UM MODELO BAYESIANO PARA ESTIMAR O TAMANHO DE UMA POPULAÇÃO FECHADA

Karolina Barone Ribeiro da SILVA[1]
José Galvão LEITE[2]
Nelson Ithiro TANAKA[3]

- RESUMO: Utilizamos o processo de captura-recaptura, sob o enfoque bayesiano, para estimar o tamanho de uma população. Apresentamos o modelo estatístico supondo independência e homogeneidade entre os elementos da população e dependência e heterogeneidade entre as épocas de amostragem. Damos exemplos com dados simulados e reais.

- PALAVRAS-CHAVE: Processo de captura-recaptura; função de verossimilhança; distribuições *a priori* e *a posteriori*; distribuição de Dirichlet; estimativas de Bayes.

1 Introdução

Suponhamos uma população fechada de N elementos portadores de uma característica de interesse, onde N é desconhecido e é do nosso interesse estimá-lo. Por população fechada subentendemos uma população cujo tamanho, N, é constante durante o processo de estimação, que neste artigo será o de captura-recaptura sob o enfoque bayesiano. Este processo, utilizado originalmente para estimar tamanhos de populações animais, consiste, inicialmente, na seleção de um número fixado ou aleatório de elementos da população. Todos os elementos capturados são marcados, suas marcas anotadas e devolvidos à população. Em seguida, um número fixado ou aleatório de elementos é selecionado em uma ou mais épocas (ocasiões) de amostragem. Em cada uma dessas épocas, todos os elementos marcados (recapturados) e não marcados (capturados pela primeira vez) selecionados são marcados, suas marcas anotadas e devolvidos à população. Desse modo, no final do processo teremos anotado o histórico de captura ou a seqüência de captura, recapturas e não recapturas de cada elemento selecionado, pelo menos uma vez da população durante as várias épocas, e baseado nos números de elementos que apresentam os possíveis históricos de captura poderemos estimar N.

Ampla é a literatura sobre as aplicações do processo de captura-recaptura na estimação de tamanhos de populações animais. Elas se iniciaram em 1896, quando Petersen fez um estudo sobre o fluxo migratório de peixes no mar Báltico embora, em

[1] Departamento de Matemática, Universidade Estadual do Centro-Oeste – UNICENTRO, CEP: 85015-230, Guarapuava, PR, Brasil. Email: *karolinabarone@yahoo.com.br*
[2] Departamento de Estatística, Centro de Ciências Exatas e Tecnologia, Universidade Federal de São Carlos – UFSCar, Caixa Postal 676, CEP: 13565-905, São Carlos. SP, Brasil. Email: *leite@power.ufscar.br*
[3] Departamento de Estatística, Instituto de Matemática e Estatística, Universidade de São Paulo – IME/USP, CEP: 05508-090, São Paulo, SP, Brasil. Email: *nitanaka@ime.usp.br*

1783, Laplace já houvesse utilizado o processo para estimar o tamanho da população da França. Em 1930 Lincoln aplicou o processo para estimar o tamanho de uma população de patos selvagens e, a partir da década de cinqüenta, diversos pesquisadores tais como, Chapman (1954), Darroch (1959), Jolly (1965), Burnham e Overton (1978), Seber (1986) e Pollock (1991) publicaram trabalhos sobre o assunto. Por outro lado, pesquisadores como Hunter e Griffiths (1978), Castledine (1981), George e Robert (1992) e Yoshida, Leite e Bolfarine (1999), por exemplo, também publicaram artigos sob o enfoque bayesiano nesta área. Com relação às aplicações desse processo em outras áreas do conhecimento destacamos, por exemplo, os artigos de Nayak (1988) e Basu e Ebrahimi (2001) na área de controle de erros de *softwares* e os de Lee, Seber, Holden e Huakau (2001) e Lee (2002) na área de controle de doenças não transmissíveis em populações.

Introduzimos o modelo estatístico supondo que cada elemento é capturado (recapturado) ou não, em qualquer época, independentemente dos demais elementos da população e dado qualquer histórico de captura, todos os elementos têm a mesma probabilidade de apresentá-lo ao longo da realização do processo, isto é, os elementos da população são independentes e homogêneos entre si. Vamos supor também que as probabilidades de que um elemento da população apresente os possíveis históricos de captura sejam diferentes e que haja uma interdependência entre as épocas de amostragem, ou equivalentemente, vamos supor que as épocas de captura sejam dependentes e heterogêneas entre si. Na seção 2 apresentamos o modelo estatístico; na seção 3 introduzimos o modelo bayesiano; na seção 4 analisamos o desempenho do modelo através de dois exemplos com dados simulados; na seção 5 damos dois exemplos com dados reais e no final apresentamos as conclusões.

2 Modelo estatístico

Denotemos por N (N desconhecido) o tamanho de uma população fechada, cujos elementos possuem uma característica de interesse. Para estimar N adotamos o processo de captura-recaptura, sob a ótica bayesiana. Seja k ($k \geq 2$ e conhecido) o número de épocas ou ocasiões de amostragem. Suponhamos que cada elemento da população, em qualquer época, seja capturado (recapturado) ou não, independentemente dos demais. Dado N associemos o vetor aleatório k-dimensional $\mathbf{X}_i = (X_{i1}, X_{i2},..., X_{ik})$ ao elemento i, onde X_{ij} assume valor 1 se o elemento i for capturado ou recapturado na j-ésima época e 0 caso contrário, $i = 1, 2,..., N$ e $j = 1, 2,..., k$. Então, os vetores aleatórios $\mathbf{X}_i$ são independentes e assumem valores no conjunto $\Delta = \{\mathbf{w} = (w_1, w_2,..., w_k): w_j = 0, 1; j = 1, 2,..., k\}$, onde # ($\Delta$) = cardinal de $\Delta = l = 2^k$. Seja $\mathbf{w}_1, \mathbf{w}_2,..., \mathbf{w}_l$, uma enumeração dos elementos de Δ, onde $\mathbf{w}_l = (0, 0, 0,..., 0)$. Notemos que $\mathbf{w}_1, \mathbf{w}_2,..., \mathbf{w}_l$, representam os possíveis históricos de captura apresentados por qualquer um dos elementos da população, quando da realização do processo, e que $\mathbf{w}_l$ representa o único histórico de captura não observável.

Então, supondo que todos os elementos têm a mesma probabilidade de apresentar um dado histórico de captura ao longo do processo, segue que a probabilidade do elemento i apresentar o histórico $\mathbf{w}_r$ é

$$p_r = P(\mathbf{X}_i = \mathbf{w}_r) = P((X_{i1}, X_{i2}, ..., X_{ik}) = (w_{r1}, w_{r2}, ..., w_{rk})) =$$
$$= P(X_{i1} = w_{r1}, X_{i2} = w_{r2}, ..., X_{ik} = w_{rk}),$$

$i = 1, 2,..., N$; $r = 1, 2,..., l$.

Sejam $n_r = \sum_{i=1}^{N} I_{\{\mathbf{w}_r\}}(\mathbf{X}_i)$ o número de elementos com o histórico de captura $\mathbf{w}_r$, $r = 1, 2,..., l$ e $n = \sum_{r=1}^{l-1} n_r$ o número de elementos distintos observados durante a realização do processo. Notemos que $n_l = N - n$ é o número de elementos que apresentam o histórico $\mathbf{w}_l$, isto é, o número de elementos não observados.

Então, a distribuição de probabilidades de $(n_1, n_2,..., n_{l-1}, N - n)$, dados N e o vetor de probabilidades $\mathbf{p} = (p_1, p_2,..., p_l)$, é dada por

$$P(n_1, n_2, \ldots n_{l-1}, N-n \mid N, \mathbf{p}) = \frac{N!}{n_1! n_2! \ldots n_{l-1}! (N-n)!} \prod_{r=1}^{l} p_r^{\,n_r}, \tag{1}$$

$N \geq n$, $0 < p_r < 1$, $r = 1, 2,..., l$.

De (1) segue que a função de verossimilhança é tal que

$$L(N, \mathbf{p} \mid n_1, n_2, \ldots n_{l-1}) = P(n_1, n_2, \ldots n_{l-1}, N-n \mid N, \mathbf{p}) = \frac{N!}{n_1! n_2! \ldots n_{l-1}! (N-n)!} \prod_{r=1}^{l} p_r^{\,n_r} \propto \frac{N!}{(N-n)!} \prod_{r=1}^{l} p_r^{\,n_r},$$

$N \geq n$, $\mathbf{p} = (p_1, p_2,..., p_l)$, $0 < p_r < 1$, $r = 1, 2,..., l$.

Na próxima seção definimos o modelo bayesiano e as distribuições *a priori* a serem atribuídas a N e $\mathbf{p}$.

3 Modelo Bayesiano

Suponhamos *a priori* que $(p_1, p_2, \ldots, p_{l-1})$ tenha distribuição de Dirichlet com vetor de parâmetros $\boldsymbol{\alpha} = (\alpha_1, \alpha_2,..., \alpha_l)$, $\alpha_i > 0$, $i = 1, 2,..., l$, e N tenha função de probabilidades $\pi(N)$, $N = 1, 2,...$. Então, a distribuição de probabilidades *a priori* de $(p_1, p_2, \ldots, p_{l-1})$ é dada por

$$\pi(p_1, p_2, \ldots, p_{l-1} \mid \boldsymbol{\alpha}) = \Gamma\left(\sum_{i=1}^{l} \alpha_i\right) \prod_{i=1}^{l} \frac{p_i^{\,\alpha_i - 1}}{\Gamma(\alpha_i)}$$

$0 < p_i < 1, i = 1, 2, \ldots, l-1, \sum_{i=1}^{l-1} p_i < 1, \; p_l = 1 - \sum_{i=1}^{l-1} p_i$ e, supondo N e $(p_1, p_2,..., p_{l-1})$ independentes, segue que a distribuição *a priori* conjunta de N e $(p_1, p_2,..., p_{l-1})$ é tal que

$$\pi(N,(p_1,p_2,\ldots,p_{l-1})\mid \boldsymbol{\alpha}) = \pi(N)\pi((p_1,p_2,\ldots,p_{l-1})\mid \boldsymbol{\alpha})$$
$$\propto \pi(N)\prod_{i=1}^{l} p_i^{\alpha_i-1},$$

$$N=1,2,\ldots,0<p_i<1, i=1,2,\ldots,l-1, \sum_{i=1}^{l-1} p_i<1,\ p_l=1-\sum_{i=1}^{l-1} p_i\ .$$

Logo, a distribuição *a posteriori* conjunta de N e $(p_1, p_2,..., p_{l-1})$ é tal que

$$\pi(N,(p_1,p_2,\ldots,p_{l-1})\mid \boldsymbol{\alpha}, n_1,n_2,\ldots n_{l-1})$$
$$\propto L(N,\mathbf{p}\mid \boldsymbol{\alpha}, n_1,n_2,\ldots n_{l-1})\pi(N,(p_1,p_2,\ldots,p_{l-1})\mid \boldsymbol{\alpha})$$
$$\propto \frac{N!}{(N-n)!}\pi(N)\prod_{r=1}^{l} p_r^{\,n_r+\alpha_r-1} \qquad (2)$$
$$\propto \frac{N!}{(N-n)!}\pi(N)\prod_{r=1}^{l-1} p_r^{\,n_r+\alpha_r-1}\, p_l^{\,N-n+\alpha_l-1},$$

$$N\geq n,\ 0<p_i<1, i=1,2,\ldots,l-1, \sum_{i=1}^{l-1} p_i<1,\ p_l=1-\sum_{i=1}^{l-1} p_i\ .$$

A prova do seguinte teorema é dada no Apêndice I.

Teorema: Suponhamos que a função de probabilidades *a priori* de N seja da forma $\pi(N) = 1/N^u$, $N = 1, 2,...; u = 0, 1$. Então,

a) para $u = 0$, a distribuição *a posteriori* conjunta de N e $(p_1, p_2,..., p_{l-1})$ existe se, e somente se, $\sum_{r=1}^{l-1}\alpha_r > 1$,

b) para $u = 1$, a distribuição *a posteriori* conjunta de N e $(p_1, p_2,..., p_{l-1})$ existe. ■

Suponhamos agora válida a condição $\sum_{r=1}^{l-1}\alpha_r > 1$ do teorema. Para u = 0 temos, de (2), que a distribuição *a posteriori* conjunta de N e $(p_1, p_2,..., p_{l-1})$ é tal que

$$\pi(N,(p_1,p_2,\ldots,p_{l-1})\mid \boldsymbol{\alpha}, n_1,n_2,\ldots n_{l-1})$$
$$\propto \frac{N!}{(N-n)!}\prod_{r=1}^{l-1} p_r^{\,n_r+\alpha_r-1}\, p_l^{\,N-n+\alpha_l-1},$$

Seja $A=\{(p_1,p_2,\ldots,p_{l-1}): 0<p_i<1,\ i=1,2,\ldots,l-1 \text{ e } \sum_{i=1}^{l-1} p_i<1\}$. Segue que a função de probabilidades *a posteriori* marginal de N é tal que

$$
\begin{aligned}
&\pi(N \mid \boldsymbol{\alpha}, n_1, n_2, \ldots n_{l-1}) \\
&= \int_A \pi(N, (p_1, p_2, \ldots, p_{l-1}) \mid \boldsymbol{\alpha}, n_1, n_2, \ldots n_{l-1})\, dp_1\, dp_2 \cdots dp_{l-1} \\
&\propto \frac{N!}{(N-n)!} \int_A \prod_{r=1}^{l} {p_r}^{n_r+\alpha_r-1}\, dp_1\, dp_2 \cdots dp_{l-1} \\
&= \frac{N!}{(N-n)!} \frac{\prod_{r=1}^{l} \Gamma(\alpha_r + n_r)}{\Gamma\left[\sum_{r=1}^{l} (\alpha_r + n_r)\right]} \\
&\propto \frac{N!}{(N-n)!} \frac{\Gamma(\alpha_l + N - n)}{\Gamma\left(\sum_{r=1}^{l} \alpha_r + N\right)},
\end{aligned}
\tag{3}
$$

$N \geq n$. Para $u = 1$, segue de (2) que a distribuição *a posteriori* conjunta de N e $(p_1, p_2, ..., p_{l-1})$ é tal que

$$
\begin{aligned}
&\pi(N, (p_1, p_2, \ldots, p_{l-1}) \mid \boldsymbol{\alpha}, n_1, n_2, \ldots n_{l-1}) \\
&\propto \frac{(N-1)!}{(N-n)!} \prod_{r=1}^{l-1} {p_r}^{n_r+\alpha_r-1} {p_l}^{N-n+\alpha_l-1},
\end{aligned}
$$

ou seja, a função de probabilidades *a posteriori* marginal de N é tal que

$$
\begin{aligned}
&\pi(N \mid \boldsymbol{\alpha}, n_1, n_2, \ldots n_{l-1}) \\
&= \int_A \pi(N, (p_1, p_2, \ldots, p_{l-1}) \mid \boldsymbol{\alpha}, n_1, n_2, \ldots n_{l-1})\, dp_1\, dp_2 \cdots dp_{l-1} \\
&\propto \frac{(N-1)!}{(N-n)!} \int_A \prod_{r=1}^{l} {p_r}^{n_r+\alpha_r-1}\, dp_1\, dp_2 \cdots dp_{l-1} \\
&= \frac{(N-1)!}{(N-n)!} \frac{\prod_{r=1}^{l} \Gamma(\alpha_r + n_r)}{\Gamma\left[\sum_{r=1}^{l} (\alpha_r + n_r)\right]} \\
&\propto \frac{(N-1)!}{(N-n)!} \frac{\Gamma(\alpha_l + N - n)}{\Gamma\left(\sum_{r=1}^{l} \alpha_r + N\right)},
\end{aligned}
\tag{4}
$$

N ≥ n.

4 Estudo do desempenho do modelo Bayesiano

Nesta seção analisamos o desempenho do modelo bayesiano descrito na seção 3, através de exemplos com dados simulados. Para determinar os valores aproximados das funções de probabilidades *a posteriori* de *N*, dadas pelas expressões (3) e (4), cálculos foram feitos de acordo com o seguinte raciocínio. Associada a cada uma dessas funções existe uma função real não negativa *f* tal que

$$\pi(N \mid \boldsymbol{\alpha}, n_1, n_2, \ldots, n_{l-1}) \propto f(N),$$

$N \geq n$. A partir da razão

$$A(N) = \frac{f(N+1)}{f(N)},$$

$N \geq n$, determinamos recursivamente as relações

$$\begin{aligned}
&\pi(n \mid \boldsymbol{\alpha}, n_1, n_2, \ldots, n_{l-1}) \propto f(n),\\
&\pi(n+1 \mid \boldsymbol{\alpha}, n_1, n_2, \ldots, n_{l-1}) \propto f(n+1) = A(n) f(n),\\
&\pi(n+2 \mid \boldsymbol{\alpha}, n_1, n_2, \ldots, n_{l-1}) \propto f(n+2) = A(n+1) f(n+1),\\
&\ldots\\
&\pi(n+s \mid \boldsymbol{\alpha}, n_1, n_2, \ldots, n_{l-1}) \propto f(n+s) = A(n+s-1) f(n+s-1),
\end{aligned}$$

onde s é um número inteiro positivo tal que $\pi(n+s+1 \mid \boldsymbol{\alpha}, n_1, n_2, \ldots, n_{l-1}) \cong 0$. Logo,

$$\pi(N \mid \boldsymbol{\alpha}, n_1, n_2, \ldots n_{l-1}) \cong \frac{f(N)}{\sum_{i=0}^{s} f(n+i)},$$

$n \leq N \leq n+s$. A média e o desvio padrão *a posteriori* de *N* são dados, respectivamente, por

$$M = \sum_{i=0}^{s} (n+i)\pi(n+i \mid \boldsymbol{\alpha}, n_1, n_2, \ldots n_{l-1})$$

e

$$D.P. = \sqrt{\sum_{i=0}^{s} \left[(n+i-M)^2 \pi(n+i \mid \boldsymbol{\alpha}, n_1, n_2, \ldots n_{l-1}) \right]}.$$

Esse método foi implementado no *software* MAPLE (versão 7.00) e no programa utilizado são determinados os demais resumos aproximados *a posteriori* de *N*, como

moda, intervalo de credibilidade e os quartis. A seguir apresentamos os exemplos 1 e 2 com dados simulados através do *software* R (versão 1.9.0).

Exemplo 1: Neste exemplo atribuímos a N o valor 100 e a k os valores 2 e 5. Geramos $(p_1, p_2,..., p_{l-1})$ a partir da distribuição de Dirichlet de parâmetros $\alpha_i = \alpha = 10$ e 5, $i = 1, 2,..., l$, e em seguida geramos as estatísticas $n_1, n_2,..., n_{l-1}$, de acordo com trajetórias específicas $\mathbf{w}_1, \mathbf{w}_2,..., \mathbf{w}_{l-1}$. Consideramos para N a distribuição *a priori* de Jeffreys, ou seja, $\pi(N) = 1/N$, $N = 1, 2,...$, e para diferentes valores de α aplicamos o procedimento descrito acima para determinar valores aproximados da função de probabilidades *a posteriori* (4) e seus resumos aproximados, como média (M), moda, quartis (Q_j, $j = 1, 2, 3$), desvio padrão (*D.P.*), intervalo de credibilidade de 95% (*I.C.* (95%)) e amplitude do intervalo de credibilidade (*Ampl. I. C.)*.

i) Considerando $k = 2$ e $\alpha = 10$ na geração dos dados, obtivemos $n = 67$. Os resultados obtidos seguem na Tabela 1.

Tabela 1 - Resumos aproximados da função de probabilidades *a posteriori* de N

α	M	*Moda*	Q_1	Q_2	Q_3	*D.P.*	*I.C. (95%)*	*Ampl.I.C.*
0,5	89,58	67	68	76	98	35,4	(67; 175)	108
1	92,59	67	72	82	102	27,05	(67; 170)	103
2	92,65	76	76	85	100	21,98	(68; 153)	85
5	90,91	83	80	87	97	14,15	(71; 125)	54
10	90,1	86	82	88	95	10,32	(74; 113)	39
20	89,71	87	83	88	94	8,12	(75; 107)	32
50	89,48	88	84	88	93	6,62	(77; 103)	26

De acordo com a Tabela 1, verificamos que a média é uma estimativa razoável para N em todos os casos e que o verdadeiro valor do parâmetro pertence a todos os intervalos de credibilidade obtidos. Além disso, a escolha do hiperparâmetro α influencia as estimavas de N.

ii) Considerando $k = 5$ e $\alpha = 5$ na geração dos dados, obtivemos $n = 98$. Os resultados obtidos seguem na Tabela 2.

Tabela 2 - Resumos aproximados da função de probabilidades *a posteriori* de N

α	M	*Moda*	Q_1	Q_2	Q_3	*D.P.*	*I.C. (95%)*	*Ampl.I.C.*
0,5	101,37	98	98	99	101	5,39	(98; 116)	18
1	101,26	98	98	99	102	3,85	(98; 111)	13
2	101,21	99	98	99	102	2,96	(98; 108)	10
5	101,18	100	98	100	101	2,32	(98; 106)	8
10	101,17	100	99	100	101	2,08	(98; 105)	7
20	101,16	100	99	100	101	1,94	(98; 105)	7
50	101,16	101	99	100	101	1,86	(98; 104)	6

De acordo com a Tabela 2, constatamos que são boas as estimativas bayesianas de *N*, independentemente do valor atribuído a α, e também podemos observar que o verdadeiro valor do parâmetro pertence a todos os intervalos de credibilidade obtidos. Notamos também que as amplitudes dos intervalos são bem menores do que aquelas obtidas no item i), quando consideramos apenas duas épocas de amostragem. É importante observar que, ao alterarmos o número de duas para cinco épocas de amostragem, houve uma melhora substancial nas estimativas obtidas. Apresentamos em seguida, Figuras 1 e 2, os gráficos das funções de probabilidades à posteriori aproximadas de N, para um valor fixo de α. O formato do gráfico se estreita à medida que α cresce.

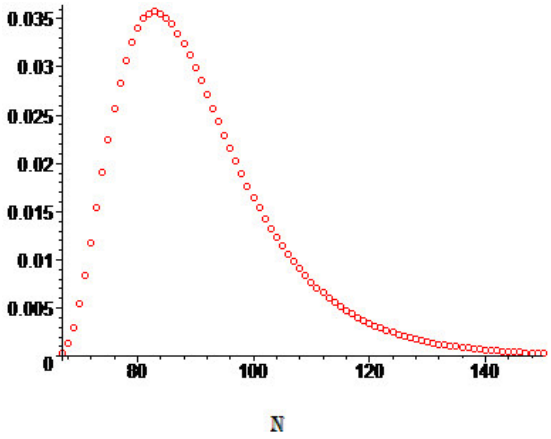

Figura 1 - Função de probabilidades marginal *a posteriori* aproximada de *N* (k = 2 e $\alpha = 5$).

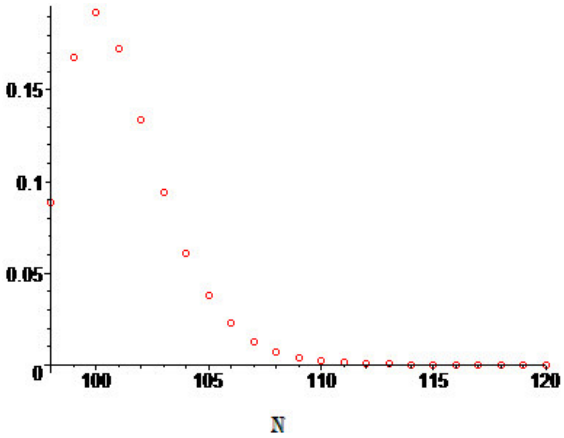

Figura 2 - Função de probabilidades marginal *a posteriori* aproximada de *N* (k = 5 e $\alpha = 5$).

Exemplo 2: Neste exemplo atribuímos aos parâmetros do modelo os mesmos valores do exemplo 1, bem como utilizamos as mesmas estatísticas. Adotamos para N a distribuição *a priori* uniforme nos inteiros não negativos, isto é, $\pi(N) = 1$, $N = 1, 2,...$, e aplicamos o procedimento anterior para a função de probabilidades *a posteriori* de, *N,* (3).

i) Considerando $k = 2$ e $n = 67$, obtivemos os resultados da Tabela 3.

Tabela 3 - Resumos aproximados da função de probabilidades *a posteriori* de *N*

α	*M*	*Moda*	Q_1	Q_2	Q_3	*D.P.*	*I.C. (95%)*	*Ampl.I.C.*
0,5	191,42	67	74	106	217	189,64	(67; 796)	729
1	126,24	67	76	94	133	94,13	(67; 396)	329
2	100,98	78	78	90	110	34,79	(68; 190)	122
5	93,12	84	81	89	100	15,48	(72; 131)	59
10	91,29	87	83	89	96	10,82	(74; 116)	42
20	90,45	88	83	89	94	8,34	(76; 108)	32
50	89,98	88	84	88	93	6,72	(77; 104)	27

De acordo com a Tabela 3, verificamos que as estimativas obtidas para N são próximas daquelas obtidas na Tabela 1 quando consideramos $\alpha = 5$, 10, 20 e 50.

ii) Considerando $k = 5$ e $n = 98$, obtivemos os resultados da Tabela 4.

Tabela 4 - Resumos aproximados da função de probabilidades *a posteriori* de *N*

α	*M*	*Moda*	Q_1	Q_2	Q_3	*D.P.*	*I.C. (95%)*	*Ampl.I.C.*
0,5	101,66	98	98	99	102	5,83	(98; 117)	19
1	101,41	98	98	99	102	4,01	(98; 111)	13
2	101,29	99	98	100	102	3,03	(98; 108)	10
5	101,23	100	98	100	102	2,35	(98; 106)	8
10	101,21	100	99	100	101	2,09	(98; 105)	7
20	101,2	101	99	100	101	1,96	(98; 105)	7
50	101,19	101	99	100	101	1,87	(98; 104)	6

De acordo com a Tabela 4, verificamos que as estimativas obtidas para N são próximas daquelas obtidas na Tabela 2. Na próxima seção apresentamos dois exemplos com dados reais. Apresentamos, nas Figuras 3 e 4, as funções de probabilidades à posteriori aproximadas de N, agora considerando o maior valor de α usado nas simulações.

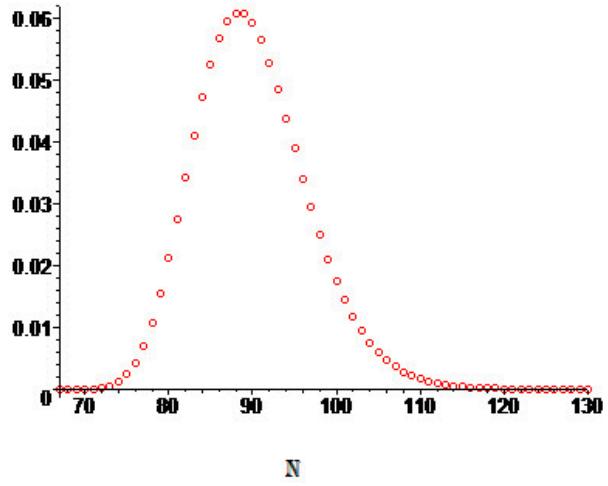

Figura 3 - Função de probabilidades marginal *a posteriori* aproximada de N (k = 2 e $\alpha = 50$).

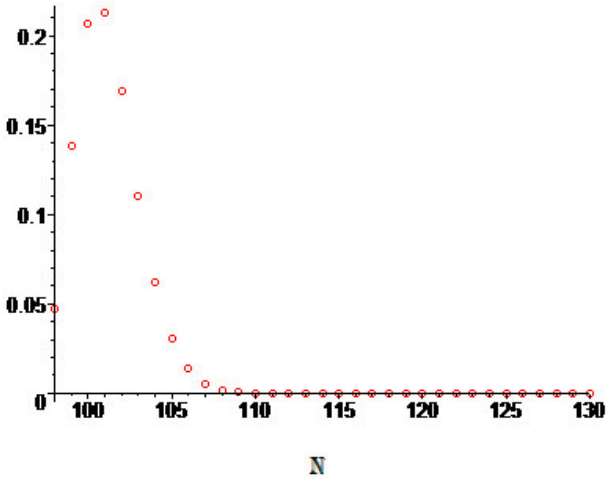

Figura 4 - Função de probabilidades marginal *a posteriori* aproximada de N (k = 5 e $\alpha = 50$).

5 Aplicações

Exemplo 3: Neste exemplo Hook *et. al* (1980) fizeram um estudo da incidência do defeito "spina bifida" em nativos do Estado de Nova York-EUA no período de 1969 a 1974. Listas foram montadas a partir de três fontes: certidão de nascimento (N), certidão de óbito (O) e prontuário médico (M). As listas foram cruzadas para determinar as intersecções de nomes. No caso, 12 nomes apareceram nas três listas e 142 apareceram nas listas N e O mas não na M. No total 626 indivíduos foram identificados nas três listas

de um total de 863.143 nascimentos vivos. A Tabela 5 apresenta os dados com uma das celas tendo dado faltante.

Tabela 5 - Dados de Hook et. al. (1980)

		M = Não	M = Sim
N = Não	O = Não	?	60
N = Não	O = Sim	49	4
N = Sim	O = Não	247	112
N = Sim	O = Sim	142	12

O dado faltante na primeira cela da Tabela 5 é igual ao número de elementos da população de portadores do defeito "spina bifida" não observados nas três listas, isto é, ele é igual a diferença entre o tamanho da população e o número de elementos distintos observados nas listas. O objetivo é estimar este dado ou o tamanho da população. Madigan e York (1999) usando esses dados aplicaram modelos gráficos de decomposição (ver Darroch, Lauritzen and Speed, 1980) e modelos log-lineares hierárquicos como em West (1985). As prioris consideradas foram a de Jeffreys e Rissanen (1983) para N e distribuições hiper-Dirichlet (ver Dawid and Lauritzen, 1993) para **p.** Suas estimativas para o tamanho da população, assumida fechada, são apresentadas na Tabela 6.

Tabela 6 - Estimativas em Madigan e York (1999)

Modelo	Média	Moda	*I.C. (95%)*
[N][OM]	731	729	(701, 767)
[NOM]	756	752	(714, 811)
[NMO]	712	709	(681, 751)
[NO][NM][OM]	697	626	(628, 934)

Neste exemplo, fazendo analogia com o processo de captura-recaptura, as três fontes de informação correspondem a três épocas de amostragem.

Apresentamos na Tabela 7 os resumos aproximados da função de probabilidades *a posteriori* do número de casos de incidência do defeito "spina bifida", *N*, segundo o modelo apresentado (aplicação do procedimento da seção 4), onde atribuímos a distribuição *a priori* uniforme para *N* e para α o valor 1, o que corresponde à distribuição de Dirichlet não informativa para $(p_1, p_2, \ldots, p_7)$.

Tabela 7 - Resumos aproximados da função de probabilidades *a posteriori* de *N*

α	*M*	*Moda*	Q_1	Q_2	Q_3	*D.P.*	*I.C. (95%)*	*Ampl.I.C.*
1	713,28	626	650	686	751	81,46	(627; 932)	305

Excetuando o intervalo de credibilidade, as estimativas bayesianas de *N* obtidas são próximas daquelas obtidas por Madigan e York (1999) segundo o modelo [NMO].

Exemplo 4:. Os dados reais considerados neste exemplo são os apresentados em Cormack (1985, 1989). Tais dados referem-se às capturas e recapturas de uma espécie de coelho (*snowshoe-hare*), que habita a região norte da América do Norte, durante 6 dias consecutivos. As freqüências dos históricos de captura observados são apresentadas na Tabela 8.

Tabela 8 - Dados em Cormack (1985, 1989)

Capturas 6, 5, 4	Capturas 3, 2, 1							
	000	001	010	011	100	101	110	111
000	?	3	6	0	5	1	0	0
001	3	2	3	0	0	1	0	0
010	4	2	3	1	0	1	0	0
011	1	0	0	0	0	0	0	0
100	4	1	1	1	2	0	2	0
101	4	0	3	0	1	0	2	0
110	2	0	1	0	1	0	1	0
111	1	1	1	0	0	0	1	2

Na Tabela anterior Captura i = 1 significa animal capturado (recapturado) na época i e Captura i = 0 significa animal não capturado (não recapturado) na época i, i = 1, 2, ..., 6. Analogamente ao exemplo anterior, o dado faltante na primeira cela da Tabela 8 é igual à diferença entre o número de elementos da população de coelhos e o número de coelhos distintos observados durante o processo. O objetivo é estimar este dado ou o tamanho populacional.

Coull e Agresti (1999) fizeram uma análise estatística clássica desses dados usando quatro modelos. As estimativas obtidas para o tamanho da população são apresentadas na Tabela 9.

Tabela 9 - Estimativas em Coull e Agresti (1999)

Modelo	E.M.V	Intervalo de Confiança (95%)
Logístico-Normal	92,0	(74,8; 153,6)
Associação de dois fatores homogêneos	90,5	(74,8: 125,1)
Classe Latente	77,1	(70,8; 87,4)
Independência Mútua	75,1	(69,9; 83,3)

E.M.V: estimativa de máxima verossimilhança de N.

Neste exemplo temos 6 épocas de amostragem e 68 coelhos distintos observados. Apresentamos na Tabela 10 os resumos aproximados da função de probabilidades *a posteriori* do número de coelhos, *N*, segundo o modelo apresentado (aplicação do procedimento da seção 4), onde atribuímos a distribuição *a priori* uniforme para *N* e para α o valor 1, o que corresponde à distribuição de Dirichlet não informativa para $(p_1, p_2,..., p_{63})$.

Tabela 10 - Resumos aproximados da função de probabilidades *a posteriori* de *N*

α	*M*	*Moda*	Q_1	Q_2	Q_3	*D.P.*	*I.C. (95%)*	*Ampl. I.C.*
1	70,1	68	68	69	72	1,54	(68; 72)	4

As estimativas bayesianas de *N* obtidas são razoavelmente próximas daquelas obtidas por Coull e Agresti (1999) segundo o modelo de independência mútua.

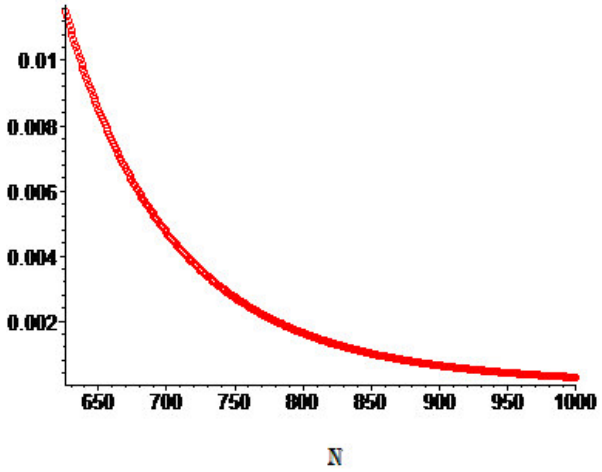

Figura 5 – Função de probabilidades marginal *a posteriori* aproximada de *N* (exemplo 3).

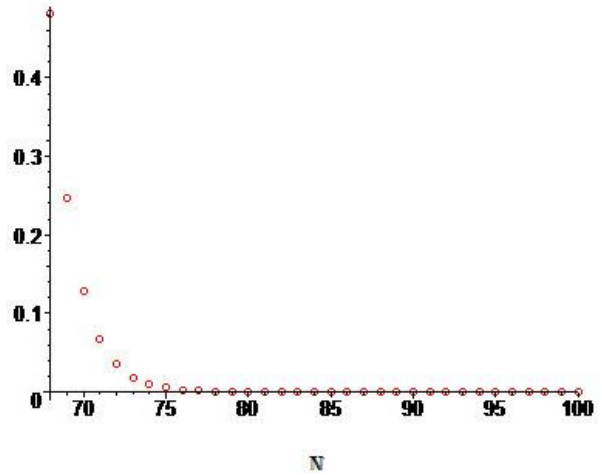

Figura 5 - Função de probabilidades marginal *a posteriori* aproximada de *N* (exemplo 4).

Conclusões

Pelos exemplos com dados simulados 1 e 2 concluímos que, para $k = 2$ e $\alpha = 5$, 10, 20 e 50, a adoção da distribuição *a priori* de Jeffreys para N é equivalente à adoção da priori uniforme nos inteiros não negativos. Para $k = 5$ essa equivalência ocorre com relação a todos os valores de α considerados, ou seja, não existem diferenças significativas nos resumos aproximados da função de probabilidades *a posteriori* de N, quando adotamos tais prioris. Além disso, a escolha do valor do hiperparâmetro α da distribuição *a priori* para $(p_1, p_2,..., p_{l-1})$ influencia os resumos aproximados da função de probabilidades *a posteriori* de N, independentemente de sua distribuição *a priori* ser a de Jeffreys ou a uniforme nos inteiros não negativos.

Agradecimentos

O segundo autor agradece o CNPq pela bolsa do programa Produtividade em Pesquisa – Processo 302813/2004-7. O terceiro autor agradece a FAPESP pelo apoio parcial através do Projeto Temático – Processo 2004/ 015304-6.

SILVA, K. B. R., LEITE, J. G. TANAKA, N. I. A Bayesian model for estimating the size of a closed population. *Rev. Bras. Biom.*, São Paulo, v.25, n.4, p.135-156, 2007.

- *ABSTRACT: We used the capture-recapture process, under Bayesian approach, for estimating the size of a closed population. We present the statistical model considering independence and homogeneity among individuals and dependence and heterogeneity among sampling times. We give examples with simulated and real data.*

- *KEYWORDS: Capture-recapture process; likelihood function; prior and posterior distributions; Dirichlet distribution; Bayes estimates.*

Referências

BASU, S.; EBRAHIMI, N. Bayesian capture-recapture methods for error detection and estimation of population size: heterogeneity and dependence. *Biometrika*, London, v.88, n.1, p.269-279, 2001.

BURHAM, K. P.; OVERTON, W. S. Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika*, London, v.65, n.3, p.625-633, 1978.

CASTLEDINE, B. A. Bayesian analysis of multiple recapture sampling from a closed population. *Biometrika*, London, v.67, n.1, p.197-210, 1978.

CHAPMAN, D. G. The estimation of biological populations. *Ann. Math. Statist.*, Ann Harbor, v.25, n.1, p.1-25, 1954.

CORMACK, R.M. Examples of the use of GLIM to analyze capture-recapture studies. In: MORGAN, B. J. T.; NORTH, P. M. (Eds.). *Statistics in ornithology.* New York: Springer, 1985. p.243-273.

CORMACK, R. M. Log-linear models for capture-recapture. *Biometrics*, Washington, v.45, n.2, p.395-413, 1989.

COULL, B. A.; AGRESTI, A. The use of mixed logit models to reflect heterogeneity in capture-recapture studies. *Biometrics*, Washington, v.55, n.1, p.294-301, 1999.

DARROCH, J. N. The multiple recapture census: estimation when there is immigration or death. *Biometrika*. London, v.46, n.3-4, p.336-351, 1959.

DARROCH, J. N.; LAURITZEN, S. L. and SPEED, T. P. Markov fields and log-linear interaction models for contingency tables. *Ann. Stat.*, Hayward, v.8, n.3, p.522-539, 1980.

DAWID, A. P.; LAURITZEN, S. L. Hyper Markov laws in the statistical analysis of decomposable graphical models. *Ann. Stat.*, Hayward, v.21, n.3, p.1272-1317, 1993.

FELLER, W. *An introduction to the theory of probability and its applications*. 3.ed. New York: John Wiley and Sons, 1967. v.1.

GEORGE, E. I.; ROBERT, C. P. Capture-recapture estimation via Gibbs Sampling. *Biometrika*, London, v.79, n.4, p.677-683, 1992.

HOOK, E. B.; ALBRIGHT, S. G.; CROSS, P. K. Use of Bernoulli census and log-linear methods for estimating the prevalence of spina bifida in live births and the completeness of vital records in New York State, *Am. J. Epidemiol.*, Baltimore, v.112, n.6, p.750-758, 1980.

HUNTER, A. J.; GRIFFITHS, H. J. Bayesian approach to estimation of insect population size. *Technometrics*, London, v. 3, n.20, p.231-234, 1978.

JOLLY, G. M. Explicit estimates from capture-recapture data with both death and immigration – stochastic model. *Biometrika*, London, v.52, n.1-2, p.225-247, 1965.

LAPLACE, P. S. Sur les naissances, les marriages et les morts. *Hist. L'Acad. R. Sci.*, p.1963, 1783.

LEE, A. J. et. al. Capture-recapture, epidemiology and list mistaches: several lists. *Biometrics*, Arlington, v.57, n.1, p.707-713, 2001.

LEE, A. J. Effect of list errors on the estimation of population size. *Biometrics*, Arlington, v.58, p.185-191, 2002.

LINCOLN, F. C. *Calculating waterfowl abundance on the basis of banding returns*. Department of Agricultural, 1930. p.1-4. (*Circular*, n.118).

MADIGAN, D.; YORK, J. C. Bayesian methods for estimation of the size of a closed population, *Biometrika*, London, v.84, n.1, p.19031, 1997.

NAYAK, T. K. Estimating population size by recapture sampling. *Biometrika*, London, v.88, n.1, p.113-120, 1988.

PETERSEN, C. G. The yearly immigration of young plaice into Limfjord from the German sea, etc. *Rept. Danish Biol. Stn.*, v.6, n.1, p.1-48, 1896.

POLLOCK, K. H. Modeling capture-recapture and removal statistics for estimation of demographic parameters for fish and wildlife populations: past, present and future. *J. Am. Stat. Assoc.*, Washington, v.86, n.1, p.225-238, 1991.

RISSANEN, J. A universal prior for integers and estimation by minimum description length. *Ann. Stat.*, Hayward, v.11, n.2, p.416-431, 1983.

SEBER, G. A. F. A review of estimating animal abundance. *Biometrics*, Washington, v.42, n.2, p.267-292, 1986.

YOSHIDA, O. S.; LEITE, J. G.; BOLFARINE, H. Stochastic monotonicity properties of Bayes estimation of the population size for capture-recapture data. *Stat. Prob. Lett.*, Amsterdanm, v.42, n.3, p.257-266, 1999.

WEST, M. Generalized linear models: parameter, outlier accomodation and prior distributions. In: BERNARDO, J. M. et al. (Ed.). *Bayesian statistics.* North-Holland: Elsevier Science, 1985. p.531-558.

Recebido em 05.10.2007.

Aprovado após revisão em 12.03.2008.

Apêndice I - Prova do teorema

Primeiramente, notemos que a constante normalizadora de (2), *C*, é tal que

$$C^{-1} = \sum_{N\geq n}\int_A \frac{N!}{(N-n)!}\pi(N)\prod_{r=1}^{l} p_r^{\,n_r+\alpha_r-1}dp_1dp_2\ldots dp_{l-1}$$

$$= \sum_{N\geq n}\frac{N!}{(N-n)!}\pi(N)\int_A\prod_{r=1}^{l} p_r^{\,n_r+\alpha_r-1}dp_1dp_2\ldots dp_{l-1},$$

onde $A=\{(p_1,p_2,\ldots,p_{l-1}): 0<p_i<1, i=1,2,\ldots,l-1 \text{ e} \sum_{i=1}^{l-1} p_i<1\}$. Então,

$$C^{-1} = \sum_{N\geq n}\frac{N!}{(N-n)!}\pi(N)\frac{\prod_{r=1}^{l}\Gamma(\alpha_r+n_r)}{\Gamma\left[\sum_{r=1}^{l}(\alpha_r+n_r)\right]}$$

$$= \sum_{N\geq n}\frac{N!}{(N-n)!}\pi(N)\frac{\prod_{r=1}^{l-1}\Gamma(\alpha_r+n_r)}{\Gamma\left(\sum_{r=1}^{l}\alpha_r+N\right)}\Gamma(\alpha_l+N-n)\propto S_n$$

onde

$$S_n = \sum_{N\geq n}\frac{N!}{(N-n)!}\pi(N)\frac{\Gamma(\alpha_l+N-n)}{\Gamma\left(\sum_{r=1}^{l}\alpha_r+N\right)} \qquad (5)$$

Inicialmente, temos

$$i)\ \frac{N!}{(N-n)!N^n} = \frac{N(N-1)\ldots[N-(n-1)]}{N^n}$$

$$= 1\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)\ldots\left(1-\frac{n-1}{N}\right)\underset{N\to\infty}{\to} 1$$

Mostremos agora que

$$ii)\ \frac{\Gamma(\alpha_l + N - n)N^{\sum_{r=1}^{l-1}\alpha_r + n}}{\Gamma(\sum_{r=1}^{l}\alpha_r + N)} \underset{N\to\infty}{\to} 1.$$

Como

$$\left(\frac{\alpha_l + N - n}{N}\right)^{\alpha_l + N - n - \frac{1}{2}} \underset{N\to\infty}{\to} \exp(\alpha_l - n),$$

$$\left(\frac{\sum_{r=1}^{l}\alpha_r + N}{N}\right)^{-\left(\sum_{r=1}^{l}\alpha_r + N - \frac{1}{2}\right)} \underset{N\to\infty}{\to} \frac{1}{\exp\left(\sum_{r=1}^{l}\alpha_r\right)}$$

e uma vez que o resultado

$$\frac{\Gamma(x)\exp(x)}{x^{x-\frac{1}{2}}} \underset{N\to\infty}{\to} \sqrt{2\pi},\ (Feller\,(1967))$$

implica

$$\frac{\Gamma(\alpha_l + N - n)\exp(\alpha_l + N - n)}{(\alpha_l + N - n)^{\alpha_l + N - n - \frac{1}{2}}} \underset{N\to\infty}{\to} \sqrt{2\pi}$$

e

$$\frac{\left(\sum_{r=1}^{l}\alpha_r + N\right)^{\sum_{r=1}^{l}\alpha_r + N - \frac{1}{2}}}{\Gamma\left(\sum_{r=1}^{l}\alpha_r + N\right)\exp\left(\sum_{r=1}^{l}\alpha_r + N\right)} \underset{N\to\infty}{\to} \frac{1}{\sqrt{2\pi}},$$

então

$$\frac{\Gamma(\alpha_l + N - n) N^{\sum_{r=1}^{l-1} \alpha_r + n}}{\Gamma(\sum_{r=1}^{l} \alpha_r + N)}$$

$$= \frac{\Gamma(\alpha_l + N - n) \exp(\alpha_l + N - n)}{(\alpha_l + N - n)^{\alpha_l + N - n - \frac{1}{2}}} \frac{\left(\sum_{r=1}^{l} \alpha_r + N\right)^{\sum_{r=1}^{l} \alpha_r + N - \frac{1}{2}}}{\Gamma\left(\sum_{r=1}^{l} \alpha_r + N\right) \exp\left(\sum_{r=1}^{l} \alpha_r + N\right)}$$

$$\times \left(\frac{\alpha_l + N - n}{N}\right)^{\alpha_l + N - n - \frac{1}{2}} \left(\frac{\sum_{r=1}^{l} \alpha_r + N}{N}\right)^{-\left(\sum_{r=1}^{l} \alpha_r + N - \frac{1}{2}\right)}$$

$$\times \exp\left(-\alpha_l + n + \sum_{r=1}^{l} \alpha_r\right) \underset{N \to \infty}{\to}$$

$$\sqrt{2\pi} \frac{1}{\sqrt{2\pi}} \exp(\alpha_l - n) \frac{1}{\exp\left(\sum_{r=1}^{l} \alpha_r\right)} \exp\left(\alpha_l - n + \sum_{r=1}^{l} \alpha_r\right)$$

$$= 1,$$

o que prova *ii*).

Logo, de *i)* e *ii)*, segue que

$$N^{\sum_{r=1}^{l-1} \alpha_r} \frac{N!}{(N-n)!} \frac{\Gamma(\alpha_l + N - n)}{\Gamma(\sum_{r=1}^{l} \alpha_r + N)} \underset{N \to \infty}{\to} 1,$$

ou seja, fixado um número real ε, $0 < \varepsilon < 1$, existe um número inteiro positivo N_0, $N_0 > n$, tal que

$$1 - \varepsilon < N^{\sum_{r=1}^{l-1} \alpha_r} \frac{N!}{(N-n)!} \frac{\Gamma(\alpha_l + N - n)}{\Gamma(\sum_{r=1}^{l} \alpha_r + N)} < 1 + \varepsilon, \tag{6}$$

para todo $N > N_0$ e S_n (dado em (5)) pode ser escrito como

$$S_n = \sum_{N=n}^{N_0} \frac{N!}{(N-n)!}\frac{1}{N^u}\frac{\Gamma(\alpha_l + N - n)}{\Gamma(\sum_{r=1}^{l}\alpha_r + N)} + W_{N_0},$$

onde

$$W_{N_0} = \sum_{N=N_0+1}^{\infty} \frac{N!}{(N-n)!}\frac{1}{N^u}\frac{\Gamma(\alpha_l + N - n)}{\Gamma(\sum_{r=1}^{l}\alpha_r + N)}.$$

Mas, de (6), segue que

$$(1-\varepsilon)\sum_{N=N_0+1}^{\infty} N^{-\left(\sum_{r=1}^{l-1}\alpha_r + u\right)} < W_{N_0} < (1+\varepsilon)\sum_{N=N_0+1}^{\infty} N^{-\left(\sum_{r=1}^{l-1}\alpha_r + u\right)}.$$

Logo, para $u = 0$ temos

$$\sum_{r=1}^{l-1}\alpha_r + u = \sum_{r=1}^{l-1}\alpha_r > 1 \Rightarrow \sum_{N=N_0+1}^{\infty} N^{-\sum_{r=1}^{l-1}\alpha_r} < \infty \Rightarrow 0 < W_{N_0} < \infty$$
$$\Rightarrow 0 < S_n < \infty,$$

o que implica na existência da distribuição *a posteriori* conjunta de N $(p_1, p_2,...,p_l)$. Reciprocamente,

$$\sum_{r=1}^{l-1}\alpha_r \le 1 \Rightarrow \sum_{N=N_0+1}^{\infty} N^{-\sum_{r=1}^{l-1}\alpha_r} \text{ é infinita} \Rightarrow W_{N_0} \text{ é infinita} \Rightarrow S_n \text{ infinita},$$

que implica na não existência da distribuição *a posteriori* conjunta de N e $(p_1, p_2,..., p_{l-1})$, o que prova a).

Por outro lado, para $u = 1$ temos

$$\sum_{r=1}^{l-1}\alpha_r + u = \sum_{r=1}^{l-1}\alpha_r + 1 > 1 \Rightarrow \sum_{N=N_0+1}^{\infty} N^{-\left(\sum_{r=1}^{l-1}\alpha_r + 1\right)} < \infty \Rightarrow 0 < W_{N_0} < \infty$$
$$\Rightarrow 0 < S_n < \infty,$$

o que prova b).

Apêndice II – Programas utilizados no estudo do desempenho do modelo (seção 4)

```
### Geração dos dados utilizando o software R ###
set.seed(3)
x<- numeric()
k<-2
l<-2^k
alpha<-10
beta<-1
for (i in 1:l)
{
x[i]<-rgamma(1,alpha,beta)
}
somax<-sum(x)
for (i in 1:l)
{
 x[i]<-x[i]/somax
 }
 N<-100
  y<-numeric()
  y<-rmultinom(1,N,x)
```

Obtenção de resultados utilizando o *software* MAPLE

Distribuição *a priori* atribuída a N: uniforme nos inteiros não negativos

```
> restart:
> f[N]:=(N!/(N-nn)!)*(GAMMA(alpha+N-nn)/GAMMA(alpha*l+N)):
> A[N]:=simplify(eval(f[N],N=N+1)/f[N]):
> kk:=2:
> l:=2^kk:
> alpha:=5:
> distintos:=67:
> u:=150:
> n[distintos]:=distintos:
> nn:=n[distintos]:
> for i from nn to u+1 do
>    n[i]:=i
> od:
> for i from nn to u do
>   f[i+1]:=evalf(eval(simplify(f[N]*A[N]),N=i)):
> od:
> with(plots):
> f[nn]:=evalf(eval(f[N],N=nn)):
> k:=1/sum('f[i]','i'=nn..u+1):
> for i from n[nn] to n[u+1] do
>    distN[i]:=k*f[i]:
```

```
> od:
> L:=[n[nn],distN[nn]]:
> for i from n[nn+1] to u do
> L:=L,[n[i],distN[i]]:
> od:
> L: = [L]:
> plot(L,n=n[nn]..n[u],style=point,symbol=circle):
> sum('distN[i]','i'=nn..u+1):
> média:=sum('distN[i]*n[i]','i'=nn..u+1):
> variância:=sum('distN[i]*((n[i]-média)^2)','i'=nn..u+1):
> dp:=sqrt(variância):
> for i from nn to u do
>   if(sum('distN[j]','j'=nn..i)<=0.025) then s:=i end if:
>   od:
> if (s=nn) then IC1:=nn else IC1:=s end if;
> IC1:
> for i from nn to u do
> if(sum('distN[j]','j'=nn..i)<=0.975) then s:=i end if:
> od:
> IC2:=s:
> for i from nn to u do
> if(sum('distN[j]','j'=nn..i)<=0.25) then s:=i end if:
> od:
> q25:=s:
> for i from nn to u do
> if(sum('distN[j]','j'=nn..i)<=0.5) then s:=i end if:
> od:
> q50:=s:
> for i from nn to u do
> if(sum('distN[j]','j'=nn..i)<=0.75) then s:=i end if:
> od:
> q75:=s:
> for i from nn to 200 do
> if(distN[i]<=distN[i+1]) then s:=i else t[i]:=i end if:
> print(t[i]):
> od:
```