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,
2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,
2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,
3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,
3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,
3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,
3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,4,4,4,4,4,4,4,4,4,4,4,4,4,
4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,
4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,
4,4,4,4,4,4,4,4,4,4,4,4,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,
5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,6,6,6,6,6,6,6,6,6,6,6,6,6,
6,6,6,6,6,6,6,6,6,6,6,6,6,7,7,7,7,7,7,7,7,8,8,8,9), n=1096)
```

INICIAIS list(theta=2, p=0.2)