Revista da Estatística UFOP, Vol III(3), 2014, ISSN 2237-8111
Edição especial: 59ª Reunião Anual da Regional Brasileira da Sociedade Internacional de Biometria - RBRAS
Trabalhos apresentados na versão poster.

Influencia de Distribuições *a priori* na Analise Bayesiana em dados de contagem

Olinda Fátima dos Santos[1]
Carla Regina Guimarães Brighenti[1]

1-Introdução

A utilização de informação *a priori* em inferência Bayesiana requer a especificação de uma distribuição *a priori* para a quantidade de interesse π, sendo esta representada por uma forma funcional, cujos parâmetros devem ser especificados de acordo com este conhecimento e são denominados de hiperparâmetros para distingui-los do parâmetro de interesse θ. Esta abordagem em geral facilita a análise e o caso mais importante é o de *prioris* conjugadas. A ideia é que as distribuições *a priori* e *a posteriori* pertençam a mesma classe de distribuições e assim a atualização do conhecimento que se tem de θ envolve apenas uma mudança nos hiperparâmetros [1].

Quando não existe informação *a priori* ou em que o conhecimento *a priori* é pouco significativo relativamente à informação amostral (o estado de conhecimento 'vago' ou 'difuso'), tem- se a chamada distribuição não-informativa [2].

Com base na distribuição marginal *a posteriori*, que contém toda a informação probabilística a respeito do parâmetro, estimativas bayesianas pontuais podem ser encontradas, como a média e mediana *a posteriori*, além de estimativas bayesianas intervalares, através da construção de intervalos de credibilidade [3].

No entanto, em diversas situações, a distribuição marginal *a posteriori* de um determinado parâmetro é difícil de ser encontrada, devido a impossibilidade de calcular analiticamente as integrais envolvidas. Dessa maneira, pode-se utilizar métodos aproximados baseados em simulação estocástica, como os métodos de Monte Carlo via cadeias de Markov (MCMC), que consistem em gerar valores de uma distribuição condicional *a posteriori* para cada parâmetro, como o algoritmo Gibbs [5]. Nos métodos de Monte Carlo via cadeias de Markov, tem-se a necessidade de diagnosticar a convergência das cadeias para a distribuição original, dos quais pode-se destacar o critério de Heidelberger & Welch do pacote CODA do R[4].

No caso de dados de contagem, duas distribuições são frequentemente utilizadas: a Binomial (n, π) e a Poisson (θ).

Assim, o objetivo deste trabalho foi estudar a influência de diferentes distribuições *a priori* na análise bayesiana de dados de contagem via simulação.

[1] DEZOO – UFSJ e-mail:carlabrighenti@ufsj.edu.br.

2-Metodologia

Para simulação da influencia das *prioris*, foram geradas amostras de ensaios de Bernoulli, seguindo os valores paramétricos fixados em π = 0,1; 0,3; 0,7; e 0,9, utilizando tamanho amostral 10, 30 ou 100. Foi utilizada a distribuição Beta como *priori*, sendo sua média e variância dadas por:

$$\mu = \frac{\alpha}{\alpha+\beta} \quad \text{e} \quad \sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

A *posteriori* será $p(x|\pi) \propto \pi^{\alpha+x-1}(1-\pi)^{\beta+n-x-1}$, ou seja, $(\pi|x) \sim Beta(\alpha+x, \beta+n-x)$ (Garthwaite et. al., 1995).

Após o processo de simulação da população, foram definidos os processos para obtenção das *prioris* sendo utilizados dois casos:

(i) *Prioris* conjugadas informativas: Foram utilizados os hiperparâmetros α = 1, 3 ou 15 e β = 1, 3 ou 15. Sendo assim, nove combinações de hiperparâmetros para cada um dos tamanhos amostrais, totalizando 27 casos.

(ii) *Prioris* conjugadas vagas: Foram utilizados hiperparâmetros $\alpha = \beta = 10^{-3}$ ou 10^{-5}, totalizando 12 casos.

No caso de dados de contagem provenientes de uma Distribuição de Poisson em que a função de verossimilhança dada por $L(\theta|x) \sim Gama(t+1, n)$ a distribuição *a priori* Gama foi utilizada como conjugada, sendo E(X) = α/β e V (X) = α/β^2. A distribuição *a posteriori* de $\theta|x$ segue uma distribuição Gama com parâmetros $(\alpha+t)$ e $(\beta+n)$, isto é, $\theta|x \sim Gama(\alpha+t, \beta+t)$,pois $\pi(\theta|x) \propto e^{-(n+\beta)\theta}\theta^{(\alpha+t)-1}$.

Gerou-se amostras da distribuição Poisson, seguindo os valores paramétricos fixados em θ = 0,1; 1; 5; e 10 e tamanhos amostrais 10, 30 ou 100. Após o processo de simulação da população, foram definidos os processos para obtenção das *prioris* sendo utilizados dois casos:

(i) *Prioris* conjugadas informativas: foi estudado o modelo da *priori* conjugada para o parâmetro que permite obter a distribuição a *posteriori* de modelo conhecido. Assim, foi utilizada a *priori* Gama (α,β) com hiperparâmetros α = 1, 3 ou 5 e β = 1, 3 ou 5. Sendo assim, nove combinações de hiperparâmetros para cada um dos tamanhos amostrais, totalizando 27 casos, em que alguns deles forneceram *prioris* Exponencial.

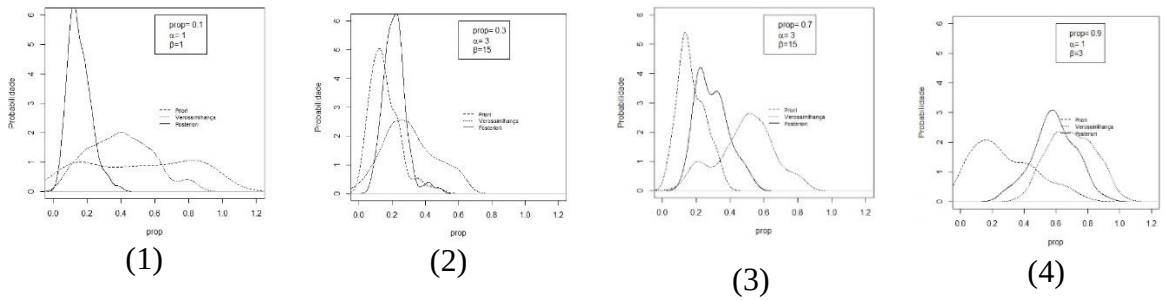
(ii) *Prioris* conjugadas vagas: Foram utilizados hiperparâmetros na distribuição Gama (α,β) sendo $\alpha = \beta = 10^{-3}$ ou 10^{-5} que forneceram distribuições vagas, com cada um dos tamanhos amostrais, totalizando 12 casos.

Para estudo do grau de informatividade das *prioris* desenvolveu-se uma rotina para utilização do pacote RBugs, especificando os comandos para o modelo e seus parâmetros e a geração das cadeias, bem como o critério de convergência. Para obtenção das distribuições marginais *a posteriori* foram gerados 11.000 valores em um processo MCMC, considerando descarte de 1.000 valores iniciais. A convergência das cadeias foi verificada usando o critério de Heidelberger & Welch do pacote CODA do R [4].

3- Resultados e Discussões

3.1- Caso Binomial

Em cada caso simulado, obteve-se a média, mediana, o Intervalo de Credibilidade a 95% e os gráficos correspondentes a *priori*, verossimilhança e *posteriori* (Figura 1).



(1) (2) (3) (4)

Figura 1 – Modelo de Gráficos gerados a partir de simulação para os seguintes valores para proporção: (1) $\pi = 0,1$; (2) $\pi = 0,3$; (3) $\pi = 0,7$; (4) $\pi = 0,9$.

Observou-se que para o tamanho amostral 10, as melhores estimativas ocorreram quando se utilizou o valor hiperparâmetro $\alpha=1$, independente do parâmetro β escolhido, que neste caso correspondem ao caso particular da *priori* Gama que gera uma distribuição Exponencial. Já para os tamanhos amostrais 30 e 100, observa-se que as piores estimativas foram obtidas quando $\alpha = 15$, sendo este fato, novamente independente do parâmetro β escolhido.

Para o caso $\pi= 0,3$ percebe-se que não houve uma regularidade entre os casos mais distantes do parâmetro real, nem das melhores estimativas. Pode-se ressaltar apenas que para o tamanho amostral 30, as piores estimativas foram obtidas quando da escolha de $\alpha = 15$, sendo este fato, novamente independente do parâmetro β escolhido.

Notou-se também que, para o caso $\pi = 0,7$, quanto maior o tamanho amostral melhores as estimativas obtidas, independente dos hiperparâmetros utilizados.

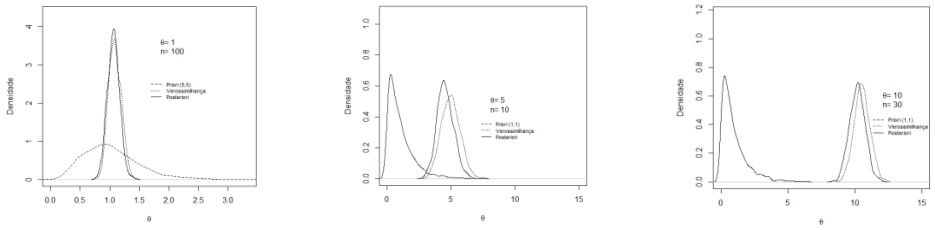
Para o caso $\pi = 0,9$ percebe-se que houve estimação mais afastada do verdadeiro valor paramétrico, para todos os tamanhos amostrais, nos casos da distribuição Gama (1,15), (3,15) e

(15,15), indicando que neste caso, a influencia da *priori* foi negativa quando utilizado $\beta = 15$ independente do parâmetro α escolhido.

Quando utilizado como hiperparâmetros os valores de α, β igual a 10^{-3} ou 10^{-5}, em apenas 1 dos 48 casos o verdadeiro valor do parâmetro não pertencia ao Intervalo de Credibilidade a 95%, evidenciando pequena influencia das *prioris* vagas.

3.2- Caso Poisson

Assim como no caso da Binomial, para cada valor paramétrico populacional (θ), em cada caso simulado, obteve-se a média, mediana, o Intervalo de Credibilidade a 95% e os gráficos correspondentes a *priori*, verossimilhança e *posteriori* (Figura 2).



(1) (2) (3)

Figura 2 – Gráficos de Estimação Bayesiana a partir de simulação computacional para os seguintes valores de θ: (1) $\theta = 1$; (2) $\theta = 5$; (3) $\theta = 10$.

Observou-se no caso $\theta = 1$, que todos os ICr a 95%, em todos os tamanhos amostrais, continham o valor paramétrico populacional simulado.

Para $\theta = 5$ apenas os ICr obtidos quando o tamanho amostral foi n = 10 e hiperparâmetro $\alpha = 1$ e quando n = 30 para $\alpha = 5$ e $\beta = 3$, não continham o verdadeiro valor paramétrico.

Quando simulou-se a população proveniente da Poisson (10) com tamanho amostral igual a 10, mais da metade dos ICr obtidos não contém ao verdadeiro valor paramétrico. Para os demais tamanhos de amostra apenas em um ICr o verdadeiro valor do parâmetro estava contido.

Pode-se concluir que o tamanho amostral foi o fator crucial para determinação de valores mais confiáveis. Percebeu-se que quanto maior o tamanho amostral, maiores foram as chances do valor paramétrico populacional pertencer aos Intervalos de Credibilidade gerados.

No caso das prioris vagas, para todos os hiperparâmetros utilizadas, evidenciou-se a pertinência do valor parametrico nos Intervalos de Credibilidade a 95%.

4 – Conclusões

A simulação dos parâmetros da Distribuição Binomial e Poisson não sofreram forte influência dos hiperparâmetros utilizados tanto para utilização das *priori* conjugadas informativas quanto para *prioris* vagas.

5 - Referências Bibliográficas

[1] BERNARDO, J. M.;A. F. M. SMITH. **Bayesian Theory**. Wiley: New York., 1994.

[2] BOX, G.E. P.; G. C. TIAO. **Bayesian Inference in Statistical Analysis**. Wiley Classics Library ed. Wiley-Interscience., 1992.

[3] GAMERMAN, D.; H. S. MIGON. **Inferência Estatística: Uma Abordagem Integrada.** Textos de Métodos Matemáticos. Instituto de Matemática, UFRJ, 1997.

[4] R DEVELOPMENT CORE TEAM. R: **A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.r-project.org. 2013.

[5] STURTZ, S., LIGGES, U., e GELMAN, A**. R2WinBUGS: A Package for Running WinBUGS from R.** Journal of Statistical Software, 12(3), 1-16, 2005.