TEORIA DA RESPOSTA AO ITEM (TRI): ESTIMAÇÃO BAYESIANA DA HABILIDADE DE INDIVÍDUOS

DÉBORA SPENASSATO[1], PAUL GERHARD KINAS[2]

Resumo – Este artigo tem por objetivo apresentar uma simulação da estimativa dos parâmetros das habilidades (θ) de indivíduos sujeitos a um teste, utilizando o método da Teoria da Resposta ao Item (TRI). Para isso, foi feito uma análise bayesiana, onde utilizou-se o método Monte Carlo via Cadeias de Markov (MCMC), implementado em linguagem R de programação e utilizando o recurso do OpenBUGS implementado nas bibliotecas R2WinBUGS e BRUGS. O desempenho do estimador bayesiano das habilidades foi avaliado comparando as estimativas com as habilidades reais utilizadas nas simulações dos testes. Apresenta-se também, conceitos básicos sobre TRI, sendo este um novo método de avaliação educacional e com aplicações em várias áreas do conhecimento.

Palavras-chaves: Inferência bayesiana, Monte Carlo via Cadeias de Markov, OpenBUGS, Teoria da Resposta ao Item.

ITEM RESPONSE THEORY (IRT): BAYESIAN ESTIMATION OF THE ABILITY OF INDIVIDUALS

[1] Mestranda do Programa de Pós-Graduação em Modelagem Computacional – FURG, deboraspenassato@hotmail.com.
[2] Professor Dr. em Estatística, Instituto de Matemática, Estatística e Física – FURG, paulkinas@furg.br.

Abstract – This article presents a simulation study for the estimation of ability measurements (θ) of individuals subjected to a test designed according to the methodology of Item Response Theory (IRT). We used a bayesian approach and the Markov Chain Monte Carlo (MCMC) procedure to obtain the simulated posterior samples via OpenBUGS. The procedure was implemented in the R language and used the R2WinBUGS and BRUGS libraries. The performance of the Bayes estimator for individual abilities was evaluated by comparison with the corresponding true values which were used to simulate the test result data. The method performs well in terms of coverage by be posterior credibility sets. Basic notions about IRT as a new method to grade education tests, and possible other applications for the methods are also included.

Key words: Bayesian inference, Markov Chain Monte Carlo, OpenBUGS, Item Response Theory.

INTRODUÇÃO

Instrumentos de avaliação são utilizados para avaliar o aprendizado dos alunos em escolas, universidades, em avaliações nacionais como ENEM, SAEB, etc. Uma metodologia de avaliação usada há muito tempo é a Teoria Clássica de Medidas (TCM), que não leva em consideração a habilidade do indivíduo, apenas verifica o seu escore final. Assim, não é possível fazer comparações entre séries diferentes ou de um ano para outro. Em contraste, a Teoria da Resposta ao Item (TRI), que ainda é recente no Brasil, e que será abordada neste artigo, se propõe a medir a habilidade dos indivíduos. Desta maneira, e sob algumas restrições, este novo método permite fazer comparações entre indivíduos que forem submetidos a diferentes provas. Porém, para que seja feito tais comparações é preciso que as estimativas estejam todas na mesma escala métrica das habilidades.

A literatura apresenta alguns modelos de TRI, que se diferenciam pelo número de parâmetros envolvidos. O modelo de um parâmetro envolve apenas a dificuldade

(*b*), o modelo de dois parâmetros envolve a dificuldade (*b*) e a discriminação (*a*) e o modelo de três parâmetros envolve a dificuldade (*b*), a discriminação (*a*) e a probabilidade de acerto ao acaso (*c*). Em todos os modelos a habilidade ou proficiência dos indivíduos é representada por θ.

O objetivo é estimar as habilidades dos indivíduos, considerando conhecidos os parâmetros *a, b* e *c* dos diferentes itens que compõe o teste. Fez-se a inferência das habilidades com enfoque bayesiano utilizando o ambiente R [10] para efetuar os cálculos estatísticos via software OpenBugs [12] e utilizando o método de Monte Carlo via Cadeias de Markov (MCMC).

Além da sua utilidade para a análise do desempenho escolar, a TRI também tem aplicabilidade em outras áreas do conhecimento como a avaliação de qualidade de vida de idosos, no desenvolvimento e validação de instrumentos em saúde mental, avaliação da usabilidade de sites de comércio eletrônico, entre outras.

A referência [1] apresenta uma proposta do uso de modelos da Teoria da Resposta ao Item (TRI) na análise de construtos, elaborados para medir a Gestão pela Qualidade Total (GQT) como uma alternativa à Teoria Clássica de Medida. É apresentado o modelo logístico de três parâmetros, assim como as interpretações dos parâmetros do modelo. Os resultados mostram que a TRI pode ser uma poderosa ferramenta na análise das práticas da GQT e da maturidade organizacional, dentro da filosofia da qualidade. O recurso computacional utilizado na análise desta pesquisa foi o programa BILOG.

Em [7], introduz-se a Teoria da Resposta ao Item (TRI) em sua forma usual, de um único grupo e para grupos múltiplos, e explica como a TRI para grupos múltiplos está sendo utilizada no Sistema Nacional de Avaliação da Educação Básica (SAEB) para a calibração dos itens e para a obtenção de uma escala única. A estimação dos parâmetros é feita pelo método de máxima verossimilhança marginal. Já em [5], poderá ser encontrado um capítulo sobre a estimativa bayesiana utilizando o método MCMC, com respectivas aplicações.

A seguir, apresentam-se alguns conceitos básicos sobre TRI de acordo com a literatura e, por intermédio de um estudo simulado, verifica-se se as habilidades estimadas reproduzem satisfatoriamente os valores reais sob os quais foram gerados.

MATERIAL E MÉTODOS

A metodologia da Teoria da Resposta ao Item (TRI) utiliza-se de modelos estatísticos que representam a relação entre a probabilidade de um indivíduo dar certa resposta a um item e seus traços latentes na área de conhecimento avaliada [2], levando em consideração que os escores não dependerão da dificuldade do teste para estimar a habilidade. Uma característica relevante da TRI é a possibilidade para comparar populações distintas, desde que os parâmetros dos itens estejam na mesma escala métrica da habilidade (θ).

Na teoria clássica de medidas (TCM), as análises são feitas com a soma dos escores do conjunto de itens como um todo. Na TRI, o interesse principal está na resposta, afirmativa ou não, que um respondente obtém para cada item, e não no escore bruto [4].

Dos modelos existentes na literatura, o que será utilizado neste artigo é o modelo logístico de três parâmetros (ML3), que tem resposta dicotômica, envolve apenas uma população e mede um único traço latente/habilidade para indivíduos.

Os três parâmetros são item-dependentes, caracterizando o seu grau de dificuldade (b), a discriminação (a) e a probabilidade de acerto ao acaso (c). A habilidade dos indivíduos é representada por θ.

A escolha do modelo ML3 deve-se ao fato dele ser o mais geral, entre os modelos dicotômicos para testes de múltipla-escolha. Sua formulação é dada por,

$$P(U_{ij}=1 \mid \theta_j) = c_i + (1-c_i)\frac{1}{1+e^{-Da_i(\theta_j - b_i)}}$$

com $i = 1, 2,...,I$ e $j = 1, 2,..., n$; onde:

U_{ij} é uma variável que assume os valores 0 ou 1, ou seja, um quando o indivíduo *j* acerta o item *i* e zero quando o indivíduo *j* erra o item *i*;

θ_j é a habilidade ou traço latente do indivíduo *j*;

$P(U_{ij}=1|\theta_j)$ é a probabilidade do indivíduo *j* com habilidade θ_j responder corretamente o item *i* ;

b_i é o parâmetro de que representa a dificuldade do item *i*, medido na mesma escala da habilidade. Quanto maior o seu valor, maior é a habilidade exigida do indivíduo para responder corretamente determinado item;

a_i é o parâmetro que corresponde a discriminação (ou inclinação) do item *i*, onde baixos valores indicam pouco poder de discriminação, não se espera que tenha valor negativo;

c_i é o parâmetro do item que representa a probabilidade de acerto ao acaso, ou seja, de indivíduos com baixa habilidade responderem corretamente o item *i*. Pode assumir valores entre zero e um;

D é um fator de escala, sendo ele constante e geralmente igual a um. Porém, para uma aproximação da função logística a uma função ogiva normal, utiliza-se o valor 1,7.

A relação existente entre $P(U_{ij} = 1|\theta_j)$ e os parâmetros do modelo é mostrada na FIGURA 1, que é chamada de Curva Característica do Item (CCI). Percebe-se a partir do gráfico que a CCI tem forma de "S" com inclinação e deslocamento na escala de habilidade definidos pelos parâmetros do item [2].

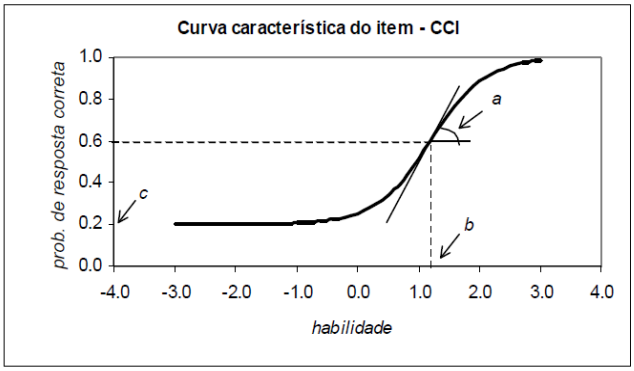


Figura 1 Exemplo de uma curva característica do item de um modelo logístico de três parâmetros [2].

Observa-se na FIFURA 1 que o eixo das ordenadas representa a probabilidade de acertar um item enquanto o eixo das abscissas indica o traço latente ou habilidade dos indivíduos. Assim, quanto maior o valor da habilidade, maior será sua proficiência, sendo que a probabilidade de resposta correta aumenta à medida que a habilidade do indivíduo aumenta. Desta forma, um indivíduo com habilidade superior a três tem probabilidade muito alta de acertar determinado item (acima de 80%), já um indivíduo com baixa habilidade (-3) tem probabilidade c (de 20%) de acertar determinado item. O acerto ao acaso pode ser observado a partir do ponto da curva mais próximo ao eixo das abscissas.

Uma curva com mais inclinação representa discriminação alta (mudança pequena na habilidade resulta em mudança acentuada na probabilidade de resposta

correta), enquanto uma curva menos inclinada indica um item com baixa discriminação.

O parâmetro θ pode assumir qualquer valor entre – ∞ e + ∞. Desta forma, precisa-se estabelecer uma origem e uma unidade de medida para a definição da escala. Esses valores são escolhidos de modo a representar, respectivamente, o valor médio (μ) e o desvio padrão (σ) dos desempenhos dos indivíduos da população em estudo. Usualmente, utiliza-se a escala com média igual a zero e desvio padrão igual a um [2].

Pode-se afirmar que um indivíduo com habilidade 1,3 na escala N(0,1) e que tem uma habilidade de 1,3 desvios padrão acima da média, apresenta um conhecimento consideravelmente maior do conteúdo em estudo do que um indivíduo com habilidade -0,3, ou seja, 0,3 desvios padrão abaixo da média dessa mesma população. No gráfico da CCI, espera-se que os valores dos parâmetros θ e *b* variem entre -3 e 3, isto é, entre três desvios padrão acima e abaixo da média.

É chamado de vetor de resposta ao item do indivíduo *j* a lista de pontuação gerada (com 1 para acerto e 0 para erro) para os *I* itens. Utiliza-se este vetor de respostas e os já conhecidos parâmetros dos itens a fim de estimar os parâmetros das habilidades θ_j [4].

Os métodos de máxima verossimilhança comumente utilizados na estimação das habilidades apresentam problemas com escores nulos e perfeitos (tanto para os itens quanto para os indivíduos), e também com padrões de resposta aberrantes (indivíduos com habilidades elevadas que respondem incorretamente itens fáceis, e vice-versa) [3]. Portanto, procura-se nos métodos de estimação bayesiana melhoria nas estimativas que apresentam estes problemas e, desta maneira, diminuir os vícios das estimativas que podem trazer problemas no processo de equalização, ou seja, quando se coloca todos os parâmetros dos itens na mesma escala métrica para efetuar comparações. A falta de resposta a algum item ou respostas com mais de uma alternativa, não são levadas em consideração na estimação dos parâmetros.

Em sua conceituação geral, a estimação bayesiana consiste em estabelecer distribuições a *priori* para os parâmetros desconhecidos e construir uma nova função denominada distribuição a *posteriori*, condicionada aos dados observados, e que representa a inferência para esses parâmetros [2]. Estimativas pontuais (por exemplo, a média, que é utilizada neste processo) são extraídas diretamente da

distribuição *posteriori* (para maiores detalhes sobre inferência bayesiana ver [6] e [9]).

Aplicação

Fez-se um estudo simulado efetuado em duas etapas: (1) geração de uma amostra simulada dos resultados de 50 testes com dez itens, para os quais os parâmetros *a, b* e *c* eram conhecidos, assim como eram conhecidas as habilidades (θ) dos 50 indivíduos; (2) estimação bayesiana das habilidades a partir dos resultados dos testes, utilizando o algoritmo MCMC. Cada uma das etapas foi efetuada conforme segue.

1) Geração dos dados simulados: inicialmente fixaram-se os valores dos parâmetros *a, b* e *c* para dez questões. Conforme a FIGURA 2,

		a		
		0.5	1	2
b	-1	1	1	1
	0	1	2	1
	1.5	1	1	1

Figura 2 Tabela de composição dos parâmetros dos itens conhecidos e da distribuição das questões

Observando essa composição mostrada na FIGURA 2, verifica-se que há questões de pouca, média e alta discriminação (a) e baixa, média e alta dificuldade (b) com c= 0.2.

Após, as habilidades de 50 indivíduos foram sorteados de uma distribuição normal, com média zero e desvio padrão um, N(0,1). Essas habilidades "reais" foram armazenadas para futura comparação conforme descrito mais abaixo. A distribuição das 50 habilidades está ilustrada no histograma das habilidades (FIGURA 3).

Para todo indivíduo *j*, os resultados dos testes foram simulados para a questão *i* conforme segue:

- Calcular $p_{ij}=P(U_{ij}=1|\theta_j)$ através do modelo ML3;
- Gerar valores 0 ou 1 de U_{ij} a partir de uma distribuição Bernoulli (p_{ij}).

Esses resultados, consistindo de 50 vetores de comprimento dez, constituídos de valores 1 (acerto) e 0 (erro) foram armazenados para servirem de base para a estimação das habilidades dos indivíduos.

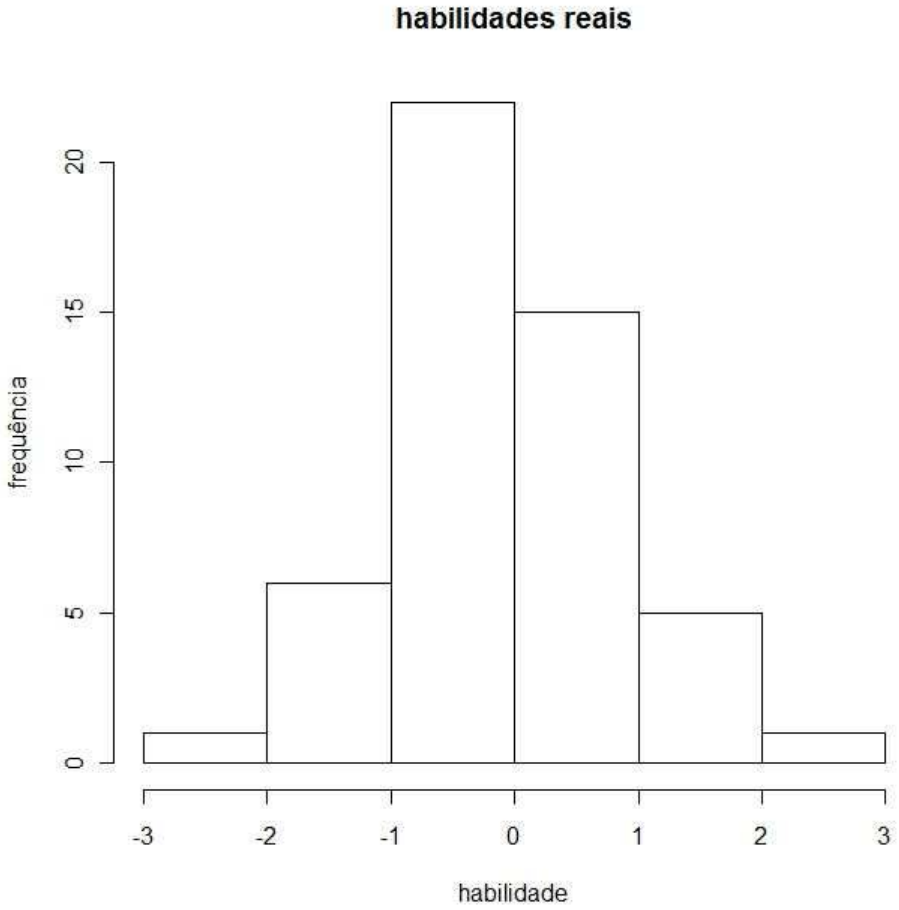


Figura 3 Histograma das habilidades reais dos indivíduos

2) Estimação das habilidades:

A partir dos resultados dos testes, estimou-se θ_j obtendo sua distribuição *posterior* através do MCMC, implementando o OpenBUGS via R e usando as bibliotecas R2WinBUGS [11] e BRUGS [12].

Quando é usado um algoritmo do tipo MCMC, os valores das iterações iniciais (*burn in*) são descartados por não ser considerada ainda uma amostra da distribuição estacionária da cadeia [8]. Para as estimativas bayesianas, adotou-se para o parâmetro θ, apenas uma das cadeias com 100000 iterações, num período de *burn in* de 50000 iterações e um salto *thin* igual a 5 entre iterações, para assegurar convergência e remover a influência dos valores iniciais, considerando

assim, uma amostra de tamanho 10000 da distribuição a *posteriori* (para maiores detalhes do método ver [6] e [9]).

RESULTADOS E DISCUSSÕES

Através dessa simulação, obtiveram-se os resultados dispostos na FIGURA 4, onde triângulos (Δ) representam os valores reais das habilidades e os círculos (○) representam os valores estimados das habilidades de cada indivíduo. Esses valores são os estimadores de Bayes, sob perda quadrática e denotam as médias das respectivas distribuições *posteriores* para as habilidades.

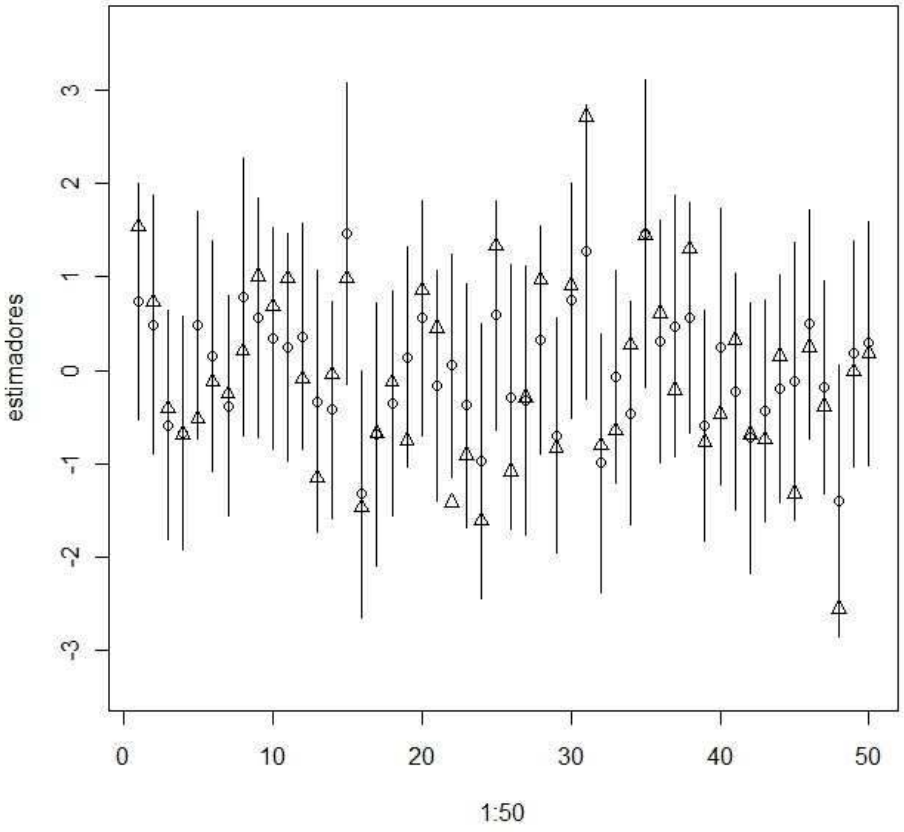


Figura 4 Resultado da simulação

Uma forma de avaliar a qualidade das estimativas é através dos intervalos de credibilidade. Este intervalo de credibilidade de 95%, também representado na FIGURA 4, cobre os valores das habilidades que totalizam os 95% mais prováveis a *posteriori* e constituem uma medida de incerteza associada à estimativa. Pode ser interpretado como a região dos 95% de valores mais prováveis para a habilidade θ_j.

Na particular simulação apresentada na FIGURA 4, o intervalo de credibilidade não conseguiu cobrir corretamente a habilidade em apenas um valor dentre os 50

indivíduos (indivíduo 22), que nesse caso, demonstrou habilidade inferior à estimativa. Houveram simulações, não mostradas aqui, com nenhum erro e com dois erros, ou seja, com estimativas fora do intervalo de 95%.

Para que se consiga fazer uma interpretação das habilidades de cada indivíduo, isto é, uma avaliação pedagógica do seu conhecimento, capacidades, etc, é preciso que sejam inseridos itens âncoras no teste. Estes itens são específicos para diagnosticar quais são essas habilidades desenvolvidas dos indivíduos sujeitos ao teste.

A partir da ferramenta de simulação descrita neste artigo, pretende-se futuramente efetuar estudos da influência da composição de itens na qualidade da inferência. O efeito do número de questões também pode ser explorado. Isto poderá servir para melhorar as estimativas das habilidades dos indivíduos.

CONCLUSÕES

A simulação de testes e a inferência bayesiana de habilidades é uma ferramenta útil no processo de estimação desses parâmetros, podendo ser utilizada para avaliar como os parâmetros *a, b* e *c*, influenciam na qualidade das estimativas.

Este novo método de avaliação, uma vez bem compreendido e aplicado, poderá auxiliar os professores na elaboração de instrumentos mais efetivos para avaliar as habilidades de seus alunos, buscando metodologias e estratégias adequadas para suprir as dificuldades apresentadas pelos alunos com diferentes habilidades.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] ALEXANDRE, J. W. C. et al.Uma proposta de análise de um construto para medição dos fatores críticos da gestão pela qualidade por intermédio da Teoria da Resposta ao Item. *Gestão &Produção* Online, São Carlos, v.9, n. 2, p.129-141, ago. 2002. ISSN 0104-530X. Disponível em: <http://www.scielo.br/scielo.php?pid=S0104-530X2002000200003&script=sci_arttext >. Acesso em: 20 set. 2009.

[2] ANDRADE, D.F.; TAVARES, H.R.; VALLE, R.C. *Teoria da Resposta ao Item:* Conceitos e Aplicações. Caxambú-MG: 14 SINAPE, 2000.

[3] AZEVEDO, C.L.N. *Métodos de estimação na teoria de resposta ao item*. Dissertação (Mestrado em Estatística) – Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2003.

[4] BAKER, F.B. *The Basics of Item Response Theory*. 2. ed. University of Wisconsin: ERIC, 2001.

[5] BERG, S. van den.; GLAS, C.A.W.; BOOMSMA, D.I. Variance Decomposition Using an IRT Measurement Model. *Journal Behavior Genetics* Online, v.37, n.4, p. 604-616, july 2007, ISSN 0001-8244. Disponível em: <http://www.springerlink.com/content/61h643426044214m/>. Acesso em: 20 set. 2009.

[6] GELMAN, A. et al. *Bayesian Data Analysis*. Texts in Statistical Science. 2. ed. Boca Raton: Chapman & Hall/CRC, 2004.

[7] KLEIN, R. Utilização da Teoria da Resposta ao Item no Sistema Nacional de Avaliação da Educação Básica (SAEB*). Revista Meta: Avaliação* Online, Rio de Janeiro, v.1, n. 2, p.125-140, mai.-ago. 2009. Disponível em: <http://metaavaliacao.cesgranrio.org.br/index.php/metaavaliacao/article/viewFile/38/17 >. Acesso em: 20 set. 2009

[8] MARQUES, K.A. Análise *bayesiana em modelos TRI de três parâmetros*. Dissertação (Mestrado em Ciência) – Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2008. Disponível em: <http://www.teses.usp.br/teses/disponiveis/45/45133/tde-02092008-214645/ >. Acesso em: 12 out. 2009.

[9] PAULINO, C.D.; TURKAN, M.A.A.; MURTEIRA, B. *Estatística Bayesiana.* Lisboa, 2003, *(*edição da Fundação Calouste Gulbenkian).

[10] R Development Core Team. R: A language and environment for statistical computing. *R Foundation for Statistical Computing* Online, Vienna, Austria, 2009, ISBN 3-900051-07-0. Disponível em: <http://www.R-project.org>. Acesso em: 5 jun. 2009.

[11] STURTZ, S.; LIGGES, U.; GELMAN, A. R2WinBUGS: A Package for Running WinBUGS from R. *Journal of Statistical Software*, v.12, n. 3, p. 1-16, 2005.

[12] THOMAS, A. et al. Making BUGS Open. *R News*, v.6, n.1, p. 12-17, 2006.