Estimativas de máxima verossimilhança e bayesianas do número de erros de um *software*.

Karolina Barone Ribeiro da Silva.

Orientador: Prof. Dr. José Galvão Leite.

Dissertação apresentada ao Departamento de Estatística da Universidade Federal de São Carlos - DEs/UFSCar, como parte dos requisitos para obtenção do título de Mestre em Estatística.

São Carlos

Fevereiro de 2006.

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

S586em

Silva, Karolina Barone Ribeiro da.
Estimativas de máxima verossimilhança e bayesianas do número de erros de um *software* / Karolina Barone Ribeiro da Silva. -- São Carlos : UFSCar, 2006.
128 p.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2006.

1. Estatística matemática. 2. Processo seqüencial de captura-recaptura. 3. MCMC. 4. Inferência bayesiana. 5. Estimativas de máxima verossimilhança. I. Título.

CDD: 519.54 (20ª)

Agradecimentos

Primeiramente aos meus pais, Véra e Agostinho pela oportunidade, pelo apoio, pela paciência e pela ajuda financeira e também ao meu irmão, Gu, pelos momentos de descontração.

Ao meu namorado, Ramirez, pelo apoio, infinita paciência, companheirismo, dedicação, carinho, amor e felicidade que vem me proporcionando desde o ano passado.

Ao meu orientador, Prof. Dr. José Galvão Leite, pelas idéias e orientação.

Ao Prof. Dr. Luis A. Milan, pelas idéias.

Às minhas amigas de graduação Cá, Glauce, Gonça e Luci pelos momentos inesquecíveis que vivemos durante quatro anos, pelo incentivo a ingressar no mestrado, nunca desistir e pela amizade que temos até hoje.

Ao meu amigo Joaquim, pela amizade e apoio.

À Ermínia de Lourdes C. Fanti, minha orientadora de iniciação científica na graduação, que foi quem me iniciou no "mundo" da pesquisa de forma brilhante e me apoiou a ingressar no Mestrado, mesmo querendo que eu ficasse por lá.

À minha amiga Simone, por vários motivos, mas principalmente por ter me apresentado um amigo muito especial até hoje.

Às minhas amigas Juliana, Kelly, Rosa, Janaina e Uliana que, mesmo longe, sempre torceram por mim.

Às minhas amigas de mestrado, Cá, Josi, Fá e Elô, pelo apoio, mas principalmente pelas "palhaçadas" e risadas.

A todos os colegas de mestrado, pelas ajudas que prestamos uns aos outros.

A todos os integrantes da extinta república "Toca da Porca" e seus respectivos agregados ou ex-agregados, por terem sempre me recebido muito bem e pelos momentos de diversão. Ao atual morador da nova república, Zanu e sua namorada, Liliam, por continuarem me "agüentando".

À Fabiana, que mesmo sem me conhecer me recebeu em sua casa durante o curso de verão.

À "D. Luiza", funcionária exemplar do departamento, que com sua simpatia, simplicidade, consideração e amizade conquistou a todos.

Aos funcionários e professores do departamento.

À Capes, pelo apoio financeiro parcial.

Resumo

Nesta dissertação apresentamos a metodologia de captura-recaptura, sob os enfoques clássico e bayesiano, para estimar o número de erros de um *software* através de sua inspeção por revisores distintos. Apresentamos o modelo estatístico geral considerando independência entre erros e entre revisores e consideramos os casos particulares de erros igualmente.detectáveis (homogêneos) e revisores não igualmente eficientes (heterogêneos) e de erros não igualmente detectáveis (heterogêneos) e revisores igualmente eficientes (homogêneos). Em seguida, sob a hipótese de heterogeneidade e independência entre erros e homogeneidade e independência entre revisores, supusemos que a heterogeneidade dos erros era expressa por uma classificação destes em fácil e difícil de detectar, admitindo conhecidas as probabilidades de detecção de um erro fácil e de um erro difícil. Finalmente, sob a hipótese de independência e homogeneidade entre erros, apresentamos um novo modelo considerando heterogeneidade e dependência entre revisores. Além disso, apresentamos exemplos com dados simulados e reais.

Palavras-chave: processo de captura-recaptura, revisão de software, estimativas de máxima verossimilhança, distribuições *a priori* e *a posteriori*, estimativas de Bayes.

Abstract

In this work we present the methodology of capture-recapture, under the classic and bayesian approach, to estimate the number of errors of software through inspection by distinct reviewers. We present the general statistical model considering independence among errors and among reviewers and consider the particular cases of equally detectable errors (homogeneous) and reviewers not equally efficient (heterogeneous) and of errors not equally detectable (heterogeneous) and equally efficient reviewers (homogeneous). After that, under the assumption of independence and heterogeneity among errors and independence and homogeneity among reviwers, we supposed that the heterogeneity of the errors was expressed by a classification of these in easy and difficult of detecting, admitting known the probabilities of detection of an easy error and of a difficult error. Finally, under the hypothesis of independence and homogeneity among errors, we presented a new model considering heterogeneity and dependence among reviewers. Besides, we presented examples with simulate and real data.

Keywords: capture-recapture process, *software* review, maximum likelihood estimates, *a priori* and *a posteriori* distributions, Bayes estimates.

Sumário

Capítulo 1

Introdução

Suponhamos uma população de N indivíduos portadores de uma característica de interesse, onde N é desconhecido e é do nosso interesse estimá-lo.

Nesta dissertação a metodologia utilizada para estimação de N será a de captura-recaptura sob os enfoques clássico e bayesiano. Esta metodologia, utilizada para estimar tamanhos de populações animais, consiste, primeiramente, na seleção de um número fixado ou aleatório de animais da população. Todos os animais capturados são marcados e devolvidos à população. Em seguida um número fixado ou aleatório de animais é selecionado em uma ou mais épocas. Em cada uma das épocas todos os animais não marcados recebem marcas e são devolvidos à população. No final do processo estimamos N, baseados nos números de animais capturados nas diversas épocas e no número de animais distintos observados.

A literatura sobre as aplicações do método de captura-recaptura na estimação do tamanho de uma população é vasta. Estudos relacionados a esta metodologia se iniciaram em 1896, quando Petersen usou-a para estudar o fluxo migratório de peixes no mar Báltico embora, em 1783, Laplace já houvesse utilizado tal método para estimar o tamanho da população da França. Em 1930 Lincoln aplicou o método para estimar o tamanho de uma população de patos selvagens e a partir da década de 50 diversos pesquisadores, tais como, Chapman (1954), Darroch (1959), Jolly(1965), Burnham e Overton (1978), Seber (1986) e Pollock (1991) publicaram trabalhos sobre o assunto. Sob o enfoque bayesiano temos publicações de Hunter e Griffiths (1978), Castledine (1981), George e Robert (1992)

e Yoshida, Leite e Bolfarine (1999), por exemplo. Além disso, várias dissertações de mestrado foram desenvolvidas sobre o tema, tais como Zacharias (2000), Rossi (2001), Shimizu (2002), Bolsoni (2002) e Missiagia (2004).

É importante salientarmos que embora o método de captura-recaptura tenha sido originalmente utilizado para estimação de tamanhos de populações animais, ele também pode ser aplicado na estimação do tamanho de qualquer população para a qual cada indivíduo tenha uma característica que o identifique em diferentes situações amostrais. Assim, as estimativas produzidas por esse método são relevantes não só nas ciências ambientais, mas também na saúde pública, no controle de erros de *software*, no controle de qualidade e em outras áreas do conhecimento.

Na saúde pública, o método pode ser utilizado para estimar N, o número de pacientes com certa doença, auxiliando no planejamento de medidas preventivas, como campanhas de vacinação, por exemplo. Neste caso, existem listas de indivíduos com a doença em hospitais e centros de saúde, e um indivíduo pode fazer ou não parte de uma ou mais destas listas. Logo, se imaginarmos que a cada indivíduo corresponde um animal, então ao observarmos o nome de um indivíduo pela primeira vez numa lista temos um evento correspondente à seleção de um animal não marcado. Por outro lado, a observação do nome de um indivíduo que já foi observado em alguma lista anterior, corresponde à seleção de um animal marcado. Assim, podemos aplicar o método de captura-recaptura na estimação de N.

Quanto ao controle de erros de *software*, o interesse é estimar o número total de erros do *software*, N. Aqui o *software* é examinado por um certo número de revisores distintos. Os erros observados são então contados e codificados. Além disso o histórico (trajetória) das observações (capturas) de cada erro é registrado. Se imaginarmos que a cada erro corresponde um animal, temos que a observação de um erro que ainda não foi observado corresponde à seleção de um animal não marcado e a observação de um erro que já foi observado anteriormente, por algum revisor, corresponde à seleção de um animal marcado. Assim o método de captura-recaputra se aplica novamente na estimação de N.

No controle de qualidade o método pode ser aplicado para estimar o número de partículas, N, contidas em um filme fotográfico que será submetido à leitura por diferentes *scanners* e cujas partículas observadas serão registradas. Analogamente aos exemplos

anteriores, se imaginarmos que a cada partícula corresponde um animal, então a observação de uma partícula que ainda não foi observada equivale à seleção de um animal não marcado e a observação de uma partícula já anteriormente observada por algum *scanner* corresponde à seleção de um animal marcado. Portanto, o método de captura-recaptura pode ser mais uma vez aplicado na estimação de N.

Entre os pesquisadores que produziram trabalhos nas áreas citadas acima temos, por exemplo, Lee *et al.* (2001) e Lee (2002) na área de saúde, Nayak (1988) e Basu e Ebrahimi (2001) na área de controle de erros de *software* e Sanathanan (1972) na área de controle de qualidade.

Nesta dissertação vamos nos ater, a título de ilustração, ao caso da inferência sobre o número de erros de um *software*, mas evidentemente os resultados também são válidos nas demais situações citadas.

Um dos aspectos motivadores do estudo do assunto revisão de *software* é que, atualmente, a detecção de erros em programas de computador é uma grande preocupação, pois muitas vezes tais erros podem causar prejuízos milionários. Foi o caso, por exemplo, de um *software* embutido na sonda Mars Climate Orbiter- da Nasa, que misturou medidas em pés com metros e provocou um prejuízo de 125 milhões de dólares. (Fonte: Revista Exame, 31/03/2004).

Além disso, segundo dados do *Instituto de Qualidade de software (Universidade do Texas), software*s com defeito custam 60 bilhões de dólares à economia americana todo ano e metade deste custo recai sobre o comprador, já que em média, mais de 15% dos erros contidos no programa só são detectados após sua venda, o que muitas vezes obriga o comprador a adquirir versões atualizadas do produto.

Ressaltamos que nosso trabalho nesta dissertação baseou-se no artigo de Basu e Ebrahimi (2001), que trataram do tema revisão de *software* sob as mais variadas condições.

No capítulo 2 discutimos o modelo estatístico geral para o problema de revisão de *software*, considerando independência entre erros e entre revisores. Particularizamos o modelo para o caso de erros igualmente detectáveis (homogêneos) e revisores não igualmente eficientes (heterogêneos), e determinamos estimativas de máxima verossimilhança e bayesianas dos parâmetros. Estudamos também o modelo para o caso de erros não igualmente detectáveis (heterogêneos) e revisores igualmente eficientes (homogêneos) e,

neste caso, apresentamos estimativas bayesianas do tamanho populacional.

No capítulo 3, mantendo as suposições de heterogeneidade e independência entre erros e homogeneidade e independência entre revisores, supomos que a heterogeneidade dos erros é expressa por uma classificação destes em fácil e difícil de detectar. Determinamos estimativas de máxima verossimilhança condicionais e bayesianas dos parâmetros do modelo, admitindo conhecidas as probabilidades de detecção de um erro fácil e de um erro difícil.

No capítulo 4, mantendo a hipótese de independência e homogeneidade entre erros, descrevemos um novo modelo para o problema de revisão de *software*, considerando heterogeneidade e dependência entre revisores. Também para este modelo determinamos estimativas bayesianas de N.

Finalmente, no capítulo 5 tecemos nossas considerações finais sobre este trabalho.

Capítulo 2

Modelo estatístico para revisão de *software*: independência entre erros e entre revisores.

Neste capítulo tratamos do problema de estimação do número de erros de um *software* com a suposição de independência entre erros e entre revisores.

Seja N (N desconhecido) o número de erros de um *software* e k ($k \geq 2$) o número de revisores que irão examiná-lo. Suponhamos que cada erro seja detectado ou não por um revisor qualquer, independemente dos demais erros, e que os revisores examinem o *software* separadamente, isto é, os revisores também atuam independentemente uns dos outros. Dado N, consideremos o vetor aleatório k-dimensional $\mathbf{X}_i = (X_{i1}, X_{i2}, ..., X_{ik})$ associado a cada erro i, onde X_{ij} assume valor 1 se o erro i for detectado pelo j-ésimo revisor e 0 caso contrário, $i = 1, 2, ..., N$, $j = 1, 2, ..., k$. Seja $\theta_{ij} = P(X_{ij} = 1), 0 < \theta_{ij} < 1$, ou seja, θ_{ij} é a probabilidade do erro i ser detectado pelo revisor j, e seja $\boldsymbol{\theta} = \{\theta_{ij}, 1 \leq i \leq N, 1 \leq j \leq k\}$. Assim as variáveis aleatórias X_{ij} são independentes e têm distribuição de Bernoulli com parâmetro θ_{ij}, isto é,

$$P(X_{ij} = x|N, \boldsymbol{\theta}) = \theta_{ij}^x (1 - \theta_{ij})^{1-x} I_{\{0,1\}}(x), \tag{2.1}$$

$1 \leq i \leq N, 1 \leq j \leq k$.

Além disso, 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 $\mathbf{w}_r = (w_{r1}, w_{r2}, ..., w_{rk})$, $r = 1, 2, ..., l-1$ e $\mathbf{w}_l = (0, 0, ..., 0)$. Notemos que cada $\mathbf{w}_r$, $1 \leq r \leq l$, representa um histórico de leitura (trajetória) das observações de um erro i, $1 \leq i \leq N$, uma vez que $\mathbf{w}_r$ é uma k-upla de elementos que assumem os valores um ou zero, conforme a detecção ou não do erro por cada revisor, e que $\mathbf{w}_l$ é um histórico de leitura não observável. Além disso, temos 2^k possíveis históricos de leitura para cada erro, pois $\#(\Delta) = 2^k$.

Então, para cada $i = 1, 2, ..., N$, a probabilidade do erro i apresentar o histórico $\mathbf{w}_r$ é

$$\begin{aligned} p_{ir}(\boldsymbol{\theta}) &= P(\mathbf{X}_i = \mathbf{w}_r | N, \boldsymbol{\theta}) = P((X_{i1}, X_{i2}..., X_{ik}) = (w_{r1}, w_{r2}..., w_{rk}) | N, \boldsymbol{\theta}) = \\ &= \prod_{j=1}^{k} P(X_{ij} = w_{rj} | N, \boldsymbol{\theta}) = \prod_{j=1}^{k} \theta_{ij}^{w_{rj}} (1 - \theta_{ij})^{1 - w_{rj}}, \end{aligned} \tag{2.2}$$

$r = 1, 2, ..., l$.

Sejam $n_r = \sum_{i=1}^{N} I_{\{\mathbf{w}_r\}}(\mathbf{X}_i)$ o número de erros com o histórico de leitura $\mathbf{w}_r$, $r = 1, 2, ..., l$ e $n = \sum_{r=1}^{l-1} n_r$ o número de erros distintos observados pelos k revisores. Notamos que $n_l = N - n$ é o número de erros que apresentam o histórico $\mathbf{w}_l$, isto é, o número de erros não observados.

Então, a distribuição de probabilidades de $(n_1, n_2, ..., n_{l-1}, N-n)$, dados N e $\boldsymbol{\theta}$, é dada por

$$P(n_1, n_2, ..., n_{l-1}, N - n | N, \boldsymbol{\theta}) =$$

$$= \sum \prod_{i_1 \epsilon A_1} p_{i_1 1}(\boldsymbol{\theta}) \prod_{i_2 \epsilon A_2} p_{i_2 2}(\boldsymbol{\theta}) ... \prod_{i_{l-1} \epsilon A_{l-1}} p_{i_{l-1}(l-1)}(\boldsymbol{\theta}) \prod_{i_l \epsilon A_l} p_{i_l l}(\boldsymbol{\theta}), \tag{2.3}$$

onde $\sum$ se extende a todos os subconjuntos disjuntos $A_1, A_2, ..., A_l$ de $\{1, 2, ..., N\}$, tais que $\#(A_r) = n_r$, $r = 1, 2, ...l$. Notamos que A_r representa o conjunto dos erros com histórico $\mathbf{w}_r$ e o número de parcelas de $\sum$ é igual a

$$\binom{N}{n_1}\binom{N - n_1}{n_2}\binom{N - (n_1 + n_2)}{n_3} \cdot \ ... \ \cdot \binom{N - (n_1 + n_2 + ... + n_{l-1})}{N - n} =$$

$$= \frac{N!}{n_1! n_2! ... (N - n)!}.$$

Exemplo 1.

Para vizualização do problema, suponhamos $N = k = 2$ (na prática N é desconhecido).

Assim, estamos na situação em que há apenas dois erros no *software* e dois revisores para examiná-lo.

Como $k = 2$, então para cada erro há $l = 2^2 = 4$ possíveis históricos de leitura, que enumeramos da seguinte forma:

$\mathbf{w}_1 = (1, 0)$: o erro é detectado pelo revisor 1, mas não pelo 2;

$\mathbf{w}_2 = (0, 1)$: o erro não é detectado pelo revisor 1, mas é pelo 2;

$\mathbf{w}_3 = (1, 1)$: o erro é detectado por ambos os revisores;

$\mathbf{w}_4 = (0, 0)$: o erro não é detectado por revisor algum.

De (2.2), a probabilidade do erro i apresentar o histórico $\mathbf{w}_r$ é

$$\begin{aligned} p_{ir}(\boldsymbol{\theta}) &= P(\mathbf{X}_i = \mathbf{w}_r | N = 2, \boldsymbol{\theta}) = P((X_{i1}, X_{i2}) = (w_{r1}, w_{r2}) | N = 2, \boldsymbol{\theta}) = \\ &= \prod_{j=1}^{2} P(X_{ij} = w_{rj} | N = 2, \boldsymbol{\theta}) = \prod_{j=1}^{2} \theta_{ij}^{w_{rj}} (1 - \theta_{ij})^{1 - w_{rj}}, \end{aligned}$$

onde $i = 1, 2$ e $r = 1, 2, 3, 4$.

Então,

$$\begin{aligned} p_{i1}(\boldsymbol{\theta}) &= \theta_{i1}(1 - \theta_{i2}), \\ p_{i2}(\boldsymbol{\theta}) &= (1 - \theta_{i1})\theta_{i2}, \\ p_{i3}(\boldsymbol{\theta}) &= \theta_{i1}\theta_{i2}, \\ p_{i4}(\boldsymbol{\theta}) &= (1 - \theta_{i1})(1 - \theta_{i2}), \end{aligned}$$

e, por exemplo,

$$P((n_1, n_2, n_3, 2 - n) = (1, 0, 0, 1) | N = 2, \boldsymbol{\theta}) =$$

$$\begin{aligned} &= \sum \prod_{i_1 \epsilon A_1} p_{i_1 1}(\boldsymbol{\theta}) \prod_{i_2 \epsilon A_2} p_{i_2 2}(\boldsymbol{\theta}) \prod_{i_3 \epsilon A_3} p_{i_3 3}(\boldsymbol{\theta}) \prod_{i_4 \ \epsilon A_4} p_{i4}(\boldsymbol{\theta}) = \\ &= p_{11}(\boldsymbol{\theta}) p_{24}(\boldsymbol{\theta}) + p_{21}(\boldsymbol{\theta}) p_{14}(\boldsymbol{\theta}) = \\ &= \theta_{11}(1 - \theta_{12})(1 - \theta_{21})(1 - \theta_{22}) + \theta_{21}(1 - \theta_{22})(1 - \theta_{11})(1 - \theta_{12}). \end{aligned}$$

Notemos que munidos das informações $N = 2$ e $(n_1, n_2, n_3, n_4) = (1, 0, 0, 1)$, temos que considerar todos os possíveis subconjuntos A_1, A_2, A_3, A_4 de $\{1, 2\}$, tais que $\#(A_r) = n_r, r = 1, 2, 3, 4$. Estes subconjuntos são dados por

$$\begin{aligned} A_1 &= \{1\}, A_2 = \{\}, A_3 = \{\}, A_4 = \{2\} \text{ e} \\ A_1' &= \{2\}, A_2' = \{\}, A_3' = \{\}, A_4' = \{1\}. \end{aligned}$$

Notemos que há duas classes de subconjuntos de {1, 2} possíveis, pois

$$\frac{N!}{n_1!n_2!...(N-n)!} = \frac{2!}{n_1!n_2!n_3!(2-n)!} = 2.$$

Em algumas situações os erros podem ser igualmente detectáveis (erros homogêneos) e os revisores não serem igualmente eficientes (revisores heterogêneos), isto é, $\theta_{ij} = \theta_j$ para todo i, $i = 1, 2, ..., N$ e $j = 1, 2, ..., k$. Também há o caso em que os erros não são igualmente detectáveis (erros heterogêneos) e os revisores são igualmente eficientes (revisores homogêneos), ou seja, $\theta_{ij} = \theta_i$ para todo i, $i = 1, 2, ..., N$ e $j = 1, 2, ..., k$. No caso em que os erros e os revisores são homogêneos, temos $\theta_{ij} = \theta$, para quaisquer i e j, $i = 1, 2, ..., N$ e $j = 1, 2, ..., k$. Nas seções 2.1 e 2.2 a seguir, trataremos dos dois primeiros casos, uma vez que o último tem pouca aplicação prática.

2.1 Homogeneidade e independência entre erros e heterogeneidade e independência entre revisores.

Nesta seção vamos supor que os revisores atuam independentemente e não são igualmente eficientes (heterogêneos) e que os erros são independentes e igualmente detectáveis (homogêneos), ou seja, que $\theta_{ij} = \theta_j, i = 1, 2, ..., N$ e $j = 1, 2, ..., k$.

Então, segue de (2.2), que para cada $i = 1, 2, ..., N$, a probabilidade do erro i apresentar o histórico $\mathbf{w}_r$ é

$$p_{ir}(\boldsymbol{\theta}) = \prod_{j=1}^{k} \theta_j^{w_{rj}} (1 - \theta_j)^{1 - w_{rj}}, r = 1, 2, ..., l. \tag{2.4}$$

Analisando o último membro de (2.4), podemos observar que a probabilidade de um

erro apresentar determinado histórico é a mesma, qualquer que seja o erro em questão, isto é, $p_{ir}(\boldsymbol{\theta})$ não depende de i. Assim, temos

$$p_r(\boldsymbol{\theta}) = p_{ir}(\boldsymbol{\theta}) = \prod_{j=1}^{k} \theta_j^{w_{rj}} (1-\theta_j)^{1-w_{rj}}, r = 1, 2, ..., l.$$

Logo, de (2.3) segue que,

$$P(n_1, n_2, ..., n_{l-1}, N-n | N, \boldsymbol{\theta}) =$$

$$\begin{aligned}
&= \sum \prod_{i_1 \epsilon A_1} p_{i_1 1}(\boldsymbol{\theta}) \prod_{i_2 \epsilon A_2} p_{i_2 2}(\boldsymbol{\theta}) ... \prod_{i_{l-1} \epsilon A_{l-1}} p_{i_{l-1}(l-1)}(\boldsymbol{\theta}) \prod_{i_l \epsilon A_l} p_{i_l l}(\boldsymbol{\theta}) = \\
&= \sum (p_1(\boldsymbol{\theta}))^{n_1} (p_2(\boldsymbol{\theta}))^{n_2} ... (p_{l-1}(\boldsymbol{\theta}))^{n_{l-1}} (p_l(\boldsymbol{\theta}))^{N-n} = \\
&= \frac{N!}{n_1! n_2! ... n_{l-1}! (N-n)!} (p_1(\boldsymbol{\theta}))^{n_1} (p_2(\boldsymbol{\theta}))^{n_2} ... (p_{l-1}(\boldsymbol{\theta}))^{n_{l-1}} (p_l(\boldsymbol{\theta}))^{N-n} = \\
&= \frac{N!}{n_1! n_2! ... n_{l-1}! (N-n)!} \prod_{r=1}^{l} (p_r(\boldsymbol{\theta}))^{n_r} = \\
&= \frac{N!}{n_1! n_2! ... n_{l-1}! (N-n)!} \prod_{r=1}^{l} \left(\prod_{j=1}^{k} \theta_j^{w_{rj}} (1-\theta_j)^{1-w_{rj}} \right)^{n_r} = \\
&= \frac{N!}{n_1! n_2! ... n_{l-1}! (N-n)!} \prod_{r=1}^{l} \prod_{j=1}^{k} \theta_j^{n_r w_{rj}} (1-\theta_j)^{n_r - n_r w_{rj}} = \\
&= \frac{N!}{n_1! n_2! ... n_{l-1}! (N-n)!} \prod_{j=1}^{k} \theta_j^{n_{(j)}} (1-\theta_j)^{N-n_{(j)}}, \qquad (2.5)
\end{aligned}$$

onde

$$\begin{aligned}
n_{(j)} &= \sum_{r=1}^{l} n_r w_{rj} = \sum_{r=1}^{l} \sum_{i=1}^{N} I_{\{\mathbf{w}_r\}}(\mathbf{X}_i) w_{rj} = \sum_{i=1}^{N} \sum_{r=1}^{l} I_{\{\mathbf{w}_r\}}(\mathbf{X}_i) w_{rj} = \\
&= \sum_{i=1}^{N} \sum_{r=1}^{l} I_{\{\mathbf{w}_r\}}(\mathbf{X}_i) \; X_{ij} = \sum_{i=1}^{N} \left(X_{ij} \sum_{r=1}^{l} I_{\{\mathbf{w}_r\}}(\mathbf{X}_i) \right) = \sum_{i=1}^{N} X_{ij}
\end{aligned}$$

é o número de erros detectados pelo revisor j, $j = 1, 2, ..., k$.

2.1.1 Função de verossimilhança e estimativas de máxima verossimilhança dos parâmetros.

Nesta seção apresentamos a função de verossimilhança para o caso $\theta_{ij} = \theta_j$ e determinamos a estimativa de máxima verossimilhança para $(N, \boldsymbol{\theta})$.

De (2.5), segue que a função de verossimilhança é dada por

$$L(N, \boldsymbol{\theta}|n_{(1)}, ..., n_{(k)}, n) = P(n_1, ..., n_{l-1}, N - n|N, \boldsymbol{\theta}) =$$

$$= \frac{N!}{n_1!n_2!...n_{l-1}!(N-n)!} \prod_{j=1}^{k} \theta_j^{n_{(j)}} (1 - \theta_j)^{N - n_{(j)}}, \tag{2.6}$$

para $N \geq n$ e $0 < \theta_j < 1$, $1 \leq j \leq k$, onde $n_{(j)}$ é o número de erros detectados pelo j-ésimo revisor.

A estimativa de máxima verossimilhança (EMV) de $(N, \boldsymbol{\theta})$ é dada pelo seguinte teorema.

Teorema 1.

A estimativa de máxima verossimilhança de $(N, \boldsymbol{\theta})$, $(\hat{N}, \hat{\boldsymbol{\theta}})$, é tal que $\hat{N}$ é aproximadamente igual à solução da equação

$$1 - \frac{n}{N} = \prod_{j=1}^{k} \left(1 - \frac{n_{(j)}}{N}\right),$$

e $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_k)$, onde $\hat{\theta}_j = \frac{n_{(j)}}{\hat{N}}$, $j = 1, 2, ..., k$.

Prova. Como $(\hat{N}, \hat{\boldsymbol{\theta}})$ é o ponto de máximo de $\ln L(N, \boldsymbol{\theta}|n_{(1)}, ..., n_{(k)}, n)$, então $(\hat{N}, \hat{\boldsymbol{\theta}})$ é o ponto de máximo de $\ln K(N, \boldsymbol{\theta}|n_{(1)}, ..., n_{(k)}, n)$, onde $K(N, \boldsymbol{\theta}|n_{(1)}, ..., n_{(k)}, n)$ é o kernel de $L(N, \boldsymbol{\theta}|n_{(1)}, ..., n_{(k)}, n)$ dado por

$$K(N, \boldsymbol{\theta}|n_{(1)}, ..., n_{(k)}, n) = \frac{N!}{(N-n)!} \prod_{j=1}^{k} \theta_j^{n_{(j)}} (1 - \theta_j)^{N - n_{(j)}},$$

o que implica

$$\begin{aligned}\ln K(N,\boldsymbol{\theta}|n_{(1)},...,n_{(k)},n) &= \ln\frac{N!}{(N-n)!}+\ln\left[\prod_{j=1}^{k}\theta_j^{n_{(j)}}(1-\theta_j)^{N-n_{(j)}}\right]= \\ &= \ln\frac{N!}{(N-n)!}+\sum_{j=1}^{k}\left[n_{(j)}\ln\theta_j+(N-n_{(j)})\ln(1-\theta_j)\right]\end{aligned}$$

e

$$\begin{aligned}\frac{\partial\ln K(N,\boldsymbol{\theta}|n_{(1)},...,n_{(2)},n)}{\partial\theta_j} &= \frac{n_{(j)}}{\theta_j}-\frac{N-n_{(j)}}{1-\theta_j}=0\Longrightarrow \\ &\Longrightarrow\ \hat{\theta}_j=\frac{n_{(j)}}{N}, 1\leq j\leq k.\end{aligned}$$

Por outro lado, $\hat{N}$ é aproximadamente igual à solução da equação $K(N,\theta|n_{(1)},...,n_{(k)},n)=K(N-1,\theta|n_{(1)},...,n_{(k)},n)$, $N\geq n+1$. Então,

$$\frac{N!}{(N-n)!}\prod_{j=1}^{k}\theta_j^{n_{(j)}}(1-\theta_j)^{N-n_{(j)}}=\frac{(N-1)!}{(N-1-n)!}\prod_{j=1}^{k}\theta_j^{n_{(j)}}(1-\theta_j)^{N-1-n_{(j)}}, N\geq n+1\Longrightarrow$$

$$\begin{aligned}&\Longrightarrow\ \left(\frac{N}{N-n}\right)\prod_{j=1}^{k}(1-\theta_j)=1, N\geq n+1\Longrightarrow \\ &\Longrightarrow\ \prod_{j=1}^{k}(1-\theta_j)=\left(\frac{N-n}{N}\right), N\geq n+1\Longrightarrow \\ &\Longrightarrow\ 1-\frac{n}{N}=\prod_{j=1}^{k}(1-\theta_j), N\geq n+1,\end{aligned}$$

o que prova o teorema.■

Exemplo 2.

Para $k=2$, $n_{(12)}=n_{(1)}+n_{(2)}-n$ é o número de erros distintos detectados pelos revisores 1 e 2 simultaneamente e, pelo teorema 1, $\hat{N}$ é aproximadamente igual à solução da equação

$$1-\frac{n}{N}=\prod_{j=1}^{2}\left(1-\frac{n_{(j)}}{N}\right)=\left(1-\frac{n_{(1)}}{N}\right)\left(1-\frac{n_{(2)}}{N}\right),$$

o que implica

$$
\begin{aligned}
1-\frac{n}{N} &= \left(1-\frac{n_{(1)}}{N}\right)\left(1-\frac{n_{(2)}}{N}\right) \Longrightarrow \\
&\Longrightarrow N(N-n) = (N-n_{(1)})(N-n_{(2)}) \Longrightarrow \\
&\Longrightarrow N^2 - Nn = N^2 - N(n_{(1)}+n_{(2)}) + n_{(1)}n_{(2)} \Longrightarrow \\
&\Longrightarrow N(n_{(1)}+n_{(2)}-n) = n_{(1)}n_{(2)} \Longrightarrow Nn_{(12)} = n_{(1)}n_{(2)} \Longrightarrow \\
&\Longrightarrow \hat{N} \cong \frac{n_{(1)}n_{(2)}}{n_{(12)}}, \text{ se } n_{(12)} > 0,
\end{aligned}
$$

e $\hat{\theta}_1 = \frac{n_{(1)}}{\hat{N}} = \frac{n_{(12)}}{n_{(2)}}$, $\hat{\theta}_2 = \frac{n_{(2)}}{\hat{N}} = \frac{n_{(12)}}{n_{(1)}}$.

Exemplo 3.

A seguir aplicamos o método de estimação do exemplo 2 a um conjunto de dados reais disponíveis em Ciordia *et al.* (1999). Tais dados são referentes a um estudo realizado na cidade de Huesca, na região de Aragón, Espanha, durante 1995, 1996 e 1997, com o objetivo de estimar os números de casos de tuberculose nesta cidade. Denotamos tais números indistintamente por N. Foram consideradas duas fontes distintas de informação sobre o número de casos da doença.

• Fonte 1: sistema EDO, cujas informações foram obtidas das fontes atendimento primário, atendimento hospitalar (Medicina Preventiva) e entrevistas recebidas de outras cidades. Foram incluídos tantos os diagnósticos de certeza quantos os casos de suspeita da doença.

• Fonte 2: informes microbiológicos, que obtiveram informação do laboratório de microbiologia do Hospital San Jorge e consideraram apenas os diagnósticos de certeza da doença.

Ciordia *et al.*(1999) aplicou o método de captura-recaptura para duas fontes de informação (Sekar e Deming (1949)) e utilizou o sistema EPI-INFO 6 para estimar o número de casos de tuberculose em Huesca. Os dados obtidos pelas duas fontes de informação e

os resultados obtidos por Ciordia *et al.* (1999) seguem na tabela 1.

Tabela 1. Resultados obtidos por Ciordia *et al.* (1999).

Ano	Dados				Estimativas		
	$n_{(1)}$	$n_{(2)}$	$n_{(12)}$	n	N	θ_1	θ_2
1995	66	49	35	80	92	0,72	0,53
1996	73	49	37	85	96	0,76	0,51
1997	72	48	41	79	84	0,86	0,57

Dados e parâmetros.

$n_{(1)}$: número de casos registrados pela fonte 1;

$n_{(2)}$: número de casos registrados pela fonte 2;

$n_{(12)}$: número de casos registrados pelas fontes 1 e 2 simultaneamente;

n : número de casos distintos registrados;

θ_1 : capacidade de detecção de casos pela fonte 1;

θ_2 : capacidade de detecção de casos pela fonte 2 .

Neste exemplo, fazendo analogia com o problema de revisão de software, as duas fontes de informação são equivalentes a dois revisores e estimar o número de casos de tuberculose é equivalente a estimar o número total de erros no *software*, N. Apresentamos na tabela 2 os resultados obtidos após estimar N, θ_1 e θ_2 pelo método descrito no exemplo 2, ou seja, através das expressões

$$\begin{aligned} \hat{N} &\cong \frac{n_{(1)}n_{(2)}}{n_{(12)}}, \\ \hat{\theta}_1 &= \frac{n_{(1)}}{\hat{N}} = \frac{n_{(12)}}{n_{(2)}} \text{ e} \\ \hat{\theta}_2 &= \frac{n_{(2)}}{\hat{N}} = \frac{n_{(12)}}{n_{(1)}}. \end{aligned}$$

Tabela 2. Estimativas de N, θ_1 e θ_2.

Ano	Estimativas		
	N	θ_1	θ_2
1995	92,4	0,714	0,53
1996	96,67	0,755	0,506
1997	84,29	0,854	0,569

De acordo com as tabelas 1 e 2, obtivemos resultados análogos aos obtidos por Ciordia *et al.* (1999).

2.1.2 Modelo bayesiano.

Nesta seção fazemos uma análise bayesiana do modelo proposto na seção 2.1. Suponhamos que $\theta_1, \theta_2, ..., \theta_k$ sejam independentes, que θ_j tenha distribuição *a priori* Beta de parâmetros α_j e $\beta_j, \alpha_j > 0, \beta_j > 0, j = 1, 2, ..., k$, que N tenha distribuição *a priori* $\pi(N), N \geq 1$, e que N e $\boldsymbol{\theta}$ sejam independentes. Sejam $\boldsymbol{\alpha} = (\alpha_1, \alpha_2..., \alpha_k)$ e $\boldsymbol{\beta} = (\beta_1, \beta_2..., \beta_k)$.

Logo, a distribuição *a priori* conjunta para N e $\boldsymbol{\theta}$ é dada por

$$\pi(N, \boldsymbol{\theta}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \pi(N)\pi(\boldsymbol{\theta}|\boldsymbol{\alpha}, \boldsymbol{\beta}) \propto \pi(N) \prod_{j=1}^{k} \theta_j^{\alpha_j - 1}(1 - \theta_j)^{\beta_j - 1},$$

$N \geq n, 0 < \theta_j < 1, 1 \leq j \leq k$, e a distribuição *a posteriori* conjunta para N e $\boldsymbol{\theta}$ é dada por

$$\begin{aligned} \pi(N, \boldsymbol{\theta}|\boldsymbol{\alpha}, \boldsymbol{\beta}, n, n_{(1)}, n_{(2)}, ..., n_{(k)}) &\propto L(N, \boldsymbol{\theta}|n, n_{(1)}, n_{(2)}, ..., n_{(k)})\pi(N, \boldsymbol{\theta}|\boldsymbol{\alpha}, \boldsymbol{\beta}) \propto \\ &\propto \pi(N)\frac{N!}{(N-n)!}\prod_{j=1}^{k} \theta_j^{n_{(j)}}(1 - \theta_j)^{N - n_{(j)}} \times \qquad (2.7) \\ \times \prod_{j=1}^{k} \theta_j^{\alpha_j - 1}(1 - \theta_j)^{\beta_j - 1} &\propto \pi(N)\frac{N!}{(N-n)!}\prod_{j=1}^{k} \theta_j^{n_{(j)} + \alpha_j - 1}(1 - \theta_j)^{N - n_{(j)} + \beta_j - 1}, \end{aligned}$$

$N \geq n, 0 < \theta_j < 1, 1 \leq j \leq k$, ou seja,

$$\pi(N, \boldsymbol{\theta}|\boldsymbol{\alpha}, \boldsymbol{\beta}, n, n_{(1)}, n_{(2)}, ..., n_{(k)}) = C_1\pi(N)\frac{N!}{(N-n)!}\prod_{j=1}^{k} \theta_j^{n_{(j)} + \alpha_j - 1}(1 - \theta_j)^{N - n_{(j)} + \beta_j - 1},$$

onde a constante normalizadora C_1 é tal que

$$\begin{aligned} C_1^{-1} &= \sum_{N=n}^{\infty} \int_0^1 ... \int_0^1 \frac{N!}{(N-n)!}\pi(N) \prod_{j=1}^{k} \theta_j^{n_{(j)}+\alpha_j-1}(1-\theta_j)^{N-n_{(j)}+\beta_j-1} d\theta_1 ... d\theta_k = \\ &= \sum_{N=n}^{\infty} \frac{N!}{(N-n)!}\pi(N) \prod_{j=1}^{k} \int_0^1 \theta_j^{n_{(j)}+\alpha_j-1}(1-\theta_j)^{N-n_{(j)}+\beta_j-1} d\theta_j = \\ &= \sum_{N=n}^{\infty} \frac{N!}{(N-n)!}\pi(N) \prod_{j=1}^{k} \frac{\Gamma(n_{(j)}+\alpha_j)\Gamma(N-n_{(j)}+\beta_j)}{\Gamma(N+\alpha_j+\beta_j)} \propto \\ &\propto S_n, \end{aligned}$$

onde

$$S_n = \sum_{N=n}^{\infty} \frac{N!}{(N-n)!}\pi(N) \prod_{j=1}^{k} \frac{\Gamma(N-n_{(j)}+\beta_j)}{\Gamma(N+\alpha_j+\beta_j)}. \tag{2.8}$$

Teorema 2.

Suponhamos que a distribuição *a priori* para N seja da forma $\pi(N) = 1/N^r$, $N = 1, 2, ...; r = 0, 1$. Então, a distribuição *a posteriori* conjunta para N e $\boldsymbol{\theta}$ existe se e somente se $\sum_{j=1}^{k} n_{(j)} - n + \sum_{j=1}^{k} \alpha_j + r > 1$.

Prova. Suponhamos, sem perda de generalidade, $k = 2$. Logo, de (2.8) temos

$$S_n = \sum_{N=n}^{\infty} \frac{N!}{(N-n)!}\frac{1}{N^r} \prod_{j=1}^{2} \frac{\Gamma(N-n_{(j)}+\beta_j)}{\Gamma(N+\alpha_j+\beta_j)}. \tag{2.9}$$

Inicialmente, mostremos que

i) $\frac{N!}{(N-n)!N^n} \underset{N\to\infty}{\to} 1$ e ii) $\frac{\Gamma(N-n_{(j)}+\beta_j)N^{n_{(j)}+\alpha_j}}{\Gamma(N+\alpha_j+\beta_j)} \underset{N\to\infty}{\to} 1, j = 1, 2.$

Como

$$\begin{aligned} \frac{N!}{(N-n)!N^n} &= \frac{N(N-1)...(N-(n-1))}{N^n} = \\ &= 1\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)...\left(1-\frac{n-1}{N}\right) \underset{N\to\infty}{\to} 1, \end{aligned}$$

segue i) e como

$$\begin{aligned}\left(\frac{N-n_{(j)}+\beta_j}{N+\alpha_j+\beta_j}\right)^{N+\beta_j-\frac{1}{2}} &= \left(\frac{N-n_{(j)}+\beta_j+\alpha_j-\alpha_j}{N+\alpha_j+\beta_j}\right)^{N+\beta_j-\frac{1}{2}} = \\ &= \left(\frac{N+\alpha_j+\beta_j-n_{(j)}-\alpha_j}{N+\alpha_j+\beta_j}\right)^{N+\beta_j-\frac{1}{2}} = \\ &= \left(1+\frac{-n_{(j)}-\alpha_j}{N+\alpha_j+\beta_j}\right)^{N+\beta_j-\frac{1}{2}+\alpha_j-\alpha_j} =\end{aligned}$$

$$\begin{aligned}&= \left(1+\frac{-n_{(j)}-\alpha_j}{N+\alpha_j+\beta_j}\right)^{N+\alpha_j+\beta_j}\left(1+\frac{-n_{(j)}-\alpha_j}{N+\alpha_j+\beta_j}\right)^{-\frac{1}{2}-\alpha_j} \underset{N\to\infty}{\to} \exp\{-n_{(j)}-\alpha_j\}\times 1 = \\ &= \exp\{-n_{(j)}-\alpha_j\},\end{aligned}$$

$$\left(\frac{N}{N-n_{(j)}+\beta_j}\right)^{n_{(j)}} \underset{N\to\infty}{\to} 1,$$

$$\left(\frac{N}{N+\alpha_j+\beta_j}\right)^{\alpha_j} \underset{N\to\infty}{\to} 1$$

e o resultado

$$\frac{\Gamma(x)\exp\{x\}}{x^{x-\frac{1}{2}}} \underset{x\to\infty}{\to} \sqrt{2\pi}, \text{ (Feller(1967), pág. 66)} \tag{2.10}$$

implica

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

e

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

então,

$$\frac{\Gamma(N-n_{(j)}+\beta_j)N^{n_{(j)}+\alpha_j}}{\Gamma(N+\alpha_j+\beta_j)} =$$

$$
\begin{aligned}
= \quad & \frac{\Gamma(N-n_{(j)}+\beta_j)\exp\{N-n_{(j)}+\beta_j\}}{(N-n_{(j)}+\beta_j)^{N-n_{(j)}+\beta_j-\frac{1}{2}}}\frac{(N+\alpha_j+\beta_j)^{N+\alpha_j+\beta_j-\frac{1}{2}}}{\Gamma(N+\alpha_j+\beta_j)\exp\{N+\alpha_j+\beta_j\}}\times \\
& \times\left(\frac{N-n_{(j)}+\beta_j}{N+\alpha_j+\beta_j}\right)^{N+\beta_j-\frac{1}{2}}\left(\frac{N}{N-n_{(j)}+\beta_j}\right)^{n_{(j)}}\left(\frac{N}{N+\alpha_j+\beta_j}\right)^{\alpha_j}\times \\
& \times\exp\{n_{(j)}+\alpha_j\}\underset{N\to\infty}{\to}\sqrt{2\pi}\frac{1}{\sqrt{2\pi}}\exp\{-n_{(j)}-\alpha_j\}\exp\{n_{(j)}+\alpha_j\}= \\
= \quad & 1, j=1,2,
\end{aligned}
$$

o que prova ii).

Logo, de i) e ii), segue que

$$
N^{\sum_{j=1}^{2}(n_{(j)}+\alpha_j)-n}\frac{N!}{(N-n)!}\prod_{j=1}^{2}\frac{\Gamma(N-n_{(j)}+\beta_j)}{\Gamma(N+\alpha_j+\beta_j)}\underset{N\to\infty}{\to}1
$$

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

$$
1-\varepsilon<N^{\sum_{j=1}^{2}(n_{(j)}+\alpha_j)-n}\frac{N!}{(N-n)!}\prod_{j=1}^{2}\frac{\Gamma(N-n_{(j)}+\beta_j)}{\Gamma(N+\alpha_j+\beta_j)}<1+\varepsilon, \tag{2.11}
$$

para todo $N>N_0$ e S_n (dado em (2.9)) pode ser escrito como

$$
S_n=\sum_{N=n}^{N_o}\frac{N!}{(N-n)!}\frac{1}{N^r}\prod_{j=1}^{2}\frac{\Gamma(N-n_{(j)}+\beta_j)}{\Gamma(N+\alpha_j+\beta_j)}+V_{N_0},
$$

onde

$$
V_{N_0}=\sum_{N=N_o+1}^{\infty}\frac{N!}{(N-n)!}\frac{1}{N^r}\prod_{j=1}^{2}\frac{\Gamma(N-n_{(j)}+\beta_j)}{\Gamma(N+\alpha_j+\beta_j)}.
$$

Mas, de (2.11), segue que

$$
(1-\varepsilon)\sum_{N=N_o+1}^{\infty}N^{-\left[\sum_{j=1}^{2}(n_{(j)}+\alpha_j)-n+r\right]}<V_{N_0}<(1+\varepsilon)\sum_{N=N_o+1}^{\infty}N^{-\left[\sum_{j=1}^{2}(n_{(j)}+\alpha_j)-n+r\right]}
$$

o que implica o resultado, pois, de um lado,

$$
\sum_{j=1}^{2}n_{(j)}-n+\sum_{j=1}^{2}\alpha_j+r>1\Longrightarrow\sum_{N=N_o+1}^{\infty}N^{-\left[\sum_{j=1}^{2}(n_{(j)}+\alpha_j)-n+r\right]}<\infty\Longrightarrow
$$

$$\Longrightarrow \quad 0 < V_{N_0} < \infty \Longrightarrow 0 < S_n < \infty \Longrightarrow$$

$$\Longrightarrow \quad \text{existe a distribuição } \textit{a posteriori} \text{ conjunta de } N \text{ e } \boldsymbol{\theta} \text{ e, de outro lado,}$$

$$\sum_{j=1}^{k} n_{(j)} - n + \sum_{j=1}^{k} \alpha_j + r \leq 1 \Longrightarrow \sum_{N=N_o+1}^{\infty} N^{-\left[\sum_{j=1}^{2}(n_{(j)}+\alpha_j)-n+r\right]} \text{ é infinita} \Longrightarrow$$

$$\Longrightarrow \quad V_{N_0} \text{ infinita} \Longrightarrow S_n \text{ infinita} \Longrightarrow$$

$$\Longrightarrow \quad \text{não existe a distribuição } \textit{a posteriori} \text{ conjunta de } N \text{ e } \boldsymbol{\theta}.\blacksquare$$

Se a *priori* $\pi(N) = 1, N \geq 1$, segue de (2.7) que a distribuição *a posteriori* conjunta para N e $\boldsymbol{\theta}$ é dada por

$$\pi(N, \boldsymbol{\theta}|\boldsymbol{\alpha}, \boldsymbol{\beta}, n, n_{(1)}, n_{(2)}, ..., n_{(k)}) \propto \frac{N!}{(N-n)!} \prod_{j=1}^{k} \theta_j^{n_{(j)}+\alpha_j-1}(1-\theta_j)^{N-n_{(j)}+\beta_j-1} \qquad (2.12)$$

Por (2.12), a distribuição condicional de N, dados $\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}, n, n_{(1)}$, $n_{(2)}, ..., n_{(k)}$, é dada por

$$\pi(N|\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}, n, n_{(1)}, n_{(2)}, ..., n_{(k)}) \propto \binom{N}{n} \prod_{j=1}^{k}(1-\theta_j)^N = \binom{N}{n} \left[\prod_{j=1}^{k}(1-\theta_j)\right]^N,$$

$N \geq n$, ou seja,

$$\pi(N|\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}, n, n_{(1)}, n_{(2)}, ..., n_{(k)}) = C_2 \binom{N}{n} \left[\prod_{j=1}^{k}(1-\theta_j)\right]^N,$$

onde a constante normalizadora C_2 é tal que

$$\begin{aligned} C_2^{-1} &= \sum_{N\geq n}\binom{N}{n}\left[\prod_{j=1}^{k}(1-\theta_j)\right]^N = \sum_{s=0}^{\infty}\binom{s+n}{n}\left[\prod_{j=1}^{k}(1-\theta_j)\right]^{s+n} = \\ &= \left[\prod_{j=1}^{k}(1-\theta_j)\right]^n \sum_{s=0}^{\infty}\binom{s+n}{n}\left[\prod_{j=1}^{k}(1-\theta_j)\right]^s = \\ &= \left[\prod_{j=1}^{k}(1-\theta_j)\right]^n \sum_{s=0}^{\infty}\binom{-n-1}{s}\left[-\prod_{j=1}^{k}(1-\theta_j)\right]^s = \\ &= \left[\prod_{j=1}^{k}(1-\theta_j)\right]^n (1-\prod_{j=1}^{k}(1-\theta_j))^{-n-1}. \end{aligned}$$

Logo,

$$\pi(N|\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\beta},n,n_{(1)},n_{(2)},...,n_{(k)}) = \binom{N}{n}\left[\prod_{j=1}^{k}(1-\theta_j)\right]^{N-n}(1-\prod_{j=1}^{k}(1-\theta_j))^{n+1}, \quad (2.13)$$

isto é, a distribuição condicional de N, dados $\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\beta},n,n_{(1)},n_{(2)},...,n_{(k)}$, é igual a distribuição de uma variável aleatória $n+Y$, onde Y tem distribuição binomial negativa de parâmetros $n+1$ e $1-\prod_{j=1}^{k}(1-\theta_j)$.

De fato, se Y tiver distribuição binomial negativa com parâmetros $n+1$ e $1-\prod_{j=1}^{k}(1-\theta_j)$, então

$$\begin{aligned} P(Y+n &= y) = P(Y=y-n) = \\ &= \binom{y}{n}\left[\prod_{j=1}^{k}(1-\theta_j)\right]^{y-n}(1-\prod_{j=1}^{k}(1-\theta_j))^{n+1}, \end{aligned}$$

$y = n, n+1, ...$

Por outro lado, segue de (2.12) que θ_1, $\theta_2,...,$ θ_k, dados N, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ $,n,n_{(1)}$, $n_{(2)},...,n_{(k)}$, são independentes onde θ_j tem distribuição Beta com parâmetros $n_{(j)}+\alpha_j$ e

$N - n_{(j)} + \beta_j$, $j = 1, 2, ..., k$, isto é,

$$\begin{aligned}\pi(\boldsymbol{\theta}|N, \boldsymbol{\alpha}, \boldsymbol{\beta}, n, n_{(1)}, n_{(2)}, ..., n_{(k)}) &= \prod_{j=1}^{k} \pi(\theta_j|N, \boldsymbol{\alpha}, \boldsymbol{\beta}, n, n_{(1)}, n_{(2)}, ..., n_{(k)}) \propto \\ &\propto \prod_{j=1}^{k} \theta_j^{n_{(j)}+\alpha_j-1}(1-\theta_j)^{N-n_{(j)}+\beta_j-1}, \qquad (2.14)\end{aligned}$$

$0 < \theta_j < 1$, $j = 1, 2, ..., k$, e a distribuição *a posteriori* marginal de N é dada por

$$\begin{aligned}\pi(N|\boldsymbol{\alpha}, \boldsymbol{\beta}, n, n_{(1)}, n_{(2)}, ..., n_{(k)}) &= \int_0^1 \int_0^1 ... \int_0^1 \pi(N, \boldsymbol{\theta}|\boldsymbol{\alpha}, \boldsymbol{\beta}, n, n_{(1)}, n_{(2)}, ..., n_{(k)})\, d\theta_1 d\theta_2 ... d\theta_k \propto \\ &\propto \binom{N}{n} \prod_{j=1}^{k} \int_0^1 \theta_j^{n_{(j)}+\alpha_j-1}(1-\theta_j)^{N-n_{(j)}+\beta_j-1}\, d\theta_j \propto \\ &\propto \binom{N}{n} \prod_{j=1}^{k} \frac{\Gamma(N - n_{(j)} + \beta_j)}{\Gamma(N + \alpha_j + \beta_j)}, \qquad (2.15)\end{aligned}$$

$N \geq n$.

Se a *priori* $\pi(N) = 1/N$, $N \geq 1$, segue de (2.7) que a distribuição *a posteriori* conjunta para N e $\boldsymbol{\theta}$ é dada por

$$\pi(N, \boldsymbol{\theta}|\boldsymbol{\alpha}, \boldsymbol{\beta}, n, n_{(1)}, n_{(2)}, ..., n_{(k)}) \propto \frac{(N-1)!}{(N-n)!} \prod_{j=1}^{k} \theta_j^{n_{(j)}+\alpha_j-1}(1-\theta_j)^{N-n_{(j)}+\beta_j-1}, \qquad (2.16)$$

De (2.16), segue que a distribuição condicional de N, dados $\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}, n, n_{(1)}$, $n_{(2)}, ..., n_{(k)}$, é dada por

$$\pi(N|\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}, n, n_{(1)}, n_{(2)}, ..., n_{(k)}) \propto \binom{N-1}{n-1} \left[\prod_{j=1}^{k}(1-\theta_j)\right]^N,$$

$N \geq n$.

Assim,

$$\pi(N|\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}, n, n_{(1)}, n_{(2)}, ..., n_{(k)}) = C_3 \binom{N-1}{n-1} \left[\prod_{j=1}^{k}(1-\theta_j)\right]^N,$$

onde a constante normalizadora C_3 é tal que

$$\begin{aligned} C_3^{-1} &= \sum_{N \geq n}^{\infty} \binom{N-1}{n-1} \left[\prod_{j=1}^{k}(1-\theta_j)\right]^N = \sum_{s=0}^{\infty} \binom{s+n-1}{n-1} \left[\prod_{j=1}^{k}(1-\theta_j)\right]^{s+n} = \\ &= \left[\prod_{j=1}^{k}(1-\theta_j)\right]^n \sum_{s=0}^{\infty} \binom{-n}{s} \left[-\prod_{j=1}^{k}(1-\theta_j)\right]^s = \\ &= \left[\prod_{j=1}^{k}(1-\theta_j)\right]^n (1-\prod_{j=1}^{k}(1-\theta_j))^{-n}, \end{aligned}$$

o que implica

$$\pi(N|\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}, n, n_{(1)}, n_{(2)}, ..., n_{(k)}) = \binom{N-1}{n-1}(1-\prod_{j=1}^{k}(1-\theta_j))^n \left[\prod_{j=1}^{k}(1-\theta_j)\right]^{N-n}, \quad (2.17)$$

$N \geq n$, isto é, a distribuição condicional de N, dados $\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}, n, n_{(1)}, n_{(2)}, ..., n_{(k)}$, é igual a distribuição de uma variável aleatória $n + Z$, onde Z tem distribuição binomial negativa de parâmetros n e $1 - \prod_{j=1}^{k}(1-\theta_j)$.

De fato, se Z tiver distribuição binomial negativa com parâmetros n e $1 - \prod_{j=1}^{k}(1-\theta_j)$, então

$$\begin{aligned} P(Z+n &= z) = P(Z = z-n) = \\ &= \binom{z-1}{n-1}(1-\prod_{j=1}^{k}(1-\theta_j))^n \left[\prod_{j=1}^{k}(1-\theta_j)\right]^{z-n}, \end{aligned}$$

$z = n, n+1, ...$

Como no caso da distribuição *a priori* uniforme para N, segue de (2.16) que θ_1, $\theta_2, ...,\ \theta_k$, dados N, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $n, n_{(1)}, n_{(2)}, ..., n_{(k)}$, são independentes onde θ_j tem distribuição

Beta com parâmetros $n_{(j)} + \alpha_j$ e $N - n_{(j)} + \beta_j$, $j = 1, 2, ..., k$, isto é,

$$\begin{aligned} \pi(\boldsymbol{\theta}|N, \boldsymbol{\alpha}, \boldsymbol{\beta}, n, n_{(1)}, n_{(2)}, ..., n_{(k)}) &= \prod_{j=1}^{k} \pi(\theta_j|N, \boldsymbol{\alpha}, \boldsymbol{\beta}, n, n_{(1)}, n_{(2)}, ..., n_{(k)}) \propto \\ &\propto \prod_{j=1}^{k} \theta_j^{n_{(j)}+\alpha_j-1}(1-\theta_j)^{N-n_{(j)}+\beta_j-1}, \end{aligned} \tag{2.18}$$

$0 < \theta_j < 1$, $j = 1, 2, ..., k$, e a distribuição *a posteriori* marginal para N é dada por

$$\begin{aligned} \pi(N|\boldsymbol{\alpha}, \boldsymbol{\beta}, n, n_{(1)}, n_{(2)}, ..., n_{(k)}) &= \int_0^1 \int_0^1 ... \int_0^1 \pi(N, \boldsymbol{\theta}|\boldsymbol{\alpha}, \boldsymbol{\beta}, n, n_{(1)}, n_{(2)}, ..., n_{(k)})\ d\theta_1 d\theta_2 ... d\theta_k \propto \\ &\propto \binom{N-1}{n-1} \prod_{j=1}^{k} \int_0^1 \theta_j^{n_{(j)}+\alpha_j-1}(1-\theta_j)^{N-n_{(j)}+\beta_j-1}\ d\theta_j \propto \\ &\propto \binom{N-1}{n-1} \prod_{j=1}^{k} \frac{\Gamma(N - n_{(j)} + \beta_j)}{\Gamma(N + \alpha_j + \beta_j)}, \end{aligned} \tag{2.19}$$

$N \geq n$.

2.1.3 Estimativas bayesianas de N.

Nesta seção utilizamos dois métodos para determinar as estimativas bayesianas do parâmetro de interesse N. O primeiro deles é um método baseado na determinação da distribuição *a posteriori* marginal quase exata de N. O segundo consiste na aplicação do algoritmo *Gibbs Sampling*.

Estimativas bayesianas de N via distribuição *a posteriori* marginal quase exata.

Suponhamos que a distribuição *a posteriori* marginal de N exista e seja tal que

$$\pi(N|dados) \propto f(N),$$

$N \geq n$, ou seja,

$$\pi(N|dados) = kf(N),$$

onde k é um número real não negativo e $N \geq n$.

Determinando

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

$N \geq n$, obtemos a relação recursiva

$$f(N+1) = f(N)A(N),$$

$N \geq n$. Logo,

$$\begin{aligned}
\pi(n|dados) &= kf(n) \\
\pi(n+1|dados) &= kf(n+1) = kf(n)A(n), \\
\pi(n+2|dados) &= kf(n+2) = kf(n+1)A(n+1), \\
&\ldots \\
\pi(n+s|dados) &= kf(n+s) = kf(n+s-1)A(n+s-1),
\end{aligned}$$

onde s é um número inteiro positivo tal que $\pi(n+s+1|dados) \simeq 0$, e somando membro a membro as relações de igualdade acima, temos

$$\sum_{i=0}^{s} \pi(n+i|dados) = k \sum_{i=0}^{s} f(n+i) \simeq 1,$$

o que implica

$$k \simeq \frac{1}{\sum_{i=0}^{s} f(n+i)}$$

e

$$\pi(N|dados) \simeq \frac{f(N)}{\sum_{i=0}^{s} f(n+i)}, \tag{2.20}$$

$N \geq n$. Além disso, temos a média *a posteriori* de N através da expressão

$$M = \sum_{i=0}^{s} (n+i)\pi(n+i|dados) \tag{2.21}$$

e o desvio padrão

$$D.P. = \sqrt{\sum_{i=0}^{s}[n+i-M]^2\pi(n+i|dados)}. \tag{2.22}$$

Esse método de resolução foi implementado no *software* MAPLE (versão 7.00) e no programa utilizado (anexado no apêndice A) são determinados os demais resumos aproximados *a posteriori* de N, como moda, intervalo de credibilidade e quartis.

A seguir, com o objetivo de analisar a performance do modelo proposto na seção 2.1, apresentamos os exemplos 4 e 5 com dados simulados através do *software R* (versão 1.9.0). Observe que nestes exemplos, apesar dos dados serem gerados através do *software* R, o método de obtenção das estimativas bayesianas de N via distribuição *a posteriori* marginal quase exata foi implementado no MAPLE.

Exemplo 4.

Neste exemplo atribuímos a N o valor 100, a k os valores 2 e 5 e às probabilidades θ_j valores maiores ou iguais a $0,5$ e baixos, em cada caso, e geramos os dados $n, n_{(1)}, ..., n_{(k)}$ de acordo com os passos a seguir.

1. Primeiramente determinamos n, o número de erros distintos detectados pelos k revisores. Para isto, para cada valor fixado de i, $i = 1, 2, ...N$, geramos valores de variáveis aleatórias independentes $X_{i1}, X_{i2}, ..., X_{ik}$, onde X_{ij} tem distribuição de Bernoulli com parâmetro θ_j, para cada $j, j = 1, 2, ..., k$. Para um erro i qualquer, lembrando que θ_j é a probabilidade do erro i ser detectado pelo revisor j, $j = 1, 2, ..., k$, temos

$$n = \sum_{i=1}^{N} I_{(\sum_{j=1}^{k} X_{ij}>0)}.$$

2. Para obter o número de erros detectados pelo revisor j, $n_{(j)}$, basta observar que

$$n_{(j)} = \sum_{i=1}^{N} X_{ij},$$

$j = 1, 2, ..., k$.

De posse dos dados, consideramos para N a distribuição uniforme nos inteiros não negativos e para θ_j distribuições *a priori* de referência (Bernardo (1979) e Smith (1991)), isto é, atribuímos para (α_j, β_j) os valores $(0,0), (0,1), (1,0)$; não informativas, ou seja, atribuí-

mos para (α_j, β_j) os valores $(1/2, 1/2), (1, 1)$ e informativas, isto é, atribuímos para (α_j, β_j) os valores $(5, 5), (10, 10), (10, 50), (50, 10), (50, 50), (100, 100)$, obtendo resumos aproximados *a posteriori* de N, 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.).

Observação: Por questão de simplificação dos cálculos, na programação deste caso utilizamos

$$f(N) = \frac{N!}{(N-n)!} \prod_{j=1}^{k} \frac{\Gamma(n_{(j)} + \alpha_j)\Gamma(N - n_{(j)} + \beta_j)}{\Gamma(N + \alpha_j + \beta_j)}$$

e não (2.15), como seria natural.

i) Considerando $k = 2, \theta_1 = 0,5$ e $\theta_2 = 0,6$, obtivemos os dados $n = 79, n_{(1)} = 48$ e $n_{(2)} = 54$. Os resultados obtidos seguem na tabela 3.

Tabela 3. Resumos aproximados *a posteriori* de N.

α_j	β_j	M	Moda	Q_1	Q_2	Q_3	D.P.	I.C(95%)	Ampl.I.C.
0	0	115,89	110	105	113	123	14,49	(93;149)	56
0	1	118,61	112	107	115	126	15,32	(94;154)	60
1	0	112,69	107	102	110	1119	13,13	(92;143)	51
0,5	0,5	115,51	110	104	112	122	14,17	(93;148)	55
1	1	115,16	109	104	112	122	13,87	(93;147)	54
5	5	113,07	109	103	111	119	20,42	(93;140)	47
10	10	111,49	108	103	109	117	10,87	(93;135)	42
10	50	180,22	172	161	176	195	25,81	(137;237)	100
50	10	87,97	87	84	87	89	3,76	(81;95)	14
50	50	107,84	106	101	106	112	7,84	(93;124)	31
100	100	106,89	106	101	105	110	7,04	(93;121)	28

De acordo com a tabela 3, verificamos que utilizando *priori*s de referência e não informativas para θ_j, obtivemos estimativas razoalmente próximas do verdadeiro valor de N e observamos uma pequena variabilidade entre elas. Porém, ao utilizarmos *priori*s informativas, verificamos uma grande variabilidade entre as estimativas de N. Logo, concluímos que as escolhas dos valores dos hiperparâmetros α_j e β_j influenciam as estimativas de N.

Em particular, para $\alpha_j = 10, \beta_j = 50$ e $\alpha_j = 50$ e $\beta_j = 10$, os respectivos intervalos de credibilidade não contiveram o verdadeiro valor do parâmetro N.

Apresentamos a seguir o gráfico da função de probabilidade aproximada *a posteriori* marginal de N considerando $\alpha_j = \beta_j = 100$.

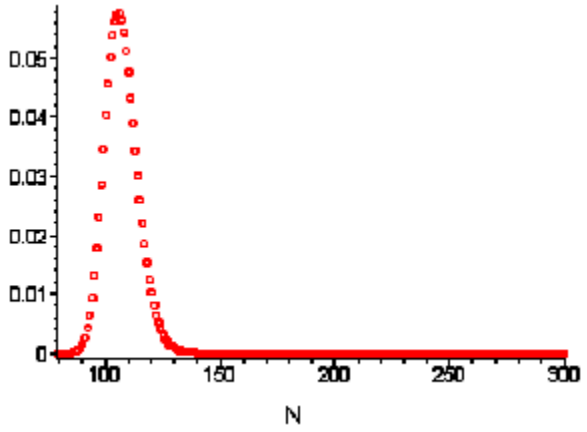

Fig.1 Gráfico da função de probabilidade aproximada *a posteriori* marginal de N.

ii) Considerando $k = 2, \theta_1 = 0,03, \theta_2 = 0,1$, obtivemos os dados $n = 11, n_{(1)} = 4$ e $n_{(2)} = 8$. Neste caso, para $(\alpha_j, \beta_j) = (0,0)$ e $(\alpha_j, \beta_j) = (0,1), j = 1,2$, a distribuição *a posteriori* marginal de N não existe, pois em ambos os casos, $n_{(1)} + n_{(2)} - n + \alpha_1 + \alpha_2 = 1$ (veja Teorema 2). Os resultados obtidos para outros valores de α_j e β_j seguem na tabela 4.

Tabela 4. Resumos aproximados *a posteriori* de N.

α_j	β_j	M	Moda	Q_1	Q_2	Q_3	D.P.	I.C(95%)	Ampl.I.C.
1	0	31,66	15	15	20	32	43,29	(11;112)	101
0,5	0,5	105,91	19	24	52	92	199,23	(12;687)	675
1	1	41,73	17	18	26	43	58,33	(12;157)	145
5	5	18,42	15	14	16	20	5,42	(11;31)	20
10	10	16,7	15	13	15	18	3,72	(11;25)	14
10	50	41,02	35	30	38	47	13,42	(20;72)	52
50	10	11,59	11	11	11	11	0,81	(11;13)	2
50	50	15,34	14	12	14	16	2,57	(11;20)	9
100	100	15,17	14	12	14	16	2,43	(11;20)	9

De acordo com a tabela 4 verificamos uma grande variabilidade das estimativas obtidas para N, com relação às *priori*s utilizadas. Isto nos mostra a influência da escolha dos valores dos hiperparâmetros α_j e β_j nas estimativas dos resumos *a posteriori* de N. É importante observar que as estimativas obtidas para N são ruins e, para a maioria dos valores de α_j e β_j, os intervalos de credibilidade correspondentes não contiveram o verdadeiro valor de N, o que poderia ser explicado pelos baixos valores dos dados.

iii) Considerando $k = 5$ e $\boldsymbol{\theta}$ =$(0,5; 0,6; 0,5; 0,5; 0,6)$, obtivemos os dados $n = 99$, $n_{(1)} = 51, n_{(2)} = 61, n_{(3)} = 48, n_{(4)} = 54$ e $n_{(5)} = 58$. Os resultados obtidos seguem na tabela 5.

Tabela 5. Resumos aproximados *a posteriori* de N.

α_j	β_j	M	Moda	Q_1	Q_2	Q_3	D.P.	I.C(95%)	Ampl.I.C.
0	0	101,1	100	99	100	101	1,56	(99;104)	5
0	1	101,24	101	99	100	101	1,62	(99;104)	5
1	0	100,98	100	99	100	101	1,51	(99;103)	4
0,5	0,5	101,11	100	99	100	101	1,57	(99;104)	5
1	1	101,12	100	99	100	101	1,57	(99;104)	5
5	5	101,2	101	99	100	101	1,6	(99;104)	5
10	10	101,28	101	99	100	101	1,62	(99;104)	5
10	50	108,95	108	105	108	110	3,7	(102;116)	14
50	10	99,48	99	99	99	99	0,7	(99;100)	1
50	50	101,66	101	99	100	102	1,71	(99;105)	6
100	100	101,85	101	100	101	102	1,75	(99;105)	6

De acordo com a tabela 5 obtivemos estimativas próximas do verdadeiro valor de N com relação às *priori*s utilizadas, exceto para $\alpha_j = 10$ e $\beta_j = 50$, cujo respectivo intervalo de credibilidade não conteve o verdadeiro valor de N. Além disso, se compararmos esta tabela com a tabela 3, podemos verificar uma melhora significativa das estimativas de N quando consideramos cinco revisores ao invés de dois.

iv) Considerando $k = 5$ e $\boldsymbol{\theta}$ =$(0,03; 0,1; 0,03; 0,03; 0,1)$, obtivemos os dados $n = 29, n_{(1)} = 6, n_{(2)} = 6, n_{(3)} = 6, n_{(4)} = 2$ e $n_{(5)} = 13$. Os resultados obtidos seguem na

tabela 6.

Tabela 6. Resumos aproximados *a posteriori* de N.

α_j	β_j	M	Moda	Q_1	Q_2	Q_3	D.P.	I.C(95%)	Ampl.I.C.
0	0	161,59	88	89	124	185	128,88	(54;493)	439
0	1	175,26	95	96	135	202	139,63	(58;539)	481
1	0	64,41	53	49	59	72	20,62	(38;115)	77
0,5	0,5	89,3	66	62	78	102	40,67	(44;192)	148
1	1	68,13	56	52	62	77	22,45	(39;124)	85
5	5	39,64	37	35	38	42	5,25	(31;51)	20
10	10	34,82	34	32	33	36	3,14	(29;41)	12
10	50	54,33	52	47	52	58	8,54	(39;73)	34
50	10	29,19	29	29	29	29	0,43	(29;30)	1
50	50	30,86	30	29	30	31	1,45	(29;33)	4
100	100	30,4	30	29	29	30	1,22	(29;32)	3

De acordo com a tabela 6 verificamos uma grande variabilidade das estimativas obtidas para N, com relação às *priori*s utilizadas. Isto nos mostra a influência da escolha dos valores dos hiperparâmetros α_j e β_j nas estimativas dos resumos *a posteriori* de N. É importante observar que as estimativas obtidas para N são ruins e, para a maioria dos valores de α_j e β_j, os intervalos de credibilidade correspondentes não contiveram o verdadeiro valor de N, o que já era esperado, pois atribuímos valores muito baixos para $\theta_1, \theta_2, ..., \theta_5$. Mesmo assim, se compararmos esta tabela com a tabela 4, podemos verificar uma melhora razoável das estimativas de N quando consideramos cinco revisores ao invés de dois.

Exemplo 5.

Neste exemplo atribuímos aos parâmetros do modelo os mesmos valores atribuídos no exemplo 4, bem como utilizamos as mesmas estatísticas. Adotamos para N a distribuição *a priori* de Jeffreys, ou seja, $\pi(N) = 1/N, N \geq 1$, e para θ_j as mesmas *prioris* do exemplo 3.

Observação: Por questão de simplificação dos cálculos, na programação deste caso

utilizamos

$$f(N) = \frac{(N-1)!}{(N-n)!} \prod_{j=1}^{k} \frac{\Gamma(n_{(j)} + \alpha_j)\Gamma(N - n_{(j)} + \beta_j)}{\Gamma(N + \alpha_j + \beta_j)}$$

e não (2.19), como seria natural.

i) Considerando $k = 2, \theta_1 = 0,5, \theta_2 = 0,6$, e os dados $n = 79, n_{(1)} = 48$ e $n_{(2)} = 54$, obtivemos os resultados da tabela 7.

Tabela 7. Resumos aproximados *a posteriori* de N.

α_j	β_j	M	Moda	Q_1	Q_2	Q_3	D.P.	I.C(95%)	Ampl.I.C.
0	0	114,39	109	103	111	121	13,78	(92;146)	54
0	1	116,8	111	105	114	124	14,56	(94;150)	56
1	0	111,27	106	105	108	1117	12,55	(91;140)	49
0,5	0,5	113,91	108	103	111	120	13,5	(92;145)	53
1	1	113,61	108	103	111	120	13,24	(92;144)	52
5	5	111,84	108	103	109	118	11,7	(92;138)	46
10	10	110,48	107	102	108	116	10,52	(92;133)	41
10	50	176,71	169	158	173	190	24,89	(135;232)	97
50	10	87,81	87	84	86	89	3,72	(81;95)	14
50	50	107,29	106	101	106	111	7,73	(93;123)	30
100	100	106,44	105	101	105	110	6,96	(93;120)	27

De acordo com a tabela 7, verificamos que utilizando *priori*s de referência e não informativas para θ_j, obtivemos estimativas para N razoalmente próximas daquelas obtidas na tabela 3.

Esta mesma situação ocorreu, como evidenciaram nossos cálculos, nos casos $k = 2, \theta_1 = 0,03$ e $\theta_2 = 0,1; k = 5$ e $\boldsymbol{\theta} = (0,5; 0,6; 0,5; 0,5; 0,6)$ e $k = 5$ e $\boldsymbol{\theta} = (0,03; 0,1; 0,03; 0,03; 0,1)$, onde as estimativas obtidas para N foram praticamente iguais as das tabelas 4, 5 e 6, respectivamente.

Dos exemplos 4 e 5 podemos concluir que, nestes casos, a utilização da *priori* uniforme nos inteiros não negativos para N é equivalente à utilização da *priori* de Jeffreys, ou seja, não existem diferenças significativas nos resumos aproximados *a posteriori* de N quando adotamos tais *priori*s.

Estimativas bayesianas de N via algoritmo *Gibbs Sampling*.

O algoritmo *Gibbs Sampling* para o modelo proposto implementa uma simulação estocástica que, utilizando cadeias de Markov e o conhecimento das distribuições condicionais de N, dados $\boldsymbol{\theta} = (\theta_1,\ \theta_2, ...\theta_k)$ e $\boldsymbol{\alpha}, \boldsymbol{\beta}, n, n_{(1)}, n_{(2)}, ..., n_{(k)}$, e de $\boldsymbol{\theta}$ dados N e $\boldsymbol{\alpha}, \boldsymbol{\beta}, n, n_{(1)}, n_{(2)}, ..., n_{(k)}$ (disponíveis sob formas conhecidas), gera uma amostra da distribuição *a posteriori* conjunta de N e $\boldsymbol{\theta}$ e conseqüentemente da distribuição *a posteriori* marginal de N.

O algoritmo *Gibbs Sampling* pode ser descrito da seguinte forma.

i) Considere um valor inicial do vetor $(N, \boldsymbol{\theta})$, $(N^{(0)}, \boldsymbol{\theta}^{(0)})$, e inicialize o contador de iterações da cadeia com $r = 1$.

ii) Obtenha um novo valor do vetor $(N, \boldsymbol{\theta})$, $(N^{(r)}, \boldsymbol{\theta}^{(r)})$, gerado de acordo com as distribuições condicionais

$$\begin{aligned} N^{(r)} &\sim \pi(N|\boldsymbol{\theta}^{(r-1)}, \boldsymbol{\alpha}, \boldsymbol{\beta}, n, n_{(1)}, n_{(2)}, ..., n_{(k)}), \\ &\text{e} \\ \boldsymbol{\theta}^{(r)} &\sim \pi(\boldsymbol{\theta}|N^{(r)}, \boldsymbol{\alpha}, \boldsymbol{\beta}, n, n_{(1)}, n_{(2)}, ..., n_{(k)}). \end{aligned}$$

iii) Altere o contador r para $r + 1$ e retorne a ii).

Após um grande número de iterações os valores gerados podem ser considerados como uma amostra da distribuição $\pi(N, \boldsymbol{\theta}|\boldsymbol{\alpha}, \boldsymbol{\beta}, n, n_{(1)}, n_{(2)}, ..., n_{(k)})$. Porém, nas primeiras iterações do algoritmo existe uma dependência devido ao valor inicial $(N^{(0)}, \boldsymbol{\theta}^{(0)})$. Assim, descartamos um certo número dos valores gerados (*burn in*) e para obter independência aproximada, devemos considerar saltos entre os elementos gerados da cadeia.

A seguir, a título de ilustração, aplicamos este algoritmo nos exemplos 6 e 7, onde os dados utilizados foram os do exemplo 4. Nestes exemplos, a convergência das cadeias geradas foi verificada utilizando a bibilioteca CODA - *Convergence Diagnostics and Ouput Analysis for Gibbs Sampling Output* (obtida via *software* R (versão 1.9.0)) que contém um conjunto de diagnósticos de convergência, dos quais utilizamos o diagnóstico de Gelman e Rubin. Esse diagnóstico baseia-se em técnicas de análise de variância, considerando que a convergência foi atingida quando a variância entre as cadeias for bem menor que a variância dentro de cada cadeia.

Além disso, em cada exemplo consideramos um *burn in* de 100 elementos, geramos duas cadeias com mil elementos cada e para garantir independência aproximada, tomamos elementos de 20 em 20. O programa utilizado para a implementação deste modelo foi feito via *software* R (versão 1.9.0) e está anexado no apêndice B.

Exemplo 6.

Neste exemplo atribuímos aos parâmetros do modelo os mesmos valores atribuídos no exemplo 4, bem como utilizamos as mesmas estatísticas. Adotamos a *priori* uniforme nos inteiros não negativos para N, para θ_j as mesmas *priori*s do exemplo 4 e implementamos o algoritmo *Gibbs Sampling* através das distribuições condicionais (2.13) e (2.14).

i) Considerando $k = 2, \theta_1 = 0,5, \theta_2 = 0,6$, e os dados $n = 79, n_{(1)} = 48$ e $n_{(2)} = 54$, obtivemos os resultados da tabela 8.

Tabela 8. Estimativas dos resumos *a posteriori* de N.

α_j	β_j	Média	Q_1	Q_2	Q_3	D.P	I.C(95%)	Ampl.I.C.	G-R
0	0	116,17	106	114	124	14,48	(95;149)	54	1
0	1	118,57	107	116	127	15,4	(95;154)	59	1
1	0	112,79	103	111	120	13,29	(93;145)	52	1
0,5	0,5	115,8	105	113	124	14,35	(94;150)	56	1
1	1	115,4	106	114	123	14,08	(94;149)	55	1,02
5	5	113,13	105	112	120	11,58	(94;139)	45	1,02
10	10	111,64	104	110	117	10,75	(94;135)	41	1,01
10	50	180,31	163	178	196	25,32	(137;236)	99	1,01
50	10	87,99	85	88	90	3,73	(82;96)	14	1
50	50	107,94	102	107	113	7,94	(95;126)	31	1
100	100	106,72	102	106	111	6,99	(94;122)	28	1,04

G.R.: valor retornado pelo diagnóstico de Gelman e Rubin (G.R.$<$1,1$\Longrightarrow$convergência)

Os resultados da tabela 8 praticamente são iguais aos da tabela 3, o que caracteriza, neste caso, a equivalência dos procedimentos *Gibbs Sampling* e distribuição *a posteriori* marginal quase exata.

Apresentamos a seguir alguns gráficos relativos a *posteriori* de N para $(\alpha_j, \beta_j) =$

$(100, 100)$.

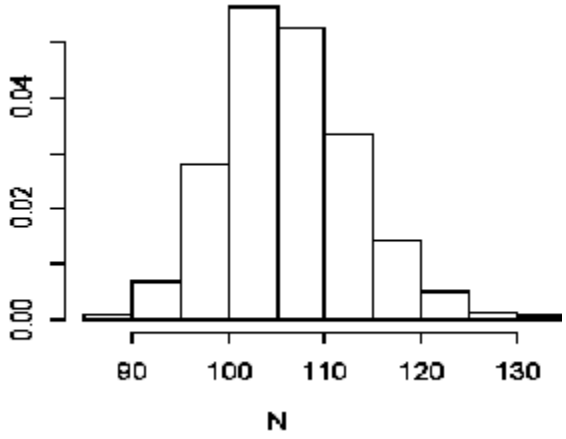

Fig. 2. Histograma da estimativa da distribuição *a posteriori* marginal de N .

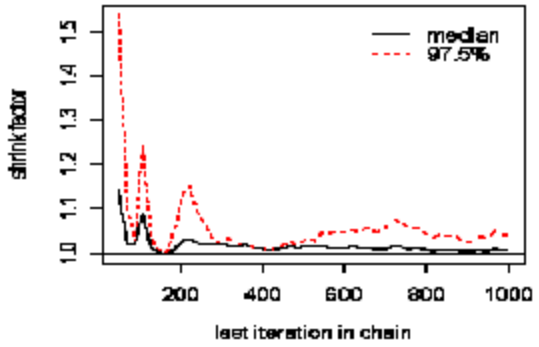

Fig. 3. Gráfico de convergência das cadeias de N (critério de Gelman-Rubin).

Pela fig. 3 temos a indicação da convergência das cadeias nas últimas iterações.

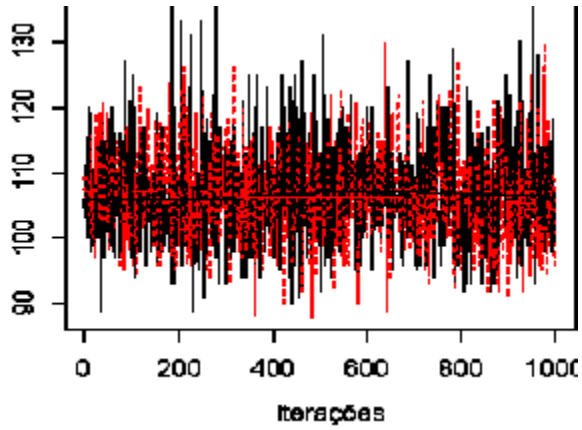

Fig. 4. Traço das cadeias de N.

Pela fig. 4 observamos uma certa uniformidade do traço para as duas cadeias, o que também indica convergência.

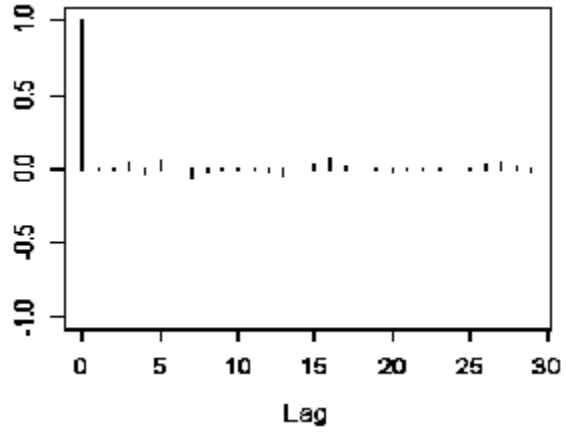

Fig. 5. Autocorrelação.

Pela fig. 5, a autocorrelação é praticamente nula, indicando independência aproximada entre os elementos da amostra.

ii) Considerando $k = 2, \theta_1 = 0,03, \theta_2 = 0,1$ e os dados $n = 11, n_{(1)} = 4$ e $n_{(2)} = 8$, obtivemos os resultados da tabela 9.

Tabela 9. Estimativas dos resumos *a posteriori* de N.

α	β	Média	Q_1	Q_2	Q_3	D.P	I.C(95%)	Ampl.I.C.	G-R
1	0	33,25	16	22	34	48,5	(12;106)	94	1,39
0,5	0,5	2814	25	43	95	8894	(14;95)	81	1,2
1	1	42,25	19	27	42	66,89	(13;42)	29	1,07
5	5	18,56	15	17	21	5,39	(12;32)	20	1
10	10	16,54	14	16	18	3,56	(12;25)	13	1,01
10	50	41,25	32	39	49	13,02	(22;73)	51	1,01
50	10	11,58	11	11	12	0,77	(11;13)	2	1,01
50	50	15,37	13	15	17	2,6	(12;22)	10	1,01
100	100	15,23	13	15	17	2,48	(11;21)	10	1,01

G.R.: valor retornado pelo diagnóstico de Gelman e Rubin (G.R.$<1,1 \Longrightarrow$convergência)

De acordo com a tabela 9 verificamos uma grande variabilidade das estimativas obtidas para N, com relação às *priori*s utilizadas. Isto nos mostra a influência da escolha dos valores dos hiperparâmetros α_j e β_j nas estimativas dos resumos *a posteriori*. É importante observar que as estimativas obtidas para N são ruins, o que já era esperado, pois atribuímos valores muito baixos para θ_1 e θ_2. Além disso, na maioria dos casos os intervalos de credibilidade não contiveram o verdadeiro valor de N.

Para os casos $k = 5$ e $\boldsymbol{\theta} = (0,5; 0,6; 0,5; 0,5; 0,6)$ e $k = 5$ e $\boldsymbol{\theta} = (0,03; 0,1; 0,03; 0,03; 0,1)$ os resultados obtidos foram próximos aos das tabelas 5 e 6, respectivamente.

Exemplo 7.

Neste exemplo atribuímos aos parâmetros do modelo os mesmos valores atribuídos no exemplo 4, bem como utilizamos as mesmas estatísticas. Adotamos a *priori* de Jeffreys para N, para θ_j as mesmas *priori*s do exemplo 4 e implementamos o algoritmo *Gibbs Sampling* através das distribuições condicionais (2.17) e (2.18).

i) Considerando $k = 2, \theta_1 = 0,5, \theta_2 = 0,6$, e os dados $n = 79, n_{(1)} = 48$ e $n_{(2)} = 54$, obtivemos os resultados da tabela 10.

Tabela 10. Estimativas dos resumos *a posteriori* de N.

α_j	β_j	Média	Q_1	Q_2	Q_3	D.P	I.C(95%)	Ampl.I.C.	G-R
0	0	114,39	104	112	122	13,96	(93;149)	56	1
0	1	117,08	107	115	125	14,52	(96;152)	56	1,01
1	0	111,14	102	109	118	12,52	(92;141)	49	1
0,5	0,5	114,33	105	112	122	13,6	(93;143)	50	1
1	1	113,90	105	112	121	12,94	(93;145)	52	1
5	5	111,83	104	110	119	11,49	(93;138)	45	1
10	10	110,68	103	109	117	10,28	(94;134)	40	1
10	50	177,15	159	174	192	25,5	(136;236)	100	1
50	10	87,92	85	87	90	3,76	(82;96)	14	1
50	50	107,36	102	107	112	7,89	(94;124)	30	1
100	100	106,75	102	106	111	6,95	(94;122)	28	1

G.R.: valor retornado pelo diagnóstico de Gelman e Rubin (G.R.<1,1$\Longrightarrow$convergência)

Os resultados da tabela 10 são próximos daqueles da tabela 3, o que evidencia, neste caso, a equivalência dos métodos "*Gibbs Sampling*"e "distribuição *a posteriori* marginal quase exata".Para os casos $k = 2, \theta_1 = 0,03$ e $\theta_2 = 0,1; k = 5$ e $\boldsymbol{\theta}$ =$(0,5; 0,6; 0,5; 0,5; 0,6)$ e $k = 5$ e $\boldsymbol{\theta}$ =$(0,03; 0,1; 0,03; 0,03; 0,1)$ os resultados obtidos foram próximos aos das tabelas 4, 5 e 6, respectivamente.

Pelos exemplos apresentados podemos concluir que a adoção da distribuição *a priori* uniforme nos inteiros não negativos para N é equivalente à adoção da *priori* de Jeffreys, o mesmo acontecendo com as metodologias "*Gibbs Sampling*"e "distribuição *a posteriori* marginal quase exata de N ".

2.1.4 Distribuição *a priori* informativa para as probabilidades de detecção dos erros.

Nesta seção, motivados pelo problema de estimação de N quando os valores de θ_j são muito baixos, conforme observado no exemplo 4 (ítens ii e iv), propomos um método de

escolha dos valores dos hiperparâmetros α_j e β_j, baseado no resultado de que se uma variável aleatória tem distribuição normal de média μ e variância σ^2, então cerca de 95% da massa desta distribuição está contida no intervalo $(\mu - 2\sigma, \mu + 2\sigma)$.

Agora, fazendo uma analogia com o nosso problema, suponhamos que, baseados em informações de especialistas, soubéssemos que a probabilidade de detecção de um erro por um certo revisor, θ_j, pertence a um intervalo (u_j, v_j) com probabilidade aproximada de 95%, $j = 1, 2, ..., k$. Esta informação é útil para escolher os valores dos hiperparâmetros α_j e β_j, pois, neste caso, motivados pelo resultado acima, seus valores são dados pela solução de um sistema onde, primeiramente igualamos o centro do intevalo (u_j, v_j) à média da distribuição *a priori* de θ_j e depois igualamos um quarto da amplitude do mesmo intervalo ao desvio padrão da distribuição *a priori* de θ_j (Coupal *et al.* (2000)). Então, fazendo

$$\begin{aligned} \frac{u_j + v_j}{2} &= \frac{\alpha_j}{\alpha_j + \beta_j} = \mu' = E(\theta_j) \text{ e} \\ \frac{v_j - u_j}{4} &= \left[\frac{\alpha_j \beta_j}{(\alpha_j + \beta_j)^2(\alpha_j + \beta_j + 1)}\right]^{1/2} = \sigma' = D.P.(\theta_j), \end{aligned} \tag{2.23}$$

$j = 1, 2, ..., k$, α_j e β_j são tais que $u_j = \mu' - 2\sigma'$ e $v_j = \mu' + 2\sigma', j = 1, 2, ..., k$.

No apêndice C implementamos a resolução deste sistema, bem como o cálculo de

$$\int_{u_j}^{v_j} \frac{\Gamma(\alpha_j + \beta_j)}{\Gamma(\alpha_j)\Gamma(\beta_j)} \theta_j^{\alpha_j - 1}(1 - \theta_j)^{\beta_j - 1} d\theta_j, \tag{2.24}$$

utilizando o *software* MAPLE (versão 7.00), para verificar se (2.24) tem valor próximo a 0,95.

A seguir apresentamos o exemplo 8 para ilustrar a aplicação do método apresentado para a escolha dos hiperparâmetros α_j e β_j. Neste exemplo estimamos N via algoritmo *Gibbs Sampling.*

Exemplo 8.

Neste exemplo atribuímos aos parâmetros do modelo os mesmos valores atribuídos no exemplo 4, bem como utilizamos as mesmas estatísticas. Adotamos para N a distribuição *a priori* uniforme nos inteiros não negativos, a *priori* de Jeffreys e implementamos o algoritmo *Gibbs Sampling* através das distribuições condicionais (2.13), (2.14), (2.17) e

(2.18).

i) Considerando $k = 2$, supusemos que, baseados em informações de especialistas, θ_1 pertencesse ao intervalo $(0; 0,06)$ e que θ_2 pertencesse ao intervalo $(0,08; 0,12)$, ambos com probabilidade aproximada de 95%. Resolvendo o sistema (2.23), obtivemos $\alpha_1 = 3,85, \beta_1 = 124,48, \alpha_2 = 89,9, \beta_2 = 809,1$. Após a aplicação do algoritmo *Gibbs Sampling* utilizando esses valores para os hiperparâmetros das *priori*s para $\theta_j, j = 1, 2$, obtivemos os resultados da tabela 11.

Tabela 11. Estimativas dos resumos *a posteriori* de N.

$\pi(N)$	Média	Q_1	Q_2	Q_3	D.P.	I.C.(95%)	Ampl. I.C.	G-R
Uniforme	92,16	72	90	109	27,67	(47;154)	107	1
Jeffreys	84,03	66	81	100	25,7	(42;143)	101	1,01

Pela tabela 11, verificamos que utilizando *priori*s informativas para θ_1 e θ_2, foi posssível corrigir a estimação de N quando os valores dessas probabilidades são muito baixos (veja tabela 4). A seguir apresentamos os gráficos das funções densidade Beta (α_1, β_1) e Beta (α_2, β_2), respectivamente.

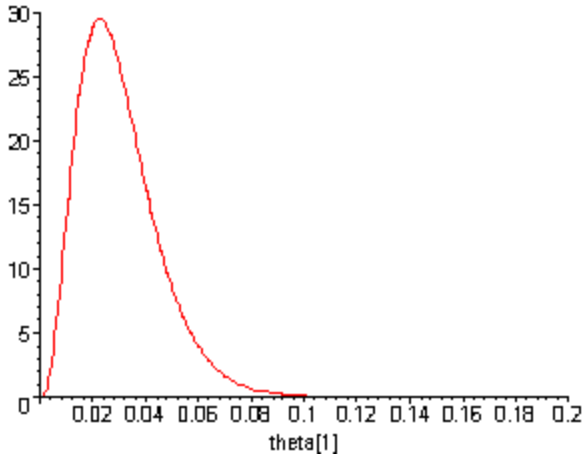

Fig.6 Função densidade Beta (3,85;124,48)

Note que aproximadamente 95% da masssa da função densidade Beta (α_1, β_1) está contida no intervalo $(0; 0,06)$, já que $\int_0^{0,06} \frac{\Gamma(\alpha_1+\beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)} \theta_1^{\alpha_1-1}(1-\theta_1)^{\beta_1-1} d\theta_1 = 0,957$.

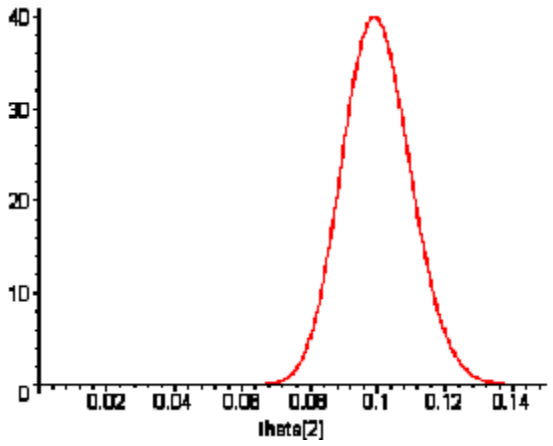

Fig.7 Função densidade Beta (89,9;809,1)

Note que aproximadamente 95% da masssa da função densidade Beta (α_2, β_2) está contida no intervalo $(0,08; 0,12)$, já que $\int_{0,08}^{0,12} \frac{\Gamma(\alpha_2+\beta_2)}{\Gamma(\alpha_2)\Gamma(\beta_2)} \theta_2^{\alpha_2-1}(1-\theta_2)^{\beta_2-1} d\theta_2 = 0,954$.

ii) Considerando $k = 5$, supusemos que, baseados em informações de especialistas, θ_1 pertencesse ao intervalo $(0; 0,06)$, θ_2 pertencesse ao intervalo $(0,08; 0,12)$, θ_3 pertencesse ao intervalo $(0; 0,1)$, θ_4 pertencesse ao intervalo $(0; 0,05)$ e que θ_5 pertencesse ao intervalo $(0,07; 0,15)$, todos com probabilidade aproximada de 95%. Resolvendo o sistema (2.23), obtivemos $\alpha_1 = 3,85, \beta_1 = 124,48, \alpha_2 = 89,9, \beta_2 = 809,1, \alpha_3 = 3,75, \beta_3 = 71,25$, $\alpha_4 = 3,87, \beta_4 = 151,12, \alpha_5 = 26,81, \beta_5 = 216,93$. Após a aplicação do algoritmo *Gibbs Sampling* utilizando esses valores para os hiperparâmetros das *priori*s para θ_j, $j = 1, 2, ..., k$, obtivemos os resultados da tabela 12.

Tabela 12. Estimativas dos resumos *a posteriori* de N.

$\pi(N)$	Média	Q_1	Q_2	Q_3	D.P.	I.C.(95%)	Ampl. I.C.	G-R
Uniforme	102,46	90	101	114	18,51	(71;143)	72	1
Jeffreys	98,87	86	98	111	18,22	(68;140)	72	1,02

Pela tabela 12, verificamos que utilizando *priori*s informativas para $\theta_j, j = 1, 2, 3, 4, 5$, foi posssível corrigir a estimação de N quando os valores dessas probabilidades são muito baixos (veja tabela 6). A seguir apresentamos os gráficos das funções densidade Beta (α_3, β_3), Beta (α_4, β_4) e Beta (α_5, β_5), respectivamente, pois os demais gráficos são idên-

ticos aos do item anterior.

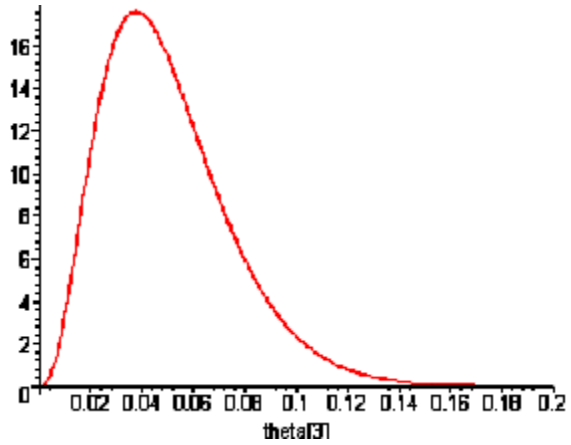

Fig.8 Função densidade Beta (3,75;71,25)

Note que aproximadamente 95% da masssa da função densidade Beta (α_3, β_3) está contida no intervalo $(0; 0,1)$, já que $\int_0^{0,1} \frac{\Gamma(\alpha_3+\beta_3)}{\Gamma(\alpha_3)\Gamma(\beta_3)} \theta_3^{\alpha_3-1}(1-\theta_3)^{\beta_3-1} d\theta_3 = 0,957$.

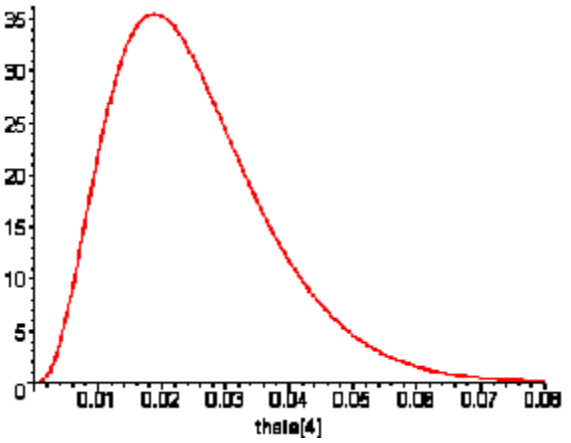

Fig.9 Função densidade Beta (3,87;151,12)

Note que aproximadamente 95% da masssa da função densidade Beta (α_4, β_4) está

contida no intervalo $(0; 0,05)$, já que $\int_0^{0,05} \frac{\Gamma(\alpha_4+\beta_4)}{\Gamma(\alpha_4)\Gamma(\beta_4)}\theta_4^{\alpha_4-1}(1-\theta_4)^{\beta_4-1}d\theta_4 = 0,957$.

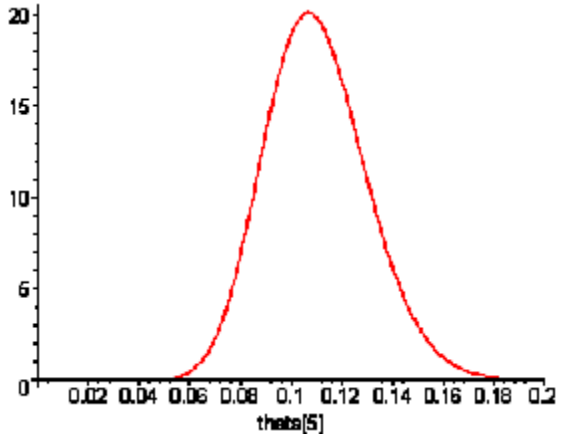

Fig.10 Função densidade Beta (26,81;216,93)

Note que aproximadamente 95% da masssa da função densidade Beta (α_5, β_5) está contida no intervalo $(0,07; 0,15)$, já que $\int_{0,07}^{0,15} \frac{\Gamma(\alpha_5+\beta_5)}{\Gamma(\alpha_5)\Gamma(\beta_5)}\theta_5^{\alpha_5-1}(1-\theta_5)^{\beta_5-1}d\theta_5 \simeq 0,956$.

2.1.5 Exemplo com dados reais.

Nesta seção aplicamos o algoritmo *Gibbs Sampling* para estimação do parâmetro de interesse N, utilizando um conjunto de dados reais disponíveis em Eick *et al.* (1993) e citados em Basu *et al.* (2001).

Exemplo 9.

Os dados utilizados neste exemplo são da AT&T 5 ESS, onde 6 revisores detectaram 43 erros distintos, sendo que o primeiro revisor detectou 25 erros, o segundo 3, o terceiro 4, o quarto 15, o quinto 7 e o sexto detectou 6 erros, ou seja, temos $n = 43, n_{(1)} = 25$, $n_{(2)} = 3, n_{(3)} = 4, n_{(4)} = 15, n_{(5)} = 7$ e $n_{(6)} = 6$. Apresentamos na tabela 13 as estimativas dos resumos *a posteriori* de N obtidos após a aplicação do algoritmo *Gibbs Sampling* e na tabela 14 as estimativas dos resumos *a posteriori* de N obtidas por Basu *et al.* (2001).

Tabela 13. Estimativas dos resumos *a posteriori* de N.

$\pi(N)$	α_j	β_j	Média	Q_1	Q_2	Q_3	D.P.	I.C.(95%)	Ampl. I.C.	G-R
Uniforme	0,5	0,5	65,13	58	64	70	10	(51;88)	37	1,01
Jeffreys	0,5	0,5	64,18	58	63	69	9,3	(50;86)	36	1
Uniforme	1	1	62,14	56	61	67	8,43	(49;81)	32	1
Jeffreys	1	1	61,15	55	60	66	8,19	(49;80)	31	1,02

Tabela 14. Estimativas dos resumos *a posteriori* de N (Basu *et al.* (2001))

$\pi(N)$	Média	Q_2	I.C.(95%)	Ampl. I.C.
Uniforme	62,805	61	(48;79)	31

Comparando as tabelas 13 e 14, observamos que nossos resultados são próximos aos obtidos por Basu *et al.* (2001) quando utilizamos *priori*s não informativas para θ_j, $j = 1, 2, ..., 6$.

2.2 Heterogeneidade e independência entre erros e homogeneidade e independência entre revisores.

Nesta seção vamos supor que os revisores atuam independentemente e são igualmente eficientes (homogêneos) e que os erros são independentes e não igualmente detectáveis (erros heterogêneos), ou seja, $\theta_{ij} = \theta_i, i = 1, 2, ..., N, j = 1, 2, ..., k$, e $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_N)$.

Sejam Y_i o número de vezes em que o erro i é detectado durante as inspeções feitas pelos k revisores, $i = 1, 2, ..., N$ e F_j o número de erros detectados exatamente j vezes durante o processo, $j = 0, 1, 2, ..., k$.

Então, dados N e $\boldsymbol{\theta}, Y_1, Y_2, ..., Y_N$ são variáveis aleatórias independentes, onde Y_i tem distribuição Binomial com parâmetros k e $\theta_i, i = 1, 2, ..., N$; e $n = \sum_{j=1}^{k} F_j$ é o número de erros distintos detectados. Note que $F_0 = N - n$ é o número de erros não detectados.

Logo, a função de verossimilhança é dada por

$$\begin{aligned} L(N,\boldsymbol{\theta}|f_1,f_2,...,f_k) &= P(F_1=f_1,F_2=f_2,...,F_k=f_k|N,\boldsymbol{\theta}) = \\ &= P\left(\bigcup_{(y_1,y_2,...,y_N)}(Y_1=y_1,Y_2=y_2,...,Y_N=y_N)|N,\boldsymbol{\theta}\right) = \\ &= \sum_{(y_1,y_2,...,y_N)}\prod_{i=1}^{N}\binom{k}{y_i}\theta_i^{y_i}(1-\theta_i)^{k-y_i}, \end{aligned} \tag{2.25}$$

$N \geq n$ e $0 < \theta_i < 1, i = 1,2,...,N$, onde $\sum_{(y_1,y_2,...,y_N)}$ se extende a todas as N-uplas $(y_1,y_2,...,y_N)$, tais que $N-n$ dos y_i's são iguais a "0"; f_1 iguais a "1"; f_2 iguais a "2";...; f_k iguais a "k".

2.2.1 Modelo bayesiano.

Nesta seção fazemos uma análise bayesiana do modelo proposto. Suponhamos a *priori* que, dado N, $\theta_1,\theta_2,...,\theta_N$ sejam independentes e identicamente distribuídos, com distribuição Beta de parâmetros α e $\beta, \alpha > 0, \beta > 0$, que N tenha distribuição *a priori* $\pi(N), N \geq 1$, e que N e $\boldsymbol{\theta}$ sejam independentes.

Logo, a distribuição *a priori* conjunta para N e $\boldsymbol{\theta}$ é dada por

$$\pi(N,\boldsymbol{\theta}|\alpha,\beta) = \pi(N)\pi(\boldsymbol{\theta}|N,\alpha,\beta) = \pi(N)\prod_{i=1}^{N}\frac{1}{B(\alpha,\beta)}\theta_i^{\alpha-1}(1-\theta_i)^{\beta-1},$$

$N \geq n$ e $0 < \theta_i < 1, i = 1,2,...,N$, e a distribuição *a posteriori* conjunta para N e $\boldsymbol{\theta}$ é dada por

$$\begin{aligned} \pi(N,\boldsymbol{\theta}|\alpha,\beta,f_1,f_2,...,f_k) &\propto L(N,\boldsymbol{\theta}|f_1,f_2,...,f_k)\pi(N,\boldsymbol{\theta}|\alpha,\beta) = \\ &= \pi(N)\sum_{(y_1,y_2,...,y_N)}\prod_{i=1}^{N}\frac{\binom{k}{y_i}}{B(\alpha,\beta)}\theta_i^{y_i+\alpha-1}(1-\theta_i)^{k-y_i+\beta-1}, \end{aligned} \tag{2.26}$$

$N \geq n$ e $0 < \theta_i < 1, i = 1,2,...,N$.

De (2.26), segue que a distribuição *a posteriori* marginal para N é

$$\pi(N|\alpha,\beta,f_1,f_2,...,f_k) = \int_0^1 ... \int_0^1 \pi(N,\boldsymbol{\theta}|\alpha,\beta,f_1,f_2,...,f_k)\ d\theta_1 d\theta_2 ... d\theta_N \propto$$

$$
\begin{aligned}
&= \pi(N) \sum_{(y_1,y_2,\ldots,y_N)} \prod_{i=1}^{N} \frac{\binom{k}{y_i}}{B(\alpha,\beta)} \int_0^1 \theta_i^{y_i+\alpha-1}(1-\theta_i)^{k-y_i+\beta-1} d\theta_i = \\
&= \pi(N) \sum_{(y_1,y_2,\ldots,y_N)} \prod_{i=1}^{N} \binom{k}{y_i} \left(\frac{B(y_i+\alpha, k-y_i+\beta)}{B(\alpha,\beta)} \right) = \\
&= \pi(N) \sum_{(y_1,y_2,\ldots,y_N)} \prod_{j=0}^{k} \left[\binom{k}{j} \left(\frac{B(j+\alpha, k-j+\beta)}{B(\alpha,\beta)} \right) \right]^{f_j} \propto \\
&\propto \pi(N) \binom{N}{n} \left(\frac{B(\alpha, k+\beta)}{B(\alpha,\beta)} \right)^{N-n} \prod_{j=1}^{k} \left(\frac{B(j+\alpha, k-j+\beta)}{B(\alpha,\beta)} \right)^{f_j} \\
&\propto \pi(N) \binom{N}{n} \left(\frac{B(\alpha, k+\beta)}{B(\alpha,\beta)} \right)^{N},
\end{aligned}
$$

$N \geq n$. Portanto,

$$
\pi(N|\alpha,\beta,n) = C_4 \pi(N) \binom{N}{n} \left(\frac{B(\alpha, k+\beta)}{B(\alpha,\beta)} \right)^{N},
$$

onde a constante normalizadora C_4 é tal que

$$
C_4^{-1} = \sum_{N \geq n} \pi(N) \binom{N}{n} \left(\frac{B(\alpha, k+\beta)}{B(\alpha,\beta)} \right)^{N}. \tag{2.27}
$$

Se a *priori* $\pi(N) = 1$, $N \geq 1$, segue de (2.27) que

$$
\begin{aligned}
C_4^{-1} &= \sum_{N \geq n} \binom{N}{n} \left(\frac{B(\alpha, k+\beta)}{B(\alpha,\beta)} \right)^{N} = \sum_{s=0}^{\infty} \binom{s+n}{n} \left(\frac{B(\alpha, k+\beta)}{B(\alpha,\beta)} \right)^{s+n} = \\
&= \left(\frac{B(\alpha, k+\beta)}{B(\alpha,\beta)} \right)^{n} \sum_{s=0}^{\infty} \binom{s+n}{n} \left(\frac{B(\alpha, k+\beta)}{B(\alpha,\beta)} \right)^{s} = \\
&= \left(\frac{B(\alpha, k+\beta)}{B(\alpha,\beta)} \right)^{n} \sum_{s=0}^{\infty} \binom{-n-1}{s} \left(-\frac{B(\alpha, k+\beta)}{B(\alpha,\beta)} \right)^{s} = \\
&= \left(\frac{B(\alpha, k+\beta)}{B(\alpha,\beta)} \right)^{n} \left(1 - \frac{B(\alpha, k+\beta)}{B(\alpha,\beta)} \right)^{-n-1}.
\end{aligned}
$$

Logo,

$$
\pi(N|\alpha,\beta,n) = \binom{N}{n} \left(\frac{B(\alpha, k+\beta)}{B(\alpha,\beta)} \right)^{N-n} \left(1 - \frac{B(\alpha, k+\beta)}{B(\alpha,\beta)} \right)^{n+1}, \tag{2.28}
$$

$N \geq n$, isto é, a distribuição *a posteriori* marginal de N é igual a distribuição de uma variável aleatória $n+W$, onde W tem distribuição binomial negativa de parâmetros $n+1$ e $1-\frac{B(\alpha,k+\beta)}{B(\alpha,\beta)}$.

De fato, se W tiver distribuição binomial negativa com parâmetros $n+1$ e $1-\frac{B(\alpha,k+\beta)}{B(\alpha,\beta)}$, então

$$\begin{aligned} P(W+n &= w) = P(W = w-n) = \\ &= \binom{w}{n}\left(\frac{B(\alpha,k+\beta)}{B(\alpha,\beta)}\right)^{w-n}\left(1-\frac{B(\alpha,k+\beta)}{B(\alpha,\beta)}\right)^{n+1}, \end{aligned}$$

$w = n, n+1, ...$

Além disso, temos que

$$E(N|\alpha,\beta,n) = E(n+W) = n + \frac{(n+1)\frac{B(\alpha,k+\beta)}{B(\alpha,\beta)}}{1-\frac{B(\alpha,k+\beta)}{B(\alpha,\beta)}} = n + \frac{n+1}{\frac{B(\alpha,\beta)}{B(\alpha,k+\beta)}-1} \tag{2.29}$$

e

$$\sigma_{N|\alpha,\beta,n} = \sqrt{Var(N|\alpha,\beta,n)} = \sqrt{Var(n+W)} = \sqrt{Var(W)} = \frac{\sqrt{(n+1)\frac{B(\alpha,k+\beta)}{B(\alpha,\beta)}}}{1-\frac{B(\alpha,k+\beta)}{B(\alpha,\beta)}}. \tag{2.30}$$

Se a *priori* $\pi(N) = 1/N$, $N \geq 1$, segue de (2.27) que

$$\begin{aligned} C_4^{-1} &= \sum_{N\geq n}\frac{1}{N}\binom{N}{n}\left(\frac{B(\alpha,k+\beta)}{B(\alpha,\beta)}\right)^N = \frac{1}{n}\sum_{N\geq n}\binom{N-1}{n-1}\left(\frac{B(\alpha,k+\beta)}{B(\alpha,\beta)}\right)^N = \\ &= \frac{1}{n}\sum_{s=0}^{\infty}\binom{s+n-1}{n-1}\left(\frac{B(\alpha,k+\beta)}{B(\alpha,\beta)}\right)^{s+n} = \\ &= \frac{1}{n}\left(\frac{B(\alpha,k+\beta)}{B(\alpha,\beta)}\right)^n\sum_{s=0}^{\infty}\binom{-n}{s}\left(-\frac{B(\alpha,k+\beta)}{B(\alpha,\beta)}\right)^s = \\ &= \frac{1}{n}\left(\frac{B(\alpha,k+\beta)}{B(\alpha,\beta)}\right)^n\left(1-\frac{B(\alpha,k+\beta)}{B(\alpha,\beta)}\right)^{-n}. \end{aligned}$$

Logo,

$$\pi(N|\alpha,\beta,n) = \binom{N-1}{n-1}\left(\frac{B(\alpha,k+\beta)}{B(\alpha,\beta)}\right)^{N-n}\left(1-\frac{B(\alpha,k+\beta)}{B(\alpha,\beta)}\right)^n, \tag{2.31}$$

$N \geq n$, isto é, a distribuição a posteriori marginal de N é igual a distribuição de uma variável aleatória $n + V$, onde V tem distribuição binomial negativa de parâmetros n e $1 - \frac{B(\alpha,k+\beta)}{B(\alpha,\beta)}$.

De fato, se V tiver distribuição binomial negativa com parâmetros n e $1 - \frac{B(\alpha,k+\beta)}{B(\alpha,\beta)}$, então

$$\begin{aligned} P(V+n &= v) = P(V = v-n) = \\ &= \binom{v-1}{n-1} \left(\frac{B(\alpha,k+\beta)}{B(\alpha,\beta)}\right)^{v-n} \left(1 - \frac{B(\alpha,k+\beta)}{B(\alpha,\beta)}\right)^{n}, \end{aligned}$$

$v = n, n+1, ...$

Além disso, temos que

$$E(N|\alpha,\beta,n) = E(n+V) = n + \frac{n\frac{B(\alpha,k+\beta)}{B(\alpha,\beta)}}{1 - \frac{B(\alpha,k+\beta)}{B(\alpha,\beta)}} = \frac{n}{1 - \frac{B(\alpha,k+\beta)}{B(\alpha,\beta)}} \tag{2.32}$$

e

$$\sigma_{N|\alpha,\beta,n} = \sqrt{Var(N|\alpha,\beta,n)} = \sqrt{Var(n+V)} = \sqrt{Var(V)} = \frac{\sqrt{n\frac{B(\alpha,k+\beta)}{B(\alpha,\beta)}}}{1 - \frac{B(\alpha,k+\beta)}{B(\alpha,\beta)}}. \tag{2.33}$$

2.2.2 Estimativas bayesianas de N.

Nesta seção, com o objetivo de analisar a performance do modelo proposto na seção 2.2, apresentamos os exemplos 10 e 11 com dados simulados através do *software R* (versão 1.9.0). O programa utilizado segue no apêndice D. Neste exemplos, apresentamos os resumos *a posteriori* de N como média e desvio padrão, determinados através de (2.29), (2.30), (2.32) e (2.33). Além disso, através de (2.28) e (2.31), obtivemos a moda, quartis e intervalo de credibilidade de 95% para N. Os cálculos de tais valores seguem no programa utilizado.

Exemplo 10.

Neste exemplo atribuímos a N o valor 100 e a k o valor 2.

Determinamos n, o número de erros distintos detectados pelos k revisores, de acordo com os passos a seguir.

1. Geramos valores das coordenadas do vetor de probabilidades $\boldsymbol{\theta} = (\theta_1, ..., \theta_N)$,

através de uma distribuição Beta (w, s), onde w e s são conhecidos.

2. Geramos valores de variáveis aleatórias independentes $Y_1, Y_2, ..., Y_N$, onde Y_i tem distribuição Binomial de parâmetros k e $\theta_i, i = 1, ..., N$. Logo,

$$n = \sum_{i=1}^{N} I_{(Y_i>0)},$$

e de posse do valor de n, consideramos para N a distribuição *a priori* uniforme nos inteiros não negativos e para θ_i distribuições a *priori* não informativas, ou seja, atribuímos para (α, β) os valores $(1/2, 1/2), (1, 1)$ e informativas, isto é, atribuímos para (α, β) os valores $(5, 5), (10, 10), (10, 50), (50, 10), (50, 50), (100, 100)$, obtendo os resumos *a posteriori* de N, como média, 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 $w = s = 1$, obtivemos $n = 70$. Os resultados obtidos seguem na tabela 15.

Tabela 15. Resumos *a posteriori* de N.

α	β	Média	Moda	Q_1	Q_2	Q_3	D.P.	I.C(95%)	Ampl.I.C.
0,5	0,5	112,6	111	106	111	117	8,25	(97;129)	32
1	1	105,5	105	99	104	109	7,29	(91;120)	29
5	5	96,62	96	91	95	100	6,04	(85;108)	23
10	10	95,19	94	90	94	98	5,84	(84;106)	22
10	50	233,11	230	216	231	247	23,19	(189;280)	91
50	10	72,2	72	71	72	75	1,5	(70;72)	2
50	50	93,98	93	89	93	97	5,66	(83;105)	22
100	100	93,82	93	89	93	96	5,64	(83;105)	22

De acordo com a tabela 15, verificamos que utilizando *priori*s não informativas para θ_i, obtivemos estimativas razoalmente próximas do verdadeiro valor de N e observamos uma pequena variabilidade entre elas. Porém, ao utilizarmos *priori*s informativas, verificamos uma grande variabilidade entre as estimativas de N. Logo, concluímos que as escolhas dos valores dos hiperparâmetros α e β influenciam as estimativas de N. Em particular,

para $\alpha = 10, \beta = 50$ e $\alpha = 50$ e $\beta = 10$, os respectivos intervalos de credibilidade não contiveram o verdadeiro valor do parâmetro N.

Apresentamos a seguir o gráfico da função de probabilidade a *posteriori* marginal de N considerando $\alpha = \beta = 1$.

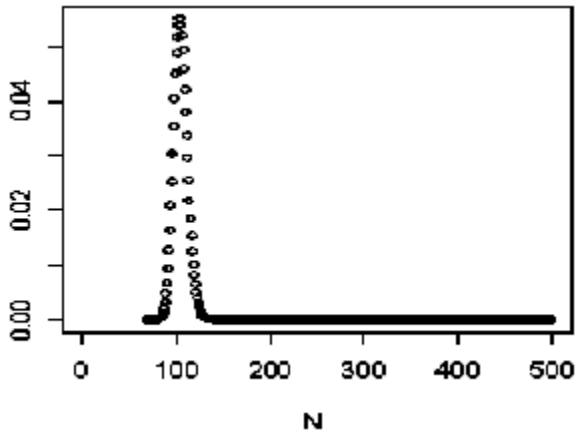

Fig. 11. Gráfico da função de probabilidade a *posteriori* marginal de N.

Observação: Vale ressaltar que refizemos este exemplo considerando $k = 5$ e $k = 8$, mas como os resultados obtidos foram, em sua maioria, análogos aos da tabela 15, optamos por omiti-los.

Exemplo 11.

Neste exemplo atribuímos aos parâmetros do modelo os mesmos valores atribuídos no exemplo 10, bem como utilizamos a mesma estatística n. Adotamos para N a distribuição *a priori* de Jeffreys e para θ_i as mesmas *priori*s do exemplo 10.

i) Considerando $w = s = 1$ e a estatística $n = 70$, obtivemos os resultados da tabela 16.

Tabela 16. Resumos *a posteriori* de N.

α	β	Média	Moda	Q_1	Q_2	Q_3	D.P.	I.C(95%)	Ampl.I.C.
0,5	0,5	112	111	105	111	116	8,19	(96;128)	32
1	1	105	104	99	104	109	7,24	(91;119)	28
5	5	96,25	95	91	95	99	6,01	(84;108)	24
10	10	94,83	94	90	94	98	5,8	(83;106)	23
10	50	230,81	228	214	229	243	23,02	(187;278)	91
50	10	72,16	72	70	71	72	1,49	(70;75)	5
50	50	93,64	93	89	92	96	5,62	(82;104)	22
100	100	93,48	93	89	92	96	5,6	(82;104)	22

De acordo com a tabela 16, verificamos que utilizando *priori*s não informativas para θ_i, obtivemos estimativas razoalmente próximas do verdadeiro valor de N e observamos uma pequena variabilidade entre elas. Porém, ao utilizarmos *priori*s informativas, verificamos uma grande variabilidade entre as estimativas de N. Logo, concluímos que as escolhas dos valores dos hiperparâmetros α e β influenciam as estimativas de N. Em particular, para $\alpha = 10, \beta = 50$ e $\alpha = 50$ e $\beta = 10$, os respectivos intervalos de credibilidade não contiveram o verdadeiro valor do parâmetro N. Por outro lado, pelos exemplos 10 e 11 podemos concluir que a adoção da distribuição *a priori* uniforme nos inteiros não negativos para N é equivalente à adoção da *priori* de Jeffreys.

Capítulo 3

Modelo estatístico para revisão de *software*: erros fáceis e difíceis de detectar.

Neste capítulo adotamos as suposições e notações da seção 2.2 e supomos que a heterogeneidade dos erros é expressa por uma classificação destes em dois tipos: fácil de detectar (que denotaremos por F) e difícil de detectar (que denotaremos por D), ou seja, $\boldsymbol{\theta}$ é um vetor cujos elementos são apenas de dois tipos: π_D ou π_F, onde π_D é a probabilidade de que um erro difícil seja detectado e π_F é a probabilidade de que um erro fácil seja detectado, $\pi_D \leq \pi_F$.

Em algumas revisões, os supervisores podem atribuir valores a π_D e π_F baseados em sua experiência. Assim, consideraremos π_D e π_F conhecidos. Dados N e α, $0 < \alpha < 1$, suponhamos $\theta_1, \theta_2, ..., \theta_N$ independentes e identicamente distribuídos com função de probabilidade

$$f_{\theta_i}(x|\alpha) = \alpha I_{\{\pi_D\}}(x) + (1-\alpha) I_{\{\pi_F\}}(x),$$

isto é,

$$f_{\theta_i}(x|\alpha) = \begin{cases} \alpha, \text{ se } x = \pi_D, \\ 1-\alpha, \text{ se } x = \pi_F, \\ 0, \text{ caso contrário,} \end{cases} \tag{3.1}$$

$i = 1, 2, ..., N$. Notemos que α é a proporção de erros difíceis no *software*.

Lembrando que f_j é o número de erros detectados exatamente j vezes durante o processo, $j = 1, 2, ..., k$, segue que a função de verossimilhança integrada sobre os $\theta_i's$, é dada por

$$\begin{aligned} L(N, \alpha | f_1, f_2, ..., f_k) &= P(f_1, f_2, ..., f_k | N, \alpha) = \int P(f_1, f_2, ..., f_k, \boldsymbol{\theta} | N, \alpha) d\boldsymbol{\theta} = \\ &= \int P(f_1, f_2, ..., f_k | \boldsymbol{\theta}, N) \pi(\boldsymbol{\theta} | N, \alpha) d\boldsymbol{\theta} = \\ &= \int L(N, \boldsymbol{\theta} | f_1, f_2, ..., f_k) \pi(\boldsymbol{\theta} | N, \alpha) d\boldsymbol{\theta}. \end{aligned}$$

Por (2.25) e (3.1) temos

$$\begin{aligned} L(N, \alpha | f_1, f_2, ..., f_k) &= \int \sum_{(y_1, y_2, ..., y_N)} \prod_{i=1}^{N} \binom{k}{y_i} \theta_i^{y_i} (1 - \theta_i)^{k - y_i} f_{\theta_i}(\theta_i | \alpha) d\boldsymbol{\theta} = \\ &= \sum_{(y_1, y_2, ..., y_N)} \prod_{i=1}^{N} \binom{k}{y_i} \int_0^1 \theta_i^{y_i} (1 - \theta_i)^{k - y_i} f_{\theta_i}(\theta_i | \alpha) d\theta_i = \\ &= \sum_{(y_1, y_2, ..., y_N)} \prod_{i=1}^{N} \binom{k}{y_i} \int_0^1 \theta_i^{y_i} (1 - \theta_i)^{k - y_i} [\alpha I_{\{\pi_D\}}(\theta_i) + (1 - \alpha) I_{\{\pi_F\}}(\theta_i)] d\theta_i = \\ &= \sum_{(y_1, y_2, ..., y_N)} \prod_{i=1}^{N} \binom{k}{y_i} [\alpha \pi_D^{y_i} (1 - \pi_D)^{k - y_i} + (1 - \alpha) \pi_F^{y_i} (1 - \pi_F)^{k - y_i}] = \end{aligned}$$

$$\begin{aligned} &= \sum_{(y_1, y_2, ..., y_N)} \prod_{y=0}^{k} \left\{ \binom{k}{y} \left[\alpha \pi_D^{y} (1 - \pi_D)^{k - y} + (1 - \alpha) \pi_F^{y} (1 - \pi_F)^{k - y} \right] \right\}^{f_y} = \\ &= \frac{N!}{f_1! f_2! ... f_k! (N - n)!} \prod_{y=0}^{k} \left\{ \binom{k}{y} \left[\alpha \pi_D^{y} (1 - \pi_D)^{k - y} + (1 - \alpha) \pi_F^{y} (1 - \pi_F)^{k - y} \right] \right\}^{f_y} \quad (3.2) \end{aligned}$$

$N \geq n$ e $0 < \alpha < 1$.

Fazendo $h(y; \alpha) = \binom{k}{y} \left[\alpha \pi_D^{y} (1 - \pi_D)^{k - y} + (1 - \alpha) \pi_F^{y} (1 - \pi_F)^{k - y} \right]$, podemos reescrever

(3.2) da seguinte forma:

$$\begin{aligned}
L(N,\alpha|f_1,f_2,...,f_k) &= \frac{N!}{f_1!f_2!...f_k!(N-n)!}h(0;\alpha)^{N-n}\prod_{y=1}^{k}[h(y;\alpha)]^{f_y} = \\
&= \binom{N}{n}\frac{n!}{f_1!f_2!...f_k!}(1-h(0;\alpha))^n h(0;\alpha)^{N-n}\frac{\prod_{y=1}^{k}[h(y;\alpha)]^{f_y}}{(1-h(0;\alpha))^n} = \\
&= \binom{N}{n}\frac{n!}{f_1!f_2!...f_k!}(1-h(0;\alpha))^n h(0;\alpha)^{N-n}\frac{\prod_{y=1}^{k}[h(y;\alpha)]^{f_y}}{(1-h(0;\alpha))^{\sum_{y=1}^{k} f_y}} = \\
&= \binom{N}{n}\frac{n!}{f_1!f_2!...f_k!}(1-h(0;\alpha))^n h(0;\alpha)^{N-n}\prod_{y=1}^{k}\left[\frac{h(y;\alpha)}{1-h(0;\alpha)}\right]^{f_y} = \\
&= \binom{N}{n}(1-h(0;\alpha))^n h(0;\alpha)^{N-n}\frac{n!}{f_1!f_2!...f_k!}\prod_{y=1}^{k}\left[\frac{h(y;\alpha)}{1-h(0;\alpha)}\right]^{f_y} = \\
&= L_1(N,\alpha|n)L_2(\alpha|f_1,f_2,...,f_k), \qquad (3.3)
\end{aligned}$$

onde

$$L_1(N,\alpha|n) = \binom{N}{n}(1-h(0;\alpha))^n h(0;\alpha)^{N-n}$$

e

$$L_2(\alpha|f_1,f_2,...,f_k) = \frac{n!}{f_1!f_2!...f_k!}\prod_{y=1}^{k}\left[\frac{h(y;\alpha)}{1-h(0;\alpha)}\right]^{f_y}, \qquad (3.4)$$

$N \geq n$ e $0 < \alpha < 1$.

Teorema 3.

i) $L_1(N,\alpha|n)$ é a função de verossimilhança correspondente à distribuição de probabilidades de n.

ii) $L_2(\alpha|f_1,f_2,...,f_k)$ é a função de verossimilhança correspondente à distribuição de probabilidades condicional de $(f_1,f_2,...,f_k)$, dado n.

Prova. i) Pelas relações (3.3) e (3.4) temos

$$\begin{aligned}
P(n|N,\alpha) &= \sum_{\substack{f_1,f_2,...,f_k \\ f_{1+}f_2+...+f_k=n}} P(f_1,f_2,...,f_k|N,\alpha) = \\
&= \sum_{\substack{f_1,f_2,...,f_k \\ f_{1+}f_2+...+f_k=n}} L(N,\alpha|f_1,f_2,...,f_k) = \\
&= L_1(N,\alpha|n) \sum_{\substack{f_1,f_2,...,f_k \\ f_{1+},f_2+...,+f_k=n}} L_2(\alpha|f_1,f_2,...,f_k) = \\
&= L_1(N,\alpha|n)\left[\underbrace{\sum_{y=1}^{k}\frac{h(y;\alpha)}{1-h(0;\alpha)}}_{=1}\right]^n = L_1(N,\alpha|n),
\end{aligned}$$

o que prova o item i).

ii) Pela relação (3.3) e o item (i), segue que

$$\begin{aligned}
P(f_1,f_2,...,f_k|n,N,\alpha) &= \frac{P(f_1,f_2,...,f_k|N,\alpha)}{P(n|N,\alpha)} = \\
&= \frac{L(N,\alpha|f_1,f_2,...,f_k)}{L_1(N,\alpha|n)} = L_2(\alpha|f_1,f_2,...,f_k), \qquad (3.5)
\end{aligned}$$

o que prova o item (ii).■

O resultado (3.3) sugere que os dados $(f_1, f_2, ..., f_k)$ podem ser considerados como resultado do seguinte experimento: dado α, correspondente a cada um dos erros, uma moeda com probabilidade de cara $1 - h(0;\alpha)$ é lançada, independentemente dos outros lançamentos, e são selecionados os erros correspondentes às ocorrências de caras. Em seguida, dado que foram selecionado n erros, gera-se uma realização do vetor $(f_1, f_2, ..., f_k)$ com distribuição de probabilidades Multinomial com parâmetros n e $(\frac{h(1;\alpha)}{1-h(0;\alpha)}, \frac{h(2;\alpha)}{1-h(0;\alpha)}, ..., \frac{h(k;\alpha)}{1-h(0;\alpha)})$.

3.1 Estimativas de N e α.

Nesta seção utilizamos dois métodos para estimação de N e α. O primeiro deles é chamado método da máxima verossimilhança condicional, proposto por Sanathanan (1972) e utilizado por Bolsoni (2002), e o outro é um método bayesiano que desenvolvemos de maneira

análoga ao de máxima verosssimilhança condicional.

3.1.1 Estimativas de máxima verossimilhança condicional de N e α.

O método da máxima verossimilhança condicional (MVC) para estimar N e α consiste em, primeiramente, determinar a estimativa de máxima verossimilhança de $\alpha, \hat{\alpha}_c$, correspondente à função de verossimilhança $L_2(\alpha|f_1, f_2, ..., f_k)$, dada em (3.4). Em seguida, determina-se a estimativa de MVC de $N, \hat{N}_c$, onde $\hat{N}_c$ é definido como sendo o ponto de máximo de $L_1(N, \hat{\alpha}_c|n)$ (veja (3.4)). Notemos que inicialmente é feita uma inferência condicional a respeito de α, baseada somente em $L_2(\alpha|f_1, f_2, ..., f_k)$ e então se infere sobre N, com base em $L_1(N, \alpha|n)$ com α substituído pela sua estimativa, $\hat{\alpha}_c$. Notemos que, como $L_2(\alpha|f_1, f_2, ..., f_k)$ envolve apenas parâmetro α, maximizar primeiramente $L_2(\alpha|f_1, f_2, ..., f_k)$ é mais simples do que maximizar $L_1(N, \alpha|n)$.

O Teorema 4 a seguir, cuja demonstração pode ser vista em Bolsoni (2002), por exemplo, indica uma forma direta de se obter $\hat{N}_c$.

Teorema 4. Uma vez determinado $\hat{\alpha}_c$, tem-se que:

i) se $\frac{n}{1-h(0;\hat{\alpha}_c)}$ não for um número inteiro, então

$$\hat{N}_c = \left[\frac{n}{1-h(0;\hat{\alpha}_c)}\right], \quad \left(\text{maior interio menor ou igual a } \frac{n}{1-h(0;\hat{\alpha}_c)}\right)$$

é o único ponto de máximo de $L_1(N, \hat{\alpha}_c|n)$;

ii) se $\frac{n}{1-h(0;\hat{\alpha}_c)}$ for um número inteiro, então

$$\hat{N}_c = \frac{n}{1-h(0;\hat{\alpha}_c)} \text{ e } \hat{N}_c - 1$$

são os pontos de máximo de $L_1(N, \hat{\alpha}_c|n)$.

Claramente $\hat{N}_c - n$ é a estimativa do número de erros não detectados do *software*.

Esse método foi implementado no *software* MAPLE (versão 7.00) e o programa utilizado segue em anexo no apêndice E.

A seguir, com o objetivo de analisar a performance do modelo proposto neste capítulo, apresentamos o exemplo 12 com dados simulados através do *software R* (versão 1.9.0).

Observe que neste exemplo, apesar dos dados serem gerados através do *software* R, o método de obtenção das estimativas de N e α foi implementado no MAPLE.

Exemplo 12.

Neste exemplo atribuímos a N o valor 100, a k os valores 2 e 5, a π_D o valor 0,3 e a π_F o valor 0,8.

Determinamos n e $(f_1, f_2, ..., f_k)$, de acordo com os passos a seguir (veja a interpretação de (3.3) após a prova do Teorema 3).

1. Atribuímos um valor para $\alpha, 0 < \alpha < 1$.

2. Para cada valor de i, $i = 1, 2, ...N$, geramos um valor de uma variável aleatória Z_i, onde Z_i tem distribuição de Bernoulli com parâmetro $1 - h(0; \alpha)$. Se observamos a interpretação de (3.3), notamos que este passo se refere ao fato de que correspondente a cada erro uma moeda com probabilidade de cara $1-h(0; \alpha)$ é lançada, independentemente dos outros lançamentos.

3. Determinamos o valor de n, o número de erros distintos detectados, através da expressão

$$n = \sum_{i=1}^{N} I_{(Z_i=1)},$$

que equivale à seleção dos erros correspondentes às ocorrências de caras.

4. De posse do valor de n, geramos um valor do vetor $(f_1, f_2, ..., f_k)$ com distribuição Multinomial de parâmetros n e $(\frac{h(1;\alpha)}{1-h(0;\alpha)}, \frac{h(2;\alpha)}{1-h(0;\alpha)}, ..., \frac{h(k;\alpha)}{1-h(0;\alpha)})$.

i) Considerando $k = 2$, obtivemos os resultados da tabela 17.

Tabela 17. Estimativas de MVC de N e α.

	Dados			Estimativas		
α (atribuído)	f_1	f_2	n	$\hat{\alpha}_c$	$\hat{N}_c$	$\hat{N}_c - n$
0,1	36	57	93	0,18	106	13
0,4	40	41	81	0,47	108	27
0,8	36	24	60	0,69	92	32

De acordo com a tabela 17, obtivemos, em geral, boas estimativas para N e α, isto é, as estimativas destes parâmetros estão próximas de seus valores verdadeiros. Além disso, notamos que a diferença $\hat{N}_c - n$ (número estimado de erros não detectados após a revisão

do *software*) assume o maior valor quando a proporção de erros difíceis no *software* é 0,8, o que é coerente com o esperado.

ii) Considerando $k = 5$, obtivemos os resultados da tabela 18.

Tabela 18. Estimativas de MVC de N e α.

	Dados						Estimativas		
α (atribuído)	f_1	f_2	f_3	f_4	f_5	n	$\hat{\alpha}_c$	$\hat{N}_c$	$\hat{N}_c - n$
0,1	8	9	16	39	25	97	0,18	100	3
0,4	20	16	13	25	15	89	0,49	97	8
0,8	26	20	15	14	10	85	0,69	96	11

De acordo com a tabela 18, obtivemos boas estimativas para N e α em todos os casos, isto é, em todos os casos as estimativas destes parâmetros estão próximas de seus valores verdadeiros. Além disso, notamos que a diferença $\hat{N}_c - n$ aumenta à medida que a proporção de erros difíceis no *software* aumenta, o que é coerente com o esperado. Podemos observar também que esta diferença sofreu uma diminuição com relação ao caso em que consideramos apenas 2 revisores, o que é um indicativo de melhor precisão do processo de revisão de *software* quando são adicionados mais 3 revisores ao processo.

3.1.2 Estimativas bayesianas de N e α.

Na seqüência desenvolvemos uma metodologia bayesiana para estimar N e α, adotando a mesma estratégia que norteou o desenvolvimento do método de máxima verossimilhança condicional, dado em 3.1.1.

A idéia é determinar, inicialmente, uma estimativa bayesiana de $\alpha, \hat{\alpha}_B$, baseada somente em uma distribuição *a priori* de α e na função de verossimilhança $L_2(\alpha|f_1, f_2, ..., f_k)$ (veja (3.4)). Em seguida, determina-se uma estimativa bayesiana de N baseada em uma distribuição *a priori* de N e na função de verossimilhança "estimada"$L_1(N, \hat{\alpha}_B|n)$ (veja (3.4)).

Suponhamos então que α tenha *a priori* distribuição Beta com parâmetros $\kappa\gamma$ e $\kappa(1-\gamma), \kappa > 0$ e $0 < \gamma < 1$, com κ e γ conhecidos. O hiperparâmetro γ representa a crença que temos sobre a proporção de erros difíceis, α, no *software* e κ representa o

quão fortemente acreditamos no valor de γ. Para entendermos melhor os significados dos hiperparâmetros γ e κ no problema basta lembrarmos que, *a priori*,

$$E(\alpha) = \frac{\kappa\gamma}{\kappa\gamma + \kappa(1-\gamma)} = \gamma \tag{3.6}$$

e

$$Var(\alpha) = \frac{\kappa^2\gamma(1-\gamma)}{[\kappa\gamma + \kappa(1-\gamma)]^2[\kappa\gamma + \kappa(1-\gamma) + 1]} = \frac{\gamma(1-\gamma)}{\kappa+1}. \tag{3.7}$$

Note que se acreditarmos que existem muitos erros difíceis no *software*, então atribuímos a α um valor alto. Conseqüentemente, devido à (3.6), γ também deverá assumir um valor alto. Fixado γ, se acreditarmos fortemente que esse valor está próximo da realidade, atribuímos a κ uma valor alto, a fim de tornarmos a variância de α pequena (veja (3.7)).

Utilizando a função de verossimilhança $L_2(\alpha|f_1, f_2, ..., f_k)$ dada em (3.4), segue que a distribuição *a posteriori* de α é tal que

$$\begin{aligned} \pi(\alpha|f_1, f_2, ..., f_k) &\propto L_2(\alpha|f_1, f_2, ..., f_k)\pi(\alpha) \propto \\ &\propto \alpha^{\kappa\gamma-1}(1-\alpha)^{\kappa(1-\gamma)-1}\prod_{y=1}^{k}\left[\frac{h(y;\alpha)}{1-h(0;\alpha)}\right]^{f_y}, \end{aligned}$$

$0 < \alpha < 1$, ou seja,

$$\pi(\alpha|f_1, f_2, ..., f_k) = C^{-1}\alpha^{\kappa\gamma-1}(1-\alpha)^{\kappa(1-\gamma)-1}\prod_{y=1}^{k}\left[\frac{h(y;\alpha)}{1-h(0;\alpha)}\right]^{f_y},$$

$0 < \alpha < 1$, onde

$$C = \int_0^1 \alpha^{\kappa\gamma-1}(1-\alpha)^{\kappa(1-\gamma)-1}\prod_{y=1}^{k}\left[\frac{h(y;\alpha)}{1-h(0;\alpha)}\right]^{f_y} d\alpha.$$

Os resumos da distribuição *a posteriori* de α, como média (M), desvio padrão ($D.P.$), quartis ($Q_j, j = 1, 2, 3$) e intervalo de credibilidade de 95% (I.C.(95%)) são dados, respec-

tivamente, por

$$\begin{aligned} \bullet M &= E(\alpha|f_1, f_2, ..., f_k) = \\ &= C^{-1} \int_0^1 \alpha^{\kappa\gamma}(1-\alpha)^{\kappa(1-\gamma)-1} \prod_{y=1}^{k} \left[\frac{h(y;\alpha)}{1-h(0;\alpha)}\right]^{f_y} d\alpha; \end{aligned} \tag{3.8}$$

$$\bullet D.P. = \sqrt{Var(\alpha|f_1, f_2, ..., f_k)} = \sqrt{E(\alpha^2|f_1, f_2, ..., f_k) - M^2}, \tag{3.9}$$

onde

$$E(\alpha^2|f_1, f_2, ..., f_k) = C^{-1} \int_0^1 \alpha^{\kappa\gamma+1}(1-\alpha)^{\kappa(1-\gamma)-1} \prod_{y=1}^{k} \left[\frac{h(y;\alpha)}{1-h(0;\alpha)}\right]^{f_y} d\alpha;$$

$$\bullet Q_1 = u, Q_2 = v, Q_3 = w, \tag{3.10}$$

onde u, v e w são números reais tais que

$$\begin{aligned} C^{-1} \int_0^u \alpha^{\kappa\gamma-1}(1-\alpha)^{\kappa(1-\gamma)-1} \prod_{y=1}^{k} \left[\frac{h(y;\alpha)}{1-h(0;\alpha)}\right]^{f_y} d\alpha &= 0,25, \\ C^{-1} \int_0^v \alpha^{\kappa\gamma-1}(1-\alpha)^{\kappa(1-\gamma)-1} \prod_{y=1}^{k} \left[\frac{h(y;\alpha)}{1-h(0;\alpha)}\right]^{f_y} d\alpha &= 0,5 \text{ e} \\ C^{-1} \int_0^w \alpha^{\kappa\gamma-1}(1-\alpha)^{\kappa(1-\gamma)-1} \prod_{y=1}^{k} \left[\frac{h(y;\alpha)}{1-h(0;\alpha)}\right]^{f_y} d\alpha &= 0,75; \end{aligned}$$

$$\bullet I.C.(95\%) = (a, b), \tag{3.11}$$

onde a e b são números reais tais que

$$\begin{aligned} C^{-1} \int_0^a \alpha^{\kappa\gamma-1}(1-\alpha)^{\kappa(1-\gamma)-1} \prod_{y=1}^{k} \left[\frac{h(y;\alpha)}{1-h(0;\alpha)}\right]^{f_y} d\alpha &= 0,025 \text{ e} \\ C^{-1} \int_0^b \alpha^{\kappa\gamma-1}(1-\alpha)^{\kappa(1-\gamma)-1} \prod_{y=1}^{k} \left[\frac{h(y;\alpha)}{1-h(0;\alpha)}\right]^{f_y} d\alpha &= 0,975. \end{aligned}$$

Em seguida, supondo que *a priori* N tenha distribuição de probabilidades $\pi(N)$, $N = 1, 2, ...$, e utilizando a "estimativa"da função de verossimilhança $L_1(N, \alpha|n)$, onde

substituímos α por M (média *a posteriori* de α), segue que a distribuição *a posteriori* de N é tal que

$$\begin{aligned} \pi(N|f_1, f_2, ..., f_k) &\propto L_1(N, M|n)\pi(N) \propto \\ &\propto \binom{N}{n} h(0; M)^N \pi(N), \end{aligned} \tag{3.12}$$

$N \geq n$.

Se *a priori* $\pi(N) = 1, N = 1, 2, ...$, então segue de (3.11) que a distribuição *a posteriori* de N é tal que

$$\pi(N|f_1, f_2, ..., f_k) \propto \binom{N}{n} h(0; M)^N,$$

$N \geq n$, o que implica, como na seção 2.2.1, que a distribuição *a posteriori* de N é igual a distribuição de uma variável aleatória $n + X$, onde X tem distribuição binomial negativa de parâmetros $n + 1$ e $1 - h(0, M)$. Conseqüentemente, temos que

$$E(N|f_1, f_2, ..., f_k) = E(n + X) = n + \frac{(n + 1)h(0; M)}{1 - h(0; M)} = \frac{n + h(0; M)}{1 - h(0; M)} \tag{3.13}$$

e

$$\begin{aligned} \sigma_{N|f_1,f_2,...,f_k} &= \sqrt{Var(N|f_1, f_2, ..., f_k)} = \sqrt{Var(n + X)} = \sqrt{Var(X)} = \\ &= \frac{\sqrt{(n + 1)h(0, M)}}{1 - h(0, M)}. \end{aligned} \tag{3.14}$$

Se *a priori* $\pi(N) = 1/N, N = 1, 2, ...$, então segue de (3.11) que a distribuição *a posteriori* de N é tal que

$$\pi(N|f_1, f_2, ..., f_k) \propto \binom{N - 1}{n - 1} [h(0; M)]^N,$$

$N \geq n$ e, como na seção 2.2.1, temos que a distribuição *a posteriori* de N é igual a distribuição de uma variável aleatória $n + Y$, onde Y tem distribuição binomial negativa de parâmetros n e $1 - h(0, M)$. Logo,

$$E(N|f_1, f_2, ..., f_k) = E(n + Y) = n + \frac{nh(0; M)}{1 - h(0; M)} = \frac{n}{1 - h(0; M)} \tag{3.15}$$

e

$$\begin{aligned} \sigma_{N|f_1,f_2,...,f_k} &= \sqrt{Var(N|f_1,f_2,...,f_k)} = \sqrt{Var(n+Y)} = \\ &= \sqrt{Var(Y)} = \frac{\sqrt{nh(0;M)}}{1-h(0;M)}. \end{aligned} \tag{3.16}$$

A seguir, a título de ilustração, apresentamos o exemplo 13, onde os dados utilizados foram simulados no exemplo 11. O programa utilizado na implementação deste método foi feito via *software* R (versão 1.9.0) e segue no apêndice F. Neste exemplo, apresentamos os resumos da distribuição *a posteriori* de α, como média, desvio padrão, quartis e intervalo de credibilidade de 95%, determinados pelas expressões (3.7), (3.8), (3.9) e (3.10), respectivamente. A distribuição *a priori* adotada para N é $\pi(N) = 1, N = 1, 2, ...$, e a média e desvio padrão *a posteriori* de N são determinados pelas expressões (3.12) e (3.13), respectivamente. Além disso, determinamos a moda, quartis e intervalo de credibilidade de 95% para N sabendo-se que a distribuição *a posteriori* de $N - n$ é binomial negativa.

Exemplo 13.

Neste exemplo atribuímos a N o valor 100, ao parâmetro α os valores $0,1; 0,4; 0,8$ e consideramos os mesmos dados amostrais gerados no exemplo 11. Atribuímos a α distribuições *a priori* não informativas, isto é, atribuímos a (κ, γ) o valor $(2; 0,5)$, e informativas, ou seja, atribuímos a (κ, γ) os valores $(2; 0,1), (2; 0,4)$ e $(2; 0,8)$. Adotamos para N a distribuição *a priori* uniforme nos inteiros não negativos, obtendo resumos aproximados das distribuições *a posteriori* de N e α, como média (M), moda (Mo), quartis ($Q_j, j = 1, 2, 3$), desvio padrão (D.P.), intervalo de credibilidade de 95% (I.C.(95%)) e sua amplitude (Ampl. I.C.).

i) Considerando $\alpha = 0,1$, $k = 2$ e os dados $f_1 = 36, f_2 = 57$ e $n = 93$, obtivemos os

resultados da tabela 19.

Tabela 19. Resumos das distribuições *a posteriori* de α e N.

(κ, γ)	Parâm.	M	Mo	Q_1	Q_2	Q_3	D.P.	I.C(95%)	Ampl.I.C.
$(2; 0, 5)$	α	0,21	0,18	0,11	0,2	0,29	0,12	(0,01;0,45)	0,44
	N	107,54	107	104	106	109	4.09	(99;115)	16
$(2; 0, 1)$	α	0,07	0,08	0,001	0,02	0,11	0,1	$(8\times10^{-9}$;0,34)	0,34
	N	100,4	100	97	99	101	2,82	(94;105)	11
$(2; 0, 4)$	α	0,18	0,14	0,08	0,17	0,26	0,12	(0,0001;0,43)	0,43
	N	106,2	106	102	105	108	3,88	(98;113)	15
$(2; 0, 8)$	α	0,26	0,26	0,17	0,26	0,34	0,11	(0,05;0,49)	0,44
	N	110,86	110	107	110	113	4,61	(102;120)	18

ii) Considerando $\alpha = 0,1$, $k = 5$ e os dados $f_1 = 8, f_2 = 9, f_3 = 16, f_4 = 39, f_5 = 25$ e $n = 97$, obtivemos os resultados da tabela 20.

Tabela 20. Resumos distribuições *a posteriori* de α e N.

(κ, γ)	Parâm.	M	Mo	Q_1	Q_2	Q_3	D.P.	I.C(95%)	Ampl.I.C.
$(2; 0, 5)$	α	0,19	0,18	0,15	0,18	0,22	0,05	(0,09;0,3)	0,21
	N	100,2	100	98	99	100	1,84	(97;103)	6
$(2; 0, 1)$	α	0,17	0,16	0,13	0,17	0,21	0,05	(0,08;0,28)	0,2
	N	100,02	99	98	99	100	1,76	(97;103)	6
$(2; 0, 4)$	α	0,18	0,17	0,15	0,18	0,22	0,05	(0,09;0,29)	0,2
	N	100,22	100	98	99	100	1,82	(97;103)	6
$(2; 0, 8)$	α	0,2	0,19	0,16	0,2	0,23	0,05	(0,1;0,31)	0,21
	N	100,48	100	98	99	101	1,89	(97;104)	7

Pela análise das tabelas 19 e 20 concluímos que, para $k = 2$ e $k = 5$, as estimativas bayesianas de α, como média, moda e mediana *a posteriori* são boas para a maioria dos valores atribuídos a (κ, γ). Particularmente, todos os intervalos de credibilidade contiveram o verdadeiro valor de α. Com relação ao parâmetro N, média, moda e mediana *a*

posteriori são boas estimativas para todos os valores atribuídos a (κ, γ). Mas para $k = 2$ e $(\kappa, \gamma) = (2, 0, 8)$, o respectivo intervalo de credibilidade não conteve o verdadeiro valor de N. Já para $k = 5$, todos os intervalos de credibilidade contiveram o verdadeiro valor do parâmetro. Observamos que ao considerarmos $k = 5$ obtivemos estimativas para N mais próximas do seu valor verdadeiro do que quando consideramos $k = 2$ e intervalos de credibilidade de menor amplitude.

iii) Considerando $\alpha = 0,4$, $k = 2$ e os dados $f_1 = 40, f_2 = 41$ e $n = 81$, obtivemos os resultados da tabela 21.

Tabela 21. Resumos das distribuições *a posteriori* de α e N.

(κ, γ)	Parâm.	M	Mo	Q_1	Q_2	Q_3	D.P.	I.C(95%)	Ampl.I.C.
$(2; 0, 5)$	α	0,44	0,47	0,36	0,45	0,54	0,13	(0,16;0,67)	0,51
	N	107,23	106	102	106	110	5,88	(96;119)	23
$(2; 0, 1)$	α	0,35	0,41	0,24	0,37	0,47	0,16	(0,00001;0,63)	0,63
	N	101,34	101	97	100	104	5,03	(91;111)	20
$(2; 0, 4)$	α	0,43	0,46	0,34	0,44	0,52	0,13	(0,13;0,67)	0,54
	N	106,18	105	101	105	109	5,73	(95;117)	22
$(2; 0, 8)$	α	0,49	0,51	0,41	0,49	0,57	0,12	(0,23;0,7)	0,47
	N	109,98	109	105	109	113	6,26	(98;122)	24

iv) Considerando $\alpha = 0,4$, $k = 5$ e os dados $f_1 = 20, f_2 = 16, f_3 = 13, f_4 = 25, f_5 = 15$

e $n = 89$, obtivemos os resultados da tabela 22.

Tabela 22. Resumos das distribuições a posteriori de α e N.

(κ, γ)	Parâm.	M	Mo	Q_1	Q_2	Q_3	D.P.	I.C(95%)	Ampl.I.C.
$(2; 0, 5)$	α	0,49	0,49	0,45	0,49	0,53	0,06	(0,37;0,61)	0,24
	N	97,19	97	94	96	98	2,99	(91;103)	12
$(2; 0, 1)$	α	0,48	0,48	0,43	0,48	0,52	0,06	(0,35;0,6)	0,25
	N	96,96	96	94	96	98	2,94	(91;102)	11
$(2; 0, 4)$	α	0,49	0,49	0,44	0,49	0,53	0,06	(0,36;0,61)	0,25
	N	97,13	97	94	96	98	2,97	(91;102)	11
$(2; 0, 8)$	α	0,5	0,5	0,46	0,5	0,54	0,06	(0,38;0,62)	0,24
	N	97,36	97	94	96	98	3,02	(91;103)	12

De acordo com as tabelas 21 e 22 concluímos que, para $k = 2$ e $k = 5$, as estimativas bayesianas de α, média, moda e mediana *a posteriori* são boas para todos os valores atribuídos a (κ, γ). Além disso, todos os intervalos de credibilidade contiveram o verdadeiro valor de α. Com relação ao parâmetro N, média, moda e mediana *a posteriori* são boas estimativas para todos os valores atribuídos a (κ, γ). Particularmente, todos os intervalos de credibilidade contiveram o verdadeiro valor de N. Observamos também, que ao considerarmos $k = 5$, obtivemos estimativas para N mais próximas do seu valor verdadeiro do que quando consideramos $k = 2$ e obtivemos intervalos de credibilidade de menor amplitude.

v) Considerando $\alpha = 0,8$, $k = 2$ e os dados $f_1 = 36, f_2 = 24$ e $n = 60$, obtivemos os

resultados da tabela 23.

Tabela 23. Resumos das distribuições *a posteriori* de α e N.

(κ, γ)	Parâm.	M	Mo	Q_1	Q_2	Q_3	D.P.	I.C(95%)	Ampl.I.C.
$(2; 0, 5)$	α	0,66	0,69	0,58	0,67	0,74	0,11	(0,4;0,85)	0,45
	N	91,01	90	85	90	94	6,83	(78;104)	26
$(2; 0, 1)$	α	0,6	0,65	0,52	0,61	0,69	0,13	(0,3;0,82)	0,52
	N	87,6	87	82	86	91	6,33	(75;100)	25
$(2; 0, 4)$	α	0,64	0,68	0,57	0,65	0,73	0,11	(0,38;0,84)	0,46
	N	90,2	89	85	89	94	6,72	(77;103)	26
$(2; 0, 8)$	α	0,69	0,72	0,62	0,7	0,77	0,11	(0,45;0,88)	0,43
	N	93,4	92	87	92	97	7,19	(79;107)	28

vi) Considerando $\alpha = 0,8$, $k = 5$ e os dados $f_1 = 26$, $f_2 = 20$, $f_3 = 15$, $f_4 = 14$, $f_5 = 10$ e $n = 85$, obtivemos os resultados da tabela 24.

Tabela 24. Resumos das distribuições *a posteriori* de α e N.

(κ, γ)	Parâm.	M	Mo	Q_1	Q_2	Q_3	D.P.	I.C(95%)	Ampl.I.C.
$(2; 0, 5)$	α	0,68	0,69	0,64	0,68	0,72	0,05	(0,56;0,79)	0,23
	N	96,2	96	93	95	97	3,55	(89;103)	14
$(2; 0, 1)$	α	0,67	0,67	0,63	0,67	0,71	0,06	(0,54;0,78)	0,24
	N	95,96	95	92	95	97	3,51	(89;102)	13
$(2; 0, 4)$	α	0,68	0,68	0,64	0,68	0,72	0,05	(0,55;0,79)	0,24
	N	96,14	96	93	95	97	3,54	(89;103)	14
$(2; 0, 8)$	α	0,69	0,7	0,65	0,69	0,73	0,05	(0,57;0,8)	0,23
	N	96,39	96	93	95	98	3,59	(89;103)	14

De acordo com as tabelas 23 e 24 concluímos que, para $k = 2$ e $k = 5$, as estimativas bayesianas de α, média, moda e mediana *a posteriori* são razoáveis para todos os valores atribuídos a (κ, γ) e que os intervalos de credibilidade não contiveram o verdadeiro valor de α. Com relação ao parâmetro N, média, moda e mediana *a posteriori* são boas

estimativas para todos os valores atribuídos a (κ, γ). Além disso, todos os intervalos de credibilidade contiveram o verdadeiro valor de N. Observamos também, que ao considerarmos $k = 5$, obtivemos estimativas para N mais próximas do seu valor verdadeiro do que quando consideramos $k = 2$ e obtivemos intervalos de credibilidade de menor amplitude.

Refizemos o exemplo 13 adotando para N *a priori* de Jeffreys, isto é, $\pi(N) = 1/N$, $N = 1, 2, ...$, e constatamos que não houve diferenças significativas entre os resumos *a posteriori* de N e aqueles obtidos no exemplo 13. Portanto, as prioris uniforme nos inteiros não negativos e a de Jeffreys para N produzem praticamente as mesmas estimativas bayesianas de N.

Além disso, pelo exemplo 13, notamos que a escolha dos hiperparâmetros κ e γ influencia as estimativas de α e N. Além disso, quando atribuímos a κ e γ os valores 2 e 0,5, respectivamente, ou seja, uma distribuição *a priori* não informativa para α, obtivemos estimativas bayesianas de α e N praticamente iguais às obtidas no exemplo 12, o que já era esperado.

Capítulo 4

Modelo estatístico para revisão de software: independência entre erros e dependência entre revisores.

Neste capítulo, mantendo a hipótese de independência entre as seleções dos erros, tratamos do problema de estimação do número de erros de um *software* com a suposição de dependência entre revisores e não mais independência, como nos capítulos 2 e 3.

4.1 Homogeneidade e independência entre erros e heterogeneidade e dependência entre revisores.

Como no capítulo 2, considere o vetor aleatório k-dimensional $\mathbf{X}_i = (X_{i1}, X_{i2}, ..., X_{ik})$, associado a cada erro i, onde X_{ij} assume valor 1 se o erro i for detectado pelo j-ésimo revisor e 0 caso contrário, $i = 1, 2, ..., N$, $j = 1, 2, ..., k$, e seja

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

a probabilidade de um erro i qualquer apresentar o histórico de leitura $\mathbf{w}_r, \mathbf{w}_r \in \Delta$, $r = 1, 2, ..., l$.

Para vizualização do problema, consideremos $k = 2$, $\Delta = \{(1,0),(0,1),(1,1),(0,0)\}$ e sejam p_1, p_2, p_3 e p_4 as probabilidades de $\mathbf{X}_i$ assumir os históricos $\mathbf{w}_1 = (1,0), \mathbf{w}_2 = (0,1), \mathbf{w}_3 = (1,1)$ e $\mathbf{w}_4 = (0,0)$, respectivamente, tais que os p_j's são distintos dois a dois, $j = 1,2,3,4$.

Assim, temos

$$P(X_{i1} = 1) = P(X_{i1} = 1, X_{i2} = 0) + P(X_{i1} = 1, X_{i2} = 1) = p_1 + p_3,$$

e

$$P(X_{i2} = 1) = P(X_{i1} = 0, X_{i2} = 1) + P(X_{i1} = 1, X_{i2} = 1) = p_2 + p_3,$$

$i = 1, 2, ..., N$.

Logo, $P(X_{i1} = 1) \neq P(X_{i2} = 1)$, o que implica na heterogeneidade dos revisores 1 e 2.

Além disso, como

$$P(X_{i1} = 1, X_{i2} = 1) = p_3 \neq (p_1 + p_3)(p_2 + p_3) = P(X_{i1} = 1)P(X_{i2} = 1),$$

$i = 1, 2, ..., N$, então os revisores 1 e 2 são dependentes.

De (2.3) segue que a função de verossimilhança é dada por

$$\begin{aligned}
L(N, \mathbf{p}|n_1, n_2, ..., n_{l-1}) &= P(n_1, n_2, ..., n_{l-1}, N - n|N, \mathbf{p}) = \\
&= \sum \prod_{i_1 \epsilon A_1} p_{i_1 1} \prod_{i_2 \epsilon A_2} p_{i_2 2} \cdots \prod_{i_{l-1} \epsilon A_{l-1}} p_{i_{l-1}(l-1)} \prod_{i_l \epsilon A_l} p_{i_l l} = \\
&= \sum {p_1}^{n_1} {p_2}^{n_2} ... {p_{l-1}}^{n_{l-1}} {p_l}^{N-n} = \\
&= \frac{N!}{n_1! n_2! ... n_{l-1}! (N-n)!} {p_1}^{n_1} {p_2}^{n_2} ... {p_{l-1}}^{n_{l-1}} {p_l}^{N-n} = \\
&= \frac{N!}{n_1! n_2! ... n_{l-1}! (N-n)!} \prod_{r=1}^{l} {p_r}^{n_r} = \\
&\propto \frac{N!}{(N-n)!} \prod_{r=1}^{l} {p_r}^{n_r},
\end{aligned}$$

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

4.1.1 Modelo bayesiano.

Nesta seção fazemos uma análise bayesiana do modelo proposto. Suponhamos *a priori* que $(p_1, p_2, ...p_{l-1})$ tenha distribuição de Dirichlet de parâmetros $\alpha_i, \alpha_i > 0$, $i = 1, 2, ..., l$, e seja $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_l)$. Assim, a distribuição *a priori* para $(p_1, p_2, ...p_{l-1})$ é dada por

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

$0 < p_i < 1$, $i = 1, ..., l-1$, $\sum_{i=1}^{l-1} p_i < 1$ e $p_l = 1 - (p_1 + p_2 + ... + p_{l-1})$.

Supondo N e $(p_1, p_2, ...p_{l-1})$ independentes, a distribuição *a priori* conjunta para N e $(p_1, p_2, ...p_{l-1})$ é dada por

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

$N \geq n, 0 < p_i < 1$, $i = 1, ..., l-1$, $\sum_{i=1}^{l-1} p_i < 1$, $p_l = 1 - (p_1 + p_2 + ... + p_{l-1})$ e a distribuição *a posteriori* conjunta para N e $(p_1, p_2, ...p_{l-1})$ é dada por

$$\begin{aligned}
\pi(N, (p_1, p_2, ...p_{l-1})|\boldsymbol{\alpha}, n_1, n_2, ..., n_{l-1}) &\propto L(N, \mathbf{p}|\boldsymbol{\alpha}, n_1, n_2, ..., n_{l-1})\pi(N, (p_1, p_2, ...p_{l-1})|\boldsymbol{\alpha}) \propto \\
&\propto \frac{N!}{(N-n)!} \prod_{r=1}^{l} {p_r}^{n_r} \pi(N) \prod_{i=1}^{l} p_i^{\alpha_i - 1} = \\
&= \frac{N!}{(N-n)!} \pi(N) \prod_{r=1}^{l} p_r^{n_r + \alpha_r - 1} = \\
&= \frac{N!}{(N-n)!} \pi(N) \prod_{r=1}^{l-1} p_r^{n_r + \alpha_r - 1} p_l^{N - n + \alpha_l - 1}, \qquad (4.1)
\end{aligned}$$

$N \geq n, 0 < p_i < 1$, $i = 1, ..., l-1$, $\sum_{i=1}^{l-1} p_i < 1$ e $p_l = 1 - (p_1 + p_2 + ... + p_{l-1})$. A constante normalizadora de (4.1), C_7, é tal que

$$
\begin{aligned}
C_7^{-1} &= \sum_{N\geq n}\int_A \frac{N!}{(N-n)!}\pi(N)\prod_{r=1}^{l} p_r^{n_r+\alpha_r-1}dp_1dp_2...dp_{l-1} = \\
&= \sum_{N\geq n}\frac{N!}{(N-n)!}\pi(N)\int_A \prod_{r=1}^{l} p_r^{n_r+\alpha_r-1}dp_1dp_2...dp_{l-1}, \text{ onde} \\
A &= \{(p_1,p_2,...p_{l-1}) : 0<p_i<1, i=1,...,l-1 \text{ e } \sum_{i=1}^{l-1} p_i<1\}. \text{ Então,}
\end{aligned}
$$

$$
\begin{aligned}
C_7^{-1} &= \sum_{N\geq n}\frac{N!}{(N-n)!}\pi(N)\frac{\prod_{r=1}^{l}\Gamma(\alpha_r+n_r)}{\Gamma[\sum_{r=1}^{l}(\alpha_r+n_r)]} = \\
&= \sum_{N\geq n}\frac{N!}{(N-n)!}\pi(N)\frac{\prod_{r=1}^{l-1}\Gamma(\alpha_r+n_r)}{\Gamma[\sum_{r=1}^{l}\alpha_r+N]}\Gamma(\alpha_l+N-n)\propto \\
&\propto S_n,
\end{aligned}
$$

onde

$$
S_n = \sum_{N\geq n}\frac{N!}{(N-n)!}\pi(N)\frac{\Gamma(\alpha_l+N-n)}{\Gamma(\sum_{r=1}^{l}\alpha_r+N)}. \tag{4.2}
$$

Teorema 5.

Suponhamos que a distribuição *a priori* para 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 para 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 para N e $(p_1, p_2, ..., p_{l-1})$ existe.

Prova. Do Teorema 2, temos

i) $\frac{N!}{(N-n)!N^n} \underset{N\to\infty}{\to} 1$. Mostremos inicialmente 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

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

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

e uma vez que o resultado (2.10) 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

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

então

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

$$\begin{aligned} = & \frac{\Gamma(\alpha_l + N - n)\exp\{\alpha_l + N - n\}}{(\alpha_l + N - n)^{\alpha_l + N - n - \frac{1}{2}}} \frac{(\sum_{r=1}^{l} \alpha_r + N)^{\sum_{r=1}^{l} \alpha_r + N - \frac{1}{2}}}{\Gamma(\sum_{r=1}^{l} \alpha_r + N)\exp\{\sum_{r=1}^{l} \alpha_r + N\}} \times \\ & \times \left(\frac{\alpha_l + N - n}{N}\right)^{\alpha_l + N - n - \frac{1}{2}} \left(\frac{\sum_{r=1}^{l} \alpha_r + N}{N}\right)^{-[\sum_{r=1}^{l} \alpha_r + N - \frac{1}{2}]} \times \\ & \times \exp\{-\alpha_l + n + \sum_{r=1}^{l} \alpha_r\} \underset{N\to\infty}{\to} \sqrt{2\pi}\frac{1}{\sqrt{2\pi}}\exp\{\alpha_l - n\}\frac{1}{\exp(\sum_{r=1}^{l} \alpha_r)}\exp\{-\alpha_l + n + \sum_{r=1}^{l} \alpha_r\} = \\ = & \ 1, \end{aligned}$$

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 $\varepsilon, 0<\varepsilon<1$, existe um número inteiro positivo N_0, $N_0>n$, tal que

$$1-\varepsilon < N^{\sum_{r=1}^{l-1}\alpha_r}\frac{N!}{(N-n)!}\frac{\Gamma(\alpha_l+N-n)}{\Gamma(\sum_{r=1}^{l}\alpha_r+N)} < 1+\varepsilon, \tag{4.3}$$

para todo $N>N_0$ e S_n (dado em 4.2) pode ser escrito como

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

onde

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

Mas, de (4.3), segue que

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

Logo, para $u=0$ temos

$$\begin{aligned}\sum_{r=1}^{l-1}\alpha_r+u &= \sum_{r=1}^{l-1}\alpha_r>1 \Longrightarrow \sum_{N=N_o+1}^{\infty} N^{-\sum_{r=1}^{l-1}\alpha_r}<\infty \Longrightarrow \\ &\Longrightarrow 0<W_{N_0}<\infty \Longrightarrow 0<S_n<\infty \Longrightarrow \\ &\Longrightarrow \text{existe a distribuição } \textit{a posteriori} \text{ conjunta de } N \text{ e } (p_1,p_2,...p_{l-1}).\end{aligned}$$

e

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

$\Longrightarrow S_n$ infinita $\Longrightarrow$ não existe a 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

$$\sum_{r=1}^{l-1} \alpha_r + u = \sum_{r=1}^{l-1} \alpha_r + 1 > 1 \Longrightarrow$$

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

$\Longrightarrow$ existe a distribuição *a posteriori* conjunta de N e $(p_1, p_2, ...p_{l-1})$,

o que prova (b). ■

Se a *priori* $\pi(N) = 1, N \geq 1$, segue de (4.1) que a distribuição *a posteriori* conjunta para N e $(p_1, p_2, ...p_{l-1})$ é dada por

$$\pi(N, (p_1, p_2, ...p_{l-1}) | \boldsymbol{\alpha}, n_1, n_2, ..., n_{l-1}) \propto \frac{N!}{(N-n)!} \prod_{r=1}^{l-1} p_r^{n_r + \alpha_r - 1} p_l^{N-n+\alpha_l - 1}. \tag{4.4}$$

De (4.4), segue que a distribuição condicional de N, dados $(p_1, p_2, ...p_{l-1})$, $\boldsymbol{\alpha}$ e $n_1, n_2, ...,$ n_{l-1} é dada por

$$\pi(N | (p_1, p_2, ...p_{l-1}), \boldsymbol{\alpha}, n_1, n_2, ..., n_{l-1}) \propto \binom{N}{n} p_l^N,$$

$N \geq n$ e $p_l = 1 - p_1 - p_2 - ... - p_{l-1}$, ou seja,

$$\pi(N | (p_1, p_2, ...p_{l-1}), \boldsymbol{\alpha}, n_1, n_2, ..., n_{l-1}) = C_8 \binom{N}{n} p_l^N,$$

onde a constante normalizadora C_8 é tal que

$$\begin{aligned} C_8^{-1} &= \sum_{N \geq n} \binom{N}{n} p_l^N = \sum_{s=0}^{\infty} \binom{s+n}{n} p_l^{s+n} = p_l^n \sum_{s=0}^{\infty} \binom{s+n}{n} p_l^s = \\ &= p_l^n \sum_{s=0}^{\infty} \binom{-n-1}{s} (-p_l)^s = p_l^n (1 - p_l)^{-n-1}. \end{aligned}$$

Logo,

$$\pi(N|(p_1,p_2,...p_{l-1}),\boldsymbol{\alpha},n_1,n_2,...,n_{l-1}) = \binom{N}{n} p_l^{N-n}(1-p_l)^{n+1}, \tag{4.5}$$

$N \geq n$, isto é, a distribuição condicional de N, dados $(p_1,p_2,...p_{l-1}),\boldsymbol{\alpha}$ e $n_1,n_2,...,n_{l-1}$, é igual a distribuição de uma variável aleatória $n+G$, onde G tem distribuição binomial negativa de parâmetros $n+1$ e $1-p_l$.

De fato, se G tiver distribuição binomial negativa com parâmetros $n+1$ e $1-p_l$, então

$$\begin{aligned} P(G+n &= g) = P(G=g-n) = \\ &= \binom{g}{n} p_l^{g-n}(1-p_l)^{n+1}, \end{aligned}$$

$g = n, n+1, ...$

De (4.4), segue que a distribuição condicional de $(p_1,p_2,...p_{l-1})$, dados N, $\boldsymbol{\alpha}$ e $n_1,n_2,...,$ n_{l-1} é dada por

$$\pi((p_1,p_2,...p_{l-1})|N,\boldsymbol{\alpha},n_1,n_2,...,n_{l-1}) \propto \prod_{r=1}^{l-1} p_r^{n_r+\alpha_r-1} p_l^{N-n+\alpha_l-1}, \tag{4.6}$$

$0 < p_i < 1, i = 1,...,l-1, \sum_{i=1}^{l-1} p_i < 1$ e $p_l = 1-(p_1+p_2+...+p_{l-1})$, isto é, $(p_1,p_2,...p_{l-1})$ dados $N,\boldsymbol{\alpha}$ e $n_1,n_2,...,n_{l-1}$ tem distribuição de Dirichlet de parâmetros $n_i+\alpha_i$, $i=1,2,...,l$, e a distribuição *a posteriori* marginal de N é dada por

$$\begin{aligned} \pi(N|\boldsymbol{\alpha},n_1,n_2,...,n_{l-1}) &= \int_A \pi(N,(p_1,p_2,...p_{l-1})|\boldsymbol{\alpha},n_1,n_2,...,n_{l-1})\ dp_1dp_2...dp_{l-1} \propto \\ &\propto \frac{N!}{(N-n)!}\int_A \prod_{r=1}^{l} p_r^{n_r+\alpha_r-1} dp_1dp_2...dp_{l-1} = \\ &= \frac{N!}{(N-n)!}\frac{\prod_{r=1}^{l}\Gamma(\alpha_r+n_r)}{\Gamma[\sum_{r=1}^{l}(\alpha_r+n_r)]} \propto \\ &\propto \frac{N!}{(N-n)!}\frac{\Gamma(\alpha_l+N-n)}{\Gamma(\sum_{r=1}^{l}\alpha_r+N)}, \end{aligned} \tag{4.7}$$

$N \geq n$.

Se a *priori* $\pi(N) = 1/N, N \geq 1$, segue de (4.1) que a distribuição *a posteriori* conjunta para N e $(p_1, p_2, ...p_{l-1})$ é dada por

$$\pi(N, (p_1, p_2, ...p_{l-1})|\boldsymbol{\alpha}, n_1, n_2, ..., n_{l-1}) \propto \frac{N!}{(N-n)!}\frac{1}{N}\prod_{r=1}^{l-1} p_r^{n_r+\alpha_r-1} p_l^{N-n+\alpha_l-1}. \quad (4.8)$$

De (4.8), segue que a distribuição condicional de N, dados $(p_1, p_2, ...p_{l-1})$, $\boldsymbol{\alpha}$ e $n_1, n_2, ...,$ n_{l-1} é dada por

$$\pi(N|(p_1, p_2, ...p_{l-1}), \boldsymbol{\alpha}, n_1, n_2, ..., n_{l-1}) \propto \binom{N-1}{n-1} p_l^N,$$

$N \geq n$.

Assim,

$$\pi(N|(p_1, p_2, ...p_{l-1}), \boldsymbol{\alpha}, n_1, n_2, ..., n_{l-1}) = C_9 \binom{N-1}{n-1} p_l^N,$$

onde C_9 é tal que

$$\begin{aligned} C_9^{-1} &= \sum_{N=n}^{\infty} \binom{N-1}{n-1} p_l^N = \sum_{s=0}^{\infty} \binom{s+n-1}{n-1} p_l^{s+n} = \\ &= p_l^n \sum_{s=0}^{\infty} \binom{-n}{s} (-p_l)^s = p_l^n (1-p_l)^{-n}, \end{aligned}$$

o que implica

$$\pi(N|(p_1, p_2, ...p_{l-1}), \boldsymbol{\alpha}, n_1, n_2, ..., n_{l-1}) = \binom{N-1}{n-1} (1-p_l)^n p_l^{N-n}, \quad (4.9)$$

$N \geq n$, isto é, a distribuição condicional de N, dados $(p_1, p_2, ...p_{l-1})$, $\boldsymbol{\alpha}$ e $n_1, n_2, ..., n_{l-1}$, é igual a distribuição de uma variável aleatória $n + J$, onde J tem distribuição binomial negativa de parâmetros n e $1 - p_l$.

De fato, se J tiver distribuição binomial negativa com parâmetros n e $1 - p_l$, então

$$\begin{aligned} P(J+n &= j) = P(J = j-n) = \\ &= \binom{j-1}{n-1} (1-p_l)^n p_l^{j-n}, \end{aligned}$$

$j = n, n+1, ...$

De (4.8) segue que a distribuição condicional de $(p_1, p_2, ...p_{l-1})$, dados $N, \boldsymbol{\alpha}$ e $n_1, n_2, ...,$ n_{l-1}, é dada por

$$\pi((p_1, p_2, ...p_{l-1})|N, \boldsymbol{\alpha}, n_1, n_2, ..., n_{l-1}) \propto \prod_{r=1}^{l-1} p_r^{n_r+\alpha_r-1} p_l^{N-n+\alpha_l-1}, \tag{4.10}$$

$0 < p_i < 1,\ i = 1, ..., l-1,\ \sum_{i=1}^{l-1} p_i < 1$ e $p_l = 1 - (p_1 + p_2 + ... + p_{l-1})$, isto é, $(p_1, p_2, ...p_{l-1})$ dados $N, \boldsymbol{\alpha}$ e $n_1, n_2, ..., n_{l-1}$ tem distribuição de Dirichlet de parâmetros $n_i + \alpha_i, i = 1, 2, ..., l$ e a distribuição *a posteriori* marginal de N é dada por

$$\begin{aligned}
\pi(N|\boldsymbol{\alpha}, n_1, n_2, ..., n_{l-1}) &= \int_A \pi(N, (p_1, p_2, ...p_{l-1})|\boldsymbol{\alpha}, n_1, n_2, ..., n_{l-1})\ dp_1 dp_2 ... dp_{l-1} \propto \\
&\propto \frac{(N-1)!}{(N-n)!} \int_A \prod_{r=1}^{l} p_r^{n_r+\alpha_r-1} dp_1 dp_2 ... dp_{l-1} = \\
&= \frac{(N-1)!}{(N-n)!} \frac{\prod_{r=1}^{l} \Gamma(\alpha_r + n_r)}{\Gamma[\sum_{r=1}^{l} (\alpha_r + n_r)]} \propto \\
&\propto \frac{(N-1)!}{(N-n)!} \frac{\Gamma(\alpha_l + N - n)}{\Gamma(\sum_{r=1}^{l} \alpha_r + N)},
\end{aligned} \tag{4.11}$$

$N \geq n$.

4.1.2 Estimativas bayesianas de N.

Nesta seção utilizamos os mesmos métodos de estimação de N apresentados na seção 2.1.3

Estimativas bayesinas de N via distribuição *a posteriori* marginal quase exata.

O método de obtenção das estimativas de N da seção 4.1.1 foi implementado no *software* MAPLE (versão 7.00) e no programa anexado no apêndice G, são determinados os resumos aproximados *a posteriori* de N, como média, desvio padrão, moda, intervalo de credibilidade e quartis.

A seguir, com o objetivo de analisar a performance do modelo proposto na seção 4.1,

apresentamos os exemplos 14 e 15 com dados simulados através do *software* R (versão 1.9.0). Observe que nestes exemplos, apesar dos dados serem gerados através do *software* R, o método de obtenção das estimativas bayesianas de N via distribuição *a posteriori* marginal quase exata foi implementado no MAPLE.

Exemplo 14.

Neste exemplo atribuímos a N o valor 100, a k os valores 2 e 5 e geramos as estatísticas $n_1, n_2, ..., n_{l-1}$(número de erros que apresentam a trajetória $\mathbf{w}_r, r = 1, 2, ..., l-1$, em que $l = 2^k$). É claro que $n_l = N - \sum_{r=1}^{l-1} n_r$. Para isto executamos os passos a seguir.

1. Primeiramente geramos $(p_1, p_2, ..., p_{l-1})$ a partir da distribuição de Dirichlet de parâmetros $\alpha_i = \alpha, i = 1, 2, ..., l$, utilizando o resultado: suponha $X_1, ..., X_l$ independentes e identicamente distribuídos, com distribuição Gama de parâmetros α e 1; e defina

$$p_j = \frac{X_j}{\sum_{j=1}^{l} X_j},$$

$j = 1, ..., l-1$. Então $(p_1, p_2, ..., p_{l-1})$ tem distribuição de Dirichlet de parâmetro α.

2. Com os valores obtidos no passo 1, geramos $n_1, n_2, ..., n_l$ a partir da distribuição multinomial de parâmetros 100 e $p_1, ..., p_{l-1}, p_l = 1 - p_1 - ... - p_{l-1}$. Vale lembrar que $n = \sum_{r=1}^{l-1} n_r$.

De posse dos dados, consideramos para N a distribuição *a priori* de Jeffreys, ou seja, $\pi(N) = 1/N, N = 1, 2, ...$ Aplicamos o método via distribuição *a posteriori* marginal (4.11), para diferentes valores de α, obtendo resumos aproximados *a posteriori* de N, 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 $\alpha = 10$ e $k = 2$ na geração dos dados, obtivemos $n_1 = 21$, $n_2 = 26, n_3 = 20, n_4 = 33$ e conseqüentemente $n = 67$. Os resultados obtidos, seguem na

tabela 25.

Tabela 25. resumos aproximados *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 25, 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 estimativas de N.

Apresentamos a seguir o gráfico da função de probabilidade aproximada a *posteriori* marginal de N considerando $\alpha = 50$.

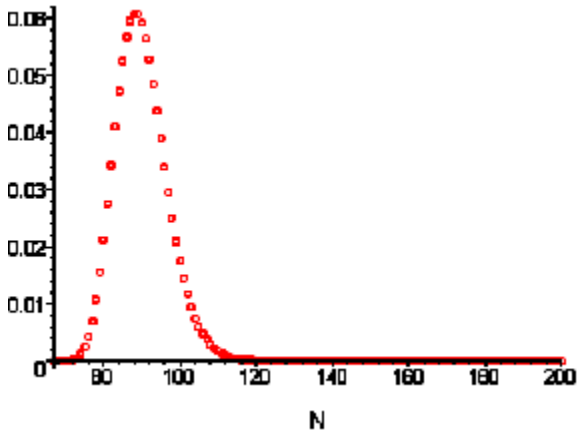

Fig. 12. Gráfico da função de probabilidade aproximada *a posteriori* marginal de N.

ii) Considerando $\alpha = 5$ e $k = 5$ na geração dos dados, obtivemos

n_1	3	n_{13}	4		
n_2	3	n_{14}	4		
n_3	1	n_{15}	4	n_{25}	5
n_4	0	n_{16}	1	n_{26}	3
n_5	6	n_{17}	1	n_{27}	1
n_6	5	n_{18}	2	n_{28}	3
n_7	3	n_{19}	1	n_{29}	2
n_8	4	n_{20}	3	n_{30}	10
n_9	3	n_{21}	3	n_{31}	5
n_{10}	3	n_{22}	6	n_{32}	2
n_{11}	0	n_{23}	2		
n_{12}	4	n_{24}	3		

e conseqüentemente $n = 98$. Os resultados obtidos seguem na tabela 26.

Tabela 26. resumos aproximados *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 26, verificamos a obtenção de boas estimativas para N, independentemente do valor atribuído a α, e também pudemos observar que o verdadeiro valor do parâmetro pertence a todos os intervalos de credibilidade obtidos. Além disso, a amplitude de tais intervalos é bem menor do que daqueles obtidos no item i), quando consideramos apenas 2 revisores. É importante observar que, ao passarmos de 2 para 5

revisores, observamos uma melhoria substancial nas estimativas obtidas.

Exemplo 15.

Neste exemplo atribuímos aos parâmetros do modelo os mesmos valores atribuídos no exemplo 14, bem como utilizamos as mesmas estatísticas. Adotamos para N a distribuição *a priori* própria uniforme nos inteiros não negativos, com suporte em $\{1, 2, ..., M\}$, $M = 200, 1000$, a fim de verificarmos a influência do valor de M nas estimativas bayesianas de N e aplicamos o método via distribuição *a posteriori* marginal (4.7).

i) Considerando $k = 2$ e os dados $n_1 = 21, n_2 = 26, n_3 = 20, n_4 = 33$ e $n = 67$, obtivemos os resultados das tabela 27 e 28 para $M = 200$ e $M = 1000$, respectivamente.

Tabela 27. Resumos aproximados *a posteriori* de N.

α	M	Moda	Q_1	Q_2	Q_3	D.P.	I.C(95%)	Ampl.I.C.
0,5	99,15	67	70	84	117	35,4	(67;188)	121
1	100,49	67	75	89	115	31,9	(67;183)	116
2	97,86	78	78	90	108	25,46	(68;167)	99
5	93,11	84	81	89	100	15,48	(72;131)	59
10	91,28	87	83	89	96	10,82	(74;116)	42
20	90,44	88	83	89	94	8,34	(76;108)	32
50	89,97	88	84	88	93	6,72	(77;104)	27

De acordo com a tabela 27, verificamos que as estimativas obtidas para N são próximas daquelas obtidas na tabela 25 quando consideramos $\alpha = 5, 10, 20$ e 50.

Tabela 28. Resumos aproximados *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

Observação: Os resultados obtidos para $\alpha = 5, 10, 20$ e 50 foram omitidos da tabela 28 por serem análogos aos da tabela 27.

Comparando as tabelas 27 e 28 com relação a $\alpha = 0,5$ e 1, verificamos uma grande

diferença entre as estimativas de N. Isto nos mostra a sensibilidade dos resumos aproximados *a posteriori* de N com relação à mudança do valor de M de 200 para 1000.

ii) Considerando $k = 5$ e os dados do exemplo 11, item ii), obtivemos os resultados da tabela 29 para para $M = 200$. Omitiremos os resultados obtidos para $M = 1000$ por serem análogos aos da tabela 29.

Tabela 29. resumos aproximados *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)	3
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

Os resulatos da tabela 29 são análogos àqueles da tabela 26.

Pelos exemplos 14 e 15 podemos concluir que

a) Para $k = 2$, a adoção da distribuição *a priori* uniforme nos inteiros não negativos para N é equivalente à adoção da *priori* de Jeffreys, se considerarmos $\alpha = 5, 10, 20$ e 50.

b) Para $k = 5$, a adoção da distribuição *a priori* uniforme nos inteiros não negativos para N é equivalente à adoção da *priori* de Jeffreys, para todos os valores de α considerados em tais exemplos, ou seja, não existem diferenças significativas nos resumos aproximados *a posteriori* de N quando adotamos tais *priori*s.

Estimativas bayesianas de N via algoritmo *Gibbs Sampling*.

A seguir, a título de ilustração, aplicamos o algoritmo *Gibbs Sampling* nos exemplos 16 e 17, onde os dados utilizados foram os do exemplo 15.

Nos exemplos 16 e 17, a convergência das cadeias geradas foi verificada utilizando a bibilioteca CODA. Além disso, em cada exemplo consideramos um *burn in* de 100 elementos, geramos duas cadeias, com mil elementos cada e para garantir independência

aproximada, tomamos elementos de 40 em 40. O programa utilizado para a implementação deste modelo foi feito via *software* R (versão 1.9.0) e está anexado no apêndice F.

Exemplo 16.

Neste exemplo atribuímos aos parâmetros do modelo os mesmo valores atribuídos no exemplo 14, bem como utilizamos as mesmas estatísticas. Adotamos a *priori* uniforme nos inteiros não negativos para N e implementamos o algoritmo *Gibbs Sampling* através das distribuições condicionais (4.5) e (4.6).

i) Considerando $k = 2$ e os dados $n_1 = 21, n_2 = 26, n_3 = 20, n_4 = 33$ e $n = 67$, obtivemos os resultados da tabela 30.

Tabela 30. Estimativas dos resumos *a posteriori* de N

α	Média	Q_1	Q_2	Q_3	D.P.	I.C(95%)	Ampl.I.C.	G-R.
0,5	8314	82	151	593	51507	(67;68149)	68082	1,02
1	136,98	77	96	138	154,6	(68;488)	420	1,02
2	98,7	79	90	108	30,93	(69;177)	108	1,01
5	92,66	82	90	100	15,13	(73;131)	58	1
10	91,08	83	89	97	10,68	(75;116)	41	1
20	90,51	85	90	96	8,4	(77;110)	33	1
50	89,88	85	89	94	6,79	(78;105)	27	1

G.R.: valor retornado pelo diagnóstico de Gelman e Rubin (G.R.<1,1$\Longrightarrow$convergência)

Apresentamos a seguir alguns gráficos relativos a *posteriori* de N para $\alpha = 20$.

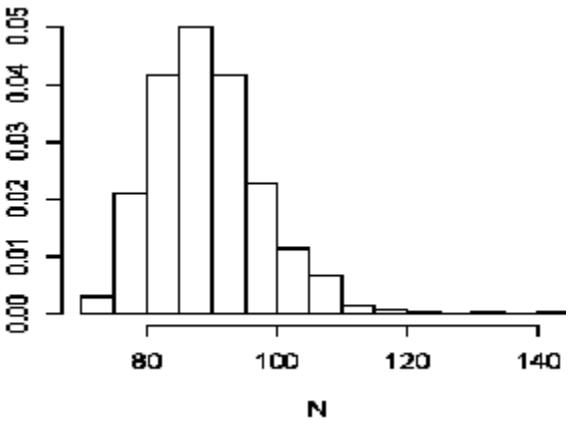

Fig. 13. Histograma da estimativa da distribuição *a posteriori* marginal de N .

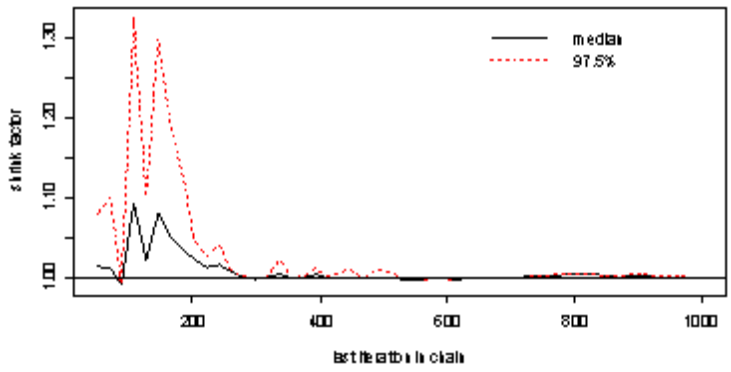

Fig. 14. Gráfico de convergência das cadeias de N (critério de Gelman-Rubin).

Pela fig. 14 temos a indicação da convergência na 600^a iteração das cadeias.

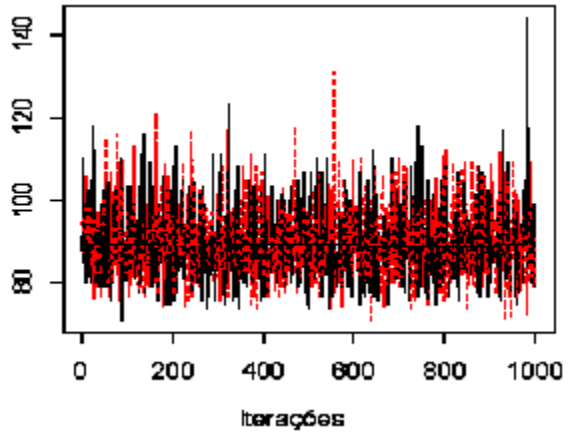

Fig. 15. Traço das cadeias de N.

Pela fig. 15 observamos uma certa uniformidade do traço para as duas cadeias, o que também indica convergência.

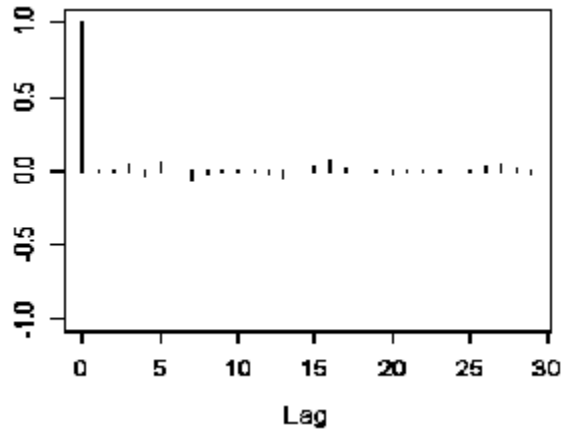

Fig. 16. Autocorrelação.

Pela fig. 16, a autocorrelação é praticamente nula, indicando independência aproximada entre os elementos da amostra.

ii) Considerando $k = 5$ e os dados do exemplo 13, item ii), obtivemos os resultados da tabela 31.

Tabela 31. Estimativas dos resumos *a posteriori* de N.

α	Média	Q_1	Q_2	Q_3	D.P.	I.C(95%)	Ampl.I.C.	G-R.
0,5	101,61	98	99	103	6	(98;117)	19	1,06
1	101,49	99	100	103	4,27	(98;113)	15	1
2	101,31	99	101	103	3,04	(98;109)	11	1
5	101,2	100	101	103	2,28	(98;106)	8	1
10	101,18	100	101	102	2,1	(98;106)	8	1
20	101,16	100	101	102	1,99	(98;105)	7	1
50	101,21	100	101	102	1,87	(98;105)	7	1

G.R.: valor retornado pelo diagnóstico de Gelman e Rubin (G.R.$<$1,1$\Longrightarrow$convergência)

Note que os resultados da tabela 30 são análogos aos da tabela 27, considerando $\alpha = 2, 5, 10, 20$ e 50 e que os resultados da tabela 31 são análogos aos da tabela 29, o que evidencia a equivalência dos métodos *Gibbs Sampling* e distribuição *a posteriori* marginal quase exata.

Exemplo 17.

Neste exemplo atribuímos aos parâmetros do modelo os mesmo valores atribuídos no exemplo 14, bem como utilizamos as mesmas estatísticas. Adotamos a *priori* de Jeffreys para N e implementamos o algoritmo *Gibbs Sampling* através das distribuições condicionais (4.9) e (4.10).

i) Considerando $k = 2$ e os dados $n_1 = 21, n_2 = 26, n_3 = 20, n_4 = 33$ e $n = 67$, obtivemos os resultados da tabela 32.

Tabela 32. Estimativas dos resumos *a posteriori* de N.

α	Média	Q_1	Q_2	Q_3	D.P.	I.C(95%)	Ampl.I.C.	G-R.
0,5	125,07	70	79	108	201,33	(67;468)	401	1,09
1	98,82	73	84	105	52,19	(67;277)	210	1,02
2	93,09	77	86	101	24,48	(69;157)	88	1
5	90,98	81	88	98	14,49	(71;129)	58	1,01
10	90,13	83	89	96	10,04	(75;113)	38	1
20	89,78	84	89	95	8,24	(76;109)	33	1
50	89,49	85	89	94	6,6	(78;104)	26	1

G.R.: valor retornado pelo diagnóstico de Gelman e Rubin (G.R.<1,1$\Longrightarrow$convergência)

Para os dados do exemplo 11, item ii), obtivemos resultados análogos aos da tabela 26 e pelos resultados da tabela 30, análogos aos da tabela 17 considerando $\alpha = 2, 5, 10, 20$ e 50, temos novamente a evidência da equivalência dos métodos *Gibbs Sampling* e distribuição *a posteriori* marginal quase exata.

4.1.3 Exemplo com dados reais.

Nesta seção aplicamos o método baseado na determinação da distribuição *a posteriori* marginal quase exata de N para estimação deste parâmetro, utilizando um conjunto de dados reais disponíveis em LaPorte *et al.* (1995). Embora não se trate propriamente da estimação do número de erros de um software, tais dados são compatíveis com a metodologia em questão. Vale salientar que neste exemplo não foi possível aplicar o algoritmo *Gibbs Sampling* para estimação de N, dado que não temos disponível no conjunto de dados as

estatísticas necessárias para aplicação deste método, ou seja, $n_1, n_2, ..., n_{16}$.

Exemplo 18.

Os dados utilizados neste exemplo referem-se a um estudo realizado numa escola em Pittsburgh, Pensilvânia, entre 1 de setembro e 31 de dezembro de 1991, com relação a casos de ferimentos dos alunos. Nesse estudo "ferimento" foi definido como qualquer evento que resulta em dano físico e conseqüente tratamento médico para o aluno. Dos 1400 alunos da escola, 1245 aceitaram participar do monitoramento. Trata-se de jovens entre 12 e 16 anos, dos sexos masculino e feminino (presentes em proporções equilibradas no estudo) e que pertecem à várias raças e classes econômica representativas da Pensilvânia. Foram consideradas quatro fontes de informação (dependentes entre si) sobre o número de ferimentos nos jovens:

• Fonte 1: entrevista mensal com todos os alunos que receberam atendimento médico.

• Fonte 2: revisão dos atestados médicos que os alunos trouxeram para a enfermeira da escola.

• Fonte 3: revisão dos registros de atendimentos diários na escola, para identificar se o aluno faltou devido a um ferimento, doença ou outra razão.

• Fonte 4: entrevista adicional ao fim do semestre para identificar todos os tratamentos médicos para ferimentos ocorridos num período de quatro meses.

Essas quatro fontes identificaram 144 (n) casos distintos de ferimentos entre os alunos pesquisados. O objetivo do monitoramento é estimar o número total de casos de ferimentos na população de alunos. LaPorte *et al.* (1995) utilizou técnicas estatísticas de modelos log-lineares e o *software* GLIM (Baker (1987)) para estimar este número e os resultados obtidos seguem na tabela 33.

Tabela 33. Resultados obtidos por LaPorte *et al.* (1995).

	Estimativa	Intervalo de Confiança (95%)	Ampl.
N	159	(148;186)	38

Neste exemplo, fazendo analogia com o problema de revisão de *software*, as quatro fontes de informação são equivalentes a quatro revisores e estimar o número total de casos de ferimentos na população torna-se equivalente a estimar o número total de erros no *software*, N. Apresentamos nas tabelas 34, 35 e 35 os resumos aproximados *a posteriori*

de N obtidos após a aplicação do método da seção 4.1.2, baseado na determinação da distribuição *a posteriori* marginal quase exata de N. Nas tabelas 34 e 35 supomos a distribuição *a priori* própria uniforme com suporte em $\{1, 2, ..., M\}$ para N, com $M = 200$ e 1000, respectivamente e na tabela 36 supomos a distribuição *a priori* de Jeffreys para N. Note que, em todos os caso, atribuímos para α o valor 1, o que corresponde à distribuição de Dirichlet não informativa para (p_1, p_2, p_3).

Tabela 34. resumos aproximados *a posteriori* de N $(M = 200)$.

α	Média	Moda	Q_1	Q_2	Q_3	D.P.	I.C.(95%)	Ampl. I.C.
1	154,48	144	146	150	158	10,68	(144;183)	39

Tabela 35. resumos aproximados *a posteriori* de N $(M = 1000)$.

α	Média	Moda	Q_1	Q_2	Q_3	D.P.	I.C.(95%)	Ampl. I.C.
1	155,15	144	146	150	158	12,57	(144;188)	44

Tabela 36. resumos aproximados *a posteriori* de N.

α	Média	Moda	Q_1	Q_2	Q_3	D.P.	I.C.(95%)	Ampl. I.C.
1	153,8	144	145	150	157	10,15	(144;181)	37

Pelas tabelas 34 a 36, podemos notar que todos os resultados obtidos são próximos daqueles obtidos por LaPorte *et al.*.

Capítulo 5

Considerações finais.

Nesta dissertação apresentamos alguns modelos estatísticos para estimar o número de erros de um software, sob os enfoques freqüentista e bayesiano. Tais modelos são descritos resumidamente no artigo Bayesian capture-recapture methods for error detection and estimation of population size: heterogeneity and dependence, publicado por Sanjib Basu e Nader Ebrahimi na Biometrika, 88, 1, p. 269-279, em 2001. Discutimos detalhadamente as metodologias apresentadas e fazemos um estudo, através de dados simulados, de suas performances, além de apresentar três exemplos com dados reais. Nosso estudo revela que as estimativas de máxima verossimilhança e máxima verossimilhança condicional dos parâmetros de certos modelos são satisfatórias. Por outro lado, com relação ao enfoque bayesiano, verificamos a equivalência da adoção da distribuição *a priori* uniforme nos inteiros não negativos e da *priori* de Jeffreys para o parâmetro de interesse, além de observar uma influência da escolha dos valores dos hiperparâmetros das distribuições *a priori* nas estimativas bayesianas dos parâmetros dos modelos. Acreditamos que uma possível alternativa para esta situação seja a adoção de um modelo hierárquico.

Apêndice A

Programa para implementação do método de estimação de N via distribuição *a posteriori* marginal quase exata para o caso $\theta_{ij} = \theta_j$.

```
### Geração dos dados utilizando o software R ###
    set.seed(100)
    teta<-nrev<-numeric()
    nn<-0
    k<-2
    N<-100
    binnteta<-matrix(0,ncol=k,nrow=N)
    ### Geração do número de erros distintos nn ###
    teta<-c(0.5,0.6)
    for(i in 1:N) {
    for(j in 1:k) {
    binnteta[i,j]<-rbinom(1,1,teta[j])
    }
    if(sum(binnteta[i,])>0) (nn<-nn+1)
    }
```

```
nn
### Geração do número de erros detectados por cada revisor ###
for(j in 1:k) nrev[j]<-sum(binnteta[,j])
nrev
```

Obtenção da distribuição *a posteriori* marginal quase exata utilizando o *software* MAPLE

priori para N: Uniforme

```
> restart;
> k:=2:
> n[1]:=48:
> n[2]:=54:
>f[N]:=(N!/(N-nn)!)*product(((GAMMA(n[j]+alpha)*
GAMMA(N-n[j]+beta))/GAMMA(N+alpha+beta)),j=1..k):
> A[N]:=simplify(eval(f[N],N=N+1)/f[N]): A(N) = f(N+1)/f(N)
> l:=2^k:
> alpha:=1:
> beta:=1:
> distintos:=79
> u:=1000
> n[distintos]:=distintos:
> nn:=n[distintos]:
> for i from nn to u+1 do criando os pontos n, n+1, n+2,..., u+1.
> n[i]:=i
> od:
```

Relação recursiva:

```
> f(N+1)=f(N)A(N)
> 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]]: # construção da lista para obter o gráfico
> 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): # verificação
> média:=sum('distN[i]*n[i]','i'=nn..u+1):
> variância:=sum('distN[i]*((n[i]-média)^2)','i'=nn..u+1):
> 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: # extremo inferior do I.C. 95%
> for i from nn to u do
> if(sum('distN[j]','j'=nn..i)<=0.975) then s:=i end if:
> od:
> IC2:=s: # extremo superior do I.C. 95%
> for i from nn to u do
> if(sum('distN[j]','j'=nn..i)<=0.25) then s:=i end if:
> od:
> q25:=s: # Q1
> for i from nn to u do
> if(sum('distN[j]','j'=nn..i)<=0.5) then s:=i end if:
> od:
> q50:=s: # Q2
> for i from nn to u do
```

```
> if(sum('distN[j]','j'=nn..i)<=0.75) then s:=i end if:
> od:
> q75:=s: # Q3
> for i from nn to 400 do # moda
> if(distN[i]<=distN[i+1]) then s:=i else t[i]:=i end if:
> print(t[i]):
> od:
```

Apêndice B

Programa para implementação do método de estimação de N via algoritmo *Gibbs Sampling* para o caso $\theta_{ij} = \theta_j$.

Os dados foram gerados como no Apêncice A.

```
### Algoritmo Gibbs Sampling (supondo priori uniforme para N) ###
M<-1000
bi<-100
salto<-20
alfa<-1
bbeta<-1
numerototal<-bi+M*salto
a<-matrix(0,ncol=k,nrow=numerototal)
teta<-matrix(0,ncol=k,nrow=numerototal+1)
N<- numeric()
N[1]<-100
teta[1,1]<-0.8
teta[1,2]<-0.1
teta.out<-N.out<-numeric()
```

```
cont<-0
### Primeira Cadeia ###
for(i in 1:numerototal) {
for(j in 1:k) a[i,j]<-(1-teta[i,j])
produto<-prod(a[i,])
N[i+1] <- (nn+rnbinom(1,nn+1,(1-produto)))
for(j in 1:k) teta[i+1,j] <- rbeta(1,nrev[j]+alfa,N[i+1]-nrev[j]+bbeta)
cont<-cont+1
print(cont)
}
for (s in 1:numerototal)
{
if ((s>bi) && ((s-bi) %% salto) == 0)
{
N.out<-rbind(N.out,N[s])
teta.out<-rbind(teta.out,teta[s,])
}
}
a11<-N.out
a21<-teta.out
### Segunda Cadeia ###
set.seed(70)
a<-matrix(0,ncol=k,nrow=numerototal)
teta<-matrix(0,ncol=k,nrow=numerototal+1)
N<-numeric()
N[1]<-90
teta[1,1]<-0.7
teta[1,2]<-0.2
teta.out<-N.out<-numeric()
produto<-numeric()
for(i in 1:numerototal) {
```

```
for(j in 1:k) a[i,j]<-(1-teta[i,j])
produto<-prod(a[i,])
N[i+1] <- (nn+rnbinom(1,nn+1,(1-produto)))
for(j in 1:k) teta[i+1,j] <- rbeta(1,nrev[j]+alfa,N[i+1]-nrev[j]+bbeta)
}
for (s in 1:numerototal)
{
if ((s>bi) && ((s-bi) %% salto) == 0)
{
N.out<-rbind(N.out,N[s])
teta.out<-rbind(teta.out,teta[s,])
}
}
a12<-N.out
a22<-teta.out
### Analisando a convergência apenas para o parâmetro de interesse N ###
library(coda)
a1<-mcmc(a11)
a2<-mcmc(a12)
a3<-mcmc.list(a1,a2)
summary(a3)
###
###
gelman.diag(a3)
##
##
gelman.plot(a3)
##
##
traceplot(a3)
##
```

```
##
densplot(a3)
##
##
autocorr.plot(a3)
##
##
hist(a11)
##
##
hist(a12)
```

Apêndice C

Programa para resolução do sistema (2.23) e cálculo da integral (2.24).

```
> restart;
    > k:=2:
    > a[1]:=0:
    > b[1]:=0.06:
    > a[2]:=0.08:
    > b[2]:=0.12:
    > for j from 1 to k do
    > eqmedia[j]:=((a[j]+b[j])/2)-(alfa[j]/(alfa[j]+beta[j]))
    > od:
    > for j from 1 to k do
    > eqdesvio[j]:=((b[j]-a[j])/4)-((alfa[j]*beta[j])/(((alfa[j]+beta[j])^2)*(alfa[j]+beta[j]+1)))^(1/2);
    > od:
    > sol:=solve({eqmedia[1],eqmedia[2],eqdesvio[1],eqdesvio[2]},{alfa[1],alfa[2],beta[1],beta[2]}):
# solução do sistema (2.23)
    > assign(sol):
    > for j from 1 to k do
    > bbeta[j]:=(GAMMA(alfa[j]+beta[j])/(GAMMA(alfa[j])*GAMMA(beta[j])))
    *(theta[j]^(alfa[j]-1))*((1-theta[j])^(beta[j]-1));
    > od:
```

```
> with(plots):
> plot(bbeta[1],theta[1]=0..0.2):
> plot(bbeta[2],theta[2]=0..0.15):
> for j from 1 to k do
> Área[j]:=int(bbeta[j],theta[j]=a[j]..b[j]): # cálculo da integral (2.24)
> od:
```

Apêndice D

Programa para do cálculo dos resumos *a posteriori* de N para o caso $\theta_{ij} = \theta_i$.

```
### Geração do número de erros distintos n ###
    set.seed(100)
    teta<-binn<-numeric()
    n<-0
    alfa<-1
    bbeta<-1
    k<-2
    N<-100
    for(i in 1:N) teta[i]<-rbeta(1,1,1)
    for (i in 1:N) {
    binn[i]<-rbinom(1,k,teta[i])
    if(binn[i]>0) (n<-n+1)
    }
    n
    ### Supondo priori uniforme para N ###
    g<-500
    j<-s<-numeric()
```

```
média<-n+(n+1)/((beta(alfa,bbeta)/beta(alfa,k+bbeta))-1)
var<-(((n+1)*((beta(alfa,k+bbeta))/(beta(alfa,bbeta))))/
((1-((beta(alfa,k+bbeta))/(beta(alfa,bbeta))))^2))
dp<-sqrt(var) # desvio padrão
marginal<-numeric()
a<-(beta(alfa,k+bbeta))/beta(alfa,bbeta)
for(i in n:g) marginal[i]<-choose(i,n)*(a^(i-n))*((1-a)^(n+1))
plot(marginal)
for(i in n:g) if(sum(marginal[n:i])<=0.025) (j<-i)
IC1<-j # extremo inferior do IC 95%
for(i in n:g) if(sum(marginal[n:i])<=0.975) (s<-i)
IC2<-s # extremo superior do IC 95%
for(i in n:g) if(sum(marginal[n:i])<=0.25) (w<-i)
q1<-w # Q1
for(i in n:g) if(sum(marginal[n:i])<=0.5) (h<-i)
mediana<-h # Q2
for(i in n:g) if(sum(marginal[n:i])<=0.75) (u<-i)
q3<-u # Q3
sumario<-matrix(c(média,dp,IC1,q1,mediana,q3,IC2),nrow=1,ncol=7)
dimnames(sumario)<-list(NULL,c("média","dp","IC1","q1","mediana","q3","IC2"))
print(sumario)
### moda ###
t<-numeric()
for (i in n:(g-1))
{
if(marginal[i]<=marginal[i+1]) (s<-i) else (t[i]<-i)
}
t<-na.omit(t)
moda<-t[1]
moda
```

Apêndice E

Programa para implementação do método de máxima verossimilhanaça condicional para estimação de α e N, para o caso de erros fáceis e difíceis de detectar.

Geração dos dados utilizando o *software* R

```
semente<-100
set.seed(semente)
k<-2
pid<-0.3
pif<-0.8
alfa<-0.1 # valor atribuído em cada caso (os demais são 0.4 e 0.8)
h_linha<-numeric()
h_y_alfa<-function(y) choose(k,y)*((alfa*(pid^y)*((1-pid)^(k-y))+(1-alfa)*(pif^y)*((1-
pif)^(k-y))))
h_0_alfa<-h_y_alfa(y=0)
h_linha_y_alfa<-function(y) choose(k,y)*((alfa*(pid^y)*((1-pid)^(k-y))+(1-alfa)*(pif^y)*((1-
```

```
pif)^(k-y))))/(1-h_0_alfa)
    for(i in 1:k) h_linha[i]<-h_linha_y_alfa(y=i)
    h_linha
    sum(hlinha) # tem que dar 1
    ### Lançamento de uma moeda com probabilidade de cara 1-h(0;alfa) e conseqüente
determinação de n ###
    bin<-numeric()
    n<-0
    N<-100
    for (i in 1:N) {
    bin[i]<-rbinom(1,1,1-h_0_alfa)
    if(bin[i]>0) (n<-n+1)
    }
    n
    ### Geração do vetor (f1,f2,...,fk) ###
    flinha<-numeric()
    flinha<-rmultinom(1,n,h_linha)# vetor (f1,f2,...,fk)
    flinha
    ### Obtenção das EMVC de α e N utilizando o software MAPLE ###
    > restart;
    > k:=2:
    > pid:=0.3:
    > pif:=0.8:
    > f[1]:=36:
    > f[2]:=57:
    > n:=sum('f[y]','y'=1..k):
    > h_y_alfa:=(k!/(y!*(k-y)!))*((alfa*(pid^y)*((1-pid)^(k-y))+(1-alfa)*(pif^y)*((1-pif)^(k-
y)))):
    > L2:=(n!/product( f[y]!, y=1..k ))*product((h_y_alfa/(1-eval(h_y_alfa,y=0)))^f[y],y=1..k):
    > deriv:=diff(L2,alfa):
    > alfa_chapéu_c:=fsolve(deriv,alfa=0..1):
```

```
> h_0_alfa_chapéu_c:=eval(h_y_alfa,[y=0,alfa=alfa_chapéu_c]):
> N_chapéu_c:=evalf(n/(1-h_0_alfa_chapéu_c),3):
```

Apêndice F

Programa para implementação da estimação bayesiana de α e N, para o caso de erros fáceis e difíceis de detectar.

Os dados foram gerados como no apêndice E.

Programa para implementação da estimação bayesiana de alfa e N para o caso de erros fáceis e difíceis de detectar

EMVC versao Bayesiana

pid<-0.3

pif<-0.8

k<-2

kapa<-2

ggama<-0.1 # valor que muda a cada caso (assume também os valores 0.4 e 0.8)

n<-93

f1<-36

f2<-57

Distribuição "a *priori*"para alfa

a<-kapa*ggama

b<-kapa*(1-ggama)

```
funcao_beta_priori<-function(alfa) (gamma(a+b)/(gamma(a)+gamma(b))*(alfa^(a-
1))*(1-alfa)^(b-1))
plot(funcao_beta_priori)
###
### h(y;alfa)=choose(k,y)*((alfa*(pid^y))*((1-pid)^(k-y))+(1-alfa)*((pif^y)*((1-pif)^(k-
y)))
h_0_alfa<-function(alfa) alfa*((1-pid)^(k))+(1-alfa)*((1-pif)^(k))
h_1_alfa<-function(alfa) choose(k,1)*(alfa*(pid^1))*((1-pid)^(k-1))+((1-alfa)*(pif^1)*((1-
pif)^(k-1)))
h_2_alfa<-function(alfa) choose(k,2)*(alfa*(pid^2))*((1-pid)^(k-2))+((1-alfa)*(pif^2)*((1-
pif)^(k-2)))
### Distribuição "a posteriori"de alfa ###
densidade_de_alfa_proporcional<-function(alfa)
(gamma(n+1)/(gamma(f1)*gamma(f2)))*((
(choose(k,1)*((alfa*(pid^1))*((1-pid)^(k-1))+((1-alfa)*(pif^1)*
((1-pif)^(k-1))))/(1-(alfa*((1-pid)^(k))+(1-alfa)*((1-pif)^(k))
))))^f1)*(((choose(k,2)*((alfa*(pid^2))*((1-pid)^(k-2))+((1-alfa)
*(pif^2)*((1-pif)^(k-2))))/(1-(alfa*((1-pid)^(k))+(1-alfa)*((1-pif)^(k))
))))^f2)*(alfa^((kapa*ggama)-1))*((1-alfa)^(kapa*(1-ggama)-1))
constante<-1/integrate(densidade_de_alfa_proporcional,lower=0,upper=1)[[1]]
densidade_de_alfa_exata<-function(alfa)constante*(gamma(n+1)/(gamma(f1)
*gamma(f2)))*(((choose(k,1)*((alfa*(pid^1))
*((1-pid)^(k-1))+((1-alfa)*(pif^1)*((1-pif)^(k-1))))/(1-(alfa*((1-pid)^(k))+(1-alfa)*((1-pif)^(k))
))))^f1)*(((choose(k,2)*((alfa*(pid^2))*((1-pid)^(k-2))+((1-alfa)*(pif^2)*
((1-pif)^(k-2))))/(1-(alfa*((1-pid)^(k))+(1-alfa)*((1-pif)^(k))
))))^f2)*(alfa^((kapa*ggama)-1))*((1-alfa)^(kapa*(1-ggama)-1))
integrate(densidade_de_alfa_exata,lower=0,upper=1) #para confirmar se integra 1
funcao_media<-function(alfa)alfa*constante*(gamma(n+1)
/(gamma(f1)*gamma(f2)))*(((choose(k,1)*((alfa*(pid^1))
*((1-pid)^(k-1))+((1-alfa)*(pif^1)*((1-pif)^(k-1))))/(1-(alfa*((1-pid)^(k))+(1-alfa)*((1-pif)^(k))
))))^f1)*(((choose(k,2)*((alfa*(pid^2))*((1-pid)^(k-2))+((1-alfa)*(pif^2)*
```

```
((1-pif)^(k-2))))/(1-(alfa*((1-pid)^(k))+(1-alfa)*((1-pif)^(k))
    ))))^f2)*(alfa^((kapa*ggama)-1))*((1-alfa)^(kapa*(1-ggama)-1))
    funcao_media_alfa_ao_quadrado<-function(alfa)(alfa^2)*constante*
(gamma(n+1)/(gamma(f1)*gamma(f2)))*(((choose(k,1)*
((alfa*(pid^1))*((1-pid)^(k-1))+((1-alfa)*(pif^1)*((1-pif)^(k-1))))/
(1-(alfa*((1-pid)^(k))+(1-alfa)*((1-pif)^(k))
    ))))^f1)*(((choose(k,2)*((alfa*(pid^2))*((1-pid)^(k-2))+((1-alfa)*(pif^2)
*((1-pif)^(k-2))))/(1-(alfa*((1-pid)^(k))+(1-alfa)*((1-pif)^(k))
    ))))^f2)*(alfa^((kapa*ggama)-1))*((1-alfa)^(kapa*(1-ggama)-1))
    mediaalfa<-integrate(funcao_media,lower=0,upper=1)[[1]]
    media_alfa_ao_quadrado<-integrate(funcao_media_alfa_ao_quadrado,lower=0,upper=1)[[1]]
    variancia<-media_alfa_ao_quadrado - (mediaalfa^2)
    dpalfa<-sqrt(variancia)
    ### Cálculo do IC (95%) e dos quartis Q1, Q2, Q3 ###
    possíveis_valores_de_alfa<-seq(0,1,by=0.01)
    ### IC1 ###
    for (i in 2 : (length (possíveis_valores_de_alfa)-1) )
    {
    if(integrate(densidade_de_alfa_exata,lower=0,upper=possíveis_valores_de_alfa[i])[[1]]
    <=0.025) (j<-possíveis_valores_de_alfa[i])
    }
    IC1alfa<-j # extremo inferior do IC aprox. de 95%
    ### IC2 ###
    for (i in 2 : (length (possíveis_valores_de_alfa)-1) )
    {
    if(integrate(densidade_de_alfa_exata,lower=0,upper=possíveis_valores_de_alfa[i])[[1]]
    <=0.975) (s<-possíveis_valores_de_alfa[i])
    }
    IC2alfa<-s
    Ampli<-IC2alfa-IC1alfa
    ### Q1 ###
```

```
for (i in 2 : (length (possíveis_valores_de_alfa)-1) )
{
if(integrate(densidade_de_alfa_exata,lower=0,upper=possíveis_valores_de_alfa[i])[[1]]
<=0.25) (w<-possíveis_valores_de_alfa[i])
}
Q1alfa<-w
### Q2 ###
for (i in 2: (length (possíveis_valores_de_alfa)-1) )
{
if(integrate(densidade_de_alfa_exata,lower=0,upper=possíveis_valores_de_alfa[i])[[1]]
<=0.50) (q<-possíveis_valores_de_alfa[i])
}
Q2alfa<-q
### Q3 ###
for (i in 2 : (length (possíveis_valores_de_alfa)-1) )
{
if(integrate(densidade_de_alfa_exata,lower=0,upper=possíveis_valores_de_alfa[i])[[1]]
<=0.75) (ç<-possíveis_valores_de_alfa[i])
}
Q3alfa<-ç
### Sumário para alfa ###
sumarioalfa<-matrix(c(mediaalfa,Q1alfa,Q2alfa,Q3alfa,dpalfa,IC1alfa,IC2alfa,Ampli),nrow=1
,ncol=8)
dimnames(sumarioalfa)<-list(NULL,c("mediaalfa","Q1alfa","Q2alfa","Q3alfa","dpalfa","IC1alfa",
"IC2alfa","Ampli"))
print(sumarioalfa)
plot(densidade_de_alfa_exata)
### Distribuição "a posteriori"de N ###
### priori uniforme para N ###
alfachapeu<-mediaalfa
# Binomial negativa n+1 e 1-h_0_alfachapeu (parametrobinneg)
```

```
h_0_alfachapeu<-h_0_alfa(alfa=alfachapeu)
parametrobinneg<-1-h_0_alfachapeu
médiaNU<-n+(((n+1)*h_0_alfachapeu)/parametrobinneg)
dpNU<-(sqrt((n+1)*h_0_alfachapeu))/parametrobinneg
g<-500
marginal<-numeric() #entenda-se por marginal a densidade exata de N
j<-s<-numeric()
for(i in n:g) marginal[i]<-choose(i,n)*(h_0_alfachapeu^(i-n))*((1-h_0_alfachapeu)^(n+1))
plot(marginal)
for(i in n:g) if(sum(marginal[n:i])<=0.025) (j<-i)
IC1NU<-j # extremo inferior do IC 95%
for(i in n:g) if(sum(marginal[n:i])<=0.975) (s<-i)
IC2NU<-s # extremo superior do IC 95%
AmpliNU<-IC2NU-IC1NU
for(i in n:g) if(sum(marginal[n:i])<=0.25) (w<-i)
q1NU<-w # quartil 25%
for(i in n:g) if(sum(marginal[n:i])<=0.5) (h<-i)
q2NU<-h # quartil 50%
for(i in n:g) if(sum(marginal[n:i])<=0.75) (u<-i)
q3NU<-u # quartil 75%
### moda1 ###
t<-numeric()
for (i in n:(g-1))
{
if(marginal[i]<=marginal[i+1]) (s<-i) else (t[i]<-i)
}
t<-na.omit(t)
modaNU<-t[1]
### Sumário para N ###
sumarioNU<-matrix(c(médiaNU,modaNU,q1NU,q2NU,q3NU,dpNU,IC1NU,
IC2NU,AmpliNU),nrow=1,ncol=9)
```

```
dimnames(sumarioNU)<-list(NULL,c("médiaNU","modaNU","q1NU","q2NU",
"q3NU","dpNU","IC1NU","IC2NU","AmpliNU"))
print(sumarioNU)
```

Apêndice G

Programa para implementação do método de estimação de N via distribuição *a posteriori* marginal quase exata no caso de dependência entre revisores.

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 da distribuição *a posteriori* marginal quase exata utilizando o *software* MAPLE

priori para N: Uniforme

```
> 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]):
> l:=2 ^k:
> alpha:=1:
> distintos:=67:
> u:=1000:
> 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):
> 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 400 do
```

```
> if(distN[i]<=distN[i+1]) then s:=i else t[i]:=i end if:
> print(t[i]);
> od:
```

Apêndice H

Programa para implementação do método de estimação de N via algoritmo *Gibbs Sampling* para o caso de dependência entre revisores.

Os dados foram gerados como no Apêndice E.

```
### Algoritmo Gibbs Sampling (supondo priori uniforme para N) ###
semente<-100
set.seed(semente)
M<-1000
bi<-100
salto<-40
numerototal<-bi+M*salto
k<-2
l<-2^k
alpha<-20
bbeta<-1
ni<-c(21,26,20,33)
n<-sum(ni)-ni[length(ni)]
Nverdadeiro<-sum(ni)
```

```
x<-matrix(0,ncol=l,nrow=numerototal)
N<- numeric()
po<- numeric()
x.out<-N.out<-numeric()
po[1]<-0.25
cont<-0
### Primeira Cadeia ###
for(i in 1:numerototal) {
N[i]<- (n+rnbinom(1,n+1,(1-po[i])))
for(j in 1:(l-1)) {
x[i,j] <- rgamma(1,alpha+ni[j],bbeta)}
x[i,l] <- rgamma(1,alpha+N[i]-n,bbeta)
somax <- sum(x[i,])
po[i+1] <- x[i,l]/somax
x[i,] <- x[i,]/somax
cont<-cont+1
print(cont)
}
for (k in 1:numerototal)
{
if ((k>bi) && ((k-bi) %% salto) == 0)
{
N.out<-rbind(N.out,N[k])
x.out<-rbind(x.out,x[k,])
}
}
a11<-N.out
a21<-x.out
### Segunda Cadeia ###
set.seed (50)
l<-2^k
```

```
alpha<-20
bbeta<-1
ni<-c(21,26,20,33)
n<-sum(ni)-ni[length(ni)]
x<-y<-matrix(0,ncol=l,nrow=numerototal)
N<- numeric()
po<- numeric()
x.out<-N.out<-numeric()
po[1]<-0.5 # mudança no valor inicial
for(i in 1:numerototal) {
N[i] <- (n+rnbinom(1,n+1,(1-po[i])))
for(j in 1:(l-1)) {
x[i,j] <- rgamma(1,alpha+ni[j],bbeta) }
x[i,l] <- rgamma(1,alpha+N[i]-n,bbeta)
somax <- sum(x[i,])
po[i+1] <- x[i,l]/somax
x[i,] <- x[i,]/somax
cont<-cont+1
print(cont)
}
for (k in 1:numerototal)
{
if ((k>bi) && ((k-bi) %% salto) == 0)
{
N.out<-rbind(N.out,N[k])
x.out<-rbind(x.out,x[k,])
}
}
a12<-N.out
a22<-x.out
### Analisando a convergência apenas para o parâmetro de interesse N ###
```

```
library(coda)
a1<-mcmc(a11)
a2<-mcmc(a12)
a3<-mcmc.list(a1,a2)
summary(a3)
###
###
gelman.diag(a3)
##
##
##
gelman.plot(a3)
##
##
##
traceplot(a3)
##
##
##
densplot(a3)
##
##
##
autocorr.plot(a3)
##
##
##
hist(a11)
hist(a12)
```

Referências Bibliográficas

[1] AARON, D. J.; CHANG, Y. F.; MARKOVIC, N.; LAPORTE, R. E. Estimating the lesbian population: a capture-recapture approach. *J. Epidemiol. Community Health*, n. 57, p. 207-209, 2003.

[2] BAKER, R. J, NELDER, J. A. GLIM manual (versão 3.77). Oxford, Inglaterra: Royal Statistical Society, 1987.

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

[4] BERNARDO, J. M. Reference posterior distributions for Bayesian inference (with discussion). *J. R. Statist. Soc.*, B n. 41, p. 113-147, 1979.

[5] BOLSONI, S. B. *Estimação dos parâmetros de uma população a partir de observações incompletas da distribuição multinomial*, Dissertação de mestrado, Centro de Ciências Exatas e Tecnologia, Departamento de Estatística, Universidade Federal de São Carlos, 2002.

[6] BRIAND, L. C.; EMAN, K. E.; FREIMUT, B.; LAITENBERGER, O. Quantitative evaluation of capture-recapture models to control *software* inspection, 8^{th} *International Symposium on software Reliability Engineering*, Albuquerque, NM, p. 234-244, 1997.

[7] BURNAM, K. P.; OVERTON, W. S. Estimation of the size of a closed population when capture probabilities vary among animals. *Biometika*, n. 65, p. 625-633, 1978.

[8] CARRENO, A. I.; CRUZ, P. M.; NAVARRO, F. M. The use of the capture-recapture method in evaluating the epidemiological meningococcal disease monitoring system in Tenerife, Spain (1999-2000). *Rev. Esp. Salud Publica*, vol. 77, n. 6, p. 701-711, nov/dez 2003.

[9] CASTLEDINE, B. A Bayesian analysis of multiple recapture sampling from a closed population. *Biometrika*, n. 67, p. 197-210, 1981, 1981.

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

[11] CIORDIA, I. P.; MOROS, A. C.; CANCER, M. F. Tuberculosis in Huesca. Aplicación del método captura-recaptura. *Rev. Esp. Salud Publica*, vol. 73, n. 3, p. 403-406, maio/junho 1999.

[12] COELI, C. M.; COUTINHO, E. S. F.; VERAS, R. P. O desafio da aplicação da metodologia de captura-recaptura na vigilância do diabetes mellitus em idosos: lições de uma experiência no Brasil. *Cad. Saúde Pública*, vol. 20, n. 6, nov/dez 2004.

[13] COUPAL, L.; GYORKOS, T. W; JOSEPH, L. Bayesian estimation of disease prevalence and parameters of diagnostic test in the absence of a gold standard. *Am. Journ. Epidem.*, n. 141, p. 263-272, 2000.

[14] DARROCH, J. N. The multiple recapture census: estimation when there is immigration or death. *Biometrika*, n. 46, p. 336-351, 1959.

[15] EICK, S.G.; LOADER , C.R.; VANDER WIEL, S.A.; VOTTA, L.G., *"How Many Errors Remain in software Design Documents after Inspection?"*Proc. 25th Symp. interface, San Diego, Calif., 1993.

[16] EMAN, K. E.; LAITENBERGER, O. Evaluating capture-recapture models with two inspectors. *IEEE Trans. Soft. Eng.*, vol. 27, n. 9, p. 851-864, 2001.

[17] FELLER, W. An introduction to the theory of probability and its applications, vol. 1, 3rd ed., New York: John Wiley and Sons, 1967.

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

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

[20] JOLLY, G. M. Explicit estimates from captur-recapture data with both death and immigration - stochastic model. *Biometrika*, n. 52, p. 225-247, 1965.

[21] LAPLACE, P. S. Sur les naissances, les marriages et les morts. *Histoire de L ´Académie Royale des Sciences*, p. 1963, 1783.

[22] LAPORTE, R. E.; DEARWATER, S. R.; CHANG, Y.; SONGER, T. J.; AARON, D. J.; ANDERSON, R. L.; OLSEN, T. Efficiency and Accuracy of Disease Monitoring Systems: Application of capture-recapture methods to injury monitoring. *Am. J. Epidemiol.*, n. 142, p. 1069-1077, 1995.

[23] LEE, A. J. Effect of list errors on the estimation of population size. *Biometrics*, n. 58, p. 185-191, 2002.

[24] LEE, A. J.; SEBER, G. A. F.; HOLDEN, J. K.; HUAKAU, J. T. Capture-recapture, epidemiology and list mismatches: serveral lists. *Biometrics*, n. 57, p. 707-713, 2001.

[25] LINCOLN, F. C. Calculating waterfowl abundance on the basis of banding returns. *U. S. Department of Agricultural Circular*, n. 118, p. 1-4, 1930.

[26] MADIGAN, D.; YORK, J. C. Bayesian methods for estimation of the size of a closed population. *Biometrika*, vol. 84, n. 1, p. 19-31, 1997.

[27] MILLER, J. On the independence of *software* inspectors. *J. Syst. Soft.*, n. 60, p. 5-10, 2001.

[28] MINGOTI, S. A. Captura-recaptura: usando estimadores bayesianos para estimar o número total de defeitos distintos em inspeções de produtos. *Rev. Produto & Produção*, vol. 5, n. 3, p. 62-70, 2001.

[29] MISSIAGIA, J. G. Estimação bayesiana do tamanho de uma população de diabéticos através de listas de pacientes., Dissertação de mestrado, Centro de Ciências Exatas e Tecnologia, Departamento de Estatística, Universidade Federal de São Carlos, 2005.

[30] NAYAK, T. K. Estimating population size by recapture sampling. *Biometrika*, n. 75, p. 113-120, 1988.

[31] OIKAWA, S. M. Introdução ao R. Apostila de Minicurso, Universidade Estadual Paulista, Faculdade de Ciências e Tecnologia, Campus de Presidente Prudente, mai/2001.

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

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

[34] ROSSI, R. M. *Estimação bayesiana para o tamanho de uma população multinomial incompleta: uma aplicação a dados obtidos via foto-identificação*, Dissertação de mestrado, Centro de Ciências Exatas e Tecnologia, Departamento de Estatística, Universidade Federal de São Carlos, 2001.

[35] SANATHANAN, L. Models and estimation methods in visual scanning experiments. *Technometrics*, v. 14, n. 4, p. 813-829, 1972.

[36] SEBER, G. A. F. A review of estimating animal abundance. *Biometrics*, n. 42, p. 267-292, 1986.

[37] SEKAR, C. C.;DEMING, W. E. On a method of estimating birth and death rates and the extent of registration. *J. Am. Stat. Assoc.*, n. 44, p. 100-115, 1949.

[38] SHIMIZU, G. Y. *Estimação de uma população fechada: método de captura-recaptura com um único estágio de marcação*, Dissertação de mestrado, Centro de Ciências Exatas e Tecnologia, Departamento de Estatística, Universidade Federal de São Carlos, 2002.

[39] SMITH, P. J. Bayesian Analysis for a multiple capture-recapture model. *Biometrika*, n. 78, p. 399-408, 1991.

[40] SPICHLER, E. R. S.; SPICHLER, D.; LESSA, I. Capture-recapture method to estimate lower extremity amputation rates in Rio de Janeiro, Brazil. *Rev. Panam. Salud Publica*, vol. 10, n. 5, p. 334-340, nov. 2001.

[41] VANDER WIEL, S. A.; VOTTA, L. G. Assessing *software* design using capture-recapture methods. *IEEE Trans. Soft. Eng.*, vol. 19, n. 11, p. 1045-1054, 1993.

[42] YOSHIDA, O. S.; LEITE, J. G.; BOLFARINE, H. Stochastic monotocity properties of Bayes estimation of the population size for capture-recapture data. *Stat. Prob. Lett.*, n. 42, p. 257-266, (1999).

[43] ZACHARIAS, H. P. *Aplicação do algoritmo Gibbs Sampling no processo de captura-recaptura*, Dissertação de mestrado, Centro de Ciências Exatas e Tecnologia, Departamento de Estatística, Universidade Federal de São Carlos, 2000.

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo