UNIVERSIDADE FEDERAL DE SÃO CARLOS – UFSCar
CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA
DEPARTAMENTO DE ESTATÍSTICA

JULIANO GALLINA MISSIAGIA

ESTIMAÇÃO BAYESIANA DO TAMANHO DE UMA POPULAÇÃO DE DIABÉTICOS ATRAVÉS DE LISTAS DE PACIENTES

SÃO CARLOS – SP

2005

ESTIMAÇÃO BAYESIANA DO TAMANHO DE UMA POPULAÇÃO DE DIABÉTICOS ATRAVÉS DE LISTAS DE PACIENTES

Juliano Gallina Missigia

Orientador: Prof. Dr. José Galvão Leite

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

São Carlos

Fevereiro de 2005

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

M678eb

Missiagia, Juliano Gallina.
Estimação Bayesiana do tamanho de uma população de diabéticos através de listas de pacientes / Juliano Gallina Missiagia. -- São Carlos : UFSCar, 2012.
113 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2005.

1. Teoria da estimativa. 2. Estimativa de população. 3. Inferência bayesiana. 4. Processo sequencial de captura-recaptura. I. Título.

CDD: 519.544 (20ª)

Programa de Pós-Graduação em Estatística
UNIVERSIDADE FEDERAL DE SÃO CARLOS
DEPARTAMENTO DE ESTATÍSTICA

Submetida à defesa pública no dia 25/02/05,
tendo sido Aprovado.

Presidente: José Galvão Leite
Membros:

Agradeço,

A Deus que me protege e me guia em cada instante da minha vida. Aos meus pais Edmar e Creusa, e a toda minha família que acreditaram em minha capacidade. Ao professor Dr. José Galvão Leite pela orientação e pelas idéias durante todo este trabalho e pela sua paixão e perseverança pelo ensino e pesquisa que sempre terei como exemplo, e ao professor Dr. Luis Aparecido Milan que me coorientou neste trabalho. À CAPES (Coordenação de Aperfeiçoamento Pessoal de Nível Superior) pela assistência financeira. Aos colegas, professores e funcionários do Departamento de Estatística da UFSCar, pela grande amizade.

A Mariana pelo seu incentivo e amor, a Érica pela amizade e carinho e aos meninos da república que sempre estiveram ao meu lado.

Resumo

Nesta dissertação apresentamos uma metodologia bayesiana para estimar o tamanho de uma população de diabéticos através de listas contendo informações sobre dados dos indivíduos. A metodologia aplicada é análoga a de captura-recaptura em população animal. Supomos corretos os registros de informações relativas aos pacientes assim como levamos em consideração registros corretos e incorretos das informações. No caso da suposição dos registros serem corretos, a metodologia é desenvolvida para duas ou mais listas e determinamos estimativas de Bayes para o tamanho populacional. Em um segundo modelo, consideramos a ocorrência de registros corretos e incorretos dos dados relativos aos pacientes, e apresentamos um método de estimação em dois estágios para os parâmetros do modelo utilizando duas listas. Para ambos os modelos, apresentamos resultados com exemplos simulados e reais.

Palavras-chave: Método de captura-recaptura, Estimativas de Bayes, Dados particionados.

Abstract

In this work, a bayesian methodology is shown to estimate the size of a diabethic-suffering population through lists containing information data of patients. The applied methodology is analogous of capture-recaptures in animal population. We assume correct the registers of relative information to the patients as well as we take in account correct and incorrect registers of the information. In case the supposed registers are correct, the methodology is developed for two or more lists and the Bayes estimate is determined for the size of a population. In a second model, the occurrency of correct and incorrect registers are considered, presenting a two-stage estimation method for the model parameters using two lists. For both models there are results with simulated and real examples.

keywords: Capture-recapture method, Bayes estimate, Partition data.

Sumário

Capítulo 1

Introdução

A estimação do tamanho de uma população de diabéticos é um problema que se enquadra na metodologia de captura-recaptura. Embora tal metodologia originalmente fosse aplicada na estimação do tamanho de uma população animal, alguns pesquisadores aplicaram-na para estimar a prevalência de doenças não transmissíveis. A idéia é considerar duas ou mais listas de indivíduos de uma população de doentes (diabéticos) como amostras selecionadas da população. Então, cada indivíduo presente em uma lista é um indivíduo capturado na amostra correspondente e desse modo podemos aplicar o método para as diversas listas, onde as informações individuais são interpretadas como marcas. Pesquisadores tais como Sekar e Deming (1949), Wittes e Sidel (1968), Wittes (1974), Fienberg (1972,1999), Seber et al (2000), Lee et al (2001), Lee (2002) e Micheletti (2003) aplicaram esta metodologia utilizando duas ou mais listas.

Em sua dissertação de mestrado Micheletti (2003) estimou, sob o enfoque estatístico clássico, o número de diabéticos de uma população utilizando duas ou mais listas de indivíduos. Especificamente, considerando registros corretos e incorretos dos dados cadastrais dos indivíduos pertencentes as listas, ela determinou estimativas de máxima verossimilhança e máxima verossimilhança condicional do tamanho populacional, bem como intervalos de confiança.

Nosso objetivo neste trabalho é apresentar uma solução bayesiana para este problema. No Capítulo 2, com a suposição de que não há erros no preenchimento dos dados cadastrais individuais e utilizando duas listas, desenvolvemos uma metodologia bayesiana. No Capítulo 3, desenvolvemos a mesma metodologia generalizando-a para três ou mais listas. No Capítulo 4, implementamos os modelos dos Capítulos 2 e 3 e apresentamos os resumos *a posteriori* para exemplos com dados simulados e reais. No Capítulo 5, desenvolvemos a mesma metodologia do Capítulo 2, considerando possíveis erros no preenchimento dos dados cadastrais individuais e damos exemplos com dados simulados e reais. No Capítulo 6, apresentamos uma proposta para

uma pesquisa futura sobre o assunto.

Capítulo 2

Estimação bayesiana do tamanho de uma população de diabéticos através de duas listas

Neste capítulo apresentamos uma solução bayesiana para o problema da estimação do tamanho de uma população de diabéticos, utilizando duas listas de indivíduos da população. Definimos o modelo estatístico adequado para esta situação e determinamos as estimativas de Bayes para o tamanho populacional, bem como para as probabilidades dos indivíduos pertencerem às listas.

Atribuímos as distribuições *a priori* uniforme nos inteiros positivos, de Jeffreys e de Poisson truncada em zero para o tamanho populacional, e a distribuição Beta, com parâmetros conhecidos para a probabilidade de um indivíduo pertencer a uma lista.

2.1 Modelo estatístico e a função de verossimilhança

Denotamos por N (N desconhecido), o tamanho da população e supomos que não haja erros no preenchimento dos dados cadastrais dos indivíduos.

A cada indivíduo i da população associamos um vetor aleatório bidimensional, $\mathbf{X}_i = (X_{i1}, X_{i2})$ definido por

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

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

Supomos que cada indivíduo pertence ou não a uma lista, independentemente dos demais indivíduos e da outra lista e que um indivíduo pertence às duas listas quando seus dados cadastrais forem idênticos. Denotamos por θ_{ij}, $0 < \theta_{ij} < 1$, a probabilidade de que o indivíduo i pertença a lista j, $i = 1, 2, ..., N$; $j = 1, 2$, e por $\boldsymbol{\theta} = (\theta_{i1}, \theta_{i2})$.

Assim as variavéis aleatórias X_{ij} são independentes com

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

$i = 1, 2, ..., N$; $j = 1, 2$, e os vetores aleatórios $\mathbf{X}_i = (X_{i1}, X_{i2})$, $i = 1, ..., N$, são independentes e assumem valores no conjunto $\Omega = \{\boldsymbol{\omega}_r = (\omega_{r1}, \omega_{r2}) : \omega_{rj} = 0, 1; r = 1, ..., 4; j = 1, 2\}$.

Para cada indivíduo i da população, existem $2^2 = 4$ possíveis trajetórias (histórias) representadas por vetores de acordo com a seguinte enumeração, por exemplo:

$\boldsymbol{\omega}_1 = (1, 0)$ se o indivíduo pertence somente à lista 1;

$\boldsymbol{\omega}_2 = (0, 1)$ se o indivíduo pertence somente à lista 2;

$\boldsymbol{\omega}_3 = (1, 1)$ se o indivíduo pertence às listas 1 e 2;

$\boldsymbol{\omega}_4 = (0, 0)$ se o indivíduo não pertence a nenhuma lista.

Notamos que $\boldsymbol{\omega}_4$ é a trajetória dos indivíduos não observados.

Supomos agora que todos indivíduos possuam a mesma probabilidade de pertencer a uma certa lista, isto é, $\theta_{ij} = \theta_j$, para $i = 1, 2, ..., N$ e $j = 1, 2$.

Logo, $\boldsymbol{\theta} = (\theta_1, \theta_2)$ e

$$\begin{aligned} p_r(\boldsymbol{\theta}) &= P(\mathbf{X}_i = \boldsymbol{\omega}_r|N, \boldsymbol{\theta}) = P[(X_{i1}, X_{i2}) = (\omega_{r1}, \omega_{r2})|N, \boldsymbol{\theta}] \\ &= \prod_{j=1}^{2} P(X_{ij} = \omega_{rj}|N, \boldsymbol{\theta}) = \prod_{j=1}^{2} \theta_j^{\omega_{rj}}(1 - \theta_j)^{1-\omega_{rj}}, \end{aligned}$$

$i = 1, 2, ..., N$ e $r = 1, ..., 4$, isto é,

$$\begin{aligned} p_1(\boldsymbol{\theta}) &= \theta_1(1 - \theta_2), \\ p_2(\boldsymbol{\theta}) &= (1 - \theta_1)\theta_2, \\ p_3(\boldsymbol{\theta}) &= \theta_1\theta_2, \\ p_4(\boldsymbol{\theta}) &= (1 - \theta_1)(1 - \theta_2). \end{aligned}$$

Sejam

$n_{(j)}$ o número de indivíduos observados somente na lista j;

n_j o número de indivíduos observados na lista j, $j = 1, 2$;

n_{12} o número de indivíduos observados em ambas as listas;

n o número de indivíduos distintos observados nas duas listas.

Logo, $N-n$ é o número de indivíduos não observados em nenhuma lista, $n = n_{(1)}+n_{(2)}+n_{12} = n_1+n_2-n_{12}$ e $(n_{(1)}, n_{(2)}, n_{12}, N-n)$, dados N e $\boldsymbol{\theta}$, tem distribuição multinomial com parâmetros N e $(p_1(\boldsymbol{\theta}), p_2(\boldsymbol{\theta}), p_3(\boldsymbol{\theta}), p_4(\boldsymbol{\theta}))$, o que implica

$$
\begin{aligned}
& P(n_{(1)}, n_{(2)}, n_{12}, N-n | N, \boldsymbol{\theta}) \\
= & \frac{N!}{n_{(1)}!n_{(2)}!n_{12}!(N-n)!}[p_1(\boldsymbol{\theta})]^{n_{(1)}}[p_2(\boldsymbol{\theta})]^{n_{(2)}}[p_3(\boldsymbol{\theta})]^{n_{12}}[p_4(\boldsymbol{\theta})]^{N-n} \\
= & \frac{N!}{n_{(1)}!n_{(2)}!n_{12}!(N-n)!}[\theta_1(1-\theta_2)]^{n_{(1)}}[(1-\theta_1)\theta_2)]^{n_{(2)}}[\theta_1\theta_2]^{n_{12}}[(1-\theta_1)(1-\theta_2)]^{N-n} \\
= & \frac{N!}{n_{(1)}!n_{(2)}!n_{12}!(N-n)!}\theta_1^{n_{(1)}+n_{12}}(1-\theta_1)^{n_{(2)}+N-n}\theta_2^{n_{(2)}+n_{12}}(1-\theta_2)^{n_{(1)}+N-n} \\
= & \frac{N!}{n_{(1)}!n_{(2)}!n_{12}!(N-n)!}\theta_1^{n_1}(1-\theta_1)^{N-n_1}\theta_2^{n_2}(1-\theta_2)^{N-n_2} \\
\propto & \frac{N!}{(N-n)!}\theta_1^{n_1}(1-\theta_1)^{N-n_1}\theta_2^{n_2}(1-\theta_2)^{N-n_2}. \qquad (2.1)
\end{aligned}
$$

De (2.1) segue que a função de verossimilhança é tal que

$$
L(N, \boldsymbol{\theta} | n_1, n_2, n) \propto \frac{N!}{(N-n)!}\prod_{j=1}^{2}\theta_j^{n_j}(1-\theta_j)^{N-n_j}, \qquad (2.2)
$$

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

Denifida a função de verossimilhança apresentamos na próxima seção o modelo bayesiano para estimação do tamanho de uma população de diabéticos.

2.2 Modelo bayesiano

Sabemos que quando n_{12} é igual a zero a estimativa de máxima verossimilhança de N não existe, e quando n_{12} é aproximadamente igual a zero, em geral, elas superestimam o verdadeiro valor do parâmetro N Micheletti (2003).

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. Com efeito, no Capítulo 4 apresentamos exemplos com dados simulados, cujos resultados produzem estimativas mais próximas do verdadeiro valor do parâmetro do que as de máxima verossimilhança.

As informações prévias são expressas através de distribuições *a priori* atribuidas aos parâmetros do modelo N e $\boldsymbol{\theta}$.

Supomos *a priori* θ_1 e θ_2 independentes, θ_j com distribuição Beta, π_j, com parâmetros α_j e β_j conhecidos ($\alpha_j > 0$ e $\beta_j > 0$), $j = 1, 2$, N com distribuição $\pi(N)$ nos inteiros estritamente positivos *e* N e $\boldsymbol{\theta}$ independentes. Com isso obtemos a distribuição *a priori* conjunta de N, θ_1 e θ_2 dada por

$$\begin{aligned}
\pi(N, \boldsymbol{\theta}) &= \pi(N)\pi_1(\theta_1)\pi_2(\theta_2) \\
&= \pi(N)\prod_{j=1}^{2}\frac{\Gamma(\alpha_j+\beta_j)}{\Gamma(\alpha_j)\Gamma(\beta_j)}\theta_j^{\alpha_j-1}(1-\theta_j)^{\beta_j-1} \\
&\propto \pi(N)\prod_{j=1}^{2}\theta_j^{\alpha_j-1}(1-\theta_j)^{\beta_j-1}, \qquad (2.3)
\end{aligned}$$

$N = 1, 2, ...,$ e $0 < \theta_j < 1, j = 1, 2$. Então, de (2.2) e (2.3), segue que a distribuição *a posteriori* conjunta de N e $\boldsymbol{\theta}$ é tal que

$$\begin{aligned}
\pi(N, \boldsymbol{\theta}|n_1, n_2, n) &\propto L(N, \boldsymbol{\theta}|n_1, n_2, n)\pi(N, \boldsymbol{\theta}) \\
&\propto \frac{N!}{(N-n)!}\prod_{j=1}^{2}\theta_j^{n_j}(1-\theta_j)^{N-n_j}\pi(N)\prod_{j=1}^{2}\theta_j^{\alpha_j-1}(1-\theta_j)^{\beta_j-1} \\
&= \pi(N)\frac{N!}{(N-n)!}\prod_{j=1}^{2}\theta_j^{n_j+\alpha_j-1}(1-\theta_j)^{N+\beta_j-n_j-1}, \qquad (2.4)
\end{aligned}$$

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

2.3 Distribuição *a priori* uniforme e de Jeffreys para o tamanho populacional

Nesta seção supomos que a distribuição *a priori* para N é da forma $\pi(N) = \frac{1}{N^r}$, para $N = 1, 2, ...$ e $r = 0, 1$. Observamos que, para $r = 0, \pi(N) = 1, N = 1, 2, ...$, isto é, $\pi(N)$ é a distribuição *a priori* uniforme nos inteiros positivos e, para $r = 1, \pi(N) = \frac{1}{N}, N = 1, 2, ...$, ou seja, $\pi(N)$ é a distribuição *a priori* de Jeffreys. Então, segue de (2.4), que a distribuição *a posteriori* conjunta de N e $\boldsymbol{\theta}$ é tal que

$$\pi(N, \boldsymbol{\theta}|n_1, n_2, n) \propto \frac{N!}{(N-n)!}\frac{1}{N^r}\prod_{j=1}^{2}\theta_j^{n_j+\alpha_j-1}(1-\theta_j)^{N+\beta_j-n_j-1}, \tag{2.5}$$

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

Notamos que as duas distribuições *a priori* utilizadas acima são impróprias, pois suas integrais não são iguais a 1.

O próximo teorema estabelece sob que condição a distribuição *a posteriori* conjunta (2.5) de N e $\boldsymbol{\theta}$ existe.

Teorema 2.1 Suponhamos que a distribuição *a priori* para N seja $\pi(N) = \frac{1}{N^r}, N = 1, 2, ..., r = 0, 1$. Se $\sum_{j=1}^{2} n_j - n + \sum_{j=1}^{2}\alpha_j + r > 1$, então a distribuição *a posteriori*, $\pi(N, \boldsymbol{\theta}|n_1, n_2, n)$, existe.

Prova

A constante normalizadora , C, de (2.5) é tal que

$$\begin{aligned}
C^{-1} &= \sum_{N=n}^{\infty}\int_0^1\int_0^1 \frac{N!}{(N-n)!}\frac{1}{N^r}\prod_{j=1}^{2}\theta_j^{n_j+\alpha_j-1}(1-\theta_j)^{N+\beta_j-n_j-1}d\theta_1 d\theta_2 \\
&= \sum_{N=n}^{\infty}\frac{\Gamma(N+1)}{\Gamma(N-n+1)}\frac{1}{N^r}\int_0^1 \theta_1^{n_1+\alpha_1-1}(1-\theta_1)^{N+\beta_1-n_1-1}d\theta_1 \times \\
&\quad \times \int_0^1 \theta_2^{n_2+\alpha_2-1}(1-\theta_2)^{N+\beta_2-n_2-1}d\theta_2 \\
&= \sum_{N=n}^{\infty}\frac{\Gamma(N+1)}{\Gamma(N-n+1)}\frac{1}{N^r}\frac{\Gamma(n_1+\alpha_1)\Gamma(N-n_1+\beta_1)}{\Gamma(N+\alpha_1+\beta_1)} \times \\
&\quad \times \frac{\Gamma(n_2+\alpha_2)\Gamma(N-n_2+\beta_2)}{\Gamma(N+\alpha_2+\beta_2)}
\end{aligned}$$

$$C^{-1} = \sum_{N=n}^{\infty} \frac{\Gamma(N+1)}{\Gamma(N-n+1)} \frac{1}{N^r} \prod_{j=1}^{2} \frac{\Gamma(N-n_j+\beta_j)\Gamma(n_j+\alpha_j)}{\Gamma(N+\alpha_j+\beta_j)}$$

$$\propto S_n, \text{ onde}$$

$$S_n = \sum_{N=n}^{\infty} \frac{\Gamma(N+1)}{\Gamma(N-n+1)} \frac{1}{N^r} \prod_{j=1}^{2} \frac{\Gamma(N-n_j+\beta_j)}{\Gamma(N+\alpha_j+\beta_j)}, \tag{2.6}$$

Vamos mostrar que C^{-1} é finita.

Inicialmente mostraremos que

(a) $\frac{\Gamma(N+1)}{\Gamma(N-n+1)} \leq O(N^n)(N \to \infty)$

e

(b)$\frac{\Gamma(N-n_j+\beta_j)}{\Gamma(N+\alpha_j+\beta_j)} \leq O(N^{-(n_j+\alpha_j)})(N \to \infty), j = 1, 2.$

(a) Da relação

$$(2\pi)^{\frac{1}{2}} x^{x-\frac{1}{2}} \exp\{-x\} \leq \Gamma(x) \leq (2\pi)^{\frac{1}{2}} x^{x-\frac{1}{2}} \exp\{-x + \frac{1}{12x}\}, \tag{2.7}$$

para todo x real, segue que

$$\Gamma(N+1) \leq (2\pi)^{\frac{1}{2}} (N+1)^{N+\frac{1}{2}} \exp\{-(N+1) + \frac{1}{12(N+1)}\}$$

e

$$\frac{1}{\Gamma(N-n+1)} \leq \frac{1}{(2\pi)^{\frac{1}{2}} (N-n+1)^{N-n+\frac{1}{2}} \exp\{-(N-n+1)\}},$$

o que implica

$$\begin{aligned} \frac{\Gamma(N+1)}{\Gamma(N-n+1)} &\leq \frac{(2\pi)^{\frac{1}{2}} (N+1)^{N+\frac{1}{2}} \exp\{-(N+1) + \frac{1}{12(N+1)}\}}{(2\pi)^{\frac{1}{2}} (N-n+1)^{N-n+\frac{1}{2}} \exp\{-(N-n+1)\}} \\ &= \left(\frac{N+1}{N-n+1}\right)^{(N+\frac{1}{2})} (N-n+1)^n \exp\{-n + \frac{1}{12(N+1)}\}, \ N \geq n. \end{aligned} \tag{2.8}$$

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

$(N-n+1)^n = O(N^n)(N \to \infty)$ e

$\exp\{-n + \frac{1}{12(N+1)}\} = O(1)(N \to \infty)$,

segue que

$$\frac{\Gamma(N+1)}{\Gamma(N-n+1)} \leq O(1)O(N^n)O(1) = O(N^n)(N \to \infty). \tag{2.9}$$

(b) Da relação (2.7), segue que

$$\Gamma(N-n_j+\beta_j) \leq (2\pi)^{\frac{1}{2}}(N-n_j+\beta_j)^{N-n_j+\beta_j-\frac{1}{2}} \exp\{-(N-n_j+\beta_j) + \frac{1}{12(N-n_j+\beta_j)}\}$$

e

$$\frac{1}{\Gamma(N+\alpha_j+\beta_j)} \leq \frac{1}{(2\pi)^{\frac{1}{2}}(N+\alpha_j+\beta_j)^{N+\alpha_j+\beta_j-\frac{1}{2}} \exp\{-(N+\alpha_j+\beta_j)\}},$$

o que implica

$$\begin{aligned}
\frac{\Gamma(N-n_j+\beta_j)}{\Gamma(N+\alpha_j+\beta_j)} &\leq \frac{(2\pi)^{\frac{1}{2}}(N-n_j+\beta_j)^{N-n_j+\beta_j-\frac{1}{2}} \exp\{-(N-n_j+\beta_j) + \frac{1}{12(N-n_j+\beta_j)}\}}{(2\pi)^{\frac{1}{2}}(N+\alpha_j+\beta_j)^{N+\alpha_j+\beta_j-\frac{1}{2}} \exp\{-(N+\alpha_j+\beta_j)\}} \\
&= \frac{(N-n_j+\beta_j)^{N-n_j+\beta_j-\frac{1}{2}} \exp\{n_j+\alpha_j + \frac{1}{12(N-n_j+\beta_j)}\}}{(N+\alpha_j+\beta_j)^{N+\alpha_j+\beta_j-\frac{1}{2}}} \\
&= (N-n_j+\beta_j)^{N-n_j+\beta_j-\frac{1}{2}}(N+\alpha_j+\beta_j)^{-\left(N+\alpha_j+\beta_j-\frac{1}{2}\right)} \times \\
&\quad \times \exp\{n_j+\alpha_j + \frac{1}{12(N-n_j+\beta_j)}\}
\end{aligned} \tag{2.10}$$

Como $(N-n_j+\beta_j)^{N-n_j+\beta_j-\frac{1}{2}} = O(N^{N-n_j+\beta_j-\frac{1}{2}})(N \to \infty)$,

$(N+\alpha_j+\beta_j)^{-\left(N+\alpha_j+\beta_j-\frac{1}{2}\right)} = O(N^{-(N+\alpha_j+\beta_j-\frac{1}{2})})(N \to \infty)$ e

$\exp\{n_j+\alpha_j + \frac{1}{12(N-n_j+\beta_j)}\} = O(1)(N \to \infty)$,

segue que

$$\begin{aligned}\frac{\Gamma(N-n_j+\beta_j)}{\Gamma(N+\alpha_j+\beta_j)} &\leq O\left(N^{N-n_j+\beta_j-\frac{1}{2}}\right)O(1)O\left(N^{-\left(N+\alpha_j+\beta_j-\frac{1}{2}\right)}\right)\\ &= O\left(N^{-(n_j+\alpha_j)}\right)(N\to\infty),\ j=1,2. \qquad (2.11)\end{aligned}$$

Logo, de (2.6) segue que

$$\begin{aligned}S_n &\leq \sum_{N=n}^{\infty}O(N^n)O(N^{-r})\prod_{j=1}^{2}O(N^{-(n_j+\alpha_j)})\\ &= \sum_{N=n}^{\infty}O(N^{n-r})O\left(N^{-\left(\sum_{j=1}^{2}(n_j+\alpha_j)\right)}\right)\\ &= \sum_{N=n}^{\infty}O\left(N^{-\left(\sum_{j=1}^{2}n_j-n+\sum_{j=1}^{2}\alpha_j+r\right)}\right)(N\to\infty), \qquad (2.12)\end{aligned}$$

isto é, $S_n \leq \sum_{N=n}^{\infty} a_N$, onde $a_N = O\left(N^{-\left(\sum_{j=1}^{2}n_j-n+\sum_{j=1}^{2}\alpha_j+r\right)}\right)(N\to\infty)$.

Então, existe uma constante $K>0$ e um número inteiro positivo N_0, $N_0>n$, tal que $a_N \leq KN^{-\left(\sum_{j=1}^{2}n_j-n+\sum_{j=1}^{2}\alpha_j+r\right)}$, para todo $N>N_0$, e

$$\begin{aligned}S_n &\leq \sum_{N=n}^{\infty}a_N = \sum_{N=n}^{N_o}a_N + \sum_{N=N_o+1}^{\infty}a_N\\ &\leq \sum_{N=n}^{N_o}a_N + \sum_{N=N_o+1}^{\infty}KN^{-\left(\sum_{j=1}^{2}n_j-n+\sum_{j=1}^{2}\alpha_j+r\right)}\\ &= \sum_{N=n}^{N_o}a_N + K\sum_{N=N_o+1}^{\infty}N^{-\left(\sum_{j=1}^{2}n_j-n+\sum_{j=1}^{2}\alpha_j+r\right)} < \infty, \qquad (2.13)\end{aligned}$$

pois, por hipótese, $\sum_{N=N_o+1}^{\infty}N^{-\left(\sum_{j=1}^{2}n_j-n+\sum_{j=1}^{2}\alpha_j+r\right)} < \infty$, o que prova o teorema.■

Doravante supomos que a hipótese $\sum_{j=1}^{2}n_j-n+\sum_{j=1}^{2}\alpha_j+r>1$ do teorema 2.1 esteja satisfeita.

Para $r = 0$ ou *a priori* $\pi(N)$ uniforme nos inteiros positivos, temos de (2.5) que a distribuição *a posteriori* conjunta de N e $\boldsymbol{\theta}$ é tal que

$$\pi(N, \boldsymbol{\theta}|n_1, n_2, n) \propto \frac{N!}{(N-n)!}\prod_{j=1}^{2}\theta_j^{n_j+\alpha_j-1}(1-\theta_j)^{N+\beta_j-n_j-1}, \tag{2.14}$$

$N \geq n$ e $0 < \theta_j < 1, j = 1, 2$. Então, a distribuição condicional de N, dados $\boldsymbol{\theta}, n_1, n_2$ e n, é dada por

$$\pi(N|\boldsymbol{\theta}, n_1, n_2, n) = C_1\binom{N}{n}\left\{\prod_{j=1}^{2}(1-\theta_j)\right\}^N, \tag{2.15}$$

$N \geq n$, onde

$$\begin{aligned}
C_1^{-1} &= \sum_{N\geq n}^{\infty}\binom{N}{n}\left\{\prod_{j=1}^{2}(1-\theta_j)\right\}^N \\
&= \sum_{s=0}^{\infty}\binom{s+n}{n}\left\{\prod_{j=1}^{2}(1-\theta_j)\right\}^{s+n} \\
&= \left\{\prod_{j=1}^{2}(1-\theta_j)\right\}^n\sum_{s=0}^{\infty}\binom{-n-1}{s}\left\{-\prod_{j=1}^{2}(1-\theta_j)\right\}^s \\
&= \left\{\prod_{j=1}^{2}(1-\theta_j)\right\}^n\left\{1-\prod_{j=1}^{2}(1-\theta_j)\right\}^{-n-1}.
\end{aligned}$$

Logo,

$$\pi(N|\boldsymbol{\theta}, n_1, n_2, n) = \binom{N}{n}\left\{1-\prod_{j=1}^{2}(1-\theta_j)\right\}^{n+1}\left\{\prod_{j=1}^{2}(1-\theta_j)\right\}^{N-n}, \tag{2.16}$$

$N \geq n$.

Na realidade a distribuição condicional de N, dados $\boldsymbol{\theta}, n_1, n_2$ e n, é igual a distribuição de uma variável aleatória $n+Y$, onde Y tem distribuição binomial negativa com parâme-tros $n+1$ e $1-\prod_{j=1}^{2}(1-\theta_j)$.

De de fato, se Y tiver distribuição binomial negativa com parâmetros $n+1$ e $1-\prod_{j=1}^{2}(1-\theta_j)$,

então

$$\begin{aligned} P(n+Y &= y|\boldsymbol{\theta}, n_1, n_2, n) = P(Y = y-n|\boldsymbol{\theta}, n_1, n_2, n) \\ &= \binom{y}{n}\left(1-\prod_{j=1}^{2}(1-\theta_j)\right)^{n+1}\left(\prod_{j=1}^{2}(1-\theta_j)\right)^{y-n}, \end{aligned}$$

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

Por outro lado, de (2.14) segue que a distribuição condicional de θ_j, dados $N, \theta_m, m \neq j$, n_1, n_2 e n, é tal que

$$\pi(\theta_j|N, \theta_m, m \neq j, n_1, n_2, n) \propto \theta_j^{n_j+\alpha_j-1}(1-\theta_j)^{N+\beta_j-n_j-1}, \tag{2.17}$$

$0 < \theta_j < 1$, $j = 1, 2$. Isto é, a distribuição condicional de θ_j, dados $N, \theta_m, m \neq j, n_1, n_2$ e n, é Beta com parâmetros $n_j + \alpha_j$ e $N + \beta_j - n_j, j = 1, 2$.

Para $r = 1$ ou *a priori* $\pi(N)$ de Jeffreys temos, de (2.5), que a distribuição *a posteriori* conjunta de N e $\boldsymbol{\theta}$ é tal que

$$\pi(N, \boldsymbol{\theta}|n_1, n_2, n) \propto \frac{N!}{(N-n)!N}\prod_{j=1}^{2}\theta_j^{n_j+\alpha_j-1}(1-\theta_j)^{N+\beta_j-n_j-1}, \tag{2.18}$$

$N \geq n$ e $0 < \theta_j < 1, j = 1, 2$. Então, a distribuição condicional de N, dados $\boldsymbol{\theta}, n_1, n_2$ e n, é dada por

$$\pi(N|\boldsymbol{\theta}, n_1, n_2, n) = C_2\frac{1}{N}\binom{N}{n}\left\{\prod_{j=1}^{2}(1-\theta_j)\right\}^{N}, \tag{2.19}$$

$N \geq n$, onde

$$
\begin{aligned}
C_2^{-1} &= \sum_{N\geq n}^{\infty} \frac{1}{N}\binom{N}{n}\left\{\prod_{j=1}^{2}(1-\theta_j)\right\}^N \\
&= \frac{1}{n}\sum_{N\geq n}^{\infty}\binom{N-1}{n-1}\left\{\prod_{j=1}^{2}(1-\theta_j)\right\}^N \\
&= \frac{1}{n}\sum_{s=0}^{\infty}\binom{s+n-1}{n-1}\left\{\prod_{j=1}^{2}(1-\theta_j)\right\}^{s+n} \\
&= \frac{1}{n}\left\{\prod_{j=1}^{2}(1-\theta_j)\right\}^n \sum_{s=0}^{\infty}\binom{-n}{s}\left\{-\prod_{j=1}^{2}(1-\theta_j)\right\}^s \\
&= \frac{1}{n}\left\{\prod_{j=1}^{2}(1-\theta_j)\right\}^n \left\{1-\prod_{j=1}^{2}(1-\theta_j)\right\}^{-n}.
\end{aligned}
$$

Logo,

$$
\pi(N|\boldsymbol{\theta}, n_1, n_2, n) = \binom{N-1}{n-1}\left\{1-\prod_{j=1}^{2}(1-\theta_j)\right\}^n \left\{\prod_{j=1}^{2}(1-\theta_j)\right\}^{N-n}, \tag{2.20}
$$

$N \geq n$.

Na realidade a distribuição condicional de N, dados $\boldsymbol{\theta}, n_1, n_2$ e n, é igual a distribuição de uma variável aleatória $n+Z$, onde Z tem distribuição binomial negativa com parâmetros n e $1-\prod_{j=1}^{2}(1-\theta_j)$.

De fato, se Z tiver distribuição binomial negativa com parâmetros n e $1-\prod_{j=1}^{2}(1-\theta_j)$, então

$$
\begin{aligned}
P(n+Z &= z|\boldsymbol{\theta}, n_1, n_2, n) = P(Z = z-n|\boldsymbol{\theta}, n_1, n_2, n) \\
&= \binom{z-1}{n-1}\left(1-\prod_{j=1}^{2}(1-\theta_j)\right)^n \left(\prod_{j=1}^{2}(1-\theta_j)\right)^{z-n},
\end{aligned}
$$

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

Por outro lado, de (2.18) segue que a distribuição condicional de θ_j, dados $N, \theta_m, m \neq j$,n_1, n_2

e n, é tal que

$$\pi(\theta_j|N,\theta_m,m\neq j,n_1,n_2,n)\propto\theta_j^{n_j+\alpha_j-1}(1-\theta_j)^{N+\beta_j-n_j-1}, \tag{2.21}$$

$0<\theta_j<1$, $j=1,2$. Isto é, a distribuição condicional de θ_j, dados $N,\theta_m,m\neq j,n_1,n_2$ e n, é Beta com parâmetros $n_j+\alpha_j$ e $N+\beta_j-n_j, j=1,2$.

2.4 Distribuição *a priori* de Poisson para o tamanho populacional

Nesta seção supomos que N tem distribuição *a priori* de Poisson truncada em zero, ou seja, $\pi(N)=\frac{e^{-\lambda}\lambda^N}{N!(1-e^{-\lambda})}$, com $\lambda>0$, conhecido e $N=1,2,....$ Assim sendo, temos de (2.4), que a distribuição *a posteriori* conjunta de N e $\boldsymbol{\theta}$ é tal que

$$\pi(N,\boldsymbol{\theta}|n_1,n_2,n)\propto\frac{\lambda^N}{(N-n)!}\prod_{j=1}^{2}\theta_j^{n_j+\alpha_j-1}(1-\theta_j)^{N+\beta_j-n_j-1}, \tag{2.22}$$

$N\geq n$ e $0<\theta_j<1, j=1,2$. Ou seja, a distribuição condicional de N, dados $\boldsymbol{\theta},n_1,n_2$ e n, é dada por

$$\pi(N|\boldsymbol{\theta},n_1,n_2,n)=C_3\frac{\lambda^N}{(N-n)!}\left\{\prod_{j=1}^{2}(1-\theta_j)\right\}^N, \tag{2.23}$$

$N\geq n$, onde

$$\begin{aligned}
C_3^{-1} &= \sum_{N\geq n}^{\infty}\frac{1}{(N-n)!}\left\{\lambda\prod_{j=1}^{2}(1-\theta_j)\right\}^N \\
&= \sum_{s=0}^{\infty}\frac{1}{(s)!}\left\{\lambda\prod_{j=1}^{2}(1-\theta_j)\right\}^{s+n} \\
&= \left\{\lambda\prod_{j=1}^{2}(1-\theta_j)\right\}^n\exp\left\{\lambda\prod_{j=1}^{2}(1-\theta_j)\right\}.
\end{aligned}$$

Logo,

$$\pi(N|\boldsymbol{\theta},n_1,n_2,n)=\frac{1}{(N-n)!}\exp\left\{-\lambda\prod_{j=1}^{2}(1-\theta_j)\right\}\left\{\lambda\prod_{j=1}^{2}(1-\theta_j)\right\}^{N-n}, \tag{2.24}$$

$N\geq n$.

A distribuição condicional de N, dados $\boldsymbol{\theta}, n_1, n_2$ e n, é igual a distribuição de uma variável aleatória $n + W$, onde W tem distribuição de Poisson com parâmetro $\lambda\prod_{j=1}^{2}(1-\theta_j)$.

De fato, se W tiver distribuição de Poisson com parâmetro $\lambda\prod_{j=1}^{2}(1-\theta_j)$, então

$$\begin{aligned} P(n+W &= w|\boldsymbol{\theta}, n_1, n_2, n) = P(W = w-n|\boldsymbol{\theta}, n_1, n_2, n) \\ &= \frac{1}{(w-n)!}\exp\left\{-\lambda\prod_{j=1}^{2}(1-\theta_j)\right\}\left\{\lambda\prod_{j=1}^{2}(1-\theta_j)\right\}^{w-n}, \end{aligned}$$

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

Por outro lado, de (2.22) segue que a distribuição condicional de θ_j, dados N, $\theta_m, m \neq j$, n_1, n_2 e n, é tal que

$$\pi(\theta_j|N, \theta_m, m \neq j, n_1, n_2, n) \propto \theta_j^{n_j+\alpha_j-1}(1-\theta_j)^{N+\beta_j-n_j-1}, \tag{2.25}$$

$0 < \theta_j < 1$, $j = 1, 2$. Isto é, a distribuição condicional de θ_j, dados N, θ_m, $m \neq j, n_1, n_2$ e n, é Beta com parâmetros $n_j + \alpha_j$ e $N + \beta_j - n_j, j = 1, 2$.

2.4.1 Distribuição *a priori* hierárquica de Poisson para o tamanho populacional

Nesta subseção supomos que N tem distribuição *a priori* de Poisson com média λ truncada em zero e atribuímos ao hiperparâmetro λ a distribuição Gama com parâmetros a e b conhecidos, $a > 0$ e $b > 0$. Supomos *a priori* θ_1 e θ_2 independentes, θ_j com distribuição Beta, π_j, com parâmetros α_j e β_j conhecidos ($\alpha_j > 0$ e $\beta_j > 0$) e N, λ e $\boldsymbol{\theta}$ independentes. Com isso obtemos a distribuição *a priori* conjunta de N, $\lambda, \boldsymbol{\theta}$ dada por

$$\begin{aligned} \pi(N, \lambda, \boldsymbol{\theta}) &= \pi(N|\lambda)\pi(\lambda)\pi_1(\theta_1)\pi_2(\theta_2) \\ &= \frac{e^{-\lambda}\lambda^N}{N!(1-e^{-\lambda})}\frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda}\prod_{j=1}^{2}\frac{\Gamma(\alpha_j+\beta_j)}{\Gamma(\alpha_j)\Gamma(\beta_j)}\theta_j^{\alpha_j-1}(1-\theta_j)^{\beta_j-1} \\ &\propto \frac{e^{-\lambda(1+b)}\lambda^{N+(a-1)}}{N!(1-e^{-\lambda})}\prod_{j=1}^{2}\theta_j^{\alpha_j-1}(1-\theta_j)^{\beta_j-1}, \end{aligned} \tag{2.26}$$

$N = 1, 2, ...; \lambda > 0$ e $0 < \theta_j < 1, j = 1, 2$.

Assim sendo, temos de (2.2) e (2.26), que a distribuição *a posteriori* conjunta de N, λ e $\boldsymbol{\theta}$ é tal que

$$\pi(N, \lambda, \boldsymbol{\theta}|n_1, n_2, n_{12}) \propto \frac{e^{-\lambda(1+b)}\lambda^{N+(a-1)}}{(1-e^{-\lambda})(N-n)!}\prod_{j=1}^{2}\theta_j^{n_j+\alpha_j-1}(1-\theta_j)^{N+\beta_j-n_j-1}, \tag{2.27}$$

$N \geq n; \lambda > 0$ e $0 < \theta_j < 1, j = 1, 2$, ou seja, como na seção 2.4 a distribuição condicional de N, dados $\boldsymbol{\theta}$,n_1, n_2 e n_{12}, é dada por

$$\pi(N|\boldsymbol{\theta}, n_1, n_2, n_{12}) = \frac{1}{(N-n)!}\exp\left\{-\lambda\prod_{j=1}^{2}(1-\theta_j)\right\}\left\{\lambda\prod_{j=1}^{2}(1-\theta_j)\right\}^{N-n}, \tag{2.28}$$

$N \geq n$. Isto é, a distribuição condicional de N, dados $\boldsymbol{\theta}$, n_1, n_2 e n_{12}, é igual a distribuição de uma variável aleatória $n + T$, onde T tem distribuição de Poisson com parâmetro $\lambda\prod_{j=1}^{2}(1-\theta_j)$.

Segue de (2.27) que a distribuição condicional de λ, dados $N, \boldsymbol{\theta}$, n_1, n_2 e n_{12}, é tal que

$$\begin{aligned}\pi(\lambda|N, \boldsymbol{\theta}, n_1, n_2, n_{12}) &= \frac{1}{(1-e^{-\lambda})}\lambda^{N+(a-1)}e^{-\lambda(1+b)} \\ &\propto \frac{(1+b)^{N+a}}{\Gamma(N+a)(1-e^{-\lambda})}\lambda^{N+(a-1)}e^{-\lambda(1+b)},\end{aligned} \tag{2.29}$$

$\lambda > 0$.

Por outro lado, de (2.27) segue que a distribuição condicional de θ_j, dados N, λ, θ_m, $m \neq j$,n_1, n_2 e n_{12}, é tal que

$$\pi(\theta_j|N, \lambda, \theta_m, m \neq j, n_1, n_2, n) \propto \theta_j^{n_j+\alpha_j-1}(1-\theta_j)^{N+\beta_j-n_j-1}, \tag{2.30}$$

$0 < \theta_j < 1$, $j = 1, 2$. Isto é, a distribuição condicional de θ_j, dados $N, \lambda, \theta_m, m \neq j$,n_1, n_2 e n_{12} é Beta com parâmetros $n_j + \alpha_j$ e $N + \beta_j - n_j, j = 1, 2$.

No Capítulo 4 verificamos o comportamento desse modelo, através de um estudo de simulação, e comparamos os resultados com os obtidos usando o modelo Poisson com hiperparâmetro conhecido.

2.5 Distribuições *a priori* não informativas e de referência para as probabilidades de um indivíduo pertencer às listas

Nesta seção vamos atribuir a $(\alpha_j, \beta_j), j = 1, 2$, os valores $(1,1), (\frac{1}{2}, \frac{1}{2}), (0,0), (1,0)$ e $(0,1)$.

Para $(\alpha_j, \beta_j) = (1,1)$, θ_j tem distribuição uniforme ou é uma *priori* vaga no intervalo (0,1), isto é,

$$\pi(\theta_j) = I_{(0,1)}(\theta_j),$$

e para $(\alpha_j, \beta_j) = (\frac{1}{2}, \frac{1}{2})$, θ_j tem distribuição de Jeffreys, ou seja,

$$\pi(\theta_j) \propto \theta_j^{-\frac{1}{2}}(1-\theta_j)^{-\frac{1}{2}} I_{(0,1)}(\theta_j).$$

Para (α_j, β_j) igual a $(0,0), (1,0)$ e $(0,1)$, temos para θ_j as chamadas *prioris* de referência (Bernardo, 1979, Smith, 1991).

Para *as prioris* de referência é possível obter transformações "um a um"de θ_j, de tal modo que suas distribuições sejam *prioris* vagas como mostra o seguinte teorema.

Teorema 2.2

(a) Se $\pi(\theta) = \frac{1}{\theta(1-\theta)} I_{(0,1)}(\theta)$ e $\eta = \log(\frac{\theta}{1-\theta})$, então $\pi(\eta) = 1$;

(b) se $\pi(\theta) = \frac{1}{1-\theta} I_{(0,1)}(\theta)$ e $\gamma = \log(\frac{1}{1-\theta})$, então $\pi(\gamma) = 1$ e

(c) se $\pi(\theta) = \frac{1}{\theta} I_{(0,1)}(\theta)$ e $\tau = \log(\theta)$, então $\pi(\tau) = 1$.

Prova

(a) A relação $\eta = \log(\frac{\theta}{1-\theta}), 0 < \theta < 1$, é equivalente a relação $\theta = \frac{e^\eta}{1-e^\eta}, \eta$ real.

Logo, pelo teorema de transformação de variáveis aleatórias,

$$\begin{aligned}
\pi_1(\eta) &= \pi\left(\frac{e^\eta}{1+e^\eta}\right)\left|\frac{d\left(\frac{e^\eta}{1+e^\eta}\right)}{d\eta}\right| \\
&= \left(\frac{e^\eta}{1+e^\eta}\right)^{-1}\left(1-\frac{e^\eta}{1+e^\eta}\right)^{-1}\frac{e^\eta}{(1+e^\eta)^2} \\
&= 1,
\end{aligned}$$

o que prova o teorema.■

As provas dos itens (b) e (c) são análogas.

Observamos que se $N > n_j > 1, j = 1, 2$, então, pela relação (2.4), as distribuições *a posteriori* conjunta de N e $\boldsymbol{\theta}$ existem nos casos em que *as prioris* para θ_j são as de referência.

Em particular o teorema 2.1 continua válido.

2.6 Distribuição *a priori* informativa para as probabilidades de um indivíduo pertencer às listas.

Suponhamos uma situação em que soubéssemos, com base em informações de pesquisadores, especialistas ou de experimentos passados, que θ_j pertence a um dado intervalo com alta probabilidade.

Então, podemos utilizar essa informação para escolher os valores dos parâmetros α_j e β_j.

Mais precisamente, suponhamos que θ_j pertence a um intervalo (a_j, b_j), $a_j < b_j$, com alta probabilidade. Neste caso, os valores de α_j e β_j são dados pela solução do sistema

$$\begin{cases} \frac{a_j+b_j}{2} = \frac{\alpha_j}{(\alpha_j+\beta_j)} = \mu\text{: média de } \theta_j, \\ \frac{b_j-a_j}{4} = \left\{\frac{(\alpha_j\beta_j)}{(\alpha_j+\beta_j)^2(\alpha_j+\beta_j+1)}\right\}^{\frac{1}{2}} = \sigma\text{: desvio padrão de } \theta_j, \end{cases}$$

ou seja, α_j e β_j são tais que $a_j = \mu - 2\sigma$ e $b_j = \mu + 2\sigma$.

A solução deste sistema está implementada computacionalmente no programa anexado no apêndice C. Notamos que utilizando este método, a massa da distribuição Beta contida no intervalo (a_j, b_j) é de aproximadamente 95%.

Apresentamos no Capítulo 4 exemplos com dados simulados e reais para os modelos aqui discutidos. Finalizando este capítulo, ressaltamos que esses estudos de simulação evidenciaram que as estimativas obtidas podem superestimar o valor de N. Isto ocorre devido ao fato de existir no modelo a tendência de não incluir em ambas as listas aqueles indivíduos que realmente pertencem às duas listas, devido a erros no preenchimento dos registros dos indivíduos.

Para sanar o problema desenvolvemos no Capítulo 5 uma metodologia que permite, mesmo para o caso dos registros incorretos, identificar um indivíduo que pertence a ambas as listas. A idéia é dividir o conjunto das informações sobre cada indivíduo em dois subconjuntos de modo que, se um indivíduo tiver pelo menos um desses subconjuntos coincidentes em ambas as listas, pode-se concluir que ele pertence a ambas as listas.

Capítulo 3

Estimação bayesiana do tamanho de uma população de diabéticos através de múltiplas listas de pacientes

Neste capítulo generalizamos a metologia descrita no capítulo anterior para três ou mais listas de indivíduos da população. Definindo o modelo estatístico e a função de verossimilhança, determinamos as estimativas de Bayes para o tamanho populacional e para as probabilidades dos indivíduos pertencerem às listas. Atribuímos as mesmas *prioris* do capítulo anterior para N e $\boldsymbol{\theta}$.

3.1 Modelo estatístico e a função de verossimilhança

Como no capítulo anterior, denotamos por N (N desconhecido) o tamanho da população, supomos que não haja erros no preenchimento dos dados cadastrais dos indivíduos e que dispomos de k listas de indivíduos, $k > 2$.

A cada idivíduo i da população, $i = 1, 2, ..., N$, associamos um vetor aleatório k-dimensional, $\mathbf{X}_i = (X_{i1}, X_{i2}, ..., X_{ik})$, tal que seu j-ésimo elemento, X_{ij}, assume o valor 1 se o indivíduo i pertencer à lista j e 0 caso contrário, $i = 1, 2, ..., N$ e $j = 1, 2, ..., k$.

Supomos que um indivíduo pertence a duas ou mais listas quando os conjuntos de informações sobre este indivíduo forem idênticos nestas listas e que cada indivíduo, independentemente dos demais e das outras listas, tenha probabilidade θ_j de pertencer à lista j, $0 < \theta_j < 1$. Seja $\boldsymbol{\theta} = (\theta_1,$

θ_2, ..., θ_k). Então, X_{ij} são variáveis aleatórias de Bernoulli independentes com

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

$i = 1, 2, ..., N$ e $j = 1, 2, ..., k$, e os vetores aleatórios $\mathbf{X}_i, i = 1, 2, ..., N$, são independentes e assumem valores no conjunto

$$\Omega = \{\boldsymbol{\omega}_r = (\omega_{r1},\ \omega_{r2},\ \ ...,\ \omega_{rk}) : \omega_{rj} = 0,\ 1;\ r = 1,\ ...,\ \ell;\ j = 1,\ 2,\ ...,\ k\},$$

onde l, $l = 2^k$, é o cardinal Ω e $\boldsymbol{\omega}_1$, $\boldsymbol{\omega}_2, ...,$ $\boldsymbol{\omega}_l$ é uma enumeração de todas as possíveis trajetórias de cada indivíduo da população, com $\boldsymbol{\omega}_1$, $\boldsymbol{\omega}_2, ...,$ $\boldsymbol{\omega}_{l-1}$, representando as trajetórias de cada indivíduo observado em pelo menos uma das listas e $\boldsymbol{\omega}_l = (0, 0, ..., 0)$ representando a trajetória de cada indivíduo não observado em nenhuma das listas. Como qualquer indivíduo pertence a uma lista independentemente dos demais indivíduos e das outras listas temos que

$$\begin{aligned} p_r(\boldsymbol{\theta}) &= P(\mathbf{X}_i = \boldsymbol{\omega}_r|N, \boldsymbol{\theta}) = P\left((X_{i1},\ X_{i2},\ ...,\ X_{ik}) = (\omega_{r1}, \omega_{r2},\ ...,\ \omega_{rk})|N, \boldsymbol{\theta}\right) \\ &= \prod_{j=1}^{k} P(X_{ij} = \omega_{rj}|N, \boldsymbol{\theta}) = \prod_{j=1}^{k} \theta_j^{\omega_{rj}}(1-\theta_j)^{1-\omega_{rj}}, \end{aligned}$$

$i = 1, 2, ..., N$ e $r = 1, 2, ..., \ell$.

A título de ilustração suponhamos, por exemplo, $k = 3$. Então, associado a cada indivíduo i da população existem $\ell = 2^3 = 8$ possíveis trajetórias (histórias) representadas, por exemplo, por vetores de acordo com a enumeração:

$\boldsymbol{\omega}_1 = (1, 0, 0)$ se o indivíduo pertence somente à lista 1;

$\boldsymbol{\omega}_2 = (0, 1, 0)$ se o indivíduo pertence somente à lista 2;

$\boldsymbol{\omega}_3 = (0, 0, 1)$ se o indivíduo pertence somente à lista 3;

$\boldsymbol{\omega}_4 = (1, 1, 0)$ se o indivíduo pertence somente as listas 1 e 2;

$\boldsymbol{\omega}_5 = (1, 0, 1)$ se o indivíduo pertence somente as listas 1 e 3;

$\boldsymbol{\omega}_6 = (0, 1, 1)$ se o indivíduo pertence somente as listas 2 e 3;

$\boldsymbol{\omega}_7 = (1, 1, 1)$ se o indivíduo pertence as três listas;

$\boldsymbol{\omega}_8 = (0, 0, 0)$ se o indivíduo não pertence à nenhuma lista e

$$
\begin{aligned}
p_1(\boldsymbol{\theta}) &= \theta_1(1-\theta_2)(1-\theta_3),\\
p_2(\boldsymbol{\theta}) &= (1-\theta_1)\theta_2(1-\theta_3),\\
p_3(\boldsymbol{\theta}) &= (1-\theta_1)(1-\theta_2)\theta_3,\\
p_4(\boldsymbol{\theta}) &= \theta_1\theta_2(1-\theta_3),\\
p_5(\boldsymbol{\theta}) &= \theta_1(1-\theta_2)\theta_3,\\
p_6(\boldsymbol{\theta}) &= (1-\theta_1)\theta_2\theta_3,\\
p_7(\boldsymbol{\theta}) &= \theta_1\theta_2\theta_3,\\
p_8(\boldsymbol{\theta}) &= (1-\theta_1)(1-\theta_2)(1-\theta_3).
\end{aligned}
$$

De um modo geral, sejam $n_{(r)} = \sum_{i=1}^{N} I_{\{\boldsymbol{\omega}_r\}}(\mathbf{X}_i)$ o número de indivíduos (da população) que apresentam a trajetória $\boldsymbol{\omega}_r$, $r = 1, 2, ..., \ell$ e $n = \sum_{r=1}^{\ell-1} n_{(r)}$ o número de indivíduos distintos (da população) observados nas k listas.

Observamos que $n_{(\ell)} = N - \sum_{r=1}^{\ell-1} n_{(r)} = N - n$ é o número de indivíduos que não foram observados em nenhuma das listas, ou seja, é o número de indivíduos que apresentam a trajetória $\boldsymbol{\omega}_\ell$ e $(n_{(1)}, n_{(2)}, ..., n_{(\ell-1)}, N-n)$, dados N e $\boldsymbol{\theta}$, tem distribuição multinomial com parâmetros N e $(p_1(\boldsymbol{\theta}), p_2(\boldsymbol{\theta}),..., p_l(\boldsymbol{\theta}))$.

Logo,

$$
\begin{aligned}
& P(n_{(1)}, n_{(2)}, ..., n_{(\ell-1)}, N-n|N, \boldsymbol{\theta}) \\
= & \frac{N!}{n_{(1)}!n_{(2)}!...n_{(\ell-1)}!(N-n)!}[p_1(\boldsymbol{\theta})]^{n_{(1)}}[p_2(\boldsymbol{\theta})]^{n_{(2)}}...[p_{\ell-1}(\boldsymbol{\theta})]^{n_{(\ell-1)}}[p_\ell(\boldsymbol{\theta})]^{N-n} \\
= & \frac{N!}{n_{(1)}!n_{(2)}!...n_{(\ell-1)}!(N-n)!}\prod_{r=1}^{\ell}[p_r(\boldsymbol{\theta})]^{n_{(r)}} \\
= & \frac{N!}{n_{(1)}!n_{(2)}!...n_{(\ell-1)}!(N-n)!}\prod_{r=1}^{\ell}\left(\prod_{j=1}^{k}\theta_j^{\omega_{rj}}(1-\theta_j)^{1-\omega_{rj}}\right)^{n_{(r)}} \\
= & \frac{N!}{n_{(1)}!n_{(2)}!...n_{(\ell-1)}!(N-n)!}\prod_{j=1}^{k}\prod_{r=1}^{\ell}\theta_j^{n_{(r)}\omega_{rj}}(1-\theta_j)^{n_{(r)}-n_{(r)}\omega_{rj}} \\
= & \frac{N!}{n_{(1)}!n_{(2)}!...n_{(\ell-1)}!(N-n)!}\prod_{j=1}^{k}\theta_j^{\sum_{r=1}^{l}n_{(r)}\omega_{rj}}(1-\theta_j)^{\sum_{r=1}^{l}n_{(r)}(1-\omega_{rj})} \\
= & \frac{N!}{n_{(1)}!n_{(2)}!...n_{(\ell-1)}!(N-n)!}\prod_{j=1}^{k}\theta_j^{n_j}(1-\theta_j)^{N-n_j} \\
\propto & \frac{N!}{(N-n)!}\prod_{j=1}^{k}\theta_j^{n_j}(1-\theta_j)^{N-n_j}, \qquad (3.1)
\end{aligned}
$$

onde

$$
\begin{aligned}
n_j = & \sum_{r=1}^{\ell}n_{(r)}\omega_{rj} = \sum_{r=1}^{\ell}\sum_{i=1}^{N}I_{\{\boldsymbol{\omega}_r\}}(\mathbf{X}_i)\omega_{rj} = \sum_{i=1}^{N}\sum_{r=1}^{\ell}I_{\{\boldsymbol{\omega}_r\}}(\mathbf{X}_i)\omega_{rj} = \\
= & \sum_{i=1}^{N}\sum_{r=1}^{\ell}I_{\{\boldsymbol{\omega}_r\}}(\mathbf{X}_i)X_{ij} = \sum_{i=1}^{N}\left(X_{ij}\sum_{r=1}^{\ell}I_{\{\boldsymbol{\omega}_r\}}(\mathbf{X}_i)\right) = \\
= & \sum_{i=1}^{N}X_{ij} : \text{número de indivíduos que pertencem a lista } j,\ j = 1, 2, ..., k.
\end{aligned}
$$

O modelo de trajetórias que adotamos no Capítulo 2 é um caso particular deste modelo onde consideramos apenas duas listas de indivíduos.

Seja $D = (n_1, ..., n_k, n)$ o vetor de dados observados. Então, segue de (3.1) que a função de

verossimilhança é tal que

$$L(N, \boldsymbol{\theta}|D) \propto \frac{N!}{(N-n)!} \prod_{j=1}^{k} \theta_j^{n_j} (1-\theta_j)^{N-n_j}, \tag{3.2}$$

$N \geq n$ e $0 < \theta_j < 1$, $1 \leq j \leq k$.

Na próxima seção definimos o modelo bayesiano e *as prioris* a serem utilizadas para N e $\boldsymbol{\theta}$.

3.2 Modelo bayesiano

Supomos *a priori* que $\theta_1, \theta_2, ..., \theta_k$, sejam independentes, θ_j tenha distribuição Beta, π_j, com parâmetros α_j e β_j conhecidos ($\alpha_j > 0$ e $\beta_j > 0$), $j = 1, 2, ..., k$, e que N tenha distribuição de probabilidades $\pi(N)$ definida nos inteiros positivos com os parâmetros N e $\boldsymbol{\theta}$ independentes.

Obtemos a distribuição *a priori* conjunta de N e $\boldsymbol{\theta}$ que é o produto das *prioris* $\pi(N)\pi_1(\theta_1)...$ $\pi_k(\theta_k)$, devido a independência de N e $\boldsymbol{\theta}$, isto é

$$\begin{aligned} \pi(N, \boldsymbol{\theta}) &= \pi(N) \prod_{j=1}^{k} \pi_j(\theta_j), j = 1, 2, ..., k \\ &\propto \pi(N) \prod_{j=1}^{k} \theta_j^{\alpha_j - 1} (1-\theta_j)^{\beta_j - 1}, \end{aligned} \tag{3.3}$$

$N = 1, 2, ...,$ e $0 < \theta_j < 1, j = 1, 2, ..., k$. Então, de (3.2) e (3.3), segue que a distribuição *a posteriori* conjunta de N e $\boldsymbol{\theta}$ é tal que

$$\begin{aligned} \pi(N, \boldsymbol{\theta}|D) &\propto L(N, \boldsymbol{\theta}|D)\pi(N, \boldsymbol{\theta}) \\ &\propto \frac{N!}{(N-n)!} \prod_{j=1}^{k} \theta_j^{n_j} (1-\theta_j)^{N-n_j} \pi(N) \prod_{j=1}^{k} \theta_j^{\alpha_j - 1} (1-\theta_j)^{\beta_j - 1} \\ &= \frac{N!}{(N-n)!} \pi(N) \prod_{j=1}^{k} \theta_j^{n_j + \alpha_j - 1} (1-\theta_j)^{N + \beta_j - n_j - 1}, \end{aligned} \tag{3.4}$$

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

Definida a distribuição *a posteriori* $\pi(N, \boldsymbol{\theta}|D)$ apresentamos nas próximas seções *as prioris* a serem adotadas para N.

3.3 Distribuição *a priori* uniforme e de Jeffreys para o tamanho populacional

Supomos nesta seção que a distribuição *a priori* para N é definida como $\pi(N) = \frac{1}{N^r}$, com $N = 1, 2, ...$ e $r = 0, 1$. Notamos que, para $r = 0, \pi(N) = 1, N = 1, 2, ...$, é a distribuição *a priori* uniforme nos inteiros positivos e, para $r = 1, \pi(N) = \frac{1}{N}, N = 1, 2, ...$, é a distribuição *a priori* de Jeffreys. Então, segue de (3.4), que a distribuição *a posteriori* conjunta de N e $\boldsymbol{\theta}$ é tal que

$$\pi(N, \boldsymbol{\theta}|D) \propto \frac{N!}{(N-n)!}\frac{1}{N^r}\prod_{j=1}^{k}\theta_j^{n_j+\alpha_j-1}(1-\theta_j)^{N+\beta_j-n_j-1}, \tag{3.5}$$

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

No próximo teorema verificamos as condições para a existência da distribuição *a posteriori* conjunta (3.5).

Teorema 3.1 Suponhamos que a distribuição *a priori* para N seja definida por $\pi(N) = \frac{1}{N^r}$, $N = 1, 2, ..., r = 0, 1$. Logo, a distribuição *a posteriori*, $\pi(N, \boldsymbol{\theta}|D)$, existe se $\sum_{j=1}^{k} n_j - n + \sum_{j=1}^{k}\alpha_j + r > 1$.

A prova deste teorema é análogo à do teorema 2.1■

Supomos que a hipótese $\sum_{j=1}^{k} n_j - n + \sum_{j=1}^{k}\alpha_j + r > 1$ do teorema 3.1 seja verificada daqui por diante.

Para $r = 0, \pi(N) = 1$, *priori* uniforme nos inteiros positivos, temos de (3.5) que a distribuição *a posteriori* conjunta de N e $\boldsymbol{\theta}$ é tal que

$$\pi(N, \boldsymbol{\theta}|D) \propto \frac{N!}{(N-n)!}\prod_{j=1}^{k}\theta_j^{n_j+\alpha_j-1}(1-\theta_j)^{N+\beta_j-n_j-1}, \tag{3.6}$$

$N \geq n$ e $0 < \theta_j < 1, j = 1, 2, ..., k$. Então, a distribuição condicional de N, dados $\boldsymbol{\theta}$ e D, é dada por

$$\pi(N|\boldsymbol{\theta}, D) = C_4\binom{N}{n}\left\{\prod_{j=1}^{k}(1-\theta_j)\right\}^N, \tag{3.7}$$

$N \geq n$, onde

$$\begin{aligned} C_4^{-1} &= \sum_{N\geq n}^{\infty}\binom{N}{n}\left\{\prod_{j=1}^{k}(1-\theta_j)\right\}^N \\ &= \sum_{s=0}^{\infty}\binom{s+n}{n}\left\{\prod_{j=1}^{k}(1-\theta_j)\right\}^{s+n} \\ &= \left\{\prod_{j=1}^{k}(1-\theta_j)\right\}^n\sum_{s=0}^{\infty}\binom{-n-1}{s}\left\{-\prod_{j=1}^{k}(1-\theta_j)\right\}^s \\ &= \left\{\prod_{j=1}^{k}(1-\theta_j)\right\}^n\left\{1-\prod_{j=1}^{k}(1-\theta_j)\right\}^{-n-1}. \end{aligned}$$

Logo,

$$\pi(N|\boldsymbol{\theta},D)=\binom{N}{n}\left\{1-\prod_{j=1}^{k}(1-\theta_j)\right\}^{n+1}\left\{\prod_{j=1}^{k}(1-\theta_j)\right\}^{N-n}, \tag{3.8}$$

$N\geq n$.

A distribuição condicional de N, dados $\boldsymbol{\theta}$ e D, é igual a distribuição de uma variável aleatória $n+R$, onde R tem distribuição binomial negativa com parâmetros $n+1$ e $1-\prod_{j=1}^{k}(1-\theta_j)$, como demonstrado na seção 2.3 do Capítulo 2.

De (3.6) segue que a distribuição condicional de θ_j, dados $N,\theta_m,m\neq j$ e D, é tal que

$$\pi(\theta_j|N,\theta_m,m\neq j,D)\propto\theta_j^{n_j+\alpha_j-1}(1-\theta_j)^{N+\beta_j-n_j-1}, \tag{3.9}$$

$0<\theta_j<1$, $j=1,2,...,k$. Isto é, a distribuição condicional de θ_j, dados $N,\theta_m,m\neq j$ e D, é Beta com parâmetros $n_j+\alpha_j$ e $N+\beta_j-n_j, j=1,2,...,k$.

Como visto anteriormente para $r=1$, $\pi(N)=\frac{1}{N}$, *priori* de Jeffreys, e de (3.5) temos que a distribuição *a posteriori* conjunta de N e $\boldsymbol{\theta}$ é tal que

$$\pi(N,\boldsymbol{\theta}|D)\propto\frac{N!}{(N-n)!N}\prod_{j=1}^{k}\theta_j^{n_j+\alpha_j-1}(1-\theta_j)^{N+\beta_j-n_j-1}, \tag{3.10}$$

$N \geq n$ e $0 < \theta_j < 1, j = 1, 2$. Então, a distribuição condicional de N, dados $\boldsymbol{\theta}$ e D, é dada por

$$\pi(N|\boldsymbol{\theta}, D) = C_5 \frac{1}{N} \binom{N}{n} \left\{ \prod_{j=1}^{k} (1 - \theta_j) \right\}^N, \tag{3.11}$$

$N \geq n$, onde

$$\begin{aligned}
C_5^{-1} &= \sum_{N \geq n}^{\infty} \frac{1}{N} \binom{N}{n} \left\{ \prod_{j=1}^{k} (1 - \theta_j) \right\}^N \\
&= \frac{1}{n} \sum_{N \geq n}^{\infty} \binom{N-1}{n-1} \left\{ \prod_{j=1}^{k} (1 - \theta_j) \right\}^N \\
&= \frac{1}{n} \sum_{s=0}^{\infty} \binom{s+n-1}{n-1} \left\{ \prod_{j=1}^{k} (1 - \theta_j) \right\}^{s+n} \\
&= \frac{1}{n} \left\{ \prod_{j=1}^{k} (1 - \theta_j) \right\}^n \sum_{s=0}^{\infty} \binom{-n}{s} \left\{ -\prod_{j=1}^{k} (1 - \theta_j) \right\}^s \\
&= \frac{1}{n} \left\{ \prod_{j=1}^{k} (1 - \theta_j) \right\}^n \left\{ 1 - \prod_{j=1}^{k} (1 - \theta_j) \right\}^{-n}.
\end{aligned}$$

Logo,

$$\pi(N|\boldsymbol{\theta}, D) = \binom{N-1}{n-1} \left\{ 1 - \prod_{j=1}^{k} (1 - \theta_j) \right\}^n \left\{ \prod_{j=1}^{k} (1 - \theta_j) \right\}^{N-n}, \tag{3.12}$$

$N \geq n$.

Segue, como na seção 2.3 do Capítulo 2 que a distribuição condicional de N, dados $\boldsymbol{\theta}$ e D, é igual a distribuição de uma variável aleatória $n + Q$, onde Q tem distribuição binomial negativa com parâmetros n e $1 - \prod_{j=1}^{k} (1 - \theta_j)$.

Novamente de (3.10) segue que a distribuição condicional de θ_j, dados $N, \theta_m, m \neq j$ **e** D, é tal que

$$\pi(\theta_j|N, \theta_m, m \neq j, D) \propto \theta_j^{n_j + \alpha_j - 1} (1 - \theta_j)^{N + \beta_j - n_j - 1}, \tag{3.13}$$

$0 < \theta_j < 1$, $j = 1, 2, ..., k$. Isto é, a distribuição condicional de θ_j, dados $N, \theta_m, m \neq j$ e D, é Beta com parâmetros $n_j + \alpha_j$ e $N + \beta_j - n_j, j = 1, 2, ..., k$.

3.4 Distribuição *a priori* de Poisson para o tamanho populacional

Nesta seção supomos que N tem distribuição *a priori* de Poisson truncada em zero, ou seja $\pi(N) = \frac{e^{-\lambda}\lambda^N}{N!(1-e^{-\lambda})}$, com $\lambda > 0$, conhecido e $N = 1, 2, ...$ Logo, temos de (3.4) que a distribuição *a posteriori* conjunta de N e $\boldsymbol{\theta}$ é tal que

$$\pi(N, \boldsymbol{\theta}|D) \propto \frac{\lambda^N}{(N-n)!}\prod_{j=1}^{k}\theta_j^{n_j+\alpha_j-1}(1-\theta_j)^{N+\beta_j-n_j-1}, \tag{3.14}$$

$N \geq n$ e $0 < \theta_j < 1, j = 1, 2, ..., k$, ou seja, a distribuição condicional de N, dados $\boldsymbol{\theta}$ e D, é dada por

$$\pi(N|\boldsymbol{\theta}, D) = C_6\frac{\lambda^N}{(N-n)!}\left\{\prod_{j=1}^{k}(1-\theta_j)\right\}^N, \tag{3.15}$$

$N \geq n$, onde

$$\begin{aligned} C_6^{-1} &= \sum_{N\geq n}^{\infty}\frac{1}{(N-n)!}\left\{\lambda\prod_{j=1}^{k}(1-\theta_j)\right\}^N \\ &= \sum_{s=0}^{\infty}\frac{1}{(s)!}\left\{\lambda\prod_{j=1}^{k}(1-\theta_j)\right\}^{s+n} \\ &= \left\{\lambda\prod_{j=1}^{k}(1-\theta_j)\right\}^n \exp\left\{\lambda\prod_{j=1}^{k}(1-\theta_j)\right\}. \end{aligned}$$

Logo,

$$\pi(N|\boldsymbol{\theta}, D) = \frac{1}{(N-n)!}\exp\left\{-\lambda\prod_{j=1}^{k}(1-\theta_j)\right\}\left\{\lambda\prod_{j=1}^{k}(1-\theta_j)\right\}^{N-n}, \tag{3.16}$$

$N \geq n$.

Notamos que a distribuição condicional de N, dados $\boldsymbol{\theta}$ e D, é igual a distribuição de uma variável aleatória $n + V$, onde V tem distribuição de Poisson com parâmetro $\lambda\prod_{j=1}^{k}(1-\theta_j)$, como demonstrado na seção 2.4 do Capítulo 2.

Da equação (3.14) segue que a distribuição condicional de θ_j, dados $N, \theta_m, m \neq j$ **e** D, é tal que

$$\pi(\theta_j|N, \theta_m, m \neq j, D) \propto \theta_j^{n_j+\alpha_j-1}(1-\theta_j)^{N+\beta_j-n_j-1}, \tag{3.17}$$

$0 < \theta_j < 1$, $j = 1, 2, ..., k$. Isto é, a distribuição condicional de θ_j, dados $N, \theta_m, m \neq j$ e D, é Beta com parâmetros $n_j + \alpha_j$ e $N + \beta_j - n_j, j = 1, 2, ..., k$.

3.4.1 Distribuição *a priori* hierárquica de Poisson para o tamanho populacional

Nesta subseção vamos supor como na subseção 2.4.1, que N tem distribuição *a priori* de Poisson com média λ truncada em zero. Atribuímos ao hiperparâmetro λ a distribuição Gama com parâmetros a e b conhecidos, $a > 0$ e $b > 0$. Supomos *a priori* θ_j, $j = 1, 2, ..., k$, com distribuição Beta, π_j, com parâmetros α_j e β_j conhecidos ($\alpha_j > 0$ e $\beta_j > 0$) e N, λ e $\boldsymbol{\theta}$ independentes. Com isso obtemos a distribuição *a priori* conjunta de N, $\lambda, \boldsymbol{\theta}$ é dada por

$$\pi(N,\ \lambda,\ \boldsymbol{\theta}) \propto \frac{e^{-\lambda(1+b)}\lambda^{N+(a-1)}}{N!(1-e^{-\lambda})} \prod_{j=1}^{k} \theta_j^{\alpha_j - 1}(1-\theta_j)^{\beta_j - 1}, \tag{3.18}$$

$N = 1, 2, ...; \lambda > 0$ e $0 < \theta_j < 1, j = 1, 2, ..., k$.

Assim sendo, temos de (3.2) e (3.18) que a distribuição *a posteriori* conjunta de N, λ e $\boldsymbol{\theta}$ é tal que

$$\pi(N, \lambda, \boldsymbol{\theta}|D) \propto \frac{e^{-\lambda(1+b)}\lambda^{N+(a-1)}}{(1-e^{-\lambda})(N-n)!} \prod_{j=1}^{k} \theta_j^{n_j + \alpha_j - 1}(1-\theta_j)^{N + \beta_j - n_j - 1}, \tag{3.19}$$

$N \geq n; \lambda > 0$ e $0 < \theta_j < 1, j = 1, 2, ..., k$. Ou seja, como na seção 3.4 a distribuição condicional de N, dados $\boldsymbol{\theta}$,λ e D, é dada por

$$\pi(N|\boldsymbol{\theta}, D) = \frac{1}{(N-n)!} \exp\left\{-\lambda \prod_{j=1}^{k}(1-\theta_j)\right\} \left\{\lambda \prod_{j=1}^{k}(1-\theta_j)\right\}^{N-n}, \tag{3.20}$$

$N \geq n$. A distribuição condicional de N, dados $\boldsymbol{\theta}$ e D, é igual a distribuição de uma variável aleatória $n + H$, onde H tem distribuição de Poisson com parâmetro $\lambda \prod_{j=1}^{k}(1-\theta_j)$.

Segue de (3.19) que a distribuição condicional de λ, dados $N, \boldsymbol{\theta}$ e D, é tal que

$$\pi(\lambda|N, \boldsymbol{\theta}, D) \propto \frac{(1+b)^{N+a}}{\Gamma(N+a)(1-e^{-\lambda})} \lambda^{N+(a-1)} e^{-\lambda(1+b)},$$

$\lambda > 0$.

Por outro lado, de (3.19) segue que a distribuição condicional de θ_j, dados $N, \lambda, \theta_m, m \neq j$

e D, é tal que

$$\pi(\theta_j|N,\lambda,\theta_m,m\neq j,n_1,n_2,n)\propto\theta_j^{n_j+\alpha_j-1}(1-\theta_j)^{N+\beta_j-n_j-1}, \tag{3.21}$$

$0<\theta_j<1$, $j=1,2,...,k$. Isto é, a distribuição condicional de θ_j, dados $N,\lambda,\theta_m,m\neq j$ e D, é Beta com parâmetros $n_j+\alpha_j$ e $N+\beta_j-n_j, j=1,2,...,k$.

3.5 Distribuições *a priori* para as probabilidades de um indivíduo pertencer às listas

Nesta seção, como nas seções 2.5 e 2.6, vamos atribuir aos hiperparâmetros (α_j,β_j) os valores $(1,1),(\frac{1}{2},\frac{1}{2}),(0,0),(1,0)$ e $(0,1)$ e os determinados pela regra estabelecida na seção 2.6.

Evidentemente, são válidos todos os resultados sobre tais *prioris,* assim como os comentários ali apresentados. Em particular, para obtermos estimativas bayesianas de N e $\boldsymbol{\theta}$ utilizamos o algoritmo *Gibbs sampling* (ver Capítulo 4). Apresentamos no próximo capítulo exemplos com dados simulados e reais e comentários sobre as estimativas obtidas.

Capítulo 4

Implementação dos modelos bayesianos

Neste capítulo implementamos os modelos bayesianos descritos nos capítulos 2 e 3 utilizando conjuntos de dados simulados e reais. A implementação de tais modelos foi feita através de métodos MCMC (Markov Chain Monte Carlo), mais especificamente, os algoritmos *Metropolis-Hastings* e *Gibbs sampling*, que proporcionaram resumos das distribuições *a posteriori*.

A idéia de simulação via cadeias de Markov é gerar valores de uma variável aleatória utilizando uma sequência de distribuições que converge fracamente para distribuição da variável. Tal método é menos eficiente do que a simulação direta, que consiste simplesmente na geração de amostras da distribuição original, embora ela seja aplicável a uma classe ampla de casos e desempenhe um papel extremamente importante em inferência bayesiana, no momento de calcular resumos de distribuições *a posteriori* nem sempre explicitas.

4.0.1 Método de Monte Carlo via Cadeia de Markov

Seja $X_0, X_1, ...$ uma cadeia de Markov homogênia com espaço de estados E, distribuição inicial $\{p(x), x \in E\}$ e probabilidade de transição

$$p_{ij} = P(X_{n+1} = x_{n+1} | X_n = x_n),$$

para $x_n, x_{n+1} \in E$.

A probabilidade de transição em $m, m \geq 1$, passos da cadeia é dada por

$$p_{ij}^{(m)} = P(X_{n+m} = x_{n+m} | X_n = x_n),$$

para x_n e $x_{n+m} \in E$.

Um estado x_j é acessível do estado x_i se $p_{ij}^{(m)} > 0$, para algum $m \geq 1$.

Dois estados x_i e x_j se comunicam se x_j for acessível de x_i e x_i for acessível de x_j. Uma cadeia de Markov é irredutível se todos estados se comunicam. Um estado x_i é recorrente se, a cadeia partindo de x_i, retorna a esse estado com probabilidade um.

Se um estado for recorrente e o tempo médio de retorno da cadeia a este estado for finito, então dizemos que ele é recorrente positivo. Caso contrário dizemos que ele é recorrente nulo. Todo estado que não seja recorrente é um estado transitório. Se todos os estados da cadeia forem recorrentes positivos, então ela é recorrente positiva.

Um estado x_i tem período d se $p_{ii}^{(n)} = 0$, a menos que n seja divisível por d e d seja o maior inteiro com essa propriedade. Um estado é aperiódico se tiver período $d = 1$. Uma cadeia de Markov aperiódica é uma cadeia de Markov em que todos os estados são aperiódicos.

Vale o seguinte resultado: se a cadeia de Markov for irredutível, aperiódica e recorrente positiva, então existe

$$\lim_{m\to\infty} p_{ij}^{(m)} = \lim_{m\to\infty} P(X_{n+m} = x_j | X_n = x_i) = \pi_j,$$

para todo $x_i \in E$. Dizemos que $\{\pi_j, j \geq 1\}$ é a distribuição de equilíbrio da cadeia.

Uma distribuição de probabilidades $\{p_j, j \geq 1\}$ é uma distribuição estacionária da cadeia se

$$p_j = \sum_{i\geq 1} p_i p_{ij}^{(m)}, j \geq 1.$$

Outro resultado válido é: se a cadeia for irredutível, aperiódica e recorrente positiva, então $\{\pi_j, j \geq 1\}$ é a distribuição estacionária da cadeia, onde

$$\pi_j = \lim_{m\to\infty} p_{ij}^{(m)}.$$

Assim, se quisermos amostrar de uma distribuição que atribui ao estado x_j a probabilidade π_j; temos de construir uma cadeia de Markov irredutível, aperiódica e recorrente positiva tal que $\{\pi_j, j \geq 1\}$ seja a distribuição de equilíbrio ou estacionária da cadeia.

4.0.2 Algoritmo *Metropolis-Hastings*

O algoritmo *Metropolis-Hastings* permite gerar uma amostra da distribuição *a posteriori* $\pi(\boldsymbol{\theta}|D)$, a partir das distribuições condicionais completas que podem possuir forma explicita ou

não (Metropolis *et al.*, 1953; Hastings, 1979; Chib and Greenberg, 1995).

O algoritmo *Metrópolis-Hantings* (Gilks et al, 1997) é dado por:

1 - Inicialize definindo um valor arbitrário θ_0.

2 - Gere θ^* de $q(\theta^{(j)}, \theta^*)$ e u de uma distribuição Uniforme (0,1).

3 - Seja $\alpha = \min\left\{1, \frac{\pi(\theta^*)q(\theta^*,\theta^{(j)})}{\pi(\theta^{(j)})q(\theta^{(j)},\theta^*)}\right\}$.

4 - Se $u < \alpha$; faça $\theta^{(j+1)} = \theta^*$, senão faça $\theta^{(j+1)} = \theta^{(j)}$.

5 - Repita os passos (2) , (3) e (4) até que a distribuição estacionária tenha sido obtida.

4.0.3 Algoritmo *Gibbs sampling*

O algoritmo *Gibbs sampling* é um caso especial de *Metropolis-Hastings.* Ele nos permite gerar uma amostra da distribuição *a posteriori* $\pi(\boldsymbol{\theta}|D) = \pi(\theta_1, \theta_2, ..., \theta_k|D)$, desde que as distribuições condicionais de $\theta_i, i = 1, 2, ..., k$, possuam forma explícita (Gelfand et al; Casella and George, 1992; Gelfand, 2000).

Suponha $\boldsymbol{\theta}$ um vetor com componentes $\theta_1, \theta_2, ..., \theta_k$ desconhecidos. Seja $\pi(\theta_i|\theta_j, j \neq i, \mathbf{y}), i = 1, ..., k$, a distribuição condicional de θ_i dados $\theta_j, j \neq i$ e D, para $i = 1, 2, ..., k$.

O algoritmo de *Gibbs sampling* pode ser descrito da seguinte forma:

1 - Inicialize o contador de iterações da cadeia com $j = 1$, atribua um valor inicial arbitrário para o vetor $\boldsymbol{\theta}$, $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, ..., \theta_k^{(0)})$, e obtenha $\theta^{(1)}$ de $\pi(\theta_1|\theta_2^{(0)}, ..., \theta_k^{(0)}, D)$.

2 - Gere $\theta_2^{(1)}$ da densidade condicional $\pi(\theta_2|\theta_1^{(1)}, \theta_3^{(0)}, ..., \theta_k^{(0)}, D)$.

.

.

.

k - Gere $\theta_k^{(1)}$ de $\pi(\theta_2|\theta_1^{(1)}, \theta_2^{(1)}, ..., \theta_{k-1}^{(1)}, D)$.

Repita o processo j vezes.

Existe nas primeiras iterações de ambos os algoritmos uma dependência devido ao valor arbitrário escolhido inicialmente para o vetor de parâmetros $\boldsymbol{\theta}^{(0)}$. Por este motivo, devemos considerar um *"burn-in"*(período de aquecimento da cadeia) que deve ser descartado dos valores restantes existe dependência entre os mesmos, logo devemos considerar saltos entre os elementos para obtermos uma independência aproximada.

Após burn-in e saltos, os elementos restantes podem ser considerados como uma amostra da distribuição $\pi(\boldsymbol{\theta}|D)$.

Para verificarmos a convergência das cadeias geradas pelos algoritmos, utilizamos o programa CODA - Convergence Diagnostics and Ouput Analysis for Gibbs Sampling Output (Best, et al.,

1995). Este software pode ser implementado via software R-Gui e contém um conjunto de diagnósticos indicativos da convergência. Mais especificamente, utilizamos o diagnóstico de convergência de Gelman Rubin disponível no software CODA.

Nos exemplos com dados simulados e reais foram consideradas três cadeias cada uma com trinta mil iterações. Destas cadeias os dez mil primeiros elementos foram descartados como "burn-in". Dos restantes foram selecionados o primeiro de cada dez (saltos) para garantir uma independência aproximada entre eles. Desse modo, obtemos amostras das distribuições *a posteriori* com dois mil elementos em cada cadeia, ou seja, somando as três cadeias obtemos uma amostra final de seis mil elementos das distribuições *a posteriori* marginais.

Os programas utilizados neste capítulo foram implementados via software R-Gui (*versão 1.9.0*) e são apresentados nos apêndices $B, B.1$, $B.2$ e G.

Adotamos, como nos capítulos 2 e 3, *as prioris* uniforme nos inteiros positivos, de Jeffreys e de Poisson truncada em zero para N, *prioris* não informativas, informativas e de referência para $\boldsymbol{\theta}$ e obtemos estimativas de Bayes utilizando duas ou mais listas de indivíduos da população.

4.1 Distribuição *a priori* uniforme para o tamanho populacional

Nesta seção utilizamos *a priori* uniforme nos inteiros positivos para N e diferentes *prioris* para θ_j, para o caso de dados simulados.

Exemplo 4.1.1 Neste exemplo consideramos duas listas de indivíduos da população e atribuímos os valores $N = 1000, \theta_1 = 0,5, \theta_2 = 0,6$. Geramos um vetor aleatório com distribuição multinomial com parâmetros N e $(p_1(\boldsymbol{\theta}), p_2(\boldsymbol{\theta}), p_3(\boldsymbol{\theta}), p_4(\boldsymbol{\theta}))$, obtendo as estatísticas $n_1 = 473, n_2 = 568, n = 772$, onde, como já visto na seção 2.1, $p_1(\boldsymbol{\theta}) = \theta_1(1-\theta_2), p_2(\boldsymbol{\theta}) = (1-\theta_1)\theta_2, p_3(\boldsymbol{\theta}) = \theta_1\theta_2, p_4(\boldsymbol{\theta}) = (1-\theta_1)(1-\theta_2)$, n_j é o número de indivíduos observados na lista j, $j = 1, 2$, e n é o número de indivíduos distintos observados nas duas listas. Utilizando *prioris* não informativas para $\boldsymbol{\theta}$, apresentamos na tabela abaixo, os resumos das distribuições a *posteriori* de N e de θ_1, θ_2, média, moda, intervalo de credibilidade de 95% e sua amplitude, quantis e desvio padrão.

Tabela 4.1.1 Estimativas dos resumos das distribuições *a posteriori* de N, θ_1, θ_2, para $N = 1000, \theta_1 = 0,5, \theta_2 = 0,6$

	α_j	β_j	$Média$	$Moda$	Q_1	$Mediana$	Q_3	DP	$IC\ 95\%$	$Am.IC$
N	1	1	1001	979	979	999	1020	29,76	(947;1065)	118
θ_1	1	1	0,47	-	0,46	0,47	0,49	0,02	(0,43;0,51)	0,08
θ_2	1	1	0,57	-	0,55	0,57	0,58	0,02	(0,52;0,61)	0,09
N	0,5	0,5	1001	1001	981	999	1019	28,98	(947;1062)	115
θ_1	0,5	0,5	0,47	-	0,46	0,47	0,49	0,02	(0,43;0,51)	0,08
θ_2	0,5	0,5	0,57	-	0,55	0,57	0,58	0,02	(0,52;0,61)	0,09

Gráficos *a posteriori* para $\alpha_j = \beta_j = 1$.

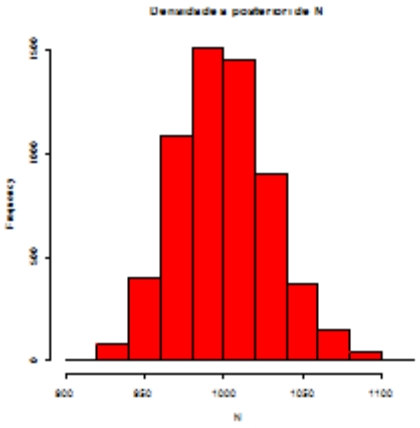

Figura 4.1.1 - Histograma da distribuição a posteriori de N

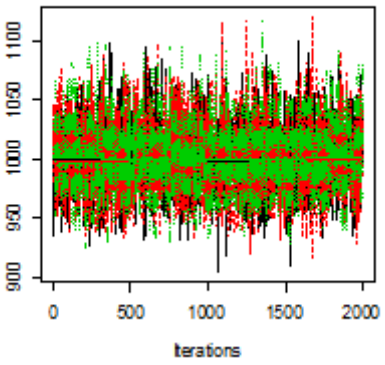

Figura 4.1.2 - Cadeias a posteriori de N

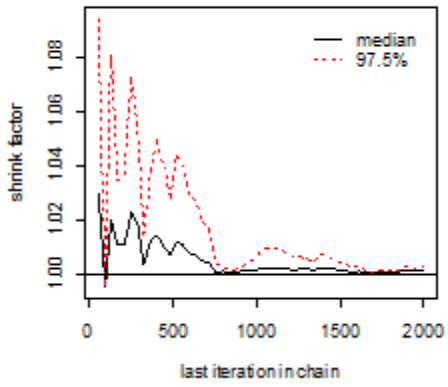

Figura 4.1.3- Gráfico do critério da convergência de Gelman Rubin das cadeias de N

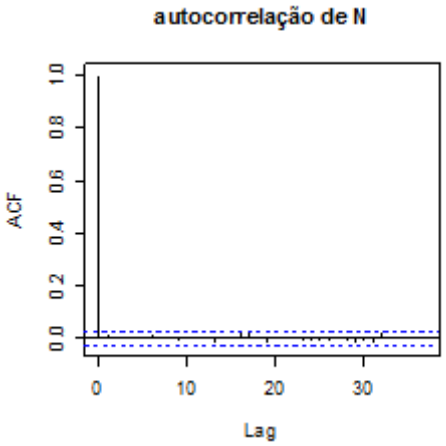

Figura 4.1.4 - autocorrelação de N

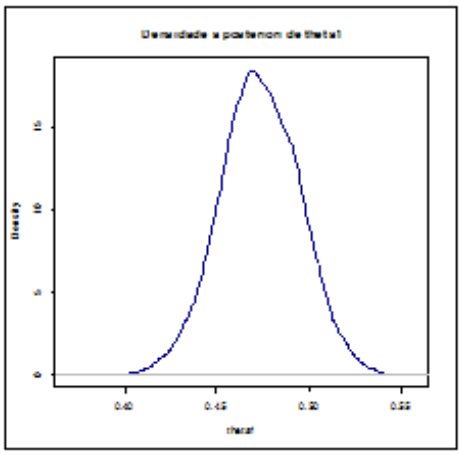

Figura 4.1.5- Densidade da distribuição a posteriori de θ_1

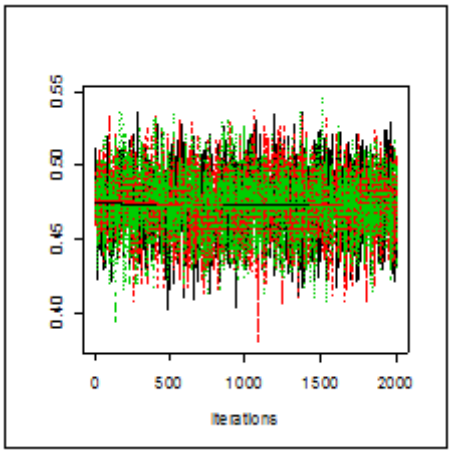

Figura 4.1.6 - cadeias a posteriori de θ_1

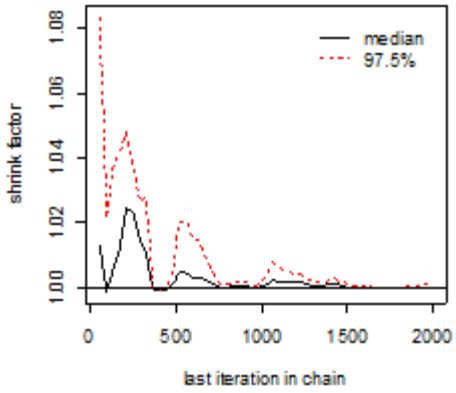

Figura 4.1.7- Gráfico do critério da convergência de Gelman Rubin das cadeias de θ_1

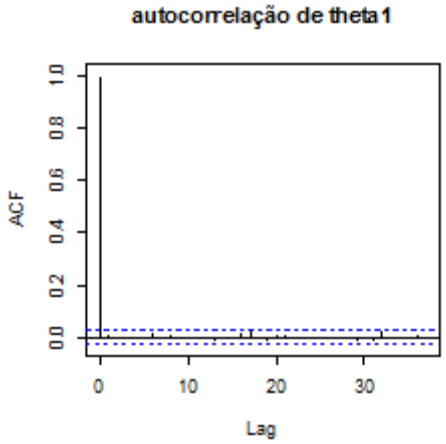

Figura 4.1.8 - autocorrelação de θ_1

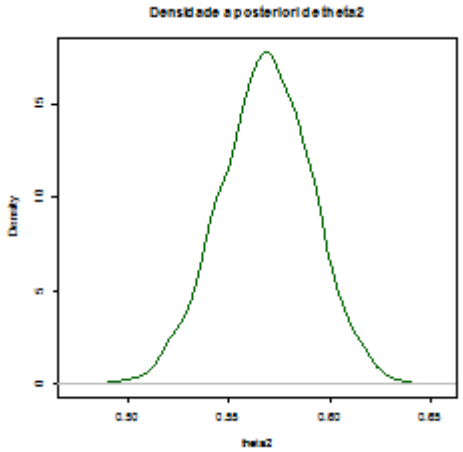

Figura 4.1.7- Gráfico do critério da convergência de Gelman Rubin das cadeias de θ_1

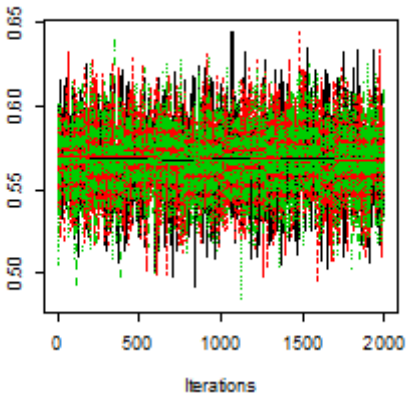

Figura 4.1.8 - autocorrelação de θ_1

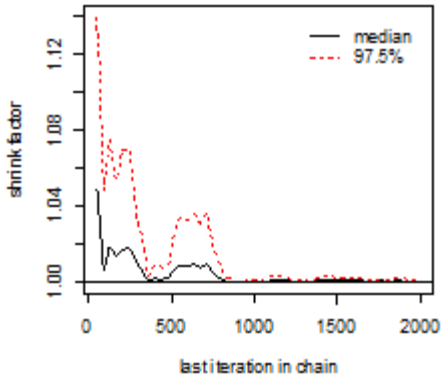

Figura 4.1.11- Gráfico do critério da convergência de Gelman Rubin das cadeias de θ_1

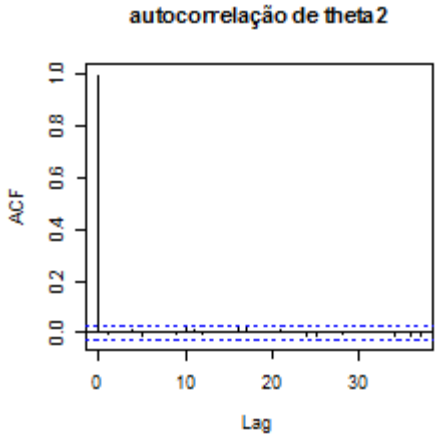

Figura 4.1.12 - autocorrelação de θ_1

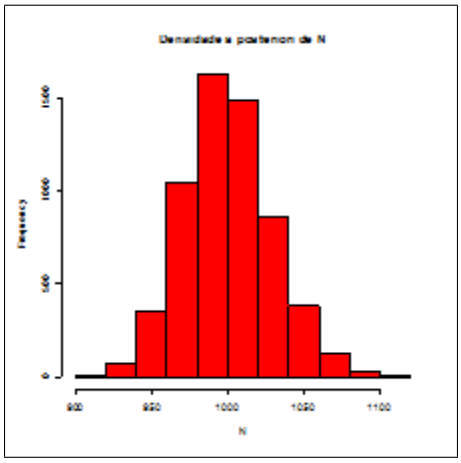

Figura 4.1.13- Histograma da distribuição a posteriori para $\alpha_j = \beta_j = 0,5$.

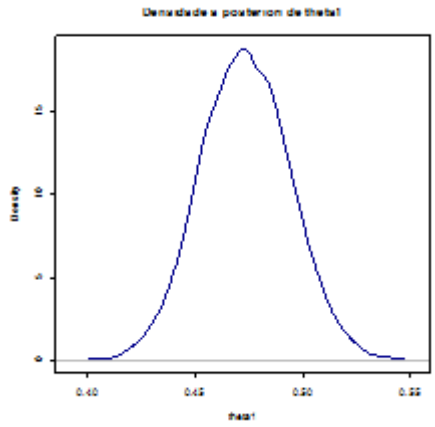

Figura 4.1.14- Densidade da distribuição a posteriori de θ_1

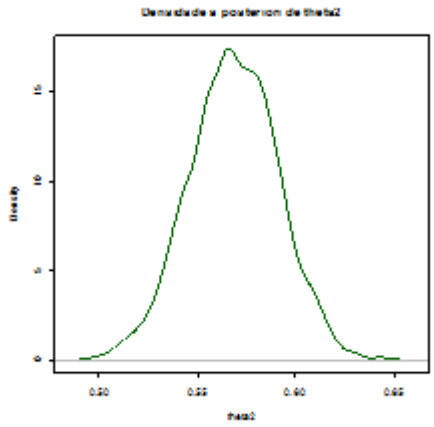

Figura 4.1.15- Densidade da distribuição a posteriori de θ_2

Como podemos observar na tabela 4.1.1 obtemos estimativas próximas dos verdadeiros valores dos parâmetros e os intervalos de credibilidade sempre contém os verdadeiros valores. Não houve problema na convergência das cadeias nem de dependência entre seus elementos, como pode ser verificado através dos gráficos do critério de convergência de Gelman Rubin e de autocorrelação acima.

Como o parâmetro N é o de interesse, apresentamos doravante somente seus resumos *a posteriori*.

Apresentamos abaixo três exemplos utilizando duas, três e quatro listas de indivíduos da população, utilizando *prioris* não informativas, informativas e de referência para $\boldsymbol{\theta}$.

Exemplo 4.1.2 Neste exemplo consideramos duas listas de indivíduos da população e atribuímos alguns valores para N e $\boldsymbol{\theta} = (\theta_1, \theta_2)$. Para cada valor atribuído geramos um vetor aleatório com distribuição multinomial com parâmetros N e $(p_1(\boldsymbol{\theta}), p_2(\boldsymbol{\theta}), p_3(\boldsymbol{\theta}), p_4(\boldsymbol{\theta}))$, obtendo as estatísticas n_1, n_2, n_3, n. Apresentamos nas tabelas abaixo os resumos das distribuições a *posteriori de* N, média, moda, intervalo de credibilidade de 95% e sua amplitude, quantis e desvio padrão.

Tabela 4.1.2 Estimativas dos resumos da distribuição *a posteriori* de N, para $N = 1000, \theta_1 = 0,5, \theta_2 = 0,6$, estatísticas $n_1 = 473, n_2 = 568, n = 772$

α_j	β_j	$Média$	$Moda$	Q_1	$Mediana$	Q_3	DP	IC 95%	$Am.IC$
1	1	1000,7	979	979	999.5	1020	29,76	(947;1065)	118
0,5	0,5	1000,6	1006	981	999	1019	28,98	(947;1062)	115
0	0	1000,4	986	980	999	1019	29,43	(946;1063)	117
0	1	1002,3	1008	982	1001	1021	29,35	(950;1063)	113
1	0	999	1017	979	998	1017	29,01	(946;1060)	114
1	4	1005,7	1004	985	1004	1025	29,55	(951;1068)	117
4	1	995,7	979	976	994	1014	28,76	(943;1058)	115
10	40	1058,6	1039	1035	1057	1081	33,87	(997;1129)	132
40	10	961,2	954	944	960	977	23,9	(917;1011)	94
100	700	1896	1872	1837	1893	1953	84,44	(1738;2066)	328
700	100	821,5	824	816	821	827	8,04	(806;839)	33

Gráficos *a posteriori* de N, para $\alpha_j = 0,5, \beta_j = 0,5$.

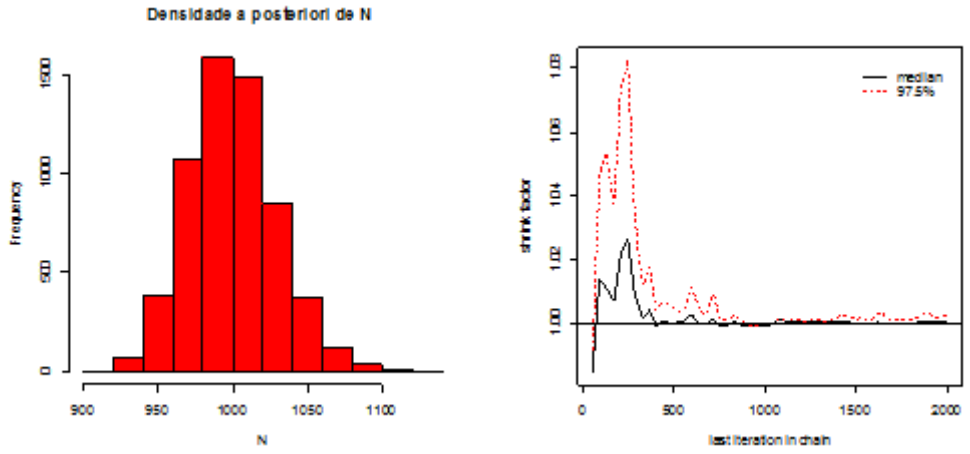

Figura 4.1.9- Histograma da dsitribuição *a posteriori* de N e gráfico do critério da convergência de Gelman Rubin das cadeias de N

Observamos através da tabela 4.1.2 que, quando utilizamos *prioris* não informativas e de referência para $\boldsymbol{\theta}$ obtemos estimativas próximas do verdadeiro valor do parâmetro e intervalos de credibilidade similares. Porém, ao utilizarmos *prioris* informativas, verificamos uma grande

variabilidade nas estimativas e na amplitude dos respectivos intervalos de credibilidade.

Tabela 4.1.3 Estimativas dos resumos da distribuição *a posteriori* de N, para $N = 100$, $\theta_1 = 0,05, \theta_2 = 0,10$, estatísticas $n_1 = 4, n_2 = 10, n = 14$

α_j	β_j	*Média*	*Moda*	Q_1	*Mediana*	Q_3	*DP*	*IC* 95%	*Am.IC*
1	1	404,5	34	45	82	184	2576	(22;2065)	2043
0,5	0,5	-	-	-	-	-	-	-	-
0	0	-	-	-	-	-	-	-	-
0	1	-	-	-	-	-	-	-	-
1	0	219,9	29	37	65	144	934,3	(19;1268)	1249
1	4	678	69	74	143	339	678	(31;3147)	3116
4	1	22,19	21	18	21	25	6,65	(15;40)	25
10	40	46,48	45	37	44	54	14,01	(26;80)	54
40	10	15,24	15	14	15	16	1,24	(14;18)	4
100	700	63,98	57	53	63	73	15,17	(38;98)	60
700	100	14,27	14	14	14	14	0,52	(14;16)	2

Gráficos *a posteriori* de N, para $\alpha_j = 100$ e $\beta_j = 700$.

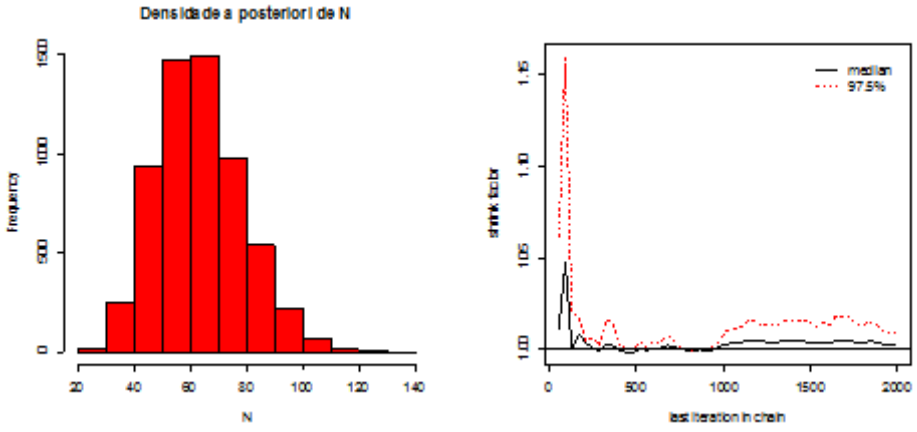

Figura 4.1.10- Histograma da distribuição *a posteriori* de N e gráfico do critério da convergência de Gelman Rubin das cadeias de N

Para $(\alpha_j = 0,5, \beta_j = 0,5)$, $(\alpha_j = 0, \beta_j = 0)$ e $(\alpha_j = 0, \beta_j = 1)$ a condição do Teorema 2.1 não se verifica, o que implica a não existência da distribuição *a posteriori* de N. Notamos que as estimativas obtidas para $(\alpha_j = 1, \beta_j = 1)$ e $(\alpha_j = 1, \beta_j = 0)$ superestimam o verdadeiro valor

de N, seus desvios padrões e as amplitudes dos intervalos de credibilidade são elevados. Isso ocorre devido às baixas probabilidades de um indivíduo qualquer pertencer a uma lista ou da pouca informação contida nos dados.

Utilizando *prioris* informativas verificamos uma grande variabilidade das estimativas e de seu respectivos intervalos de credibilidade, sendo que em nenhum dos casos o parâmetro está contido em seu respectivo intervalo de credibilidade. Logo, devemos ser cuidadosos na escolha dos hiperparâmetros α_j e β_j.

Exemplo 4.1.3 Neste exemplo consideramos três listas de indivíduos da população e atribuímos alguns valores para N e $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$. Para cada valor atribuído geramos um vetor aleatório com distribuição multinomial de parâmetros N e $(p_1(\boldsymbol{\theta}), p_2(\boldsymbol{\theta}), ..., p_8(\boldsymbol{\theta}))$, obtendo as estatísticas n_1, n_2, n_3, n, onde $p_1(\boldsymbol{\theta}) = \theta_1(1-\theta_2)(1-\theta_3), p_2(\boldsymbol{\theta}) = (1-\theta_1)\theta_2(1-\theta_3), p_3(\boldsymbol{\theta}) = (1-\theta_1)(1-\theta_2)\theta_3, p_4(\boldsymbol{\theta}) = \theta_1\theta_2(1-\theta_3), p_5(\boldsymbol{\theta}) = \theta_1(1-\theta_2)\theta_3, p_6(\boldsymbol{\theta}) = (1-\theta_1)\theta_2\theta_3, p_7(\boldsymbol{\theta}) = \theta_1\theta_2\theta_3, p_8(\boldsymbol{\theta}) = (1-\theta_1)(1-\theta_2)(1-\theta_3)$, n_j é o número de indivíduos observados na lista j, $j = 1, 2, 3$, e n é o número de indivíduos distintos observados nas três listas. Apresentamos nas tabelas abaixo os resumos das distribuições a *posteriori* de N, média, moda, intervalo de credibilidade de 95% e sua amplitude, quantis e desvio padrão.

Tabela 4.1.4 Estimativas dos resumos da distribuição *a posteriori* de N, para $N = 500$, $\theta_1 = 0,3, \theta_2 = 0,2, \theta_3 = 0,4$, estatísticas $n_1 = 143, n_2 = 85, n_3 = 187, n = 324$

α_j	β_j	$Média$	$Moda$	Q_1	$Mediana$	Q_3	DP	IC 95%	$Am.IC$
1	1	487,5	478	468	486	505	26,98	(440;545)	105
0,5	0,5	489,1	475	470	487	507	27,31	(442;547)	105
0	0	490,8	483	471	489	509	27,91	(441;550)	109
0	1	492,3	490	473	491	540	27,82	(444;552)	108
1	0	485,5	485	467	484	502	26,55	(439;542)	103
1	4	494,6	475	475	493	512	27,62	(446;555)	109
4	1	475,6	473	458	474	491	24,71	(431;528)	97
10	40	522	491	501	520	541	29,79	(470;585)	115
40	10	408,2	404	398	408	417	13,98	(383;438)	55
100	700	842,9	841	810	842	872	48,26	(754;874)	120
700	100	333	332	331	333	335	2,33	(329;338)	9

Observamos através da tabela 4.1.4 que, quando utilizamos *prioris* não informativas e de referência para $\boldsymbol{\theta}$, obtemos estimativas próximas do verdadeiro valor, bem como intervalos de credibi-lidade similares. Porém, ao utilizarmos *prioris* informativas, verificamos uma grande variabilidade nas estimativas e na amplitude de seus intervalos de credibilidade.

Tabela 4.1.5 Estimativas dos resumos da distribuição *a posteriori* de N, para $N = 800$, $\theta_1 = 0,15, \theta_2 = 0,1, \theta_3 = 0,05$, estatísticas $n_1 = 132, n_2 = 90, n_3 = 48, n = 239$

α_j	β_j	$Média$	$Moda$	Q_1	$Mediana$	Q_3	DP	IC 95%	$Am.IC$
1	1	680,7	606	608	669	739	101,4	(521;914)	393
0,5	0,5	698,4	650	623	685	760	105,81	(530;942)	412
0	0	721,9	785	643	708	786	112,14	(545;982)	437
0	1	728,3	688	649	714	794	113,93	(548;992)	444
1	0	675,3	627	604	663	734	99,29	(516;901)	385
1	4	701,3	639	628	688	762	105,28	(532;943)	411
4	1	580,6	532	529	572	625	73,01	(459;743)	284
10	40	612,2	585	563	606	654	67,64	(499;762)	263
40	10	329,94	338	318	329	340	17,06	(300;367)	67
100	700	727,6	695	692	725	762	51,12	(635;833)	198
700	100	241,3	241	240	241	242	1,83	(238;245)	7

Como podemos observar na tabela 4.1.5, não obtemos boas estimativas quando utilizamos *as prioris* não informativa e de referência, mas os intervalos de credibilidade contém seu verdadeiro valor. Utilizando *as prioris* informativas citadas acima não obtemos boas estimativas.

Exemplo 4.1.4 Neste exemplo consideramos quatro listas de indivíduos da população e atribuímos alguns valores para N e $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4)$. Para cada valor atribuído geramos um vetor aleatório com distribuição multinomial de parâmetros N e $(p_1(\boldsymbol{\theta}), p_2(\boldsymbol{\theta}), ..., p_{16}(\boldsymbol{\theta}))$, obtendo as estatísticas n_1, n_2, n_3, n_4, n, onde $p_1(\boldsymbol{\theta}) = \theta_1(1-\theta_2)(1-\theta_3)(1-\theta_4), p_2(\boldsymbol{\theta}) = (1-\theta_1)\theta_2(1-\theta_3)(1-\theta_4), ..., p_{16}(\boldsymbol{\theta}) = (1-\theta_1)(1-\theta_2)(1-\theta_3)(1-\theta_4)$, n_j é o número de indivíduos observados na lista j, $j = 1, 2, 3, 4$, e n é o número de indivíduos distintos observados nas quatro listas. Apresentamos nas tabelas abaixo os resumos das distribuições a *posteriori* de N, média, moda, intervalo de credibilidade de 95% e sua amplitude, quantis e desvio padrão.

Tabela 4.1.6 Estimativas dos resumos da distribuição *a posteriori* de N, para $N = 1000$, $\theta_1 = 0,65, \theta_2 = 0,5, \theta_3 = 0,4, \theta_4 = 0,3$,
estatísticas $n_1 = 642, n_2 = 492, n_3 = 430, n_4 = 318, n = 936$

α_j	β_j	$Média$	$Moda$	Q_1	$Mediana$	Q_3	DP	IC 95%	$Am.IC$
1	1	1008,8	1016	1001	1009	1016	10,6	(990;1031)	41
0,5	0,5	1008,8	1020	1001	1009	1016	10,67	(989;1031)	42
0	0	1009	1013	1001	1009	1016	10,73	(989;1032)	43
0	1	1009,6	1013	1002	1009	1017	10,74	(990;1032)	42
1	0	1008,6	1012	1001	1008	1016	10,92	(988;1031)	43
1	4	1010,5	10005	1003	1010	1018	10,85	(990;1033)	43
4	1	1007,9	999	1000	1008	1015	10,66	(988;1030)	43
10	40	1022,4	1020	1015	1022	1030	11,68	(1001;1046)	45
40	10	997,15	992	991	997	1003	9,47	(980;1017)	37
100	700	1293	1285	1275	1294	1312	27,72	(1241;1350)	109
700	100	945,2	942	943	945	947	3,26	(940;952)	12

Observamos através da tabela 4.1.6 que, quando utilizamos *prioris* não informativas e de refe-rência para $\boldsymbol{\theta}$, obtemos estimativas próximas do verdadeiro valor do parâmetro e intervalo de credibilidade similares. Porém, ao utilizarmos *prioris* informativas, verificamos uma grande variabilidade nas estimativas e em seus intervalos de credibilidade, sendo que em alguns casos o intervalo de credibilidade não contém o verdadeiro valor de N.

Tabela 4.1.7 Estimativas dos resumos da distribuição *a posteriori* de N,
para $N = 10000, \theta_1 = 0,4, \theta_2 = 0,1, \theta_3 = 0,05, \theta_4 = 0,2$,
estatísticas $n_1 = 3960, n_2 = 976, n_3 = 445, n_4 = 1970, n = 5836$

α_j	β_j	$Média$	$Moda$	Q_1	$Mediana$	Q_3	DP	IC 95%	$Am.IC$
1	1	10120	10200	10001	10116	10237	176,3	(9783;10472)	689
0,5	0,5	10124	10134	10004	10124	10242	176,1	(9783;10471)	688
0	0	10126	10086	10007	10127	10246	177,2	(9783;10475)	692
0	1	10133	10128	10011	10131	10252	180	(9784;10494)	710
1	0	10117	10088	9996	10116	10235	177,9	(9771;10475)	704
1	4	10129	10021	10011	10123	10250	175,7	(9793;10477)	684
4	1	10083	10041	9966	10082	10198	175,6	(9747;10441)	694
10	40	10135	10150	10019	10133	10251	174	(9801;10486)	685
40	10	9710	9666	9604	9710	9815	158,1	(9405;10031)	626
100	700	10953	10885	10827	10951	11078	185,8	(10589;11323)	734
700	100	7181,4	7186	7142	7180	7219	56,9	(7072;7295)	223

Como podemos observar na tabela 4.1.7, obtemos estimativas razoáveis quando utilizamos *as prioris* não informativas e de referência, os respectivos intervalos de credibilidade contém o verdadeiro valor do parâmetro.

Utilizando *prioris* informativas com $(\alpha_j = 1, \beta_j = 4)$ e $(\alpha_j = 4, \beta_j = 1)$, obtemos boas estimativas e para *as prioris* informativas restante não obtemos boas estimativas.

4.2 Distribuição *a priori* de Jeffreys para o tamanho populacional

Nesta seção utilizamos *a priori* de Jeffreys para N e *as prioris* não informativas, informativas e de referência para os θ_j, para o caso de dados simulados.

Apresentamos abaixo três exemplos utilizando duas, três e quatro listas de indivíduos da po-pulação e apresentamos, nas tabelas a seguir, os resumos das distribuições a *posteriori* de N, média, moda, intervalo de credibilidade de 95% e sua amplitude, quantis e desvio padrão.

Exemplo 4.2.1 Os dados utilizados nesses exemplos são os mesmos gerados no exemplo 4.1.2.

Tabela 4.2.1 Estimativas dos resumos da distribuição *a posteriori* de N, para $N = 1000$, $\theta_1 = 0,5, \theta_2 = 0,6$, estatísticas $n_1 = 473, n_2 = 568, n = 772$

α_j	β_j	$Média$	$Moda$	Q_1	$Mediana$	Q_3	DP	IC 95%	$Am.IC$
1	1	999,9	991	979	999	1019	29,19	(948;1060)	112
0,5	0,5	1000	995	980	999	1020	29,48	(948;1061)	113
0	0	999,2	990	978	998	1018	29,08	(947;1059)	112
0	1	1001,7	988	981	1000	1021	29,4	(948;1063)	115
1	0	997,8	992	977	997	1017	29,07	(946;1058)	112
1	4	1005,1	998	984	1004	1024	30,14	(951;1068)	117
4	1	994,9	969	975	994	1014	28,71	(943;1055)	112
10	40	1056,8	1065	1033	1055	1079	34,75	(993;1129)	136
40	10	960,6	941	944	960	976	23,7	(917;1009)	92
100	700	1892,3	1866	1833	1888	1946	85,18	(1736;2064)	328
700	100	821,54	823	816	821	827	8,13	(806;839)	33

Tabela 4.2.2 Estimativas dos resumos da distribuição *a posteriori* de N, para $N = 100$, $\theta_1 = 0,05, \theta_2 = 0,1$, estatísticas $n_1 = 4, n_2 = 10, n = 14$

α_j	β_j	$Média$	$Moda$	Q_1	$Mediana$	Q_3	DP	IC 95%	$Am.IC$
1	1	63,9	29	29	42	68	93	(18;236)	218
0,5	0,5	553,12	32	40	73	160	8874	(20;1704)	1684
0	0	-	-	-	-	-	-	-	-
0	1	-	-	-	-	-	-	-	-
1	0	54,55	20	24	35	57	162,3	(16;191)	175
1	4	102,66	37	43	66	114	127,6	(24;401)	377
4	1	20,87	16	17	20	23	5,48	(14;35)	21
10	40	42,77	41	34	41	49	12,54	(24;73)	49
40	10	15,13	14	14	15	16	1,15	(14;18)	4
100	700	60,27	53	50	59	69	14,61	(36;92)	56
700	100	14,24	15	14	14	14	0,51	(14;16)	2

Os resultados obtidos são similares aos das tabelas 4.1.2 e 4.1.3.

Exemplo 4.2.2 Os dados utilizados nesses exemplos são os mesmos gerados no exemplo 4.1.3.

Tabela 4.2.3 Estimativas dos resumos da distribuição *a posteriori* de N, para $N = 500$, $\theta_1 = 0,3, \theta_2 = 0,2, \theta_3 = 0,4$, estatísticas $n_1 = 143, n_2 = 85, n_3 = 187, n = 324$

α_j	β_j	$Média$	$Moda$	Q_1	$Mediana$	Q_3	DP	IC 95%	$Am.IC$
1	1	489,4	477	471	487	506	27,04	(441;548)	107
0,5	0,5	490,2	469	471	489	507	26,99	(442;548)	106
0	0	491,9	489	472	490	509	27,43	(443;552)	109
0	1	493,9	494	474	492	511	27,64	(446;554)	108
1	0	487,9	472	469	486	505	26,81	(441;546)	105
1	4	495,5	470	476	493	513	27,62	(447;555)	108
4	1	476,5	479	459	475	492	24,85	(433;530)	97
10	40	523,5	497	503	521	542	29,82	(471;588)	117
40	10	409,7	405	400	409	419	13,76	(385;438)	53
100	700	841,7	833	809	840	873	48,25	(751;940)	189
700	100	334,1	335	333	334	336	2,31	(330;339)	9

Tabela 4.2.4 Estimativas dos resumos da distribuição *a posteriori* de N, para $N = 800$, $\theta_1 = 0,15, \theta_2 = 0,1, \theta_3 = 0,05$, estatísticas $n_1 = 132, n_2 = 90, n_3 = 48, n = 239$

α_j	β_j	M	Mod	Q_1	Med	Q_3	DP	$I.C.$	$Ampl.$
1	1	674,31	658	605	664	730	97,61	(518;891)	373
0,5	0,5	691,38	638	619	680	750	102,5	(526;922)	396
0	0	711,69	682	633	698	775	109,5	(537;964)	427
0	1	719,53	631	640	706	781	111,9	(542;979)	437
1	0	668,87	627	600	658	725	97,49	(512;725)	213
1	4	694,11	605	622	682	753	102,1	(529;926)	397
4	1	577,03	541	527	569	620	70,39	(461;732)	271
10	40	609,27	611	563	604	650	66,08	(496;755)	259
40	10	331,34	343	319	330	342	17,02	(301;366)	65
100	700	725,2	730	690	723	758	50,62	(632;830)	198
700	100	242,32	241	241	242	243	1,83	(239;246)	7

Os resultados obtidos nas tabelas 4.2.3 e 4.2.4 são similares aos das tabelas 4.1.2 e 4.1.3.

Exemplo 4.2.3 Os dados utilizados nesses exemplos são os mesmos gerados no exemplo 4.1.3.

Tabela 4.2.5 Estimativas dos resumos da distribuição *a posteriori* de N,
para $N = 1000, \theta_1 = 0,65, \theta_2 = 0,5, \theta_3 = 0,4, \theta_4 = 0,3$,
estatísticas $n_1 = 642, n_2 = 492, n_3 = 430, n_4 = 318, n = 936$

α_j	β_j	$Média$	$Moda$	Q_1	$Mediana$	Q_3	DP	IC 95%	$Am.IC$
1	1	1010,3	1011	1003	1010	1017	10,81	(990;1033)	43
0,5	0,5	1010,5	1011	1003	1010	1017	10,78	(990;1033)	43
0	0	1010,4	1011	1003	1010	1017	10,8	(990;1033)	43
0	1	1010,8	1014	1003	1011	1018	11,02	(990;1033)	43
1	0	1009,9	1018	1003	1009	1017	10,76	(990;1033)	43
1	4	1011,7	1002	1004	1011	1019	11,01	(992;1035)	43
4	1	1009,2	1007	1002	1009	1016	10,64	(989;1031)	42
10	40	1023,8	1016	1015	1024	1032	12,06	(1002;1048)	46
40	10	998,5	994	992	998	1005	9,53	(981;1018)	37
100	700	1294,4	1316	1276	1294	1313	27,81	(1241;1350)	109
700	100	946,2	944	944	946	948	3,27	(940;953)	13
700	100	242,32	241	241	242	243	1,83	(239;246)	7

Tabela 4.2.6 Estimativas dos resumos da distribuição *a posteriori* de N,
para $N = 10000, \theta_1 = 0,4, \theta_2 = 0,1, \theta_3 = 0,05, \theta_4 = 0,2$,
estatísticas $n_1 = 3960, n_2 = 976, n_3 = 445, n_4 = 1970, n = 5836$

α_j	β_j	$Média$	$Moda$	Q_1	$Mediana$	Q_3	DP	IC 95%	$Am.IC$
1	1	10122,7	10163	10002	10119	10241	175,0	(9795;10469)	674
0,5	0,5	10125,8	10062	10007	10122	10242	176,0	(9789;10480)	691
0	0	10129,1	10070	10011	10128	10246	176,2	(9790;10477)	687
0	1	10133,6	10090	10013	10131	10253	177,3	(9792;10486)	694
1	0	10118,1	10005	10001	10116	10235	175,6	(9775;10475)	700
1	4	10130	10145	10010	10129	10248	177,7	(9789;10484)	695
4	1	10084,3	10134	9965	10082	10201	172,6	(9760;10429)	669
10	40	10135,8	10151	10019	10134	10252	173,9	(9802;10486)	684
40	10	9711,2	9656	9605	9707	9819	157,2	(9411;10022)	611
100	700	10952,8	10954	10829	10950	11079	185,8	(10588;11317)	729
700	100	7182,7	7167	7145	7182	7221	56,5	(7072;7295)	22323
700	100	242,32	241	241	242	243	1,83	(239;246)	7

Os resultados obtidos nas tabelas 4.2.5 e 4.2.6 são similares aos das tabelas 4.1.6, 4.1.7.

4.3 Distribuição *a priori* de Poisson para o tamanho populacional

Nesta seção utilizamos *a priori* de Poisson truncada em zero para N e *as prioris* não informativas, informativas e de referência para os θ_j, para o caso de dados simulados.

A escolha do hiperparâmetro λ da distribuição *a priori* de Poisson truncada em zero é importante, pois ele influencia as estimativas bayesianas dos parâmetros do modelo como pode ser visto no exemplo a seguir, onde atribuímos a N os valores 100 e 1000, a θ_1 o valor 0,5 e a θ_2

o valor 0,6.

Tabela4.3.1-Resumos da distribuição *a posteriori* de N para diferentes valores de λ.

N	θ_1	θ_2	λ	$Média$	$Moda$	DP	IC 95%
100	0,5	0,6	10	72,24	72	1,13	(71;75)
			50	79,35	78	3,35	(74;87)
			100	97,65	98	7,18	(85;113)
			250	227	242	15,72	(197;258)
			500	474,8	480	22,22	(431;519)
			1000	976,2	972	31,48	(914;1038)
1000	0,5	0,6	10	773	773	0,99	(772;775)
			50	777,2	778	2,32	(773;782)
			100	782,9	779	3,33	(777;790)
			250	802	801	5,91	(791;814)
			500	844	837	10,12	(825;864)
			1000	999,6	996	21,38	(959;1042)

.

Pela tabela 4.3.1, para este conjunto de dados, verificamos que obtemos boas estimativas para N somente quando o valor de λ está próximo do verdadeiro valor de N.

Contudo, uma metodologia bayesiana empírica baseada em estudos de simulação nos possibilitou definir um critério para sua escolha. A idéia para escolhermos o valor de λ é dar-lhe um valor inicial igual o da estatística n e atribuir aos seguintes o valor da correspondente média *a posteriori* de N, até que a módulo da diferença entre a respectiva média *a posteriori* de N e o correspondente valor de λ seja menor ou igual a $0,15$, por exemplo.

Esse critério pode ser resumido pelo seguinte algoritmo.

Algoritmo:

1- escolha o valor inicial para $\lambda = n$ e através de simulação (algoritmo *Gibbs sampling*) determine $E(N|\boldsymbol{\theta}, D)$;

2- se $\lambda = E(N|\boldsymbol{\theta}, D)$, a menos de um erro a ser definido, finalize;

3- senão atribua a λ o valor $E(N|\boldsymbol{\theta}, D)$, do item anterior;

4- repita (1) , (2) e (3) até que o algoritmo seja finalizado.

Para aplicarmos o algoritmo para escolha do hiperparâmetro λ vamos considerar o seguinte conjunto de dados do exemplo 4.1.1 Utilizando *a priori* uniforme para $\boldsymbol{\theta}$ $((\alpha_j, \beta_j) = (1, 1))$,

$j = 1, 2$.

Tabela 4.3.2 Escolha do valor do parâmetro λ da distribuição *a priori* de Poisson

λ	772	912,63	962,73	982,80	991,92	995,60	997,52	998,55	998,42
$E(N/\theta, D)$	912,63	962,73	982,80	991,92	995,60	997,52	998,55	998,42	998,62

Logo, através desta metodologia o valor adequado para λ é $998,55$.

Estudos de simulação evidenciaram que esta regra de escolha do valor de λ pode ser aplicada com sucesso quando $\theta_j > 0,15$. Na seção 4.3.1 vamos amenizar este problema, com a adoção de *priori* de Poisson hierárquica para N.

Apresentamos abaixo três exemplos utilizando duas, três e quatro listas de indivíduos da população e também apresentamos, nas tabelas a seguir, os resumos das distribuições *a posteriori* de N, média, moda intervalo de credibilidade de 95% e sua amplitude, quantis e desvio padrão.

Exemplo 4.3.1 Os dados utilizados nesses exemplos são os mesmos gerados no exemplo 4.1.2.

Tabela 4.3.3 Estimativas dos resumos da distribuição *a posteriori* de N, para $N = 1000$, $\theta_1 = 0,5, \theta_2 = 0,6$, estatísticas $n_1 = 473, n_2 = 568, n = 772$

α_j	β_j	$Média$	$Moda$	Q_1	$Mediana$	Q_3	DP	IC 95%	$Am.IC$
1	1	999	998	984	999	1014	21,55	(958;1042)	84
0,5	0,5	998,6	999	984	998	1012	21,29	(958;1042)	84
0	0	998,8	991	984	998	1013	21,83	(959;1043)	84
0	1	1000,7	993	987	1000	1015	21,23	(960;1044)	84
1	0	996,1	1000	982	996	1010	21,26	(956;1039)	83
1	4	1004,3	997	990	1004	1018	21,46	(963;1048)	85
4	1	993,87	983	980	993	1008	20,81	(954;1036)	82
10	40	1055,7	1048	1039	1055	1071	23,72	(1011;1103)	92
40	10	959,6	951	947	959	972	18,69	(924;997)	73
100	700	1893,4	1875	1868	1892	1919	38,08	(1820;1970)	150
700	100	821,8	814	817	822	827	7,89	(807;838)	31

Tabela 4.3.4 Estimativas dos resumos da distribuição *a posteriori* de N, para $N = 100$, $\theta_1 = 0,05, \theta_2 = 0,1$, estatísticas $n_1 = 4, n_2 = 10, n = 14$

α_j	β_j	$Média$	$Moda$	Q_1	$Mediana$	Q_3	DP	IC 95%	$Am.IC$
1	1	37,89	35	34	38	42	5,89	(27;50)	23
0,5	0,5	45	44	41	45	49	6,44	(33;58)	25
0	0	55,5	56	51	55	60	7,39	(42;71)	29
0	1	74,54	74	69	74	80	8,45	(58;91)	33
1	0	29	26	26	29	32	4,98	(26;39)	13
1	4	52	56	47	52	57	7,02	(39;66)	27
4	1	19,8	21	18	19	22	3,13	(15;27)	12
10	40	40,9	40	37	41	45	5,59	(30;52)	22
40	10	15,14	14	14	15	16	1,13	(14;18)	4
100	700	59,95	57	55	60	64	6,78	(47;74)	27
700	100	14,24	14	14	14	14	0,48	(14;15)	198

Os resultados obtidos são similares aos obtidos no exemplo 4.1.2 e 4.2.1 nas tabelas 4.1.2, 4.1.3 e 4.2.1, 4.2.2, logo vamos suprimir seus comentários.

Exemplo 4.3.2 Os dados utilizados nesses exemplos são os mesmos gerados no exemplo 4.1.3.

Tabela 4.3.5 Estimativas dos resumos da distribuição *a posteriori* de N, para $N = 500$, $\theta_1 = 0,3, \theta_2 = 0,2, \theta_3 = 0,4$, estatísticas $n_1 = 143, n_2 = 85, n_3 = 187, n = 324$

α_j	β_j	$Média$	$Moda$	Q_1	$Mediana$	Q_3	DP	IC 95%	$Am.IC$
1	1	487	484	475	486	498	16,99	(455;522)	67
0,5	0,5	489	492	478	489	500	16,85	(457;523)	66
0	0	489,9	482	478	490	501	16,96	(458;524)	66
0	1	490	485	478	489	501	17,5	(457;528)	71
1	0	486	481	475	486	497	16,86	(454;520)	66
1	4	493,8	510	482	493	506	17,29	(461;528)	67
4	1	474,9	480	464	475	486	16,31	(443;508)	65
10	40	522	514	510	522	534	17,72	(488;557)	69
40	10	409,8	405	402	410	417	11,47	(388;433)	45
100	700	839,8	844	823	839	856	24,58	(791;889)	98
700	100	334,1	331	333	334	336	2,3	(330;339)	9

Tabela 4.3.6 Estimativas dos resumos da distribuição *a posteriori* de N, para $N = 800$, $\theta_1 = 0,15, \theta_2 = 0,1, \theta_3 = 0,05$, estatísticas $n_1 = 132, n_2 = 90, n_3 = 48, n = 239$

α_j	β_j	$Média$	$Moda$	Q_1	$Mediana$	Q_3	DP	IC 95%	$Am.IC$
1	1	649,08	638	633	649	666	24,33	(599;697)	98
0,5	0,5	669,97	669	653	670	686	24,78	(621;718)	94
0	0	701,96	696	684	701	720	25,79	(652;752)	100
0	1	714,7	717	697	715	732	26,06	(665;767)	102
1	0	649,02	659	632	649	665	24,82	(601;698)	97
1	4	692,75	686	675	693	710	25,74	(642;744)	102
4	1	566,97	578	552	567	582	22,49	(523;611)	88
10	40	599,02	593	583	599	614	23,11	(554;646)	92
40	10	329,85	328	321	329	338	12,28	(307;355)	48
100	700	721,27	707	705	721	737	23,74	(675;768)	93
700	100	242,37	240	241	242	244	1,89	(239;246)	7

Suprimimos os comentários sobre as tabelas 4.3.5 e 4.3.6, devido a similaridade dos resultados

obtidos nas tabelas 4.1.4 , 4.1.5 e 4.2.3, 4.2.4..

Exemplo 4.3.3 Os dados utilizados nesses exemplos são os mesmos gerados no exemplo 4.1.4.

Tabela 4.3.7 Estimativas dos resumos da distribuição *a posteriori* de N,
para $N = 1000, \theta_1 = 0.65, \theta_2 = 0,5, \theta_3 = 0,4, \theta_4 = 0,3$,
estatísticas $n_1 = 642, n_2 = 492, n_3 = 430, n_4 = 318, n = 936$

α_j	β_j	$Média$	$Moda$	Q_1	$Mediana$	Q_3	DP	IC 95%	$Am.IC$
1	1	1010,2	1010	1003	1010	1017	10,23	(991;1031)	40
0,5	0,5	1010,3	1010	1003	1010	1017	9,97	(991;1031)	40
0	0	1010,3	1011	1004	1010	1017	10,10	(991;1031)	40
0	1	1010,8	1004	1004	1011	1018	10,14	(991;1031)	40
1	0	1010	1009	1003	1010	1017	10,22	(991;1031)	40
1	4	1011,7	1002	1005	1012	1018	10,22	(993,1033)	40
4	1	1009	1001	1002	1009	1015	9,98	(990;1029)	39
10	40	1023,5	1033	1016	1023	1031	11,03	(1002;1045)	43
40	10	998,6	999	992	998	1005	9,14	(981;1017)	36
100	700	1294	1304	1279	1294	1309	22,01	(1251;1336)	85
700	100	946,2	946	944	946	948	3,3	(940;953)	13
700	100	242,37	240	241	242	244	1,89	(239;246)	7

Tabela 4.3.8 Estimativas dos resumos da distribuição a posteriori de N,
para $N = 10000, \theta_1 = 0,4, \theta_2 = 0,1, \theta_3 = 0,05, \theta_4 = 0,2$,
estatísticas $n_1 = 3960, n_2 = 976, n_3 = 445, n_4 = 1970, n = 5836$

α_j	β_j	$Média$	$Moda$	Q_1	$Mediana$	Q_3	DP	IC 95%	$Am.IC$
1	1	10122	10082	10062	10122	10182	87,42	(9953;10293)	340
0,5	0,5	10123	10120	10063	10124	10181	87,2	(9953;10295)	342
0	0	10130	10127	10070	10129	10189	87,82	(9963;10302)	339
0	1	10131,9	10162	10072	10131	10190	86,94	(9961;10303)	342
1	0	10122,6	10142	10063	10122	10182	87,30	(9956;10295)	339
1	4	10130	10099	10069	10129	10188	87,69	(9961;10304)	343
4	1	10084,7	10093	10027	10086	10143	86,36	(9916;10253)	337
10	40	10139	10163	9973	10080	10197	86,92	(9973;10309)	336
40	10	9710,7	9717	9655	9711	9768	84,58	(9546;9876)	330
100	700	10951,3	10910	10891	10951	11013	90,50	(10770;11126)	356
700	100	7182,6	7170	7152	7183	7214	45,92	(7092;7271)	180
700	100	242,37	240	241	242	244	1,89	(239;246)	7

Os resultados obtidos são similares aos obtidos no exemplo 4.1.4 e 4.2.3 nas tabelas 4.1.6, 4.1.7 e 4.2.5, 4.2.6, logo vamos suprimir seus comentários.

4.3.1 Distribuição *a priori* de Poisson hierárquica para o tamanho populacional

Nesta subseção vamos aplicar a metodologia desenvolvida na subseção 2.4.1 do Capítulo 2 para estimar N. Inicialmente supomos $\alpha_j = \beta_j = 1$ e $a = b = 10^{-3}$, isto é, atribuímos aos hiperparâmetros α_j, β_j e λ *prioris* não informativas.

Apresentamos abaixo três exemplos utilizando duas, três e quatro listas de indivíduos da população e também apresentamos, nas tabelas a seguir, os resumos das distribuições *a posteriori* de N, média, moda intervalo de credibilidade de 95% e sua amplitude, quantis e desvio padrão.

Exemplo 4.3.4 Os dados utilizados nesses exemplos são os do exemplo 4.1.2.

Tabela 4.3.9 Estimativas dos resumos da distribuição *a posteriori* de N, para $N = 1000$, $\theta_1 = 0,5, \theta_2 = 0,6$, estatísticas $n_1 = 473, n_2 = 568, n = 772$

Média	*Moda*	Q_1	*Mediana*	Q_3	*DP*	*IC* 95%	*Am.IC*
999.7	980	979	998	1018	29.07	(947;1061)	114

Verificamos através da tabela 4.3.9 que os resumos obtidos *a posteriori* são semelhantes aos da tabela 4.3.3.

Tabela 4.3.10 Estimativas dos resumos da distribuição *a posteriori* de N, para $N = 100$, $\theta_1 = 0,05, \theta_2 = 0,1$, estatísticas $n_1 = 4, n_2 = 10, n = 14$

Média	*Moda*	Q_1	*Mediana*	Q_3	*DP*	*IC* 95%	*Am.IC*
61.9	26	29	42	67	63.96	(18;223)	205

Como podemos observar na tabela 4.3.10 as estimativas *a posteriori* são melhores do que as obtidas na tabela 4.3.4.

Exemplo 4.3.5 Os dados utilizados nesse exemplos são os do exemplo 4.1.3.

Tabela 4.3.11 Estimativas dos resumos da distribuição *a posteriori* de N, para $N = 800$, $\theta_1 = 0,15, \theta_2 = 0,1, \theta_3 = 0,05$, estatísticas $n_1 = 132, n_2 = 90, n_3 = 48, n = 239$

Média	*Moda*	Q_1	*Mediana*	Q_3	*DP*	*IC* 95%	*Am.IC*
679.2	630	609	669	735	98.8	(520;908)	388

Como podemos observar na tabela 4.3.11 as estimativas *a posteriori* são mais precisas do que as obtidas na tabela 4.3.6.

Exemplo 4.3.6 Os dados utilizados nesse exemplo são os do exemplo 4.1.4.

Tabela 4.3.12 Estimativas dos resumos da distribuição *a posteriori* de N, para $N = 10000, \theta_1 = 0,4, \theta_2 = 0,1, \theta_3 = 0,05, \theta_4 = 0,2$, estatísticas $n_1 = 3960, n_2 = 976, n_3 = 445, n_4 = 1970, n = 5836$

Média	*Moda*	Q_1	*Mediana*	Q_3	*DP*	*IC* 95%	*Am.IC*
10117	10175	10005	10113	10226	164	(9812;10453)	641

Observamos que as estimativas obtidas na tabela 4.3.12 são mais precisas do que as obtidas

na tabela 4.3.8.

Logo para $\theta_j < 0,15$ devemos utilizar *a priori* de Poisson hierárquica, para $\theta_j \geq 0,15$ obtemos intervalos de credibilidade com amplitude menores.

4.4 Distribuição *a priori* fornecida por especialistas

Nesta seção vamos aplicar a metodologia desenvolvida na seção 2.6 para as escolhas dos valores dos hiperparâmetros α_j e β_j. Para os casos de duas e três listas.

Suponhamos então, por exemplo, que soubéssemos com base em informações de pesquisadores, especialistas ou de experimentos passados que θ_1 pertence ao intervalo $(0; 0,1)$, e que θ_2 pertence ao intervalo $(0,05; 0,15)$, ambos com probabilidade de 0,95. Em seguida colhemos os dados sobre esses respectivos locais (listas) obtendo as estatística $n_1 = 4, n_2 = 10$ e $n = 14$ como os descritos na tabela 4.1.3.

Resolvendo o sistema de equações da seção 2.6, através do programa anexado no apêndice C, obtemos os seguintes valores $\alpha_1 = 3,75, \beta_1 = 71,25, \alpha_2 = 14,3, \beta_2 = 128,7$.

Tabela 4.4.1 Estimativas dos resumos da distribuição *a posteriori* de N

$\pi(N)$	*Média*	*Moda*	Q_1	*Mediana*	Q_3	*DP*	*IC* 95%	*Am.IC*
Uniforme	117,59	86	88	110	139	41,67	(57;219)	162
Jeffreys	104,98	90	79	98	124	36,23	(52;192)	140
Poisson	100,91	99	94	101	107	9,64	(94;120)	26

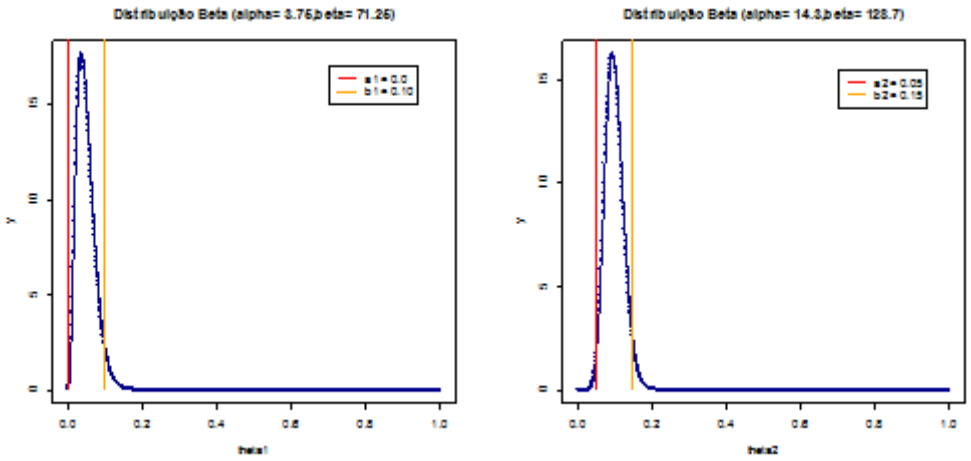

Figura 4.4.1- Distribuição Beta $(3,75; 71,25)$ e $(14,3; 128,7)$

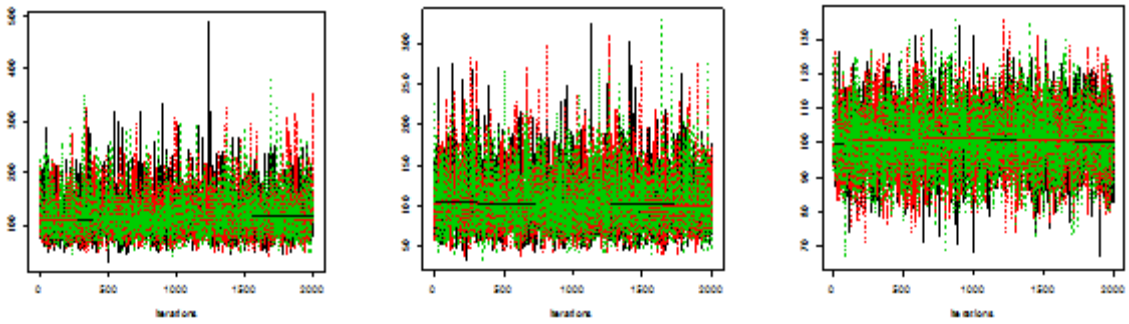

Figura 4.4.2 - Cadeias para *priori* uniforme nos inteiros positivos, de Jeffreys e de Poisson truncada em zero para N

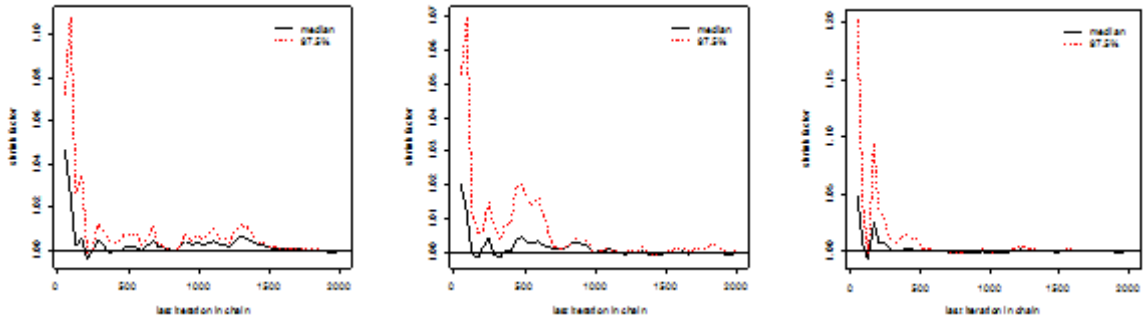

Figura 4.4.3 - Gráfico do critério da convergência de Gelman Rubin para *priori* uniforme nos inteiros positivos, de Jeffreys e de Poisson truncada em zero para N

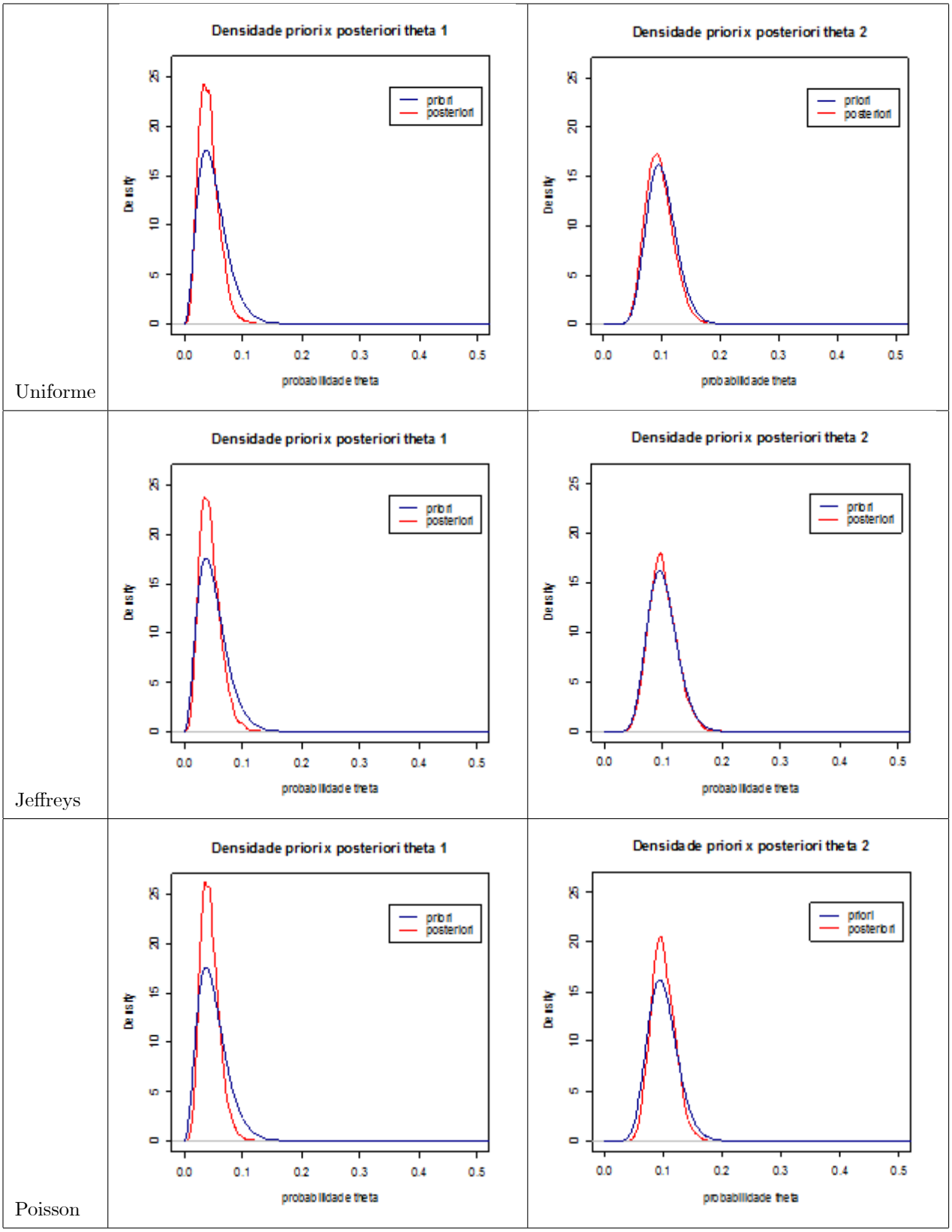

Figura 4.4.4 - Gráficos das densidades *a priori* e densidade *a posteriori* de N, para as *prioris* uniforme, de Jeffreys e de Poisson

Como pode ser verificado na tabela 4.4.1, para este conjunto de dados e utilizando *prioris*

informativas para θ_j, conseguimos corrigir o problema de estimação do parâmetro quando há baixa probabilidade de um indivíduo pertencer a uma lista. Utilizando *a priori* de Poisson truncada em zero, obtemos um intervalo de credibilidade menor que os demais.

Novamente, suponhamos por exemplo, que soubéssemos com base em informações de especialistas, que θ_1 pertence ao intervalo $(0,075; 0,225)$, θ_2 pertence ao intervalo $(0,05; 0,15)$ e θ_3 pertence ao intervalo $(0; 0,1)$, ambos com probabilidade de 0,95. Em seguida colhemos os dados sobre esses respectivos locais (listas) e obtemos as estatísticas $n_1 = 132, n_2 = 90, n_3 = 48$ e $n = 239$, como os obtidos na tabela 4.1.5.

Resolvendo o sistema de equações da seção 2.6, através do programa anexado no apêndice C, obtemos os valores $\alpha_1 = 13,45, \beta_1 = 76,216, \alpha_2 = 14,3, \beta_2 = 128,7, \alpha_3 = 3,75, \beta_3 = 72,25$.

Tabela 4.4.2 Estimativas dos resumos da distribuição *a posteriori* de N

$\pi(N)$	$Média$	$Moda$	Q_1	$Mediana$	Q_3	DP	IC 95%	$Am.IC$
Uniforme	797,31	752	730	790	857	95,62	(633;1006)	373
Jeffreys	789,66	821	724	783	847	91,54	(634;987)	353
Poisson	787,6	784	769	787	806	26,79	(736;840)	104

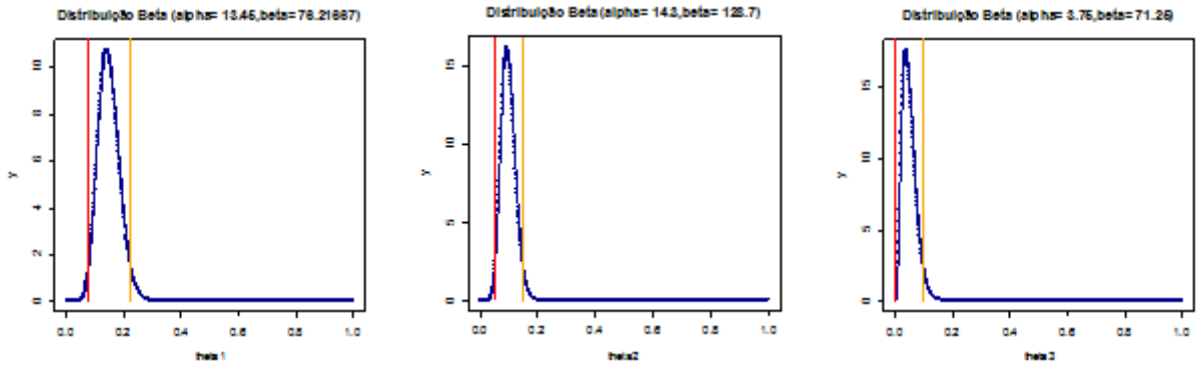

Figura 4.4.5- Distribuição Beta $(13,45; 76,216)$, $(14,3; 128,7)$ e $(3,75; 72,25)$

Como no exemplo anterior corrigimos o problema da estimação do parâmetro.

4.5 Exemplos com dados reais

Nesta seção aplicamos a metodologia desenvolvida nos Capítulos 2 e 3 para dados dos exemplos 2.3.2 e 3.2.2 de (Micheletti, 2003).

Exemplo 4.5.1 Neste exemplo consideramos os dados do exemplo 2.3.2 do Capítulo 2 de

Micheletti. Trata-se da estimação do número de novos casos, N, de DMI (Diabetes Mellitus Insulino Dependente), diagnosticados em indivíduos menores de 15 anos de idade e residentes há pelo menos um ano em Londrina, PR. Os dados foram obtidos de um estudo desenvolvido por Campos et al (1998). Foram utilizadas duas listas de pacientes, uma constituída pelas notificações feitas por médicos endocrinologistas e pediatras da cidade durante o período de 1990 à 1996 e outra constituída por inquéritos escolares de 1992 e 1993 e dados relativos a prescrições de insulina recolhidas junto às farmácias do município, em 1994. Foram identificados $n_1 = 50$ casos da doença na primeira lista, $n_2 = 24$ casos na segunda lista e $n_{12} = 21$ casos em ambas.

Tabela 4.5.1 Estimativas dos resumos da distribuição *a posteriori* de N.

$\pi(N)$	α_j	β_j	$Média$	$Moda$	Q_1	$Mediana$	Q_3	DP	IC 95%	$Am.IC$
Uniforme	1	1	58,77	55	56	58	61	4,16	(53;69)	16
Jeffreys	1	1	58,46	56	56	58	60	3,90	(53;68)	15
Poisson	1	1	58,13	58	56	58	60	3,31	(53;66)	13
Uniforme	0,5	0,5	58,16	55	55	57	60	3,96	(53;68)	15
Jeffreys	0,5	0,5	57,91	56	55	57	60	3,74	(53;67)	14
Poisson	0,5	0,5	57,50	55	55	57	59	3,06	(53;65)	12

Micheletti obteve as seguintes estimativas para N: estimativa de máxima verossimilhança $\widehat{N}_v \cong 57$ e estimativa de máxima verossimilhança condicional $\widehat{N}_c \cong 57$, intervalo de confiança aproximado de 95% de (49;67) e intervalo de confiança assintótico de 95% de (50;63). Notamos que os resultados obtidos por Micheletti são similares aos obtidos pelo método bayesiano utilizando *prioris* não informativas, como pode ser visto na tabela 4.5.1.

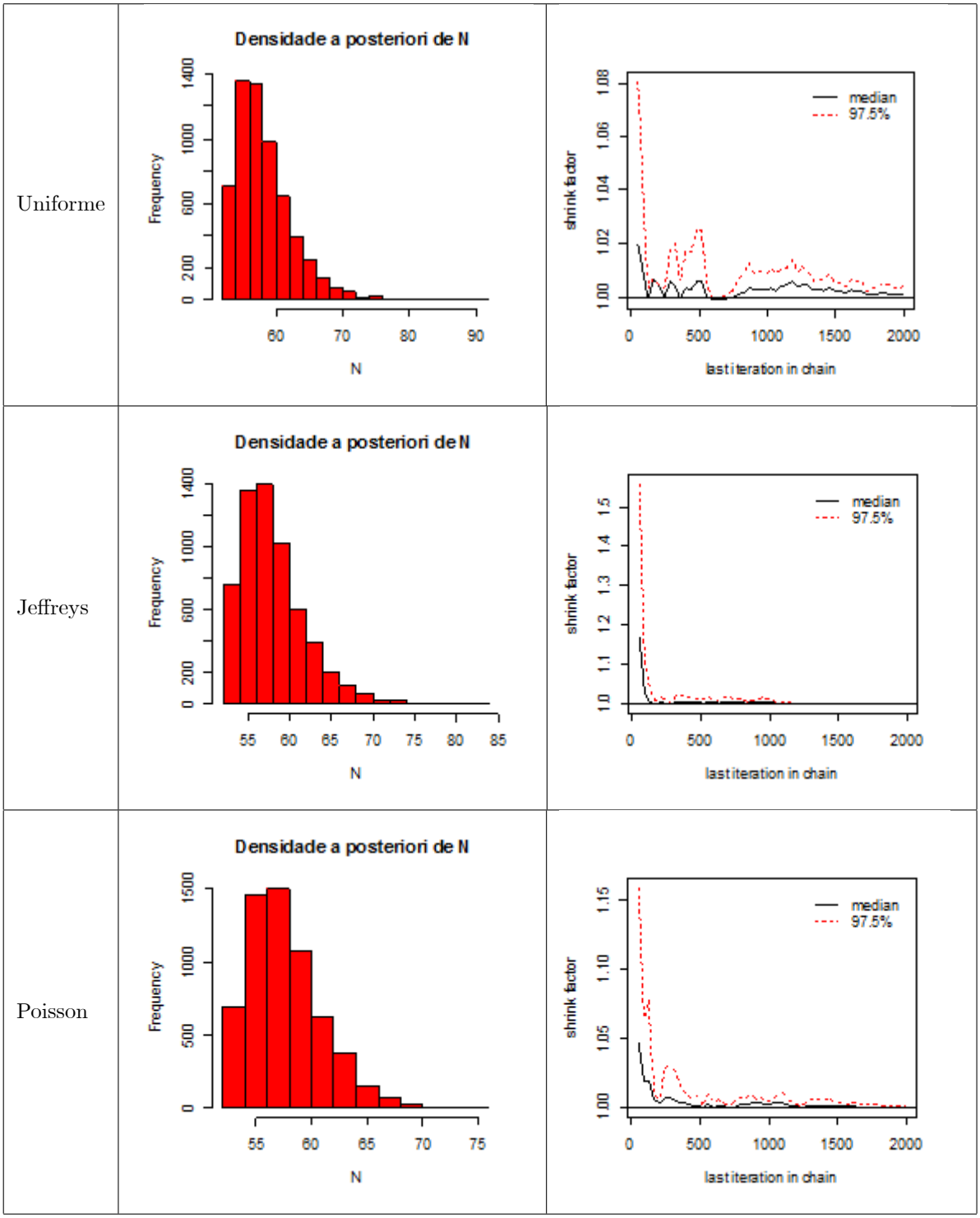

Figura 4.5.1- Histogramas das distribuições *a posteriori* de N e gráficos do critério da convergência de Gelman Rubin das cadeias N, para $\pi(N) = 1, \alpha_j = \beta_j = 1$

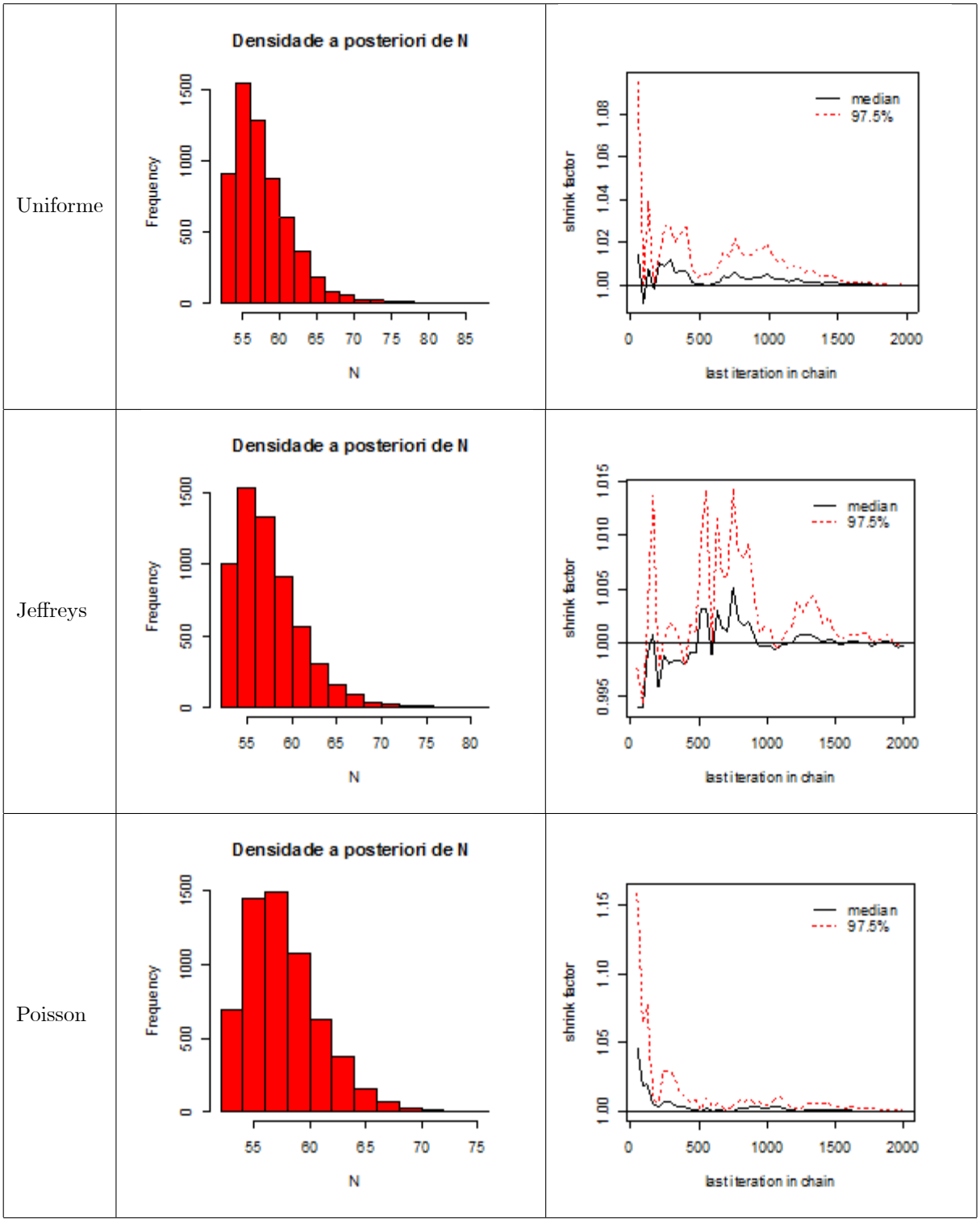

Figura 4.5.2- Histogramas das distribuições *a posteriori* de N e gráficos do critério da convergência de Gelman Rubin das cadeias de N, para $\pi(N) = 1, \alpha_j = \beta_j = 0,5$

Exemplo 4.5.2. Neste exemplo consideramos os dados do exemplo 3.2.2 de Micheletti

(2003). Onde analisamos um conjunto de dados relativo a um estudo realizado em Casale Monferrato, no Norte da Itália, em outubro de 1988 Fienberg et al (1999), para estimar a prevalência de diabetes na população local. Foram utilizadas quatro listas. A primeira lista foi obtida via clínicas da região, a segunda lista foi obtida junto a hospitais públicos e privados, a terceira lista foi constituída de prescrições de insulina e hipoglicemia oral e a quarta lista foi constituída de pacientes que receberam reembolsos para insulina e outros medicamentos. Na Tabela 4.5.2 abaixo descrevemos o conjunto de dados obtidos neste estudo. Na primeira e na terceira colunas da tabela temos todas as possíveis trajetórias (histórias) para as 4 listas e, na segunda e na quarta colunas temos o número de indivíduos que apresentaram as trajetórias correspondentes.

Tabela 4.5.2. Dados reais de um estudo sobre prevalência de diabetes (Fienberg *et al*, (1999))

$(1,0,0,0)$	709	$(0,1,0,1)$	7
$(0,1,0,0)$	74	$(0,0,1,1)$	8
$(0,0,1,0)$	182	$(1,1,1,0)$	157
$(0,0,0,1)$	10	$(1,1,0,1)$	18
$(1,1,0,0)$	104	$(1,0,1,1)$	46
$(1,0,1,0)$	650	$(0,1,1,1)$	14
$(1,0,0,1)$	12	$(1,1,1,1)$	58
$(0,1,1,0)$	20	$(0,0,0,0)$	?

Pela Tabela 4.5.2. acima obtemos $n_1 = 1754$, $n_2 = 452$, $n_3 = 1135$, $n_4 = 173$ e $n = 2069$.

Tabela 4.5.3 Estimativas dos resumos da distribuição *a posteriori* de N.

$\pi(N)$	α_j	β_j	$Média$	$Moda$	Q_1	$Mediana$	Q_3	DP	IC 95%	$Am.IC$
Uniforme	1	1	2249,5	2252	2237	2249	2262	18,68	(2215;2287)	72
Jeffreys	1	1	2250,72	2248	2238	2250	2263	18,53	(2216;2288)	72
Poisson	1	1	2250,8	2248	2239	2250	2262	17,35	(2217;2286)	69
Uniforme	0,5	0,5	2250,8	2243	2238	2250	2263	18,7	(2216;2289)	73
Jeffreys	0,5	0,5	2250,7	2241	2238	2250	2263	18,6	(2216;2289)	73
Poisson	0,5	0,5	2250,7	2252	2239	2250	2262	17,1	(2218;2285)	67

Micheletti obteve as seguintes estimativas para N: estimativa de máxima verossimilhança

$\widehat{N}_v \cong 2250$ e estimativa de máxima verossimilhança condicional $\widehat{N}_c \cong 2250$ com intervalo de confiança assintótico de 95% de $(2222; 2277)$. Notamos que os resultados obtidos por Micheletti são praticamente iguais aos obtidos pelo método bayesiano utilizando *prioris* não informativas como pode ser visto na tabela 4.5.3.

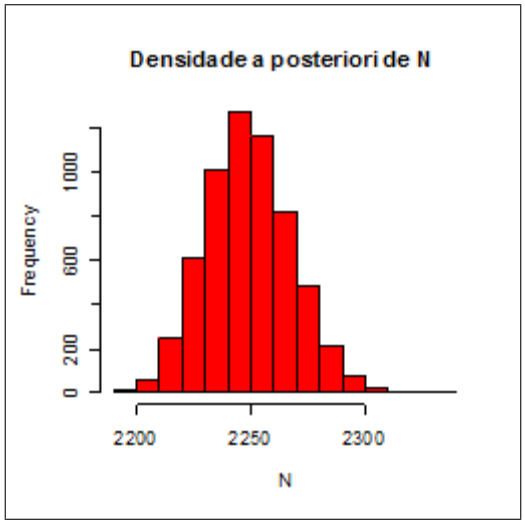

Figura 4.5.3 - Histograma da distribuição *a posteriori* de N, para $\alpha_j = \beta_j = 1$

4.6 Conclusões

Segundo os resultados obtidos através dos exemplos deste capítulo observamos que, para $\theta_j \geq 0,15$, todas *as prioris* adotadas para N produziram boas estimativas *a posteriori*. Em particular, *as prioris* uniforme e de Jeffreys forneceram praticamente os mesmos resultados. Para o caso em que a distribuição *a priori* de N é a de Poisson, obtemos intervalos de credibilidade menores do que aqueles produzidos pelas *prioris* uniforme e de Jeffreys. Para $\theta_j < 0,15$ e *prioris* não informativas e de referência para θ_j, as estimativas são muito diferentes do verdadeiro valor do N, o que não se verifica quando para *as prioris* adotadas para θ_j são informativas.

Capítulo 5

Estimação bayesiana do tamanho de uma população com dados particionados

No Capítulo 2 verificamos a existência de um problema na estimação do tamanho populacional, N, pois consideramos que um indivíduo pertence as duas listas somente quando todos seus dados cadastrais forem coincidentes. Tal fato nos leva a perda de coincidências de indivíduos nas listas, pois é possível ocorrer erro no preenchimento dos dados individuais ou omissão dos mesmos. Neste capítulo consideramos a possibilidade de um indivíduo pertencer a ambas as listas, mesmo no caso de não existir uma total "coincidência" entre seus dados cadastrais mas, como veremos, o número de indivíduos pertencentes às duas listas passa a ser um parâmetro desconhecido. Deste modo, vamos considerar dois modelos bayesianos que nos permitirão estimar o número de indivíduos pertencentes à ambas as listas, levando em conta subconjuntos dos dados cadastrais dos indivíduos e em seguida estimar N.

Determinamos neste capítulo, estimativas de Bayes e intervalo de credibilidade para os parâme-tros dos modelos e apresentamos exemplos com dados simulados e reais.

5.1 Modelo estatístico e as funções de verossimilhança

Inicialmente dividimos os dados cadastrais de cada indivíduo pertencente as duas listas, como por exemplo, nome, sexo, idade, endereço, número do CIC, número do RG, número da carteira de trabalho, número do título eleitoral, etc, em dois subconjuntos que chamamos de fichas e supomos que tais fichas possam apresentar ou não erros quando de seus preenchimentos.

Logo, a cada indivíduo, em qualquer lista, correspondem duas fichas que denotamos por A e B e que contém informações corretas ou incorretas sobre o indivíduo.

As fichas devem ser construídas de tal maneira que, se uma ou outra ficha estiver correta, o indivíduo é identificado de modo único. Descartamos a possibilidade de ocorrer uma falsa coincidência, quer se trate de um mesmo indivíduo ou indivíduos diferentes. Isto é, a "coincidência" de uma ficha em ambas às listas significa que ela foi corretamente preenchida nas duas listas e se refere, portanto, a um mesmo indivíduo.

Deste modo cada ficha deve ser composta por pelo menos dois registros, pois se ela fosse cons-tituída apenas pelo número do RG, por exemplo, o indivíduo registrado incorretamente poderia conduzir a uma falsa "coincidência". Suponhamos que para cada indivíduo, as fichas A e B sejam preenchidas independentemente. Logo, para garantirmos essa independência devemos tomar cuidado na composição das fichas, pois se entre os dados cadastrais existirem, por exemplo, informações que consideramos mais fáceis de serem digitadas ou registradas incorretamente, então elas deveriam estar em uma mesma ficha. Finalmente, suponhamos que cada indivíduo pertença ou não a uma lista qualquer independentemente dos demais indivíduos e da outra lista.

O conjunto das trajetórias possíveis 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

AO, AO significa que as fichas A's do indivíduo foram preenchidas corretamente em ambas as listas e as fichas B's foram preenchidas incorretamente em ambas as listas;

AO, OB significa que a ficha A do indivíduo foi preenchida corretamente na lista 1 e preenchida incorretamente na lista 2, a ficha B foi preenchida incorretamente na lista 1 e preenchida corretamente na lista 2;

.

.

.

OO, OO significa que as fichas A's e B's do indivíduo foram preenchidas incorretamente em ambas as listas.

Observamos que AB, AB é a única trajetória observável e AO, AB; AB, AO e $AB, OB; OB, AB$ são trajetórias não observáveis e indistinguíveis daquelas que apresentam a trajetória AO, AO e OB, OB, respectivamente.

Denotemos por

n_1 o número de indivíduos observados na lista 1;

n_2 o número de indivíduos observados na lista 2;

$m = \min\{n_1, n_2\}$;

n_{12} o número (não observado) de indivíduos presentes em ambas as listas;

m_{AO} o número de indivíduos que apresentam as trajetórias $AB, AO; AO, AB; AO, AO$, isto é, m_{AO} é o número de indivíduos que possuem apenas as fichas $A's$ coincidentes em ambas as listas;

m_{OB} o número de indivíduos que apresentam as trajetórias $AB, OB; OB, AB; OB, OB$, ou seja, m_{OB} é o número de indivíduos que possuem apenas as fichas $B's$ coincidentes em ambas as listas;

m_{AB} o número de indivíduos que apresentam as trajetórias AB, AB, ou m_{AB} é o número de indivíduos que tem as fichas $A's$ e $B's$ coincidentes em ambas as listas e

$m_T = m_{AO} + m_{OB} + m_{AB}$ o número total de indivíduos distintos observados nas duas listas. Diferentemente do Capítulo 2, n_{12} é um parâmetro desconhecido e $n_{12} - m_T = n_{12} - (m_{AO} + m_{OB} + m_{AB})$ é o número de indivíduos de ambas as listas para os quais nenhuma das fichas é coincidente. Notamos que $n_{12} - m_T$ é não observável.

Na sequência vamos estimar bayesianamente n_{12} e posteriormente vamos utilizar sua estimativa para estimar N como no Capítulo 2.

Sejam

p_1 a probabilidade de que apenas as fichas $A's$ sejam preenchidas corretamente em ambas as listas;

p_2 a probabilidade de que apenas as fichas $B's$ sejam preenchidas corretamente em ambas as listas;

p_3 a probabilidade de que as fichas $A's$ e $B's$ sejam preenchidas corretamente em ambas as listas;

p_4 a probabilidade de que as fichas $A's$ e $B's$ sejam preenchidas incorretamente em ambas as listas;

θ_1 a probabilidade de que um indivíduo qualquer da população pertença a lista 1;

θ_2 a probabilidade de que um indivíduo qualquer da população pertença a lista 2, $\boldsymbol{\theta} = (\theta_1, \theta_2)$

e $\mathbf{p} = (p_1, p_2, p_3, p_4)$. Notamos que $\sum_{j=1}^{4} p_j = 1$.

Logo, dados n_{12} e $\mathbf{p}$, $(m_{AO},\, m_{OB},\, m_{AB},\, n_{(12)} - m_T)$ tem distribuição Multinomial com parâme-tros n_{12} e $\mathbf{p}$, o que implica

$$\begin{aligned} & P(m_{AO},\, m_{OB}, m_{AB},\, n_{12} - m_T | n_{12},\, \mathbf{p}) \\ = \;& \frac{n_{12}!}{m_{AO}! m_{OB}! m_{AB}! (n_{12} - m_T)!} {p_1}^{m_{AO}} {p_2}^{m_{OB}} {p_3}^{m_{AB}} {p_4}^{n_{12} - m_T}. \end{aligned} \tag{5.1}$$

Seja $Ð = (m_{AO},\, m_{OB}, m_{AB}, m)$ os dados observados. De (5.1) segue que a função de verossimilhança de n_{12} e $\mathbf{p}$ é tal que

$$L(n_{12},\, \mathbf{p} |\, Ð) \propto \frac{n_{12}!}{(n_{12} - m_T)!} {p_1}^{m_{AO}} {p_2}^{m_{OB}} {p_3}^{m_{AB}} {p_4}^{n_{12} - m_T}, \tag{5.2}$$

$m_T \leq n_{12} \leq m$ e $0 < p_j < 1, j = 1, 2, 3, 4$.

Uma vez obtida uma estimativa de n_{12}, $\widehat{n_{12}}$, a idéia é estimar, em seguida, o parâmetro N. Então, de (2.2) segue que uma estimativa da função de verossimilhança de N e $\boldsymbol{\theta}$ é tal que

$$L^*(N, \theta | n_1, n_2, \widehat{n_{12}}) \propto \binom{N}{\widehat{n}} \prod_{j=1}^{2} \theta_j^{n_j} (1 - \theta_j)^{N - n_j}, \tag{5.3}$$

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

5.2 Modelo bayesiano

Suponhamos *a priori* n_{12} e $\mathbf{p}$ independentes, $\mathbf{p}$ com distribuição de Dirichlet, $\pi(\mathbf{p})$, com parâmetros $\delta_j, \delta_j > 0$, conhecido, $j = 1, 2, 3, 4$ e n_{12} com distribuição $\pi(n_{12})$, cujo suporte é o conjunto $\{1, 2, ..., m\}$. Então, a distribuição *a priori* conjunta de n_{12} e $\mathbf{p}$, é tal que

$$\begin{aligned} \pi(n_{12},\, \mathbf{p}) &= \pi(n_{12}) \pi(\mathbf{p}) \\ &= \pi(n_{12}) \Gamma\left(\sum_{j=1}^{4} \delta_j \right) \prod_{j=1}^{4} \frac{p_j^{\delta_j - 1}}{\Gamma(\delta_j)} \\ &\propto \pi(n_{12}) \prod_{j=1}^{4} p_j^{\delta_j - 1}, \end{aligned} \tag{5.4}$$

$1 \leq n_{12} \leq m$ e $0 < p_j < 1, j = 1, 2, 3, 4$. Temos de (5.2) e (5.4), segue que a distribuição *a posteriori* conjunta de n_{12} e $\mathbf{p}$, é tal que

$$\begin{aligned}
\pi(n_{12}, \mathbf{p}|\ Đ) &\propto L(n_{12}, \mathbf{p}|\ Đ)\pi(n_{12}, \mathbf{p}) \\
&\propto \frac{n_{12}!}{(n_{12}-m_T)!} {p_1}^{m_{AO}} {p_2}^{m_{OB}} {p_3}^{m_{AB}} {p_4}^{n_{12}-m_T} \pi(n_{12}) \prod_{j=1}^{4} p_j^{\delta_j - 1} \\
&= \pi(n_{12}) \frac{n_{12}!}{(n_{12}-m_T)!} {p_1}^{m_{AO}+\delta_1-1} {p_2}^{m_{OB}+\delta_2-1} {p_3}^{m_{AB}+\delta_3-1} {p_4}^{n_{12}-m_T+\delta_4-1} \quad (5.5)
\end{aligned}$$

$m_T \leq n_{12} \leq m$ e $0 < p_j < 1, j = 1, 2, 3, 4$.

Por outro lado, de (2.4) segue que uma estimativa da distribuição *a posteriori* conjunta de N e $\boldsymbol{\theta}$ é tal que

$$\pi(N, \theta | n_1, n_2, \widehat{n}) \propto \pi(N) \binom{N}{\widehat{n}} \prod_{j=1}^{2} \theta_j^{n_j+\alpha_j-1} (1-\theta_j)^{N-n_j+\beta_j-1}, \quad (5.6)$$

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

5.3 Distribuições *a priori* não informativas para os parâmetros do modelo

Nesta seção supomos que n_{12} tem distribuição *a priori* uniforme no conjunto $\{1, 2, ..., m\}$, ou seja, $\pi(n_{12}) = \frac{1}{m}$, $n_{12} = 1, 2, ..., m$, N tem distribuição uniforme nos inteiros positivos, θ_1 e θ_2 tem distribuição Beta com parâmetros $\alpha_1 = \alpha_2 = \beta_1 = \beta_2 = 1$ e $\delta_1 = \delta_2 = \delta_3 = \delta_4 = \delta$. Segue de (5.5) que a distribuição *a posteriori* conjunta de n_{12} e $\mathbf{p}$ é tal que

$$\pi(n_{12}, \mathbf{p}|\ Đ) \propto \binom{n_{12}}{m_T} {p_1}^{m_{AO}+\delta-1} {p_2}^{m_{OB}+\delta-1} {p_3}^{m_{AB}+\delta-1} {p_4}^{n_{12}-m_T+\delta-1}, \quad (5.7)$$

$m_T \leq n_{12} \leq m$ e $0 < p_j < 1, j = 1, 2, 3, 4$.

Determinamos, na seqüência as distribuições condicionais que são necessárias para obtermos resumos *a posteriori* de n_{12} e $\mathbf{p}$.

A distribuição condicional de n_{12}, dados $\mathbf{p}$ e Đ, é tal que

$$\pi(n_{12}|\mathbf{p}, Đ) \propto \binom{n_{12}}{m_T} {p_4}^{n_{12}}, \quad (5.8)$$

$m_T \leq n_{12} \leq m$, e a distribuição condicional de $\mathbf{p}$, dados n_{12} e Ð, é tal que

$$\pi(\mathbf{p}|n_{12},\ Ð) \propto {p_1}^{m_{AO}+\delta-1}{p_2}^{m_{OB}+\delta-1}{p_3}^{m_{AB}+\delta-1}{p_4}^{n_{12}-m_T+\delta-1}, \tag{5.9}$$

$0 < p_j < 1, j = 1, 2, 3, 4$. Isto é, $\mathbf{p}$ dados n_{12} e Ð, tem distribuição de Dirichlet com parâmetros $m_{AO} + \delta$, $m_{OB} + \delta$, $m_{AB} + \delta$ e $n_{12} - m_T + \delta$.

Por outro lado, de (2.14) segue que uma estimativa da distribuição *a posteriori* conjunta de N e $\boldsymbol{\theta}$ é tal que

$$\pi(N, \theta|n_1, n_2, \widehat{n}) \propto \binom{N}{\widehat{n}} \prod_{j=1}^{2} \theta_j^{n_j}(1-\theta_j)^{N-n_j}, \tag{5.10}$$

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

De acordo com o Capítulo 2 e (5.10) segue que a distribuição condicional de N, dados $\boldsymbol{\theta}, n_1, n_2$ e $\widehat{n}$, é igual à distribuição de uma variável aleatória $\widehat{n} + H$, onde H tem distribuição binomial negativa com parâmetros $\widehat{n} + 1$ e $1 - \prod_{j=1}^{2}(1-\theta_j)$ e a distribuição condicional de θ_j, dados $\theta_h, h \neq j, N, n_1, n_2$ e $\widehat{n}$ tem distribuição Beta com parâmetros $n_j + 1$ e $N - n_j + 1, j = 1, 2$.

Na próxima seção apresentamos exemplos do modelo acima.

5.4 Exemplos com dados simulados

Nesta seção apresentamos exemplos com dados simulados. Na implementação do modelo des-crito acima utilizamos métodos de simulação estocástica MCMC, mais especificamente, o algoritmo *Gibbs sampling* e o algoritmo de busca em tabela estática, para determinarmos os resumos das distribuições marginais *a posteriori* dos parâmetros.

A convergência das cadeias geradas foi verificada através do software CODA, utilizando o critério da convergência de Gelman Rubin.

Consideramos cadeias com quarenta e três mil elementos. Destas cadeias os trinta mil elementos iniciais foram descartados como *"burn-in"* e dos restantes foram considerados um em cada duzentos, para obtermos independência aproximada entre seus elementos.

Nos exemplos com dados simulados atribuímos valores aos parâmetros N, θ_1 e θ_2 e obtivemos as estatísticas n_1, n_2 e n_{12}, gerando uma amostra de tamanho um da distribuição (2.1). O vetor $\mathbf{p}$ foi gerado da distribuição de Dirichlet com parâmetros $\delta_1 = \delta_2 = \delta_3 = \delta_4 = 1$ (caso não informativo) e, em seguida, obtivemos as estatísticas m_{AO}, m_{OB} e m_{AB}, gerando uma amostra de tamanho um da distribuição (5.1).

Os programas utilizados para gerar os resumos *a posteriori* dos parâmetros de interesse foram implementados via software R-gui (versão 1.9.0) e está disponível no apêndice D.

Exemplo 5.4.1 - Neste exemplo os valores atribuídos aos parâmetros foram $N = 500, \theta_1 = 0,7$ e $\theta_2 = 0,5$. Obtivemos as estatísticas $n_1 = 344, n_2 = 265$ e $n_{12} = 174$, e em seguida, obtivemos as estatísticas $m_{AO} = 52, m_{OB} = 27$ e $m_{AB} = 85$.

Os resumos das distribuições *a posteriori* de n_{12} e N são dados na tabela 5.4.1 e 5.4.2, respectivamente.

Tabela 5.4.1 Estimativas dos resumos da distribuição *a posteriori* de n_{12}

δ	*Média*	*Moda*	Q_1	*Mediana*	Q_3	*DP*	*IC* 95%	*Am.IC*
0,1	157,56	145	145	145	153	26,50	(145;248)	103
0,5	179,16	146	150	167	201	33,84	(145;257)	112
1	187,95	147	159	180	213	33,58	(146;259)	113
5	197,18	198	178	193	213	24,78	(158;253)	95
10	196,96	175	182	194	210	20,77	(165;247)	82

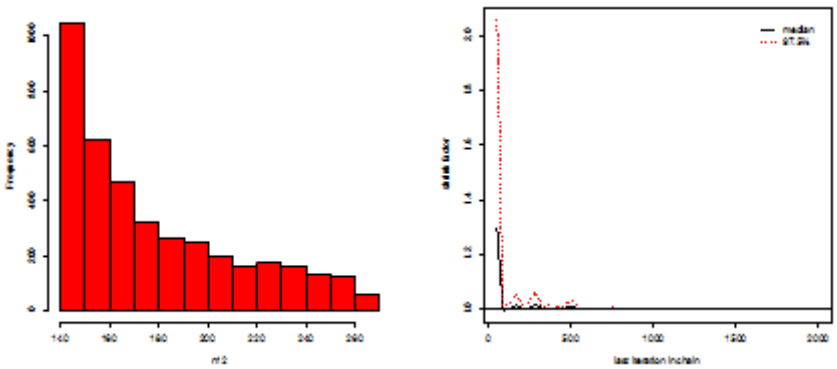

Figura 5.4.1- Histograma da distribuição *a posteriori* de n_{12} e gráfico do critério da convergência de Gelman Rubin das cadeias de n_{12}, para $\delta = 0,5$

510,33	509	500	509	520	15,27	(484;543)	59
486,06	472	477	485	494	12,95	(463;514)	51
463,93	459	456	463	471	11,07	(445;487)	42

463,93 459 456 463 471 11,07 (445;487) 42

Tabela 5.4.2 Estimativas dos resumos da distribuição *a posteriori* de N

δ	$Média$	$Moda$	Q_1	$Mediana$	Q_3	DP	IC 95%	$Am.IC$
0,1	580,69	574	564	579	594	21,74	(541;627)	86
0,5	510,33	509	500	509	520	15,27	(484;543)	59
1	486,06	472	472	485	494	12,95	(463;514)	51
5	463,93	459	459	463	471	11,07	(445;487)	42
10	463,93	459	459	463	471	11,07	(445;487)	42

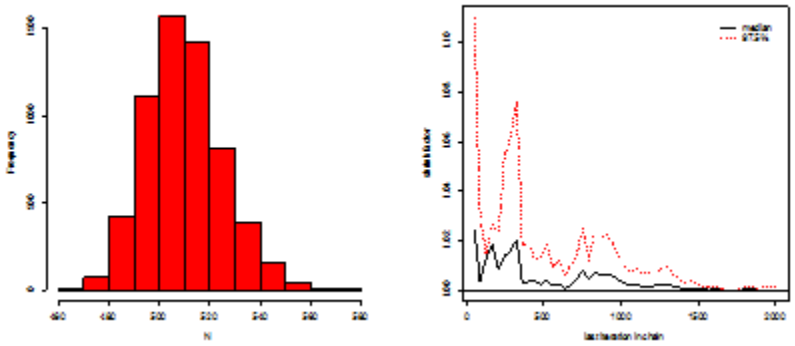

Figura 5.4.2- Histograma da distribuição *a posteriori* de N e gráfico do critério da convergência de Gelman Rubin das cadeias de N, para $\delta = 0,5$

Notamos, pelos dados das tabelas 5.4.1 e 5.4.2, que os valores atribuídos ao parâmetro δ sensibilizam as estimativas de Bayes de n_{12} e consequentemente, de N. No exemplo notamos que o valor $\delta = 0,5$ produziu as melhores estimativas. Notamos que para $\delta = 5$ e 10, os respectivos intervalos de credibilidade não contém o verdadeiro valor do parâmetro.

Exemplo 5.4.2 - Neste exemplo os valores atribuídos aos parâmetros foram $N = 1000, \theta_1 = 0,6$ e $\theta_2 = 0,7$. Obtivemos as estatísticas $n_1 = 576, n_2 = 722$ e $n_{12} = 414$, e em seguida, obtivemos as estatísticas $m_{AO} = 16, m_{OB} = 73$ e $m_{AB} = 242$.

Os resumos das distribuições *a posteriori* de n_{12} e N são dados na tabela 5.4.3 e 5.4.4, respectivamente.

Tabela 5.4.3 Estimativas dos resumos da distribuição *a posteriori* de n_{12}

δ	$Média$	$Moda$	Q_1	$Mediana$	Q_3	DP	IC 95%	$Am.IC$
0,1	356,61	331	331	331	331	52,87	(331,534)	203
0,5	394,96	331	340	369	436	65,54	(331;546)	215
1	414,68	362	359	398	462	65,08	(333;553)	220
5	443,76	397	406	438	477	49,53	(365;549)	184
10	444,99	407	416	441	470	40,81	(379;536)	157

Tabela 5.4.4 Estimativas dos resumos da distribuição *a posteriori* de N

δ	$Média$	$Moda$	Q_1	$Mediana$	Q_3	DP	IC 95%	$Am.IC$
0,1	1166,51	1154	1148	1165	1184	27,16	(1117;1222)	105
0,5	1053,95	1042	1040	1053	1067	20,13	(1017;1096)	79
1	1003,33	995	991	1003	1015	17,19	(972;1039)	67
5	937,55	928	928	937	946	13,20	(913;965)	52
10	937,59	937	927	935	944	13,04	(911;963)	52

Como no exemplo anterior, os valores atribuídos ao parâmetro δ sensibilizam as estimativas de Bayes de n_{12} e N. Notamos que o valor $\delta = 1$ produziu as melhores estimativas dos parâmetros. Observamos que fazendo $\delta = 5$ e 10, seus respectivos intervalos de credibilidade não contém o verdadeiro valor de N.

Exemplo 5.4.3 - Neste exemplo os valores atribuídos aos parâmetros foram $N = 100, \theta_1 = 0,2$ e $\theta_2 = 0,4$. Obtivemos as estatísticas $n_1 = 23, n_2 = 45$ e $n_{12} = 10$, e em seguida, obtivemos as estatísticas $m_{AO} = 0, m_{OB} = 2$ e $m_{AB} = 7$.

Os resumos das distribuições *a posteriori* de n_{12} e N são dados na tabela 5.4.5 e 5.4.6,

respectivamente.

Tabela 5.4.5 Estimativas dos resumos da distribuição *a posteriori* de $n_{12} = 10$

δ	$Média$	$Moda$	Q_1	$Mediana$	Q_3	DP	IC 95%	$Am.IC$
0,1	10,74	9	9	9	11	3,39	(9;21)	12
0,5	12,53	9	9	11	15	3,97	(9;22)	13
1	12,84	10	10	12	15	3,8	(9;22)	13
5	12,62	11	11	12	14	2,82	(9;20)	11
10	12,56	11	11	12	14	2,53	(9;19)	10

Tabela 5.4.6 Estimativas dos resumos da distribuição *a posteriori* de N

δ	$Média$	$Moda$	Q_1	$Mediana$	Q_3	DP	IC 95%	$Am.IC$
0,1	100,53	95	85	96	111	21,62	(71;156)	85
0,5	86,33	75	75	83	93	15,42	(66;125)	59
1	84,11	73	74	81	91	14,63	(64;121)	57
5	85,77	71	75	83	93	15,22	(64;123)	59
10	86,33	75	75	83	93	15,42	(66;125)	59

Como nos exemplos anteriores a escolha do valor de δ influencia os valores das estimativas dos parâmetros. Os melhores resultados ocorreram para $\delta = 0,1$.

5.5 Um modelo bayesiano alternativo

Nesta seção apresentamos um modelo bayesiano alternativo, onde reparametrizamos o parâmetro $\mathbf{p}$. Seja ϕ_k a probabilidade de que as fichas k's de um indivíduo sejam preenchidas corretamente em ambas as listas, $k = A$ e B. Isto implica que

$$\begin{aligned} p_1 &= \phi_A(1-\phi_B), \\ p_2 &= (1-\phi_A)\phi_B, \\ p_3 &= \phi_A\phi_B \text{ e} \\ p_4 &= (1-\phi_A)(1-\phi_B), \end{aligned} \tag{5.11}$$

e de (5.2) segue que a função de verossimilhança de n_{12} e $\boldsymbol{\phi} = (\phi_A, \phi_B)$, é tal que

$$\begin{aligned} L(n_{12}, \boldsymbol{\phi}|\ \text{Đ}) &\propto \frac{n_{12}!}{(n_{12}-m_T)!}[\phi_A(1-\phi_B)]^{m_{AO}}[(1-\phi_A)\phi_B)]^{m_{OB}}[\phi_A\phi_B]^{m_{AB}} \times \\ &\quad \times[(1-\phi_A)(1-\phi_B)]^{n_{12}-m_T} \\ &= \frac{n_{12}!}{(n_{12}-m_T)!}\phi_A^{m_{AO}+m_{AB}}\phi_B^{m_{OB}+m_{AB}}(1-\phi_A)^{m_{OB}+n_{12}-m_T} \times \\ &\quad \times(1-\phi_B)^{m_{AO}+n_{(12)}-m_T} \\ &= \frac{n_{12}!}{(n_{12}-m_T)!}\phi_A^{m_A}(1-\phi_A)^{n_{12}-m_A}\phi_B^{m_B}(1-\phi_B)^{n_{12}-m_B}, \end{aligned} \tag{5.12}$$

$m_T \leq n_{12} \leq m$ e $0 < \phi_A, \phi_B < 1$.

Supomos *a priori* n_{12}, ϕ_A e ϕ_B independentes, ϕ_A com distribuição Beta, π_A, com parâmetros α_A^* e β_A^* conhecidos ($\alpha_A^* > 0$ e $\beta_A^* > 0$), ϕ_B com distribuição Beta, π_B, com parâmetros α_B^* e β_B^* conhecidos ($\alpha_B^* > 0$ e $\beta_B^* > 0$) e n_{12} com distribuição $\pi(n_{12})$, cujo suporte é o conjunto $\{1, 2, ..., m\}$. Com isso a distribuição *a priori* conjunta de n_{12} e $\boldsymbol{\phi}$, é dada por

$$\begin{aligned} \pi(n_{12},\, \boldsymbol{\phi}) &= \pi_{_{12}}(n_{12})\pi_A(\phi_A)\pi_B(\phi_B) \\ &= \pi(n_{12})\frac{\Gamma(\alpha_A^*+\beta_A^*)}{\Gamma(\alpha_A^*)\Gamma(\beta_A^*)}\phi_A^{\alpha_A^*-1}(1-\phi_A)^{\beta_A^*-1}\frac{\Gamma(\alpha_B^*+\beta_B^*)}{\Gamma(\alpha_B^*)\Gamma(\beta_B^*)}\phi_B^{\alpha_B^*-1}(1-\phi_B)^{\beta_B^*-1} \\ &\propto \pi(n_{12})\phi_A^{\alpha_A^*-1}(1-\phi_A)^{\beta_A^*-1}\phi_B^{\alpha_B^*-1}(1-\phi_B)^{\beta_B^*-1}, \end{aligned} \tag{5.13}$$

$1 \leq n_{12} \leq m$ e $0 < \phi_A, \phi_B < 1$. Então, de (5.12) e (5.13), segue que a distribuição *a posteriori* conjunta de $\boldsymbol{\phi}$ e n_{12}, é tal que

$$\begin{aligned}
\pi(n_{12},\, \boldsymbol{\phi}|\, \text{Đ}) &\propto L(n_{12},\, \boldsymbol{\phi}|\, \text{Đ})\pi(n_{12},\, \boldsymbol{\phi}) \\
&= \pi(n_{12})\frac{n_{12}!}{(n_{12}-m_T)!}\phi_A^{m_A}(1-\phi_A)^{n_{12}-m_A} \times \phi_B^{m_B}(1-\phi_B)^{n_{12}-m_B} \times \\
&\quad \times \phi_A^{\alpha_A^*-1}(1-\phi_A)^{\beta_A^*-1}\phi_B^{\alpha_B^*-1}(1-\phi_B)^{\beta_B^*-1} \\
&= \pi(n_{12})\frac{n_{12}!}{(n_{12}-m_T)!} \times \phi_A^{m_A+\alpha_A^*-1}(1-\phi_A)^{n_{12}-m_A+\beta_A^*-1} \times \\
&\quad \times \phi_B^{m_B+\alpha_B^*-1}(1-\phi_B)^{n_{12}-m_B+\beta_B^*-1}, \qquad (5.14)
\end{aligned}$$

$m_T \leq n_{12} \leq m$ e $0 < \phi_A, \phi_B < 1$.

Notamos que a distribuição *a posteriori* conjunta de N e $\boldsymbol{\theta}$ é igual a dada em (5.6).

5.6 Distribuições *a priori* não informativas para os parâmetros do modelo

Nesta seção supomos que n_{12} tem distribuição *a priori* uniforme no conjunto $\{1, ..., m\}$, N tem distribuição uniforme nos inteiros positivos e θ_1 e θ_2 tem distribuição Beta com parâmetros $\alpha_1 = \alpha_2 = \beta_1 = \beta_2 = 1$. Logo, de (5.14), segue que a distribuição *a posteriori* conjunta de n_{12} e $\boldsymbol{\phi}$ é tal que

$$\pi(n_{12},\, \boldsymbol{\phi}|\, \text{Đ}) \propto \frac{n_{12}!}{(n_{12}-m_T)!m}\phi_A^{m_A+\alpha_A^*-1}(1-\phi_A)^{n_{12}-m_A+\beta_A^*-1}\phi_B^{m_B+\alpha_B^*-1}(1-\phi_B)^{n_{12}-m_B+\beta_B^*-1}, \qquad (5.15)$$

$m_T \leq n_{12} \leq m$ e $0 < \phi_A, \phi_B < 1$, o que implica que as distribuições condicionais são tais que

$$\pi(n_{12}|\boldsymbol{\phi},\, \text{Đ}) \propto \binom{n_{12}}{m_T}(1-\phi_A)^{n_{12}}(1-\phi_B)^{n_{12}}, \qquad (5.16)$$

$m_T \leq n_{12} \leq m$, e

$$\pi(\phi_k|n_{12}, \phi_w, w \neq k,\, \text{Đ}) \propto \phi_k^{m_k+\alpha_k^*-1}(1-\phi_A)^{n_{12}-m_k+\beta_k^*-1}, \qquad (5.17)$$

$0 < \phi_k < 1$. Isto é, ϕ_k dados $n_{12}, \phi_w, w \neq k$ e Ð, tem distribuição Beta com parâmetros $m_k + \alpha_k^*$ e $n_{12} - m_k + \beta_k^*$.

Novamente, do Capítulo 2 e de (5.6) segue que a distribuição condicional de N, dados $\boldsymbol{\theta}, n_1, n_2$ e $\widehat{n}$ e a distribuição condicional de θ_j, dados $\theta_h, h \neq j, N, n_1, n_2$ e $\widehat{n}$, são as mesmas de (2.16) e (2.17).

5.7 Exemplos com dados simulados

Nesta seção apresentamos exemplos com dados simulados. Na implementação do modelo des-crito acima utilizamos métodos de simulação estocástica MCMC, mais especificamente, o algoritmo *Gibbs sampling* e o algoritmo de busca em tabela estática, para determinarmos os resumos das distribuições marginais *a posteriori* dos parâmetros.

A convergência das cadeias geradas foi verificada através do software CODA, utilizando o critério da convergência de Gelman Rubin.

Consideramos cadeias com dezessete mil elementos. Destas cadeias os três mil elementos iniciais foram descartados como *"burn-in"* e dos restantes foram considerados um em cada seis, para obtermos independência aproximada entre seus elementos.

Nos exemplos com dados simulados atribuímos valores aos parâmetros N, θ_1 e θ_2 e obtivemos as estatísticas n_1, n_2 e n_{12}, gerando uma amostra de tamanho um da distribuição (2.1). Definimos valores para ϕ_A e ϕ_B, em seguida, obtivemos as estatísticas m_{AO}, m_{OB} e m_{AB}, gerando uma amostra de tamanho um da distribuição (5.1).

Os programas utilizados para gerar os resumos *a posteriori* dos parâmetros de interesse foram implementados via software R-gui (versão 1.9.0) e estão desponíveis no apêndice E.

Exemplo 5.7.1 - Neste exemplo os valores atribuídos aos parâmetros foram $N = 500, \theta_1 = 0,7$ e $\theta_2 = 0,5$. Obtivemos as estatísticas , $n_1 = 344, n_2 = 265$ e $n_{12} = 174$. Em seguida, fizemos $\phi_A = 0,81$ e $\phi_B = 0,64$ obtendo as estatísticas $m_A = 137, m_B = 112$ e $m_{AB} = 85$.

Os resumos das distribuições *a posteriori* de n_{12} e N são dados na tabela 5.7.1 e 5.7.2,

respectivamente.

Tabela 5.7.1 Estimativas dos resumos da distribuição *a posteriori* de n_{12}

α_j^*	β_j^*	*Média*	*Moda*	Q_1	*Mediana*	Q_3	*DP*	*IC* 95%	*Am.IC*
1	1	181,57	176	177	181	185	5,99	(171;195)	24
0,5	0,5	181,31	176	177	181	185	6,03	(171;195)	24
0	0	181,05	175	177	180	184	5,93	(171;194)	23
0	1	182,05	176	178	181	186	6,19	(172;196)	24
1	0	180,62	175	176	180	184	5,82	(171;193)	22

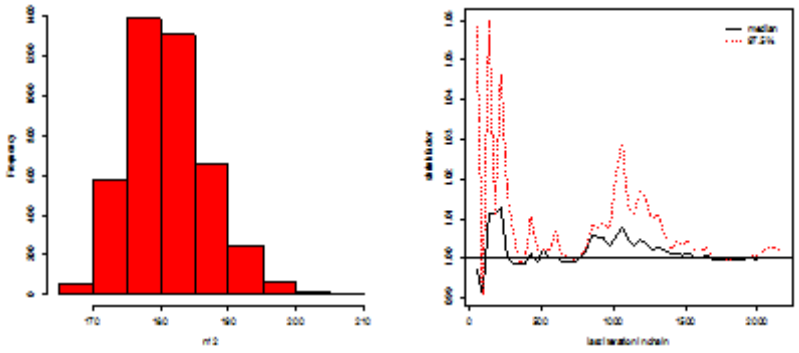

Figura 5.7.1- Histograma da densidade *a posteriori* de n_{12} e gráfico do critério da convergência de Gelman Rubin das cadeias de n_{12}, para $\alpha_j^* = \beta_j^* = 1$

Tabela 5.7.2 Estimativas dos resumos da distribuição *a posteriori* de N

Média	*Moda*	Q_1	*Mediana*	Q_3	*DP*	*IC* 95%	*Am.IC*
503,76	508	493	502	513	14,95	(477;535)	58
504,06	499	493	504	514	14,55	(479;536)	57
504,97	504	495	504	514	14,70	(479;536)	57
502,42	496	492	501	512	14,77	(476;533)	57
506,11	503	495	505	515	15,03	(480;539)	59

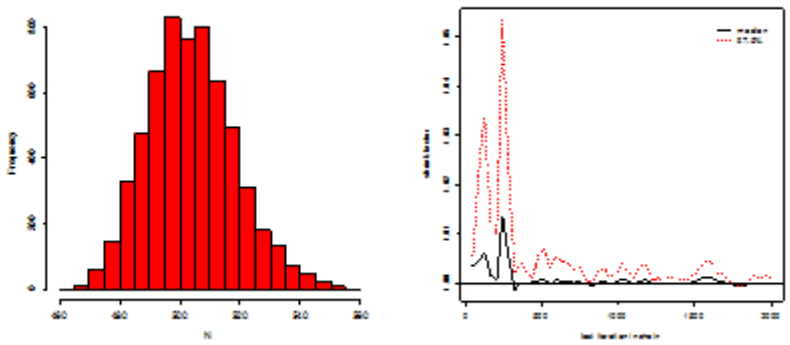

Figura 5.7.2- Histograma da densidade *a posteriori* de N e gráfico do critério da convergência de Gelman Rubin das cadeias de N

Notamos que utilizando *prioris* não informativas obtemos boas estimativas dos parâmetros.

Exemplo 5.7.2 - Neste exemplo os valores atribuídos aos parâmetros foram $N = 1000, \theta_1 = 0,6$ e $\theta_2 = 0,7$. Obtivemos as estatísticas , $n_1 = 576, n_2 = 722$ e $n_{12} = 414$. Em seguida, fizemos $\phi_A = 0,81$ e $\phi_B = 0,81$ obtendo as estatísticas $m_A = 334, m_B = 340$ e $m_{AB} = 272$.

Os resumos das distribuições *a posteriori* de n_{12} e N são dados na tabela 5.7.3 e 5.7.4, respectivamente.

Tabela 5.7.3 Estimativas dos resumos da distribuição *a posteriori* de n_{12}

α_j^*	β_j^*	*Média*	*Moda*	Q_1	*Mediana*	Q_3	*DP*	*IC* 95%	*Am.IC*
1	1	415,04	411	412	415	418	4,71	(407;426)	19
0,5	0,5	414,86	411	412	414	418	4,67	(407;425)	18
0	0	414,68	410	411	414	417	4,66	(407,425)	18
0	1	415,17	411	412	415	418	4,72	(407;426)	19
1	0	414,57	410	411	414	417	4,64	(407;425)	18

Tabela 5.7.4 Estimativas dos resumos da distribuição *a posteriori* de N

Média	*Moda*	Q_1	*Mediana*	Q_3	*DP*	*IC* 95%	*Am.IC*
1003,33	995	991	1003	1015	17,19	(972;1039)	67
1003,77	996	992	1002	1014	14,14	(973;1040)	67
1004,20	1003	992	1003	1015	17,27	(972;1039)	67
1002,77	1006	991	1002	1014	16,90	(971;1038)	67
1004,28	1000	992	1003	1015	17,02	(973;1040)	67

Notamos através das tabelas acima que obtemos boas estimativas dos parâmetros.

Exemplo 5.7.3 - Neste exemplo os valores atribuídos aos parâmetros foram $N = 1000, \theta_1 = 0,2$ *e* $\theta_2 = 0,3$ obtivemos as estatísticas , $n_1 = 313, n_2 = 219$ e $n_{12} = 77$. Em seguida, fizemos $\phi_A = 0,64$ e $\phi_B = 0,81$ obtendo as estatísticas $m_A = 48, m_B = 63$ e $m_{AB} = 38$.

Os resumos das distribuições *a posteriori* de n_{12} e N são dados na tabela 5.7.5 e 5.7.6, respectivamente.

Tabela 5.7.5 Estimativas dos resumos da distribuição *a posteriori* de n_{12}

α_j^*	β_j^*	*Média*	*Moda*	Q_1	*Mediana*	Q_3	*DP*	*IC* 95%	*Am.IC*
1	1	80,54	77	78	80	83	3,93	(74;90)	16
0,5	0,5	80,26	77	77	80	82	3,86	(74;89)	15
0	0	79,96	77	77	79	82	3,78	(74,89)	15
0	1	80,95	77	78	80	83	4,14	(75;91)	16
1	0	79,65	77	77	79	82	3,62	(74;88)	14

Tabela 5.7.6 Estimativas dos resumos da distribuição *a posteriori* de N

Média	*Moda*	Q_1	*Mediana*	Q_3	*DP*	*IC* 95%	*Am.IC*
855,28	825	808	851	897	65,91	(740;100)	260
857	823	810	852	900	65,87	(743;1002)	259
860,67	819	813	857	903	66,62	(743;1005)	264
849,97	840	804	845	890	65,72	(734;992)	258
864,16	835	817	859	906	67,22	(744;1009)	265

Utilizando *prioris* não informativas obtemos estimativas razoáveis dos parâmetros. Notamos que fazendo $\alpha_j^* = 0$ e $\beta_j^* = 1$, seu respectivo intervalo de credibilidade não contém o verdadeiro

valor de N.

5.8 Distribuição *a priori* binomial para o número de indivíduos coincidentes e uniforme para o tamanho populacional

Nesta seção supomos $\theta_1, \theta_2, \phi_A$ e ϕ_B distribuídos como na seção (5.6), mas agora n_{12} tem distribuição *a priori* binomial truncada em zero com parâmetros m e p, p conhecido ou seja, $\pi(n_{12}) = \frac{\binom{m}{n_{12}} p^{n_{12}}(1-p)^{m-n_{12}}}{1-(1-p)^m}$, $n_{12} = 1, 2, ..., m$. Assim sendo, temos de (5.14) que a distribuição *a posteriori* conjunta de n_{12} e $\boldsymbol{\phi}$ é tal que

$$\pi(n_{12}, \boldsymbol{\phi} | \text{ Ð}) \propto \frac{n_{12}!}{(n_{12}-m_T)!}\binom{m}{n_{12}} p^{n_{12}}(1-p)^{m-n_{12}} \phi_A^{m_A+\alpha_A^*-1} \times$$

$$\times (1-\phi_A)^{n_{12}-m_A+\beta_A^*-1} \phi_B^{m_B+\alpha_B^*-1}(1-\phi_B)^{n_{12}-m_B+\beta_B^*-1}, \qquad (5.18)$$

$m_T \leq n_{12} \leq m$ e $0 < \phi_A, \phi_B < 1$.

Novamente, notamos que a distribuição *a posteriori* conjunta de N e $\boldsymbol{\theta}$ é igual a dada em (5.6).

Determinamos na seqüência as distribuições condicionais, que são necessárias para obtermos resumos *a posteriori* dos parâmetros do modelo.

Temos de (5.18) que a distribuição condicional de n_{12}, dados $\boldsymbol{\phi}$ e Ð, é tal que

$$\pi(n_{12}|\boldsymbol{\phi}, \text{ Ð}) \propto \binom{n_{12}}{m_T}\binom{m}{n_{12}}(p(1-\phi_A)(1-\phi_B))^{n_{12}}(1-p)^{-n_{12}}, \qquad (5.19)$$

$m_T \leq n_{12} \leq m$, e que a distribuição condicional de ϕ_k, dados $n_{12}, \phi_w, w \neq k$ e Ð, é tal que

$$\pi(\phi_k|n_{12}, \phi_w, w \neq k, \text{ Ð}) \propto \phi_k^{m_k+\alpha_k^*-1}(1-\phi_A)^{n_{12}-m_k+\beta_k^*-1}, \qquad (5.20)$$

$0 < \phi_k < 1$. Isto é, ϕ_k dados $n_{12}, \phi_w, w \neq k$ e Ð, tem distribuição Beta com parâmetros $m_k + \alpha_k^*$ e $n_{12} - m_k + \beta_k^*$.

Do Capítulo 2 e de (5.6) segue que a distribuição condicional de N, dados $\boldsymbol{\theta}, n_1, n_2$ e $\widehat{n}$ e a distribuição condicional de θ_j, dados $\theta_h, h \neq j, N, n_1, n_2$ e $\widehat{n}$ são as mesmas dadas em (2.16) e (2.17).

5.9 Exemplos com dados simulados

Nesta seção implementamos o modelo como descrito anteriormente na seção 5.7.

Exemplo 5.9.1 - Neste exemplo os valores atribuídos aos parâmetros foram $N = 500, \theta_1 = 0,7$ e $\theta_2 = 0,5$. Obtivemos as estatísticas , $n_1 = 344, n_2 = 265$ e $n_{12} = 174$. Em seguida, fizemos $\phi_A = 0,81$ e $\phi_B = 0,64$ obtendo as estatísticas $m_A = 137, m_B = 112$ e $m_{AB} = 85$.

Os resumos das distribuições *a posteriori* de n_{12} e N são dados na tabela 5.9.1 e 5.9.2, respectivamente.

Tabela 5.9.1 Estimativas dos resumos da distribuição *a posteriori* de n_{12}

α_j^*	β_j^*	p	$Média$	$Moda$	Q_1	$Mediana$	Q_3	DP	IC 95%	$Am.IC$
1	1	0,10	165	164	164	165	166	0,99	(164;167)	3
1	1	0,25	166,93	166	166	167	168	1,67	(164;171)	7
1	1	0,50	172,32	170	170	172	174	2,72	(167;178)	11
1	1	0,75	185,48	182	183	185	188	4,04	(178;194)	16
1	1	0,90	209,21	206	206	209	213	4,95	(200;219)	19

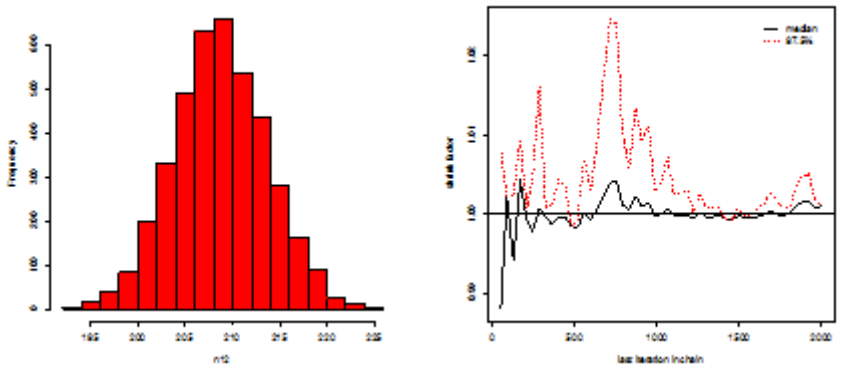

Figura 5.9.1- Histograma da densidade *a posteriori* de n_{12}, gráfico do critério da convergência de Gelman Rubin das cadeias de n_{12}, para $\alpha_j^* = \beta_j^* = 1$ e $p = 0,5$

Tabela 5.9.2 Estimativas dos resumos da distribuição *a posteriori* de N

p	$Média$	$Moda$	Q_1	$Mediana$	Q_3	DP	$IC\ 95\%$	$Am.IC$
0,10	554,18	559	541	553	566	19,31	(520;596)	76
0,25	547,96	533	534	547	560	19,02	(514;588)	74
0,50	530,70	523	519	530	542	17,15	(500;568)	68
0,75	492,88	488	483	492	501	13,66	(470;522)	52
0,90	436,65	426	430	435	442	8,73	(422;456)	34

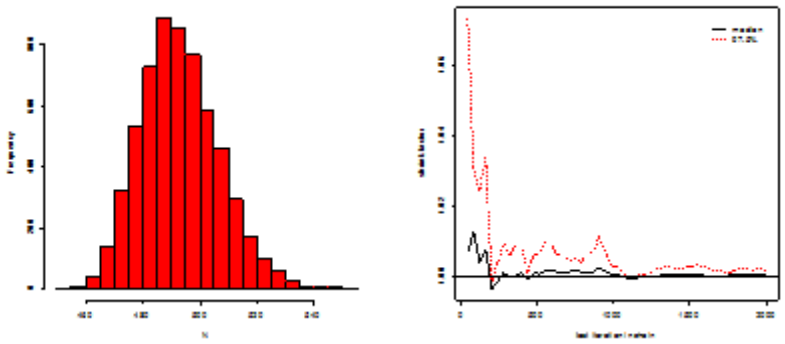

Figura 5.9.2- Histograma da densidade *a posteriori* de N e gráfico do critério da convergência de Gelman Rubin das cadeias de N

Os valores atribuídos ao parâmetro p sensibilizam as estimativas de Bayes de n_{12} e consequentemente, de N. No exemplo, notamos que o valor $p = 0,75$ produziu os melhores resultados. Através da tabela acima verificamos que para $p = 0,1$ e 0,25 seus respectivos intervalos de credibilidade não contém o verdadeiro valor do parâmetro N.

Exemplo 5.9.2 - Neste exemplo os valores atribuídos aos parâmetros foram $N = 1000, \theta_1 = 0,6$ e $\theta_2 = 0,7$. Obtivemos as estatísticas, $n_1 = 576, n_2 = 722$ e $n_{12} = 414$. Em seguida, fizemos $\phi_A = 0,81$ e $\phi_B = 0,81$ e obtivemos as estatísticas $m_A = 334, m_B = 340$ e $m_{AB} = 272$.

Os resumos das distribuições *a posteriori* de n_{12} e N são dados na tabela 5.9.3 e 5.9.4,

respectivamente.

Tabela 5.9.3 Estimativas dos resumos da distribuição *a posteriori* de n_{12}

α_j^*	β_j^*	p	$Média$	$Moda$	Q_1	$Mediana$	Q_3	DP	IC 95%	$Am.IC$
1	1	0,10	401,72	401	401	402	402	1,28	(400;405)	5
1	1	0,25	405,11	403	404	405	407	2,17	(401;410)	9
1	1	0,50	414,54	412	412	414	417	3,59	(408;422)	14
1	1	0,75	437,42	439	434	437	441	5,32	(427;448)	21
1	1	0,90	478,80	474	474	479	483	6,53	(466;492)	26

Tabela 5.9.4 Estimativas dos resumos da distribuição *a posteriori* de N

p	$Média$	$Moda$	Q_1	$Mediana$	Q_3	DP	IC 95%	$Am.IC$
0,10	1036,19	1039	1023	1035	1048	19,01	(1001;1076)	75
0,25	1027,97	1019	1015	1027	1040	18,64	(994;1067)	73
0,50	1004,56	1007	992	1003	1015	17,18	(973;1039)	66
0,75	951,79	948	942	951	961	14,08	(927;981)	54
0,90	869,43	871	863	869	875	9,46	(852;889)	37

Os valores atribuídos ao parâmetro p sensibilizam as estimativas de Bayes de n_{12} e de N. Neste exemplo, o valor $p = 0,5$ produziu os melhores resultados. Observamos que fazendo $p = 0,1; 0,75$ e 0,90 seus respectivos intervalos de credibilidade não contém o verdadeiro valor do parâmetro N.

Exemplo 5.9.3 - Neste exemplo os valores atribuídos aos parâmetros foram $N = 1000, \theta_1 = 0,2$ e $\theta_2 = 0,3$. Obtivemos as estatísticas , $n_1 = 313, n_2 = 219$ e $n_{12} = 77$. Em seguida, atribuímos $\phi_A = 0,64$ e $\phi_B = 0,81$ e obtivemos as estatísticas $m_A = 48, m_B = 63$ e $m_{AB} = 38$.

Os resumos das distribuições *a posteriori* de n_{12} e N são dados na tabela 5.9.5 e 5.9.6,

respectivamente.

Tabela 5.9.5 Estimativas dos resumos da dsitribuição *a posteriori* de n_{12}

α_j^*	β_j^*	p	$Média$	$Moda$	Q_1	$Mediana$	Q_3	DP	IC 95%	$Am.IC$
1	1	0,10	74,46	74	73	74	74	1,19	(73;77)	4
1	1	0,25	77,23	76	76	77	79	2	(74;82)	8
1	1	0,50	85,03	83	83	85	87	3,27	(79;92)	13
1	1	0,75	103,96	100	101	104	107	4,78	(95;114)	19
1	1	0,90	138,35	138	134	138	142	5,91	(127;150)	23

Tabela 5.9.6 Estimativas dos resumos da dsitribuição *a posteriori* de N

p	$Média$	$Moda$	Q_1	$Mediana$	Q_3	DP	IC 95%	$Am.IC$
0,10	924,02	890	870	919	972	76,07	(788;1088)	300
0,25	891,37	860	842	887	937	70,61	(764;1040)	276
0,50	809,76	784	768	806	846	59,73	(704;938)	234
0,75	661,54	654	634	659	686	38,46	(595;743)	148
0,90	497,93	490	484,7	496,7	510,7	19,56	(468;540)	72

Como nos exemplos anteriores a escolha do valor de p influencia os valores das estimativas dos parâmetros. Os melhores resultados ocorreram para $p = 0,10$. Para $p = 0,50; 0,75$ e 0,90, seus respectivos intervalos de credibilidade para N não contém seu verdadeiro valor.

5.10 Exemplo com dados reais

Nesta seção aplicamos a metodologia desenvolvida no Capítulos 5 para o exemplo com dados reais, dado por Saber et al (2000) e que foi citada por (Micheletti, 2003).

Exemplo 5.10.1 Neste exemplo obtemos estimativas para $n_{(12)}$ e N através de dados reais obtidos de um estudo realizado no Sul da Nova Zelândia (Seber et al, 2000). Foram identificados na primeira lista 4186 diabéticos junto a médicos da região e 2203 na segunda lista, por um estudo caseiro. Para cada indivíduo registrou-se as seguintes informações: primeiro nome, sobrenome, idade, data de nascimento, sexo, rua e bairro. Eles incluíram 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 $m_{AB} = 116$, $m_A = 298$ e $m_B = 231$.

Apresentamos nas tabelas 5.10.1 e 5.10.2 os resumos da distribuição *a posteriori* de n_{12} utilizando o modelo bayesiano, onde $\mathbf{p}$ tem distribuição de Dirichlet com parâmetro δ e, em seguida, estimamos N como descrito anteriormente na seção 5.2 e 5.3.

Tabela 5.10.1 Estimativas dos resumos da distribuição *a posteriori* de n_{12}

δ	$Média$	$Moda$	Q_1	$Mediana$	Q_3	DP	IC 95%	$Am.IC$
1	588,17	587	466	545,5	681	149,69	(418;944)	526

Tabela 5.10.2 Estimativas dos resumos da distribuição *a posteriori* de N

$Média$	$Moda$	Q_1	$Mediana$	Q_3	DP	IC 95%	$Am.IC$
15685,47	15422	15336	15672	16025	516,45	(14703;16728)	2025

Nas tabelas 5.10.3 e 5.10.4 apresentamos os resumos da distribuição *a posteriori* de n_{12} e N, utilizando o modelo bayesiano para $\boldsymbol{\phi}$ e *a priori* uniforme para n_{12}, como descrito na seção 5.6.

Tabela 5.10.3 Estimativas dos resumos da distribuição *a posteriori* de n_{12}

α_j^*	β_j^*	$Média$	$Moda$	Q_1	$Mediana$	Q_3	DP	IC 95%	$Am.IC$
1	1	595,61	589	575	594	614	29,72	(542;660)	118
0,5	0,5	595,84	590	576	594	614	29,87	(542;660)	118

Tabela 5.10.4 Estimativas dos resumos da distribuição *a posteriori* de N

$\alpha_j^*= \beta_j^*$	$Média$	$Moda$	Q_1	$Mediana$	Q_3	DP	IC 95%	$Am.IC$
1	15488,6	15400	15144	15470	15821	502	(14547;16505)	1958
0,5	15483,5	15492	15144	15470	15804	501	(14551;16510)	1959

Apresentamos nas tabelas 5.10.5 e 5.10.6 os resumos das distribuições *a posteriori* de n_{12} e N, utilizando o modelo bayesiano para $\boldsymbol{\phi}$ e *a priori* binomial truncada em zero para n_{12}, como descrito na seção 5.8.

Tabela 5.10.5 Estimativas dos resumos da distribuição *a posteriori* de n_{12}

α_j^*	β_j^*	p	$Média$	$Moda$	Q_1	$Mediana$	Q_3	DP	IC 95%	$Am.IC$
1	1	0,75	538	531	531	538	544	9,87	(519;558)	39

Tabela 5.10.6 Estimativas dos resumos da distribuição *a posteriori* de N

p	$Média$	$Moda$	Q_1	$Mediana$	Q_3	DP	IC 95%	$Am.IC$
0,75	17141,2	17141	16011	16722	17540	604,2	(16011;18367)	2356

Micheletti obteve as seguintes estimativas para n_{12}. Estimativas de máxima verossimilhança: $\widehat{n_{12}} \cong 593$ e $\widehat{N} \cong 15551$; intervalo de confiança de 95%: (14549; 16700). Notamos que os resultados obtidos por Micheletti são similares aos obtidos pelos dois modelos bayesianos utilizados.

5.10.1 Conclusão

Comparando os modelos utilizados acima, através dos exemplos com dados simulados, notamos que, utilizando o modelo da seção 5.5 obtemos estimativas mais próximas dos verdadeiros valores dos parâmetros, intervalos de credibilidade com amplitudes menores do que os modelos das seções 5.2 e 5.8.

5.11 Proposta para uma Pesquisa Futura

Como visto nesta dissertação, apresentamos uma solução bayesiana para estimação do número de diabéticos de uma população, através de listas de pacientes.

Uma proposta para uma pesquisa futura seria realizar um estudo de um modelo bayesiano hierárquico pleno para os modelos apresentados nos Capítulos 2 e 3. Além disso, poderíamos pensar em generalizar o modelo com erros apresentado no Capítulo 5 para três ou mais listas e, finalmente, poderíamos adotar uma metodologia bayesiana supondo n_{12} uma variável latente.

Apêndices

5.12 A - Programa para gerar valores para n_1, n_2 e n - utilizando a distribuição multinomial via software R

```
set.seed(50)
    #Dados
    N<-100; theta1<-0.08; theta2<-0.05
    p1<-theta1*(1-theta2); p2<-(1-theta1)*theta2; p3<-theta1*theta2; p4<-(1-theta1)*(1-theta2)
    x<-numeric()
    p.out<-p1+p2+p3+p4
    n1<-n2<-n12<-N0<-0
    for (i in 1:N)
    {
    x[i]<-runif(1)
    if ((x[i] >= 0) && (x[i]<= p1))
    {
    n1<-n1+1
    }
    if ((x[i] > p1) && (x[i]<= (p1+p2)))
    {
    n2<-n2+1
    }
    if ((x[i] > (p1+p2)) && (x[i]<= (p1+p2+p3)))
    {
    n12<-n12+1
    }
```

```
if ((x[i] > (p1+p2+p3)) && (x[i]<= 1))
{
N0<-N0+1
}
}
n1.out<-n1+n12; n2.out<-n2+n12; n<-n1+n2+n12
n1.out; n2.out; n
```

5.13 B - Implementação do algoritmo *Gibbs sampling* utilizando *a priori* uniforme para N

```
set.seed(50)
bi<-10000 # burn-in
na<-30000 # tamanho da amostra a ser gerada
s<-10 # salto entre os valores amostrados para obter indep.
alpha1<- # chute do valor de alpha1
alpha2<-   ; beta1<-  ; beta2<-
n1<-      ; n2<-    ; n<-
#valores iniciais
theta1o<-c(0.2,0.5,0.8); theta2o<-c(0.2,0.5,0.8); No<-c(800,1000,1500)
theta1<-matrix(0,na,length(theta1o)); theta2<-matrix(0,na,length(theta2o)); N<-matrix(0,na,length(No))
theta1.out<-theta2.out<-N.out<-numeric()
N[1,]<-(n+rnbinom(3,n+1,(1-((1-theta1o)*(1-theta2o)))))
theta1[1,]<-rbeta(3,n1+alpha1,N[1,]-n1+beta1)
theta2[1,]<-rbeta(3,n2+alpha2,N[1,]-n2+beta2)
for      (i in 2:na)
{
     N[i,]<-(n+rnbinom(3,n+1,(1-((1-theta1[i-1,])*(1-theta2[i-1,])))))
     theta1[i,]<-rbeta(3,n1+alpha1,N[i,]-n1+beta1)
     theta2[i,]<-rbeta(3,n2+alpha2,N[i,]-n2+beta2)
}
for (k in 1:na)
{
```

```
        if ((k > bi) && ((k-bi) %% s) == 0)
        {
            theta2.out<-rbind(theta2.out,theta2[k,])
            theta1.out<-rbind(theta1.out,theta1[k,])
            N.out<-rbind(N.out,N[k,])
        }
    }
    m<-(na-bi)/s
    N1.out<-N2.out<-N3.out<-theta11.out<-theta12.out<-theta13.out<-theta21.out<-theta22.out<-theta23.out<-
rep(0,2000)
    for (k in 1:2000)
    {
        theta21.out[k]<-theta2.out[k]
        theta11.out[k]<-theta1.out[k]
        N1.out[k]<-N.out[k]
    }
    for (k in 2001:4000)
    {
        theta22.out[k-2000]<-theta2.out[k]
        theta12.out[k-2000]<-theta1.out[k]
        N2.out[k-2000]<-N.out[k]
    }
    for (k in 4001:6000)
    {
        theta23.out[k-4000]<-theta2.out[k]
        theta13.out[k-4000]<-theta1.out[k]
        N3.out[k-4000]<-N.out[k]
    }
    ##############################################
    theta1.out<-c(theta1.out[,1],theta1.out[,2],theta1.out[,3])
    theta2.out<-c(theta2.out[,1],theta2.out[,2],theta2.out[,3])
    quantile(theta1.out,c(0.025,0.25,0.50,0.75,0.975))
    quantile(theta2.out,c(0.025,0.25,0.50,0.75,0.975))
```

```
mean(theta1.out);mean(theta2.out)
N.out<-c(N.out[,1],N.out[,2],N.out[,3]) #moda
a<-max(N.out); b<-a-n+1; freq<-rep(0,b)
for(j in 1:b) freq[N.out[j]-n+1]<-freq[N.out[j]-n+1]+1
m<-1
for(j in 2:b) if(freq[j]>freq[m]) m<-j
moda<-n+m-1
moda
quantile(N.out,c(0.025,0.25,0.50,0.75,0.975))
mean(N.out); sd(N.out)
hist(N.out,main="Densidade a posteriori de N ",col="red",xlab="N")
### diagósticos de convergencia
library(coda)
N1m<-mcmc(N1.out); N2m<-mcmc(N2.out); N3m<-mcmc(N3.out)
N123m<-mcmc.list(N1m,N2m,N3m)
summary(N123m); traceplot(N123m); gelman.diag(N123m); gelman.plot(N123m); autocorr(N123m);
autocorr.plot(N123m)
```

5.13.1 B.1 - Implementação do algoritmo *Gibbs sampling* utilizando *a priori* de Jeffreys para N

O programa é o mesmo utilizado no apêndice B com a seguinte alteração:

```
N[1,]<-(n+rnbinom(3,n,(1-((1-theta1o)*(1-theta2o)))))
theta1[1,]<-rbeta(3,n1+alpha1,N[1,]-n1+beta1)
theta2[1,]<-rbeta(3,n2+alpha2,N[1,]-n2+beta2)

for     (i in 2:na)
{
    N[i,]<-(n+rnbinom(3,n,(1-((1-theta1[i-1,])*(1-theta2[i-1,])))))
    theta1[i,]<-rbeta(3,n1+alpha1,N[i,]-n1+beta1)
    theta2[i,]<-rbeta(3,n2+alpha2,N[i,]-n2+beta2)
}
```

5.13.2 B.2 - Implementação do algoritmo *Gibbs sampling* utilizando *a priori* de Poisson para N

O programa é o mesmo utilizado no apêndice B com a seguinte alteração:

```
N[1,]<-(n+rpois(3,(lambda*(1-theta1o)*(1-theta2o))))
theta1[1,]<-rbeta(3,n1+alpha1,N[1,]-n1+beta1)
theta2[1,]<-rbeta(3,n2+alpha2,N[1,]-n2+beta2)

for     (i in 2:na)
{
    N[i,]<-(n+rpois(3,(lambda*(1-theta1[i-1,])*(1-theta2[i-1,]))))
    theta1[i,]<-rbeta(3,n1+alpha1,N[i,]-n1+beta1)
    theta2[i,]<-rbeta(3,n2+alpha2,N[i,]-n2+beta2)
    }
```

5.14 C - Gerador da distribuição Beta adequada método subjetivo via software R

```
a<-0.05; b<-0.15
    c<-a+b; d<-b-a
    alpha<-(c/(2*(d^2)))*((8*c)-(4*(c^2))-(d^2))
    beta<-((2-c)/(2*(d^2)))*((8*c)-(4*(c^2))-(d^2))
    almbe<-alpha+beta
    alvbe<-alpha*beta
    media<-alpha/almbe
    variancia<-((alvbe)/((almbe^2)*(almbe+1)))^0.5
    x<-seq(0,1,0.001)
    y<-dbeta(x,alpha,beta)
    a1<-rep(a,10); b1<-rep(b,10)
    a2<-seq(0,90,10); b2<-seq(0,90,10)
    plot(x,y,type="o",col="darkblue",main="Distribuição Beta (alpha= 22.7,beta= 450)",xlab="theta1")
    lines(a1,a2,col="red"); lines(b1,b2,col="orange")
    legend(0.60,27,c("a1 = 0.02","b1 = 0.08"),col=c("red","orange"),lty=1)
    alpha; beta; media; variancia
```

5.15 D - Programa para geração de n_{12} e p via *Gibbs sampling* e busca em tabela estática utilizando *a priori* de Dirichlet para p

```
#definições da cadeia
    burn<-30000; tamanho<-430000; salto<-200
    ########
    #dirichilet
    set.seed(22222)
    x<-numeric()
    l<-4; alpha<-1; beta<-1
    x<-rgamma(l,alpha,beta)
    somax<-sum(x)
    x<-x/somax
    n12<-414
    y<-numeric()
    y<-rmultinom(1,n12,x)
    ####################
    # parâmetros e estatísticas
    set.seed(54321)
    l<-4; alpha<-0.1; beta<-1, ni<-y[,1]
    n<-sum(ni)-ni[l]
    ######################
    ### valores iniciais do Gibbs
    po<-0.25
    p<-rep(0,tamanho); n12<-rep(0,tamanho); p.out<-n12.out<-numeric()
    x<-matrix(0,l,3)
    na<-tamanho
    ########################
    # primeira iteração do gibbs
    ########################
    ###Valores das estatísticas
    n1<-28; n2<-60;
```

```
m<-min(n1,n2)
mao<-ni[1]; mob<-ni[2]; mab<-ni[3]
mT<-mao+mob+mab
c<-1
n12.test<-seq(mT,m,1)
lfx<-rep(0,length(n12.test))
fx<-rep(0,length(n12.test))
fx.out<-rep(0,tamanho)
for (i in 1:length(n12.test))
{
    lfx[i]<-log(choose(n12.test[i],mT))+n12.test[i]*log(po)
}
lfx<-lfx-max(lfx)
fx<-exp(lfx)
fx.out<-fx/sum(fx)
n12.trans<-rep(0,length(fx.out))
n12.trans[1]<-fx.out[1]
for (i in 2:length(fx.out))
{
    n12.trans[i]<-n12.trans[i-1]+fx.out[i]
}
n12.f<-0
prop<-0
uniforme<-runif(1)
if(uniforme < n12.trans[1])
    {
        prop<-n12.trans[1]
        n12.f<-n12.test[1]
        break
    }
for (j in 2:length(fx.out))
{
    if((n12.trans[j-1] <= uniforme) && (uniforme < n12.trans[j]))
```

```
    {
        prop<-n12.trans[j]
        n12.f<-n12.test[j]
        break
    }
}
n12[1]<-n12.f
n12[1]
############
## Dirichilet
############
for(j in 1:(l-1))
{
x[j]<-rgamma(1,alpha+ni[j],beta)
}
x[l]<-rgamma(1,alpha+(n12[1]-mT),beta)
somax<-sum(x)
x<-x/somax
p[1]<-x[l]
###############
### Loop Gibbs
###############
cont<-1
for     (i in 2:tamanho)
{
###
###busca em tabela estática para n12
    lfx<-rep(0,length(n12.test))
    fx<-rep(0,length(n12.test))
    fx.out<-rep(0,na)
    for (w in 1:length(n12.test))
    {
        lfx[w]<-log(choose(n12.test[w],mT))+n12.test[w]*log(p[i-1])
```

```
}
lfx<-lfx-max(lfx)
fx<-exp(lfx)
fx.out<-fx/sum(fx)
n12.trans<-rep(0,length(fx.out))
n12.trans[1]<-fx.out[1]
for (k in 2:length(fx.out))
{
    n12.trans[k]<-n12.trans[k-1]+fx.out[k]
}
n12.f<-0
prop<-0
uniforme<-runif(1,0,1)
if(uniforme < n12.trans[1])
        {
            prop<-n12.trans[1]
            n12.f<-n12.test[1]

        }
for (j in 2:length(fx.out))
{
    if((n12.trans[j-1] <= uniforme) && (uniforme < n12.trans[j]))
    {
        prop<-n12.trans[j]
        n12.f<-n12.test[j]
        break
    }
}
n12[i]<-n12.f
##### gibbs para dirichlet
for(j in 1:(l-1))
{
    x[j]<-rgamma(1,alpha+ni[j],beta)
```

```
        }
        x[l]<-rgamma(1,alpha+(n12[i]-mT),beta)
        somax<-sum(x)
        x<-x/somax
    p[i]<-x[l]
    cont<-cont+1
    print(cont)
    }
    ##################################################
    for (k in 1:tamanho)
    {
        if ((k > burn) && ((k-burn) %% salto) == 0)
        {
            p.out<-rbind(p.out,p[k])
            n12.out<-rbind(n12.out,n12[k])
        }
    }
    a11<-n12.out; a21<-p.out
    n<-mT
    n12.final<-c(a11,a12) #moda
    a<-max(n12.final)
    b<-a-n+1
    freq<-rep(0,b)
    for(j in 1:b) freq[n12.final[j]-n+1]<-freq[n12.final[j]-n+1]+1
    m<-1
    for(j in 2:b) if(freq[j]>freq[m]) m<-j
    moda<-n+m-1
    moda
```

5.16 E - Implementação estimação de n_{12} utilizando o parâmetro ϕ

```
set.seed(100)
```

```
bi<-5000 # burn-in
na<-18000 # tamanho da amostra a ser gerada
s<-6 # salto entre os valores amostrados para obter indep.
###Valores das estatísticas
alpha1<-1; alpha2<-1; beta1<-1; beta2<-1
n1<-576; n2<-722
m<-min(n1,n2)
ma<-334; mb<-340; mab<-274
mT<-ma+mb-mab
c<-1
#valores iniciais
fi1o<-0.7; fi2o<-0.7
fi1<-rep(0,na); fi2<-rep(0,na); n12<-rep(0,na)
fi1.out<-fi2.out<-n12.out<-numeric()
n12.test<-seq(mT,m,1)
lfx<-rep(0,length(n12.test))
fx<-rep(0,length(n12.test))
fx.out<-rep(0,na)
for (i in 1:length(n12.test))
{
    lfx[i]<-log(choose(n12.test[i],mT))+n12.test[i]*log(1-fi1o)+n12.test[i]*log(1-fi2o)
}
lfx<-lfx-max(lfx)
fx<-exp(lfx)
fx.out<-fx/sum(fx)
n12.trans<-rep(0,length(fx.out))
n12.trans[1]<-fx.out[1]
for (i in 2:length(fx.out))
{
    n12.trans[i]<-n12.trans[i-1]+fx.out[i]
}
n12.f<-0; prop<-0
uniforme<-runif(1)
```

```
if(uniforme < n12.trans[1])
    {
        prop<-n12.trans[1]
        n12.f<-n12.test[1]
        break
    }
for (j in 2:length(fx.out))
{
    if((n12.trans[j-1] <= uniforme) && (uniforme < n12.trans[j]))
    {
        prop<-n12.trans[j]
        n12.f<-n12.test[j]
        break
    }
}
n12[1]<-n12.f
fi1[1]<-rbeta(1,ma+alpha1,n12[1]-ma+beta1)
fi2[1]<-rbeta(1,mb+alpha2,n12[1]-mb+beta2)
n12[1];fi1[1];fi2[1]
cont<-1
for     (i in 2:na)
{
    set.seed(i)
    lfx<-rep(0,length(n12.test))
    fx<-rep(0,length(n12.test))
    fx.out<-rep(0,na)
    for (w in 1:length(n12.test))
    {
        lfx[w]<-log(choose(n12.test[w],mT))+n12.test[w]*log(1-fi1[i-1])+n12.test[w]*log(1-fi2[i-1])
    }
    lfx<-lfx-max(lfx)
    fx<-exp(lfx)
    fx.out<-fx/sum(fx)
```

```
n12.trans<-rep(0,length(fx.out))
n12.trans[1]<-fx.out[1]
for (k in 2:length(fx.out))
{
    n12.trans[k]<-n12.trans[k-1]+fx.out[k]
}
n12.f<-0
prop<-0
uniforme<-runif(1,0,1)
if(uniforme < n12.trans[1])
        {
            prop<-n12.trans[1]
            n12.f<-n12.test[1]

        }
for (j in 2:length(fx.out))
{
    if((n12.trans[j-1] <= uniforme) && (uniforme < n12.trans[j]))
    {
        prop<-n12.trans[j]
        n12.f<-n12.test[j]
        break
    }
}
n12[i]<-n12.f
fi1[i]<-rbeta(1,ma+alpha1,n12[i]-ma+beta1)
fi2[i]<-rbeta(1,mb+alpha2,n12[i]-mb+beta2)
cont<-cont+1
print(cont)
}
for (k in 1:na)
{
    if ((k > bi) && ((k-bi) %% s) == 0)
```

```
        {
            fi2.out<-rbind(fi2.out,fi2[k])
            fi1.out<-rbind(fi1.out,fi1[k])
            n12.out<-rbind(n12.out,n12[k])
        }
    }
    n12.out1<-n12.out
```

5.17 F - Programa para estimação de n_{12} via *Gibbs sampling* e busca em tabela estática, utilizando *a priori* binomial truncada em zero para n_{12}

O programa é o mesmo utilizado no apêndice B com a seguinte alteração:

```
    set.seed(60)
    bi<-5000        # burn-in
    na<-17000       # tamanho da amostra a ser gerada
    s<-6        # salto entre os valores amostrados para obter indep.
    p<-0.90
    for (i in 1:length(n12.test))
    {
        lfx[i]<-log(choose(n12.test[i],mT))+log(choose(m,n12.test[i]))+n12.test[i]*log(p)+n12.test[i]*log(1-
fi1o)+n12.test[i]*log(1-fi2o)-(n12.test[i])*log(1-p)
    }
        for (w in 1:length(n12.test))
        {
    lfx[w]<-log(choose(n12.test[w],mT))+log(choose(m,n12.test[w]))+n12.test[w]*log(p)+n12.test[w]*log(1-
fi1o)+n12.test[w]*log(1-fi2o)-(n12.test[w])*log(1-p)
        }
```

5.18 G - Implementação *Gibbs sampling* para distribuição *a priori* de Poisson hierárquica

```
set.seed(100)
```

```
bi<-10000
na<-30000
s<-10
alpha1<-1; alpha2<-1; beta1<-1; beta2<-1
n1<-143; n2<-85; n<- # valor de n
c1<-0.0001; d1<-0.0001
#valores iniciais
theta1o<-c(0.5)
theta2o<-c(0.5)
No<-c(1000)
lambdao<-c(1000)
theta1<-matrix(0,na,length(theta1o))
theta2<-matrix(0,na,length(theta2o))
lambda<-matrix(0,na,length(lambdao))
N<-matrix(0,na,length(No))
s2<-theta1.out<-theta2.out<-N.out <-lambda.out<-numeric()
N[1,]<-(n+rpois(1,(lambdao*(1-theta1o)*(1-theta2o))))
theta1[1,]<-rbeta(1,n1+alpha1,N[1,]-n1+beta1)
theta2[1,]<-rbeta(1,n2+alpha2,N[1,]-n2+beta2)
## introduzir o metropolis
int<-0
cont<-0
lambda.test<-0
alpha11<-0
lambda.test<-rgamma(1,N[1,]+c1,d1+1)
u<-runif(1)
alpha11<-min(1,((1-exp(-lambda.test))/(1-exp(-lambdao))))
if (u <= alpha11)
{
cont<-cont+1
lambda[1,]<-lambda.test
}
else
```

```
 {
lambda[1,]<-lambdao
 }
 int<-int+1
 print(cont)
 print(int)

 for (i in 2:na)
 {
 N[i,]<-(n+rpois(1,(lambda[i-1,]*(1-theta1[i-1,])*(1-theta2[i-1,]))))
 theta1[i,]<-rbeta(1,n1+alpha1,N[i,]-n1+beta1)
 theta2[i,]<-rbeta(1,n2+alpha2,N[i,]-n2+beta2)
 ## introduzir o metropolis
 lambda.test<-0
 alpha11<-0
 lambda.test<-rgamma(1,N[i,]+c1,d1+1)
 u<-runif(1)
 alpha11<-min(1,((1-exp(-lambda.test))/(1-exp(-lambda[i-1,]))))
 if (u <= alpha11)
 {
cont<-cont+1
lambda[i,]<-lambda.test
 }
 else
 {
lambda[i,]<-lambda[i-1,]
 }
 int<-int+1
 print(cont)
 print(int)
}
```

Referências Bibliográficas

[1] BEST, N. G.; COWLES, M. K.; VINES, S. K. *CODA manual version 0.30.* MRC, Cambridge, UK: Biostatistics Unit, 1995.

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

[3] BOX, G.E.P. Bayesian inference in statistical analysis. New York: John Wiley and Sons, 1992.

[4] CAMPOS, J. J. B.; ALMEIDA, H. G. G.; IOCHIDA, L. C. Incidência de diabetes mellitus dependente (tipo I) na cidade de Londrina, PR-Brasil. *Arq. Bras. Endocrinal. Metab.,* n. 42, p. 36-44, 1998.

[5] CASELLA, G.; GEORGE, E. Explaining the Gibbs Sampler. *Amer. Statistician,* v. 46, p. 167-174,1992.

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

[7] FIENBERG, S. E. The multiple recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika,* n. 59, p. 591-603, 1972.

[8] FIENBERG, S. E.; JOHNSON, M. S.; JUNKER, B. W. Classical multilevel and bayesian approaches to population size estimation using multiple lists. *J. R. Statist. Soc.*, v. 162A,n. 3, p. 383-405, 1999.

[9] GELMAN, A. et al. *Bayesian data analysis.* London: Chapman and Hall, 2000.

[10] GEORGE, E. I.; ROBERT, C. P. Capture-recapture estimation via Gibbs sampling. *Biometrika*, v.79, n. 4, p. 677-83, 1992.

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

[12] LEE, A.J. et al. Capture-recapture, epidemiology and list mismatches: several lists. *Biometrics*, n. 57, p. 707-713, 2001.

[13] LEITE, J. G.; OISHI, J.; PEREIRA, C. A. B. Anote on the exact maximum likelihood estimation of the size of a finite and closed population. *Biometrika*, n. 75, p. 178-180, 1988.

[14] LEITE, J. G.; SINGER J.M., *Métodos assintótico em estatística fundamentos e aplicações.* São Paulo: Associação Brasileira de Estatística - ABE, 1990.

[15] MICHELETTI, L. R. *Aplicação da metodologia de verossimilhança na prevalência do diabetes.* 2003. Dissertação de mestrado em estatística, Centro de Ciências Exatas e Tecnologia Departamento de Estatísica, Universidade Federal de São Carlos, 2003.

[16] RAO, C.R. *Linear statistical inference and its applications.* 2. ed., New York: John Wiley, 1973.

[17] SEBER, G. A. F.; HUAKAU, J. T.; SIMMONS, D. Capture-recapture, epidemiology, and list mismatches: two lists. *Biometrics*, n. 56, p. 1227-1232, 2000.

[18] SEKAR, C.; DEMING, W. E. On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*, n. 44, p. 101-115, 1949.

[19] SMITH, P. J. Bayesian analyses for a multiple capture-recapture model. *Biometrika*, n. 78, p. 399-407, 1991.

[20] WANG, X. Bayesian Analysis of Capture-Recapture Models. 2002. Tese de doutorado em estatística, University of Missouri at Columbia, 2002.

[21] WITTES, J. T. On the bias and estimated variance of Chapman's two-sample capture-recapture population estimate. *Biometrics*, n. 28, p. 592-597, 1972.

[22] WITTES, J. T. Applications of a multinomial capture-recapture method to epidemiological data. *Journal of the American Statistical Association*, n. 69, p. 93-97, 1974.

[23] WITTES, J. T.; SIDEL, V. W. A generalization of the simple capture-recapture model with applications to epidemiological research. *Journal of Chronic Diseases*, n. 21, p. 287-301, 1968.

[24] WITTES, J. T.; COLTON, T.; SIDEL, V. W. Capture-recapture methods for assessing the completeness of cases ascertainment when using multiple information sources. *Journal of Chronic Diseases*, n. 27, p. 25-36, 1974.