Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Estatística
Departamento de Estatística

Inferência Bayesiana para o Tamanho de uma População Fechada com Erros de Registros de Dados Amostrais

Fausto Hideki Oda

Orientador: José Galvão Leite
Coorientador: Luis Aparecido Milan

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
Janeiro de 2008

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

O22ib

Oda, Fausto Hideki.
Inferência bayesiana para o tamanho de uma população fechada com erros de registros de dados amostrais / Fausto Hideki Oda. -- São Carlos : UFSCar, 2008.
67 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2008.

1. Inferência bayesiana. 2. Variável latente. 3. Processo sequencial de captura-recaptura. 4. MCMC. I. Título.

CDD: 519.542 (20ª)

"A viagem da descoberta consiste não em achar novas paisagens, mas em ver com novos olhos."

Agradecimentos

Aos meus pais, Koji Oda e Dalva Fukushima Oda, minhas irmãs, Denise Sayuri Oda Nampo e Cristiante Satie Oda, a minha tia, Mitie Fukushima e meu cunhado Fernado Kenji Nampo, que estiveram em todos os momentos me apoiando, incentivando e orando por mim e foram partes fundamentais nesta jornada.

À toda minha família pela força e apoio.

Aos meus amigos Adriano, Erlandson, Luis Ernesto, Sabrina, Mateus, Danilo, Claudinei, Tiago, pela amizade ajuda e apoio.

À minha turma do mestrado, onde tive o prazer de conviver neste tempo, tanto em momentos de alegrias e festas como em momentos de angústia.

Aos amigos que conheci durante a graduação no IBILCE (UNESP - São José do Rio Preto) onde aprendi muito com eles. Em especial para Chela, Pedrão, Pedrinho, Rodrigão, Fernanda, Luisão e Pitangui.

Aos meus grandes amigos de Tupã, Ivan, Luis Carlos (Pioio) e Lari, Leonardo (Treta) , Valdemar, Adriana, Gustavo, Juliano e Fernando, que apesar da distância sempre estiveram me apoiando e incentivando.

Aos funcionários do departamento de estatística que sempre estiveram prontos a me ajudar com muita vontade.

À CAPES pelo auxílio financeiro.

Aos meus professores do mestrado e graduação, em especial ao professor Dr. Luís Aparecido Milan que sempre me ajudou com muita sabedoria e disposição.

Ao meu orientador José Galvão Leite que me orientou e me ensinou com muita sabedoria e paciência.

Resumo

Nesta dissertação determinamos estimativas de máxima verossimilhança e bayesianas do tamanho de uma população fechada, a partir de duas listas de dados de elementos da população. Supomos que os registros das informações individuais nas listas são passíveis de erros e, com relação ao método bayesiano, as distribuições *a priori* adotadas para os parâmetros são não informativas e de máxima entropia. Apresentamos também um o modelo bayesiano, onde consideramos o número de elementos coincidentes nas duas listas como uma variável latente. Comparamos estes três modelos através de exemplos com dados simulados e reais.

Abstract

In this dissertation we determine maximum likelihood and bayesian estimates of the size of a closed population, from two lists of data of elements of the population. It has been supposed that the registers of the individual information in the lists are capable of mismatches and, with relation to the bayesian method, the prioris distributions are noninformative and they have maximum enthropy for the parameters. We also present the bayesian model, witch has considered the numbers elements of the two lists as a latent variable. We compare these models through examples with simulated and real data.

Sumário

Capítulo 1

Introdução

A estimação do tamanho de uma população através de listas de seus elementos é um assunto extensivamente discutido na literatura estatística. A metodologia aplicada se assemelha à da captura recaptura em população animal. Diversos autores tais como Fienberg *et al* (1999), Seber *et al* (2000), Wang (2002), Micheletti (2003), Missiagia (2005), utilizaram esta metodologia na área da saúde, em especial para estudar a prevalência de uma doença não transmissível como, por exemplo, o diabetes. A idéia é considerar duas ou mais listas de indivíduos portadores da doença como amostras selecionadas da população. Todo indivíduo em qualquer lista é identificado segundo um conjunto de informações tais como, nome, idade, sexo, endereço, número do R.G., número do C.P.F., número da carteira de trabalho, número do plano de saúde, etc. A estimação do tamanho populacional é feita considerando os números de indivíduos das listas e coincidentes em ambas as listas.

Supomos válidas as seguinte condições:

(1) a população é fechada, isto é, não há nascimentos, mortes, emigrações e imigrações durante o período em estudo;

(2) os indivíduos pertencem ou não a qualquer lista independentemente dos demais e das outras listas.

Na prática, a primeira condição é satisfeita se o tempo de aplicação do método for suficientemente curto. A segunda condição pode não se verificar na área médica, por exemplo, se um indivíduo indica a outros pacientes um determinado médico, ou um médico indica a seus pacientes sempre um determinado hospital. Nesta dissertação supomos que

isto não ocorra.

Em suas dissertações de mestrado Micheletti (2003) abordou, sob o enfoque clássico, o problema da estimação do tamanho populacional, com a suposição de registros corretos e incorretos das informações individuais, para duas e múltiplas listas enquanto Missiagia (2005) apresentou uma abordagem bayesiana para o mesmo problema, também considerando registros corretos e incorretos para duas listas. Nessas dissertações, para o caso de registros incorretos, os autores consideraram o número de indivíduos de ambas as listas como uma variável não observável. Desse modo, primeiramente determinaram uma estimativa para este número e daí inferiram sobre o tamanho populacional.

Em particular, em nosso trabalho consideramos duas listas e adotamos um modelo bayesiano, onde o número de indivíduos das duas listas é considerado uma variável latente. Com isto desenvolvemos um modelo que possibilitou estimar sob o enfoque bayesiano o tamanho populacional, sem a necessidade de estimar o número de indivíduos coincidentes das duas listas. Além disso estimamos todos os parâmetros conjuntamente. Por outro lado, determinamos estimativas clássicas e bayesianas para o tamanho populacional e para o número de indivíduos coincidentes, como em Micheletti (2003) e Missiagia (2007), com o objetivo de comparar tais estimativas com as obtidas através do modelo de variável latente.

No Capítulo 2 definimos o modelo probabilístico e determinamos estimativas de máxima verossimilhança (EMV) e intervalos de confiança para os parâmetros do modelo. No Capítulo 3 apresentamos uma solução bayesiana para o problema, onde primeiramente estimamos o número de indivíduos coincidentes nas listas e em seguida inferimos sobre o tamanho populacional. Definimos distribuições *a priori* não informativas e informativas para os parâmetros e determinamos as distribuições condicionais necessárias para a aplicação do algoritmo MCMC (Markov Chain Monte Carlo). No Capítulo 4 aplicamos a metodologia bayesiana de variável latente estimando, deste modo, conjuntamente os parâmetros do modelo. Definimos também distribuições *a priori* não informativas e informativas para os parâmetros e determinamos as probabilidades condicionais. Em cada capítulo da dissertação apresentamos exemplos com dados simulados e reais. Finalmente no Capítulo 5 apresentamos os programas utilizados nas simulações dos exemplos.

Capítulo 2

Estimação do Tamanho Populacional: Duas Listas com Dados Particionados

Neste capítulo, considerando duas listas de indivíduos de uma população fechada de tamanho desconhecido, N, como amostras aleatórias selecionadas da população, propomos um modelo estatístico que leva em conta, em ambas as listas, possíveis erros dos registros dos dados individuais na estimação do parâmetro N.

Na seção 2.1 definimos o modelo estatístico; na seção 2.2 apresentamos as estimativas de máxima verossimilhança (EMV) dos parâmetros do modelo; na seção 2.3 construímos um intervalo de confiança assintótico para N e na seção 2.4 apresentamos exemplos com dados simulados e reais.

2.1 Modelo Estatístico

Suponhamos duas listas de indivíduos de uma população fechada de tamanho desconhecido. Cada indivíduo petencente às listas é identificado através dos registros dos dados individuais como, por exemplo, nome, sobrenome, data de nascimento, sexo, endereço, número do C.P.F., número do R.G. e outros.

Denotamos por

N o tamaho da população;

n_j, o número de indivíduos pertencentes a lista j, $j = 1, 2$;

n_{12}, o número de indivíduos pertencentes a ambas listas 1 e 2;

n, o número de indivíduos distintos pertencentes as duas listas, isto é, $n = n_1 + n_2 - n_{12}$.

Notamos que $N - n$ é o número de indivíduos da população não observados nas listas.

Suponhamos que todos os indivíduos têm a mesma probabilidade θ_j de pertencer a uma lista j, $0 < \theta_j < 1$, $j = 1, 2$ e que cada indivíduo pertence a uma dada lista, independentemente dos demais indivíduos e de pertencer ou não a outra lista.

Ao indivíduo i da população associamos o vetor aleatório $\mathbf{X}_i = (X_{i1}, X_{i2})$, onde

$$X_{ij} = \begin{cases} 1, \text{ se o indivíduo } i \text{ pertence a lista } j, \\ 0, \text{ caso contrário,} \end{cases}$$

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

Então, dado N, os vetores $\mathbf{X}_1$, $\mathbf{X}_2$, ..., $\mathbf{X}_N$, são independentes e assumem valores no conjunto

$$\Omega = \{\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \boldsymbol{\omega}_3, \boldsymbol{\omega}_4\},$$

onde $\boldsymbol{\omega}_1 = (\omega_{11}, \omega_{12}) = (1, 0)$, $\boldsymbol{\omega}_2 = (\omega_{21}, \omega_{22}) = (0, 1)$, $\boldsymbol{\omega}_3 = (\omega_{31}, \omega_{32}) = (1, 1)$ e $\boldsymbol{\omega}_4 = (\omega_{41}, \omega_{42}) = (0, 0)$.

Notamos que $\boldsymbol{\omega}_1$, $\boldsymbol{\omega}_2$, $\boldsymbol{\omega}_3$ e $\boldsymbol{\omega}_4$ representam as trajetórias (histórias) dos indivíduos que pertencem somente a lista 1, somente a lista 2, as duas listas e a nenhuma lista, respectivamente. Denotando por $\boldsymbol{\theta} = (\theta_1, \theta_2)$, a distribuição de probabilidades de $\mathbf{X}_i$, dados N e $\boldsymbol{\theta}$, $i = 1, 2, ..., N$, é dada por

$$p_r(\boldsymbol{\theta}) = P(\mathbf{X}_i = \boldsymbol{\omega}_r \mid N, \boldsymbol{\theta}) = P[(X_{i1}, X_{i2}) = (\omega_{r1}, \omega_{r2}) \mid N, \boldsymbol{\theta}]$$

$$= \prod_{j=1}^{2} P(X_{ij} = \omega_{rj} \mid N, \boldsymbol{\theta}) = \prod_{j=1}^{2} \theta_j^{\omega_{rj}} (1 - \theta_j)^{1 - \omega_{rj}},$$

$r = 1, 2, 3, 4$.

Denotamos por $n_{(r)}$ o número de indivíduos que apresentam a trajetória $\boldsymbol{\omega}_r$, $r = 1, 2, 3, 4$. Notemos que $n = \sum_{r=1}^{3} n_{(r)}$ e $n_{(4)} = N - n$. Logo, dados N e $\boldsymbol{\theta}$, a distribuição de probabilidades do vetor $\left(n_{(1)}, n_{(2)}, n_{(3)}, N - n\right)$ é multinomial com parâmetros N e

$(p_1(\boldsymbol{\theta}), p_2(\boldsymbol{\theta}), p_3(\boldsymbol{\theta}), p_4(\boldsymbol{\theta}))$, isto é,

$$
\begin{aligned}
& P(n_{(1)}, n_{(2)}, n_{(3)}, N-n \mid N, \boldsymbol{\theta}) \\
&= \frac{N!}{n_{(1)}!n_{(2)}!n_{(3)}!\,(N-n)!}\,[p_1(\boldsymbol{\theta})]^{n_{(1)}}\,[p_2(\boldsymbol{\theta})]^{n_{(2)}}\,[p_3(\boldsymbol{\theta})]^{n_{(3)}}\,[p_4(\boldsymbol{\theta})]^{N-n} \\
&= \frac{N!}{n_{(1)}!n_{(2)}!n_{(3)}!\,(N-n)!}\prod_{r=1}^{4}[p_r(\boldsymbol{\theta})]^{n_{(r)}} \\
&= \frac{N!}{n_{(1)}!n_{(2)}!n_{(3)}!\,(N-n)!}\prod_{r=1}^{4}\left[\prod_{j=1}^{2}\theta_j^{\omega_{rj}}(1-\theta_j)^{1-\omega_{rj}}\right]^{n_{(r)}} \qquad (2.1) \\
&= \frac{N!}{n_{(1)}!n_{(2)}!n_{(3)}!\,(N-n)!}\prod_{j=1}^{2}\prod_{r=1}^{4}\theta_j^{n_{(r)}\omega_{rj}}(1-\theta_j)^{n_{(r)}(1-\omega_{rj})} \\
&= \frac{N!}{n_{(1)}!n_{(2)}!n_{(3)}!\,(N-n)!}\prod_{j=1}^{2}\theta_j^{\sum_{r=1}^{4}n_{(r)}\omega_{rj}}(1-\theta_j)^{\sum_{r=1}^{4}n_{(r)}(1-\omega_{rj})}.
\end{aligned}
$$

Como $n_j = \sum_{r=1}^{4} n_{(r)}\omega_{rj}$ e $N - n_j = \sum_{r=1}^{4} n_{(r)}(1-\omega_{rj})$, $j = 1, 2$, segue de (2.1) que

$$
\begin{aligned}
& P(n_{(1)}, n_{(2)}, n_{(3)}, N-n \mid N, \boldsymbol{\theta}) \\
&= \frac{N!}{n_{(1)}!n_{(2)}!n_{(3)}!\,(N-n)!}\prod_{j=1}^{2}\theta_j^{n_j}(1-\theta_j)^{N-n_j}.
\end{aligned} \qquad (2.2)
$$

Por outro lado, como $n_{(1)} = n_1 - n_{12}$, $n_{(2)} = n_2 - n_{12}$ e $n_{(3)} = n_{12}$, temos de (2.2) que

$$
\begin{aligned}
& P(n_1, n_2, n_{12} \mid N, \boldsymbol{\theta}) \\
& = P(n_{(1)} = n_1 - n_{12}, n_{(2)} = n_2 - n_{12}, n_{(3)} = n_{12}, n_{(4)} = N - n \mid N, \boldsymbol{\theta}) \\
& = \frac{N!}{(n_1 - n_{12})!\,(n_2 - n_{12})! n_{12}!\,(N-n)!} \prod_{j=1}^{2} \theta_j^{n_j} (1-\theta_j)^{N-n_j} \\
& = \binom{n_1}{n_{12}} \binom{N}{n_1} \binom{N-n_1}{n_2 - n_{12}} \prod_{j=1}^{2} \theta_j^{n_j} (1-\theta_j)^{N-n_j}.
\end{aligned}
\tag{2.3}
$$

Portanto, denotando por $D_1 = (n_1, n_2, n_{12})$, segue de (2.3) que a função de verossimilhança é tal que

$$
L_1\left(N, \boldsymbol{\theta} \mid D_1\right) \propto \frac{N!}{(N-n)!} \prod_{j=1}^{2} \theta_j^{n_j} (1-\theta_j)^{N-n_j}, \tag{2.4}
$$

$N \geq n$ e $0 < \theta_j < 1$, $j = 1, 2$.

Em uma situação ideal, ou seja, em uma situação em que as informações dos indivíduos pertencentes as listas são registradas corretamente, o parâmetro n_{12} é conhecido e, de (2.4), podemos determinar as EMV de N e θ_j, $j = 1, 2$.

Contudo, nesta dissertação vamos admitir a possibilidade de que alguns dados sejam registrados incorretamente, bem como o caso onde alguns indivíduos respondam de forma intencional ou erroneamente algumas questões formuladas. Logo n_{12} passa a ser um valor não observado de uma variável aleatória, ou uma variável latente, e o vetor de dados $D_1 = (n_1, n_2, n_{12})$ um vetor de "dados ampliados". Neste caso a estratégia que adotamos neste capítulo para estimar N e θ_j, $j = 1, 2$, é determinar, segundo o modelo estatístico denominado modelo com dados particionados (Seber *et. al* (2000)), uma estimativa, $\widehat{n_{12}}$, de n_{12} e substituir n_{12} por $\widehat{n_{12}}$ em (2.4) obtendo a função de verossimilhança estimada $\widehat{L}_1\left(N, \boldsymbol{\theta}\right) = L_1\left(N, \boldsymbol{\theta} \mid n_1, n_2, \widehat{n_{12}}\right)$ e, em seguida, estimar N e θ_j, $j = 1, 2$, a partir de $\widehat{L}_1$. Na seqüência descrevemos o modelo com dados particionados que nos permitirá estimar n_{12}. Suponhamos que o conjunto dos dados identificadores de cada indivíduo, em qualquer lista, seja dividido em dois subconjuntos A e B, que denominaremos de fichas. Uma

vez definidas as fichas A e B, vamos supor que se as fichas A ou B forem preenchidas corretamente para um indivíduo, então ele é identificado de modo único e para decidir se dois indivíduos, um da lista 1 e um da lista 2, são coincidentes ou não, comparamos os dados das respectivas fichas, A e B. Se estes dados forem iguais para pelo menos uma das fichas, então decidimos que as fichas coincidentes foram preenchidas corretamente e, portanto, os indivíduos são coincidentes.

Notamos que com esta suposição descartamos a possibilidade da ocorrência de uma coincidência casual de dois indivíduos, como por exemplo, dois indivíduos distintos com mesmo nome acabam sendo diferenciados pelo R.G., ou, no caso em que algum dado é registrado incorretamente outra informação pode identificar este indivíduo. Isto implica que a ficha A ou B não deve ser constituída, por exemplo, apenas pelo número do R.G., cujo registro errado de um de seus algarismos poderia implicar em uma falsa coincidência. Suponhamos também que

(1) as fichas A e B sejam preenchidas independentemente entre os indivíduos de cada lista e entre as listas;

(2) para cada indivíduo em qualquer lista o preenchimento da ficha A ou B é independente do preenchimento da outra ficha, B ou A.

Observamos pela suposição (2), que as composições das fichas A e B devem ser tais que os dados individuais mais prováveis de serem registrados incorretamente, porque são complicados, devem pertencer a uma mesma ficha, A ou B.

Suponhamos que a probabilidade de a ficha X ser preenchida corretamente seja a mesma para todos os indivíduos em qualquer lista, $X = \{A,\ B\}$. Isto implica que a probabilidade de que a ficha X seja preenchida corretamente em ambas as listas é a mesma para todos os indivíduos de ambas as listas. Denotemos então por ϕ_X a probabilidade de que a ficha X de um indivíduo pertencente às duas listas seja preenchida corretamente nas duas listas. O conjunto das possíveis trajetórias associadas a cada indivíduo pertencente a ambas as listas pode ser descrito como o conjunto

$$\{AO, AO;\ AO, OB;\ AO, AB;\ AO, OO;\ AB, AO;\ AB, OB;\ AB, AB;\ AB, OO;$$
$$OB, AO;\ OB, OB;\ OB, AB;\ OB, OO;\ OO, AO;\ OO, OB;\ OO, AB;\ OO, OO\},$$

onde, o par (AO, AO) significa que as fichas A's do indivíduo foram preenchidas corretamete em ambas as listas e as fichas B's foram preenchidas incorretamente em ambas as listas; (AO, OB), significa que a ficha A foi preenchida corretamente na lista 1 e incorretamente na lista 2, ao passo que a ficha B foi preenchida incorretamente na lista 1 e corretamente na lista 2 e assim por diante. Notamos que as freqüências dos indivíduos que apresentam certas trajetórias não são observáveis como, por exemplo, as freqüências dos indivíduos que apresentam as trajetórias (AO, AO) e (AO, AB), pois os indivíduos que apresentam trajetória (AO, AO) são indistingüíveis dos indivíduos que apresentam a trajetória (AO, AB).

Sejam $n_{AB,AO}$, $n_{AO,AB}$, $n_{AO,AO}$, $n_{AB,OB}$, $n_{OB,AB}$, $n_{OB,OB}$ e $n_{AB,AB}$ o número de indivíduos que apresentam as respectivas trajetórias AB, AO; AO, AB; AO, AO; AB, OB; OB, AB; OB, OB e AB, AB. Denotamos por

$n_A = n_{AB,AO} + n_{AO,AB} + n_{AO,AO}$, o número (observado) de indivíduos que possuem apenas as fichas A's coincidentes em ambas as listas;

$n_B = n_{AB,OB} + n_{OB,AB} + n_{OB,OB}$ o número (observado) de indivíduos que possuem apenas as fichas B's coincidentes em ambas as listas;

$n_{AB} = n_{AB,AB}$ o número (observado) de indivíduos que possuem as fichas A's e B's coincidentes em ambas as listas;

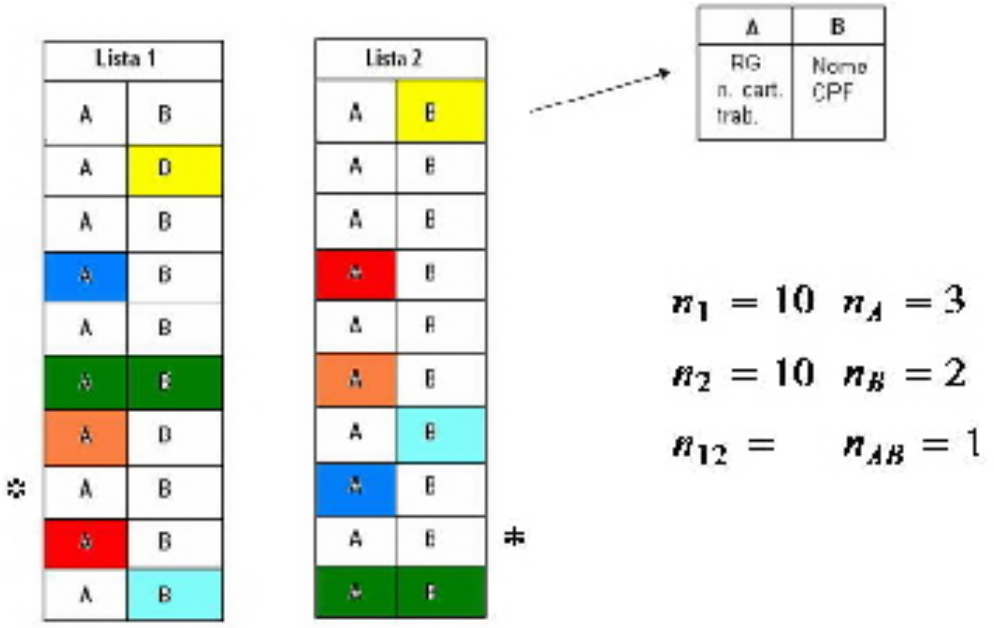
$n_T = n_A + n_B + n_{AB}$ o número (observado) de indivíduos coincidentes nas listas.

Notamos que, $n_{12} - n_T$ é o número (não observado) de indivíduos que estão em ambas as listas para os quais nenhuma das fichas são coincidentes, ou seja, são indivíduos que pertencem as duas listas, mas não foram identificados devido à erros de preenchimento das fichas.

A seguir apresentamos um exemplo para melhor compreender as notações de números de elementos nas listas e fichas.

Exemplo 2.1.1. Consideremos duas listas de indivíduos da população onde cada lista possui 10 elemnetos e para cada elemento dividimos as informações cadastrais em duas fichas A e B. Na ficha A colocamos o R.G. e o número da carteira de trabalho e na ficha B o nome e o CPF. Na figura abaixo temos que os indivíduos que possuem as fichas A's e B's coloridas de mesma cor são os indivíduos que são identificados por terem as fichas A's e B's preenchidas corretamente. Analogamente para os que possuem somente as fichas

A's ou B's coloridas. O indivíduo que está marcado com uma estrela ao lado representa um indivíduo perdido, isto é, possui ambas as fichas preenchidas incorretamente. Deste modo, o número de indivíduos pertencentes a ambas as listas se torna uma variável não observável.



Pelas suposições feitas anteriormente temos que $(n_A, n_B, n_{AB}, n_{12} - n_T)$, dados n_{12} e $\boldsymbol{\phi} = (\phi_A, \phi_B)$, tem distribuição multinomial com parâmetros n_{12} e $(\phi_A\,(1-\phi_B)\,,\,(1-\phi_A)\,\phi_B,$ $\phi_A\phi_B,\ (1-\phi_A)\,(1-\phi_B))$, isto é,

$$
\begin{aligned}
& P(n_A, n_B, n_{AB}, n_{12} - n_T \mid n_{12}, \boldsymbol{\phi}) \\
& = \frac{n_{12}!}{n_A! n_B! n_{AB}! (n_{12} - n_T)!} \left[\phi_A (1 - \phi_B)\right]^{n_A} \left[(1 - \phi_A) \phi_B\right]^{n_B} \left[\phi_A \phi_B\right]^{n_{AB}} \\
& \times \left[(1 - \phi_A)(1 - \phi_B)\right]^{n_{12} - n_T} \\
& = \frac{n_{12}!}{n_A! n_B! n_{AB}! (n_{12} - n_T)!} \phi_A^{n_A + n_{AB}} (1 - \phi_A)^{n_B + n_{12} - n_T} \phi_B^{n_B + n_{AB}} (1 - \phi_B)^{n_A + n_{12} - n_T} \\
& = \frac{n_{12}!}{n_A! n_B! n_{AB}! (n_{12} - n_T)!} \phi_A^{m_A} (1 - \phi_A)^{n_{12} - m_A} \phi_B^{m_B} (1 - \phi_B)^{n_{12} - m_B},
\end{aligned} \tag{2.5}
$$

onde $m_A = n_A + n_{AB}$ é o número de indivíduos que possuem as fichas A 's coincidentes em ambas as listas e $m_B = n_B + n_{AB}$ é o número de indivíduos que possuem as fichas B 's coincidentes em ambas as listas. Denotando por $D_2 = (m_A, m_B, n_{AB})$ segue, de (2.5), que D_2 é uma estatística suficiente para $(n_{12}, \boldsymbol{\phi})$ e a função de verossimilhança é tal que

$$
L_2(n_{12}, \boldsymbol{\phi} \mid D_2) \propto \frac{n_{12}!}{(n_{12} - n_T)!} \phi_A^{m_A} (1 - \phi_A)^{n_{12} - m_A} \phi_B^{m_B} (1 - \phi_B)^{n_{12} - m_B}, \tag{2.6}
$$

para $m_A + m_B - n_{AB} = n_T \leq n_{12} \leq \min\{n_1, n_2\}$ e $0 < \phi_X < 1$, $X = A, B$.

Na próxima seção apresentamos as EMV de n_{12}, ϕ_X, $X = A, B$; N e θ_j, $j = 1, 2$.

2.2 Estimadores de Máxima Verossimilhança dos Parâmetros do Modelo

Nesta seção determinamos a EMV de n_{12}, $\widehat{n_{12}}$, substituímos n_{12} por $[\widehat{n_{12}}]$ em (2.4), onde $[\widehat{n_{12}}]$ é igual ao maior número inteiro menor ou igual a $\widehat{n_{12}}$, determinando assim a função de verossimilhança estimada $\widehat{L}_1(N, \boldsymbol{\theta}) = L_1(N, \boldsymbol{\theta} \mid n_1, n_2, [\widehat{n_{12}}])$, a partir da qual determinamos as EMV de N e θ_j, $j = 1, 2$.

As EMV de n_{12}, ϕ_A e ϕ_B são dadas pelo seguinte teorema.

Teorema 2.2.1. As EMV de n_{12}, ϕ_A e ϕ_B são, respectivamente, $\widehat{n_{12}} = \dfrac{m_A m_B}{n_{AB}}$, $\widehat{\phi_A} = \dfrac{n_{AB}}{m_B}$ e $\widehat{\phi_B} = \dfrac{n_{AB}}{m_A}$.

Prova: Seja $K(n_{12}, \boldsymbol{\phi})$ o *Kernel* de $L_2(n_{12}, \boldsymbol{\phi} \mid D_2)$. Segue, de (2.6), que

$$K(n_{12}, \boldsymbol{\phi}) = \frac{n_{12}!}{(n_{12} - n_T)!} \phi_A^{m_A} (1 - \phi_A)^{n_{12} - m_A} \phi_B^{m_B} (1 - \phi_B)^{n_{12} - m_B},$$

$n_T \leq n_{12} \leq \min\{n_1, n_2\}$ e $0 < \phi_X < 1$, $X = A, B$, o que implica

$$\ln K(n_{12}, \boldsymbol{\phi}) = \ln(n_{12}!) - \ln((n_{12} - n_T)!) + m_A \ln \phi_A + (n_{12} - m_A) \ln(1 - \phi_A)$$

$$+ (m_B) \ln \phi_B + (n_{12} - m_B) \ln(1 - \phi_B).$$

O ponto de máximo, $\left(\widehat{n_{12}}, \widehat{\boldsymbol{\phi}}\right)$, de $K(n_{12}, \boldsymbol{\phi})$ ou de $\ln K(n_{12}, \boldsymbol{\phi})$ satisfaz o sistema de equações

$$\frac{\partial \ln K(n_{12}, \boldsymbol{\phi})}{\partial \phi_A} = 0,$$

$$\frac{\partial \ln K(n_{12}, \boldsymbol{\phi})}{\partial \phi_B} = 0,$$

$$\ln K(n_{12}, \boldsymbol{\phi}) - \ln K(n_{12} - 1, \boldsymbol{\phi}) = 0.$$

Resolvendo este sistema, temos

$$\frac{\partial \ln K(n_{12}, \boldsymbol{\phi})}{\partial \phi_A} = \frac{m_A}{\phi_A} - \frac{n_{12} - m_A}{1 - \phi_A} = 0 \Rightarrow \phi_A = \frac{m_A}{n_{12}},$$

$$\frac{\partial \ln K(n_{12}, \boldsymbol{\phi})}{\partial \phi_B} = \frac{m_B}{\phi_B} - \frac{n_{12} - m_B}{1 - \phi_B} = 0 \Rightarrow \phi_B = \frac{m_B}{n_{12}},$$

e n_{12} satisfaz a equação

$$\ln K\left(n_{12}, \boldsymbol{\phi}\right) - \ln K\left(n_{12}-1, \boldsymbol{\phi}\right) = \ln n_{12} - \ln\left(n_{12}-n_T\right) + \ln\left(1-\phi_A\right) + \ln\left(1-\phi_B\right) = 0$$

$$\Rightarrow \ln\left(\left(\frac{n_{12}}{n_{12}-n_T}\right)\left(1-\phi_A\right)\left(1-\phi_B\right)\right) = 0$$

$$\Rightarrow \left(\frac{n_{12}}{n_{12}-n_T}\right)\left(1-\phi_A\right)\left(1-\phi_B\right) = 1.$$

Substituindo-se ϕ_A e ϕ_B por $\frac{m_A}{n_{12}}$ e $\frac{m_B}{n_{12}}$, respectivamente, na equação acima temos

$$\left(\frac{n_{12}}{n_{12}-n_T}\right)\left(1-\frac{m_A}{n_{12}}\right)\left(1-\frac{m_B}{n_{12}}\right) = 1$$

$$\Rightarrow \left(n_{12}-m_A\right)\left(n_{12}-m_B\right) = n_{12}\left(n_{12}-n_T\right)$$

$$\Rightarrow n_{12}\left(m_A+m_B-n_T\right) = m_A m_B$$

$$\Rightarrow n_{12} = \frac{m_A m_B}{n_{AB}}.$$

Portanto,

$$\widehat{n_{12}} = \frac{m_A m_B}{n_{AB}}$$

$$\widehat{\phi_A} = \frac{m_A}{\widehat{n_{12}}} = \frac{n_{AB}}{m_B}$$

$$\widehat{\phi_B} = \frac{m_B}{\widehat{n_{12}}} = \frac{n_{AB}}{m_A},$$

o que prova o teorema. ■

Substituindo-se n_{12} por $[\widehat{n_{12}}]$ determinado no Teorema 2.2.1 em (2.4) , temos a função de verossimilhança estimada

$$\begin{aligned}\widehat{L}_1(N,\boldsymbol{\theta}) &= L_1(N,\boldsymbol{\theta}\,|n_1,n_2,[\widehat{n_{12}}]) \\ &\propto \frac{N!}{(N-n_1-n_2+[\widehat{n_{12}}])!}\prod_{j=1}^{2}\theta_j^{n_j}(1-\theta_j)^{N-n_j},\end{aligned} \tag{2.7}$$

$N \geq n_1+n_2-[\widehat{n_{12}}]$ e $0<\theta_j<1$, $j=1,2$.

As EMV de N, θ_1 e θ_2 em relação a $\widehat{L}_1(N,\boldsymbol{\theta})$ são dadas pelo seguinte teorema.

Teorema 2.2.2. As EMV de N, θ_1 e θ_2 em relação a (2.7) são, respectivamente, $\widehat{N}=\dfrac{n_1n_2}{[\widehat{n_{12}}]}$, $\widehat{\theta_1}=\dfrac{[\widehat{n_{12}}]}{n_2}$, $\widehat{\theta_2}=\dfrac{[\widehat{n_{12}}]}{n_1}$.

Prova: A prova deste teorema é análoga ao do Teorema 2.2.1. ■

Observamos que a estimativa $\widehat{n_{12}}$ não está definida se $m_A=0$, $m_B=0$ ou $m_{AB}=0$, o que implica que a estimativa $\widehat{N}$ também não está definida. Esta situação pode ser solucionada se considerarmos o estimador

$$N^*=\frac{(n_1+1)(n_2+1)}{(n_T+1)}\left\{1-\frac{n_An_B}{n_T(n_{AB}+1)}\right\}-1, \tag{2.8}$$

que é aproximadamente não viciado se ϕ_A e ϕ_B forem suficientemente grandes. O estimador N^* foi proposto por Seber e Felton (1981).

2.3 Construção de um Intervalo de Confiança Assintótico para o Tamanho Populacional

Nesta seção construímos um intervalo de confiança assintótico para N. De (2.3) segue que

$$P(n_{12}\mid n_1,n_2,N,\boldsymbol{\theta})=\frac{P(n_1,n_2,n_{12}\mid N,\boldsymbol{\theta})}{P(n_1,n_2\mid N,\boldsymbol{\theta})}, \tag{2.9}$$

onde a distribuição de probabilidades de n_1 e n_2 dados N e $\boldsymbol{\theta}$, é dada por

$$P(n_1, n_2 \mid N, \boldsymbol{\theta}) = \sum_{n_{12}} P(n_1, n_2, n_{12} \mid N, \boldsymbol{\theta})$$

$$= \sum_{n_{12}} \binom{n_1}{n_{12}} \binom{N}{n_1} \binom{N-n_1}{n_2-n_{12}} \prod_{j=1}^{2} \theta_j^{n_j} (1-\theta_j)^{N-n_j}$$

$$= \binom{N}{n_1} \prod_{j=1}^{2} \theta_j^{n_j} (1-\theta_j)^{N-n_j} \sum_{n_{12}} \binom{n_1}{n_{12}} \binom{N-n_1}{n_2-n_{12}}$$

$$= \binom{N}{n_2} \binom{N}{n_1} \prod_{j=1}^{2} \theta_j^{n_j} (1-\theta_j)^{N-n_j} \sum_{n_{12}} \frac{\binom{n_1}{n_{12}} \binom{N-n_1}{n_2-n_{12}}}{\binom{N}{n_2}}$$

$$= \binom{N}{n_2} \binom{N}{n_1} \prod_{j=1}^{2} \theta_j^{n_j} (1-\theta_j)^{N-n_j}.$$

Logo, (2.9) é dada por

$$P(n_{12} \mid n_1, n_2, N, \boldsymbol{\theta}) = \frac{\binom{n_1}{n_{12}} \binom{N-n_1}{n_2-n_{12}}}{\binom{N}{n_2}},$$

isto é, a função de probabilidades condicional de n_{12}, dados n_1, n_2, N e $\boldsymbol{\theta}$ é hipergeométrica de parâmetros N, n_1 e n_2.

Suponhamos agora que $\frac{n_2}{N} \simeq 0$, isto é, a magnitude de n_2 seja pequena quando comparado à de N. Então,

$$P(n_{12} \mid n_1, n_2, N, \boldsymbol{\theta}) \simeq \binom{n_2}{n_{12}} \left(\frac{n_1}{N}\right)^{n_{12}} \left(1 - \frac{n_1}{N}\right)^{n_2-n_{12}},$$

ou seja, a função de probabilidades condicional de n_{12}, dados n_1, n_2, N e $\boldsymbol{\theta}$, é aproximadamente igual a função de probabilidade binomial com parâmetros n_2 e $\frac{n_1}{N}$. Na prática podemos considerar a lista com menor número de indivíduos como sendo a lista 2.

Pelo Teorema Central do Limite sabemos que, dados n_1 e n_2 com $\frac{n_2}{N} \simeq 0$,

$$\frac{n_{12} - \frac{n_2 n_1}{N}}{\sqrt{\frac{n_2 n_1}{N}\left(1 - \frac{n_1}{N}\right)}} \mid n_1, n_2$$

tem aproximadamente distribuição normal com média 0 e variância 1. Logo, um intervalo de confiança com coeficiente de confiança $(1-\alpha)\,100\%$ para o parâmetro $\frac{n_1}{N}$ é dado por

$$\left(\frac{n_{12}}{n_2} + z_{\alpha/2}\sqrt{\frac{n_1}{n_2 N}\left(1 - \frac{n_1}{N}\right)}, \frac{n_{12}}{n_2} - z_{\alpha/2}\sqrt{\frac{n_1}{n_2 N}\left(1 - \frac{n_1}{N}\right)}\right), \tag{2.10}$$

onde $z_{\alpha/2}$ é o quantil $\frac{\alpha}{2}$ da distribuição normal padrão. Substituindo-se n_{12} por $\widehat{n_{12}}$ e N por $\widehat{N} \cong \frac{n_1 n_2}{\widehat{n_{12}}}$ em (2.10), segue que

$$\left(\frac{\widehat{n_{12}}}{n_2} + z_{\alpha/2}\sqrt{\frac{\widehat{n_{12}}\left(n_2 - \widehat{n_{12}}\right)}{\left(n_2\right)^3}}, \frac{\widehat{n_{12}}}{n_2} - z_{\alpha/2}\sqrt{\frac{\widehat{n_{12}}\left(n_2 - \widehat{n_{12}}\right)}{\left(n_2\right)^3}}\right)$$

é um intervalo de confiança aproximado com coeficiente de confiança $(1-\alpha)100\%$ para $\frac{n_1}{N}$ o que implica que

$$\left(\frac{n_1 n_2}{\widehat{n_{12}} - z_{\alpha/2}\sqrt{\frac{\widehat{n_{12}}\left(n_2 - \widehat{n_{12}}\right)}{n_2}}}, \frac{n_1 n_2}{\widehat{n_{12}} + z_{\alpha/2}\sqrt{\frac{\widehat{n_{12}}\left(n_2 - \widehat{n_{12}}\right)}{n_2}}}\right), \tag{2.11}$$

é um intervalo de confiança aproximado com coeficiente de confiança $(1-\alpha)100\%$ para N, onde $\widehat{n_{12}} = \frac{m_A m_B}{n_{AB}}$.

Na próxima seção apresentamos exemplos com dados simulados e reais. Através dos exemplos com dados simulados fazemos um estudo da performance das estimativas dos parâmetros e do intervalo de confiança (2.11).

2.4 Exemplos com Dados Simulados e Reais

Nesta seção apresentamos exemplos com dados simulados e reais e analisamos as performances das estimativas dos parâmetros do modelo e do intervalo de confiança para N. Os programas utilizados nesta dissertação para gerar dados e determinar as estimativas dos parâmetros foram implementados via *software R-Gui* (versão 2.3.0) e são apresentados nos apêndices.

2.4.1 Exemplo com Dados Simulados

Neste exemplo atribuímos valores para N, θ_1 e θ_2; geramos o valor de um vetor aleatório com distribuição multinomial de parâmetros N e $(\theta_1(1-\theta_2)$, $(1-\theta_1)\theta_2$, $\theta_1\theta_2$, $(1-\theta_1)(1-\theta_2))$, obtendo as quantidades n_1, n_2 e n_{12}. Atribuímos valores para ϕ_A e ϕ_B gerando novamente da distribuição multinomial, agora com parâmetros n_{12} e $(\phi_A(1-\phi_B)$, $(1-\phi_A)\phi_B$, $\phi_A\phi_B$, $(1-\phi_A)(1-\phi_B))$, as quantidades n_A, n_B e n_{AB}.

Exemplo 2.4.1 A Tabela 2.4.1 contém os valores obtidos pelas simulações para diferentes valores de N, θ_1, θ_2, ϕ_A e ϕ_B.

Tabela 2.4.1 Quantidades geradas da distribuição multinomial

Valores fixados					Valores gerados					
N	θ_1	θ_2	ϕ_A	ϕ_B	n_1	n_2	n_{12}	n_A	n_B	n_{AB}
50	0,7	0,8	0,7	0,9	33	38	25	2	7	15
50	0,8	0,7	0,2	0,1	37	33	22	2	4	0
50	0,1	0,3	0,7	0,7	3	19	2	1	1	0
50	0,1	0,3	0,2	0,1	3	19	2	1	1	0
500	0,6	0,8	0,7	0,8	299	403	235	38	73	115
500	0,9	0,7	0,3	0,2	453	344	307	74	45	20
500	0,1	0,4	0,8	0,6	46	203	18	4	1	11
500	0,2	0,3	0,1	0,2	101	159	30	3	3	1
2000	0,6	0,8	0,7	0,9	1173	1576	905	67	238	576
2000	0,5	0,6	0,4	0,3	947	1168	528	157	95	68
2000	0,2	0,4	0,7	0,8	424	827	156	19	37	93
2000	0,2	0,1	0,3	0,4	411	209	39	9	11	1

A Tabela 2.4.2 apresenta valores aproximados das estimativas dos parâmetros do modelo e a Tabela 2.4.3 contém intervalos de confiança aproximados com coeficiente de confiança de 95% para N.

Tabela 2.4.2 Valores aproximados das estimativas dos parâmetros

Valores fixados e gerados						Estimativas aproximadas						
N	θ_1	θ_2	ϕ_A	ϕ_B	n_{12}	$\widehat{N}$	$[\widehat{n_{12}}]$	$\widehat{\theta_1}$	$\widehat{\theta_2}$	$\widehat{\phi_A}$	$\widehat{\phi_B}$	N^*
50	0,7	0,8	0,7	0,9	25	52	24	0,632	0,727	0,682	0,882	50
50	0,8	0,7	0,2	0,1	22	–	–	–	–	–	–	−63
50	0,1	0,3	0,7	0,7	2	–	–	–	–	–	–	12
50	0,1	0,3	0,2	0,1	2	–	–	–	–	–	–	11
500	0,6	0,8	0,7	0,8	235	482	250	0,62	0,836	0,612	0,752	476
500	0,9	0,7	0,3	0,2	307	511	305	0,887	0,673	0,308	0,213	−159
500	0,1	0,4	0,8	0,6	18	584	16	0,079	0,348	0,917	0,733	551
500	0,2	0,3	0,1	0,2	30	1004	16	0,101	0,158	0,25	0,25	728
2000	0,6	0,8	0,7	0,9	905	2036	908	0,576	0,774	0,708	0,896	2032
2000	0,5	0,6	0,4	0,3	528	2052	539	0,461	0,569	0,417	0,302	1119
2000	0,2	0,4	0,7	0,8	156	2248	156	0,189	0,368	0,715	0,83	2227
2000	0,2	0,1	0,3	0,4	39	716	120	0,574	0,292	0,083	0,1	−5338

Tabela 2.4.3 Intervalos de confiança aproximados de 95% de confiança para N

N	θ_1	θ_2	ϕ_A	ϕ_B	$\widehat{N}$	I.C.(95%)	Ampl.
50	0,7	0,8	0,7	0,9	52	(42; 69)	27
50	0,8	0,7	0,2	0,1	–	–	–
50	0,1	0,3	0,7	0,7	–	–	–
50	0,1	0,3	0,2	0,1	–	–	–
500	0,6	0,8	0,7	0,8	482	(448; 522)	74
500	0,9	0,7	0,3	0,2	511	(492; 531)	39
500	0,1	0,4	0,8	0,6	584	(397; 1102)	705
500	0,2	0,3	0,1	0,2	1004	(685; 1875)	1190
2000	0,6	0,8	0,7	0,9	2036	(1953; 2126)	173
2000	0,5	0,6	0,4	0,3	2052	(1932; 2188)	255
2000	0,2	0,4	0,7	0,8	2248	(1969; 2618)	648
2000	0,2	0,1	0,3	0,4	716	(641; 810)	169

Na Tabela 2.4.3 denotamos por

I.C.(95%), o intervalo de confiança com 95% de confiança;

Ampl., a amplitude do intervalo de confiança.

Notamos, pelas Tabelas 2.4.2 e 2.4.3, que para valores pequenos de N, θ_1, θ_2, ϕ_A e ϕ_B as estimativas dos parâmetros e os intervalos de confiança de N não são razoáveis e, em alguns casos, as EMV dos parâmetros e os intervalos de confiança não existem.

Na próxima seção damos um exemplo com dados reais.

2.4.2 Exemplo com Dados Reais

Nesta seção apresentamos um exemplo com dados reais obtidos de um estudo realizado no Sul da Nova Zelândia (Seber *et al*, (2000)), sobre pacientes portadores do diabetes.

Exemplo 2.4.2 Foram identificados na primeira lista 4186 pacientes junto a médicos da região e 2203 pacientes na segunda lista identificados independentemente por um estudo caseiro. Deste modo temos $n_1 = 4186$ e $n_2 = 2203$. Para cada indivíduo foi registrado as

seguintes informações: primerio nome, sobrenome, idade, data de nascimento, sexo, rua e bairro. Foram incluídos o primeiro nome, sobrenome e idade na ficha A e o restante na ficha B. Com esta divisão das informações eles esperavam ter fichas independentes com relação aos erros. Foram observados $m_A = 298$, $m_B = 231$ e $n_{AB} = 116$. Assim, obtemos as seguintes estimativas aproximadas dos parâmetros.

Tabela 2.4.4 Valores aproximados das estimativas dos parâmetros e intervalo de confiança aproximado de 95% para N

$\widehat{N}$	$[\widehat{n_{12}}]$	$\widehat{\theta_1}$	$\widehat{\theta_2}$	$\widehat{\phi_A}$	$\widehat{\phi_B}$	N^*	I.C. de N	Ampl.
15551	593	0.269	0.142	0.502	0.389	12634	$(14550; 16700)$	2150

Micheletti (2003) determinou estimativas muito próximas dos valores da Tabela 2.4.4 e, em Seber *et al* (2000), encontramos $\widehat{N} = 15540$ e $(13601; 17666)$ como um intervalo de confiança de 95% de confiança para N. Notamos que o intervalo de confiança aproximado de 95% de confiança para N da Tabela 2.4.4 é mais preciso do que este último intervalo.

No próximo capítulo apresentamos uma abordagem Bayesiana para o problema, onde atribuímos distribuições *a priori* não informativas e informativas para o tamanho populacional.

Capítulo 3

Estimação Bayesiana do Tamanho Populacional: Duas Listas com Dados Particionados

Neste capítulo apresentamos uma solução Bayesiana para o problema de estimação do tamanho de uma população fechada. Vimos que a EMV de n_{12}, $\widehat{n_{12}}$, pode não existir implicando que a EMV de N, $\widehat{N}$, também não existe e quando $\widehat{n_{12}}$ é aproximadamente igual a zero, em geral, $\widehat{N}$ superestima o valor de N. Neste contexto, o modelo bayesiano é uma alternativa para resolver este problema, uma vez que informações prévias (*a priori*) do pesquisador, especialista ou de estudos passados podem ser incorporadas ao modelo, melhorando assim as estimativas. Analogamente ao capítulo 1, vamos primeiramente estimar n_{12} e, em seguida, estimar o tamanho populacional.

Na seção 3.1 definimos o modelo bayesiano; na seção 3.2 consideramos distribuições *a priori* não informativas para os parâmetros do modelo; na seção 3.3 introduzimos a distribuição *a priori* de máxima entropia para N e na seção 3.4 apresentamos exemplos com dados simulados e reais.

3.1 Modelo Bayesiano

Supomos *a priori* ϕ_X com distribuição beta de parâmetros α_X e β_X conhecidos e ϕ_A e ϕ_B independentes, $X = \{A, B\}$; n_{12} com função de probabilidades $\pi(n_{12})$ com suporte

$\{0, 1, 2, ..., m\}$, onde $m = \min\{n_1, n_2\}$, e n_{12} e $\boldsymbol{\phi} = (\phi_A, \phi_B)$ independentes. Então, a distribuição *a priori* conjunta de n_{12} e $\boldsymbol{\phi}$ é dada por

$$\begin{aligned}\pi(n_{12}, \boldsymbol{\phi}) &= \pi(n_{12})\,\pi(\phi_A)\,\pi(\phi_B) \\ &= \pi(n_{12}) \prod_{X\in\{A,B\}} \frac{\Gamma(\alpha_X + \beta_X)}{\Gamma(\alpha_X)\,\Gamma(\beta_X)} \phi_X^{\alpha_X - 1} (1 - \phi_X)^{\beta_X - 1} \\ &\propto \pi(n_{12}) \prod_{X\in\{A,B\}} \phi_X^{\alpha_X - 1} (1 - \phi_X)^{\beta_X - 1},\end{aligned} \tag{3.1}$$

$n_{12} = 0, 1, 2, ..., m$ e $0 < \phi_X < 1$, $X = \{A, B\}$. Logo, de (2.6) e (3.1) temos que a distribuição *a posteriori* conjunta de n_{12} e $\boldsymbol{\phi}$ é tal que

$$\begin{aligned}\pi(n_{12}, \boldsymbol{\phi}\,|D_2) &\propto L_2(n_{12}, \boldsymbol{\phi} \mid D_2)\,\pi(n_{12}, \boldsymbol{\phi}) \\ &\propto \pi(n_{12}) \frac{n_{12}!}{(n_{12} - n_T)!} \prod_{X\in\{A,B\}} \phi_X^{m_X + \alpha_X - 1} (1 - \phi_X)^{n_{12} - m_X + \beta_X - 1},\end{aligned} \tag{3.2}$$

$n_T \leq n_{12} \leq m$ e $0 < \phi_X < 1$, $X = \{A, B\}$.

Por outro lado, supomos *a priori* θ_i com distribuição beta de parâmetros φ_i e δ_i conhecidos e θ_1 e θ_2 independentes, $i = 1, 2$; N com distribuição $\pi(N)$, $N = 1, 2, ...$, e N e $\boldsymbol{\theta} = (\theta_1, \theta_2)$ independentes. Então, a distribuição *a priori* conjunta de N e $\boldsymbol{\theta}$ é dada por

$$\begin{aligned}\pi(N, \boldsymbol{\theta}) &= \pi(N)\,\pi(\theta_1)\,\pi(\theta_2) \\ &= \pi(N) \prod_{i=1}^{2} \frac{\Gamma(\varphi_i + \delta_i)}{\Gamma(\varphi_i)\,\Gamma(\delta_i)} \theta_i^{\varphi_i - 1} (1 - \theta_i)^{\delta_i - 1} \\ &\propto \pi(N) \prod_{i=1}^{2} \theta_i^{\varphi_i - 1} (1 - \theta_i)^{\delta_i - 1},\end{aligned} \tag{3.3}$$

$N = 1, 2, ...$ e $0 < \theta_i < 1$, $i = 1, 2$. Logo, de (2.4) e (3.3) temos que a distribuição *a posteriori* conjunta de N e $\boldsymbol{\theta}$ é tal que

$$\begin{aligned}\pi\left(N,\boldsymbol{\theta}\mid D_1\right) &\propto L_1\left(N,\boldsymbol{\theta}\mid D_1\right)\pi\left(N,\boldsymbol{\theta}\right)\\ &\propto \pi\left(N\right)\frac{N!}{(N-n)!}\prod_{i=1}^{2}\theta_i^{n_i+\varphi_i-1}(1-\theta_i)^{N-n_i+\delta_i-1},\end{aligned} \tag{3.4}$$

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

Adotando uma estratégia semelhante a do capítulo 2, determinamos, a partir de (3.2), a média *a posteriori* de n_{12}, $E\left(n_{12} \mid D_2\right)$, substituímos n_{12} por $\left[E\left(n_{12} \mid D_2\right)\right]$ (maior número inteiro não superior a $E\left(n_{12} \mid D_2\right)$) em (3.4), assim, temos que a distribuição *a posteriori* conjunta de N e $\boldsymbol{\theta}$ é dada por

$$\begin{aligned}&\pi\left(N,\boldsymbol{\theta}\mid n_1,n_2,\left[E\left(n_{12}\mid D_2\right)\right]\right)\\ &\propto \pi\left(N\right)\frac{N!}{(N-n_1-n_2+\left[E\left(n_{12}\mid D_2\right)\right])!}\prod_{i=1}^{2}\theta_i^{n_i+\varphi_i-1}(1-\theta_i)^{N-n_i+\delta_i-1},\end{aligned} \tag{3.5}$$

$N \geq n_1 + n_2 - \left[E\left(n_{12} \mid D_2\right)\right]$, $0 < \theta_i < 1$, $i = 1, 2$, a partir da qual determinamos as estimativas bayesianas de N e θ_i, $i = 1, 2$.

Na próxima seção consideramos o modelo bayesiano sob o contexto das distribuições *a priori* não informativa.

3.2 Distribuições *a priori* Não Informativas para os Parâmetros do Modelo

Nesta seção atribuímos distribuições *a priori* não informativas aos parâmetros n_{12} e N e determinamos as distribuições condicionais dos parâmetros do modelo. As distribuições condicionais são necessárias para a implementação dos algoritmos amostrador de Gibbs e Metropolis Hastings, que possibilitarão a determinação dos resumos aproximados (média, moda, mediana e quartis) das distribuições *a posteriori* dos parâmetros. Suponhamos *a priori* que n_{12} tem função de probabilidades uniforme no conjunto $\{0, 1, 2, ..., m\}$, isto é, $\pi\left(n_{12}\right) = \dfrac{1}{m+1}$, $n_{12} = 0, 1, ..., m$ e N tem distribuições $\pi\left(N\right) = \dfrac{1}{N^r}$, $N = 1, 2, ...$, $r = 0, 1$. Para $r = 0$, $\pi\left(N\right) = 1$, $N = 1, 2, ...$, isto é, N tem distribuição imprópria

uniforme nos inteiros estritamente positivos e para $r = 1$, $\pi(N) = \frac{1}{N}$, $N = 1, 2, ...$, ou seja, N tem distribuição de Jeffreys. Denotando por $\widehat{D_1} = (n_1, n_2, [E(n_{12} \mid D_2)])$ e por $n^* = n_1 + n_2 - [E(n_{12} \mid D_2)]$, temos de (3.2) e (3.5), que as distribuições *a posteriori* conjunta de n_{12} e $\boldsymbol{\phi}$ e de N e $\boldsymbol{\theta}$ são tais que

$$\pi(n_{12}, \boldsymbol{\phi} \mid D_2) \propto \frac{n_{12}!}{(n_{12} - n_T)!} \prod_{X \in \{A,B\}} \phi_X^{m_X + \alpha_X - 1} (1 - \phi_X)^{n_{12} - m_X + \beta_X - 1}, \tag{3.6}$$

$n_T \leq n_{12} \leq m$ e $0 < \phi_X < 1$, $X = \{A, B\}$, e

$$\pi\left(N, \boldsymbol{\theta} \mid \widehat{D_1}\right) \propto \frac{N!}{N^r (N - n^*)!} \prod_{i=1}^{2} \theta_i^{n_i + \varphi_i - 1} (1 - \theta_i)^{N - n_i + \delta_i - 1}, \tag{3.7}$$

$N \geq n^*$, $0 < \theta_i < 1$, $i = 1, 2$, $r = 0, 1$. Uma vez que $\pi(N) = \frac{1}{N^r}$, $r = 0, 1$, são distribuições impróprias a questão que se coloca é se as distribuições *a posteriori* de N e $\boldsymbol{\theta}$, dadas em (3.7), existem. O próximo teorema especifica sob qual condição elas existem.

Teorema 3.2.1. A distribuição $\pi\left(N, \boldsymbol{\theta} \mid \widehat{D_1}\right)$ existe

1) para $r = 1$;

2) para $r = 0$ se $[E(n_{12} \mid D_2)] + \varphi_1 + \varphi_2 > 1$.

Prova: Seja C^{-1} a constante normalizadora de (3.7). Então,

$$C = \sum_{N=n^*}^{\infty} \int_0^1 \int_0^1 \frac{N!}{N^r (N - n^*)!} \prod_{i=1}^{2} \theta_i^{n_i + \varphi_i - 1} (1 - \theta_i)^{N - n_i + \delta_i - 1} d\theta_1 d\theta_2$$

$$= \sum_{N=n^*}^{\infty} \frac{N!}{N^r (N - n^*)!} \frac{\Gamma(n_1 + \varphi_1)\Gamma(N - n_1 + \delta_1)}{\Gamma(N + \varphi_1 + \delta_1)} \frac{\Gamma(n_2 + \varphi_2)\Gamma(N - n_2 + \delta_2)}{\Gamma(N + \varphi_2 + \delta_2)}$$

$$\propto A_{n^*},$$

onde

$$A_{n^*} = \sum_{N=n^*}^{\infty} \frac{N!}{N^r (N - n^*)!} \frac{\Gamma(N - n_1 + \delta_1)}{\Gamma(N + \varphi_1 + \delta_1)} \frac{\Gamma(N - n_2 + \delta_2)}{\Gamma(N + \varphi_2 + \delta_2)}. \tag{3.8}$$

Vamos mostrar que (3.8) é convergente para $r = 1$ e para $r = 0$ se $[E(n_{12} \mid D_2)] +$

$\varphi_1 + \varphi_2 > 1$. Inicialmente provaremos que

$$(i)\ \frac{N!}{(N-n^*)!} = O\left(N^{n^*}\right)(N \to \infty),$$

$$(ii)\ \frac{\Gamma(N - n_i + \delta_i)}{\Gamma(N + \varphi_i + \delta_i)} \leq O\left(N^{-(n_i+\varphi_i)}\right)(N \to \infty), i = 1, 2.$$

A relação

$$\frac{N!}{(N-n^*)!} = N(N-1)...(N-n^*+1) < N^{n^*}$$

$$\Rightarrow \frac{N!}{(N-n^*)!N^{n^*}} < 1, N > n^*,$$

o que implica (i); por outro lado, da relação

$$\sqrt{2\pi}x^{x-\frac{1}{2}}\exp\{-x\} \leq \Gamma(x) \leq \sqrt{2\pi}x^{x-\frac{1}{2}}\exp\left\{-x + \frac{1}{12x}\right\},$$

para todo x real, temos

$$\Gamma(N - n_i + \delta_i) \leq \sqrt{2\pi}(N - n_i + \delta_i)^{N-n_i+\delta_i-\frac{1}{2}}\exp\left\{-(N - n_i + \delta_i) + \frac{1}{12(N - n_i + \delta_i)}\right\}$$

e

$$\Gamma(N + \varphi_i + \delta_i)^{-1} \leq \left[\sqrt{2\pi}(N + \varphi_i + \delta_i)^{N+\varphi_i+\delta_i-\frac{1}{2}}\exp\{-(N + \varphi_i + \delta_i)\}\right]^{-1},$$

o que implica

$$\frac{\Gamma(N - n_i + \delta_i)}{\Gamma(N + \varphi_i + \delta_i)} \leq \frac{(N - n_i + \delta_i)^{N-n_i+\delta_i-\frac{1}{2}}\exp\left\{-(N - n_i + \delta_i) + \frac{1}{12(N-n_i+\delta_i)}\right\}}{(N + \varphi_i + \delta_i)^{N+\varphi_i+\delta_i-\frac{1}{2}}\exp\{-(N + \varphi_i + \delta_i)\}}$$

$$= (N - n_i + \delta_i)^{N-n_i+\delta_i-\frac{1}{2}}(N + \varphi_i + \delta_i)^{-\left(N+\varphi_i+\delta_i-\frac{1}{2}\right)}\exp\left\{n_i + \varphi_i + \frac{1}{12(N-n_i+\delta_i)}\right\}. \tag{3.9}$$

Uma vez que

$$(N - n_i + \delta_i)^{N-n_i+\delta_i-\frac{1}{2}} = O\left(N^{N-n_i+\delta_i-\frac{1}{2}}\right)(N \to \infty),$$

$$(N + \varphi_i + \delta_i)^{-\left(N+\varphi_i+\delta_i-\frac{1}{2}\right)} = O\left(N^{-\left(N+\varphi_i+\delta_i-\frac{1}{2}\right)}\right)(N \to \infty),$$

e

$$\exp\left\{n_i + \varphi_i + \frac{1}{12\,(N - n_i + \delta_i)}\right\} = O(1)\,(N \to \infty),$$

segue, de (3.9), que

$$\frac{\Gamma(N - n_i + \delta_i)}{\Gamma(N + \varphi_i + \delta_i)} \leq O\left(N^{N-n_i+\delta_i-\frac{1}{2}}\right) O\left(N^{-\left(N+\varphi_i+\delta_i-\frac{1}{2}\right)}\right) O(1)$$

$$= O\left(N^{-(n_i+\varphi_i)}\right)(N \to \infty), i = 1, 2,$$

o que implica (ii).

Logo, de (3.8), temos

$$A_{n^*} \leq \sum_{N=n^*}^{\infty} O\left(N^{-r}\right) O\left(N^{n^*}\right) \prod_{i=1}^{2} O\left(N^{-(n_i+\varphi_i)}\right)$$

$$= \sum_{N=n^*}^{\infty} O\left(N^{-(n_1+n_2-n^*+\varphi_1+\varphi_2+r)}\right)$$

$$= \sum_{N=n^*}^{\infty} O\left(N^{-([E(n_{12}|D_2)]+\varphi_1+\varphi_2+r)}\right)(N \to \infty),$$

isto é, $A_{n^*} \leq \sum_{N=n^*}^{\infty} a_N$, onde $a_N = O\left(N^{-([E(n_{12}|D_2)]+\varphi_1+\varphi_2+r)}\right)(N \to \infty)$. Então, existe uma constante $M > 0$ e um número inteiro positivo n_0, $n_0 > n^*$, tal que $a_N \leq M.N^{-([E(n_{12}|D_2)]+\varphi_1+\varphi_2+r)}$, para todo $N \geq n_0$, e

$$A_{n^*} \leq \sum_{N=n^*}^{n_0} a_N + \sum_{N=n_0+1}^{\infty} a_N \leq \sum_{N=n^*}^{n_0} a_N + \sum_{N=n_0+1}^{\infty} M.N^{-([E(n_{12}|D_2)]+\varphi_1+\varphi_2+r)}$$
$$= \sum_{N=n^*}^{n_0} a_N + M \sum_{N=n_0+1}^{\infty} N^{-([E(n_{12}|D_2)]+\varphi_1+\varphi_2+r)}. \tag{3.10}$$

Para $r = 1$ segue que $[E(n_{12} \mid D_2)]+\varphi_1+\varphi_2+r > 1$, o que implica $\sum_{N=n_0+1}^{\infty} N^{-([E(n_{12}|D_2)]+\varphi_1+\varphi_2+r)} < \infty$ e, de (3.10), temos que $A_{n^*} < \infty$, o que prova (1).

Para $r = 0$ se $[E(n_{12} \mid D_2)]+\varphi_1+\varphi_2 > 1$, segue que $\sum_{N=n_0+1}^{\infty} N^{-([E(n_{12}|D_2)]+\varphi_1+\varphi_2+r)} < \infty$ e, de (3.10), temos $A_{n^*} < \infty$, o que prova (2). ■

Apresentamos na seqüência as distribuições condicionais dos parâmetros. Segue, de (3.6), que a função de probabilidades condicional de n_{12}, dados $\boldsymbol{\phi}$ e D_2, é tal que

$$\pi(n_{12} \mid \boldsymbol{\phi}, D_2) \propto \binom{n_{12}}{n_T} \left[\prod_{X=A}^{B} (1-\phi_X)\right]^{n_{12}}, \tag{3.11}$$

$n_T \leq n_{12} \leq m$, e a distribuição condicional de $\boldsymbol{\phi}$ dados n_{12} e D_2, é tal que

$$\pi(\boldsymbol{\phi} \mid n_{12}, D_2) \propto \prod_{X\in\{A,B\}} \phi_X^{m_X+\alpha_X-1} (1-\phi_X)^{n_{12}-m_X+\beta_X-1}, \tag{3.12}$$

$0 < \phi_X < 1$, $X = \{A, B\}$. Segue imediatamente de (3.12) que ϕ_A e ϕ_B são condicionalmente independentes, dados n_{12} e D_2, com distribuições beta de parâmetros $m_A + \alpha_A$, $n_{12} - m_A + \beta_A$ e $m_B + \alpha_B$, $n_{12} - m_B + \beta_B$, respectivamente. Para $r = 0$, segue de (3.7), que a função de probabilidades condicional de N, dados $\boldsymbol{\theta}$ e $\widehat{D_1}$, é tal que

$$\pi\left(N \mid \boldsymbol{\theta}, \widehat{D_1}\right) \propto \binom{N}{n^*} \left[\prod_{i=1}^{2}(1-\theta_i)\right]^{N}, \tag{3.13}$$

$N \geq n^*$. O inverso da constante normalizadora de (3.13) é dado por

$$\sum_{N\geq n^*} \binom{N}{n^*} \left[\prod_{i=1}^{2}(1-\theta_i)\right]^N$$

$$= \sum_{S\geq 0} \binom{S+n^*}{n^*} \left[\prod_{i=1}^{2}(1-\theta_i)\right]^{S+n^*}$$

$$= \left[\prod_{i=1}^{2}(1-\theta_i)\right]^{n^*} \sum_{S\geq 0} \binom{S+n^*}{n^*} \left[\prod_{i=1}^{2}(1-\theta_i)\right]^{S}$$

$$= \left[\prod_{i=1}^{2}(1-\theta_i)\right]^{n^*} \left[1-\prod_{i=1}^{2}(1-\theta_i)\right]^{-(n^*+1)},$$

o que implica

$$\pi\left(N \mid \boldsymbol{\theta}, \widehat{D_1}\right) = \binom{N}{n^*} \left[1-\prod_{i=1}^{2}(1-\theta_i)\right]^{n^*+1} \left[\prod_{i=1}^{2}(1-\theta_i)\right]^{N-n^*}, \tag{3.14}$$

$N \geq n^*$. Notamos que (3.14) é a função de probabilidades de uma variável aleatória $Y + n^*$, onde Y tem função de probabilidades binomial negativa com parâmetros $n^* + 1$ e $1 - \prod_{i=1}^{2}(1-\theta_i)$. De fato, se Y tiver função de probabilidades binomial negativa com parâmetros $n^* + 1$ e $1 - \prod_{i=1}^{2}(1-\theta_i)$, então

$$P\left(Y + n^* = y \mid \boldsymbol{\theta}, \widehat{D_1}\right) = P\left(Y = y - n^* \mid \boldsymbol{\theta}, \widehat{D_1}\right)$$

$$= \binom{y}{n^*} \left[1-\prod_{i=1}^{2}(1-\theta_i)\right]^{n^*+1} \left[\prod_{i=1}^{2}(1-\theta_i)\right]^{y-n^*},$$

$y = n^*, n^* + 1, \ldots$. Para $r = 1$ segue, de (3.7), que a função de probabilidades condicional de N, dados $\boldsymbol{\theta}$ e $\widehat{D_1}$, é tal que

$$\pi\left(N \mid \boldsymbol{\theta}, \widehat{D_1}\right) \propto \binom{N-1}{n^*-1} \left[\prod_{i=1}^{2}(1-\theta_i)\right]^N, \tag{3.15}$$

$N \geq n^*$. Analogamente ao caso $r = 0$, notamos que (3.15) é igual a função de probabilidades de uma variável aleatória $Z + n^*$, onde Z tem função de probabilidades binomial negativa com parâmetros n^* e $1 - \prod_{i=1}^{2}(1-\theta_i)$. Finalmente segue, de (3.7), que a dis-

tribuição condicional de $\boldsymbol{\theta}$ dados N e $\widehat{D_1}$, é tal que

$$\pi\left(\boldsymbol{\theta} \mid N, \widehat{D_1}\right) \propto \prod_{i=1}^{2} \theta_i^{n_i+\varphi_i-1}(1-\theta_i)^{N-n_i+\delta_i-1}, \tag{3.16}$$

$0 < \theta_i < 1$, $i = 1, 2$, isto é, θ_1 e θ_2 são condicionalmente independentes, dados N e $\widehat{D_1}$, com distribuições beta com parâmetros $n_1 + \varphi_1$, $N - n_1 + \delta_1$ e $n_2 + \varphi_2$, $N - n_2 + \delta_2$, respectivamente.

Na próxima seção apresentamos a distribuição *a priori* informativa de máxima entropia.

3.3 Distribuição *a priori* de Máxima Entropia para o Tamanho Populacional

Nesta seção consideramos ainda uma distribuição *a priori* uniforme para n_{12} e vamos atribuir como *priori* para N, uma distribuição que seja de máxima entropia e que leve em conta a informação fornecida pelo pesquisador ou especialista sobre N. Mais especificamente, vamos atribuir a N uma distribuição *a priori* que expresse a maior variabilidade possível (menor informação possível) e ao mesmo tempo contemple a informação fornecida pelo pesquisador como, por exemplo, a média ou a variância de N. Na seqüência definimos uma distribuição de probabilidades de máxima entropia, determinamos a distribuição *a priori* de máxima entropia para N, dados seus momentos até uma certa ordem e, finalmente, determinamos as distribuições *a posteriori* e condicionais dos parâmetros do modelo.

Suponhamos $f(x)$ uma função de probabilidades com suporte $\{x_1, x_2, ...\}$ finito ou infinito. A entropia de $f(x)$, denotada por $\varepsilon(f)$, é definida por

$$\varepsilon(f) = -\sum_{i\geq 1} f(x_i) \ln f(x_i).$$

A idéia da entropia é quanto maior o valor de $\varepsilon(f)$ maior é a variabilidade (caos) do experimento subjacente à função de probabilidades $f(x)$ ou menos informação temos sobre $f(x)$. A distribuição de probabilidades *a priori* de máxima entropia para N, dados

seus momentos μ_1, μ_2,..., μ_p é dado pelo seguinte teorema.

Teorema 3.3.1. A função de probabilidades *a priori* de máxima entropia para N, $\pi(N)$, $N = 1, 2, ...$, cujos momentos até a ordem p são iguais a μ_1, μ_2,..., μ_p, respectivamente, é dada por

$$\pi(N) = \frac{\exp\left\{-\sum_{k=1}^{p} \lambda_k N^k\right\}}{\sum_{N\geq 0} \exp\left\{-\sum_{k=1}^{p} \lambda_k N^k\right\}}, \qquad N \geq 0,$$

onde λ_k são constantes positivas a serem determinadas em função de μ_k.

Prova: Seja $f(i)$, $i = 0, 1, ...$, uma função de probabilidades de N tal que

$$(1) \qquad \mu_k = E\left(N^k\right) = \sum_{i\geq 1} i^k f(i), \qquad 1 \leq k \leq p.$$

Nosso objetivo é determinar a função de probabilidades $\widetilde{f}(i)$, $i \geq 0$, que satisfaça a condição (1) e maximize $\varepsilon(f) = -\sum_{i\geq 1} f(x_i) \ln f(x_i)$. Utilizando a técnica de multiplicadores de Lagrange, devemos maximizar a função

$$h(f) = \varepsilon(f) + \sum_{k=1}^{p} \lambda_k \left[\mu_k - E\left(N^k\right)\right]$$

$$= -\sum_{i\geq 0} f(i) \ln f(i) + \sum_{k=1}^{p} \lambda_k \left[\mu_k - \sum_{i\geq 1} i^k f(i)\right]$$

$\lambda_k > 0$, $1 \leq k \leq p$. Então, $\widetilde{f}(i)$ satisfaz o sitema de equações

$$\begin{cases} \dfrac{\partial h(f)}{\partial f(i)} = -1 - \ln f(i) - \sum_{k=1}^{p} \lambda_k i^k = 0, \; i \geq 0; \\ \\ \dfrac{\partial h(f)}{\partial \lambda_k} = \mu_k - \sum_{i\geq 1} i^k f(i) = 0, \; 1 \leq k \leq p \end{cases}$$

$$\Leftrightarrow \begin{cases} \ln f(i) = -\left(1 + \sum_{k=1}^{p} \lambda_k i^k\right), \; i \geq 0, \\ \\ \mu_k = \sum_{i\geq 1} i^k f(i), \; 1 \leq k \leq p. \end{cases}$$

Logo, $\ln \widetilde{f}(i) = -\left(1 + \sum_{k=1}^{p} \lambda_k i^k\right)$, $i \geq 0$, ou $\widetilde{f}(i) = \exp\left\{-\left(1 + \sum_{k=1}^{p} \lambda_k i^k\right)\right\} \propto$

$\exp\left\{-\sum_{k=1}^{p}\lambda_k i^k\right\}$, $i \geq 0$, e $\mu_k = \sum_{i\geq 1} i^k \widetilde{f}(i)$, $1 \leq k \leq p$.

Como $\sum_{i\geq 0}\widetilde{f}(i) = 1$, segue que

$$\widetilde{f}(i) = \frac{\exp\left\{-\sum_{k=1}^{p}\lambda_k i^k\right\}}{\sum_{i\geq 0}\exp\left\{-\sum_{k=1}^{p}\lambda_k i^k\right\}},$$

$i \geq 0$, o que prova o teorema. ■

Suponhamos que o pesquisador possua a informação (subjetiva) somente sobre a média *a priori*, μ, de N. Pelo Teorema 3.3.1 temos $k = 1$ e a distribuição *a priori* para N de máxima entropia, cuja média é igual a μ, é dada por

$$\pi(N) = \frac{\exp\{-\lambda_1 N\}}{\sum_{N\geq 0}\exp\{-\lambda_1 N\}} = \frac{e^{-\lambda_1 N}}{(1-e^{-\lambda_1})^{-1}} = \left(1-e^{-\lambda_1}\right)\left(e^{-\lambda_1}\right)^N,$$

$N = 0, 1, ...$, isto é, $\pi(N)$ é a função de probabilidades geométrica de parâmetro $1 - e^{-\lambda_1}$, $\lambda_1 > 0$. Como $E(N) = \dfrac{e^{-\lambda_1}}{1-e^{-\lambda_1}} = \mu$, temos $e^{-\lambda_1} = \dfrac{\mu}{1+\mu}$ ou

$$\pi(N) = \left(\frac{1}{1+\mu}\right)\left(\frac{\mu}{1+\mu}\right)^N, \tag{3.17}$$

$N = 0, 1, ...$, ou seja, a distribuição *a priori* de máxima entropia para N de média μ é a distribuição geométrica de parâmetro $\dfrac{1}{1+\mu}$.

Apresentamos na seqüência as distribuições condicionais dos parâmetros. De (3.6) temos que a função de probabilidades condicional de n_{12}, dados $\boldsymbol{\phi}$ e D_2, é tal que

$$\pi(n_{12}|\,\boldsymbol{\phi}, D_2) \propto \binom{n_{12}}{n_T}\left[\prod_{X\in\{A,B\}}(1-\phi_X)\right]^{n_{12}}, \tag{3.18}$$

$n_T \leq n_{12} \leq m$, e que ϕ_A e ϕ_B são condicionalmente independentes, dados n_{12} e D_2, com distribuições beta de parâmetros $m_A + \alpha_A$, $n_{12} - m_A + \beta_A$ e $m_B + \alpha_B$, $n_{12} - m_B + \beta_B$, respectivamente.

Para N, segue de (3.5) e (3.17) que a que a função de probabilidades condicional de N, dados $\boldsymbol{\theta}$ e $\widehat{D_1}$, é tal que

$$\pi\left(N \mid \boldsymbol{\theta}, \widehat{D_1}\right) \propto \binom{N}{n^*}\left[\left(\frac{\mu}{1+\mu}\right)\prod_{i=1}^{2}(1-\theta_i)\right]^N, \tag{3.19}$$

$N \geq n^*$. O inverso da constante normalizadora de (3.19) é

$$\sum_{N\geq n^*}\binom{N}{n^*}\left[\left(\frac{\mu}{1+\mu}\right)\prod_{i=1}^{2}(1-\theta_i)\right]^N$$

$$=\sum_{S\geq 0}\binom{S+n^*}{n^*}\left[\left(\frac{\mu}{1+\mu}\right)\prod_{i=1}^{2}(1-\theta_i)\right]^{S+n^*}$$

$$=\left[\left(\frac{\mu}{1+\mu}\right)\prod_{i=1}^{2}(1-\theta_i)\right]^{n^*}\sum_{S\geq 0}\binom{S+n^*}{n^*}\left[\left(\frac{\mu}{1+\mu}\right)\prod_{i=1}^{2}(1-\theta_i)\right]^{S}$$

$$=\left[\left(\frac{\mu}{1+\mu}\right)\prod_{i=1}^{2}(1-\theta_i)\right]^{n^*}\left[1-\left(\frac{\mu}{1+\mu}\right)\prod_{i=1}^{2}(1-\theta_i)\right]^{-(n^*+1)},$$

o que implica

$$\pi\left(N \mid \boldsymbol{\theta}, \widehat{D_1}\right) = \binom{N}{n^*}\left[1-\left(\frac{\mu}{1+\mu}\right)\prod_{i=1}^{2}(1-\theta_i)\right]^{n^*+1}\left[\left(\frac{\mu}{1+\mu}\right)\prod_{i=1}^{2}(1-\theta_i)\right]^{N-n^*}, \tag{3.20}$$

$N \geq n^*$. Notamos, como em (3.14), que (3.20) é a função de probabilidades de uma variável aleatória $W + n^*$, onde W tem função de probabilidades binomial negativa com parâmetros $n^* + 1$ e $1 - \left(\frac{\mu}{1+\mu}\right)\prod_{i=1}^{2}(1-\theta_i)$.

Finalmente, de (3.5) segue que que as distribuições condicionais de θ_1 e θ_2, dados N e $\widehat{D_1}$, são beta com parâmetros $n_1+\varphi_1$, $N-n_1+\delta_1$ e $n_2+\varphi_2$, $N-n_2+\delta_2$, respectivamente.

Na próxima seção apresentamos exemplos com dados simulados e reais, onde determinamos resumos aproximados das distribuições *a posteriori*.

3.4 Exemplos com Dados Simulados e Reais

Nesta seção apresentamos exemplos com dados simulados e reais utilizando as distribuições *a priori* vistas neste capítulo. Nos exemplos consideramos duas cadeias cujos

elementos foram simulados das distribuições condicionais dos parâmetros, onde cada uma contém 50000 elementos. Destas cadeias os 10000 primeiros elementos foram descartados como aquecimento e dos restantes foram selecionados os primeiros de cada dez (saltos) para garantir a independência entre os valores. Desse modo obtivemos amostras das distribuições *a posteriori* com 4000 elementos em cada cadeia, ou seja, somando as duas cadeias obtivemos uma amostra de 8000 elementos das distribuições *a posteriori* marginais.

Na próxima seção apresentamos exemplos com dados simulados.

3.4.1 Exemplo com Dados Simulados

Nesta seção utilizamos as mesmas quantidades geradas no exemplo 2.4.1. A Tabela 3.4.1 mostra estes valores.

Tabela 3.4.1 Quantidades geradas da distribuição multinomial

Valores fixados					Valores gerados					
N	θ_1	θ_2	ϕ_A	ϕ_B	n_1	n_2	n_{12}	n_A	n_B	n_{AB}
50	0,7	0,8	0,7	0,9	33	38	25	2	7	15
50	0,8	0,7	0,2	0,1	37	33	22	2	4	0
50	0,1	0,3	0,7	0,7	3	19	2	1	1	0
50	0,1	0,3	0,2	0,1	3	19	2	1	1	0
500	0,6	0,8	0,7	0,8	299	403	235	38	73	115
500	0,9	0,7	0,3	0,2	453	344	307	74	45	20
500	0,1	0,4	0,8	0,6	46	203	18	4	1	11
500	0,2	0,3	0,1	0,2	101	159	30	3	3	1
2000	0,6	0,8	0,7	0,9	1173	1576	905	67	238	576
2000	0,5	0,6	0,4	0,3	947	1168	528	157	95	68
2000	0,2	0,4	0,7	0,8	424	827	156	19	37	93
2000	0,2	0,1	0,3	0,4	411	209	39	9	11	1

O exemplo a seguir apresenta os resumos aproximados das distribuições *a posteriori* dos parâmetros, utilizando a distribuição *a priori* uniforme para o tamanho populacional.

Exemplo 3.4.1 Neste exemplo supomos $\boldsymbol{\phi}$ e $\boldsymbol{\theta}$ com distribuição uniforme no intervalo

$(0,1)$, ou seja, $\alpha_X = \beta_X = 1$, $X = A, B$ e $\varphi_j = \delta_j = 1$, $j = 1,2$ e tomamos $r = 0$. A Tabela 3.4.2 mostra os resumos aproximados das distribuições *a posteriori* de n_{12} e $\boldsymbol{\phi}$, onde consideramos as estatísticas correspondentes a $N = 2000$, $\theta_1 = 0,6$, $\theta_2 = 0,8$, $\phi_A = 0,7$ e $\phi_B = 0,9$ na Tabela 3.4.1.

Tabela 3.4.2 Resumos aproximados das distribuições *a posteriori* de n_{12}, ϕ_A e ϕ_B

	Média	Moda	Q_1	Mediana	Q_3	D.P.	I.C.(95%)	Ampl.	Conv.
n_{12}	910,9	910	906	911	915	6,65	(899; 925)	26	1
ϕ_A	0,70	0,71	0,69	0,71	0,77	0,02	(0,67; 0,74)	0,07	1
ϕ_B	0,89	0,89	0,88	0,89	0,90	0,01	(0,87; 0,92)	0,05	1

Na Tabela 3.4.2 e nas que seguem denotamos por

Q_j, o j-ésimo quartil, $j = 1,3$;

D.P., o desvio padrão;

I.C.(95%), o intervalo de credibilidade com 95% de credibilidade;

Ampl., a amplitude do intervalo de credibilidade;

Conv., a medida de convergência do diagnóstico de Gelman Rubin, isto é, as cadeias estão convergindo aproximadamente para um mesmo valor.

Assumindo a média da distribuição *a posteriori* como estimativa bayesiana paramétrica, notamos da Tabela 3.4.2 que a estimativa de cada parâmetro é aproximadamente igual ao seu valor verdadeiro, bem como o intervalo de credibilidade contém seu verdadeiro valor.

A Figura 3.4.1 mostra os gráficos de autocorrelação e do critério de convergência de

Gelman Rubin para a convergência das cadeias.

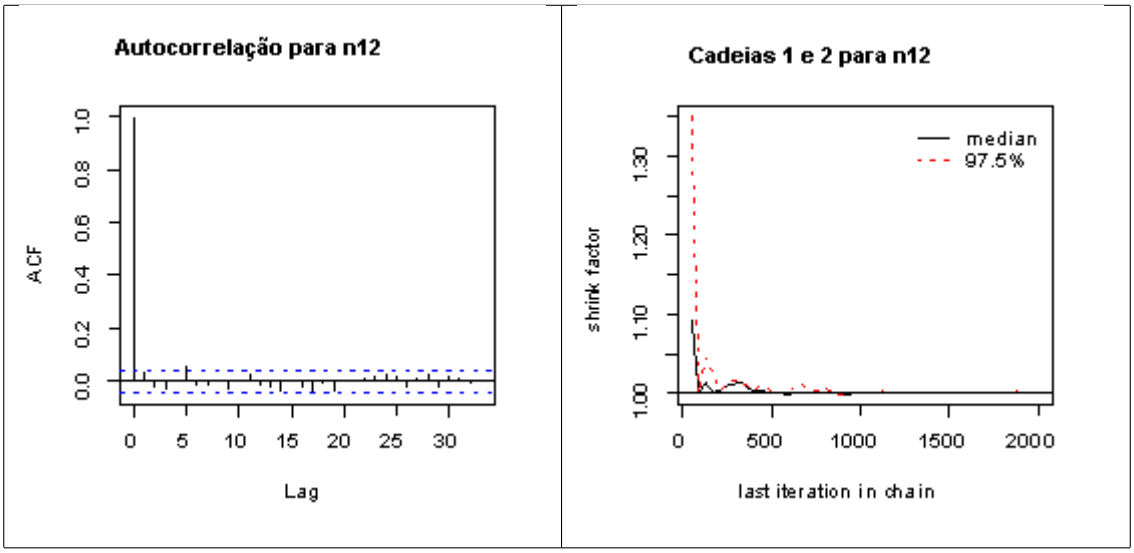


Figura 3.4.1 Gráficos de autocorrelação e do critério de convergência de Gelman Rubin para n_{12}

Na Figura 3.4.1 observamos que utilizando saltos de dez resolvemos o problema de autocorrelação e que ocorre a convergência das cadeias para o parâmetro n_{12}. A mesma análise gráfica para os parâmetros ϕ_A e ϕ_B se comportaram de forma semelhante.

A Figura 3.4.2 mostra o histograma da distribuição *a posteriori* de n_{12} e o comportamento das cadeias da distribuição *a posteriori* de n_{12}.

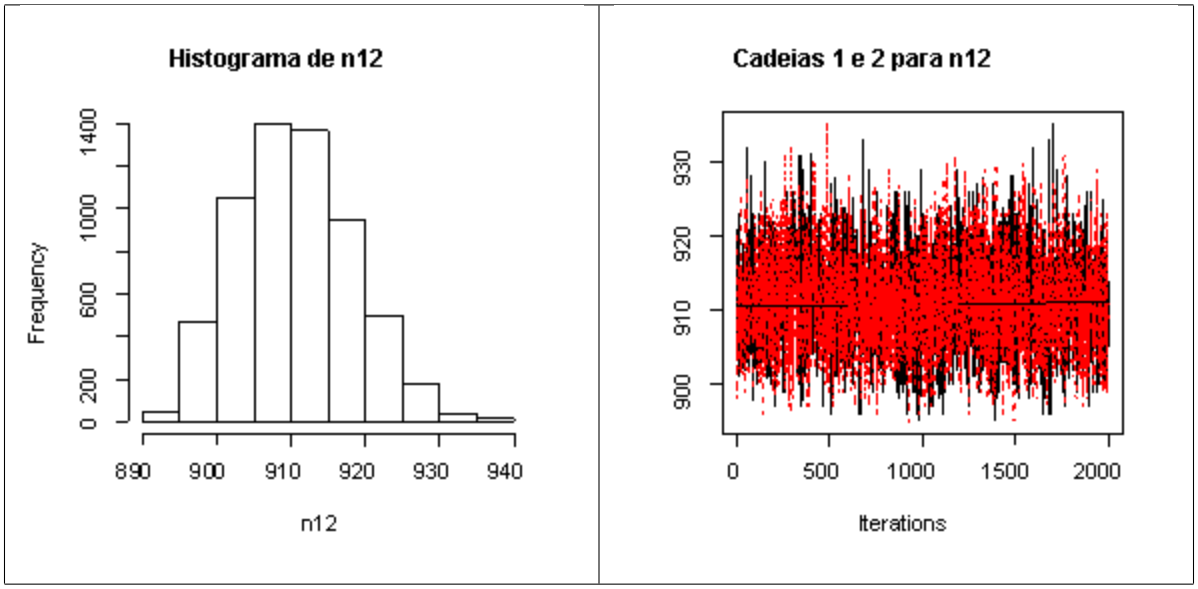


Figura 3.4.2 Histograma e cadeias da distribuição *a posteriori* de n_{12}

A Figura 3.4.3 apresenta as densidades das distribuições *a posteriori* de ϕ_A e ϕ_B.

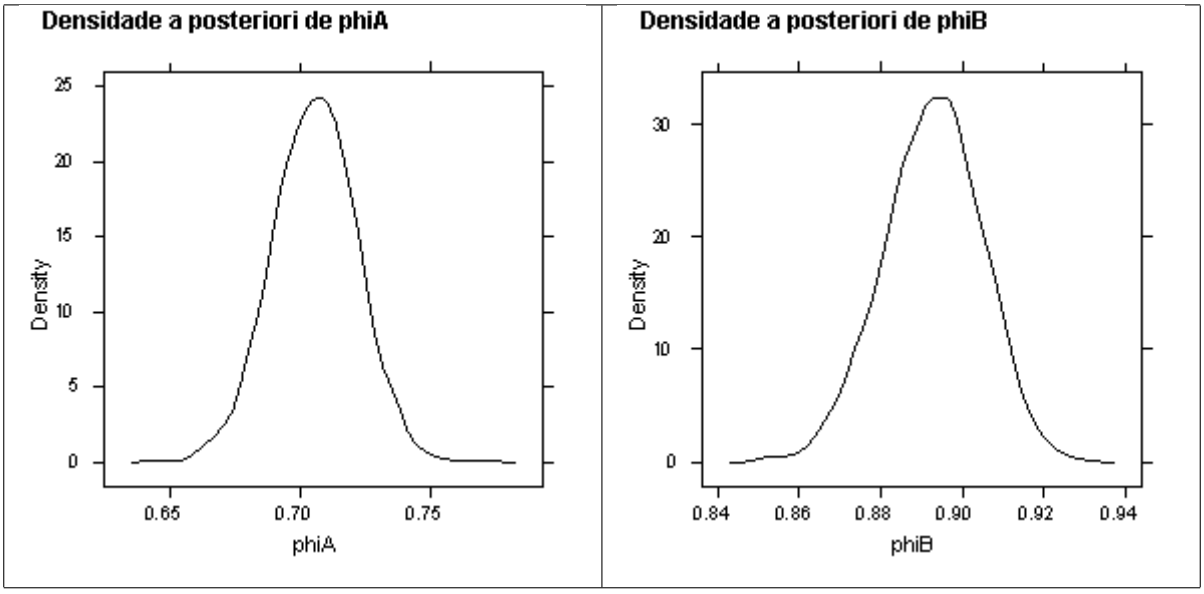


Figura 3.4.3 Densidades das distribuições *a posteriori* de ϕ_A e ϕ_B

Utilizando a média *a posteriori* como estimativa de n_{12}, vamos agora determinar estimativas para N, θ_1 e θ_2. A Tabela 3.4.3 mostra os resultados obtidos.

Tabela 3.4.3 Resumos aproximados das distribuições *a posteriori* de N, θ_1 e θ_2

	Média	Moda	Q_1	Mediana	Q_3	D.P.	I.C.(95%)	Ampl.	Conv.
N	2033	2032	2018	2032	2046	20,70	(1994; 2075)	81	1
θ_1	0,58	0,58	0,57	0,58	0,58	0,01	(0,55; 0,60)	0,05	1
θ_2	0,77	0,78	0,77	0,77	0,78	0,01	(0,75; 0,81)	0,06	1

Notamos na Tabela 3.4.3 que as estimativas bayesianas (médias) de N, θ_1 e θ_2 estão próximas dos seus respectivos verdadeiros valores e seus intervalos de credibilidade contêm estes valores. Observamos também que a estimativa (média) de N da Tabela 3.4.3 está mais próxima do verdadeiro valor quando comparado com o EMV da Tabela 2.4.2 e seu intervalo de credibilidade possui uma amplitude menor do que o intervalo de confiança determinado na Tabela 2.4.3.

A análise de convergência para os parâmetros N, θ_1 e θ_2 é feita de maneira análoga à

de n_{12}, ϕ_A e ϕ_B.

A Figura 3.4.4 mostra o histograma da distribuição *a posteriori* de N e o gráfico do comportamento das cadeias da distribuição *a posteriori* de N.

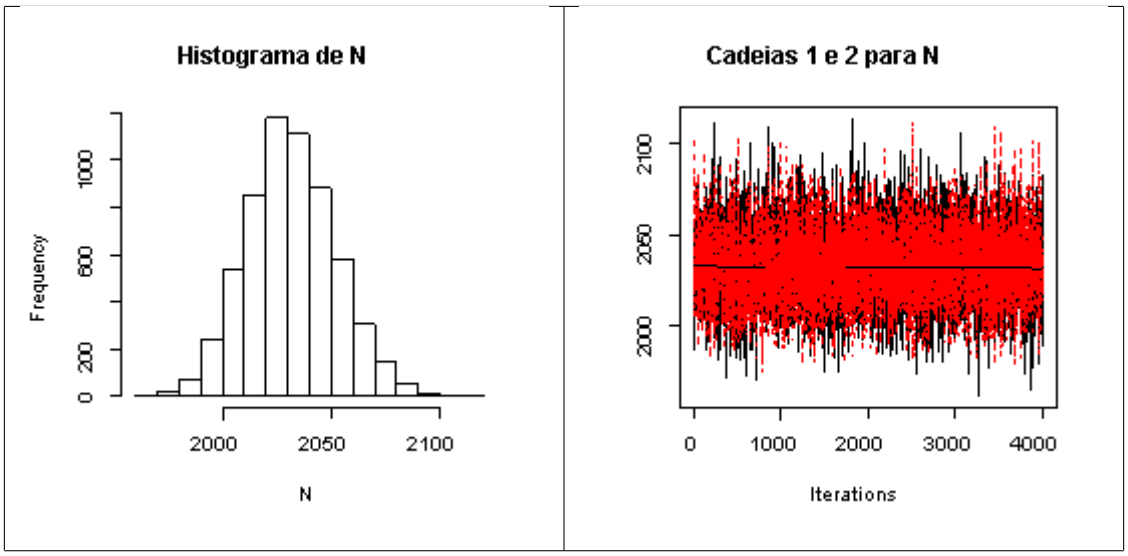


Figura 3.4.4 Histograma da distribuição *a posteriori* de N e cadeias da distribuição *a posteriori* de N

Na Figura 3.4.5 temos as densidades das distribuições *a posteriori* de θ_1 e θ_2.

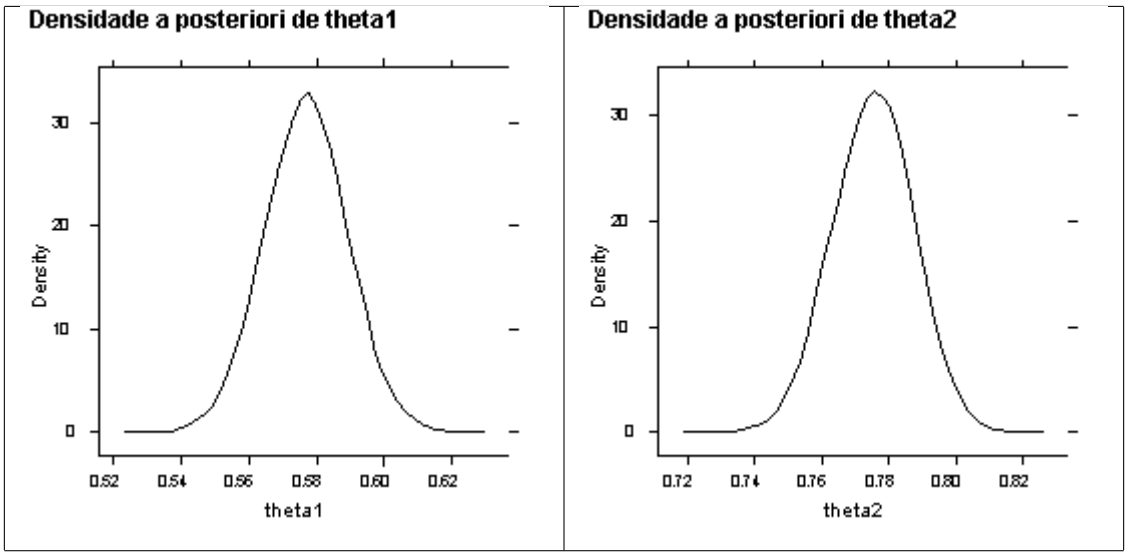


Figura 3.4.5 Densidades *a posteriori* de θ_1 e θ_2

A seguir apresentamos somente os resumos aproximados das distribuições *a posteriori* de n_{12} e N, onde n_{12} é o parâmetro que devemos estimar primeiramente para, em seguida, estimar N, nosso parâmetro de interesse.

Doravante não mostraremos a coluna da tabela que avalia a convergência das cadeias através do diagnóstico de Gelman Rubin, pois, para todos os valores simulados obtivemos convergência.

A Tabela 3.4.4 apresenta os resumos aproximados das distribuições *a posteriori* de n_{12} para estatísticas variando como na Tabela 3.4.1.

Tabela 3.4.4 Resumos aproximados das distribuições *a posteriori* de n_{12}

n_{12}	θ_1	θ_2	ϕ_A	ϕ_B	Média	Moda	Q_1	Mediana	Q_3	D.P.	I.C.(95%)	Ampl.
25	0,7	0,8	0,7	0,9	27,31	26	26	27	28	1,67	(25; 31)	6
22	0,8	0,7	0,2	0,1	18,46	14	13	18	24	6,73	(8; 31)	23
2	0,1	0,3	0,7	0,7	3	3	–	3	–	–	–	–
2	0,1	0,3	0,2	0,1	3	3	–	3	–	–	–	–
235	0,6	0,8	0,7	0,8	238,6	237	236	238	246	7,81	(230; 260)	30
307	0,9	0,7	0,3	0,2	288,7	270	266	290	315	31,91	(225; 339)	114
18	0,1	0,4	0,8	0,6	18,62	18	17	18	19	1,88	(17; 23)	6
30	0,2	0,3	0,1	0,2	24,25	15	13	19	30	15,87	(9; 70)	61
905	0,6	0,8	0,7	0,9	910,9	910	906	911	915	6,65	(899; 925)	26
528	0,5	0,6	0,4	0,3	549,5	529	515	546	579	47,54	(469; 658)	189
156	0,2	0,4	0,7	0,8	159,2	158	156	159	162	4,27	(153; 170)	17
39	0,2	0,1	0,3	0,4	93,39	66	59	85	121	44,73	(36; 194)	158

A Tabela 3.4.5 apresenta os resumos aproximados das distribuições *a posteriori* de N, utilizando as $[E(n_{12} \mid D_2)]$ determinadas na Tabela 3.4.4.

Tabela 3.4.5 Resumos aproximados das distribuições *a posteriori* de N, utilizando a distribuição *a priori* não informativa

N	Média	Moda	Q_1	Mediana	Q_3	D.P.	I.C.(95%)	Ampl.
50	47, 08	46	45	47	48	2, 40	(44; 53)	9
50	68, 83	64	63	68	73	8, 16	(57; 88)	31
50	24, 13	19	20	21	25	5, 70	(19; 39)	20
50	24, 78	19	20	21	25	5, 70	(19; 47)	28
500	507, 3	504	501	507	513	9, 64	(490; 528)	38
500	539, 9	540	534	539	545	7, 80	(526; 557)	31
500	514, 2	454	451	500	561	90, 47	(377; 735)	358
500	624, 9	638	553	614	682, 2	99, 58	(465; 845)	380
2000	2033	2032	2018	2032	2046	20, 70	(1994; 2075)	81
2000	2016	2006	1988	2015	2042	40, 06	(1943; 2099)	156
2000	2208	2206	2115	2205	2286	125, 30	(1985; 2472)	487
2000	925, 1	934	881	920	963	63, 56	(816; 1065)	249

Notamos pela Tabela 3.4.5 que as estimativas bayesianas de N são influenciadas pelas probabilidades $\boldsymbol{\theta}$ e $\boldsymbol{\phi}$. Assim, para valores baixos de $\boldsymbol{\theta}$ ou $\boldsymbol{\phi}$ não obtivemos boas estimativas e em alguns casos o intervalo de credibilidade não contém o verdadeiro valor do parâmetro. O mesmo acontece quando o tamanho populacional é relativamente pequeno.

A seguir apresentamos um exemplo, onde consideramos a distribuição *a priori* não informativa para o tamanho populacional com $r = 1$.

Exemplo 3.4.2 Neste exemplo continuamos assumindo distribuições *a priori* não informativas para os parâmetros, mas agora consideramos $r = 1$, ou seja, a *priori* de Jeffreys para o tamanho populacional. Utilizamos as estimativas bayesianas do parâmetro n_{12}, $[E(n_{12} \mid D_2)]$, como sendo aquelas dadas na Tabela 3.4.4.

A Tabela 3.4.6 apresenta os resumos aproximados das distribuições *a posteriori* de N.

Tabela 3.4.6 Resumos aproximados das distribuições *a posteriori* de N, utilizando a distribuição *a priori* de Jeffreys

N	Média	Moda	Q_1	Mediana	Q_3	D.P.	I.C.(95%)	Ampl.
50	47, 08	46	45	47	48	2, 40	(44; 53)	9
50	68, 76	64	63	67	73	8, 24	(57; 89)	32
50	23, 14	19	20	21	24	5, 66	(19; 39)	20
50	23, 1	19	20	21	24	5, 48	(19; 38)	19
500	506, 7	504	500	506	513	9, 46	(490; 526)	36
500	540	541	535	540	545	7, 63	(527; 556)	29
500	514, 2	454	451	500	561	90, 47	(377; 735)	358
500	611, 4	623	543	598	666, 2	96, 11	(460; 836)	376
2000	2032	2032	2019	2032	2046	20, 39	(1994; 2074)	80
2000	2016	2006	1988	2015	2042	40, 06	(1943; 2099)	156
2000	2208	2207	2115	2205	2286	125, 30	(1984; 2472)	487
2000	923, 8	892	879	919	963	63, 60	(812; 1062)	226

Observamos na Tabela 3.4.6 que as estimativas bayesianas de N são praticamente iguais as da Tabela 3.4.5, ou seja, praticamente as mesmas quando utilizamos a distribuição *a priori* de N para $r = 0$.

No exemplo a seguir consideramos a distribuição *a priori* de máxima entropia para o tamanho populacional.

Exemplo 3.4.3 Neste exemplo utilizamos a distribuição *a priori* de máxima entropia para o tamanho populacional e supomos distribuições *a priori* não informativas para o restante dos parâmetros. Consideramos as estatísticas correspondentes a $N = 2000$, $\theta_1 = 0,6$, $\theta_2 = 0,8$, $\phi_A = 0,7$, $\phi_B = 0,9$ e a $N = 2000$, $\theta_1 = 0,2$, $\theta_2 = 0,1$, $\phi_A = 0,3$, $\phi_B = 0,4$, dadas na Tabela 3.4.1 e supomos que o pesquisador possui a informação (subjetiva) de que a média populacional é igual a μ. As Tabelas 3.4.7 e 3.4.8 apresentam os resumos aproximados das distribuições *a posteriori* de N, para diferentes valores de μ.

Tabela 3.4.7 Resumos aproximados das distribuições *a posteriori* de N, correspondentes a $N = 2000$, $\theta_1 = 0,6$, $\theta_2 = 0,8$, $\phi_A = 0,7$ e $\phi_B = 0,9$

N	μ	Média	Moda	Q_1	Mediana	Q_3	D.P.	I.C.(95%)	Ampl.
2000	10	1998	2002	1986	1997	2009	17,51	(1965; 2033)	68
2000	20	2013	2017	2000	2013	2026	18,80	(1978; 2051)	73
2000	50	2024	2028	2010	2024	2037	19,75	(1987; 2065)	78
2000	100	2029	2037	2015	2028	2042	20,22	(1991; 2070)	79
2000	1000	2032	2031	2018	2031	2045	20,51	(1993; 2074)	81
2000	2000	2032	2030	2018	2031	2046	20,67	(1994; 2075)	81
2000	4000	2033	2030	2019	2032	2046	20,49	(1994; 2074)	80

Tabela 3.4.8 Resumos aproximados das distribuições *a posteriori* de N, correspondentes a $N = 2000$, $\theta_1 = 0,2$, $\theta_2 = 0,1$, $\phi_A = 0,3$ e $\phi_B = 0,4$

N	μ	Média	Moda	Q_1	Mediana	Q_3	D.P.	I.C.(95%)	Ampl.
2000	10	747,6	738	726	745	767	30,52	(692; 811)	119
2000	20	806,7	787	778	804	832	40,29	(735; 892)	157
2000	50	864,8	859	829	860	896	51,87	(774; 977)	203
2000	100	892,8	864	854	890	926	55,25	(795; 1009)	214
2000	1000	924,1	929	879	919	961	64,05	(815; 1068)	253
2000	2000	925,2	916	881	920	966	64,15	(815; 1067)	252
2000	4000	927,7	904	882	924	967	64,21	(818; 1069)	251

Notamos nas tabelas acima que a medida que o valor atribuído a μ aumenta. Os resumos da distribuição *a posteriori* de N são aproximadamente iguais aqueles obtidos utilizando as distribuições *a priori* não informativas vistas nas Tabelas 3.4.5 e 3.4.6, e para valores baixos de μ estas estimativas se apresentaram um pouco menores

Na próxima seção apresentamos um exemplo com dados reais.

3.4.2 Exemplo com dados Reais

Utilizamos os dados do exemplo 2.4.2. Lembremos que $n_1 = 4186$, $n_2 = 2203$, $m_A = 298$, $m_B = 231$ e $n_{AB} = 116$.

Exemplo 3.4.4 Neste exemplo consideramos novamente que n_{12}, N, $\boldsymbol{\theta}$ e $\boldsymbol{\phi}$ possuem distribuições *a priori* não informativas e $r = 0$.

A Tabela 3.4.9 mostra os resumos aproximados das distribuições *a posteriori* de n_{12}, ϕ_A e ϕ_B.

Tabela 3.4.9 Resumos aproximados das distribuições *a posteriori* de n_{12}, ϕ_A e ϕ_B

	Média	Moda	Q_1	Mediana	Q_3	D.P.	I.C.(95%)	Ampl.	Conv.
n_{12}	$599,4$	575	574	597	622	$33,68$	$(541; 673)$	132	1
ϕ_A	$0,50$	$0,52$	$0,47$	$0,50$	$0,52$	$0,03$	$(0,43; 0,57)$	$0,14$	1
ϕ_B	$0,39$	$0,38$	$0,37$	$0,39$	$0,41$	$0,03$	$(0,33; 0,44)$	$0,11$	1

A Figura 3.4.6 mostra o histograma da distribuição *a posteriori* de n_{12} e o gráfico do critério de convergência de Gelman Rubin para as cadeias, com relação ao parâmetro n_{12}.

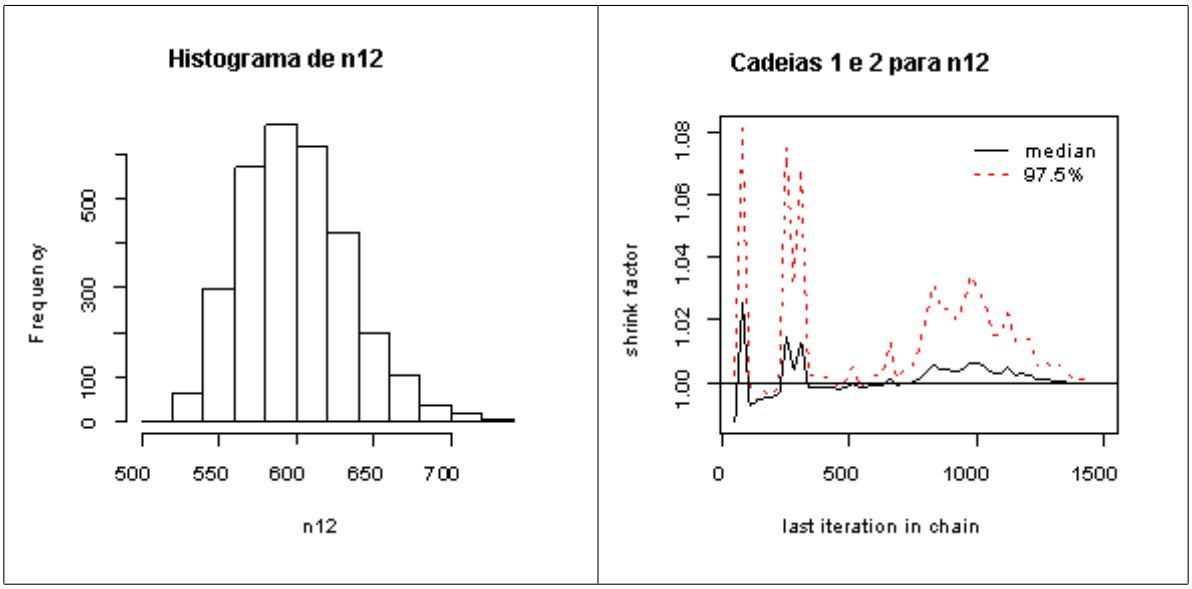


Figura 3.4.6 Histograma da distribuição *a posteriori* de n_{12} e gráfico do critério de convergência de Gelman Rubin das cadeias para o parâmetro n_{12}

A Figura 3.4.7 apresenta as densidades das distribuições *a posteriori* de ϕ_A e ϕ_B.

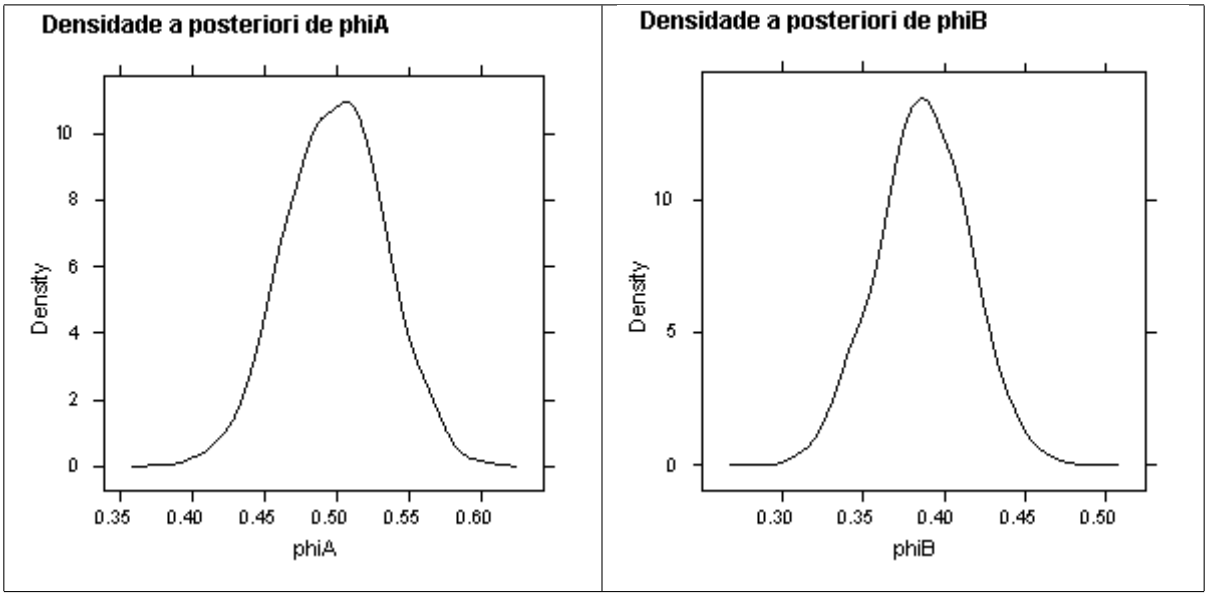


Figura 3.4.7 Densidades das distribuições *a posteriori* de ϕ_A e ϕ_B

Em seguida estimamos os parâmetros N, θ_1 e θ_2. A Tabela 3.4.10 apresenta os resumos aproximados das distribuições *a posteriori* destes parâmetros.

Tabela 3.4.10 Resumos aproximados das distribuições *a posteriori* de N, θ_1 e θ_2

	Média	Moda	Q_1	Mediana	Q_3	D.P.	I.C.(95%)	Ampl.	Conv.
N	15400	15258	15070	15380	15730	500,1	(14458; 16440)	1981	1
θ_1	0,27	0,27	0,27	0,27	0,28	0,01	(0,25; 0,29)	0,04	1
θ_2	0,14	0,14	0,14	0,14	0,15	0,01	(0,13; 0,15)	0,02	1

Notamos da Tabela 3.4.10 que obtivemos esitmativas dos parâmetros próximas as da Tabela 2.4.4, mas a amplitude do intervalo de credibilidade de N resultou menor do que do intervalo de confiança da Tabela 2.4.4.

A Figura 3.4.8 apresenta o histograma e o gráfico das seqüências da distribução *a posteriori* de N.

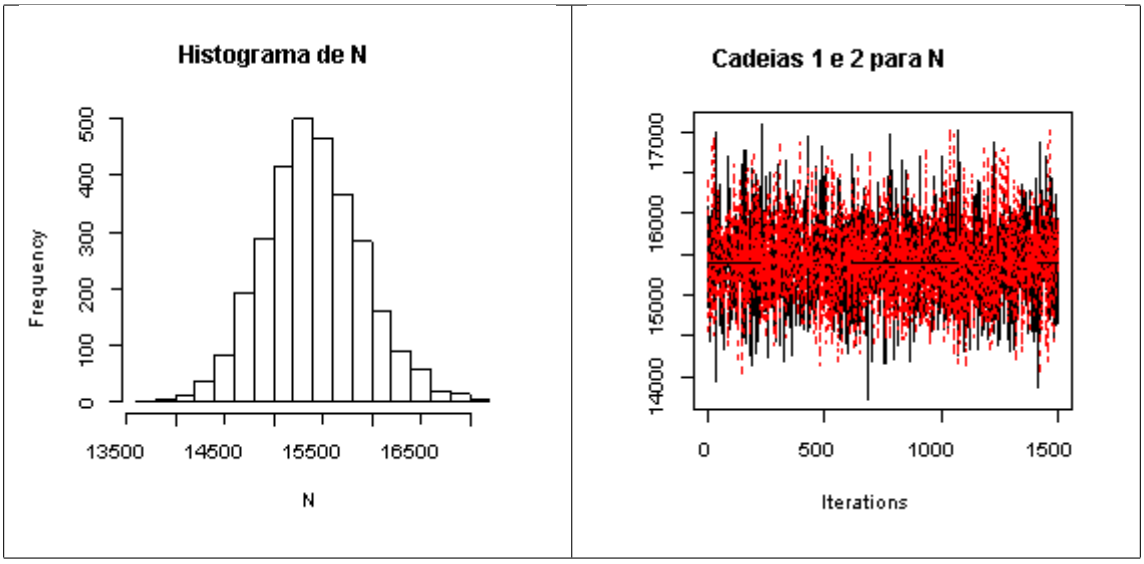


Figura 3.4.8 Histograma e gráfico das seqüências da distribução *a posteriori* do parâmetro N

A Figura 3.4.9 mostra as densidades das distribuições *a posteriori* de θ_1e θ_2.

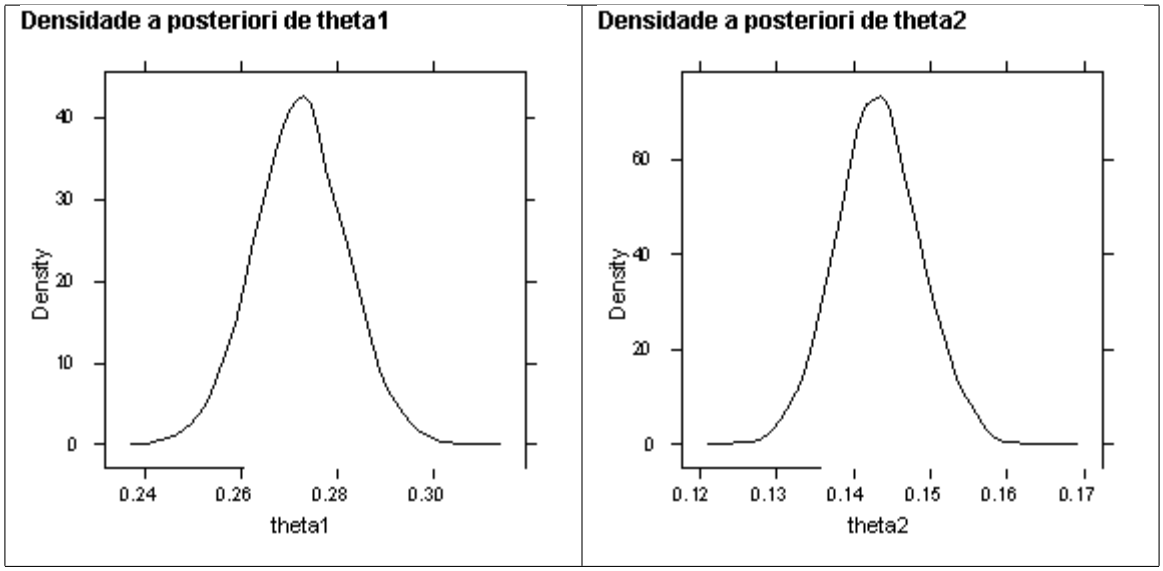


Figura 3.4.9 Densidades das distribuições *a posteriori* de θ_1e θ_2

Capítulo 4

Estimação Bayesiana do Tamanho Populacional: Duas Listas com Dados Particionados e Variável Latente

Neste capítulo apresentamos uma abordagem bayesiana para o problema de estimação do tamanho de uma população fechada, mas sob um enfoque diferente daquele apresentado no Capítulo 3. Como vimos, a estimação do tamanho populacional, N, segundo o modelo apresentado no Capítulo 3, depende da estimação do número de indivíduos coincidentes em ambas as listas, n_{12}. Isto é, primeiro temos de estimar n_{12} para, em seguida, estimar N. Neste capítulo, adotando a chamada técnica da "variável latente"definimos um modelo bayesiano que permite estimar n_{12} e N simultaneamente.

Na seção 4.1 definimos o modelo bayesiano com variável latente; na seção 4.2 tratamos de distribuições *a priori* não informativas para o tamanho populacional; na seção 4.3 consideramos uma distribuição *a priori* de máxima entropia para o tamanho populacional e na seção 4.4 apresentamos exemplos com dados simulados e reais.

4.1 Modelo Bayesiano com Variável Latente

Lembramos que

n_1 é o número de indivíduos pertencentes a lista 1;

n_2 é o número de indivíduos pertencentes a lista 2;

m_A é o número de indivíduos que possuem as fichas A's coincidentes em ambas as listas;

m_B é o número de indivíduos que possuem as fichas B's coincidentes em ambas as listas;

n_{AB} é o número de indivíduos que possuem as fichas A's e B's coincidentes nas duas listas;

n_{12} é uma variável latente que denota o número de indivíduos pertencentes às duas listas;

$n_T = m_A + m_B - n_{AB}$ e $n = n_1 + n_2 - n_{12}$.

Seja $D = (n_1, n_2, m_A, m_B, n_{AB})$ o vetor de dados. Então, de (2.3) e (2.5), segue que

$$P(D, n_{12} \mid N, \boldsymbol{\theta}, \boldsymbol{\phi}) = P(m_A, m_B, n_{AB} \mid n_{12}, \boldsymbol{\phi}) \times P(n_1, n_2, n_{12} \mid N, \boldsymbol{\theta})$$

$$= \frac{1}{(m_A - n_{AB})!\,(m_B - n_{AB})! n_{AB}!\,(n_{12} - n_T)!} \prod_{X=A}^{B} \phi_X^{m_X} (1 - \phi_X)^{n_{12} - m_X} \tag{4.1}$$

$$\times \frac{N!}{(n_1 - n_{12})!\,(n_2 - n_{12})!\,(N - n)!} \prod_{j=1}^{2} \theta_j^{n_j} (1 - \theta_j)^{N - n_j}.$$

Supomos *a priori* N, ϕ_A, ϕ_B, θ_1 e θ_2 independentes; N com função de probabilidades $\pi(N)$, $N = 1, 2, ...$; ϕ_X com distribuição beta de parâmetros α_X e β_X conhecidos, $X = A, B$ e θ_j com distribuição beta de parâmetros φ_j e δ_j conhecidos, $j = 1, 2$, isto é, a distribuição *a priori* conjunta de N, $\boldsymbol{\theta}$, $\boldsymbol{\phi}$ é dada por

$$\pi(N, \boldsymbol{\theta}, \boldsymbol{\phi}) = \pi(N)\,\pi(\boldsymbol{\theta})\,\pi(\boldsymbol{\phi})$$

$$= \pi(N) \prod_{j=1}^{2} \frac{\Gamma(\varphi_j + \delta_j)}{\Gamma(\varphi_j)\,\Gamma(\delta_j)} \theta_j^{\varphi_j - 1} (1 - \theta_j)^{\delta_j - 1} \tag{4.2}$$

$$\times \prod_{X=A}^{B} \frac{\Gamma(\alpha_X + \beta_X)}{\Gamma(\alpha_X)\,\Gamma(\beta_X)} \phi_X^{\alpha_X - 1} (1 - \phi_X)^{\beta_X - 1}.$$

Logo, de (4.1) e (4.2), temos que a distribuição *a posteriori* conjunta de N, n_{12}, $\boldsymbol{\theta}$, e $\boldsymbol{\phi}$ é tal que

$$\begin{aligned}
&\pi\left(N, n_{12}, \boldsymbol{\theta}, \boldsymbol{\phi} \mid D\right) = \frac{P\left(N, n_{12}, \boldsymbol{\theta}, \boldsymbol{\phi}, D\right)}{P\left(D\right)} \\
&\propto P\left(D, n_{12} \mid N, \boldsymbol{\theta}, \boldsymbol{\phi}\right) \times \pi\left(N, \boldsymbol{\theta}, \boldsymbol{\phi}\right) \\
&\propto \frac{\pi\left(N\right) N!}{(n_1 - n_{12})!\,(n_2 - n_{12})!\,(n_{12} - n_T)!\,(N-n)!} \prod_{X=A}^{B} \phi_X^{m_X+\alpha_X-1}\left(1-\phi_X\right)^{n_{12}-m_X+\beta_X-1} \\
&\times \prod_{j=1}^{2} \theta_j^{n_j+\varphi_j-1}(1-\theta_j)^{N-n_j+\delta_j-1},
\end{aligned} \tag{4.3}$$

Para determinar a variação de n_{12} em (4.3) notamos que $N - n_1 - n_2 + n_{12} \geq 0$ implica $n_{12} \geq n_1 + n_2 - N$. Logo, $t = \max\left\{n_T, n_1 + n_2 - N\right\} \leq n_{12} \leq m = \min\left\{n_1, n_2\right\}$.

Na próxima seção atribuímos *prioris* não informativas ao parâmetro N.

4.2 Distribuição *a priori* Não Informativa para o Tamanho Populacional

Nesta seção atribuímos a N as distribuições *a priori* uniforme e de Jeffreys nos inteiros estritamente positivos e determinamos as distribuições condicionais dos parâmetros. Suponhamos então que N tem distribuição *a priori* $\pi\left(N\right) = \dfrac{1}{N^r}$, $N = 1, 2, ...$, $r = 0, 1$. Logo, de (4.3), segue que a distribuição *a posteriori* conjunta de $(N, n_{12}, \boldsymbol{\theta}, \boldsymbol{\phi})$ é tal que

$$\begin{aligned}
&\pi\left(N, n_{12}, \boldsymbol{\theta}, \boldsymbol{\phi} \mid D\right) \propto \frac{N!}{(n_1 - n_{12})!\,(n_2 - n_{12})!\,(n_{12} - n_T)!\,(N-n)!N^r} \\
&\times \prod_{X=A}^{B} \phi_X^{m_X+\alpha_X-1}\left(1-\phi_X\right)^{n_{12}-m_X+\beta_X-1} \prod_{j=1}^{2} \theta_j^{n_j+\varphi_j-1}(1-\theta_j)^{N-n_j+\delta_j-1},
\end{aligned} \tag{4.4}$$

$N \geq n$, $t \leq n_{12} \leq m$, $0 < \theta_j < 1$, $j = 1, 2$, $0 < \phi_X < 1$, $X = A, B$.

O próximo teorema especifica sobre qual condição a distribuição *a posteriori* (4.4) existe.

Teorema 4.2.1 A distribuição $\pi\left(N, n_{12}, \boldsymbol{\theta}, \boldsymbol{\phi} \mid D\right)$ existe

1) para $r = 1$;

2) para $r = 0$ se $n_{12} + \varphi_1 + \varphi_2 > 1$, onde $t \leq n_{12} \leq m$.

Prova: Seja S^{-1} a constante normalizadora de (4.4). Logo,

$$S = \sum_{n_{12}=n_T}^{m} \sum_{N=n_1+n_2-n_{12}}^{\infty} \int_0^1 \int_0^1 \int_0^1 \int_0^1 \frac{N!}{(n_1-n_{12})!\,(n_2-n_{12})!\,(n_{12}-n_T)!\,(N-n_1-n_2+n_{12})!N^r}$$

$$\times \prod_{X=A}^{B} \phi_X^{m_X+\alpha_X-1}\left(1-\phi_X\right)^{n_{12}-m_X+\beta_X-1} \prod_{j=1}^{2} \theta_j^{n_j+\varphi_j-1}(1-\theta_j)^{N-n_j+\delta_j-1} d\phi_A d\phi_B d\theta_1 d\theta_2$$

$$\propto \sum_{n_{12}=n_T}^{m} \left[\frac{1}{(n_1-n_{12})!\,(n_2-n_{12})!\,(n_{12}-n_T)!} \frac{\Gamma\left(n_{12}-m_A+\beta_A\right)}{\Gamma\left(n_{12}+\alpha_A+\beta_A\right)} \frac{\Gamma\left(n_{12}-m_B+\beta_B\right)}{\Gamma\left(n_{12}+\alpha_B+\beta_B\right)} \right.$$

$$\left. \times \sum_{N=n_1+n_2-n_{12}}^{\infty} \frac{N!}{(N-n_1-n_2+n_{12})!N^r} \frac{\Gamma\left(N-n_1+\delta_1\right)}{\Gamma\left(N+\varphi_1+\delta_1\right)} \frac{\Gamma\left(N-n_2+\delta_2\right)}{\Gamma\left(N+\varphi_2+\delta_2\right)} \right]. \tag{4.5}$$

Para $t \leq n_{12} \leq m$, seja

$$B_{n_{12}} = \sum_{N=n_1+n_2-n_{12}}^{\infty} \frac{N!}{(N-n_1-n_2+n_{12})!N^r} \frac{\Gamma\left(N-n_1+\delta_1\right)}{\Gamma\left(N+\varphi_1+\delta_1\right)} \frac{\Gamma\left(N-n_2+\delta_2\right)}{\Gamma\left(N+\varphi_2+\delta_2\right)}.$$

De maneira análoga à prova do Teorema 3.2.1 pode-se provar que, para $t \leq n_{12} \leq m$,

$$B_{n_{12}} \leq \sum_{N=n_1+n_2-n_{12}}^{\infty} O\left(N^{-(n_{12}+\varphi_1+\varphi_2+r)}\right)\left(N \to \infty\right),$$

isto é, $B_{n_{12}} \leq \sum\limits_{N=n_1+n_2-n_{12}}^{\infty} b_N\left(n_{12}\right)$, onde $b_N\left(n_{12}\right) = O\left(N^{-(n_{12}+\varphi_1+\varphi_2+r)}\right)\left(N \to \infty\right)$. Então, existe um número real $I\left(n_{12}\right) > 0$ e um número inteiro positivo $n_0\left(n_{12}\right)$, $n_0\left(n_{12}\right) > n_1 + n_2 - n_{12}$, tal que $b_N\left(n_{12}\right) \leq I\left(n_{12}\right).N^{-(n_{12}+\varphi_1+\varphi_2+r)}$, para todo $N \geq n_0\left(n_{12}\right)$, e

$$B_{n_{12}} \leq \sum_{N=n_1+n_2-n_{12}}^{n_0(n_{12})} b_N(n_{12}) + \sum_{N=n_0(n_{12})+1}^{\infty} b_N(n_{12}) \tag{4.6}$$

$$\leq \sum_{N=n_1+n_2-n_{12}}^{n_0(n_{12})} b_N(n_{12}) + I(n_{12}) \sum_{N=n_0(n_{12})+1}^{\infty} N^{-(n_{12}+\varphi_1+\varphi_2+r)}.$$

Para $r = 1$ segue que $n_{12} + \varphi_1 + \varphi_2 + r > 1$, para todo n_{12} tal que $t \leq n_{12} \leq m$, o que implica $\sum_{N=n_0(n_{12})+1}^{\infty} N^{-(n_{12}+\varphi_1+\varphi_2+r)} < \infty$ e, de (4.6), temos que $B_{n_{12}} < \infty$, o que por sua vez implica, por (4.5), que $S^{-1} < \infty$ e (1) esta provado.

Para $r = 0$ se $n_{12} + \varphi_1 + \varphi_2 > 1$, para todo n_{12} tal que $t \leq n_{12} \leq m$, segue que $\sum_{N=n_0(n_{12})+1}^{\infty} N^{-(n_{12}+\varphi_1+\varphi_2+r)} < \infty$ e, de (4.6), temos $B_{n_{12}} < \infty$, o que por sua vez implica, por (4.5), que $S^{-1} < \infty$ e (2) está provado. ■

Apresentamos na seqüência as distribuições condicionais dos parâmetros. Segue, de (4.4) , que a função de probabilidades condicional de n_{12}, dados N, $\boldsymbol{\theta}$, $\boldsymbol{\phi}$ e D é tal que

$$\pi(n_{12} \mid N, \boldsymbol{\theta}, \boldsymbol{\phi}, D) \propto \frac{\left[\prod_{X=A}^{B}(1-\phi_X)\right]^{n_{12}}}{(n_1-n_{12})!\,(n_2-n_{12})!\,(n_{12}-n_T)!\,(N-n_1-n_2+n_{12})!}, \tag{4.7}$$

para $t \leq n_{12} \leq m$. A distribuição condicional de $\boldsymbol{\phi}$, dados N, n_{12}, $\boldsymbol{\theta}$ e D, é tal que

$$\pi(\boldsymbol{\phi} \mid N, n_{12}, \boldsymbol{\theta}, D) \propto \prod_{X=A}^{B} \phi_X^{m_X+\alpha_X-1}(1-\phi_X)^{n_{12}-m_X+\beta_X-1}, \tag{4.8}$$

$0 < \phi_X < 1$, $X = A, B$, ou seja, ϕ_A e ϕ_B são condicionalmente independentes, dados N, n_{12}, $\boldsymbol{\theta}$ e D, com distribuições beta de parâmetros $m_A + \alpha_A$, $n_{12} - m_A + \beta_A$ e $m_B + \alpha_B$, $n_{12} - m_B + \beta_B$, respectivamente. Ainda de (4.4), temos que a distribuição condicional de $\boldsymbol{\theta}$, dados N, n_{12}, $\boldsymbol{\phi}$ e D é tal que

$$\pi(\boldsymbol{\theta} \mid N, n_{12}, \boldsymbol{\phi}, D) \propto \prod_{j=1}^{2} \theta_j^{n_j+\varphi_j-1}(1-\theta_j)^{N-n_j+\delta_j-1}, \tag{4.9}$$

$0 < \theta_j < 1$, $j = 1, 2$, isto é, de (4.9) segue que θ_1 e θ_2 são condicionalmente independentes dados N, n_{12}, $\boldsymbol{\phi}$ e D, com distribuições beta de parâmetros $n_1 + \varphi_1$, $N - n_1 + \delta_1$ e $n_2 + \varphi_2$,

$N - n_2 + \delta_2$, respectivamente.

Para $r = 0$, segue de (4.4), que a função de probabilidades condicional de N, dados n_{12}, $\boldsymbol{\theta}$, $\boldsymbol{\phi}$ e D é tal que

$$\pi\left(N \mid n_{12}, \boldsymbol{\theta}, \boldsymbol{\phi}, D\right) \propto \binom{N}{n} \left[\prod_{j=1}^{2}(1-\theta_j)\right]^N, \tag{4.10}$$

$N \geq n$. O inverso da constante normalizadora de (4.10) é dada por

$$\begin{aligned}
&\sum_{N \geq n} \binom{N}{n} \left[\prod_{j=1}^{2}(1-\theta_j)\right]^N \\
&= \sum_{S \geq 0} \binom{S+n}{n} \left[\prod_{j=1}^{2}(1-\theta_j)\right]^{S+n} \\
&= \left[\prod_{j=1}^{2}(1-\theta_j)\right]^n \sum_{S \geq 0} \binom{S+n}{n} \left[\prod_{j=1}^{2}(1-\theta_j)\right]^S \\
&= \left[\prod_{j=1}^{2}(1-\theta_j)\right]^n \left[1 - \prod_{j=1}^{2}(1-\theta_j)\right]^{-(n+1)},
\end{aligned}$$

o que implica

$$\pi\left(N \mid n_{12}, \boldsymbol{\theta}, \boldsymbol{\phi}, D\right) = \binom{N}{n} \left[1 - \prod_{j=1}^{2}(1-\theta_j)\right]^{n+1} \left[\prod_{j=1}^{2}(1-\theta_j)\right]^{N-n}, \tag{4.11}$$

$N \geq n$. Notamos, como em (3.14), que (4.11) é igual a função de probabilidades de uma variável aleatória $T + n$, onde T tem função de probabilidades binomial negativa com parâmetros $n + 1$ e $1 - \prod_{j=1}^{2}(1-\theta_j)$.

Para $r = 1$, segue de (4.4), que a função de probabilidades condicional de N, dados n_{12}, $\boldsymbol{\theta}$, $\boldsymbol{\phi}$ e D, é tal que

$$\pi\left(N \mid n_{12}, \boldsymbol{\theta}, \boldsymbol{\phi}, D\right) \propto \binom{N-1}{n-1} \left[\prod_{j=1}^{2}(1-\theta_j)\right]^N, \tag{4.12}$$

$N \geq n$. Analogamente ao caso $r = 0$, temos que (4.12) é igual a função de probabilidades de uma variável aleatória $U + n$, onde U tem função de probabilidades binomial negativa com parâmetros n e $1 - \prod_{j=1}^{2} (1 - \theta_j)$.

Na próxima seção atribuímos uma distribuição *a priori* de máxima entropia para o tamanho populacional.

4.3 Distribuição *a priori* de Máxima Entropia para o Tamanho Populacional

Nesta seção supomos que N possui distribuição *a priori* de máxima entropia e cuja média é conhecida. Vimos na seção 3.3 que ao considerar somente a média de N, μ, como informação subjetiva do pesquisador obtemos a seguinte distribuição *a priori* de máxima entropia para N :

$$\pi(N) = \left(\frac{1}{1+\mu}\right)\left(\frac{\mu}{1+\mu}\right)^{N}, \tag{4.13}$$

$N = 0, 1, 2, ...$

Então, de (4.3) , temos que a distribuição *a posteriori* conjunta de N, n_{12}, $\boldsymbol{\theta}$ e $\boldsymbol{\phi}$ é tal que

$$\begin{aligned}\pi(N, n_{12}, \boldsymbol{\theta}, \boldsymbol{\phi} \mid D) \propto & \frac{N!\left(\frac{\mu}{1+\mu}\right)^{N}}{(n_1 - n_{12})!\,(n_2 - n_{12})!\,(n_{12} - n_T)!\,(N-n)!} \\ & \times \prod_{X=A}^{B} \phi_X^{m_X+\alpha_X-1}(1-\phi_X)^{n_{12}-m_X+\beta_X-1} \prod_{j=1}^{2} \theta_j^{n_j+\varphi_j-1}(1-\theta_j)^{N-n_j+\delta_j-1},\end{aligned} \tag{4.14}$$

$N \geq n$, $t \leq n_{12} \leq m$, $0 < \theta_j < 1$, $j = 1, 2$, $0 < \phi_X < 1$, $X = A, B$.

Em seguida apresentamos somente a distribuição condicional de N, dados n_{12}, $\boldsymbol{\theta}$, $\boldsymbol{\phi}$ e D, pois as distribuições condicionais de n_{12}, $\boldsymbol{\phi}$ e $\boldsymbol{\theta}$ são iguais à (4.7), (4.8) e (4.9) respectivamente.

Temos de (4.14), que a distribuição condicional de N, dados n_{12}, $\boldsymbol{\theta}$, $\boldsymbol{\phi}$ e D, é tal que

$$\pi\left(N \mid n_{12}, \boldsymbol{\theta}, \boldsymbol{\phi}, D\right) = \binom{N}{n}\left[\left(\frac{\mu}{1+\mu}\right)\prod_{j=1}^{2}(1-\theta_j)\right]^N, \tag{4.15}$$

$N \geq n$. O inverso da constante normalizadora de (4.15) é dado por

$$\sum_{N\geq n}\binom{N}{n}\left[\left(\frac{\mu}{1+\mu}\right)\prod_{j=1}^{2}(1-\theta_j)\right]^N$$

$$= \sum_{S\geq 0}\binom{S+n}{n}\left[\left(\frac{\mu}{1+\mu}\right)\prod_{j=1}^{2}(1-\theta_j)\right]^{S+n}$$

$$= \left[\left(\frac{\mu}{1+\mu}\right)\prod_{j=1}^{2}(1-\theta_j)\right]^n \sum_{S\geq 0}\binom{S+n}{n}\left[\left(\frac{\mu}{1+\mu}\right)\prod_{j=1}^{2}(1-\theta_j)\right]^S$$

$$= \left[\left(\frac{\mu}{1+\mu}\right)\prod_{j=1}^{2}(1-\theta_j)\right]^n \left[1-\left(\frac{\mu}{1+\mu}\right)\prod_{j=1}^{2}(1-\theta_j)\right]^{-(n+1)},$$

o que implica

$$\pi\left(N \mid n_{12}, \boldsymbol{\theta}, \boldsymbol{\phi}, D\right) = \binom{N}{n}\left[1-\left(\frac{\mu}{1+\mu}\right)\prod_{j=1}^{2}(1-\theta_j)\right]^{n+1}\left[\left(\frac{\mu}{1+\mu}\right)\prod_{j=1}^{2}(1-\theta_j)\right]^{N-n}, \tag{4.16}$$

$N \geq n$. Isto é, como em (3.14), (4.16) é igual a distribuição de uma variável aleatória $n+V$, onde V tem uma distribuição binomial negativa com parâmetros $(n+1)$ e $[1-\left(\frac{\mu}{1+\mu}\right)\prod_{i=1}^{2}(1-\theta_i)]$.

Na próxima seção atribuímos uma distribuição *a priori* hierárquica de Poisson para o parâmetro N.

4.4 Distribuição *a priori* Hierárquica de Poisson para o Tamanho Populacional

Nesta seção atribuímos a N a distribuição *a priori* de Poisson com média λ e atribuímos ao hiperparâmetro λ a distribuição gama com parâmetros a e b conhecidos, $a > 0$ e $b > 0$. Logo, a distribuição *a priori* conjunta de N e λ é dada por

$$\pi(N, \lambda) \propto \frac{e^{-\lambda(b+1)}\lambda^{N+a-1}}{N!} \tag{4.17}$$

$N = 0, 1, 2, ...$ e $\lambda > 0$.

Então, de (4.3), segue que a distribuição *a posteriori* conjunta de N, n_{12}, λ, $\boldsymbol{\theta}$ e $\boldsymbol{\phi}$ é tal que

$$\begin{aligned}\pi(N, n_{12}, \lambda, \boldsymbol{\theta}, \boldsymbol{\phi} \mid D) &\propto \frac{e^{-\lambda(b+1)}\lambda^{N+a-1}}{(n_1 - n_{12})!\,(n_2 - n_{12})!\,(n_{12} - n_T)!\,(N-n)!} \\ &\times \prod_{X=A}^{B} \phi_X^{m_X+\alpha_X-1}(1-\phi_X)^{n_{12}-m_X+\beta_X-1} \prod_{j=1}^{2} \theta_j^{n_j+\varphi_j-1}(1-\theta_j)^{N-n_j+\delta_j-1},\end{aligned} \tag{4.18}$$

$N \geq n$, $t \leq n_{12} \leq m$, $0 < \theta_j < 1$, $j = 1, 2$, $0 < \phi_X < 1$, $X = A, B$, $\lambda > 0$.

Em seguida, apresentamos somente as distribuições condicionais de N, dados n_{12}, λ, $\boldsymbol{\theta}$, $\boldsymbol{\phi}$, D e de λ dados N, n_{12}, $\boldsymbol{\theta}$, $\boldsymbol{\phi}$ e D, pois as distribuições condicionais de n_{12}, $\boldsymbol{\phi}$ e $\boldsymbol{\theta}$ são dados por (4.7), (4.8) e (4.9) respectivamente.

Temos de (4.18) que a distribuição condicional de N, dados n_{12}, λ, $\boldsymbol{\theta}$, $\boldsymbol{\phi}$ e D é tal que

$$\pi(N \mid n_{12}, \lambda, \boldsymbol{\theta}, \boldsymbol{\phi}, D) \propto \frac{1}{(N-n)!}\left[\lambda \prod_{j=1}^{2}(1-\theta_j)\right]^N, \tag{4.19}$$

$N \geq n$. O inverso da constante normalizadora de (4.19) é dado por

$$\sum_{N\geq n} \frac{1}{(N-n)!}\left[\lambda \prod_{j=1}^{2}(1-\theta_j)\right]^N$$

$$= \sum_{S=0}^{\infty} \frac{1}{S!}\left[\lambda \prod_{j=1}^{2}(1-\theta_j)\right]^{S+n}$$

$$= \left[\lambda \prod_{j=1}^{2}(1-\theta_j)\right]^n \exp\left\{\lambda \prod_{j=1}^{2}(1-\theta_j)\right\},$$

o que implica

$$\pi\left(N \mid n_{12}, \lambda, \boldsymbol{\theta}, \boldsymbol{\phi}, D\right) = \frac{1}{(N-n)!}\exp\left\{-\lambda \prod_{j=1}^{2}(1-\theta_j)\right\}\left[\lambda \prod_{j=1}^{2}(1-\theta_j)\right]^{N-n}, \quad (4.20)$$

$N \geq n$. Isto é, (4.20) é igual a distribuição de uma variável aleatória $n+W$, onde W tem distribuição de Poisson com parâmetro $\lambda \prod_{j=1}^{2}(1-\theta_j)$.

Por outro lado, de (4.18) segue que a distribuição condicional de λ dados N, n_{12}, $\boldsymbol{\theta}$, $\boldsymbol{\phi}$ e D é tal que

$$\pi\left(\lambda \mid N, n_{12}, \boldsymbol{\theta}, \boldsymbol{\phi}, D\right) \propto \lambda^{(N+a)-1} e^{-\lambda(b+1)}, \quad (4.21)$$

$\lambda > 0$, isto é, (4.21) tem distribuição gama com parâmetros $N+a$ e $b+1$.

Na próxima seção apresentamos exemplos com dados simulados e reais.

4.5 Exemplos com Dados Simulados e Reais

Nesta seção apresentamos exemplos com dados simulados e reais, onde utilizamos o modelo e as distribuições *a priori* tratados neste capítulo. A metodologia utilizada na simulação é análoga aquela dada na seção 3.4.

4.5.1 Exemplo com Dados Simulados

Nesta seção utilizamos novamente as mesmas estatísticas da seção 2.4.1. A Tabela 4.4.1 mostra os valores simulados de n_1, n_2, n_{12}, n_A, n_B e n_{AB}.

Tabela 4.4.1 Quantidades geradas da distribuição multinomial

Valores fixados					Valores gerados					
N	θ_1	θ_2	ϕ_A	ϕ_B	n_1	n_2	n_{12}	n_A	n_B	n_{AB}
50	0,7	0,8	0,7	0,9	33	38	25	2	7	15
50	0,8	0,7	0,2	0,1	37	33	22	2	4	0
50	0,1	0,3	0,7	0,7	3	19	2	1	1	0
50	0,1	0,3	0,2	0,1	3	19	2	1	1	0
500	0,6	0,8	0,7	0,8	299	403	235	38	73	115
500	0,9	0,7	0,3	0,2	453	344	307	74	45	20
500	0,1	0,4	0,8	0,6	46	203	18	4	1	11
500	0,2	0,3	0,1	0,2	101	159	30	3	3	1
2000	0,6	0,8	0,7	0,9	1173	1576	905	67	238	576
2000	0,5	0,6	0,4	0,3	947	1168	528	157	95	68
2000	0,2	0,4	0,7	0,8	424	827	156	19	37	93
2000	0,2	0,1	0,3	0,4	411	209	39	9	11	1

No exemplo 4.4.1 apresentamos os resumos aproximados da distribuição *a posteriori*, onde utilizamos a distribuição *a priori* uniforme para o tamanho populacional.

Exemplo 4.4.1 Neste exemplo consideramos $\boldsymbol{\phi}$ e $\boldsymbol{\theta}$ com distribuição uniforme, ou seja, uma *priori* uniforme no intervalo $(0,1)$. Então, temos $\alpha_X = \beta_X = 1$, $X = A, B$; $\varphi_j = \delta_j = 1$, $j = 1,2$, e tomamos $r = 0$, isto é, a distribuição *a priori* uniforme para o tamanho populacional. A Tabela 4.4.2 apresenta os resumos aproximados das distribuições *a posteriori* de N, $\boldsymbol{\theta}$ e $\boldsymbol{\phi}$ para as estatísticas correspondentes a $N = 2000$, $\theta_1 = 0,6$, $\theta_2 = 0,8$, $\phi_A = 0,7$ e $\phi_B = 0,9$.

Tabela 4.4.2 Resumos aproximados da distribuição *a posteriori* de N, $\boldsymbol{\theta}$ e $\boldsymbol{\phi}$

	Média	Moda	Q_1	Mediana	Q_3	D.P.	I.C.(95%)	Ampl..
N	2034	2031	2017	2034	2052	25,84	(1985; 2087)	102
θ_1	0,58	0,58	0,57	0,58	0,59	0,01	(0,55; 0,60)	0,05
θ_2	0,77	0,78	0,77	0,77	0,78	0,01	(0,75; 0,80)	0,05
ϕ_A	0,71	0,70	0,70	0,71	0,72	0,02	(0,67; 0,74)	0,07
ϕ_B	0,89	0,90	0,89	0,89	0,90	0,01	(0,87; 0,92)	0,05

Da tabela acima observamos que as estimativas aproximadas (médias) de N, $\boldsymbol{\theta}$ e $\boldsymbol{\phi}$ estão próximas de seus respectivos valores verdadeiros e seus intervalos de credibilidade contém estes valores.

A Figura 4.4.1 apresenta o gráfico do comportamento das cadeias da distribuição *a posteriori* de N e o histograma da distribuição *a posteriori* de N.

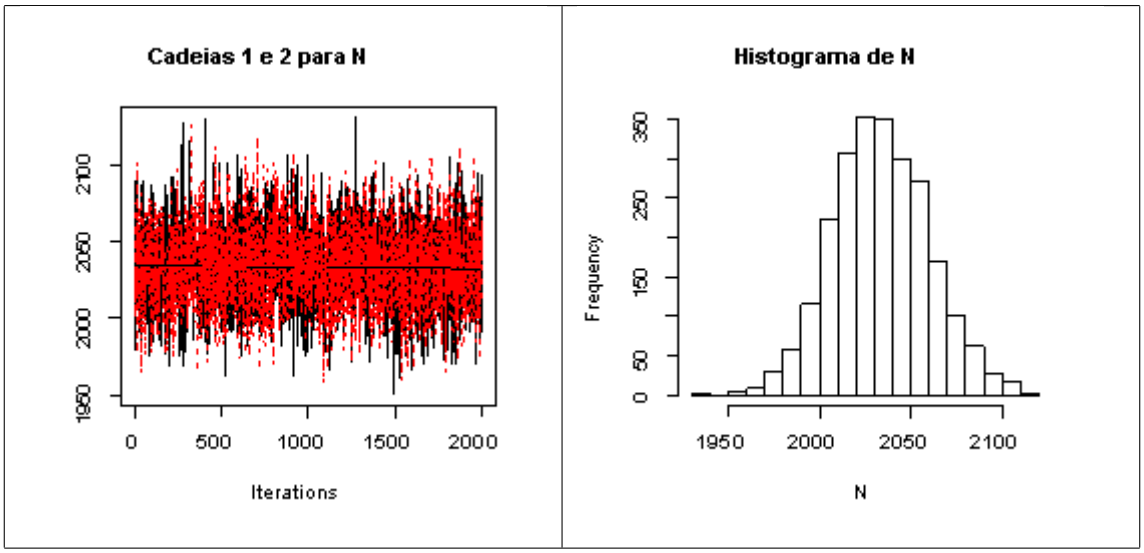


Figura 4.4.1 Gráfico do comportamento das cadeias e histograma da distribuição *a posteriori* de N

Na Figura 4.4.2 temos as densidades da distribuição *a posteriori* de $\boldsymbol{\theta}$ e $\boldsymbol{\phi}$.

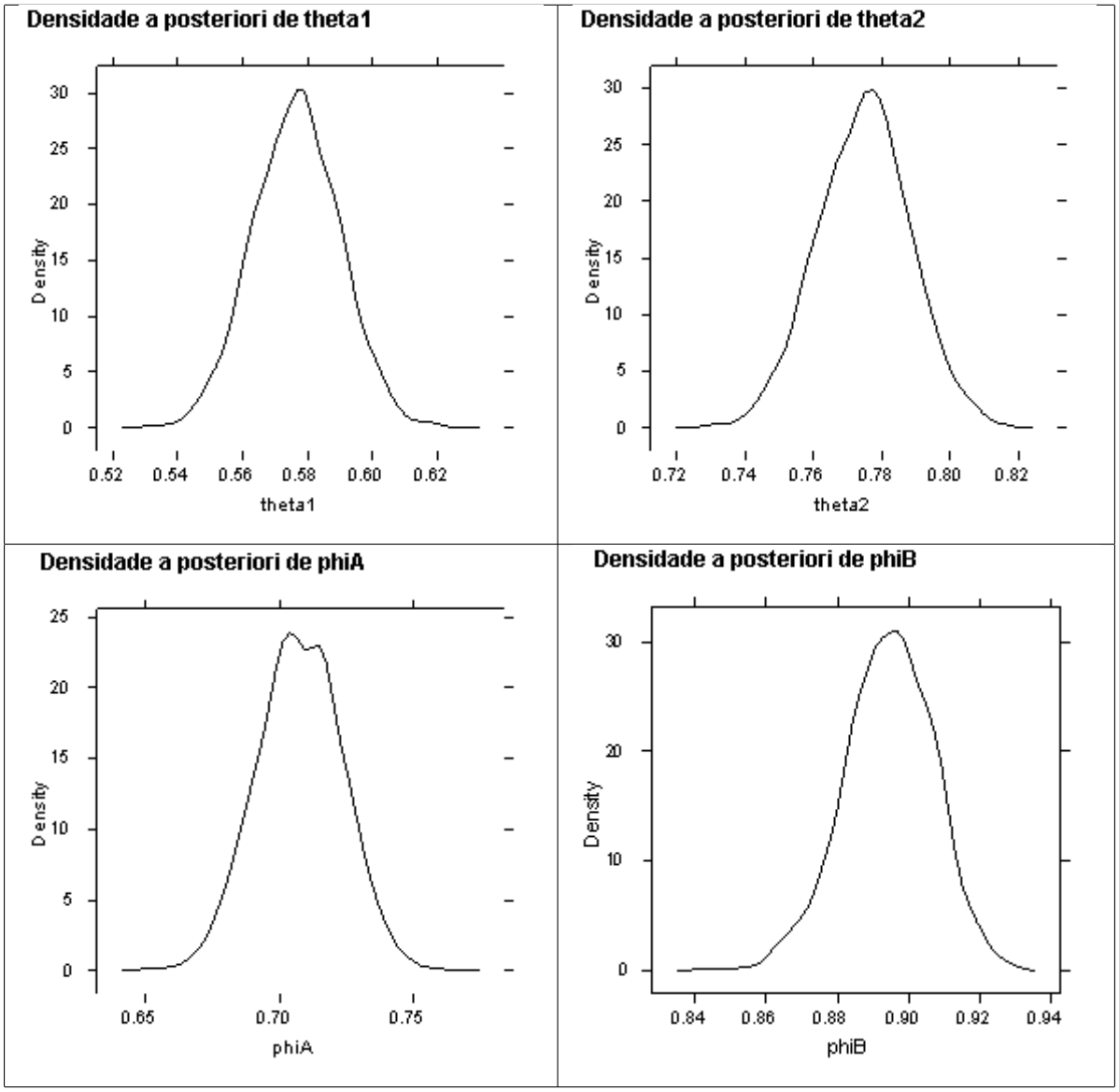


Figura 4.4.2 Densidades das distribuições *a posteriori* de $\boldsymbol{\theta}$ e $\boldsymbol{\phi}$

Na seqüência mostramos somente os resumos aproximados das distribuições *a posteriori* de N para as estatísticas da Tabela 4.4.1.

A Tabela 4.4.3 apresenta os resumos aproximados das distribuições *a posteriori* de N.

Tabela 4.4.3 Resumos aproximados das distribuições *a posteriori* de N

N	Média	Moda	Q_1	Mediana	Q_3	D.P.	I.C.(95%)	Ampl.
50	48, 18	48	46	48	50	3, 85	(41; 57)	16
50	80, 2	73	54	74	97	34, 14	(40; 162)	122
50	34, 76	30	24	28	37	22, 86	(19; 80)	61
50	34, 76	30	24	28	37	22, 86	(19; 80)	61
500	493	488	470	483	493	15, 30	(465; 512)	47
500	530, 9	530	482	516	562	58, 71	(465; 663)	198
500	550, 6	579	479	542, 5	626	113, 22	(386; 825)	439
500	1265	998	754, 5	1117	1626	721, 52	(332; 2851)	2518
2000	2034	2031	2017	2034	2052	25, 84	(1985; 2087)	102
2000	2040	2038	1939	2042	2160	165, 06	(1719; 2360)	641
2000	2235	2210	2141	2231	2325	142, 07	(1978; 2539)	561
2000	1151	894	684	970	1427	650, 333	(458; 2908)	2450

Notamos da Tabela 4.4.3 que obtivemos boas estimativas e intervalos de credibilidade para o tamanho populacional como os obtidos na seção 3.4.1. Notamos também que para valores baixos das probabilidades $\boldsymbol{\theta}$ ou $\boldsymbol{\phi}$ não obtivemos boas estimativas.

Comparando estes resultados com os do exemplo 3.4.1, onde, utilizamos as mesmas estatísticas geradas notamos que obtivémos estimativas muito parecidas, porém o intervalo de credibilidade para o modelo bayesiano com variável latente apresentou amplitudes maiores.

Nesta seção não apresentamos exemplos utilizando as distribuições *a priori* de Jeffreys e de máxima entropia, pois através de um estudo que fizemos obtivemos valores próximos aos da Tabela 4.4.3.

A seguir apresentamos um exemplo utilizando a distribuição *a priori* hierárquica de poisson para o tamanho populacional.

Exemplo 4.4.2 Neste exemplo utilizamos a distribuição *a priori* hierárquica de Poisson para o tamanho populacional. Consideramos as mesmas quantidades geradas da Tabela 4.4.1. Supomos que $a = 0, 01$ e $b = 0, 001$, isto é, atribuímos ao hiperparâmetro λ

a distriuição *a priori* não informativa e continuamos assumindo $\alpha_X = \beta_X = 1$, $X = A, B$; $\varphi_j = \delta_j = 1$, $j = 1, 2$.

Apresentamos a seguir as estimativas aproximadas para o tamanho populacional.

Tabela 4.4.4 Resumos aproximados das distribuições *a posteriori* de N

N	Média	Moda	Q_1	Mediana	Q_3	D.P.	I.C.(95%)	Ampl.
50	57, 66	56	54	57	59	7, 16	(44; 76)	32
50	61, 83	62	56	61	70	25, 76	(45; 93)	52
50	42	40	32	41	43	21, 86	(29; 93)	64
50	43	40	32	41	36	21, 85	(29; 92)	63
500	514, 15	513	480	515	491	15, 68	(471; 531)	61
500	540	539	491	533	582	68, 71	(472; 669)	197
500	572	580	498	570	623	123, 02	(481; 802)	321
500	791	794	754, 5	789	862	431, 31	(332; 901)	569
2000	2043, 1	2041	2016	2042	2052	31, 13	(1986; 2079)	93
2000	2048, 03	2048	1947	2047	2165	160, 34	(1739; 2384)	645
2000	2145	2144	2138	2144	2301	132, 24	(1981; 2513)	532
2000	1143	903	688	970	1430	580, 5	(329; 2009)	1680

Notamos na Tabela 4.4.4 que as estimativas aproximadas de N estão próximas dos seus respectivos valores verdadeiros, mas comparando estas estimativas com as da Tabela 4.4.3, observamos que elas se apresentaram com valores um pouco mais elevados.

4.5.2 Exemplo com Dados Reais

Utilizamos novamente os dados do exemplo 2.4.2. Lembremos que $n_1 = 4186$, $n_2 = 2203$, $m_A = 298$, $m_B = 231$ e $n_{AB} = 116$. Nesta seção utilizamos somente a distribuição *a priori* de Jeffreys para o tamanho populacional.

Exemplo 4.4.3 Neste exemplo consideramos novamente que $\boldsymbol{\theta}$ e $\boldsymbol{\phi}$ possuem distribuições *a priori* não informativas e assumimos que N tem distribuição *a priori* de Jeffreys. A Tabela 4.4.5 apresenta os resumos da distribuição *a posteriori* para os parâmetros

do modelo.

Tabela 4.4.5 Resumos aproximados das distribuições *a posteriori* de N, $\boldsymbol{\theta}$ e $\boldsymbol{\phi}$

	Média	Moda	Q_1	Mediana	Q_3	D.P.	I.C.(95%)	Ampl..
N	15356	15311	14762	15360	16101	934,04	(13638; 17200)	3562
θ_1	0,27	0,27	0,26	0,270	0,28	0,02	(0,55; 0,60)	0,05
θ_2	0,15	0,14	0,14	0,14	0,15	0,03	(0,12; 0,16)	0,04
ϕ_A	0,50	0,51	0,48	0,50	0,52	0,03	(0,43; 0,56)	0,13
ϕ_B	0,39	0,40	0,37	0,39	0,41	0,03	(0,33; 0,45)	0,12

Notamos da tabela acima que obtivemos estimativas próximas as da Tabela 2.4.4 e da Tabela 3.4.10, mas o intervalo de credibilidade de N se apresentou com amplitude maior do que o intervalo visto na Tabela 3.4.10. Missiagia (2005) obteve estimativas muito próximas as da Tabela 4.4.5.

A Figura 4.4.4 mostra o histograma da distribuição *a posteriori* de N.

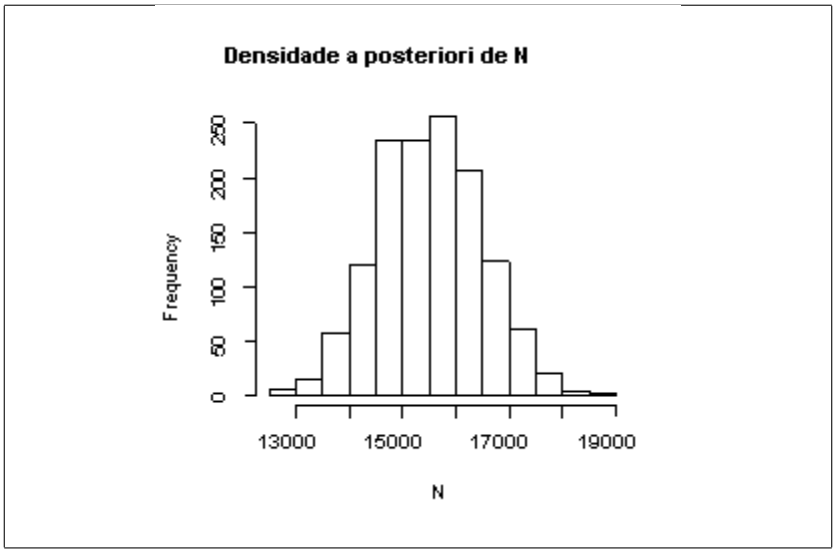


Figura 4.4.4 Histograma da distribuição *a posteriori* de N

Na Figura 4.4.5 apresentamos as densidades da distribuição *a posteriori* de $\boldsymbol{\theta}$ e $\boldsymbol{\phi}$.

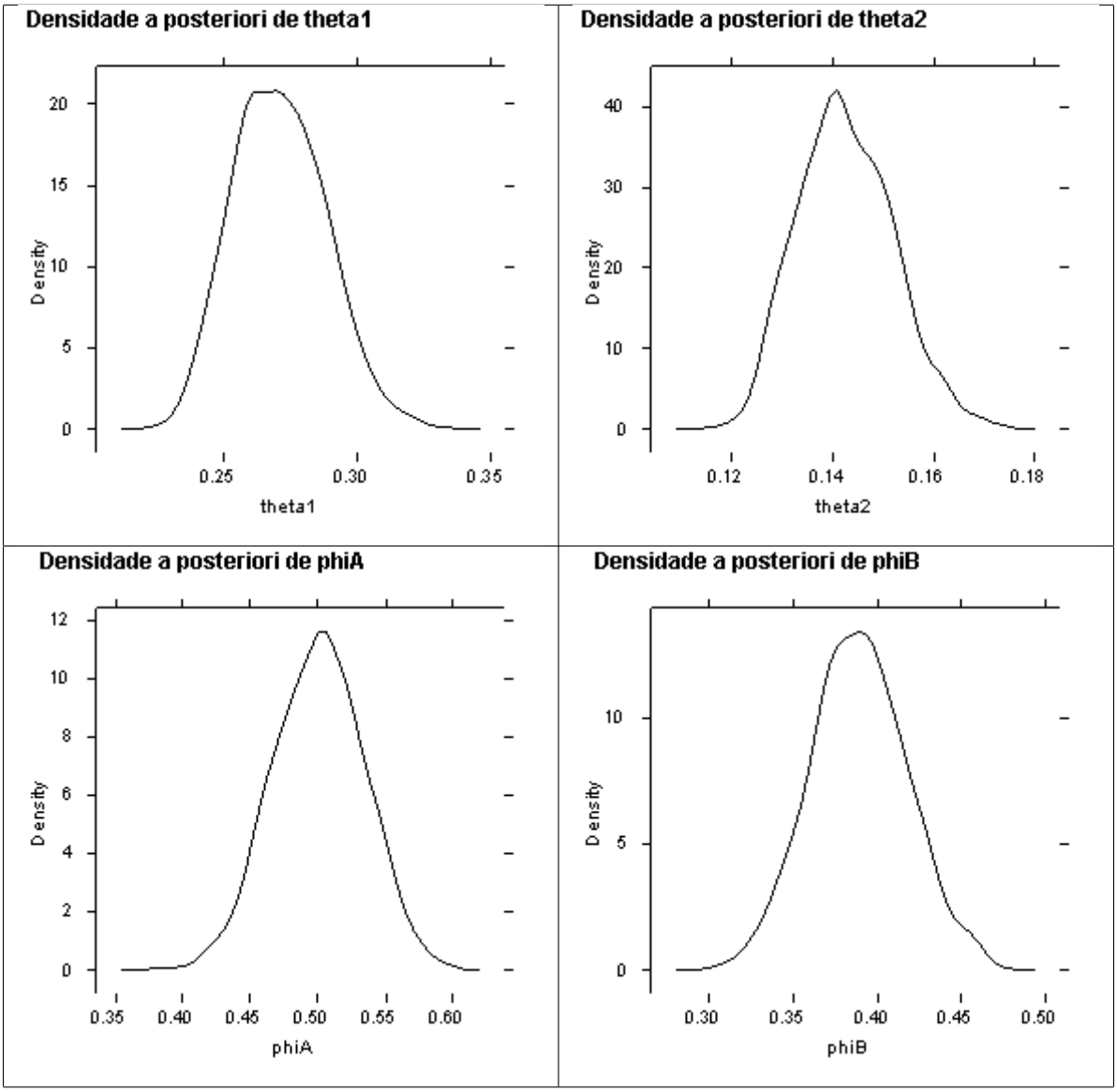


Figura 4.4.5 Densidades da distribuição *a posteriori* de $\boldsymbol{\theta}$ e $\boldsymbol{\phi}$

4.6 Conclusão

Comparando os desenhos dos modelos utilizados através dos dados simulados, notamos que os modelos bayesianos e bayesiano com variável latente produzem estimativas de N próximas umas das outras, mas os intervalos de credibilidade para os modelos bayesianos se apresentaram com amplitudes menores do que os correspondentes ao modelo bayesiano com variável latente. Por outro lado, quando o número de elementos identificados nas duas listas é pequeno, tanto os modelos bayesianos e clássico (capítulo 2) produzem péssimas

estimativas ou simplesmente não produzem estimativas, respectivamente, para o tamanho populacional.

Uma observação importante a ser feita em relação aos modelos bayesianos é que ao utilizarmos a estratégia de primeiramente estimar n_{12} para assim inferir sobre N, estamos determinando um valor aproximado da estimativa (média) de n_{12} e com este valor, determinamos uma estimativa aproximada para o tamanho populacional, ou seja, estamos realizando duas aproximações. Por outro lado, no modelo bayesiano com variável latente estimamos conjuntamente os parâmetros do modelo. Deste modo, apesar do modelo bayesiano apresentar intervalos de credibilidade com amplitudes menores, podemos dizer que o modelo bayesino com variável latente apresenta resultados mais fidedignos.

Capítulo 5

Apêndices

A - Geração de Quantidades da Distribuição Multinomial

```
N<-1000; theta1<-.7; theta2<-.8; phiA<-.6; phiB<-.8
w<-matrix(c(1,0,1,0,0,1,1,0),4,2)
p<-numeric()
for (i in 1:4)
{
p[i]<-(theta1^w[i,1])*((1-theta1)^(1-w[i,1]))*(theta2^w[i,2])*((1-theta2)^(1-w[i,2]))
}
amostra<-rmultinom(N,1, c(p[1],p[2],p[3],p[4]))
n1<-sum(amostra[1,])+sum(amostra[3,])
n2<-sum(amostra[2,])+sum(amostra[3,])
n12<-sum(amostra[3,])
n<-n1+n2-n12
w1<-matrix(c(1,0,1,0,0,1,1,0),4,2)
pi<-numeric()
for(i in 1:4)
{
pi[i]<-(phiA^w1[i,1])*((1-phiA)^(1-w1[i,1]))*(phiB^w1[i,2])*((1-phiB)^(1-w1[i,2]))
}
amostra1<-rmultinom(n12,1, c(pi[1],pi[2],pi[3],pi[4]))
```

```
nA<-sum(amostra1[1,])
nB<-sum(amostra1[2,])
nAB<-sum(amostra1[3,])
nT<-nA+nB+nAB
mA<-nA+nAB
mB<-nB+nAB
```

B - Algoritmo Gibbs Sampling com Metropolis-Hastings em Dois Estágios

```
c1phiA<-numeric()
c1phiB<-numeric()
c1n12<-numeric()
li<-(nT)
ls<-(min(n1,n2))
l<-30000
c1n12[1]<-900
for(i in 2:l){
c1phiA[i]<-round(rbeta(1,mA+1,c1n12[i-1]-mA+1),4)
c1phiB[i]<-round(rbeta(1,mB+1,c1n12[i-1]-mB+1),4)
x<-round(runif(1,li,ls))
p<-lgamma(c1n12[i-1]-nT)-lgamma(x-nT)+lgamma(x)-lgamma(c1n12[i-1])+(x-c1n12[i-
1])*
(log(1-c1phiA[i])+log(1-c1phiB[i]))
alpha<-min(1,exp(p))
u<-runif(1)
if(u<alpha) c1n12[i]<-x else c1n12[i]<-c1n12[i-1]
cat('\n',i,"C1",c1n12[i],c1phiA[i],c1phiB[i])
}
c1theta1<-numeric()
c1theta2<-numeric()
c1N<-numeric()
n12e<-599
```

```
c1N[1]<-6000
l<-40000
for(i in 2:l){
c1theta1[i]<-rbeta(1,n1+1,c1N[i-1]-n1+1)
c1theta2[i]<-rbeta(1,n2+1,c1N[i-1]-n2+1)
W<-rnbinom(1,n1+n2-n12e,1-(1-c1theta1[i])*(1-c1theta2[i]))
c1N[i]<-(W+n1+n2-n12e)
cat('\n',i,"C1",c1N[i],c1theta1[i],c1theta2[i])
}
```

C - Algoritmo Gibbs Sampling com Metropolis-Hastings e Variável Latente

```
set.seed(400)
c1phiA<-numeric()
c1phiB<-numeric()
c1theta1<-numeric()
c1theta2<-numeric()
c1N<-numeric()
c1n12<-numeric()
l<-50000
ls<-(min(n1,n2)-1)
c1n12[1]<-28
c1N[1]<-47
for(i in 2:l){
c1phiA[i]<-rbeta(1,mA+1,c1n12[i-1]-mA+1)
c1phiB[i]<-rbeta(1,mB+1,c1n12[i-1]-mB+1)
c1theta1[i]<-rbeta(1,n1+1,c1N[i-1]-n1+1)
c1theta2[i]<-rbeta(1,n2+1,c1N[i-1]-n2+1)
ls<-t<-max(nT,n1+n2-c1N[i])
W<-rnbinom(1,(n1+n2-c1n12[i-1]+1),(1-(1-c1theta1[i])*(1-c1theta2[i])))
c1N[i]<-W+n1+n2-c1n12[i-1]
```

```
x<-round(runif(1,li,ls))  p<-lgamma(n1-c1n12[i-1]+1)-lgamma(n1-x+1)+lgamma(n2-
c1n12[i-1]+1)
-lgamma(n2-x+1)+lgamma(c1n12[i-1]-nT+1)-lgamma(x-nT+1)+lgamma(N-n1-n2+c1n12[i-
1]+1)
-lgamma(N-n1-n2+x+1)+(x-c1n12[i-1])*(log(1-c1phiA[i])+log(1-c1phiB[i]))
alpha<-min(1,exp(p))
u<-runif(1)
if(u<alpha) c1n12[i]<-x else c1n12[i]<-c1n12[i-1]
cat('\n',i,"C1",c1N[i],c1n12[i],c1theta1[i],c1theta2[i],c1phiA[i],c1phiB[i])
}
```

Referências Bibliográficas

[1] Agresti A. (1994) . Simple capture-recapture models permiting unequal catchability and variable sampling effort. *Biometrics*, **50**: 494-500.

[2] Chapman, D. G. (1951). Some properties of the hypergeometric distribution with applications to zoological sample censures. *U. Cal. Publ. Statist.*, **1**, 131-160.

[3] Dorazio R. M. Royle J. A.(2003). Mixture models for estimating the size of a closed population when capture rates vary among individuals. *Biometrics*, **59**: 351-364.

[4] Feller, W. An introduction to the theory of probability and its applications. New York: John Wiley and Sons, 1967.

[5] Fienberg, S. E.; Johnson, M. S.; Junker, B. W. (1999). Classical multilevel and bayesian approaches to population size estimation using multiple lists. *J. R. Statist. Soc.*, **162A**, n.3, 383-405.

[6] Micheletti L. R. (2003). Aplicação da Metodologia da Verossimilhança na Prevalência do Diabetes. Dissertação de Mestrado em Estatística, Centro de Ciências Exatas e Tecnologia, Departamento de Estatística, Universidade Federal de São Carlos.

[7] Missiagia, J. G.(2005) . Estimação Bayesiana do tamanho de uma população de diabéticos através de listas de pacientes. Dissertação de Mestrado em Estatística, Centro de Ciências Exatas e Tecnologia, Departamento de Estatística, Universidade Federal de São Carlos.

[8] Sanathanan L. (1972). Estimating the size of a multinomial population. *Annals of Mathematical Statistics*, **43**, 142-152.

[9] Seber, G. A. F.; Huakau, J. T.; Simmons, D. (2000). Capture-recapture, epidemiology and list mismatches: two lists. *Biometrics*, **56**, 1227-1232.

[10] Seber, G. A. F.; Felton, R. (1981). Tag loss and the Petersen mark-recapture experiment. *Biometrika*, **68**, 211-219.

[11] Smith, P. J.(2002). Bayesian analyses for a multiple capture-recapture model. *Biometrika*, **78**, 399-407.

[12] Wang, X., He C. Z., Sum D. (2005) . Bayesian inference on the patient population size given list mismatches. *Statistics in Medicine*, **24**, 249-267.

[13] Wittes, J. T.; Sidel, V. W.(1968). A generalization of the simple capture-recapture model with applications to epidemiological research. *Journal of Chronic Diseases*, **21**, 287-301.