UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE MATEMÁTICA, ESTATÍSTICA E COMPUTAÇÃO CIENTÍFICA
DEPARTAMENTO DE ESTATÍSTICA

# Estimação Bayesiana e por Máxima Verossimilhança de Modelos SIR Estocásticos

Rodrigo Bonato Manfredini

Dissertação de Mestrado orientada pelo

Prof. Dr. Luiz Koodi Hotta

ESTIMAÇÃO BAYESIANA E POR MÁXIMA VEROSSIMILHANÇA DE MODELOS SIR ESTOCÁSTICOS

Este exemplar corresponde à redação final da dissertação devidamente corrigida e defendida por Rodrigo Bonato Manfredini e aprovada pela comissão julgadora.

Campinas, 5 de Outubro de 2009

Prof. Dr: Luiz Koodi Hotta
Orientador

Banca Examinadora:

1 Luiz Koodi Hotta (IMECC - Unicamp)
2 Jorge Alberto Achcar (FMRP - USP)
3 Edwin Moises Marcos Ortega (ESALQ-USP)

Dissertação apresentada ao Instituto de Matemática, Estatística e Computação Científica, UNICAMP, como requisito parcial para obtenção do Título de MESTRE em Estatística.

**FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA DO IMECC DA UNICAMP**
Bibliotecária: Maria Fabiana Bezerra Müller – CRB8 / 6162

Manfredini, Rodrigo Bonato
M313e Estimação bayesiana e por máxima verossimilhança de modelos sir estocásticos/Rodrigo Bonato Manfredini -- Campinas, [S.P. : s.n.], 2009.

Orientador : Luiz Koodi Hotta.
Dissertação (mestrado) - Universidade Estadual de Campinas, Instituto de Matemática, Estatística e Computação Científica.

1.Modelagem de dados. 2.Epidemiologia. 3. Método de Monte Carlo. I. Hotta, Luiz Koodi. II. Universidade Estadual de Campinas. Instituto de Matemática, Estatística e Computação Científica. III. Título.

Título em inglês: Bayesian and maximum likelihood estimation of sir stochastic models

Palavras-chave em inglês (Keywords): 1. Data modelling. 2. Epidemics. 3. Monte Carlo method.
Área de concentração: Inferência

Titulação: Mestre em Estatística

Banca examinadora: Luiz Koodi Hotta (IMECC-UNICAMP)
Jorge Alberto Achcar (FMRP-USP)
Edwin Moises Marcos Ortega (ESALQ-USP)

Data da defesa: 13/08/2009

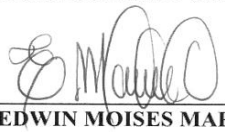
Programa de Pós-Graduação: Mestrado em Estatística

**Dissertação de Mestrado defendida em 13 de agosto de 2009 e aprovada**

**Pela Banca Examinadora composta pelos Profs. Drs.**

**Prof(a). Dr(a). LUIZ KOODI HOTTA**

**Prof(a). Dr(a). JORGE ALBERTO ACHCAR**



**Prof(a). Dr(a). EDWIN MOISES MARCOS ORTEGA**

# *Agradecimentos*

Agradeço à minha família: minha mãe Regina, minha irmã Juliana e meu sobrinho João Lucas por todo apoio que sempre me deram. Agradeço também à minha namorada, Gabriela, por toda a paciência e ajuda que me deu durante o curso de mestrado. Certamente se não fossem eles eu não teria conseguido este título.

Ao meu orientador Prof. Dr. Luiz Koodi Hotta, por todo o seu conhecimento e experiência transmitidos durante a relização da dissertação. À Capes, por ter financiado uma parte do mestrado.

Aos meus amigos em geral, por todo apoio dado durante a realização do curso. Não gostaria de citar nomes, pois a lista é bastante extensa.

# *Resumo*

Os modelos compartimentais têm sido amplamente utilizados para modelar epidemias. Vários métodos têm sido propostos na literatura para estimar os modelos, sendo os mais aplicados, estimadores pelo método de mínimos quadrados, estimadores de máxima verossimilhança e estimadores bayesianos baseados em simulação de Monte Carlo. Na maioria dos casos reais, os dados são apenas parcialmente observáveis. O trabalho considera o caso em que todos os dados são observados e o caso em que apenas os tempos de remoção são disponíveis. As propriedades amostrais dos estimadores de máxima verossimilhança e dos estimadores Bayesianos para dados completos e incompletos são investigadas através de simulação.

# *Abstract*

Compartmental models have been widely used in order to model epidemics. Several methods have been proposed in the literature to estimate the models, specially, the least squares method, maximum likelihood estimation and Bayes estimators based on Monte Carlo simulation. In the most of real cases, the data are only partially observable. The work considers the case that all the data are observed and the case that only the removal times are available. The sampling properties of the maximum likelihood and Bayes estimators for complete and incomplete data are investigated through simulation.

# *Sumário*

# *Lista de Figuras*

# *Lista de Tabelas*

# 1 Introdução

## 1.1 Considerações Iniciais

Esta dissertação de mestrado tem por objetivo estudar a estimação na modelagem de fenômenos epidemiológicos, mais especificamente o modelo SEIR.

A modelagem determinística tem sido uma ferramenta muito importante para se entender a dinâmica da dispersão de doenças infecciosas. Tais doenças podem ser causadas por vírus ou bactérias, como por exemplo: sarampo, varíola, rubéola, catapora, algumas doenças sexualmente transmissíveis, vírus influenza, entre outras. A ferramenta teórica mais comumente utilizada para a modelagem de epidemias é baseada na divisão da população humana (hospedeira) em quatro categorias, contendo suscetíveis, infectados mas não infectantes ainda (expostos), infectantes e removidos. Estes modelos, denominados SEIR, são geralmente expressados como um sistema de equações diferenciais (Anderson e May, 1991), em que as taxas de transição entre os compartimentos são determinadas por parâmetros especificados de acordo com a história natural da doença. Também temos outra abordagem para a modelagem de epidemias: a abordagem estocástica. Esta também é de fundamental importância para se modelar epidemias, principalmente devido ao fato da dispersão de epidemias ser um fenômeno estocástico. Também utiliza-se a modelagem estocástica quando a aproximação determinística não é válida. Por exemplo, quando o tamanho da população e/ou o número de infectados não é suficientemente grande para utilizar a lei dos grandes números.

Algumas outras doenças, apesar de terem algumas características em comum com as doenças infecciosas, não podem ser descritas através destes modelos, como as doenças

causadas por algum parasita. Há outras utilizações destes modelos de epidemia, com algumas modificações, como por exemplo, os trabalhos desenvolvidos nas ciências sociais por Daley e Kendall (1965) e Maki e Thompson (1973).

Há duas características principais que distinguem as doenças infecciosas das doenças não transmissíveis. A primeira, e mais importante característica, é que o indivíduo depende dos outros indivíduos em seu redor para se tornar infectado, enquanto que para doenças não transmissíveis esta dependência geralmente não ocorre. A ocorrência de doenças não transmissíveis é comumente modelada utilizando análise de sobrevivência, em que a função de risco descreve o risco do indivíduo se tornar doente através da idade. Um importante exemplo desta modelagem é a modelagem de riscos proporcionais (Cox, 1972).

A outra característica é que a grande maioria das epidemias é apenas parcialmente observada na prática. Por exemplo, dificilmente se sabe por quem o indivíduo foi infectado, nem o tempo exato em que o indivíduo ficou infeccioso e nem por quanto tempo este indívíduo ficou infectado.

O trabalho está organizado da seguinte forma: Neste primeiro capítulo será apresentada uma introdução a alguns conceitos de epidemiologia, fundamentais para a compreensão da dissertação. Também será apresentado um breve histórico da modelagem de epidemias e serão discutidas algumas vantagens e desvantagens dos modelos estocásticos em comparação com os determinísticos. Em seguida, será apresentado o modelo SEIR. No final do capítulo, uma simulação estocástica e determinística do modelo serão realizadas e as características destas simulações serão discutidas. O capítulo 2 apresentará a estimativa por máxima verossimilhança dos parâmetros do modelo para dados completos e incompletos assim como apresentará uma estimativa Bayesiana para os parâmetros do modelo para dados completos e incompletos. O capítulo 3 tratará de simulações e uma aplicação com dados reais. Por fim, serão abordadas as conclusões finais bem como os trabalhos futuros.

Todos os algoritmos da dissertação foram escritos na linguagem R, exceto em alguns casos em que será comentada a linguagem na qual o programa foi escrito.

## 1.2 Conceitos Fundamentais em Epidemiologia

Primeiramente, alguns conceitos epidemiológicos, fundamentais para o entendimento do projeto, serão definidos:

**Indivíduo Suscetível:** É aquele indivíduo que não teve qualquer contato com o vírus, ou seja, não tem vírus e anticorpos.

**Indivíduo Latente ou Exposto:** É o indivíduo que teve um contato recente com o vírus e este replica-se no interior das células parasitadas, por isso, o indivíduo não pode transmiti-lo. Ou seja, o indivíduo possui vírus e não possui anticorpos. As modelagens utilizadas no projeto não levam em conta a latência do indivíduo, ou seja, o indivíduo passa diretamente do estado suscetível para o infectado.

**Indivíduo Infectante:** É aquele indivíduo que apresenta uma concentração razoável de vírus em seu organismo de tal forma que as células parasitadas expelem o vírus para o meio extracelular e, consequentemente, o indivíduo começa a eliminar o vírus para o meio ambiente. Isto é, os indivíduos infectantes possuem o vírus e os anticorpos em seu organismo.

**Indivíduo Imune:** É aquele indivíduo que não possui o vírus e possui anticorpos. Tal condição pode ser alcançada pelo declínio completo do vírus no meio extracelular devido aos mecanismos de defesa do organismo ou através da indução da produção de anticorpos através de vacinas. Tal imunidade pode ser perene, ou seja, o indivíduo não adquire mais a doença, ou temporária, quando o indivíduo pode voltar a adquirir a doença.

**Doenças de Transmissão Direta:** As infecções de transmissão direta são causadas por vírus ou bactérias em que a disseminação ocorre diretamente, através do meio físico, quando há um contato apropriado entre os indivíduos suscetíveis e infectantes.

**Período de Latência:** É o período em que o indivíduo é dito ser latente ou exposto.

**Período de Imunidade:** Compreende o período em que o indivíduo é dito ser imune. A imunidade pode ou não ser induzida por uma vacina.

**Período de Recuperação:** É o período em que o indivíduo é dito ser infectante.

Então, o indivíduo elimina o vírus através de excreções diversas.

**Endemia:** Nível de infecção estável em uma comunidade.

**Epidemia:** Nível de infecção crescente em uma comunidade.

**Reprodutibilidade Basal:** É o número médio de novos casos secundários que um caso primário é capaz de produzir em uma população totalmente suscetível.

**Tamanho Final da Epidemia:** É definido como o número total de indivíduos que eventualmente contraíram a doença.

## 1.3 Breve Histórico da Modelagem de Epidemias

O primeiro registro de aplicação da matemática no estudo de epidemias é datado de 1760. Neste ano, Daniel Bernoulli utilizou um método matemático para avaliar os efeitos da técnica da variolação no controle da epidemia de varíola. Nesta época, o desenvolvimento de modelos matemáticos aplicados a fenômenos mais amplos era muito limitado devido ao fato da ausência de conhecimento médico sobre os agentes causadores das infecções.

Somente a partir do nascimento da bacteriologia com Louis Pasteur (1822-95) e Robert Koch (1843-1910), e da descoberta dos vírus neste século, foi possível identificar as causas das doenças infecciosas e, consequentemente, aplicar à epidemiologia modelos matemáticos mais gerais e mais próximos da realidade. Assim, um dos primeiros estudos da não linearidade de um modelo epidemiológico pode ser encontrado em Hamer(1906). Ele postulou que o desenvolvimento de uma epidemia depende da taxa de contato entre indivíduos suscetíveis e infecciosos. Tal princípio é denominado princípio da ação de massa. O princípio de Hamer foi formulado considerando o tempo discreto. Dois anos depois, Ross, o generalizou para o tempo contínuo.

O primeiro modelo estocástico foi proposto por McKendrik (1926). Tal modelo é uma versão em tempo contínuo do modelo determinístico de Kermack e McKendrik (1927). Um modelo que recebeu mais atenção naquele momento foi o modelo de Reed e Frost, que será apresentado na seção 1.6.

Após estes estudos, o crescimento da literatura voltada à epidemiologia matemática cresceu bastante, como por exemplo Dietz (1988) e Becker(1978). Em modelos estocásticos, a literatura cresceu bastante também, como por exemplo, Bartlett (1955 e 1960) e Bailey (1975).

Dois livros clássicos na área de modelagem de epidemias são: Bailey (1957) e Anderson e May (1991). O primeiro livro aborda modelos estocásticos e determinísticos com diversas aplicações a dados reais. O segundo livro também possui o foco na estimação. Contudo, a estimação é realizada do ponto de vista determinístico.

Um histórico mais detalhado da modelagem de epidemias pode ser encontrado em Bailey (1975) e Anderson e May (1991). A seção 1.8 tratará de bibliografias acerca de estimação de modelos estocásticos para estimação de epidemias.

## 1.4 Modelos Estocásticos vs. Determinísticos

Esta dissertação de mestrado foca apenas em modelos estocásticos, mesmo que os modelos determinísticos também sejam de fundamental importância para a modelagem de epidemias.

A principal vantagem dos modelos determinísticos é que estes são, em geral, mais fáceis de lidar, permitindo assim a utilização de modelos mais complexos para se modelar uma epidemia. Geralmente, para um modelo estocástico ser matematicamente fácil de lidar, este tem que ser mais simples e isto pode tornar o modelo irreal. Por exemplo, os estudos desta dissertação de mestrado não consideram o período em que o indivíduo fica exposto, ou seja, os indivíduos passam diretamente do estado suscetível para infectado.

Contudo, há diversas razões para se preferir os modelos estocásticos. Primeiramente, a dispersão de uma doença é um fenômeno estocástico, pois há uma certa probabilidade da doença se transmitir de um indivíduo para outro ao invés de simplesmente definir deterministicamente se a transmissão vai ou não ocorrer.

Outra razão para se preferir os modelos estocásticos é que há fenômenos que são genuinamente estocásticos e que não podem ser aproximados pelo método determinístico.

Por exemplo, em uma comunidade grande, muitos modelos conduzirão a uma epidemia pequena, infectando poucos indivíduos, ou conduzirão a uma epidemia grande, infectando diversos indivíduos. O cálculo da probabilidade desses dois casos somente é possível na modelagem estocástica.

A terceira importante vantagem dos modelos estocásticos é a estimação. O conhecimento sobre a incerteza nas estimativas requer um modelo estocástico, e uma estimativa não tem muito fundamento sem algum conhecimento de sua incerteza.

Portanto, pode-se dizer que os dois tipos de modelagem são de fundamental importância para se entender a dispersão de uma determinada doença.

## 1.5 Modelos Determinísticos

Os modelos SEIR (Suscetíveis - Expostos - Infectantes - Recuperados) foram introduzidos por Kermack e Mckendrik em 1927 e tem sido amplamente utilizados por ajustarem-se a diversas epidemias. Tais modelos consideram uma população que se divide em quatro compartimentos: suscetíveis, expostos, infectantes e recuperados.

O modelo consiste em um sistema de equações integro-diferencial. Uma versão com população fechada (sem entradas) é dada a seguir:

$$\begin{aligned}
&\frac{\partial S(a,t)}{\partial a} + \frac{\partial S(a,t)}{\partial t} = -\left[\lambda(a,t) + v(a,t) + \mu\right] S(a,t) + \pi R(a,t) \\
&\frac{\partial E(a,t)}{\partial a} + \frac{\partial E(a,t)}{\partial t} = \lambda(a,t)\, S(a,t) - (\sigma + \mu)\, E(a,t) \\
&\frac{\partial I(a,t)}{\partial a} + \frac{\partial I(a,t)}{\partial t} = \sigma E(a,t) - (\gamma + \mu)\, I(a,t) \\
&\frac{\partial R(a,t)}{\partial a} + \frac{\partial R(a,t)}{\partial t} = v(a,t)\, S(a,t) + \gamma I(a,t) - (\pi + \mu)\, R(a,t),
\end{aligned}$$

em que:

- $\lambda(a,t) = \int_0^L \beta(a,a')\, I(t,a')\, da'$.

- $S(a,t), E(a,t), I(a,t), R(a,t)$ são, respectivamente, a distribuição etária dos indivíduos suscetíveis, expostos, infectantes e recuperados no instante t e idade a.

- $\sigma$ é a taxa de incubação, ou seja, a proporção de latentes que tornam-se infectantes por unidade de tempo.

- $\gamma$ é a taxa de recuperação, ou seja, a proporção de infectantes que tornam-se recuperados por unidade de tempo.

- $\pi$ é a taxa de perda de imunidade, ou seja, a proporção de recuperados que tornam-se suscetíveis por unidade de tempo.

- $\lambda(a,t)$ é a força de infecção na idade a no instante t.

- $\mu$ é a taxa de mortalidade natural. O modelo considera que a taxa de mortalidade natural é igual em todos os compartimentos.

- $v(a,t)$ é a taxa de vacinação na idade a no instante t.

O modelo é apresentado na figura 1.

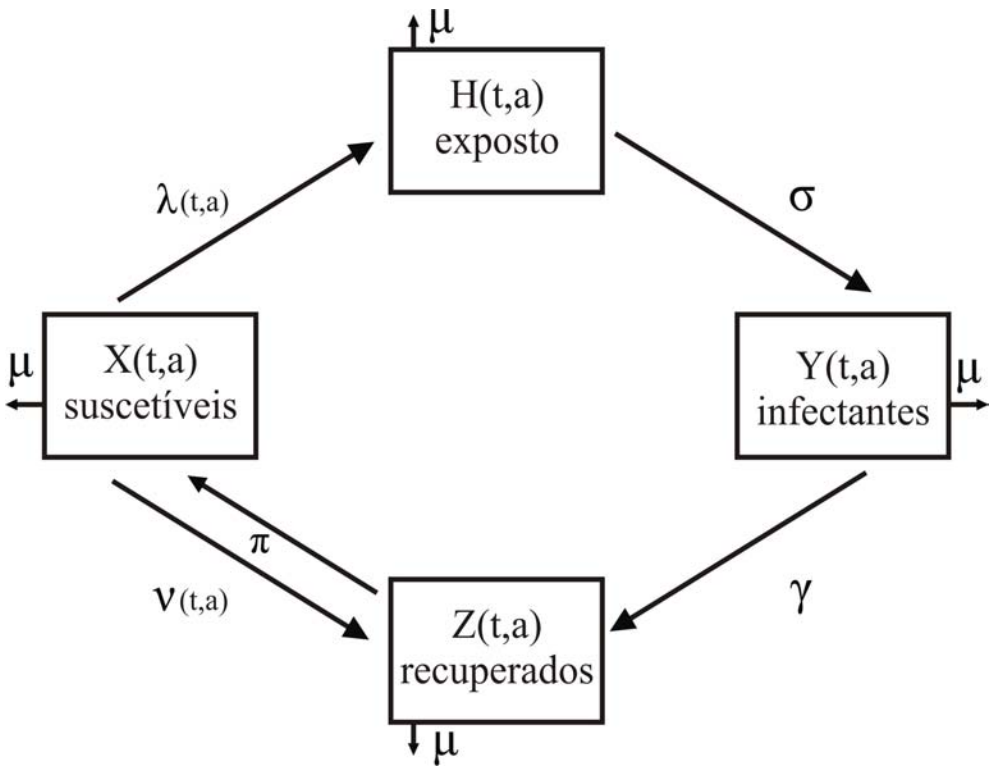



Figura 1: Representação Esquemática do Modelo SEIR sem Entradas.

Note que $\sigma^{-1}$, $\gamma^{-1}$ e $\pi^{-1}$ definem, respectivamente, os períodos de incubação, de recuperação e de imunidade.

Considere agora um modelo mais simplificado, o modelo SIR. O modelo consiste também em um sistema de equações integro-diferencial. Uma versão que assume população constante e entrada de novos suscetíveis é dada a seguir:

$$\frac{dS(t)}{dt} = \mu N - \mu S(t) - \lambda(t)S(t)$$
$$\frac{dI(t)}{dt} = \lambda(t)S(t) - \mu I(t) - \gamma I(t)$$
$$\frac{dR(t)}{dt} = \gamma I(t) - \mu R(t),$$

sendo $\lambda(t) = \beta I(t)$.

No modelo, $\mu$, a taxa de mortalidade, é considerada igual à taxa de natalidade para que a população seja constante de tamanho N. Uma maneira esquemática de se ver o modelo é apresentada na figura 2.

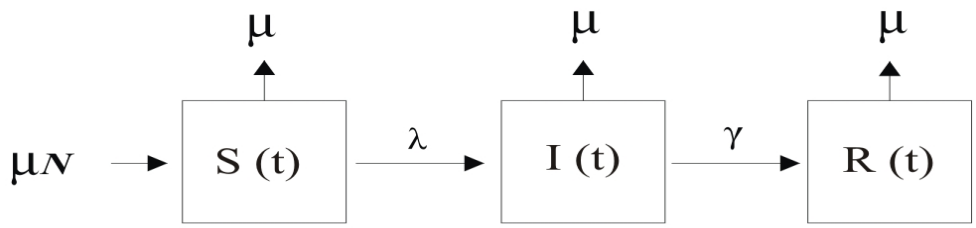



Figura 2: Representação Esquemática do Modelo SIR com População Total Constante.

E para gerar as trajetórias de S(t), I(t) e R(t) através do sistema acima, aplica-se métodos numéricos, como por exemplo, Runge-Kutta (ver Boyce e Diprima, 2006, capítulo 8), com valores de $\mu$, $\gamma$ e $\beta$ pré-determinados.

Uma análise importante nos modelos determinísticos é encontrar os pontos de equilíbrio do sistema e verificar a sua estabilidade. Para encontrar tal ponto iguala-se as equações do sistema a zero e encontra-se os valores de S(t), I(t) e R(t). Por exemplo, no sistema encontra-se o ponto de equilíbrio trivial dado por $(S_1, I_1, R_1) = (N, 0, 0)$ e o ponto de equilíbrio não trivial dado por:

$$(S_2, I_2, R_2) = \left( \frac{\mu + \gamma}{\beta}, \frac{\mu[\beta N - \mu - \gamma]}{\beta[\mu + \gamma]}, \frac{\gamma[\beta N - \mu - \gamma]}{\beta[\mu + \gamma]} \right).$$

Para maiores detalhes, consultar Yang (2001).

Observe que nos modelos determinísticos o número de indivíduos em cada compartimento é real. Contudo, na prática eles são inteiros.

## 1.6 O Modelo de Reed-Frost

Nesta seção será apresentado o modelo estocástico mais simples para a dispersão de uma doença infecciosa. Este modelo é um modelo de cadeia binomial de tempo discreto e

é denominado modelo Reed-Frost em homenagem a seus criadores. Em algumas palestras ocorridas em 1928 nos Estados Unidos, Lowell J. Reed e Wade Hampton Frost propuseram o uso de idéias probabilísticas para se modelar epidemias.

No modelo estocástico, o número de indivíduos em cada compartimento é inteiro. Para deixar claro essa diferença, nos modelos estocásticos a seguinte notação é utilizada:

- $X_t$ é o número de suscetíveis no tempo t;

- $Y_t$ é o número de infectados no tempo t.

Este modelo é da classe SIR, ou seja, há, a princípio, indivíduos suscetíveis na população. Se um indivíduo torna-se infectado, ele será infectante por algum tempo e depois se recuperará e se tornará imune, estado denominado removido.

Neste modelo, o tempo é considerado discreto. Com o tempo discreto, é natural pensar que o período infeccioso é curto e é precedido por um período de latência maior, pois a taxa de infecção é a mesma durante o mesmo tempo ou geração. Então, novas infecções irão ocorrer em gerações e tais gerações serão separadas pelo período de latência com a unidade de tempo discreta.

As probabilidades em um dado tempo somente dependem do tempo passado (isto é, um modelo Markoviano), e estes eventos são especificados por probabilidades binomiais. Um indivíduo suscetível no tempo t se mantém suscetível no tempo t+1 se ele não for infectado pelos indivíduos infectados no tempo t. Estes eventos são independentes e cada um ocorre com probabilidade q. A probabilidade 1-q é a probabilidade de um indivíduo suscetível encontrar um infectante. Além disso, os indivíduos infectados são removidos no próximo tempo. Então, o modelo de cadeia-binomial Reed-Frost tem as seguintes probabilidades condicionais:

$$P(Y_{t+1} = y_{t+1}/X_0 = x_0, Y_0 = y_0, ..., X_t = x_t, Y_t = y_t)$$

$$= P(Y_{t+1} = y_{t+1}/X_t = x_t, Y_t = y_t)$$

$$= \begin{pmatrix} x_t \\ y_{t+1} \end{pmatrix} (1 - q^{y_t})^{y_{t+1}} \, (q^{y_t})^{x_t - y_{t+1}},$$

com $X_{t+1} = X_t - Y_{t+1}$.

Dada as condições iniciais $X_0 = n$ e $Y_0 = m$, a probabilidade da cadeia completa $y_1, ..., y_k, y_{k+1} = 0$ é obtida condicionando seqüencialmente e usando a propriedade da cadeia de Markov. Se $x_{t+1} = x_t - y_{t+1}$, temos:

$$P(Y_1 = y_1, ..., Y_k = y_k, Y_{k+1} = 0/X_0 = n, Y_0 = m)$$
$$= P(Y_1 = y_1/X_0 = n, Y_0 = m) \times ... \times P(Y_{k+1} = 0/X_k = x_k, Y_k = y_k)$$

$$= \begin{pmatrix} n \\ y_1 \end{pmatrix} (1-q^m)^{y_1} (q^m)^{n-y_1} \times ... \times \begin{pmatrix} x_k \\ 0 \end{pmatrix} (1-q^{y_k})^0 (q^{y_k})^{x_k}. \tag{1.1}$$

A fórmula 1.1 pode ser utilizada para calcular o tamanho final da epidemia ($Z = \sum_{j\geq 1} Y_j$). Para se calcular $P(Z = z/X_0 = n, Y_0 = m)$ soma-se as probabilidades de todas as cadeias nas quais $|y| = \sum_{j\geq 1} Y_j = z$. Das equações acima, verifica-se que $Y_j = 0$ implica em $Y_{j+1} = 0$. Isto significa que uma nova infecção somente ocorrerá se há algum infectado no instante anterior, o que implica que o tamanho da cadeia não poderá ser maior do que o total de indivíduos infectados, fazendo com que o número de cadeias possíveis seja finito. Então, a função de probabilidade para o tamanho final da epidemia é dada por:

$$P(Z = z/X_0 = n, Y_0 = m)$$
$$= \sum_{y:|y|=z} P(Y_1 = y_1, ..., Y_k = y_k, Y_{k+1} = 0/X_0 = n, Y_0 = m).$$

Por exemplo, se tomarmos m=1 (como a maioria dos modelos adota) e n=3, calcu-

lamos as seguintes probabilidades do tamanho final da epidemia:

$$\begin{aligned}
P(Z=0/X_0=3) &= P(Y_1=0/X_0=3)=q^3 \\
P(Z=1/X_0=3) &= P(Y_1=1,Y_2=0/X_0=3) \\
&= P(Y_1=1/X_0=3)\ P(Y_2=0/X_1=2,Y_1=1) \\
&= \left(\begin{array}{c}3\\1\end{array}\right)(1-q)^1\,(q)^{3-1}\ \left(\begin{array}{c}2\\0\end{array}\right)(1-q)^0\,(q)^{2-0} \\
&= \left(\begin{array}{c}3\\1\end{array}\right)pq^2\ q^2,
\end{aligned}$$

onde $p=1-q$

$$\begin{aligned}
P(Z=2/X_0=3) &= P(Y_1=2,Y_2=0/X_0=3)+P(Y_1=1,Y_2=1,Y_3=0/X_0=3) \\
&= \left(\begin{array}{c}3\\2\end{array}\right)p^2q\ q^2+\left(\begin{array}{c}3\\1\end{array}\right)pq^2\ \left(\begin{array}{c}2\\1\end{array}\right)pq\ q.
\end{aligned}$$

$$P(Z=3/X_0=3)=1-P(Z=2/X_0=3)-P(Z=1/X_0=3)-P(Z=0/X_0=3).$$

Como vemos acima, o cálculo destas probabilidades torna-se muito complicado mesmo para tamanhos amostrais não muito grandes (n>10). Este modelo é um caso especial de modelo SIR, no qual a duração do período infeccioso é determinística. Também assume-se que a população é homogênea e que os indivíduos se misturam de maneira homogênea também.

Este modelo foi utilizado, por exemplo, para se modelar epidemias em famílias (O'Neill e Roberts(1999)) .

## 1.7 O Modelo SIR Estocástico

A dissertação lidará com os modelos SIR, que é um modelo compartimental com 3 estados: suscetível, infectado e recuperado (Cliff e Haggett, 1993), conforme já visto anteriormente. Um indivíduo suscetível se torna um transmissor da doença se o seu estado da infecção se altera de suscetível para infectado. Quando este indivíduo infectado se cura da doença, ele não contribui mais para a dispersão dela e assim passa para o estado de recuperado. Em alguns casos considera-se que o indivíduo que morreu ou que foi isolado do resto da população é do estado recuperado. O modelo assume que a população é homogênea, ou seja, dois indivíduos escolhidos ao acaso tem probabilidades iguais de se encontrarem. Para uma população que vive em uma área extensa esta não é uma boa suposição, visto que as pessoas possuem chance maior de encontrar seus vizinhos do que as pessoas que moram longe. O modelo SIR pode ser visto sob duas abordagens, estocástica ou determinística, conforme visto na seção anterior. Podemos observar os indivíduos como quantidades discretas ou contínuas e podemos assumir que o tempo é contínuo ou discreto, ver Diekmann e Heesterbeek (2000) e Andersson e Britton (2000). A estimação dos parâmetros do modelo estocástico depende dos dados disponíveis. Se o tempo e o tipo de cada transição entre os compartimentos é completamente observado, a modelagem torna-se mais fácil. Contudo, isto é muito raro de se acontecer na prática devido ao fato dos dados serem colhidos de maneira retrospectiva e não com um planejamento prévio. O princípio da epidemia é, em geral, difícil de se determinar e geralmente só há informações secundárias, como por exemplo, sintomas clínicos. Em geral, é mais fácil de se detectar as recuperações, pois elas se referem ao período de tempo no qual o indivíduo se torna inativo com respeito à dispersão da doença.

Considere uma população inicial de $N$ indivíduos suscetíveis e $a$ indivíduos infectantes. Em geral, $a$ pode ser grande, mas para simplificar a maioria dos modelos adota que $a$=1. $X(t)$ e $Y(t)$ denotam, respectivamente, o número de indivíduos suscetíveis e infectantes no tempo $t \geq 0$. A população é considerada fechada quando não temos nascimentos ou mortes. Nesta dissertação, é razoável considerar a população como sendo fechada, visto que, a epidemia não é letal, ocorrerá em uma população pequena e será observada durante um curto período de tempo. Neste caso, para todo t, a igualdade

$X(t) + Y(t) + R(t) = N + a$ é satisfeita, onde $R(t)$ denota o número de indivíduos recuperados no tempo t. A evolução do processo epidemiológico $(X(t), Y(t))$ pode ser descrita com as seguintes taxas de transição:

$$(X(t), Y(t)) \rightarrow (X(t) - 1, Y(t) + 1) : \beta X(t)Y(t)$$

$$(X(t), Y(t)) \rightarrow (X(t), Y(t) - 1) : \gamma Y(t)$$

Ou seja, um indivíduo passa de suscetível para infectado com a taxa de $\beta Y(t)$ e um indivíduo passa de infectado para recuperado com a taxa de $\gamma$. Diversas bibliografias, como Andersson e Britton (2000), utilizam $\frac{\beta Y(t)}{N(t)}$ ao invés de $\beta Y(t)$ para a taxa de transição de suscetíveis para infectantes, onde $N(t)$ é o tamanho total da população no instante t. Nesta dissertação, esta diferença ocorre apenas na interpretação do parâmetro, pois estamos considerando uma população fechada.

A epidemia começa com um único indivíduo infectado e ela acontece até que não se tenha mais nenhum indivíduo infectado na população. Se o indivíduo infectado no início da epidemia se recupera antes que outro indivíduo da população se torne infectado a epidemia termina.

O modelo considerado é Markoviano. O processo $(X, Y) = \{(X(t), Y(t)); t \geq 0\}$ é um processo Markoviano se e somente se o período infeccioso tiver a propriedade de falta de memória (Anderson e Britton, 2000). Ver propriedade 1 do apêndice.

Para um dado instante t, o tempo da próxima infecção possui distribuição exponencial com taxa de falha $\beta X(t)Y(t)$ e o tempo da próxima recuperação terá distribuição exponencial com taxa de falha $\gamma Y(t)$. Ou seja, a função densidade de probabilidade do tempo da próxima infecção $T_1$ é dada por:

$$f(t_1) = \beta X(t)Y(t)e^{-\beta X(t)Y(t)t_1}, \text{ com } E(t_1) = \frac{1}{\beta X(t)Y(t)}$$

e a do tempo da próxima recuperação $T_2$ é dada por:

$f(t_2) = \gamma Y(t) e^{-\gamma Y(t) t_2}$, com $E(t_2) = \dfrac{1}{\gamma Y(t)}$ .

**Exemplo 1: Abordagem Determinística vs. Estocástica**

O modelo utilizado será o modelo SIR, cujo sistema de equações é dado por:

$$\frac{dX(t)}{dt} = -\beta X(t) Y(t)$$

$$\frac{dY(t)}{dt} = \beta X(t) Y(t) - \gamma Y(t)$$

$$\frac{dR(t)}{dt} = \gamma Y(t),$$

em que $\beta = 0,005$ e $\gamma = 0,08$. Tem-se 49 indivíduos suscetíveis ($X(0) = 49$) e 1 indivíduo infectante ($Y(0) = 1$) no início.

As trajetórias para X(t) e Y(t) serão geradas pelo software Berkeley Madonna e o método numérico utilizado é o Runge-Kutta de ordem 4.

Também será realizada uma simulação estocástica da epidemia, equivalente ao modelo determinístico descrito acima.

O algoritmo (algoritmo 1) para a simulação estocástica é o seguinte:

1. Seja $t_0$ o início da epidemia. Então temos que $X(t_0) = 49$ e $Y(t_0) = 1$.

2. Gerar uma observação da distribuição exponencial de taxa de falha $\beta X(t_0) Y(t_0)$ e denotar por $t_1'$. Gerar uma observação da distribuição exponencial de taxa de falha $\gamma Y(t_0)$ e denotar por $t_2'$.

3. Seja $t_1 = min\left(t_1^{'}, t_2^{'}\right)$.

   No intervalo $[t_0, t_0+t_1)$ os valores de número de suscetíveis e infectados permanecem constantes, isto é, neste intervalo $X(t) = X(t_0)$ e $Y(t) = Y(t_0)$.

   Se $t_1^{'} < t_2^{'}$, então $X(t_0 + t_1) = X(t_0) - 1$ e $Y(t_0 + t_1) = Y(t_0) + 1$, ou seja, ocorreu uma infecção. Caso contrário, $X(t_0 + t_1) = X(t_0)$ e $Y(t_0 + t_1) = Y(t_0) - 1$, ou seja, ocorreu uma remoção.

4. Fazer $t_0 = t_0 + t_1$.

5. Repetir os passos 2-4 até que não se tenha mais nenhum indivíduo infectante na população.

A figura 3 apresenta uma comparação das trajetórias dos suscetíveis e infectados geradas pela abordagem determinística e estocástica. Observando estes gráficos, vemos que, mesmo utilizando os mesmos parâmetros, podemos ter trajetórias diferentes para os suscetíveis e infectantes na abordagem estocástica, fato que não acontece na abordagem determinística. Também podemos ter tamanhos finais (m) e tempos finais diferentes na abordagem estocástica, para um mesmo conjunto de parâmetros $\beta$ e $\gamma$.

Em Andersson e Britton (2000), capítulo 2, pode-se encontrar como calcular a probabilidade de todos os tamanhos finais da epidemia, para valores fixos de $\beta$, $\gamma$ e número de suscetíveis e infectantes iniciais na população. Seja N o tamanho da população fechada e a o número de infectantes iniciais da população. Denote por $P_k^N$, a probabilidade do tamanho final da epidemia ser igual a k, $0 \leq k \leq N$. Então, para uma população fechada, temos:

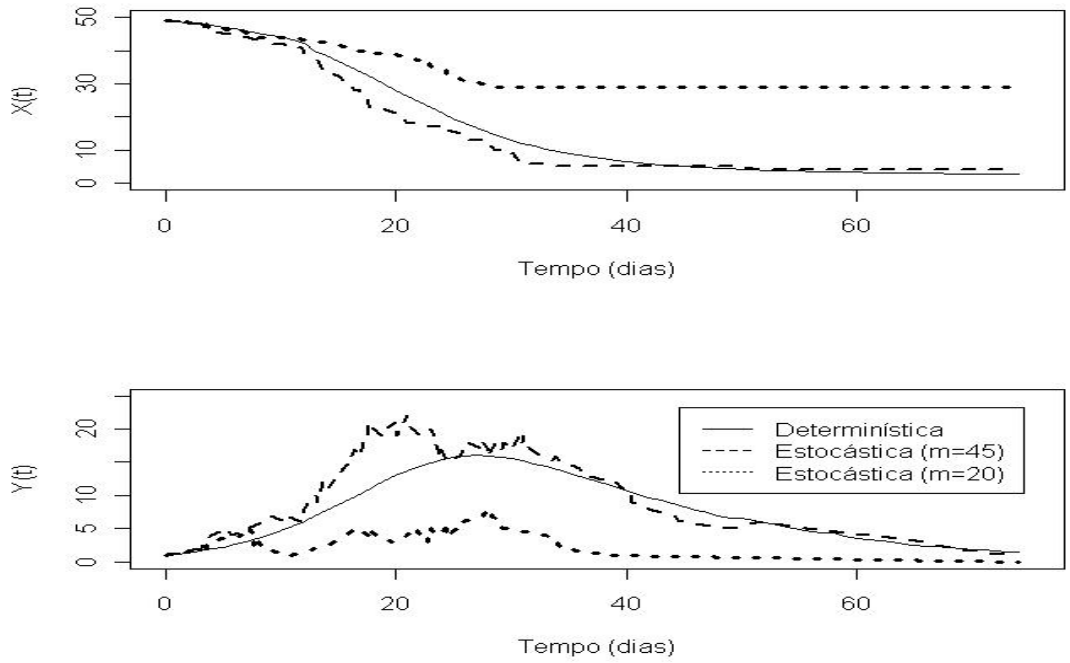



Figura 3: Comparação das trajetórias do número de suscetíveis X(t) e infectados Y(t) entre a abordagem determinística e duas amostras estocásticas.

$$\sum_{k=0}^{L} \binom{N-k}{L-k} P_k^N / \left[\phi\left(\beta(N-L)\right)\right]^{k+a} = \binom{n}{L}, \; 0 \leq L \leq N,$$

onde $\phi(\theta) = E\left(e^{-\theta I}\right)$, onde I é uma variável aleatória com distribuição do tempo de transição de infectante para recuperado. No caso Markoviano possui distribuição exponencial de taxa de falha $\gamma$.

Para a epidemia descrita no exemplo, as probabilidades de tamanhos finais da epidemia são apresentadas na figura 4. Verifica-se que, para o nosso conjunto de parâmetros e indivíduos infectantes e suscetíveis iniciais na população, a maior probabilidade existente é de não haver nenhum caso da doença na população. A segunda maior probabilidade é que 49 dos 50 indivíduos se infectem. Tendo estas probabilidades, facilmente pode-se calcular o valor esperado do número de infectados. O valor esperado é de 31,7 indivíduos infectados.

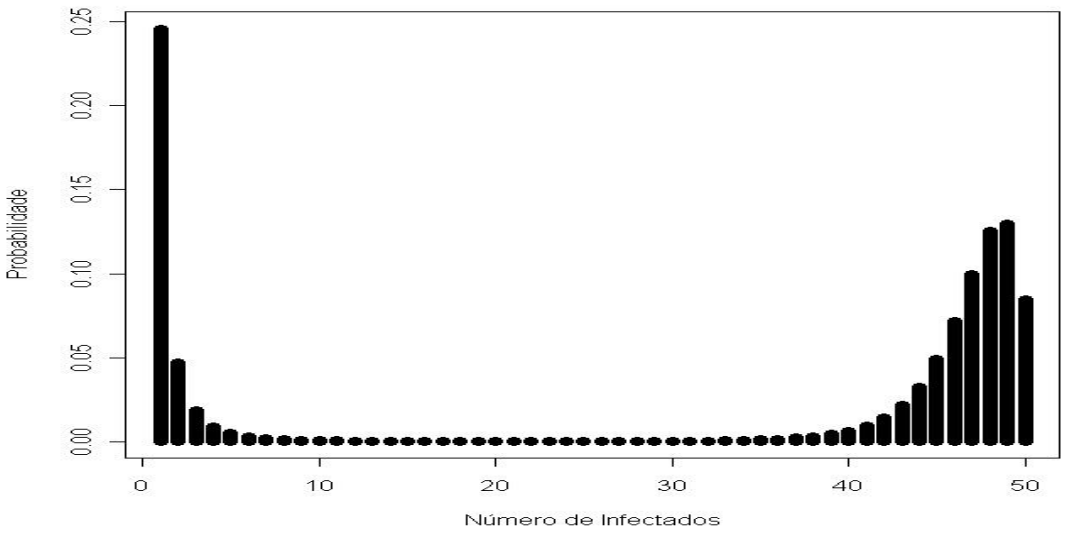



Figura 4: Probabilidades do tamanho final da epidemia. Os parâmetros utilizados para gerar a epidemia foram $\beta = 0{,}005$ e $\gamma = 0{,}08$ com 49 suscetíveis e 1 infectante no início.

Para mostrarmos a variabilidade dos tempos finais da epidemia na abordagem estocástica, foi feito um histograma para 5000 simulações do modelo, apresentado na figura 5.

Uma estimativa importante para se fazer durante a modelagem de epidemias é a estimativa da taxa de reprodutibilidade basal. Tal taxa, conforme já dito, é o número esperado de casos secundários que um caso primário pode produzir. Segundo Höhle e Jorgensen (2002), a taxa de reprodutibilidade basal deste modelo é dada por:

$$R_0 = \frac{\beta N}{\gamma}.$$

Tem-se o interesse nesta medida devido ao fato que, para grandes populações, grandes epidemias ocorrem se e somente se $R_0 > 1$ (Andersson e Britton (2000)).

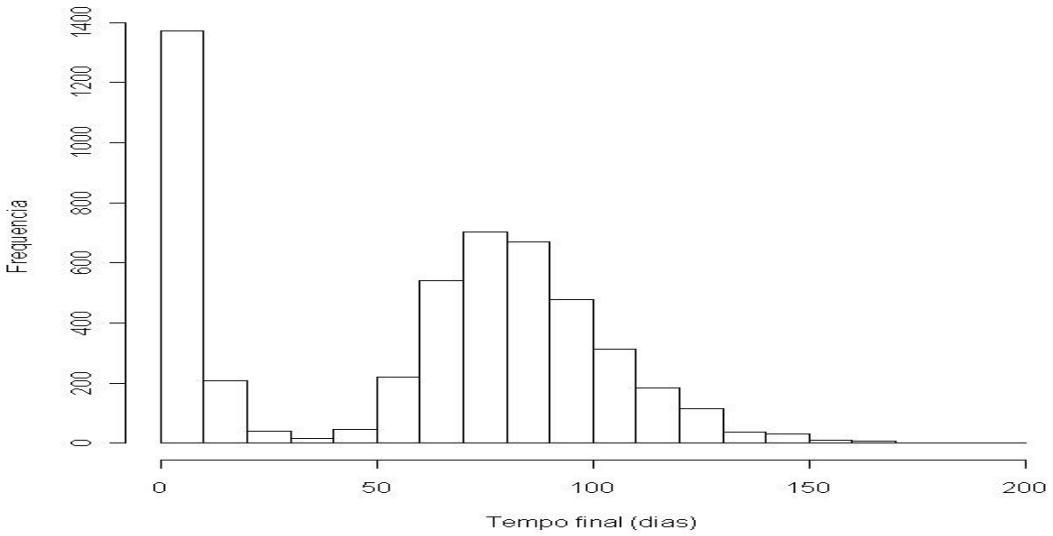



Figura 5: Histograma do tempo final da epidemia. Os parâmetros utilizados para gerar a epidemia foram $\beta = 0{,}005$ e $\gamma = 0{,}08$ com 49 suscetíveis e 1 infectante no início.

## 1.8 Estimação de Modelos SEIR Estocásticos

Nos últimos anos, os estudos de modelos estocásticos compartimentais têm crescido bastante devido ao fato da grande melhora dos recursos computacionais e devido ao desenvolvimento de diversas técnicas como, por exemplo, os métodos MCMC - Monte Carlo em Cadeias de Markov (Gelfand e Smith, 1990). Portanto, a lista de trabalhos relacionados a técnicas estatísticas para inferência e estimação de modelos dinâmicos para populações tem crescido bastante.

Na abordagem clássica, Bailey (1975) realizou estimativas de máxima verossimilhança em modelos SIR com dados incompletos de uma epidemia de varíola; Brookhart *et al.* (2002) utilizaram a abordagem de verossimilhança perfilada com restrições para estimar um modelo de transmissão de epidemias ajustado à epidemia de *Cryptosporidium parvum* ocorrida em Milwaukee, EUA, em 1993; Em Andersson e Britton (2000) pode se encontrar diversas técnicas de estimação como martingais, algoritmo EM e máxima verossimilhança. Neste trabalho é considerado o cenário de dados completos e incompletos; Höhle e Jorgensen (2002) estudaram estimativas por máxima verossimilhança em modelos SIR

estocásticos com dados incompletos; Chowell *et al.* realizaram estimação por minímos quadrados em modelos SEIR estocásticos com aplicação a dados de ebola.

Na abordagem bayesiana, os métodos MCMC têm sido amplamente utilizados. Por exemplo, Cancré *et al.* (2000) estimaram um modelo dinâmico de transmissão de malária utilizando o algoritmo de Metropolis Hastings; Ades e Cliffe (2002) utilizaram o método Monte Carlo em Cadeias de Markov (MCMC) para estimar parâmetros quando existem informações provenientes de várias fontes; O'Neill e Roberts (1999) utilizaram o algoritmo MCMC para estimação com dados com observação incompleta; O'Neill e Becker (2001) utilizaram os métodos MCMC para se realizar inferências para os parâmetros de modelos SEIR quando a suscetibilidade dos indivíduos varia; Streftaris e Gibson (2004) realizaram estimações de modelos SEIR estocásticos utilizando modelos não markovianos; Lekone e Finkenstädt (2006) estudaram inferências estatísticas em modelos SEIR estocásticos com intervenção aplicados a dados de ebola; Clancy e O'Neill (2008) realizaram estimativas bayesianas para a taxa de reprodutibilidade basal; Höhle e Jorgensen (2002) também estudaram estimativas bayesianas em modelos SIR estocásticos; o livro de Andersson e Britton (2000) reúne também diversas estimativas bayesianas de modelos estocásticos para a modelagem de epidemias.

Na abordagem bayesiana, um outro método que vem sendo bastante estudado é o "Bayesian Melding"(ver Raftery *et al.* (1995) e Poole e Raftery (2000)). Tal método propõe uma forma de combinar diferentes formas de informações, qualitativas ou quantitativas, sobre diferentes variáveis do modelo determinístico. Nesta abordagem podem ser combinadas informações a priori das entradas e saídas do modelo, bem como as verossimilhanças relacionas a diferentes observações. Hotta (2009) realizou estimativas para o modelo SEIR utilizando o Bayesian Melding; Spear *et al.* (2002) também utilizaram esta técnica na modelagem da esquistossomose.

O algoritmo EM também é bastante utilizado na estimação de modelos SIR, devido ao fato de geralmente lidarmos com observações incompletas do processo epidêmico. Por exemplo, Becker (1993b) realizou inferências utilizando este algoritmo; Becker (1997) utilizou o algoritmo EM para analisar dados de HIV; Meester *et al.* (2002) utilizou o algoritmo EM modificado para modelar uma epidemia de gripe suína. Outro método

bastante utilizado em estimação de modelos SIR estocásticos é a estimação através de martingais. Picard (1980) mostrou que os martingais podem ser utilizados na modelagem de epidemias; Becker (1989), capítulo 7, aplicou a técnica dos martingais para fazer estimações com dados de varíola; Becker (1993a) estudou sobre as propriedades das estimativas por martingais; Höhle (2003) utilizou os martingais para realizar estimativas da taxa de reprodutibilidade basal.

# 2 *Estimativa Bayesiana e de Máxima Verossimilhança para Dados Completos e Incompletos*

Neste capítulo será descrita a verossimilhança do modelo e serão calculados os estimadores de máxima verossimilhança quando temos os dados completos e incompletos. Também será apresentado um estimador bayesiano quando temos dados completos e incompletos.

Nos modelos SIR com tempos de transição de distribuição exponencial, os dados são completos quando temos todos os tempos de infecção e remoção completamente observados. No caso das distribuições dos tempos de transição não serem exponenciais, isto é, sem memória, é necessário que seja especificado quais indivíduos sofreram a transição.

Nesta dissertação, será estudado modelos SIR cujos tempos de transição possuem distribuição exponencial com informação completa e incompleta. No caso de informação incompleta, será considerado o caso em que temos os tempos de remoção conhecidos e os tempos de infecção desconhecidos.

## 2.1 Descrição do Modelo

Inicialmente, será descrito o modelo para dados completos. Assume-se que temos uma população fechada de tamanho N. Denote por $I_0 < 0$ o tempo da primeira infecção. Assume-se que a epidemia é observada no tempo $[I_0, T]$. Se no tempo T todos os infectados são recuperados, a epidemia é considerada observada até o final. Os dados consistem nos

tempos de remoção $\boldsymbol{\tau}$ e tempos de infecção $\mathbf{I}$, observados durante $[I_0, T]$. O vetor de removidos $\boldsymbol{\tau}$=$(\tau_1, ..., \tau_n)$ está definido de maneira ordenada na qual a primeira remoção está definida no tempo 0. Então, $\tau_1 \equiv 0$ e $\tau_i \leq \tau_{i+1}$. Similarmente, $\mathbf{I} = (I_1, ..., I_{m-1})$ com $I_j \leq I_{j+1}$. Como a epidemia é observada desde o início, temos que $m \leq n \leq N$ e se a epidemia é observada até o final temos que $m = n$. Para esta dissertação só foram consideradas epidemias observadas do início até o final.

Como a duração do período infectante é estocástica, $I_i$ e $\tau_i$ não tem que necessariamente se referir ao mesmo indivíduo. Quando não se tem mais indivíduos infectantes a epidemia acaba. Este conhecimento combinado com os dados observados pode ser transmitido para termos uma relação entre os I's e $\tau's$. Define-se que I(t) e R(t) são o número de infectados e recuperados no tempo t, respectivamente. Para cada recuperação observada $\tau_i$, a infecção correspondente tem que ocorrer antes de $\tau_i$, isto é, para todo $t \in [I_0, \tau_n[$ deve-se ter $Y(t) = I(t) - R(t) > 0$. Isto é equivalente à condição:

$$I_i < \tau_i \text{ para todo } 1 \leq i \leq m-1. \tag{2.1}$$

A figura 6 mostra um exemplo de uma epidemia observada até o final, com três infecções e consequentemente, três recuperações.

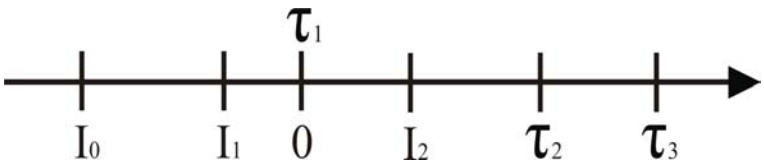



Figura 6: Exemplo de evolução de uma epidemia no tempo até a extinção com três infecções e três remoções.

## 2.2 Estimativa por Máxima Verossimilhança

### 2.2.1 Estimativa por Máxima Verossimilhança para Dados Completos

No caso de dados completos observamos $(\boldsymbol{\tau}, I_0, \mathbf{I})$. Para facilitar a notação será modelado o tempo de espera entre os dois eventos consecutivos, isto é, $t_i = t_i^{'} - t_{i-1}^{'}$, em que $t_i^{'}$ denota os tempos dos eventos. A verossimilhança de cada tempo de espera pode ser encontrada utilizando análise de sobrevivência com estrutura para diversos modelos de falha (Fahrmeir e Tutz, 1994). Dois eventos de transição (falha) são possíveis: um dos $X(t_i)$ suscetíveis infecta-se com taxa individual $\beta Y(t_i)$, ou um dos $Y(t_i)$ indivíduos infectados se recupera com taxa individual $\gamma$. Logo, as taxas de transição de um indivíduo qualquer da população para infectante $(\lambda_{inf}(t_i))$ ou recuperado $(\lambda_{rec}(t_i))$ são dadas por:

$$\lambda_{inf}(t_i) = \beta X(t_i) Y(t_i)$$
$$\lambda_{rec}(t_i) = \gamma Y(t_i),$$

onde $X(t_i)$ e $Y(t_i)$ são o número de suscetíveis e infectados no tempo $t_i$, respectivamente.

Então, de acordo com a propriedade 2 do apêndice, a função de risco total (risco de ocorrer uma infecção ou remoção) será dada por:

$$\lambda_{total}(t_i) = \beta X(t_i) Y(t_i) + \gamma Y(t_i).$$

Logo, entre dois eventos consecutivos, a função risco é constante, pois o número de indivíduos em cada compartimento não se altera durante este período. Esta simplificação vem do fato de se adotar distribuições exponenciais para os tempos de transição.

O conjunto de dados da epidemia é descrito por $\mathbf{D}=\{(t_i, e_i), e_i \in \{inf, rec\}\}$ denota o tipo do evento. $\mathbf{D}$ consiste em n eventos de recuperação e m eventos de infecção, com m=n quando epidemia é observada até o final. O interesse agora é na densidade de

$\boldsymbol{\tau}, \mathbf{I}$ dado $\beta, \gamma, I_0$. Considere então a verossimilhança de um evento arbitrário $e_i$ em $\mathbf{D}$ numa configuração com função risco constante e censura não informativa. Neste caso, a verossimilhança é descrita por:

$$L_i = \lambda_{e_i}(t_i)\, P(T_i \geq t_i/\beta, \gamma) = \lambda_{e_i}(t_i)\, S(t_i/\beta, \gamma),$$

em que $T_i$ é uma variável estocástica denotando o tempo de falha de $e_i$ e

$$S(t/\beta, \gamma) = \exp\left(-\int_0^t \lambda(u/\beta, \gamma)\, du\right)$$

é a função sobrevivência baseada no risco total. Assumindo independência dos intervalos de tempo entre ocorrências, a verossimilhança total é dada olhando-se para todos os eventos de infecção e remoção, exceto $I_0$. $I_0$ não faz parte da verossimilhança, pois uma epidemia é sempre condicionada na existência do primeiro indivíduo infectante. Então, temos que a função de verossimilhança L, considerando m=n, é igual a:

$$\left[\prod_{i=1}^{n} \gamma Y\left(\tau_i^-\right) \exp\left(-\int_{\tau_{i-1}}^{\tau_i} \lambda(t/\beta, \gamma)\, dt\right)\right]\left[\prod_{i=1}^{n-1} \beta X\left(I_i^-\right) Y\left(I_i^-\right) \exp\left(-\int_{I_{i-1}}^{I_i} \lambda(t/\beta, \gamma)\, dt\right)\right]$$

$$= \prod_{i=1}^{n} \gamma Y\left(\tau_i^-\right) \prod_{i=1}^{n-1} \beta X\left(I_i^-\right) Y\left(I_i^-\right)\left[\prod_{i=1}^{n} \exp\left(-\int_{\tau_{i-1}}^{\tau_i} \lambda(t/\beta, \gamma)\, dt\right)\right]\left[\prod_{i=1}^{n-1} \exp\left(-\int_{I_{i-1}}^{I_i} \lambda(t/\beta, \gamma)\, dt\right)\right]$$

$$= \prod_{i=1}^{n} \gamma Y\left(\tau_i^-\right) \prod_{i=1}^{n-1} \beta X\left(I_i^-\right) Y\left(I_i^-\right)\left[\exp\left(-\sum_{i=1}^{n}\left(\int_{\tau_{i-1}}^{\tau_i} \lambda(t/\beta, \gamma)\, dt\right) - \sum_{i=1}^{n-1}\left(\int_{I_{i-1}}^{I_i} \lambda(t/\beta, \gamma)\, dt\right)\right)\right],$$

onde $X(t^-)$, $Y(t^-)$ denotam, respectivamente, o número de suscetíveis e infectantes no instante logo antes de t, isto é, $Y(t^-) = lim_{t \to t^-} Y(t)$. Como os eventos $i$ e $\tau$ são consecutivos no tempo a soma das integrais podem ser simplificadas em só uma integral. Então, temos que:

$$L = f(\boldsymbol{\tau}, \mathbf{I}) = \prod_{i=1}^{n} \gamma Y\left(\tau_i^-\right) \prod_{i=1}^{n-1} \beta X\left(I_i^-\right) Y\left(I_i^-\right) \exp\left(-\int_{I_0}^{T} \beta X(t) Y(t) + \gamma Y(t)\, dt\right). \tag{2.2}$$

Reescrevendo a função de verossimilhança, temos que:

$$L = \gamma^n \prod_{i=1}^{n} Y\left(\tau_i^-\right) \beta^{n-1} \prod_{i=1}^{n-1} X\left(I_i^-\right) Y\left(I_i^-\right) \exp\left(-\int_{I_0}^{T} \beta X(t) Y(t) + \gamma Y(t)\, dt\right).$$

O logaritmo da função de verossimilhança é dada por:

$$l = \ln L = n \ln \gamma + \sum_{i=1}^{n} \ln Y\left(\tau_i^-\right) + (n-1) \ln \beta + \sum_{i=1}^{n-1} \ln X\left(I_i^-\right) + \sum_{i=1}^{n-1} \ln Y\left(I_i^-\right) - \int_{I_0}^{T} \beta X(t) Y(t)\, dt - \int_{I_0}^{T} \gamma Y(t)\, dt.$$

Diferenciando a equação acima com relação a $\beta$ e igualando a 0 temos:

$$\frac{\partial l}{\partial \beta} = \frac{n-1}{\beta} - \int_{I_0}^{T} X(t) Y(t)\, dt = 0 \Rightarrow \hat{\beta} = \frac{n-1}{\int_{I_0}^{T} X(t) Y(t)\, dt}.$$

Calculando-se a segunda derivada temos que:

$\frac{\partial^2 l}{\partial \beta^2} = \frac{-(n-1)}{\beta^2}$, que é negativa quando temos $n \geq 2$, pois $\beta > 0$.

Agora, diferenciando a função de log-verossimilhança com relação a $\gamma$ e igualando a 0 temos:

$$\frac{\partial l}{\partial \gamma} = \frac{n}{\gamma} - \int_{I_0}^{T} Y(t)\, dt = 0 \Rightarrow \hat{\gamma} = \frac{n}{\int_{I_0}^{T} Y(t)\, dt}.$$

Calculando a segunda derivada temos que:

$\frac{\partial^2 l}{\partial \gamma^2} = \frac{-n}{\gamma^2}$, que é sempre negativa, pois $\gamma > 0$.

Como $\frac{\partial^2 l}{\partial \beta \partial \gamma} = \frac{\partial^2 l}{\partial \gamma \partial \beta} = 0$, a matriz hessiana é diagonal e os termos de sua diagonal principal são negativos, temos que os estimadores de máxima verossimilhança encontrados são realmente pontos de máximo.

Logo, as estimativas de máxima verossimilhança para $\beta$ e $\gamma$ são dadas por:

$$\hat{\beta} = \frac{n-1}{\int_{I_0}^{T} X(t)\, Y(t)\, dt} \tag{2.3}$$

e

$$\hat{\gamma} = \frac{n}{\int_{I_0}^{T} Y(t)\, dt}. \tag{2.4}$$

Assim, temos que a matriz de informação de Fisher é dada por:

$$\begin{pmatrix} \frac{n-1}{\beta^2} & 0 \\ 0 & \frac{n}{\gamma^2} \end{pmatrix}.$$

Podemos obter facilmente sua inversa, dada por:

$$\begin{pmatrix} \frac{\beta^2}{n-1} & 0 \\ 0 & \frac{\gamma^2}{n} \end{pmatrix}.$$

Para n suficientemente grande, pelas propriedades dos estimadores de máxima verossimilhança, o estimador $\hat{\beta}$ possui distribuição aproximadamente normal com média $\beta$ e variância $\frac{\beta^2}{n-1}$ e $\hat{\gamma}$ possui distribuição aproximadamente normal com média $\gamma$ e variância $\frac{\gamma^2}{n}$.

As convergências dos dois estimadores acima para a distribuição normal são demonstradas em Anderson e Britton (2000), capítulo 9.

Um intervalo de confiança aproximado 95% para $\beta$ é dado por:

$$\left(\hat{\beta} - 1,96\sqrt{\frac{\hat{\beta}^2}{(n-1)}}; \hat{\beta} + 1,96\sqrt{\frac{\hat{\beta}^2}{(n-1)}}\right)$$

e um intervalo de confiança 95% aproximado para $\gamma$ é dado por:

$$\left(\hat{\gamma} - 1,96\sqrt{\frac{\hat{\gamma}^2}{n}}; \hat{\gamma} + 1,96\sqrt{\frac{\hat{\gamma}^2}{n}}\right).$$

Para calcular a variância de $R_0$, utiliza-se o fato de que se $\underline{\hat{\theta}} = \left(\hat{\beta}, \hat{\gamma}\right)$, a variância de $g\left(\underline{\hat{\theta}}\right) = \frac{\hat{\beta}\ N}{\hat{\gamma}}$ é dada por $g'\left(\underline{\hat{\theta}}\right)^T Var\left(\underline{\hat{\theta}}\right) g'\left(\underline{\hat{\theta}}\right)$, onde $g'\left(\underline{\hat{\theta}}\right)$ é o vetor de derivadas e $Var\left(\underline{\hat{\theta}}\right)$ é a inversa da matriz de informação de Fisher. Assim, tem-se que a variância aproximada de $R_0$ é dada por:

$$\frac{N^2\hat{\beta}^2}{\hat{\gamma^2}\,(n-1)} + \frac{N^2\hat{\beta}^2\hat{\gamma^2}}{n\hat{\gamma^4}}.$$

### 2.2.2 Estimativa por Máxima Verossimilhança para Dados Incompletos

Nesta seção será apresentado um método de estimação dos parâmetros $\beta$ e $\gamma$ por máxima verossimilhança quando temos dados incompletos. Nesse caso, os dados são incompletos devido ao fato de não termos os tempos de infecção e somente termos os tempos de remoção. O algoritmo descrito a seguir foi proposto por Höhle e Jorgensen(2002). A diferença do algoritmo proposto por eles e o descrito a seguir está na forma de simular os tempos de infecção desconhecidos. Eles propuseram, em cada passo do algoritmo, a realização de um sorteio uniforme de todos os tempos de infecção no espaço de todas as configurações válidas segundo a condição 2.1. O algoritmo proposto a seguir parte de uma configuraçào válida do vetor de infectados e somente altera duas posições dos tempos de infecção para cada passo do algoritmo.

Em cada passo do algoritmo, ocorrerá uma tentativa de alteração no vetor de infectados e, após a tentativa, será realizada uma estimativa por máxima verossimilhança para $\beta$ e $\gamma$ utilizando o vetor de infectados do passo atual e o vetor de removidos já conhecido. Tais estimativas serão feitas segundo as equações 2.3 e 2.4.

Assim, o algoritmo para estimativa por máxima verossimilhança com dados incompletos (algoritmo 2) é o seguinte:

1. Inicializar o vetor de infectados com uma configuração viável segundo a condição 2.1. Isto pode ser feito, por exemplo, colocando no tempo I(0) os m casos de infecções e 0 nos outros tempos de infecção.

2. Estimar $\beta$ e $\gamma$ por máxima verossimilhança de acordo com as equações 2.3 e 2.4 usando o vetor já conhecido $\boldsymbol{\tau}$ e o vetor de infectados do passo atual.

3. Sortear $t_1$ e $t_2$ uniformes em $(I_0, T)$. Fazer $Y(t_1) = Y(t_1) - 1$ e $Y(t_2) = Y(t_2) + 1$ . Aceitar esta mudança somente se a condição 2.1 for satisfeita.

4. Repetir os passos 2 e 3 k vezes (k suficientemente grande) até que se tenha uma distribuição estacionária para as estimativas de $\beta$ e $\gamma$.

Para tomarmos amostras independentes de $\beta$ e $\gamma$, após um burn-in conveniente, tomamos amostras espaçadas das k estimativas de modo que essas estimativas não possuam autocorrelação significativa. Da amostra independente, tomamos a média para termos a amostra requerida.

No passo 3 assume-se que os tempos de infecção são discretos. Isto ocorre devido ao fato de termos, na prática, os tempos de remoção observados de maneira discreta também. Esta suposição foi feita pois, em geral, somente é observado o dia em que o indivíduo foi removido. Para calcular um intervalo de confiança aproximado para os parâmetros, toma-se os quantis da amostra final.

Neste algoritmo, temos o problema de $I_0$ não ser limitado à esquerda. Por isso, assumiu-se que $-I_0$ possui distribuição U(0,$\theta$). O valor de $\theta$ será discutido posteriormente no capítulo 3.

## 2.3 Estimação Bayesiana

Nesta seção será proposta uma estimativa Bayesiana para os parâmetros da epidemia. Conforme já mencionado anteriormente, os tempos de infecção são raramente observados na prática e estamos interessados em estimar $\beta$, $\gamma$ e/ou $R_0$ faltando as observações de $I_0$ e **I**. Baseado apenas no tempo de remoção observado, os métodos MCMC (Markov Chain Monte Carlo) podem ser utilizados para lidar com a epidemia parcialmente observada.

Para completar a especificação do modelo, as seguintes distribuições a priori são assumidas para os parâmetros desconhecidos $\beta$, $\gamma$ e $I_0$:

$\beta \sim \Gamma(a, b)$,

$\gamma \sim \Gamma(c, d)$,

$-I_0 \sim Exp\,(\theta)$,

em que $\Gamma\,(\nu, \lambda)$ denota distribuição gamma com média $\nu/\lambda$ e variância $\nu/\lambda^2$ e $Exp\,(\theta)$ denota a distribuição exponencial de média $1/\theta$ e $a,\ b,\ c,\ d,\ \theta$ são hiperparâmetros, cujos valores serão discutidos posteriormente. Usando a equação 2.2 as posterioris condicionais podem ser obtidas para $\beta$, $\gamma$ e $I_0$, o que permite a utilização do algoritmo amostrador de Gibbs para obter estas densidades. Para $\mathbf{I}$ as posterioris condicionais não são obtidas facilmente, então um amostrador Metropolis será utilizado. O cálculo da densidade a posteriori é obtido através da seguinte fórmula:

Posteriori $\propto$ Verossimilhança X Priori

As posterioris condicionais necessárias para o amostrador de Gibbs/Metropolis são:

## i) Obtenção de $\beta/\boldsymbol{\tau}, I_0, \mathbf{I}, \gamma$

$$\pi\,(\beta/\boldsymbol{\tau}, I_0, \mathbf{I}, \gamma) \propto \beta^{n-1} \exp\left(-\int_{I_0}^{T} \beta X\,(t)\, Y\,(t)\, dt\right) \beta^{a-1} e^{-b\beta}$$

$$= \beta^{n+a-2} \exp\left\{-\beta\left(b + \int_{I_0}^{T} \beta X\,(t)\, Y\,(t)\, dt\right)\right\}.$$

Então, $\pi\,(\beta/\boldsymbol{\tau}, I_0, \mathbf{I}, \gamma) \sim \Gamma\left(a + n - 1, b + \int_{I_0}^{T} X\,(t)\, Y\,(t)\, dt\right)$.

## ii) Obtenção de $\gamma/\boldsymbol{\tau}, I_0, \mathbf{I}, \beta$

$$\pi\,(\gamma/\boldsymbol{\tau}, I_0, \mathbf{I}, \beta) \propto \gamma^{n} \exp\left(-\gamma \int_{I_0}^{T} Y\,(t)\, dt\right) \gamma^{c-1} e^{-d\gamma}$$

$$= \gamma^{n+c-1} \exp\left\{-\gamma\left(d + \int_{I_0}^{T} Y(t)\, dt\right)\right\}.$$

Então, $\pi(\gamma/\boldsymbol{\tau}, I_0, \mathbf{I}, \beta) \sim \Gamma\left(n + c, d + \int_{I_0}^{T} Y(t)\, dt\right)$.

### iii) Obtenção de $-I_0/\boldsymbol{\tau}, \beta, \mathbf{I}, \gamma$

No caso de termos dados incompletos, ou seja, não temos o vetor de infectados, temos que achar a posteriori condicional para $I_0$. Desse modo, utiliza-se o fato de $-I_0$ ter distribuição $Exp(\theta)$.

$$\pi(-I_0/\boldsymbol{\tau}, \beta, \mathbf{I}, \gamma) \propto f(\boldsymbol{\tau}, \mathbf{I}/\beta, \gamma, I_0)\, \pi(-I_0)$$
$$= \prod_{i=1}^{n} \gamma Y\left(\tau_i^-\right) \prod_{i=1}^{n-1} \beta X\left(I_i^-\right) Y\left(I_i^-\right) \exp\left(-\int_{I_0}^{T} \left[\beta X(t) Y(t) + \gamma Y(t)\right] dt\right) \theta e^{\theta I_0}$$
$$\propto \exp\left(-\int_{I_0}^{I_1} \left[\beta X(t) Y(t) + \gamma Y(t)\right] dt\right) \theta e^{\theta I_0}.$$

Como $X(t) = N$ e $Y(t) = 1$ para $I_0 \leq t < I_1$, as integrais acima tornam:

$$\int_{I_0}^{I_1} Y(t)\, dt = I_1 - I_0 \text{ e } \int_{I_0}^{I_1} X(t) Y(t)\, dt = N(I_1 - I_0).$$

resultando em $\pi(-I_0/\boldsymbol{\tau}, \beta, \mathbf{I}, \gamma) \propto \exp\{-\beta N(I_1 - I_0) - \gamma(I_1 - I_0) - \theta(-I_0)\}$.

Ou seja, temos que:

$$\pi(-I_0/\boldsymbol{\tau}, \beta, \mathbf{I}, \gamma) \sim \text{Exp}(\beta N + \gamma + \theta).$$

iv) Obtenção de $\mathbf{I}/\boldsymbol{\tau}, \beta, -I_0, \gamma$

Sortear $t_1$ e $t_2$ uniformes em $(I_0, T)$. Fazer $Y(t_1) = Y(t_1) - 1$ e $Y(t_2) = Y(t_2) + 1$. A mudança somente ocorrerá se $I_i \leq \tau_i$ para $1 \leq i \leq m-1$ e será aceita com probabilidade $min\left[1, \frac{L(\mathbf{Inovo})}{L(\mathbf{Iantigo})}\right]$. $L(\mathbf{Inovo})$ é o cálculo da equação 2.2 utilizando o vetor de infectados depois dos 2 sorteios e $L(\mathbf{Iantigo})$ é o cálculo da equação 2.2 utilizando o vetor de infectados antes dos 2 sorteios.

No passo iv, consideramos o tempo de maneira discreta, devido ao fato de que, na prática, temos os tempos de remoção em dias, ou seja, da forma discreta.

Se temos dados completos e tomamos a média a posteriori como estimador bayesiano, os estimadores bayesianos para $\beta$ e $\gamma$ são dados por:

$$\tilde{\beta} = \frac{\nu_\beta + n - 1}{\lambda_\beta + \int_{I_0}^{T} X(t)\, Y(t)\, dt},$$

e

$$\tilde{\gamma} = \frac{n + \nu_\gamma}{\lambda_\gamma + \int_{I_0}^{T} Y(t)\, dt}.$$

No caso de prioris difusas, ou seja, $a = b = c = d = 0$ os estimadores de máxima verossimilhança e os estimadores bayesianos são idênticos.

Então, o algoritmo (algoritmo 3) de estimação bayesiana para $\beta$ e $\gamma$, quando temos o vetor de infectados desconhecido e apenas o vetor de removidos conhecido, será dado por:

1. Inicializar o vetor de infectados com uma configuração viável segundo a condição 2.1.

Isto pode ser feito, por exemplo, colocando no tempo I(0) os m+1 casos de infecções e 0 nos outras tempos de infecção. Dar um chute inicial para $-I_0$.

2. Gerar as posterioris condicionais para $\beta$ e $\gamma$ conforme i e ii.

3. Gerar a posteriori condicional para $-I_0$ conforme iii, utilizando as amostras de $\beta$ e $\gamma$ do passo 2.

4. Gerar a posteriori condicional para $\mathbf{I}$ conforme iv.

5. Repetir os passos 2 a 4 até a convergência da cadeia.

Depois da convergência da cadeia, tomar amostras espaçadas das posterioris condicionais de $\beta$ e $\gamma$ de modo que as amostras não tenham autocorrelação significativa, para termos amostras independentes.

Há outras formas de estimação dos parâmetros quando temos dados incompletos. Por exemplo, em Andersson e Britton (2000) encontra-se estimativas baseadas no algoritmo EM e em martingais.

# 3 Simulações e Exemplo de Aplicação

## 3.1 Simulações

Neste capítulo, serão realizadas diversas simulações de uma epidemia com os parâmetros $\beta$ e $\gamma$ conhecidos. Depois de simular a epidemia, os parâmetros serão estimados de acordo com as duas técnicas já discutidas: máxima verossimilhança e bayesiana. Primeiramente, as estimativas serão feitas baseadas no fato de que temos os dados completos, ou seja, observamos todos os tempos de remoção e infecção. Em seguida, as estimativas serão feitas partindo do pressuposto de que temos somente os dias de remoção e a quantidade de removidos por dia, ou seja, teremos dados incompletos devido ao fato de não termos a observação dos tempos de infecção. No final do capítulo terá uma aplicação com dados reais, os dados de varíola (Bailey, 1975).

Para estimarmos os parâmetros com dados completos, serão simuladas as mesmas epidemias do exemplo 1. Conforme já visto, os parâmetros utilizados são $\beta = 0,005$, $\gamma = 0,08$ e $R_0 = \frac{\beta N}{\gamma} = 3,125$, com 49 indivíduos suscetíveis e 1 indivíduo infectante no início. A escolha dos parâmetros foi baseada no gráfico de probabilidades do tamanho final do exemplo 1 (figura 4). Os parâmetros escolhidos apresentam razoável probabilidade em diversos tamanhos finais da epidemia. É necessário que se tenha diversos tamanhos finais nas simulações, pois neste capítulo será feito um estudo das qualidades das estimativas de acordo com o tamanho final. Quando consideramos o tempo contínuo, os resultados dependem apenas da relação $\frac{\beta}{\gamma}$. Adotamos $\beta = 0,005$ e $\gamma = 0,08$ pois estes valores são da mesma ordem de grandeza dos valores estimados no exemplo com dados reais. O valor

absoluto dos parâmetros funciona apenas como escala temporal. Na prática, quando são observados apenas o número de infectados no final do dia, temos uma discretização e a escala temporal tem pouca influência nos resultados.

Para cada simulação foi calculado um intervalo de confiança assintótico 95% para $\beta$, $\gamma$ e $R_0$ na estimativa por máxima verossimilhança para dados completos e, na abordagem bayesiana, foi calculado o intervalo de credibilidade 95% para os parâmetros. Para a abordagem de máxima verossimilhança para dados incompletos, toma-se os quantis da amostra final para se construir um intervalo de confiança.

Para as estimativas bayesianas, duas prioris foram utilizadas:

- Priori imprópria: $\beta \sim \Gamma(0;0)$ e $\gamma \sim \Gamma(0;0)$.

- Priori informativa: $\beta \sim \Gamma(0,5;100)$ - Média: 0,005 e desvio padrão: 0,0071 e $\gamma \sim \Gamma(0,8;10)$ - Média: 0,08 e desvio-padrão: 0,089.

em que $\Gamma(\nu;\lambda)$ denota distribuição gamma com média $\nu/\lambda$. A partir de agora, a priori imprópria será chamada de priori não informativa, mesma nomenclatura utilizada por O'Neill e Roberts (1999).

Quando temos dados incompletos, a priori utilizada para $-I_0$ foi $-I_0 \sim exp(0,1)$. Então, a média a priori para $-I_0$ é 10 e o desvio-padrão é 10. Para a estimação de máxima verossimilhança, considerou que $-I_0$ possui distribuição U(0,4), ou seja, $\theta$=4. Fez-se algumas estimações para $\beta$ e $\gamma$ com este valor de $\theta$ e percebeu-se que as estimações apresentaram valores razoáveis. Höhle e Jorgensen (2002) também utilizaram distribuição uniforme para $-I_0$ no algoritmo de máxima verossimilhança para dados incompletos.

Para as estimativas bayesianas e de máxima verossimilhança com dados incompletos, foi utilizado um burn-in de 200 foi realizado e depois tomou-se amostras espaçadas de 20 em 20 para termos amostras com pequena dependência. Para se verificar a convergência das estimativas foram feitos gráficos de séries temporais, conforme mostrado na figura 27 e para verificar se a independência da amostra final, fez-se gráficos das autocorrelações, conforme mostrado na figura 28.

As figuras 7 a 12 mostram gráficos de dispersão para as três estimativas (máxima

verossimilhança, bayesiana informativa e bayesiana não informativa), considerando dados completos e incompletos. Para construir os gráficos, foram considerados 3490 casos nas estimações com dados completos e 300 casos com dados incompletos. Só foram considerados casos com tamanhos finais da epidemia maiores que cinco.

Com o intuito de se verificar a qualidade das estimativas de acordo com o tamanho final da epidemia, dividiu-se os dados em dez categorias distintas de acordo com a faixa de tamanho final (m). Tais categorias foram divididas baseadas nas probabilidades teóricas dos tamanhos finais, que foram calculadas no exemplo 1, capítulo 1, figura 4. As faixas são:

- Tamanho final de 5 a 9;
- Tamanho final de 10 a 19;
- Tamanho final de 20 a 29;
- Tamanho final de 30 a 39;
- Tamanho final de 40 a 44;
- Tamanho final de 45;
- Tamanho final de 46;
- Tamanho final de 47;
- Tamanho final de 48;
- Tamanho final de 49.

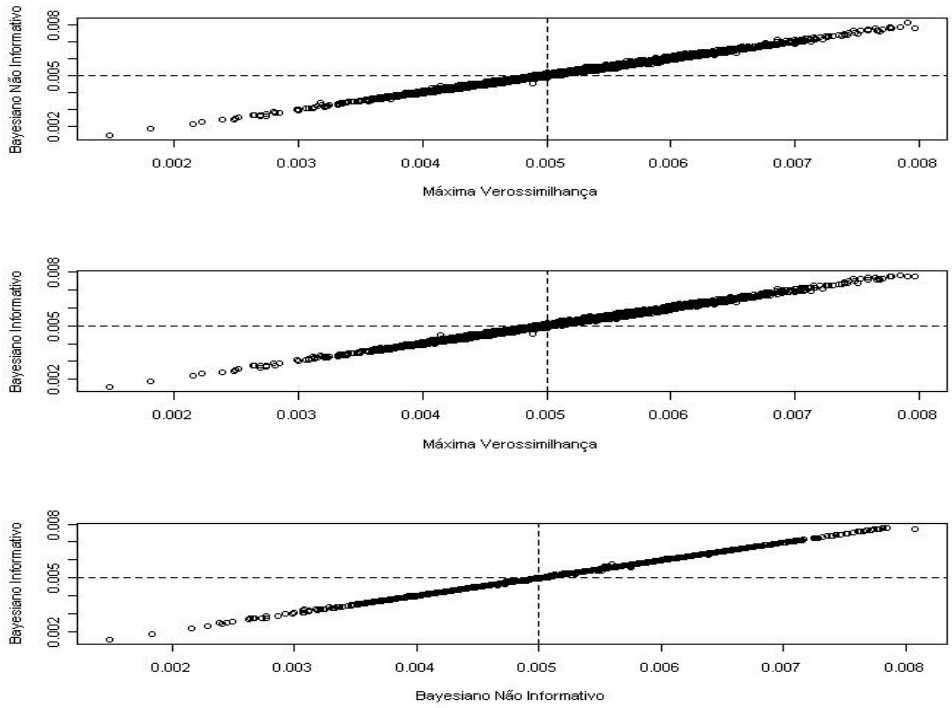



Figura 7: Estimativas para $\beta$ com dados completos. O valor verdadeiro do parâmetro é 0,005 e é a linha pontilhada nos gráficos.

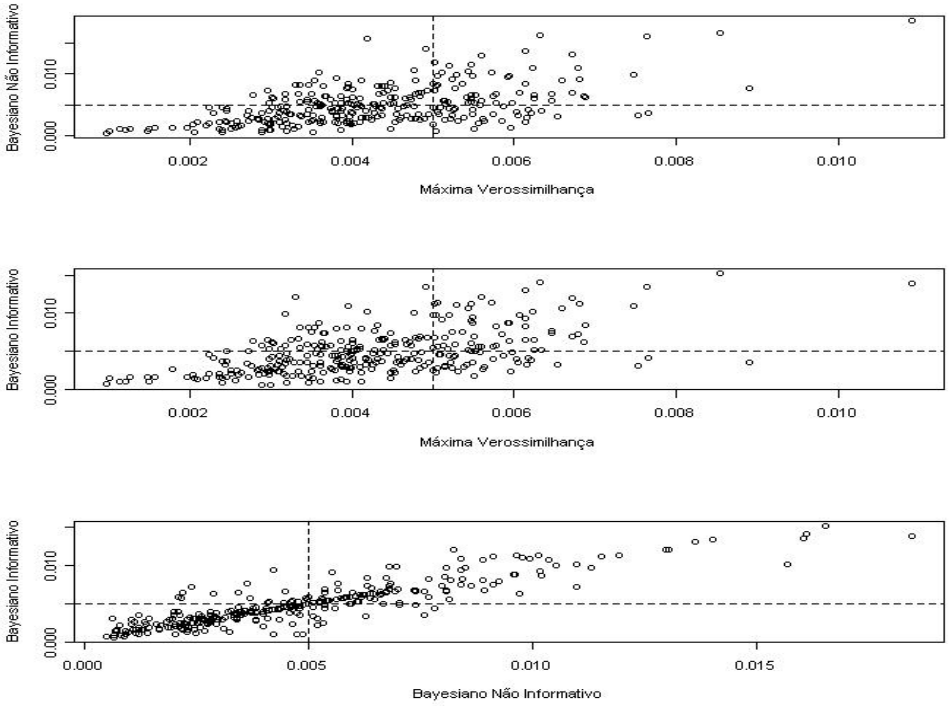



Figura 8: Estimativas para $\beta$ com dados incompletos. O valor verdadeiro do parâmetro é 0,005 e é a linha pontilhada nos gráficos.

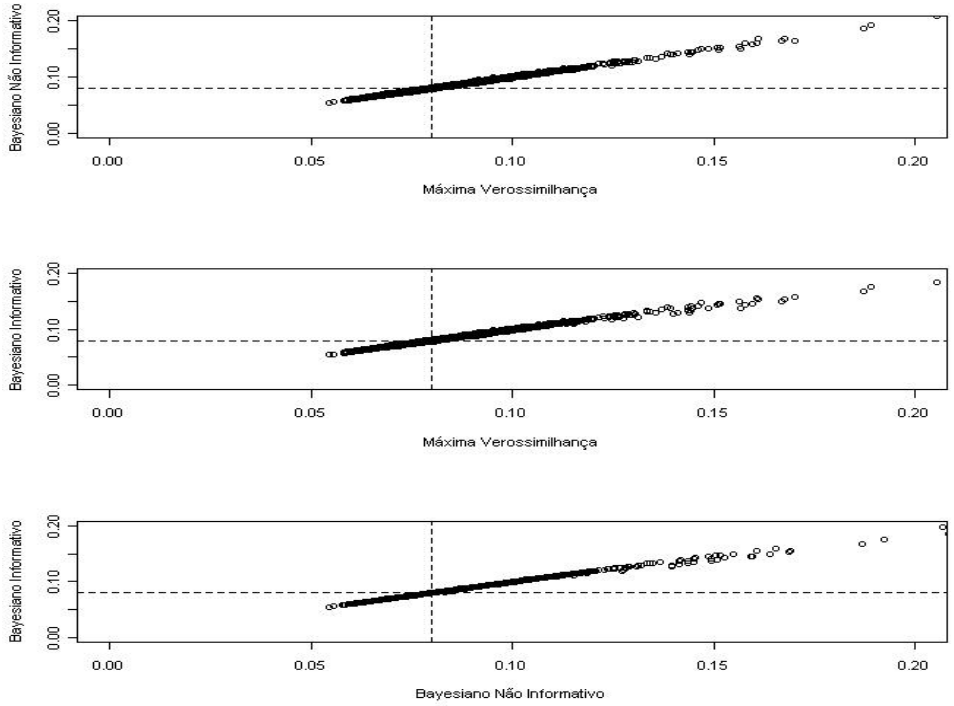



Figura 9: Estimativas para $\gamma$ com dados completos. O valor verdadeiro do parâmetro é 0,08 e é a linha pontilhada nos gráficos. Há alguns outliers nas estimativas. Os gráficos compreendem o intervalo de 0 a 0,2.

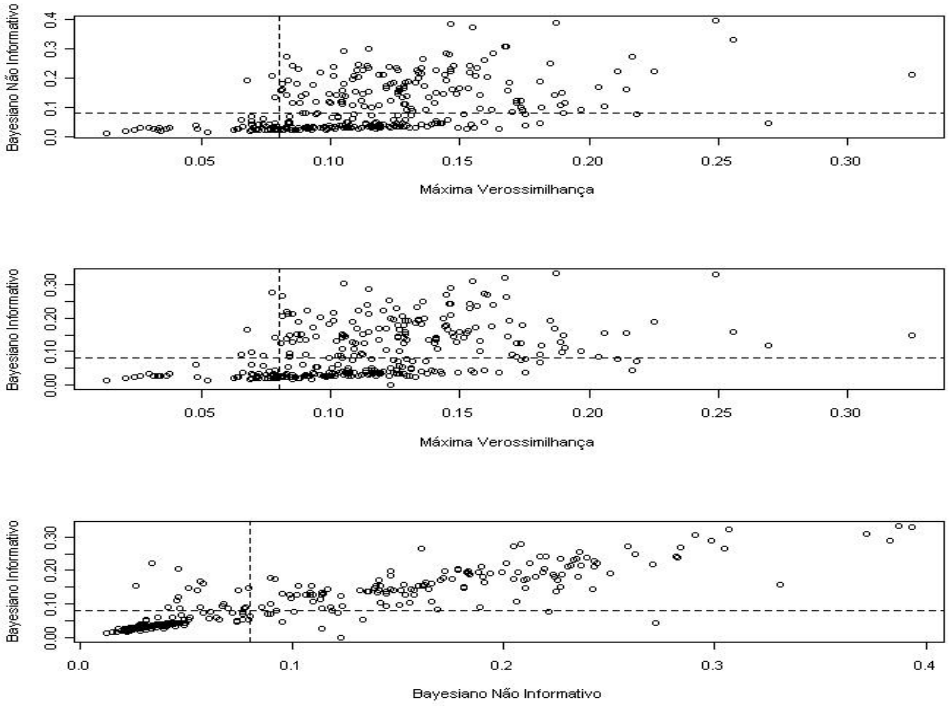



Figura 10: Estimativas para $\gamma$ com dados incompletos. O valor verdadeiro do parâmetro é 0,08 e é a linha pontilhada nos gráficos.

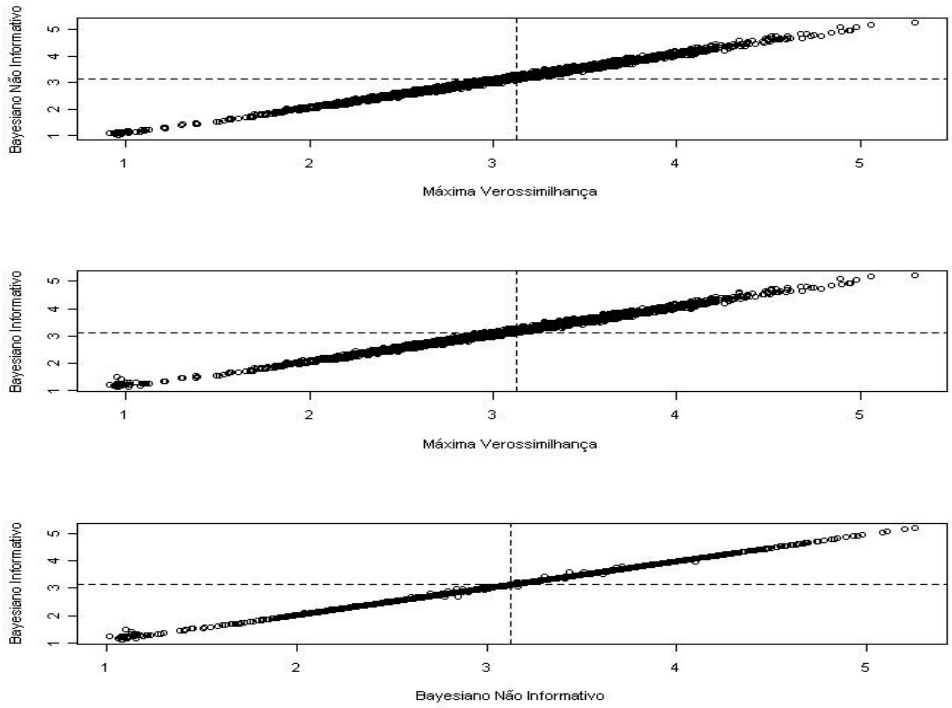



Figura 11: Estimativas para $R_0$ com dados completos. O valor verdadeiro do parâmetro é 3,125 e é a linha pontilhada nos gráficos.

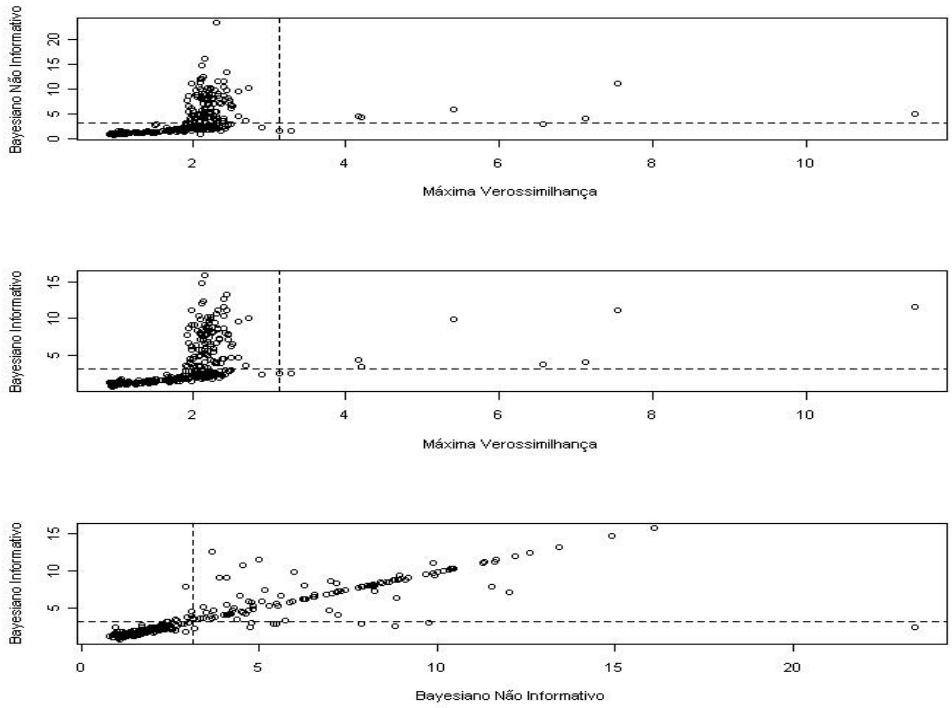



Figura 12: Estimativas para $R_0$ com dados incompletos. O valor verdadeiro do parâmetro é 3,125 e é a linha pontilhada nos gráficos.

As figuras 7, 9 e 11 apresentam gráficos de dispersão das estimativas por máxima verossimilhança, bayesiana informativa e não informativa para dados completos. A análise dos gráficos mostra que as três estimativas são muito parecidas para os três parâmetros. As figuras 8, 10 e 12 comparam as três metodologias de estimação para o caso de dados incompletos. Ao observamos os gráficos, vemos que as estimativas para $\beta$ são parecidas entre as metodologias, contudo há algumas estimativas para $\gamma$ e $R_0$ que são diferentes ao compararmos os métodos bayesianos com o de máxima verossimilhança. As estimativas bayesianas para $R_0$ , em geral, produzem valores mais altos do que as de máxima verossimilhança quando temos dados incompletos.

A figura 13 mostra gráficos de dispersão das estimativas com dados completos vs. estimativas com dados incompletos para os três parâmetros utilizando as três metodologias. Observa-se que os gráficos de dispersão estão próximos a uma reta, indicando que há correlação entre as estimativas com dados completos e incompletos para as três metodolo-

gias utilizadas. Para gerar estes gráficos, 300 estimativas com dados completos e 300 estimativas com dados incompletos foram utilizadas.

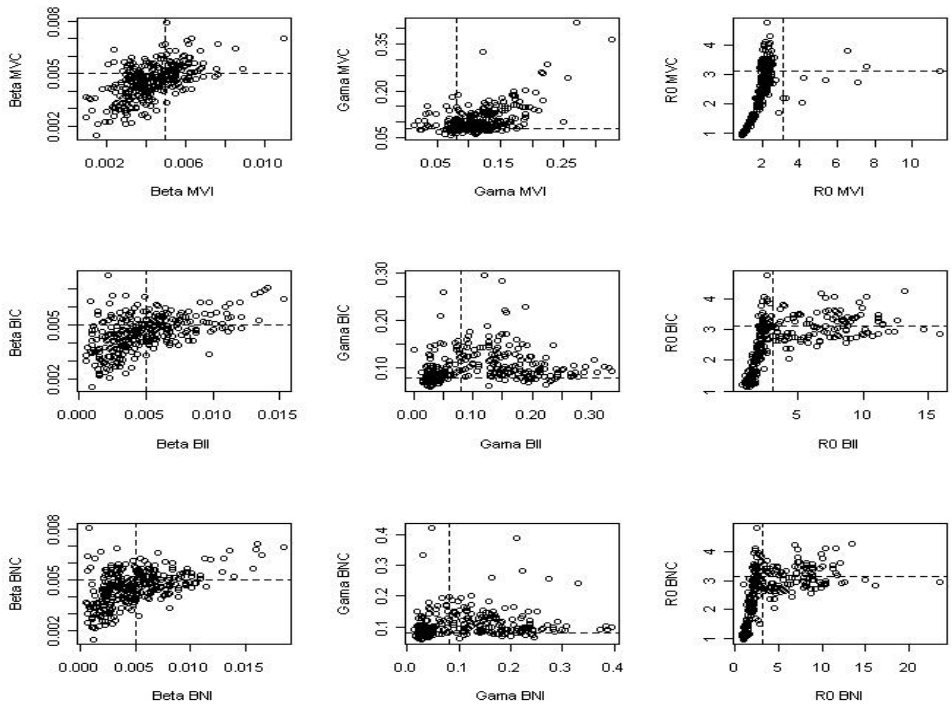



Figura 13: Estimativas para dados completos vs. estimativas para dados incompletos. As linhas pontilhadas são os valores reais dos parâmetros.

Como todas as estimativas dependem do tamanho final da epidemia (m), serão feitos gráficos de dispersão das estimativas versus o valor do tamanho final. Pelas figuras 14, 15, 16, 17, 18 e 19 vemos que as estimativas estão muito relacionadas com o tamanho final da epidemia. Para $\beta$ e $R_0$, conforme aumentamos o valor do tamanho final, aumentamos o valor das estimativas dos parâmetros. Para $\gamma$, conforme aumentamos o valor do tamanho final, diminuímos o valor de sua estimativa. Observa-se também que o uso de prioris informativas produziu estimativas melhores para os parâmetros em tamanhos finais da epidemia menores. Muitos outliers foram retirados através destas prioris informativas. Também pode-se notar que nas estimativas bayesianas temos $R_0$ maiores que nas estimativas por máxima verossimilhança. Para as estimativas com dados completos, 3490 casos foram utilizados para se construir e para as estimativas com dados incompletos, 300 casos

foram utilizados.

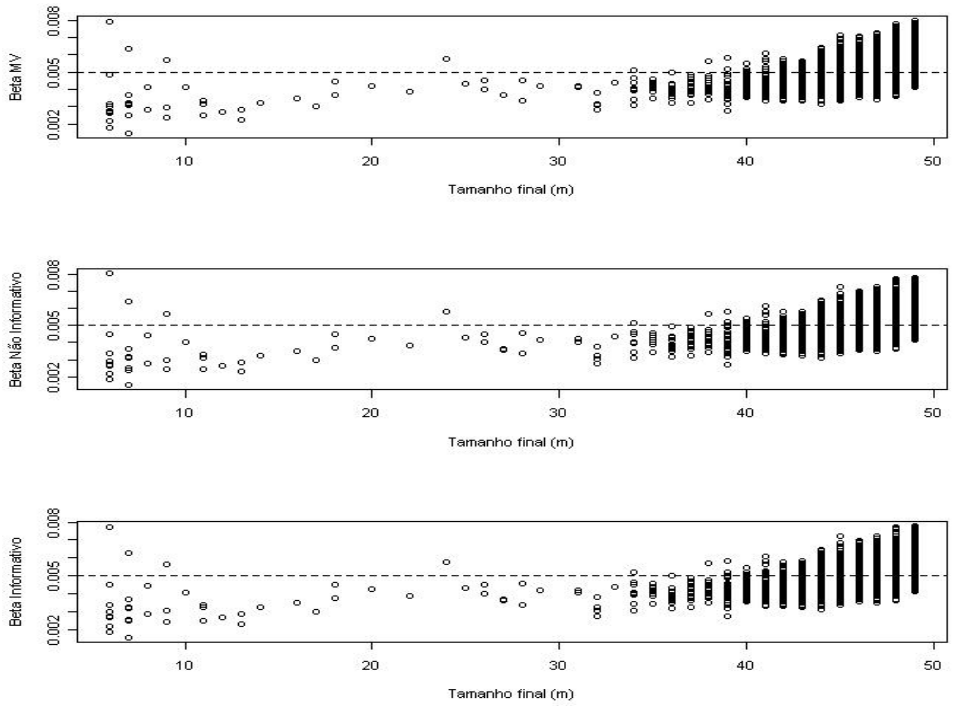



Figura 14: Estimativas para $\beta$ com dados completos vs. Tamanho final (m). O valor verdadeiro do parâmetro é 0,005 e é a linha pontilhada nos gráficos.

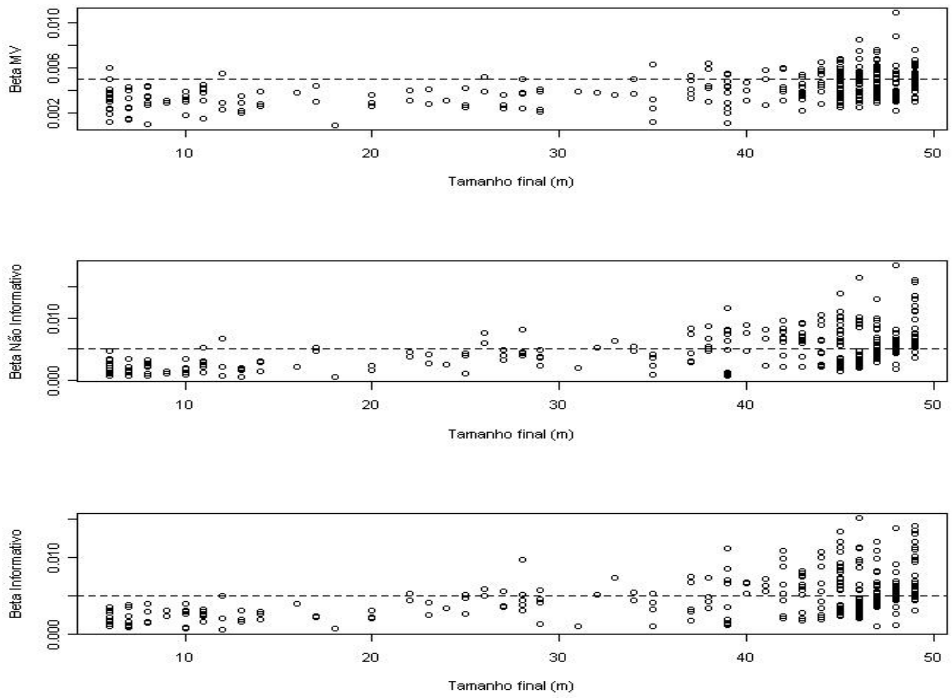



Figura 15: Estimativas para $\beta$ com dados incompletos vs. Tamanho final (m). O valor verdadeiro do parâmetro é 0,005 e é a linha pontilhada nos gráficos.

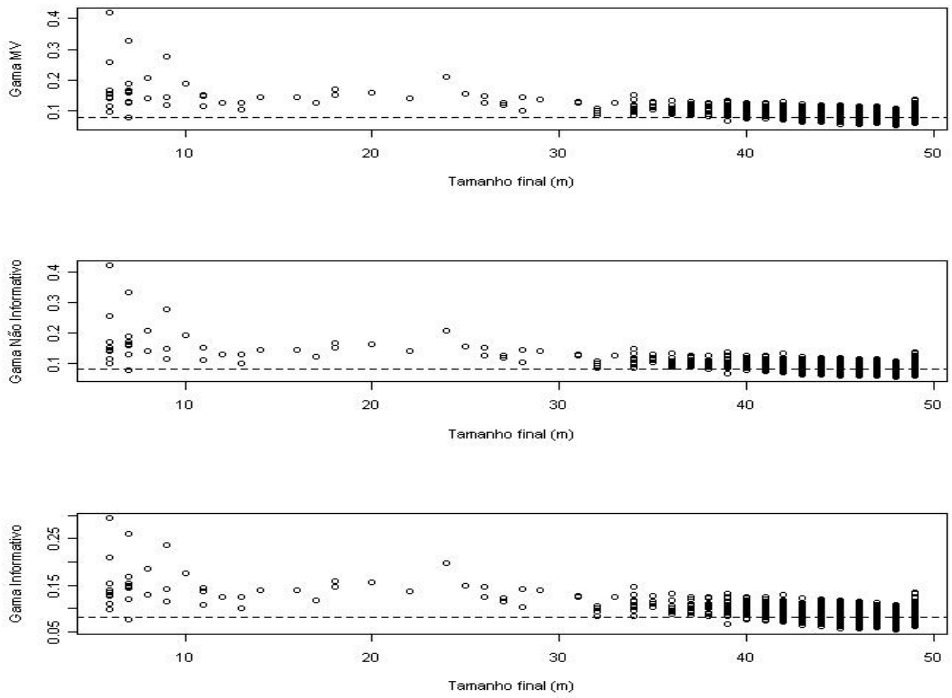



Figura 16: Estimativas para $\gamma$ com dados completos vs. Tamanho final (m). O valor verdadeiro do parâmetro é 0,08 e é a linha pontilhada nos gráficos.

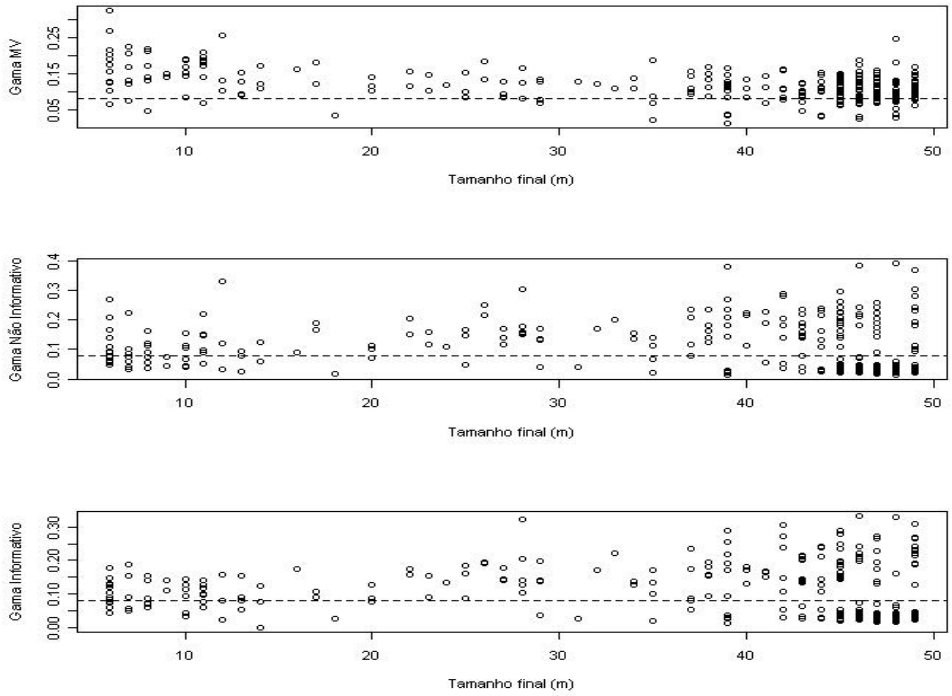



Figura 17: Estimativas para $\gamma$ com dados incompletos vs. Tamanho final (m). O valor verdadeiro do parâmetro é 0,08 e é a linha pontilhada nos gráficos.

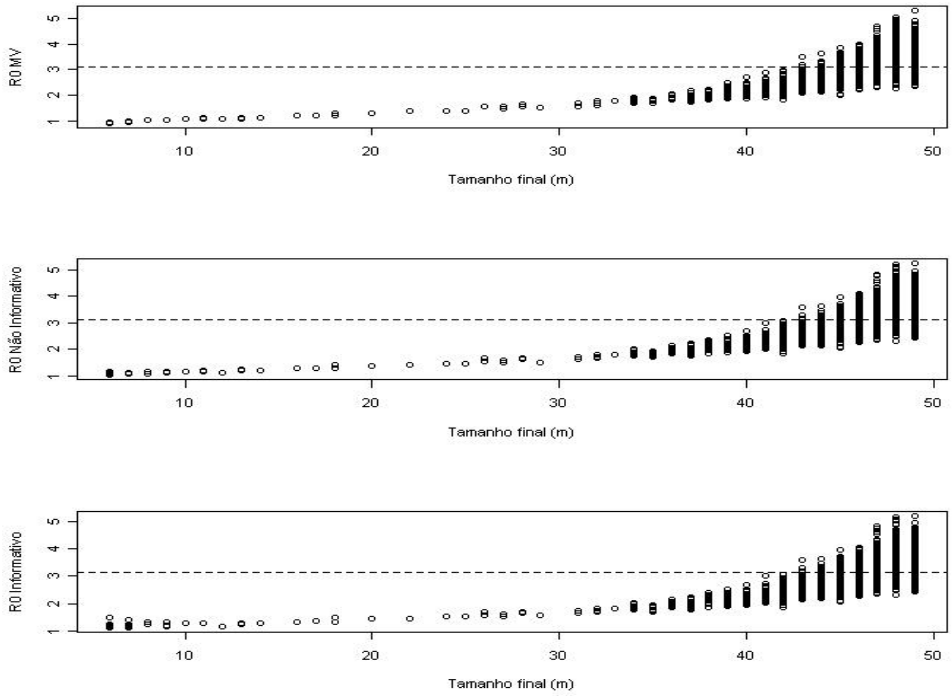



Figura 18: Estimativas para $R_0$ com dados completos vs. Tamanho final (m). O valor verdadeiro do parâmetro é 3,125 e é a linha pontilhada nos gráficos.

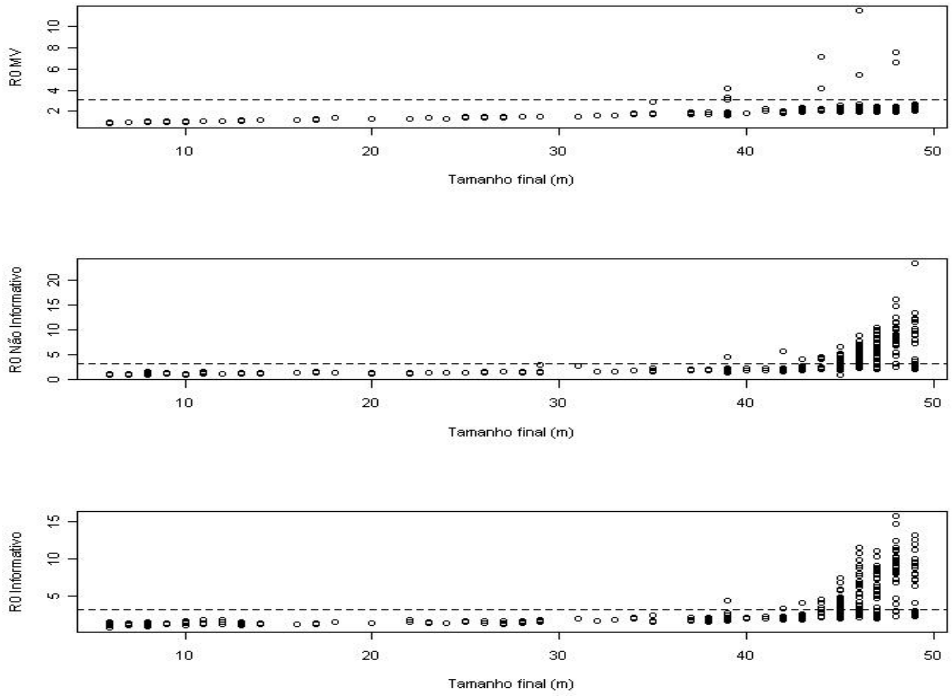



Figura 19: Estimativas para $R_0$ com dados incompletos vs. Tamanho final (m). O valor verdadeiro do parâmetro é 3,125 e é a linha pontilhada nos gráficos.

As estatísticas descritivas das estimativas por faixas de tamanho final e seus respectivos boxplots serão apresentados nas tabelas 1 a 6. A coluna % Cobert nas tabelas corresponde ao percentual de casos em que o verdadeiro valor do parâmetro está contido no intervalo de confiança ou credibilidade 95 % . A coluna EQM nas tabelas corresponde à média dos erros quadráticos médios das estimativas presentes na faixa de tamanho final. A coluna N. Rep. corresponde ao número de epidemias simuladas pelo método na respectiva faixa de tamanho final. Na faixa total das tabelas, a cobertura dos intervalos de confiança ou credibilidade, a média e o EQM médio estão ponderados pelas probabilidades calculadas no exemplo 1, capítulo 1, figura 4.

Os gráficos 20, 22 e 24 correspondem a boxplots com 30 observações para cada faixa de tamanho final tanto para dados completos, como para dados incompletos. Os gráficos 21, 23 e 25 são boxplots com 30 séries simuladas em cada faixa para dados incompletos, e com os resultados de todas as 3490 séries simuladas para dados completos (o número de séries simuladas por faixa de tamanho final é dado nas tabelas 1, 3 e 5). Nestes últimos gráficos, não serão mostrados os outliers (whiskers).

Tabela 1: Resumo das estimativas de $\beta$: Máxima Verossimilhança Completa (MVC) e Incompleta (MVI), Bayesiana Não Informativa Completa (BNC) e Incompleta (BNI) e Bayesiana Informativa Completa (BIC) e Incompleta. (BII).* Os valores de EQM estão multiplicados por $10^5$.

| m | Método | N. Rep. | Mínimo | $1^0$ Quart | Mediana | Média | $3^0$ Quart | Máximo | D.P. | % Cobert | EQM* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Total | MVC | 3490 | 0,00148 | 0,00446 | 0,00496 | 0,005 | 0,00551 | 0,00797 | 0,00084 | 94 | 0,12 |
| | MVI | 300 | 0,001 | 0,0033 | 0,0041 | 0,0047 | 0,0052 | 0,0109 | 0,0014 | 52 | 0,26 |
| | BNC | 3490 | 0,00149 | 0,00446 | 0,00496 | 0,005 | 0,00551 | 0,00807 | 0,00084 | 93 | 0,13 |
| | BIC | 3490 | 0,00157 | 0,00447 | 0,00496 | 0,005 | 0,0055 | 0,0078 | 0,00083 | 93 | 0,13 |
| | BNI | 300 | 0,0005 | 0,0026 | 0,0044 | 0,0058 | 0,0064 | 0,0185 | 0,0031 | 56 | 1,33 |
| | BII | 300 | 0,0005 | 0,0028 | 0,0044 | 0,0057 | 0,0065 | 0,0152 | 0,0029 | 57 | 1,13 |
| 5-9 | MVC | 52 | 0,00148 | 0,00265 | 0,00301 | 0,00348 | 0,00368 | 0,00791 | 0,00148 | 60 | 0,66 |
| | MVI | 30 | 0,00102 | 0,00242 | 0,00314 | 0,00315 | 0,00392 | 0,00604 | 0,00114 | 37 | 0,6 |
| | BNC | 52 | 0,00149 | 0,00262 | 0,00304 | 0,00348 | 0,00365 | 0,00807 | 0,00151 | 69 | 0,69 |
| | BIC | 52 | 0,00157 | 0,00272 | 0,00315 | 0,00352 | 0,00372 | 0,00774 | 0,00142 | 71 | 0,63 |
| | BNI | 30 | 0,00065 | 0,0011 | 0,00159 | 0,00187 | 0,00255 | 0,0047 | 0,00101 | 10 | 1,21 |
| | BII | 30 | 0,00088 | 0,00148 | 0,00163 | 0,00217 | 0,00303 | 0,00399 | 0,00096 | 33 | 1,05 |
| 10-19 | MVC | 38 | 0,00209 | 0,0028 | 0,00315 | 0,00324 | 0,00356 | 0,00539 | 0,0007 | 42 | 0,44 |
| | MVI | 30 | 0,00098 | 0,00259 | 0,00319 | 0,00326 | 0,00397 | 0,00553 | 0,00102 | 30 | 0,47 |
| | BNC | 38 | 0,00213 | 0,00283 | 0,00311 | 0,00324 | 0,00353 | 0,0053 | 0,0007 | 50 | 0,45 |
| | BIC | 38 | 0,00218 | 0,00288 | 0,00326 | 0,00329 | 0,00356 | 0,00529 | 0,00069 | 50 | 0,43 |
| | BNI | 30 | 0,00049 | 0,00131 | 0,00209 | 0,00243 | 0,00305 | 0,00666 | 0,00154 | 22 | 1 |
| | BII | 30 | 0,00054 | 0,00189 | 0,00244 | 0,00236 | 0,00291 | 0,00498 | 0,00103 | 11 | 0,88 |
| 20-29 | MVC | 35 | 0,00236 | 0,00349 | 0,00389 | 0,00386 | 0,00421 | 0,0058 | 0,00062 | 74 | 0,23 |
| | MVI | 30 | 0,0021 | 0,00257 | 0,00318 | 0,00336 | 0,00402 | 0,00527 | 0,00087 | 24 | 0,38 |
| | BNC | 35 | 0,00236 | 0,00337 | 0,0039 | 0,00386 | 0,00423 | 0,0058 | 0,00062 | 71 | 0,23 |
| | BIC | 35 | 0,00239 | 0,00339 | 0,00391 | 0,00388 | 0,00425 | 0,00579 | 0,00062 | 71 | 0,23 |
| | BNI | 30 | 0,0011 | 0,00266 | 0,00404 | 0,00398 | 0,00458 | 0,00806 | 0,00162 | 56 | 0,46 |
| | BII | 30 | 0,00127 | 0,00304 | 0,00414 | 0,00417 | 0,00506 | 0,00969 | 0,00168 | 52 | 0,44 |
| 30-39 | MVC | 117 | 0,00274 | 0,00378 | 0,00409 | 0,00413 | 0,0044 | 0,00587 | 0,00057 | 76 | 0,15 |
| | MVI | 30 | 0,00113 | 0,00317 | 0,00405 | 0,00401 | 0,00495 | 0,00646 | 0,00135 | 43 | 0,32 |
| | BNC | 117 | 0,00273 | 0,0038 | 0,00414 | 0,00413 | 0,00446 | 0,00585 | 0,00057 | 80 | 0,16 |
| | BIC | 117 | 0,00275 | 0,00381 | 0,00415 | 0,00414 | 0,00446 | 0,00584 | 0,00057 | 80 | 0,16 |
| | BNI | 30 | 0,00066 | 0,00225 | 0,00472 | 0,00473 | 0,00692 | 0,01152 | 0,0028 | 40 | 0,88 |
| | BII | 30 | 0,00095 | 0,00177 | 0,00434 | 0,00455 | 0,00684 | 0,01117 | 0,00267 | 30 | 0,81 |
| 40-44 | MVC | 638 | 0,00313 | 0,00413 | 0,00451 | 0,00455 | 0,00491 | 0,00645 | 0,00058 | 90 | 0,1 |
| | MVI | 30 | 0,00224 | 0,00368 | 0,00433 | 0,00436 | 0,00524 | 0,00659 | 0,00107 | 51 | 0,2 |
| | BNC | 638 | 0,00312 | 0,00414 | 0,0045 | 0,00455 | 0,00492 | 0,00649 | 0,00058 | 92 | 0,11 |
| | BIC | 638 | 0,00313 | 0,00415 | 0,0045 | 0,00456 | 0,00492 | 0,00647 | 0,00058 | 92 | 0,11 |
| | BNI | 30 | 0,0021 | 0,004 | 0,0066 | 0,00614 | 0,00813 | 0,0105 | 0,00253 | 43 | 0,95 |
| | BII | 30 | 0,00172 | 0,00329 | 0,0061 | 0,00585 | 0,0076 | 0,01096 | 0,00267 | 46 | 0,95 |

Tabela 2: Resumo das estimativas de $\beta$: Máxima Verossimilhança Completa (MVC) e Incompleta (MVI), Bayesiana Não Informativa Completa (BNC) e Incompleta (BNI) e Bayesiana Informativa Completa (BIC) e Incompleta. (BII).* Os valores de EQM estão multiplicados por $10^5$.

| m | Método | N. Rep. | Mínimo | $1^0$ Quart | Mediana | Média | $3^0$ Quart | Máximo | D.P. | % Cobert | EQM* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 45 | MVC | 394 | 0,00335 | 0,00438 | 0,00476 | 0,0048 | 0,00517 | 0,00716 | 0,0006 | 96 | 0,09 |
| | MVI | 30 | 0,00258 | 0,00359 | 0,00484 | 0,00467 | 0,00568 | 0,00688 | 0,00127 | 59 | 0,23 |
| | BNC | 394 | 0,00339 | 0,00438 | 0,00476 | 0,0048 | 0,00515 | 0,00726 | 0,0006 | 96 | 0,1 |
| | BIC | 394 | 0,0034 | 0,00439 | 0,00476 | 0,00481 | 0,00516 | 0,00723 | 0,0006 | 96 | 0,1 |
| | BNI | 30 | 0,0014 | 0,00279 | 0,00367 | 0,00563 | 0,00855 | 0,01401 | 0,00347 | 28 | 1,47 |
| | BII | 30 | 0,00233 | 0,00301 | 0,00579 | 0,00604 | 0,00834 | 0,01344 | 0,00329 | 28 | 1,36 |
| 46 | MVC | 499 | 0,00357 | 0,00453 | 0,00494 | 0,005 | 0,00542 | 0,00707 | 0,00064 | 97 | 0,1 |
| | MVI | 30 | 0,0025 | 0,00362 | 0,00497 | 0,00477 | 0,00535 | 0,00853 | 0,00128 | 58 | 0,23 |
| | BNC | 499 | 0,00357 | 0,00454 | 0,00493 | 0,00501 | 0,00543 | 0,007 | 0,00065 | 96 | 0,1 |
| | BIC | 499 | 0,00358 | 0,00454 | 0,00493 | 0,00501 | 0,00542 | 0,00697 | 0,00064 | 96 | 0,1 |
| | BNI | 30 | 0,002 | 0,00253 | 0,00336 | 0,00414 | 0,00395 | 0,0165 | 0,003 | 33 | 1,09 |
| | BII | 30 | 0,00201 | 0,00289 | 0,0037 | 0,00485 | 0,00579 | 0,01523 | 0,0031 | 36 | 1,08 |
| 47 | MVC | 643 | 0,00344 | 0,00468 | 0,00507 | 0,00514 | 0,00558 | 0,00728 | 0,00067 | 98 | 0,1 |
| | MVI | 30 | 0,00297 | 0,0038 | 0,00467 | 0,00487 | 0,00596 | 0,00765 | 0,00129 | 53 | 0,23 |
| | BNC | 643 | 0,00347 | 0,00468 | 0,00505 | 0,00513 | 0,00556 | 0,00728 | 0,00067 | 96 | 0,11 |
| | BIC | 643 | 0,00348 | 0,00468 | 0,00505 | 0,00513 | 0,00556 | 0,00724 | 0,00066 | 97 | 0,11 |
| | BNI | 30 | 0,00281 | 0,00373 | 0,00469 | 0,00539 | 0,00626 | 0,013 | 0,00231 | 76 | 0,8 |
| | BII | 30 | 0,00094 | 0,00386 | 0,0044 | 0,00518 | 0,00627 | 0,01202 | 0,00239 | 79 | 0,81 |
| 48 | MVC | 642 | 0,00359 | 0,0049 | 0,00535 | 0,00539 | 0,0058 | 0,00778 | 0,00069 | 97 | 0,12 |
| | MVI | 30 | 0,00224 | 0,00346 | 0,00399 | 0,00459 | 0,00546 | 0,0109 | 0,00178 | 30 | 0,4 |
| | BNC | 642 | 0,00364 | 0,00491 | 0,00534 | 0,00539 | 0,00578 | 0,00778 | 0,0007 | 96 | 0,13 |
| | BIC | 642 | 0,00365 | 0,00491 | 0,00534 | 0,00538 | 0,00577 | 0,00773 | 0,00069 | 96 | 0,13 |
| | BNI | 30 | 0,00183 | 0,005 | 0,00561 | 0,00599 | 0,00621 | 0,01847 | 0,0027 | 87 | 1,37 |
| | BII | 30 | 0,00105 | 0,00483 | 0,00557 | 0,00575 | 0,00618 | 0,01387 | 0,00217 | 83 | 1,01 |
| 49 | MVC | 432 | 0,00411 | 0,00518 | 0,00563 | 0,00571 | 0,00619 | 0,00797 | 0,00077 | 92 | 0,18 |
| | MVI | 30 | 0,00302 | 0,00436 | 0,00514 | 0,00509 | 0,00586 | 0,00764 | 0,00106 | 79 | 0,2 |
| | BNC | 432 | 0,00414 | 0,0052 | 0,00566 | 0,00572 | 0,0062 | 0,00784 | 0,00077 | 88 | 0,18 |
| | BIC | 432 | 0,00415 | 0,0052 | 0,00565 | 0,00571 | 0,00619 | 0,0078 | 0,00076 | 88 | 0,18 |
| | BNI | 30 | 0,00361 | 0,00546 | 0,00691 | 0,00825 | 0,01115 | 0,01613 | 0,00371 | 59 | 2,94 |
| | BII | 30 | 0,00307 | 0,00542 | 0,00667 | 0,00774 | 0,01018 | 0,01409 | 0,00306 | 59 | 2,05 |

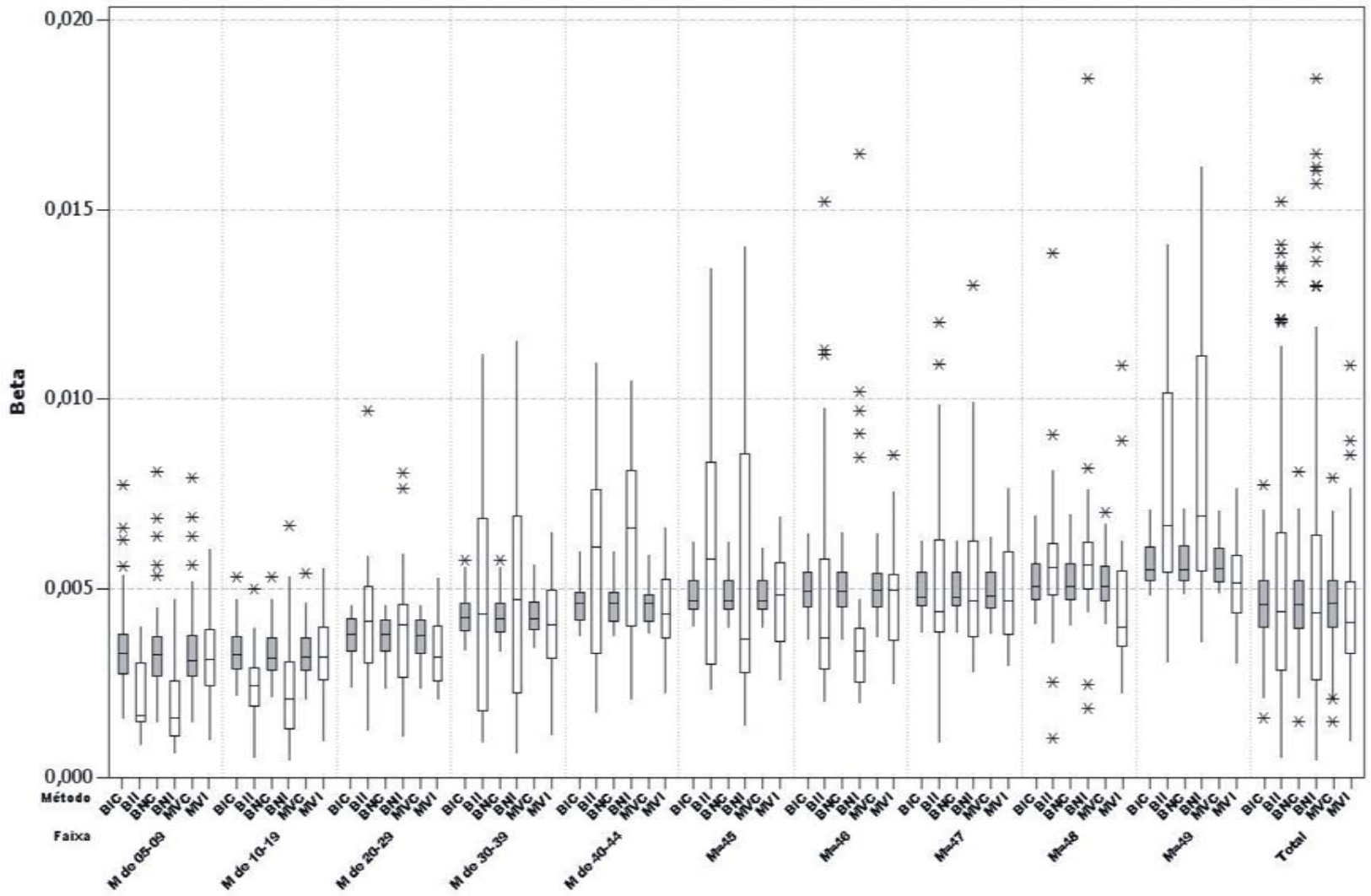



Figura 20: Boxplot para as estimativas de $\beta$ por faixa de tamanho final(30 replicações para cada faixa). O valor verdadeiro de $\beta$ é 0,005.

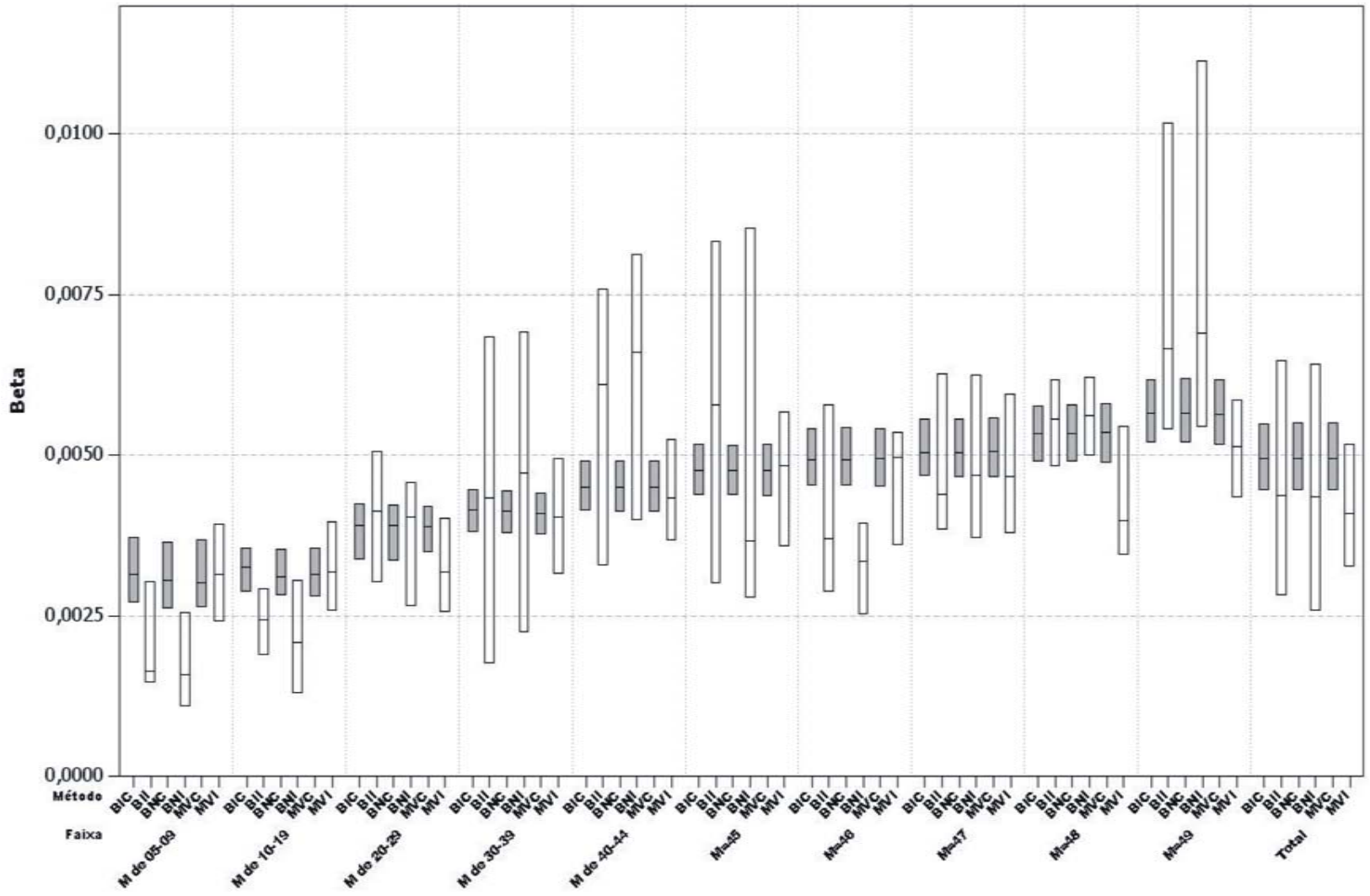



Figura 21: Boxplots com apenas os quantis, considerando todas as séries simuladas para $\beta$. O valor verdadeiro de $\beta$ é 0,005.

Tabela 3: Resumo das estimativas de $\gamma$: Máxima Verossimilhança Completa (MVC) e Incompleta (MVI), Bayesiana Não Informativa Completa (BNC) e Incompleta (BNI) e Bayesiana Informativa Completa (BIC) e Incompleta. (BII).* Os valores de EQM estão multiplicados por $10^2$.

| m | Método | N. Rep. | Mínimo | $1^0$ Quart | Mediana | Média | $3^0$ Quart | Máximo | D.P. | % Cobert | EQM* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Total | MVC | 3490 | 0,0554 | 0,0766 | 0,084 | 0,0865 | 0,0935 | 0,4175 | 0,0208 | 97 | 0,02 |
| | MVI | 300 | 0,0136 | 0,0877 | 0,1134 | 0,1072 | 0,1381 | 0,3248 | 0,042 | 41 | 0,05 |
| | BNC | 3490 | 0,0554 | 0,0768 | 0,084 | 0,0866 | 0,0935 | 0,4212 | 0,0209 | 94 | 0,07 |
| | BIC | 3490 | 0,0556 | 0,0768 | 0,0839 | 0,0861 | 0,0932 | 0,2946 | 0,0176 | 95 | 0,05 |
| | BNI | 300 | 0,0126 | 0,0322 | 0,0707 | 0,0928 | 0,1631 | 0,3933 | 0,0854 | 12 | 0,08 |
| | BII | 300 | 0,0159 | 0,0333 | 0,0764 | 0,0912 | 0,1557 | 0,3332 | 0,0784 | 8 | 0,07 |
| 5-9 | MVC | 52 | 0,0775 | 0,1401 | 0,1566 | 0,1786 | 0,1873 | 0,4175 | 0,0771 | 85 | 0,5 |
| | MVI | 30 | 0,0481 | 0,1272 | 0,164 | 0,1931 | 0,2078 | 1,1074 | 0,1828 | 7 | 4,69 |
| | BNC | 52 | 0,0768 | 0,1395 | 0,158 | 0,1792 | 0,1871 | 0,4212 | 0,0791 | 69 | 1,02 |
| | BIC | 52 | 0,0772 | 0,1275 | 0,145 | 0,1564 | 0,167 | 0,2946 | 0,0515 | 81 | 1,19 |
| | BNI | 30 | 0,0313 | 0,0574 | 0,0827 | 0,1324 | 0,1269 | 1,1074 | 0,1931 | 73 | 4,13 |
| | BII | 30 | 0,0446 | 0,0714 | 0,0909 | 0,1361 | 0,1429 | 1,1074 | 0,1878 | 63 | 3,87 |
| 10-19 | MVC | 38 | 0,0961 | 0,1252 | 0,1435 | 0,1443 | 0,1528 | 0,2426 | 0,0292 | 84 | 0,17 |
| | MVI | 30 | 0,0345 | 0,1087 | 0,1525 | 0,1465 | 0,1814 | 0,2557 | 0,0488 | 7 | 0,81 |
| | BNC | 38 | 0,099 | 0,1253 | 0,1435 | 0,1439 | 0,1543 | 0,2423 | 0,0293 | 58 | 0,27 |
| | BIC | 38 | 0,0965 | 0,1216 | 0,1374 | 0,1368 | 0,1471 | 0,2169 | 0,0248 | 68 | 0,52 |
| | BNI | 30 | 0,0179 | 0,0612 | 0,0988 | 0,108 | 0,1481 | 0,3307 | 0,0674 | 59 | 0,71 |
| | BII | 30 | 0,0181 | 0,0606 | 0,094 | 0,0934 | 0,1264 | 0,1759 | 0,0444 | 67 | 0,32 |
| 20-29 | MVC | 35 | 0,0954 | 0,1198 | 0,1285 | 0,1333 | 0,1471 | 0,21 | 0,0232 | 49 | 0,07 |
| | MVI | 30 | 0,0695 | 0,0893 | 0,1185 | 0,1182 | 0,138 | 0,1851 | 0,0305 | 36 | 0,29 |
| | BNC | 35 | 0,0967 | 0,1196 | 0,1279 | 0,1339 | 0,1507 | 0,2069 | 0,0226 | 29 | 0,14 |
| | BIC | 35 | 0,096 | 0,1175 | 0,1259 | 0,131 | 0,1468 | 0,197 | 0,0208 | 51 | 0,37 |
| | BNI | 30 | 0,0404 | 0,1156 | 0,1518 | 0,1489 | 0,1719 | 0,3065 | 0,0584 | 24 | 0,94 |
| | BII | 30 | 0,0351 | 0,1142 | 0,1419 | 0,1477 | 0,1815 | 0,3211 | 0,056 | 20 | 0,87 |
| 30-39 | MVC | 117 | 0,066 | 0,0966 | 0,1052 | 0,1064 | 0,1167 | 0,1517 | 0,0153 | 76 | 0,03 |
| | MVI | 30 | 0,0136 | 0,0884 | 0,1122 | 0,1099 | 0,1397 | 0,1867 | 0,0429 | 27 | 0,31 |
| | BNC | 117 | 0,0656 | 0,0955 | 0,1045 | 0,1064 | 0,1171 | 0,1505 | 0,0152 | 67 | 0,1 |
| | BIC | 117 | 0,0659 | 0,0951 | 0,1039 | 0,1055 | 0,1159 | 0,1478 | 0,0146 | 70 | 0,12 |
| | BNI | 30 | 0,0126 | 0,0608 | 0,1403 | 0,1383 | 0,2032 | 0,3831 | 0,0891 | 13 | 1,2 |
| | BII | 30 | 0,0126 | 0,0491 | 0,1359 | 0,1291 | 0,1839 | 0,2901 | 0,0782 | 17 | 0,93 |
| 40-44 | MVC | 638 | 0,0627 | 0,0834 | 0,0913 | 0,0918 | 0,0991 | 0,1302 | 0,0114 | 97 | 0,02 |
| | MVI | 30 | 0,0318 | 0,0853 | 0,1049 | 0,1044 | 0,1261 | 0,1626 | 0,0315 | 37 | 0,19 |
| | BNC | 638 | 0,0623 | 0,0836 | 0,0913 | 0,0918 | 0,0991 | 0,1312 | 0,0114 | 92 | 0,07 |
| | BIC | 638 | 0,0625 | 0,0835 | 0,0911 | 0,0915 | 0,0987 | 0,1297 | 0,0111 | 93 | 0,05 |
| | BNI | 30 | 0,0265 | 0,0798 | 0,158 | 0,1506 | 0,2173 | 0,291 | 0,0784 | 20 | 1,23 |
| | BII | 30 | 0,0251 | 0,0623 | 0,1452 | 0,1423 | 0,2035 | 0,3048 | 0,0765 | 11 | 1,05 |

Tabela 4: Resumo das estimativas de $\gamma$: Máxima Verossimilhança Completa (MVC) e Incompleta (MVI), Bayesiana Não Informativa Completa (BNC) e Incompleta (BNI) e Bayesiana Informativa Completa (BIC) e Incompleta. (BII).* Os valores de EQM estão multiplicados por $10^2$.

| m | Método | N. Rep. | Mínimo | $1^0$ Quart | Mediana | Média | $3^0$ Quart | Máximo | D.P. | % Cobert | EQM* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 45 | MVC | 394 | 0,0582 | 0,079 | 0,0861 | 0,0868 | 0,0935 | 0,1199 | 0,0105 | 98 | 0,02 |
| | MVI | 30 | 0,0639 | 0,0856 | 0,1177 | 0,1106 | 0,1336 | 0,1492 | 0,027 | 41 | 0,21 |
| | BNC | 394 | 0,0577 | 0,0792 | 0,0863 | 0,0868 | 0,0934 | 0,1184 | 0,0105 | 97 | 0,06 |
| | BIC | 394 | 0,058 | 0,0792 | 0,0861 | 0,0867 | 0,093 | 0,1169 | 0,0103 | 97 | 0,03 |
| | BNI | 30 | 0,0215 | 0,0342 | 0,0659 | 0,1165 | 0,2039 | 0,2987 | 0,0912 | 21 | 1,05 |
| | BII | 30 | 0,0214 | 0,0361 | 0,1445 | 0,1172 | 0,1845 | 0,287 | 0,0873 | 7 | 0,96 |
| 46 | MVC | 499 | 0,0582 | 0,0769 | 0,0833 | 0,0841 | 0,0908 | 0,1164 | 0,0102 | 98 | 0,02 |
| | MVI | 30 | 0,0244 | 0,0829 | 0,1155 | 0,1064 | 0,1235 | 0,1874 | 0,0343 | 42 | 0,23 |
| | BNC | 499 | 0,0577 | 0,077 | 0,0832 | 0,0842 | 0,0909 | 0,1155 | 0,0103 | 98 | 0,06 |
| | BIC | 499 | 0,0579 | 0,077 | 0,0832 | 0,0841 | 0,0907 | 0,1147 | 0,0101 | 98 | 0,03 |
| | BNI | 30 | 0,0212 | 0,0255 | 0,0333 | 0,064 | 0,0451 | 0,3871 | 0,081 | 6 | 0,71 |
| | BII | 30 | 0,0216 | 0,028 | 0,034 | 0,0678 | 0,0456 | 0,3331 | 0,079 | 3 | 0,65 |
| 47 | MVC | 643 | 0,0583 | 0,0738 | 0,08 | 0,0806 | 0,0865 | 0,1151 | 0,0096 | 98 | 0,01 |
| | MVI | 30 | 0,069 | 0,0907 | 0,1033 | 0,1093 | 0,1316 | 0,1594 | 0,0264 | 50 | 0,19 |
| | BNC | 643 | 0,0582 | 0,0737 | 0,08 | 0,0806 | 0,0862 | 0,1138 | 0,0096 | 98 | 0,05 |
| | BIC | 643 | 0,0585 | 0,0738 | 0,08 | 0,0806 | 0,0861 | 0,113 | 0,0095 | 98 | 0,02 |
| | BNI | 30 | 0,0171 | 0,0284 | 0,0356 | 0,0733 | 0,1036 | 0,2587 | 0,0749 | 6 | 0,61 |
| | BII | 30 | 0,0159 | 0,0275 | 0,0367 | 0,0716 | 0,0849 | 0,2714 | 0,0764 | 6 | 0,63 |
| 48 | MVC | 642 | 0,0554 | 0,0723 | 0,0785 | 0,0788 | 0,0843 | 0,1086 | 0,0094 | 97 | 0,01 |
| | MVI | 30 | 0,0265 | 0,0767 | 0,0905 | 0,0981 | 0,1235 | 0,2489 | 0,0418 | 53 | 0,25 |
| | BNC | 642 | 0,0554 | 0,072 | 0,0783 | 0,0789 | 0,0842 | 0,1077 | 0,0095 | 97 | 0,05 |
| | BIC | 642 | 0,0556 | 0,0722 | 0,0783 | 0,0789 | 0,084 | 0,1071 | 0,0093 | 98 | 0,02 |
| | BNI | 30 | 0,0152 | 0,0258 | 0,0296 | 0,0435 | 0,0364 | 0,3933 | 0,0667 | 3 | 0,58 |
| | BII | 30 | 0,0154 | 0,0261 | 0,0297 | 0,046 | 0,0361 | 0,3303 | 0,0595 | 7 | 0,49 |
| 49 | MVC | 432 | 0,0609 | 0,076 | 0,0831 | 0,0842 | 0,091 | 0,137 | 0,0119 | 96 | 0,01 |
| | MVI | 30 | 0,0624 | 0,099 | 0,1106 | 0,1125 | 0,1273 | 0,1678 | 0,0238 | 21 | 0,21 |
| | BNC | 432 | 0,0607 | 0,0759 | 0,0832 | 0,0842 | 0,0911 | 0,1366 | 0,012 | 95 | 0,05 |
| | BIC | 432 | 0,0609 | 0,076 | 0,0832 | 0,0841 | 0,0909 | 0,1351 | 0,0117 | 95 | 0,03 |
| | BNI | 30 | 0,0222 | 0,0314 | 0,0397 | 0,1103 | 0,2131 | 0,372 | 0,1084 | 14 | 1,31 |
| | BII | 30 | 0,0225 | 0,0317 | 0,0391 | 0,1061 | 0,2176 | 0,3095 | 0,0991 | 3 | 1,07 |

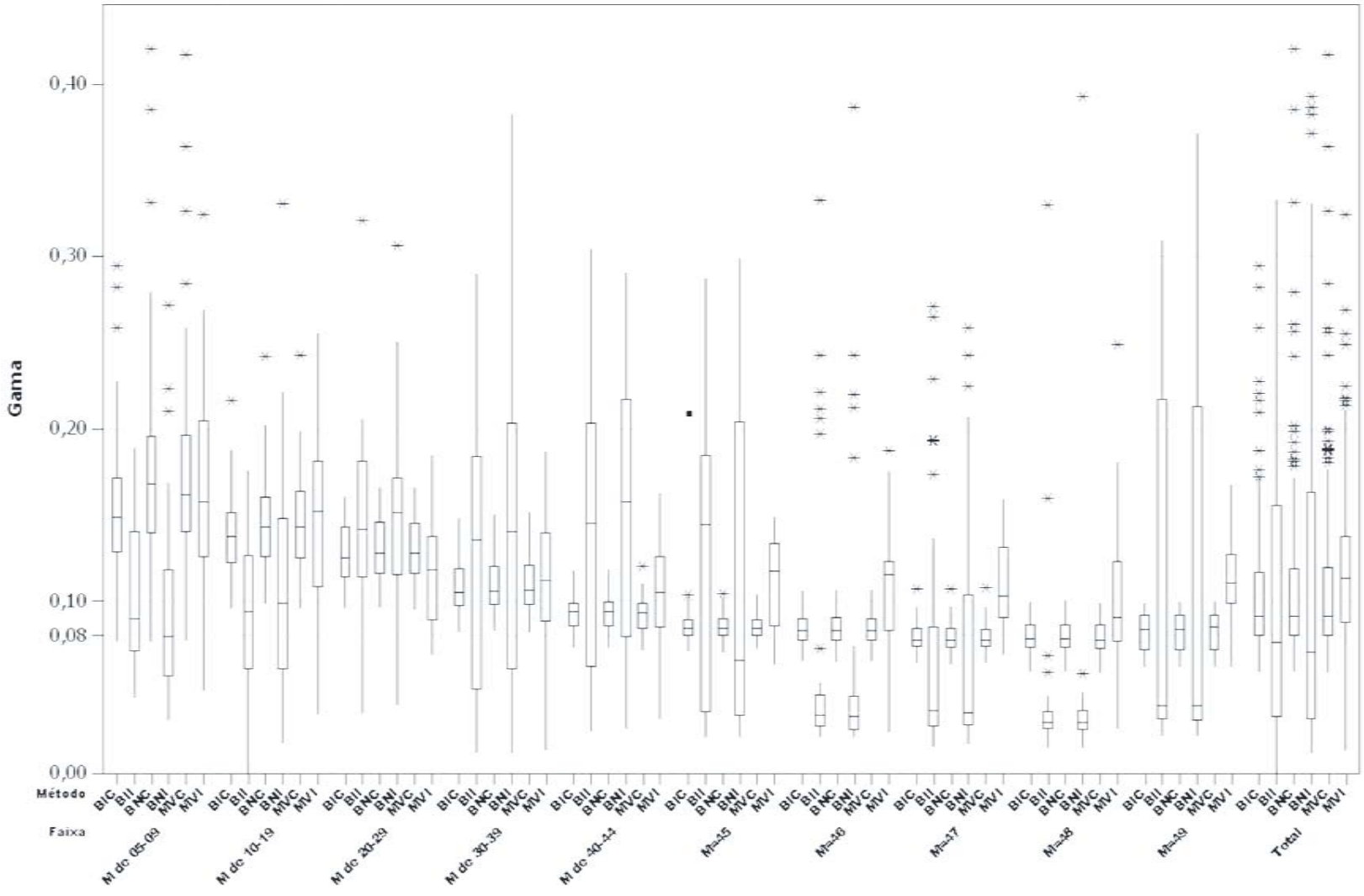



Figura 22: Boxplot para as estimativas de $\gamma$ por faixa de tamanho final(30 replicações para cada faixa). O valor verdadeiro de $\gamma$ é 0,08.

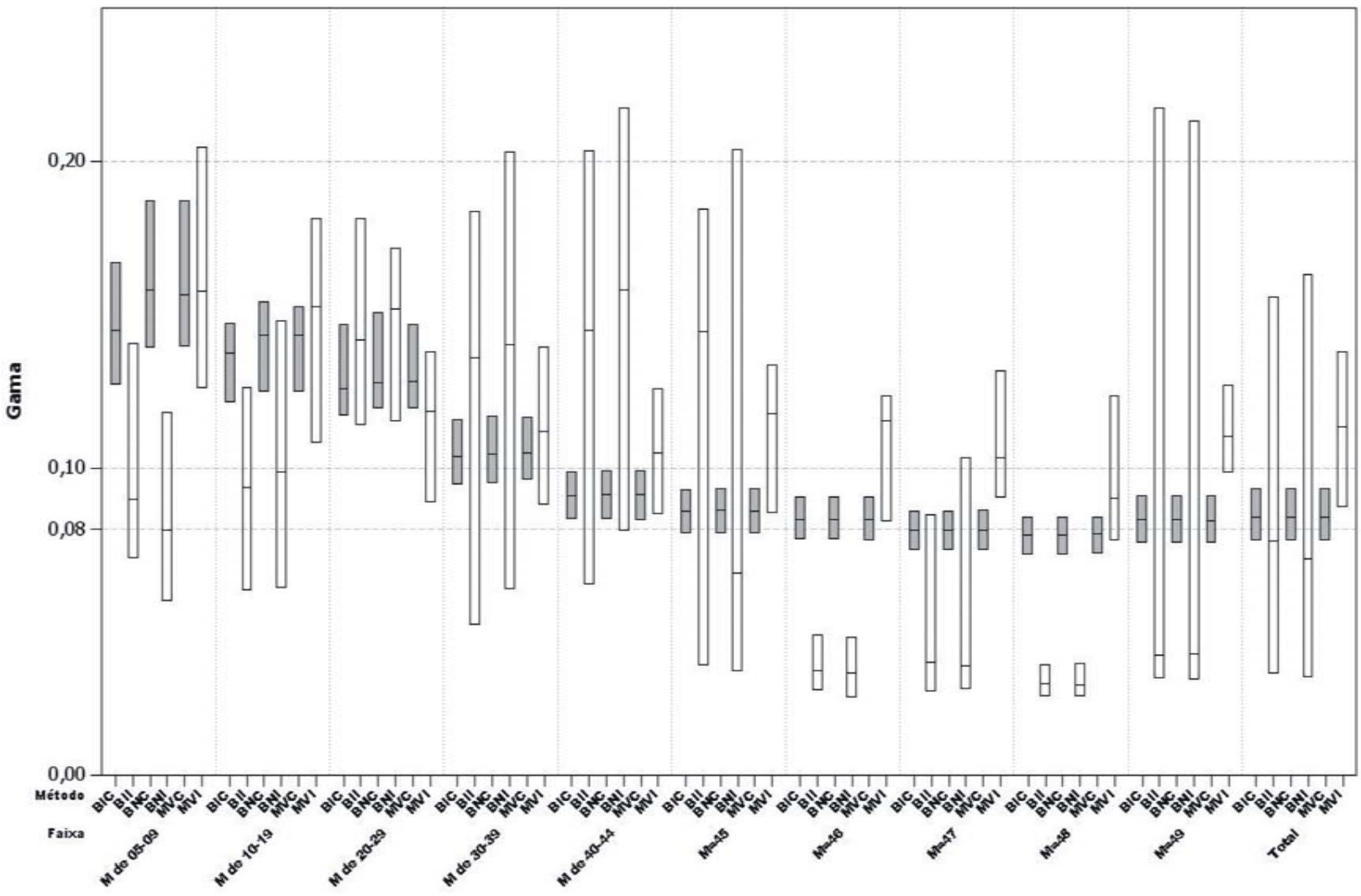



Figura 23: Boxplots com apenas os quantis, considerando todas as séries simuladas para $\gamma$. O valor verdadeiro de $\gamma$ é 0,08.

Tabela 5: Resumo das estimativas de $R_0$: Máxima Verossimilhança Completa (MVC) e Incompleta (MVI), Bayesiana Não Informativa Completa (BNC) e Incompleta (BNI) e Bayesiana Informativa Completa (BIC) e Incompleta. (BII).

| m | Método | N. Rep. | Mínimo | $1^0$ Quart | Mediana | Média | $3^0$ Quart | Máximo | D.P. | % Cobert | EQM |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Total | MVC | 3490 | 0,912 | 2,607 | 2,973 | 2,982 | 3,345 | 5,296 | 0,638 | 80 | 0,57 |
| | MVI | 300 | 0,931 | 1,506 | 2,057 | 2,304 | 2,23 | 11,406 | 0,932 | 0 | 1,61 |
| | BNC | 3490 | 0,979 | 2,66 | 3,037 | 3,048 | 3,419 | 5,252 | 0,643 | 95 | 0,81 |
| | BIC | 3490 | 1,121 | 2,668 | 3,039 | 3,051 | 3,416 | 5,216 | 0,627 | 95 | 0,79 |
| | BNI | 300 | 0,816 | 1,477 | 2,335 | 5,253 | 5,385 | 23,428 | 3,335 | 16 | 21,61 |
| | BII | 300 | 0,703 | 1,579 | 2,324 | 5,174 | 5,751 | 15,829 | 3,145 | 15 | 19,51 |
| 5-9 | MVC | 52 | 0,912 | 0,948 | 0,974 | 0,974 | 1 | 1,039 | 0,029 | 0 | 4,8 |
| | MVI | 30 | 0,931 | 0,95 | 0,973 | 0,982 | 1,014 | 1,054 | 0,038 | 0 | 4,77 |
| | BNC | 52 | 0,979 | 1,08 | 1,109 | 1,111 | 1,146 | 1,237 | 0,047 | 4 | 4,47 |
| | BIC | 52 | 1,121 | 1,199 | 1,239 | 1,249 | 1,278 | 1,488 | 0,08 | 23 | 3,97 |
| | BNI | 30 | 0,816 | 1,018 | 1,071 | 1,081 | 1,133 | 1,611 | 0,154 | 0 | 4,72 |
| | BII | 30 | 0,703 | 1,065 | 1,183 | 1,17 | 1,306 | 1,426 | 0,166 | 7 | 4,47 |
| 10-19 | MVC | 38 | 1,036 | 1,081 | 1,104 | 1,122 | 1,162 | 1,307 | 0,064 | 0 | 4,14 |
| | MVI | 30 | 1,032 | 1,069 | 1,1 | 1,127 | 1,161 | 1,424 | 0,084 | 0 | 4 |
| | BNC | 38 | 1,112 | 1,17 | 1,208 | 1,216 | 1,255 | 1,414 | 0,066 | 0 | 3,91 |
| | BIC | 38 | 1,165 | 1,249 | 1,272 | 1,288 | 1,329 | 1,479 | 0,06 | 0 | 3,65 |
| | BNI | 30 | 0,91 | 1,088 | 1,146 | 1,209 | 1,357 | 1,552 | 0,172 | 4 | 3,94 |
| | BII | 30 | 0,979 | 1,155 | 1,256 | 1,296 | 1,43 | 1,777 | 0,208 | 0 | 3,67 |
| 20-29 | MVC | 35 | 1,236 | 1,381 | 1,449 | 1,454 | 1,549 | 1,65 | 0,106 | 0 | 2,89 |
| | MVI | 30 | 1,269 | 1,36 | 1,44 | 1,431 | 1,498 | 1,562 | 0,086 | 0 | 2,88 |
| | BNC | 35 | 1,284 | 1,445 | 1,493 | 1,5 | 1,583 | 1,677 | 0,099 | 0 | 2,83 |
| | BIC | 35 | 1,307 | 1,472 | 1,524 | 1,535 | 1,611 | 1,697 | 0,095 | 0 | 2,72 |
| | BNI | 30 | 1,063 | 1,252 | 1,367 | 1,413 | 1,486 | 2,941 | 0,341 | 4 | 3,22 |
| | BII | 30 | 1,218 | 1,319 | 1,486 | 1,463 | 1,59 | 1,812 | 0,159 | 0 | 2,95 |
| 30-39 | MVC | 117 | 1,57 | 1,819 | 1,94 | 1,951 | 2,071 | 2,472 | 0,186 | 6 | 1,49 |
| | MVI | 30 | 1,518 | 1,715 | 1,825 | 1,999 | 1,909 | 4,157 | 0,583 | 0 | 1,61 |
| | BNC | 117 | 1,613 | 1,877 | 2,009 | 2,007 | 2,137 | 2,513 | 0,192 | 45 | 1,52 |
| | BIC | 117 | 1,646 | 1,893 | 2,024 | 2,025 | 2,15 | 2,52 | 0,189 | 46 | 1,47 |
| | BNI | 30 | 1,406 | 1,706 | 1,786 | 1,916 | 1,95 | 4,534 | 0,557 | 3 | 1,97 |
| | BII | 30 | 1,416 | 1,653 | 1,864 | 1,98 | 2,076 | 4,327 | 0,545 | 0 | 1,82 |
| 40-44 | MVC | 638 | 1,836 | 2,288 | 2,47 | 2,492 | 2,685 | 3,629 | 0,275 | 43 | 0,57 |
| | MVI | 30 | 1,79 | 1,895 | 2,063 | 2,251 | 2,163 | 7,126 | 0,934 | 0 | 1,63 |
| | BNC | 638 | 1,851 | 2,343 | 2,525 | 2,55 | 2,743 | 3,648 | 0,284 | 96 | 0,73 |
| | BIC | 638 | 1,873 | 2,355 | 2,533 | 2,558 | 2,753 | 3,638 | 0,28 | 96 | 0,71 |
| | BNI | 30 | 1,475 | 1,915 | 2,164 | 2,441 | 2,411 | 5,728 | 0,946 | 23 | 1,77 |
| | BII | 30 | 1,79 | 1,951 | 2,131 | 2,359 | 2,352 | 4,478 | 0,703 | 20 | 1,39 |

Tabela 6: Resumo das estimativas de $R_0$: Máxima Verossimilhança Completa (MVC) e Incompleta (MVI), Bayesiana Não Informativa Completa (BNC) e Incompleta (BNI) e Bayesiana Informativa Completa (BIC) e Incompleta. (BII).

| m | Método | N. Rep. | Mínimo | $1^0$ Quart | Mediana | Média | $3^0$ Quart | Máximo | D.P. | % Cobert | EQM |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 45 | MVC | 394 | 2,005 | 2,576 | 2,755 | 2,781 | 2,956 | 3,835 | 0,285 | 78 | 0,33 |
| | MVI | 30 | 1,926 | 2,009 | 2,063 | 2,117 | 2,204 | 2,602 | 0,152 | 0 | 1,07 |
| | BNC | 394 | 2,057 | 2,64 | 2,804 | 2,841 | 3,018 | 3,972 | 0,294 | 99 | 0,53 |
| | BIC | 394 | 2,074 | 2,647 | 2,81 | 2,845 | 3,018 | 3,958 | 0,289 | 99 | 0,52 |
| | BNI | 30 | 0,951 | 2,31 | 2,577 | 3,11 | 3,815 | 6,562 | 1,273 | 38 | 2,57 |
| | BII | 30 | 1,926 | 2,304 | 2,614 | 3,453 | 4,429 | 7,454 | 1,523 | 31 | 3,4 |
| 46 | MVC | 499 | 2,224 | 2,767 | 2,947 | 2,987 | 3,193 | 3,979 | 0,322 | 94 | 0,29 |
| | MVI | 30 | 1,909 | 2,083 | 2,167 | 2,52 | 2,294 | 11,406 | 1,624 | 3 | 2,96 |
| | BNC | 499 | 2,257 | 2,822 | 3,023 | 3,054 | 3,274 | 4,096 | 0,332 | 100 | 0,54 |
| | BIC | 499 | 2,267 | 2,825 | 3,023 | 3,055 | 3,274 | 4,03 | 0,327 | 100 | 0,53 |
| | BNI | 30 | 2,126 | 3,05 | 4,509 | 4,581 | 5,59 | 8,792 | 1,807 | 25 | 7,24 |
| | BII | 30 | 2,082 | 2,914 | 4,504 | 5,255 | 6,92 | 11,528 | 2,642 | 28 | 13,76 |
| 47 | MVC | 643 | 2,306 | 2,927 | 3,185 | 3,204 | 3,431 | 4,677 | 0,384 | 97 | 0,36 |
| | MVI | 30 | 1,98 | 2,12 | 2,198 | 2,236 | 2,366 | 2,515 | 0,15 | 0 | 0,85 |
| | BNC | 643 | 2,363 | 2,994 | 3,241 | 3,269 | 3,5 | 4,844 | 0,396 | 100 | 0,65 |
| | BIC | 643 | 2,372 | 2,993 | 3,239 | 3,267 | 3,495 | 4,816 | 0,39 | 100 | 0,63 |
| | BNI | 30 | 2,062 | 3,003 | 6,085 | 5,846 | 8,237 | 10,459 | 2,726 | 15 | 18,79 |
| | BII | 30 | 1,985 | 2,546 | 5,84 | 5,605 | 7,899 | 11,135 | 2,674 | 6 | 17,26 |
| 48 | MVC | 642 | 2,259 | 3,137 | 3,378 | 3,44 | 3,707 | 5,059 | 0,436 | 99 | 0,56 |
| | MVI | 30 | 1,948 | 2,137 | 2,224 | 2,557 | 2,41 | 7,536 | 1,237 | 0 | 1,83 |
| | BNC | 642 | 2,298 | 3,207 | 3,449 | 3,509 | 3,76 | 5,202 | 0,447 | 99 | 0,88 |
| | BIC | 642 | 2,307 | 3,207 | 3,448 | 3,504 | 3,762 | 5,168 | 0,44 | 99 | 0,86 |
| | BNI | 30 | 2,358 | 7,258 | 8,804 | 8,961 | 10,414 | 16,124 | 2,961 | 3 | 41,01 |
| | BII | 30 | 2,172 | 7,104 | 8,692 | 8,531 | 10,276 | 15,829 | 3,336 | 13 | 50,69 |
| 49 | MVC | 432 | 2,353 | 3,08 | 3,391 | 3,429 | 3,706 | 5,296 | 0,486 | 98 | 0,59 |
| | MVI | 30 | 2,021 | 2,136 | 2,288 | 2,28 | 2,389 | 2,731 | 0,171 | 0 | 0,78 |
| | BNC | 432 | 2,435 | 3,152 | 3,462 | 3,502 | 3,802 | 5,252 | 0,498 | 98 | 0,91 |
| | BIC | 432 | 2,444 | 3,148 | 3,457 | 3,497 | 3,796 | 5,216 | 0,491 | 98 | 0,891 |
| | BNI | 30 | 1,977 | 2,609 | 7,199 | 7,031 | 9,927 | 23,428 | 4,921 | 0 | 46,7 |
| | BII | 30 | 2,139 | 2,615 | 7,153 | 6,431 | 8,964 | 13,201 | 3,593 | 0 | 29,66 |

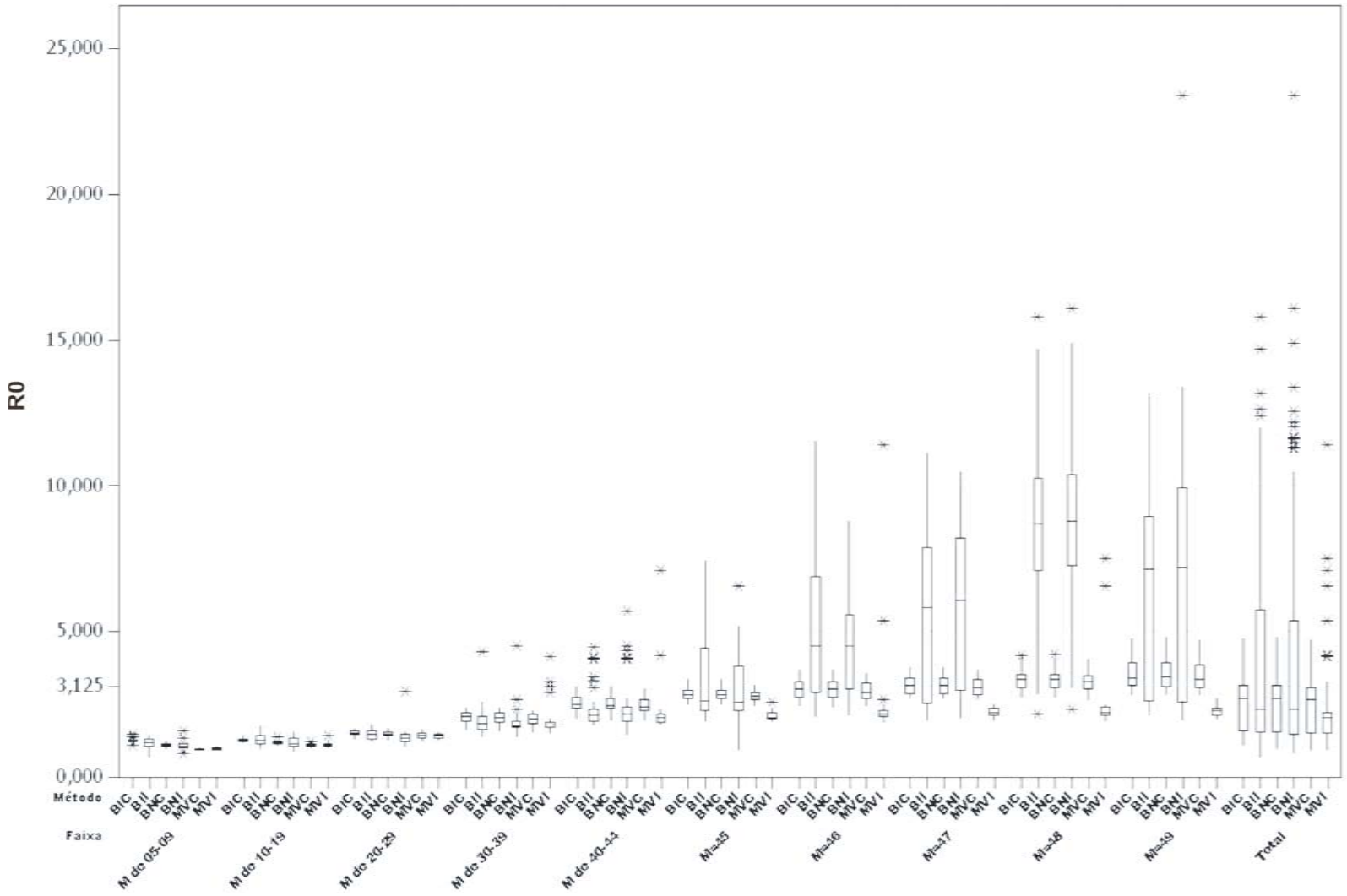



Figura 24: Boxplot para as estimativas de $R_0$ por faixa de tamanho final(30 replicações para cada faixa). O valor verdadeiro de $R_0$ é 3,125.

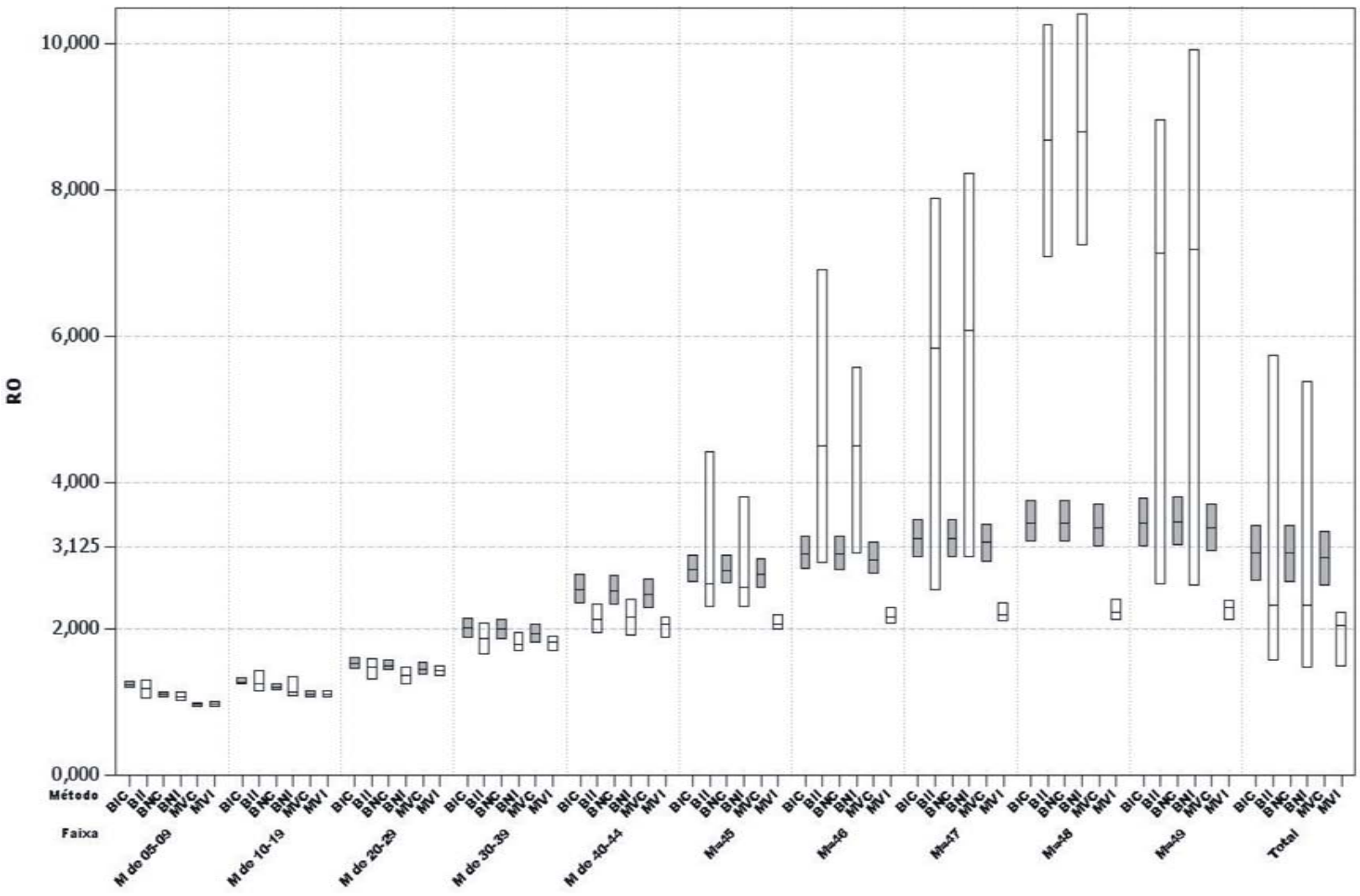



Figura 25: Boxplots com apenas os quantis, considerando todas as séries simuladas para $R_0$. O valor verdadeiro de $R_0$ é 0,08.

Através dos boxplots (figuras 20 a 25) e tabelas (tabelas 1 a 6) , podemos concluir que temos uma maior variabilidade nas estimativas quando temos dados incompletos ao compararmos com dados completos. Também pode-se notar que as coberturas dos intervalos de confiança/credibilidade 95 % foram baixas para algumas faixas de tamanhos finais, mesmo quando se tem dados completos.

Podemos notar também que, quando aumentamos o tamanho final da epidemia(m), as estimativas para $\beta$ e $R_0$ aumentam, sobretudo para dados incompletos. Para $\gamma$, quando temos aumento no tamanho final, temos estimativas menores, especialmente quando temos dados incompletos.

Observando-se as médias dos erros quadráticos médios nas faixas de tamanho final, verifica-se que o EQM das três metodologias com dados completos é muito semelhante. Nota-se também que, para dados incompletos, temos EQM menor para a estimativa por máxima verossimilhança para quase todas as faixas de tamanho final, sobretudo nas faixas de tamanho final grandes para $\beta$ e $R_0$, em que estes parâmetros foram superestimados pelas estimativas bayesianas.

Observa-se também que as estimações utilizando prioris informativas foram, em geral, mais próximas do valor real dos parâmetros do que as estimações com prioris não informativas, especialmente quando temos informação incompleta. Contudo, as prioris informativas não foram necessárias para corrigir o vício da estimação devido ao tamanho final.

Logo, podemos concluir que o tamanho final da epidemia afeta bastante na qualidade das estimativas, principalmente quando se trata de dados incompletos.

Para saber em quais tamanhos finais temos as maiores diferenças entre as estimativas, fez-se gráficos do desvio vs. o tamanho final da epidemia (m). Denomina-se desvio a diferença entre a estimativa com dados completos e a estimativa com dados incompletos.

Através da figura 26, podemos concluir que os maiores desvios estão nos maiores tamanhos finais da epidemia, ou seja, para valores grandes de tamanhos finais da epidemia que temos as maiores diferenças entre as estimativas com dados completos e incompletos.

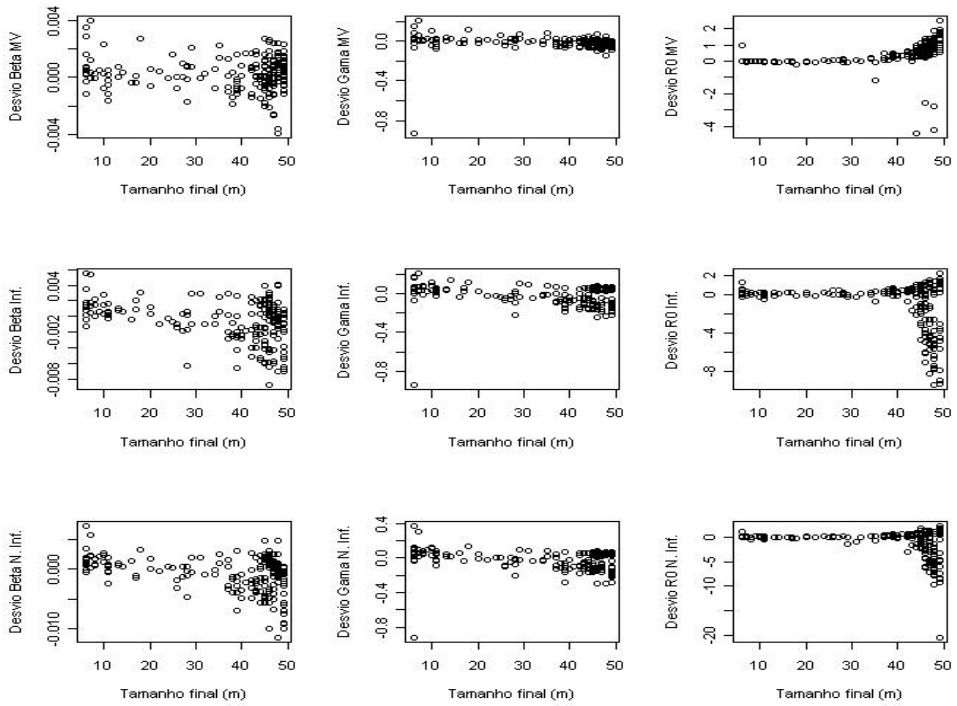



Figura 26: Desvio vs. Estimativas.

## 3.2 Exemplo de Aplicação com Dados Reais - Dados de Varíola

Nesta seção será feita uma aplicação com dados reais. O principal objetivo desta aplicação é verificar se o algoritmo para estimação com dados incompletos utilizado na seção anterior produz resultados semelhantes aos obtidos nas literaturas. Outro objetivo é mostrar como foi detectada a convergência das estimativas da seção anterior (figura 27) e como foi tomada a amostra final com pouca dependência (figura 27) na seção anterior.

Os dados foram obtidos de uma epidemia de varíola em uma comunidade de 120 indivíduos em Abakaliki, Nigéria (ver Bailey, 1975). Os dados consistem em 29 tempos de remoção. Não temos os tempos de infecção. Se o dia de início da epidemia for o dia zero, os 29 dias em que as remoções foram observadas são: 0, 13, 20, 22, 25, 25, 25, 26, 30, 35, 38, 40, 40, 42, 42, 47, 50, 51, 55, 55, 56, 57, 58, 60, 60, 61, 66, 66, 71, 76.

Para todas as estimativas foram feitos gráficos de séries temporais com a finalidade de verificar a convergência (figura 27). Um burn-in de 5000 foi utilizado e foram tomadas amostras com intervalos de 100 cada uma para tomarmos amostras independentes. No final, obteve-se uma amostra de tamanho 250.

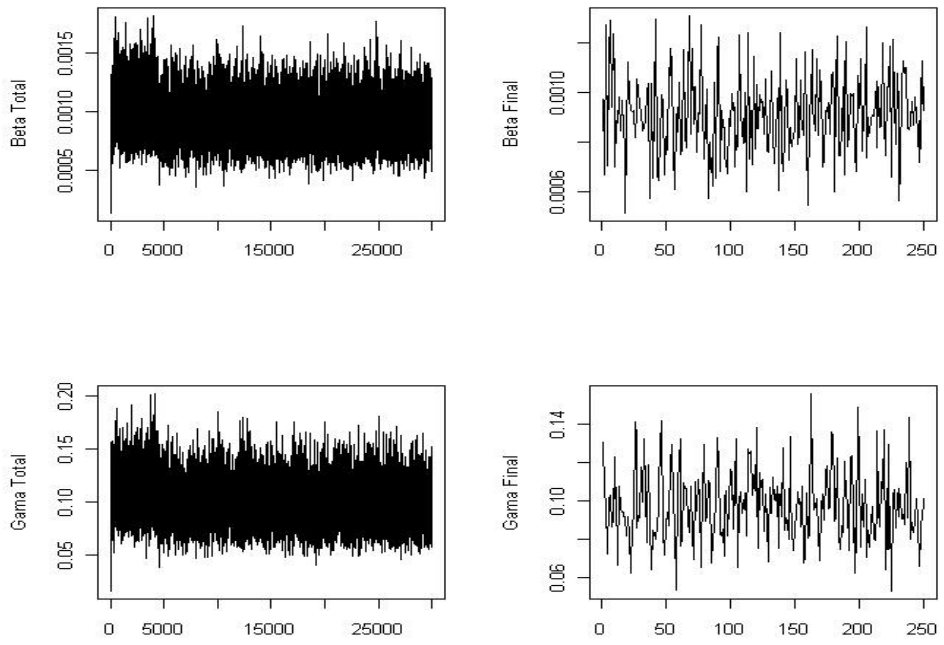



Figura 27: Convergência das estimativas (estimativa bayesiana utilizando prioris informativas)

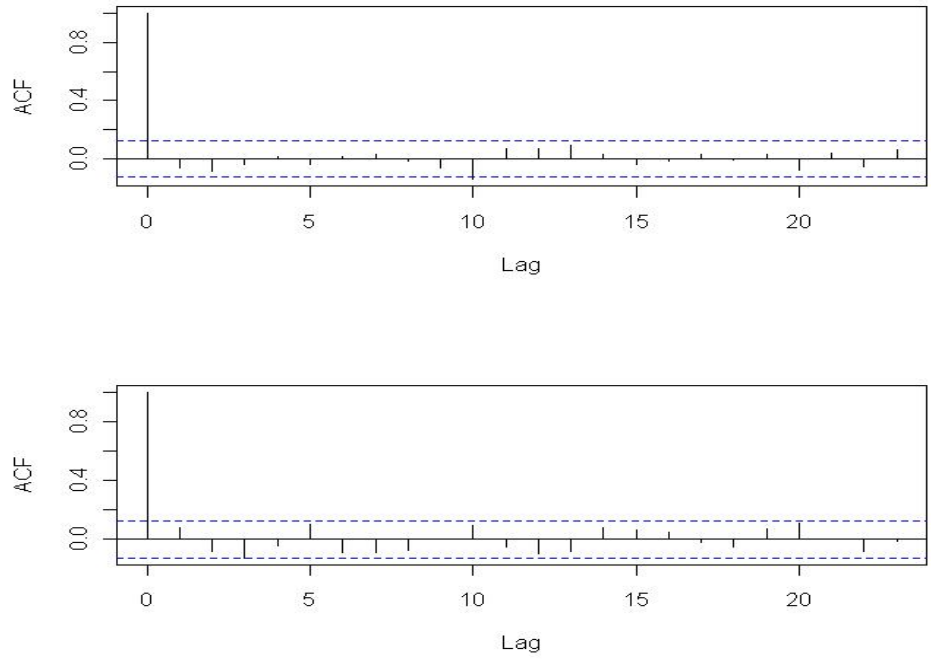



Figura 28: Gráfico das autocorrelações. Em a, temos as autocorrelações para $\beta$ e em b temos a autocorrelações para $\gamma$. Esta foi a estimativa bayesiana utilizando prioris informativas.

Tabela 7: Estimativas para os dados de varíola e comparação com as estimativas obtidas na literatura. MV= Máxima Verossimilhança, N. Inf. = Bayesiana com prioris não informativas, Inf. = Bayesiana com prioris informativas.

| Parâmetro | Método | **Valor** | **Desvio-padrão** | **Int. cred./conf. 0,95** |
|---|---|---|---|---|
| $\beta$ | MV | 0,00087 | 0,00018 | (0,00058;0,00127) |
| | MV-Bailey(1975) | 0,00088 | 0,00025 | - |
| | N. Inf. | 0,00088 | 0,00021 | (0,00049;0,00131) |
| | N. Inf.-O'Neill e Roberts(1999) | 0,00090 | 0,00019 | - |
| | Inf. | 0,00090 | 0,00016 | (0,00060;0,00124) |
| | Inf.-O'Neill e Roberts(1999) | 0,0011 | 0,0001 | - |
| $\gamma$ | MV | 0,0916 | 0,0188 | (0,0606;0,1306) |
| | MV-Bailey(1975) | 0,091 | 0,031 | - |
| | N. Inf. | 0,0950 | 0,0223 | (0,0572;0,1452) |
| | N. Inf.-O'Neill e Roberts(1999) | 0,098 | 0,0207 | - |
| | Inf. | 0,0974 | 0,0187 | (0,0649;0,1366) |
| | Inf.-O'Neill e Roberts(1999) | 0,107 | 0,009 | - |

As prioris não informativas são dadas por: $\beta \sim \Gamma(0;0)$ e $\gamma \sim \Gamma(0;0)$ e $-I_0 \sim \exp(0,1)$. As prioris informativas são as mesmas utilizadas por O'Neill e Roberts(1999), dadas por: $\beta \sim \Gamma(10;10000)$ e $\gamma \sim \Gamma(10;100)$ e $-I_0 \sim \exp(0,1)$. Então, as médias a priori para $\beta$, $\gamma$ e $-I_0$ são 0,0001, 0,1 e 10, respectivamente.

Através da tabela 7, vemos que as estimativas estão bastante próximas comparadas à literatura. A estimativa de máxima verossimilhança está bem próxima com a estimativa de máxima verossimilhança encontrada em Bailey(1975) e as estimativas bayesianas estão muito próximas das estimativas bayesianas encontradas em O'Neill e Roberts(1999).

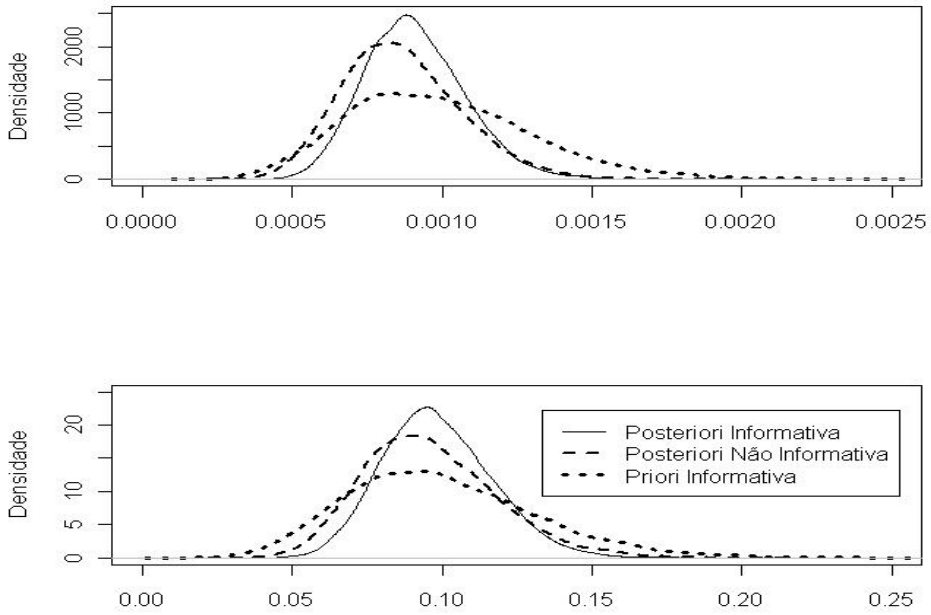



Figura 29: Distribuição a posteriori para os parâmetros $\beta$ (painel superior) e $\gamma$ (painel inferior). As legendas para os dois gráficos são iguais.

# 4 Conclusões e Trabalhos Futuros

Neste trabalho, estivemos interessados em estudar a estimação de parâ-metros em modelos SIR estocásticos. Nas simulações, verificamos que as estimativas são bastante dependentes do tamanho final da epidemia. Tal vício pode ser notado através do fato de termos cobertura dos intervalos de confiança/credibilidade 95 % diferentes para cada faixa de tamanho final. Também podemos ver este viés através das médias diferentes de erros quadráticos médios nas diferentes faixas de tamanho final. O problema da baixa cobertura dos intervalos de confiança/credibilidade e diferença nas médias de erros quadráticos médios pôde ser notado tanto para dados completos quanto para dados incompletos.

Através das simulações, pôde-se concluir também que, na abordagem bayesiana, a utilização de prioris mais informativas em geral traz estimativas mais próximas ao valor real do parâmetro, contudo o vício na estimativa devido ao tamanho final da epidemia não é corrigido através do uso destas prioris.

Portanto, como trabalhos futuros, tentar-se-á propor estimadores menos viciados para os parâmetros do modelo SIR. Por exemplo, pode-se tomar outras distribuições de probabilidade para os tempos de transição do mo-delo, como já foi proposto por Streftaris e Gibson (2004) e verificar o vício do estimador com relação ao tamanho final. Uma outra melhoria que pode ser realizada nas estimações é não utilizar independência a priori para os parametros $\beta$ e $\gamma$ do modelo. No modelo bayesiano estudado, supomos independência a priori dos parâmetros, suposição que pode ter feito com que a qualidade das estimativas diminuísse. Outra modificação no modelo que poderá ser estudada é utilizar o modelo SEIR ao invés do modelo SIR. É muito importante que se considere o compartimento dos expostos, pois há a necessidade de se considerar o período de latência na modelagem

de epidemias. No futuro, também serão estudados outros métodos de estimação, como baseadas em martingais e baseadas no algoritmo EM.

# *Referências*

ADES, A.E.; CLIFFE, S. Markov chain Monte Carlo estimation of a multiparameter decision model: consistency of evidence and the accurate assessment of uncertainty. *Medical Decision Making*, Vol. 22, p. 359-371, 2002.

ANDERSON, R.; MAY, R. *Infectious Diseases of Humans; Dynamics and Control*. Oxford University Press, Oxford, 1991.

ANDERSSON, H.; BRITTON, T. Stochastic Epidemic Models and Their Statistical Analysis. *Lecture Notes in Statistics*, Vol. 151, 2000.

BAILEY, N. *The Mathematical Theory of Infectious Diseases and Its Applications*. Griffin, Londres, 1975.

BARTLETT, M. S. *Stochastic Processes*. Cambridge University Press, 1955.

BARTLETT, M. S. *Stochastic Population Models in Ecology and Epidemiology*. Methuen, Londres, 1960.

BECKER, N. G. The use of epidemic models. *Biometrics*, Vol. 35, p. 295-305, 1978.

BECKER, N. G. *Analysis of Infectious Disease Data*. Editora Chapman e Hall, Londres, 1989.

BECKER, N. G. Martingale methods for the analysis of epidemic data. *Statistical Methods in Medical Research*, Vol. 2, N. 1,p. 93-112, 1993.

BECKER, N. G. Parametric inference for epidemic models.*Mathematical Biosciences*, Vol. 117, p. 239-251, 1993.

BECKER, N. G. Uses of the EM algorithm in the analysis of data on HIV/AIDS and other infectious diseases.*Statistical Methods in Medical Research*, Vol. 6, N. 1, p. 24-37, 1997.

BERNOULLI, D. Essai d'une nouvelle analyse de la mortalité causée par la petite vérole et des advantages de l'inoculation pour la prévenir. *Mémoires de Mathématiques et de Physique, Académie Royale des Sciences*. Paris, p. 1-45, 1760.

BICKEL, P.,J.; DOKSUM, K., A. *Mathematical Statistics*. Prentice-Hall, New Jersey. Segunda edição, 2001.

BOYCE, W.; DIPRIMA, R. *Equações Diferenciais Elementares e Problemas de Valores de Contorno*. Oitava edição, Editora LTC, Rio de Janeiro, 2006.

BROOKHART, M.A.; HUBBARD, A.E.; VAN DER LAAN, M.J.; COLFORD, J.M.; EISENBERG, N.S. Statistical estimation of parameters in a disease transmission model: analysis of a *Cryptosporidium* outbreak. *Statistics in Medicine*, Vol. 21, p. 3627-3638, 2002.

CANCRÉ, N.; TALL, A.; ROGIER, C.; FAYE, J.; SARR, O.; TRAPE, J-F.; SPIEGEL, A. ; BOIS, F. Bayesian analysis of a epidemiological model of *Plamodium falciparum* malaria infect in Ndiop, Senegal. *American Journal of Epidemiology*, Vol. 152 (8), p. 760-770, 2000.

CHOWELL, G.; HENGARTNER, N.W.; CASTILLO-CHAVEZ, C.; FENIMORE, P.W.; HYMAN, J.M. The basic reproductive number of ebola and the effects of public health measures: The Cases of Congo and Uganda. *Journal of Theoretical Biology*, Vol. 229, p. 119-126, 2004.

CLANCY, D.; O'NEILL, P.D. Bayesian estimation of the basic reproduction number in stochastic epidemic models. *Bayesian Analysis*, Vol. 3, p. 737-758, 2008.

CLIFF, A.D.; HAGGETT, P. Statistical modelling of measles and influenza outbreaks. *Statistical Methods in Medical Research*, Vol. 2, p. 42-73, 1993.

COX, D. R. Regression models and life tables(with discussion). *Journal of the Royal Statistical Society B*, Vol. 34, p. 187-220, 1992.

DIETZ, K. Density-dependence in parasite transmission dynamics. *Parasitology Today*, Vol. 4, 91-7, 1988.

DALEY, D. J; KENDALL, D. G. *Journal of the Institute of Mathematics and Its Applications*, Vol.1, p. 42-55, 1965.

DIEKMANN, O.; HEESTERBEEK, J. A. P. *Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis and Interpretation*. Wiley Series in Mathematical and Computational Biology, 2000.

FAHRMEIR, L. ; TUTZ, G. *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer-Verlag, 1994.

GELFAND, A.E; SMITH, A.F.M. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, Vol. 85, p. 398-409, 1990.

HAMER, W. H. The Milroy lectures on epidemic disease in England - The evidence of variability and persistence of type. *The Lancet*, Vol. 1, p. 733-739, 1906.

HÖHLE, M. R0 estimation by the martingale method. *Biometry Research Unit Internal Report*, 2003.

HÖHLE, M.; JORGENSEN, E. Estimating parameters for stochastic epidemics. *Dina Research Report*, Dinamarca, n.102, 2002.

HOTTA, L. K. Bayesian melding estimation of the stochastic SEIR model. A aparecer em *Mathematical Population Studies*, 2009.

JAMES, B. *Probabilidade: Um Curso Em Nível Intermediário*. Projeto Euclides. Terceira edição, 2004.

KERMACK, W.O; McKENDRICK, A.G. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London A*, Vol. 115, p. 700-721, 1927.

LEKONE, P. E.; FINKENSTADT, B. F. ; Statistical inference in a stochastic epidemic SEIR model with control intervention: ebola as a case study. *Biometrics*, Vol. 62, p. 1170-1177, 2006.

MAKI, D. P.; THOMPSON, M. *Mathematical Models and Applications*. Prentice-Hall, Englewood Cliffs, NJ, 1973.

McKENDRICK, A.G. Applications of mathematics to medical problems. *Proceedings of the Edinburgh Mathematical Society*, Vol. 14, p. 98-130, 1926.

MEESTER, R.; DE KONING, J.; DE JONG, M.; DIEKMANN, O. Modeling and real-time prediction of classical swine fever epidemics. *Biometrics*, Vol. 58, N. 1, p 178-184, 2002.

O'NEILL, P. D.; BECKER, N.G. Inference for an epidemic when susceptibility varies. *Biostatistics*, Vol. 2, p. 99-108, 2001.

O'NEILL, P. D.; ROBERTS, G. O. Bayesian inference for partially observed stochastic epidemics. *Journal of The Royal Statistical Society A*, Vol. 162, p. 121-129, 1999.

PICARD, P. Applications of martingale theory to some epidemic models. *Journal of Applied Probability*, Vol. 17, N. 3, p. 583-599, 1980.

POOLE, D.; RAFTERY, A.E. Inference for deterministic simulation models: the Bayesian melding approach. *Journal of the American Statistical Association*, Vol. 95 (452), p. 1244-1255, 2000.

RAFTERY, A.E.; GIVENS, G.H.; ZEH, J.E. Inference from a deterministic population dynamics model for bowhead whales. *Journal of the American Statistical Association*, Vol. 90 (430), p. 402-416, 1995.

ROSS, R. *Report on the Prevention of Malaria in Mauritius*. Londres, 1908.

SPEAR. R.C.; HUBBARD, A.; LIANG, S.; SETO, E. Disease transmission models for public health decision making: toward an approach for designing intervention strategies for *Schistomiasis Japonica*. *Environmental Health Perspectives*, 110 (9), p. 907-915, 2002.

STREFTARIS, G.; GIBSON, G. Bayesian inference for stochastics epidemics in closed populations. *Statistical Modelling*, Vol. 4, p. 63-75, 2004.

YANG, H. M. *Epidemiologia Matemática: Estudo dos Efeitos da Vacinação em Doenças de Transmissão Direta*. Editora Unicamp, Campinas-SP, 2001.

# *Apêndice*

**Propriedade 1 - Falta de memória:**

$\forall s, t \geq 0,\ P(T > t + s/T > t) = P(t > s)$

A distribuição exponencial de taxa de falha $\lambda$ possui esta propriedade, pois:

$P(T > t + s/T > t) = \dfrac{P(T > t + s, T > t)}{P(T > t)} =^{s,t>0} \dfrac{P(T > t + s)}{P(T > t)} = \dfrac{e^{-\lambda(t+s)}}{e^{-\lambda t}} = e^{-\lambda s} = P(T > s)$.

**Propriedade 2 - Mínimo entre duas exponenciais:**

O mínimo entre 2 exponenciais de taxa de falha $\lambda_1$ e $\lambda_2$ possui distribuição exponencial de taxa de falha $\lambda_1 + \lambda_2$.

*Demonstração:*

Seja $Z = \text{mínimo}(X_1, X_2)$ com $X_1$ *e* $X_2$ independentes de distribuição exponencial com taxas de falha $\lambda_1$ e $\lambda_2$ respectivamente.

$P(Z > z) = P(X_1 > z, X_2 > z) = P(X_1 > z)P(X_2 > z) = e^{-\lambda_1 z}e^{-\lambda_2 z} = e^{-(\lambda_1+\lambda_2)z}$.

**Definição da matriz de informação de Fisher:** Seja $\boldsymbol{\theta} = (\theta_1, .., \theta_d) \in \Theta$ aberto em $\Re^d$. Se $\{p(x/\theta), \theta \in \Theta\}$ é a família de probabilidades onde as condições de regularidade a seguir valem para cada coordenada $\theta_j$, j=1,...,d:

**Condição 1:**

i) $A = \{x : p(x/\theta) > 0\}$ não depende de $\theta$

ii) $\forall x \in A,\ \forall \theta \in \Theta,\ \dfrac{\partial \ln p(x/\theta)}{\partial \theta}$ existe e é finita.

**Condição 2:** Se T é qualquer estatística tal que $E_\theta(|T|) < \infty,\ \forall \theta \in \Theta$, então $\dfrac{\partial}{\partial \theta}\left\{\int T(X)\, p(x/\theta)\, dx\right\} = \int T(X)\, \dfrac{\partial}{\partial \theta} p(x/\theta)\, dx$

Então, a matriz de informação de Fisher é dada por:

$$I_{pxp} = (I_{jk}(\theta)), \text{ onde } \ (I_{jk}(\theta)) = E\left(\frac{\partial}{\partial \theta_j} \ln p(x/\theta)\, \frac{\partial}{\partial \theta_k} \ln p(x/\theta)\right)$$

**Teorema 1:** Se $X_1, ..., X_n$ é i.i.d. $P$ onde $P \in Q$, um modelo contendo $P \equiv \left\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\right\}$ tal que:

(i) $\Theta$ aberto $\subset \Re^p$.

(ii) Densidades de $P_\theta$ são $p(., \boldsymbol{\theta}), \theta \in \Theta$.

Se as condições a seguir valem para $\rho \equiv logp(x, \boldsymbol{\theta})$ :

**A0.** $\boldsymbol{\Psi} \equiv (\psi_1, ..., \psi_p)^T$, onde $\psi_j = \frac{\partial \rho}{\partial \theta_j}$ é bem definida e

$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\Psi}(X_i, \overline{\boldsymbol{\theta}}_{\boldsymbol{n}}) = \mathbf{0}$, onde $\overline{\boldsymbol{\theta}}_{\boldsymbol{n}}$ é a estimativa de mínimo contraste.

**A1.** O parâmetro $\boldsymbol{\theta}(P)$ dado pela solução de

$$\int \boldsymbol{\Psi}(x, \boldsymbol{\theta}) dP(x) = \mathbf{0} \tag{4.1}$$

é bem definida em Q tal que $\boldsymbol{\theta}(P)$ é a única solução da equação 4.1.

**A2.** $E_p \left|\boldsymbol{\Psi}(X_1, \boldsymbol{\theta}(P))\right|^2 < \infty$, onde $|\cdot|$ é a norma Euclidiana.

**A3.** $\psi_i(\cdot, \boldsymbol{\theta})$, $1 \leq i \leq p$, tem derivadas parciais de primeira ordem com respeito a todas as coordenadas e

$$E_P \left|D\boldsymbol{\Psi}(X_1, \boldsymbol{\theta})\right| < \infty$$

onde

$$E_P D\boldsymbol{\Psi}(X_1, \boldsymbol{\theta}) = \left\| E_P \frac{\partial \psi_i}{\partial \theta_j}(X_1, \boldsymbol{\theta}) \right\|_{pxp}$$

é não singular.

**A4.** $sup\left\{\left|\frac{1}{n}\sum_{i=1}^{n}(D\boldsymbol{\Psi}(X_i,\boldsymbol{t})-\boldsymbol{\Psi}(X_i,\boldsymbol{\theta}(P)))\right| : |\boldsymbol{t}-\boldsymbol{\theta}(P)|\leq\epsilon_n\right\}\xrightarrow{P}0$ se $\epsilon_n\to 0$.

**A5.** $\overline{\boldsymbol{\theta}}_{\boldsymbol{n}}\xrightarrow{P}\boldsymbol{\theta}(P)$ para todo $P\in Q$.

**A6.** Se $l(\cdot,\boldsymbol{\theta})$ é diferenciável,

$E_{\boldsymbol{\theta}}D\boldsymbol{\Psi}(X_1,\boldsymbol{\theta})=-E_{\boldsymbol{\theta}}\boldsymbol{\Psi}(X_1,\boldsymbol{\theta})Dl(X_1,\boldsymbol{\theta})=-Cov_{\boldsymbol{\theta}}(\boldsymbol{\Psi}(X_1,\boldsymbol{\theta}),Dl(X_1,\boldsymbol{\theta}))$.

Então, o estimador de máxima verossimilhança $\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ satisfaz

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}=\boldsymbol{\theta}+\frac{1}{n}\sum_{i=1}^{n}I^{-1}(\boldsymbol{\theta})Dl(X_1,\boldsymbol{\theta})+o_p(n^{-1/2}).$$

Logo

$\sqrt{n}(\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}-\boldsymbol{\theta})\xrightarrow{d}N(\mathbf{0},I^{-1}(\boldsymbol{\theta}))$.

**Algumas Distribuições Importantes na Análise de Sobrevivência**

**Função densidade de probabilidade:** $f(t) = lim_{\Delta t \to 0} \frac{1}{\Delta t} P[t \leq T \leq t + \Delta t]$

**Função de distribuição acumulada:** $F(t) = P(T \leq t) = \int_0^t f(s)\, ds$

**Função de sobrevivência:** $S(t) = P(T \geq t) = 1 - F(t) = \int_t^\infty f(s)\, ds$

**Função risco:** $\lambda(t) = lim_{\Delta t \to 0} \frac{1}{\Delta t} P[t \leq T \leq t + \Delta t / T \geq t] = \frac{f(t)}{S(t)}$

**Função risco acumulada:** $\Lambda(t) = \int_0^t \lambda(s)\, ds = \int_0^t \frac{f(s)}{1 - F(s)} ds$. Fazendo a substituição simples $u = 1 - F(s)$ temos que: $\Lambda(t) = -\ln(S(t))$