Yasmin Ferreira Cavaliere

Estimando o tamanho de populações de difícil acesso usando o método *Network Scale-up*

Niterói - RJ, Brasil

22 de dezembro de 2016

Universidade Federal Fluminense

Yasmin Ferreira Cavaliere

Estimando o tamanho de populações de difícil acesso usando o método *Network Scale-up*

Trabalho de Conclusão de Curso

Monografia apresentada para obtenção do grau de Bacharel em Estatística pela Universidade Federal Fluminense.

Orientador: Prof.ª Mariana Albi de Oliveira Souza

Coorientador: Prof. Leonardo Soares Bastos

Niterói - RJ, Brasil

22 de dezembro de 2016

Universidade Federal Fluminense

Yasmin Ferreira Cavaliere

Estimando o tamanho de populações de difícil acesso usando o método *Network Scale-up*

Monografia de Projeto Final de Graduação sob o título *"Estimando o tamanho de populações de difícil acesso usando o método Network Scale-up"*, defendida por Yasmin Ferreira Cavaliere e aprovada em 22 de Dezembro de 2016 22 de dezembro de 2016, na cidade de Niterói, no Estado do Rio de Janeiro, pela banca examinadora constituída pelos professores:

Prof.ª Dra. Mariana Albi de Oliveira Souza
Departamento de Estatística – UFF

Prof. Dr. Wilson Calmon Almeida dos Santos
Departamento de Estatística – UFF

Prof. Dr. Jony Arrais Pinto Junior
Departamento de Estatística – UFF

Niterói, 22 de dezembro de 2016

C376 Cavaliere, Yasmin Ferreira

Estimando o tamanho de populações de difícil acesso usando o método network scale-up / Yasmin Ferreira Cavaliere. – Niterói, RJ: [s.n.], 2016

51f.

Orientador: Profª.Drª. Mariana Albi de Oliveira Souza

Coorientador: Prof. Dr. Leonardo Soares Bastos

TCC (Graduação de Bacharelado em Estatística) – Universidade Federal Fluminense, 2016.

1.Amostragem (Estatística). 2.Saúde Publica. 3.Cognição. I.Título.

CDD 519.54

Resumo

Para promover políticas de saúde pública, há uma grande necessidade em conhecer o tamanho de populações de difícil acesso, tais como profissionais do sexo, usuários de drogas, homens que fazem sexo com outros homens, entre outras. Populações essas, muitas vezes, estigmatizadas e até mesmo criminalizadas, mas que tem um papel muito grande na dinâmica de várias doenças transmissíveis. O método *Network Scale-up* (NSUM) usa informações indiretas obtidas da população geral para estimar o tamanho de tais populações. Para ilustrar este método, pode-se pensar no seguinte exemplo: pergunta-se a uma pessoa aleatoriamente selecionada da população geral "quantos dos seus amigos pertencem a população X?". Além da população de difícil acesso de interesse, pergunta-se também sobre outras populações conhecidas. Essas perguntas são usadas, em um primeiro passo, para estimar o total de "amigos"de cada indivíduo, também chamado de grau do indivíduo, e, em um segundo passo, estima-se o tamanho da população de difícil acesso. Neste trabalho, foram simuladas populações conectadas em rede, usando um modelo de rede aleatória, e, para cada nó gerado, características foram aleatoriamente alocadas, conhecendo-se assim tanto as características dos indivíduos quanto de seus contatos. Amostras foram selecionadas dessas populações e estimativas para o tamanho dessas populações foram calculadas usando as abordagens frequentista e bayesiana dando indícios de que o método *Network Scale-up* têm se mostrado eficiente. Além disso, foram avaliados o tamanho da amostra, a prevalência de uma dada subpopulação, e outros parâmetros associados à modelagem. Nesse trabalho, foi possível validar o método NSUM em condições totalmente controladas e, em seguida, discutidos os problemas e vieses na aplicação aos dados reais.

Palavras-chaves: *Network Scale-up*; Populações de difícil acesso; Amostragem em rede; Epidemiologia ; Saúde pública.

Abstract

In order to promote public health policies, it is of utmost importance to get to know the size of hard-to-count populations such as sex workers, illicit drug users, men who have sex with other men, among others. These populations are often stigmatized and even criminalized. However, they have an important role in the dynamics of several transmissible diseases. The network scale-up method (NSUM) uses indirect information about the personal networks of respondents to make population size estimates. In order to illustrate the behavior of the NSUM, think of the following example: respondents are asked questions of the type "How many X do you know?". Besides being asked about hard-to-count populations, respondents are also asked about populations of known size. These questions are used to estimate the total number of people known by a respondent, that is, its degree or personal network size. The size of the unknown subpopulation is then estimated by using responses to questions about the number of people known in the unknown subpopulation, combined with the degree estimate. In this study, network populations were simulated using a random network model. For each generated node, characteristics were randomly assigned; this way, the characteristics of each individual as well as their contacts will be known. Random samples were selected from these populations. Later, estimates for their size were calculated through frequentist and Bayesian models, giving evidence that the Network Scale-up Method has proved to be efficient. The size of the sample, the prevalence of a given subpopulation and other parameters associated to the modeling were also evaluated. In this work, it is was possible to validate the NSUM method under fully controlled conditions and, then, discussed the problems and biases in the application to real data.

Key words: Network Scale-up; Hidden populations; Network sampling; Epidemiology; Public health.

Agradecimentos

Agradeço, em primeiro lugar, a Deus por sempre iluminar a minha trajetória e tornar possível a conclusão de mais uma jornada em minha vida. Por estar ao meu lado em todos os momentos me protegendo e confortando.

Aos meus pais, Lúcia e José, por sempre se sacrificarem para me proporcionar uma boa educação e priorizarem minha vida para que eu pudesse alcançar meus objetivos. Não menos importante, às minhas irmãs, Priscila, Bianca e Leinha (de coração), que mesmo implicando por eu ser a mais nova, sempre me apoiaram e acreditaram no meu potencial. Vocês, sem dúvidas, são as pessoas mais importantes da minha vida.

Aos amigos que fiz durante esses anos de graduação. Em especial, Thaylla Carolina e Luisa Santoro (seguimos invictas) pelos incontáveis dias de diversão e muito estudo a ponto de nunca me deixarem esquecer os conteúdos do curso por me fazerem explicar toda a matéria novamente... Camila Simões por ser minha dupla do vôlei preferida na faculdade. Isabella Philbert e Rosana Gayer que tornaram-se grandes amigas no meu último ano de graduação. Com certeza, são amizades que vou levar pro resto da vida. É impossível listar todos os nomes, pois são tantos os colegas que dividiram essa estrada comigo.

Ao meu (co)orientador, Leo Bastos, que é uma das minhas referências em Estatística, por me apresentar o mundo da pesquisa científica que culminou na realização deste projeto. Tão importante quanto, agradeço, imensamente, a minha orientadora, Mariana Albi, por toda a disponibilidade e dedicação para o desenvolvimento deste trabalho. Você, sem dúvidas, é uma motivação com a qual posso me espelhar para seguir a minha carreira acadêmica.

Aos professores do GET-UFF por todo o aprendizado que me concederam, por toda a paciência e sensibilidade em ensinar. Vocês foram verdadeiros mestres e cada um, de seu modo único, preencheram alguma lacuna do meu conhecimento. Em especial, aos professores, Jony Arrais e Wilson Calmon, por serem excelentes profissionais, tão solicitos e aceitarem o convite para participar da minha banca. Quem sabe um dia nós sejamos companheiros de trabalho...

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

Populações de difícil acesso são populações sobre as quais tem-se dificuldade de obter informações. Como exemplos de populações de difícil acesso pode-se citar populações criminalizadas e/ou estigmatizadas pela sociedade, tais como número de mulheres que fazem aborto, número de usuários de drogas ilícitas, profissionais do sexo, entre tantas outras que são rotuladas negativamente pela sociedade (REIS, 2014). Poder mensurar o tamanho dessas populações é de suma importância principalmente em Saúde Pública e Ciências Sociais, permitindo subsidiar futuras ações que possam contornar o problema que tais populações podem causar. Por exemplo, pensando em uma população de usuários de crack, é interessante identificar o perfil desses usuários, as regiões mais frequentes de utilização de tal substância, entre outras perguntas necessárias para o financiamento de programas para tratar a propagação desse uso inadequado.

O problema de estimar o tamanho de populações de difícil acesso justifica-se pelo fato de os indivíduos não se sentirem à vontade em responder se pertencem ou não às populações problemáticas, isto é, se possuem hábitos e/ou comportamentos que caracterizam tais populações. Sendo assim, utilizar o método direto de pesquisa, que consiste em perguntar a indivíduos da população geral se eles fazem ou não parte da população-alvo a ser estimada, tende a subestimar o tamanho da população em questão.

Existem diversos métodos para estimar o tamanho de populações de difícil acesso, como, por exemplo, o método *Multiplier* (UNAIDS, 2003) que produz estimativas utilizando informações de duas ou mais fontes distintas. A primeira é proveniente de uma amostra da população-alvo e outra que é, habitualmente, obtida em instituições ou programas a qual esta população tem acesso, de forma que os indivíduos desta população tenham chances de serem incluídos em ambas (ou em mais de duas) as fontes de captura de informação. Outra técnica também bastante utilizada é a técnica de captura-recaptura, principalmente na área da ecologia para a estimação do tamanho de populações de animais selvagens (WHITE, 1982). Essa técnica consiste na coleta de duas amostras independentes em dois momentos distintos (amostras sequenciais) de uma população de animais fechada,

ou seja, na qual não são observados nascimentos, mortes ou migrações durante o período de estudo. Além disso, todos os elementos dessa população devem apresentar a mesma probabilidade de captura (COELI; VERAS; COUTINHO, 2000).

Neste trabalho, será utilizado um método indireto nomeado *Network Scale-up* representado pela sigla NSUM (do inglês *Network Scale-up Method*), desenvolvido inicialmente para estimar o número de mortos em um terremoto no México, em 1985, (BERNARD et al., 1991) em uma tentativa de usar o conhecimento dos entrevistados sobre os seus contatos sociais para estimar o número de pessoas que morreram nesta ocasião. Bernard e seus colegas perceberam que a informação que um indivíduo possui sobre os outros em sua rede social pode ser utilizado para estimar as populações que são de difícil acesso (JOHNSEN; KILLWORTH; ROBINSON, 1989; BERNARD et al., 1991; JOHNSEN et al., 1995; KILLWORTH et al., 1998a; KILLWORTH et al., 1998b).

Os dados necessários para o método *Network Scale-up* vem de inquéritos/entrevistas com uma amostra aleatória da população-alvo. Além de questões demográficas básicas, os respondentes nessas entrevistas são perguntados sobre quantas pessoas eles conhecem na população de interesse. Esse método é caracterizado por dois estágios: primeiro estima-se o tamanho da rede de contatos dos indivíduos, utilizando informações/dados de populações conhecidas, ou seja, populações com tamanhos conhecidos de antemão, para as quais é possível contar com cadastros. Posteriormente, condicionado ao tamanho da rede de contatos do indivíduo, pergunta-se também sobre subpopulações de tamanho desconhecido, para geração das estimativas do tamanho destas, o que constitui o objetivo primário da aplicação do método.

Apesar do NSUM ter surgido para um fim muito específico (estimar o número de mortos no terremoto no México), estudos vêm mostrando que este método está sendo utilizado para estimar tamanhos de diferentes populações de difícil acesso. Ainda no México, o método foi utilizado para estimar o número de mulheres que haviam sofrido violência sexual (BERNARD et al., 1991). Em um momento subsequente, foi utilizado para estimar o número de crianças vítimas de experiências de sufocamento/asfixia na Itália (SNIDERO et al., 2007). Nos Estados Unidos esta metodologia vem sendo utilizada em diferentes campos, como, por exemplo, para a estimação da prevalência do HIV, da população de rua em diferentes comunidades (KILLWORTH et al., 1998a) e do uso de heroína (KADUSHIN et al., 2006). No Brasil, o único estudo com uso do método *Network Scale-up* em Saúde pública, realizado no Paraná, estimou o número de usuários de drogas

ilícitas (que não a maconha) no município de Curitiba, em 2011 (SALGANIK et al., 2011a).

Mais recentemente, novos estudos têm sido feitos no sentido de melhorar o método NSUM. McCormick, Salganik e Zheng (2010) desenvolveram estratégias para melhorar a estimativa do grau, isto é, do número de conhecidos do indivíduo. Posteriormente, McCormick e Zheng (2007) propuseram uma curva de calibração para ajustar o erro de transmissão (quando o indivíduo desconhece que um determinado conhecido da sua rede tem a característica de interesse) que foi incorporado por McCormick, Salganik e Zheng (2010). Ezoe et al. (2012) na pesquisa de homens que fazem sexo com outros homens (MSM's), utilizaram a rede de contato dos indivíduos para estimar o número de MSM's.

Neste trabalho, serão simuladas populações em rede aleatória e para cada indivíduo serão atribuídas características. Dessas populações, amostras serão selecionadas para estimar o tamanho de subpopulações de interesse, aqui chamadas de populações-alvo. No Capítulo 2, os objetivos do trabalho serão definidos. Em seguida, no Capítulo 3, serão apresentados métodos diretos e indiretos para estimar o tamanho de populações de difícil acesso sob duas abordagens: frequentista e bayesiana; esta última incluindo estimação via Monte Carlo via Cadeias de Markov (MCMC). Além disso, a abordagem indireta será tratada de duas maneiras diferentes: considerando o grau dos indivíduos conhecido e considerando o grau dos indivíduos desconhecido (NSUM). No Capítulo 4, serão apresentados os resultados baseados em um estudo simulado e uma aplicação em dados reais. Por fim, no Capítulo 5, conclusões a respeito do método *Network Scale-up* e comparações com os outros métodos citados no Capítulo 3 serão apresentadas.

2 Objetivos

O objetivo geral deste trabalho é estimar o tamanho de populações de difícil acesso utilizando métodos diretos e indiretos sob as abordagens frequentista e bayesiana. Os objetivos específicos se apresentam da seguinte forma:

- Estimar e comparar as estimativas do tamanho de populações de difícil acesso através de três metodologias distintas: método de estimação direta, método de estimação indireta com grau conhecido e método *Network Scale-up*.
- Elaborar um estudo simulado como forma de teste para a aplicação das metodologias.
- Avaliar a importância do tamanho da amostra, a prevalência de uma dada subpopulação e outros parâmetros associados à modelagem.
- Utilizar o método Monte Carlo Via Cadeias de Markov como alternativa ao método *Network Scale-up*.
- Descrever a aplicação da metodologia *Network Scale-up* em um estudo de representatividade nacional, comparando-a com outras metodologias abordadas neste trabalho.

3 Metodologia

Em linhas gerais, a Inferência Estatística objetiva estudar a população através de evidências fornecidas pela amostra. É a amostra que contém os elementos que podem ser observados e é onde as quantidades de interesse podem ser medidas.

Na abordagem clássica paramétrica, θ é considerado como uma quantidade desconhecida, mas fixa. Uma amostra aleatória $X_1, \ldots, X_n$ é obtida a partir de uma população indexada por θ e, com base nos valores observados na amostra, o conhecimento sobre o valor de θ é obtido. Na abordagem bayesiana, θ é considerado como uma quantidade aleatória cuja variabilidade pode ser descrita por uma distribuição de probabilidade, chamada de distribuição *a priori*. Esta é uma distribuição subjetiva, baseado na crença do experimentador, e é formulada antes que os dados sejam vistos. Uma amostra é então obtida a partir de uma população indexada por θ, e a distribuição *a priori* é atualizada com esta informação sobre a amostra, dando origem à chamada distribuição *a posteriori* do parâmetro. Esta atualização é feita com uso do Teorema de Bayes, por isso o nome de Inferência bayesiana (CASELLA; BERGER, 2002).

Em cada caso existem diversos possíveis estimadores pontuais para θ, porém, neste trabalho, será utilizado o de máxima verossimilhança no caso frequentista e a média a *posteriori* no caso bayesiano.

Ao longo deste capítulo, as diferentes abordagens para estimação de populações de difícil acesso utilizadas neste trabalho serão apresentadas. Inicialmente, será tratado o método direto, que consiste em perguntar diretamente aos indivíduos da população em geral se eles fazem ou não parte da população de interesse a ser estimada. A seguir, serão tratados os métodos indiretos, que utilizam informações das redes de contatos dos indivíduos para obter informações de forma indireta sobre a população de interesse.

3.1 Método direto

O método direto consiste em perguntar se o indivíduo pertence ou não à população de interesse, ou seja, se eles (os respondentes) têm determinados hábitos ou comportamentos que caracterizam estas populações.

Seja W a variável aleatória (v.a.) indicadora da característica de interesse da população. Para uma amostra aleatória (a.a.) $W_1, \ldots, W_n$ da população W, segue-se que $\forall i$, $i = 1, \ldots, n$,

$$W_i \sim Ber(\theta),$$

onde θ é a proporção de indivíduos com a característica de interesse. Neste caso, para uma população geral de tamanho N, o tamanho da população-alvo a ser estimada é $\theta \times N$.

3.1.1 Estimação frequentista

Sob o ponto de vista frequentista, o estimador de máxima verossimilhança para θ neste caso é a média amostral, que consiste na proporção de indivíduos da amostra que pertencem à população a ser estimada:

$$\hat{\theta}_F = \frac{\sum_{i=1}^{n} W_i}{n}. \tag{3.1}$$

Além disso, pelo Teorema Central do Limite tem-se que, para um tamanho de amostra suficientemente grande, pode-se considerar a proporção amostral $\hat{\theta}_F$ como tendo, aproximadamente, distribuição normal com média θ e variância $\frac{\theta(1-\theta)}{n}$. Observe que a média e a variância de $\hat{\theta}_F$ dependem do parâmetro desconhecido θ. No entanto, pelo fato de n ser grande, pode-se aproximar θ por $\hat{\theta}_F$ na variância da distribuição amostral. Dessa forma, um intervalo de confiança de $100(1-\alpha)\%$ para θ pode ser aproximado por:

$$IC_{100(1-\alpha)\%}(\theta) = \left[\hat{\theta}_F - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{\theta}_F(1-\hat{\theta}_F)}{n}}; \hat{\theta}_F + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{\theta}_F(1-\hat{\theta}_F)}{n}}\right], \tag{3.2}$$

onde $z_{1-\frac{\alpha}{2}}$ é o quantil $1-\frac{\alpha}{2}$ da distribuição normal padrão.

3.1.2 Estimação bayesiana

Atribuindo uma distribuição *a priori* Beta com parâmetros (a_θ,b_θ) para θ, segue da conjugação, uma vez que a família de distribuições Beta é conjugada ao modelo Bernoulli, que $\theta|\underset{\sim}{W} = \underset{\sim}{w} \sim Beta\left(\sum_{i=1}^{n} w_i + a_\theta, n - \sum_{i=1}^{n} w_i + b_\theta\right)$. Neste caso, um estimador de Bayes para θ será sua média *a posteriori* dada por

$$\hat{\theta}_B = \mathbb{E}[\theta|\underset{\sim}{W}] = \frac{\sum_{i=1}^{n} W_i + a_\theta}{n + a_\theta + b_\theta}. \tag{3.3}$$

Um intervalo de $100(1-\alpha)\%$ de credibilidade para θ será dado pelos quantis $\frac{\alpha}{2}$ e $1-\frac{\alpha}{2}$ da distribuição *a posteriori* de θ, isto é, da distribuição Beta $\left(\sum_{i=1}^{n} w_i + a_\theta, n - \sum_{i=1}^{n} w_i + b_\theta\right)$.

3.2 Método indireto com grau conhecido

Este método considera que o tamanho da rede de contatos do indivíduo (grau do indivíduo) é conhecido, portanto só é utilizado em dados simulados no qual é possível determinar com exatidão o tamanho desta. Faz-se isso pois, na prática, é dificil ou até mesmo impossível, enumerar com precisão o tamanho da rede de contato das pessoas que, em geral, é "grande". Portanto, para esse método, só é necessário contabilizar o número de conhecidos de um determinado indivíduo que possua a característica da população de interesse e, baseado nessa informação associada ao grau do indivíduo i, $i = 1, \ldots, n$, conhecido, é possível determinar a estimativa de θ relacionada à variável $Y_i \sim Bin(\delta_i, \theta)$.

Mais especificamente, seja Y_i a variável que conta o número de indivíduos que i conhece na população de interesse desconhecida, para $i = 1, \ldots, n$,

$$Y_i \sim Bin(\delta_i, \theta),$$

onde δ_i é o grau (número de conhecidos) do indivíduo i e θ é a proporção desconhecida de indivíduos com a característica de interesse. Assim como na Seção 3.1, supondo que a população de interesse é subpopulação de uma população geral de tamanho N, o tamanho da população a ser estimado será dado por $\theta \times N$.

3.2.1 Estimação frequentista

Seja $Y_1, ..., Y_n$ amostra aleatória tal que $Y_i \sim Bin(\delta_i, \theta)$, $i = 1, \ldots, n$. Logo, a função de verossimilhança para θ será dada por

$$L(\theta|\underset{\sim}{y}) = \left[\prod_{i=1}^{n} \binom{\delta_i}{y_i}\right] \theta^{\sum_{i=1}^{n} y_i} (1-\theta)^{\sum_{i=1}^{n} (\delta_i - y_i)},$$

donde segue que o estimador de máxima verossimilhança para θ será dado por

$$\hat{\theta}_F = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} \delta_i}. \tag{3.4}$$

Como $Y_1, ..., Y_n$ são independentes com distribuição $Bin(\delta_i, \theta)$, $i = 1, \ldots, n$, então $\sum_{i=1}^{n} Y_i \sim Bin\left(\sum_{i}^{n} \delta_i, \theta\right)$. Com isso, a variância de $\hat{\theta}_F$ será dada por:

$$Var(\hat{\theta}_F) = \frac{1}{\left(\sum_{i=1}^{n} \delta_i\right)^2} Var\left(\sum_{i=1}^{n} Y_i\right) = \frac{\theta(1-\theta)}{\sum_{i=1}^{n} \delta_i}.$$

Logo, um estimador para a variância de $\hat{\theta}_F$ será $\widehat{Var}(\hat{\theta}_F) \approx \frac{\hat{\theta}_F(1-\hat{\theta}_F)}{\sum_{i=1}^{n} \delta_i}$. Novamente, pelo Teorema Central do Limite, pode-se considerar que $\hat{\theta_F} \sim N\left(\theta, \frac{\theta(1-\theta)}{\sum_{i=1}^{n} \delta_i}\right)$ e, portanto, construir um intervalo de $100(1-\alpha)\%$ de confiança para θ da forma:

$$IC_{100(1-\alpha)\%}(\theta) = \left[\hat{\theta}_F - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\theta}_F(1-\hat{\theta}_F)}{\sum_{i=1}^{n} \delta_i}}; \hat{\theta}_F + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\theta}_F(1-\hat{\theta}_F)}{\sum_{i=1}^{n} \delta_i}}\right], \tag{3.5}$$

onde $z_{1-\frac{\alpha}{2}}$ é o quantil $1-\frac{\alpha}{2}$ da distribuição normal padrão.

3.2.2 Estimação bayesiana

Atribuindo uma distribuição *a priori* Beta (a_θ,b_θ) para θ, a distribuição *a posteriori* é $Beta\left(\sum_{i=1}^{n} y_i + a_\theta, \sum_{i=1}^{n} \delta_i - \sum_{i=1}^{n} y_i + b_\theta\right)$. Então, tomando como estimador pontual a média *a posteriori* de θ, segue que,

$$\hat{\theta}_B = \mathbb{E}[\theta|\underset{\sim}{Y}] = \frac{\sum_{i=1}^{n} Y_i + a_\theta}{\sum_{i=1}^{n} \delta_i + a_\theta + b_\theta}, \tag{3.6}$$

e, um intervalo de $100(1-\alpha)\%$ de credibilidade para θ será dado pelos quantis $\frac{\alpha}{2}$ e $1-\frac{\alpha}{2}$ da distribuição *a posteriori* de θ, isto é, da distribuição Beta $\left(\sum_{i=1}^{n} y_i + a_\theta, \sum_{i=1}^{n} \delta_i - \sum_{i=1}^{n} y_i + b_\theta\right)$.

3.3 Método indireto com grau desconhecido - NSUM

A metodologia *Network Scale-up* produz estimativas de tamanhos populacionais valendo-se de informações das redes de contatos dos respondentes de inquéritos realizados com uma amostra aleatória da população geral, tendo como pressupostos básicos: (a) todos têm a mesma chance de conhecer alguém de uma dada subpopulação; (b) o tamanho da rede de contatos é constante, isto é, não são observados nascimentos, mortes ou migrações durante o período de estudo; (c) todos conhecem bem os comportamentos dos membros de sua rede de contatos (KILLWORTH et al., 1998b).

O método *Network Scale-up* é caracterizado em duas etapas. A primeira etapa consiste em estimar a rede de contatos de cada indivíduo, isto é, o número de conhecidos (grau) do indivíduo baseando-se em m subpopulações conhecidas. Dessa forma, pergunta-se a uma pessoa aleatoriamente selecionada da população geral "quantos amigos você conhece na população X?". McCarty et al. (2001) define um conhecido como uma pessoa que vive na região da população X a qual se conhece de vista e de nome, desde que esta pessoa também conheça o indivíduo e que ambos tenham tido algum contato (por e-mail, telefone, rede social ou pessoalmente) nos últimos 2 anos. Esse passo é necessário pois como o tamanho da rede de contatos das pessoas, em geral, é "grande"e frequentemente difícil de precisar, o respondente tende a não enumerar/contar de fato cada um de seus contatos, e sim "chutar" (intuir) esse número, o que poderia gerar viés na estimativa NSUM. A segunda etapa consiste em estimar o tamanho da população-alvo utilizando

as estimativas acerca do tamanho da rede de contatos do respondente obtida na etapa anterior.

Para estimar o grau do indivíduo i (δ_i) é preciso definir a variável X_{ik} que conta o número de conhecidos de i na k-ésima subpopulação, para $i = 1, 2, \ldots, n$ e $k = 1, 2, \ldots, m$. Sem perda de generalidade, suponha que m subpopulações de uma população geral de tamanho N têm seus tamanhos conhecidos, denotados por N_k, $k = 1, ..., m$, e que outra subpopulação (de difícil acesso), aqui chamada população de interesse ou população-alvo, tem seu tamanho desconhecido, por exemplo, usuários de drogas pesadas. A estimativa do *Network Scale-up* assume que $X_{ik} \sim Bin\left(\delta_i, \frac{N_k}{N}\right)$ (KILLWORTH et al., 1998a; KILLWORTH et al., 1998b). Entretanto, a fim de encontrar analiticamente um estimador para δ_i, se a rede de contatos é grande, e a prevalência da característica da subpopulação k é pequena, propõe-se, neste trabalho, a utilização da Lei dos Pequenos Números (Anexo A), que permite a aproximação do modelo Binomial pelo modelo Poisson. Mais especificamente, nestas condições, pode-se considerar

$$X_{ik} \approx Pois\left(\delta_i \frac{N_k}{N}\right),$$

onde X_{ik} é o número de conhecidos do respondente i na subpopulação k, δ_i é o grau do indivíduo i, N_k é o tamanho da subpopulação k conhecida, $k = 1, 2, \ldots, m$, e N é o tamanho da população geral. Normalmente, utilizam-se 20 subpopulações conhecidas para estimar o tamanho da rede de contato dos respondentes (REIS, 2014).

Uma vez estimado o grau do indivíduo i, ao definir a variável Y_i como no método indireto com grau conhecido (Seção 3.2) basta substituir δ_i por seu respectivo estimador $\hat{\delta}_i$; portanto, para $i = 1, \ldots, n$:

$$Y_i \sim Bin(\hat{\delta}_i, \theta),$$

onde $\hat{\delta}_i$ é o número estimado da rede de contatos do respondente i e θ é a proporção desconhecida de indivíduos com a característica de interesse.

Dessa forma, o processo de estimação via NSUM consistirá nas seguintes etapas:

- **etapa 1**: utilizar a aproximação do modelo Binomial pelo Poisson, estimando o grau δ_i de cada indivíduo i; e
- **etapa 2**: estimar a proporção θ da população de difícil acesso levando em consideração o grau do indivíduo estimado na etapa anterior.

3.3.1 Estimação frequentista

Utilizando a aproximação $X_{ik} \approx Poisson\left(\delta_i \frac{N_k}{N}\right)$, e considerando $\underset{\sim}{x_{i.}} = (x_{i1}, \ldots, x_{im})$, pode-se aproximar a função de verossimilhança de δ_i por

$$L(\delta_i | \underset{\sim}{x_i}) = \prod_{k=1}^{m} \frac{\left(\delta_i \frac{N_k}{N}\right)^{x_{ik}} e^{-\delta_i \frac{N_k}{N}}}{x_{ik}!} = \frac{\left[\prod_{k=1}^{m} \delta_i^{x_{ik}}\right] \left[\left(\frac{N_k}{N}\right)^{\sum_{k=1}^{m} x_{ik}}\right] e^{-\delta_i \sum_{k=1}^{m} \frac{N_k}{N}}}{\prod_{k=1}^{m} x_{ik}!}. \tag{3.7}$$

Utilizando a expressão 3.7, pode-se provar que o EMV para δ_i neste caso é

$$\hat{\delta}_{iF} = \frac{\sum_{k=1}^{m} X_{ik}}{\frac{\sum_{k=1}^{m} N_k}{N}}. \tag{3.8}$$

Dada a estimativa $\hat{\delta}_{iF}$ e, em analogia à estimativa indireta com grau conhecido (conforme discutido na Seção 3.2.1), pode-se assumir que

$$\hat{\theta}_F = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} \hat{\delta}_{iF}} \quad e, \tag{3.9}$$

$$IC_{100(1-\alpha)\%}(\theta) = \left[\hat{\theta}_F - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\theta}_F(1-\hat{\theta}_F)}{\sum_{i=1}^{n} \hat{\delta}_{iF}}}; \hat{\theta}_F + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\theta}_F(1-\hat{\theta}_F)}{\sum_{i=1}^{n} \hat{\delta}_{iF}}}\right], \tag{3.10}$$

onde $z_{1-\frac{\alpha}{2}}$ é o quantil $1-\frac{\alpha}{2}$ da distribuição normal padrão.

3.3.2 Estimação bayesiana

Assim como na estimação frequentista (Seção 3.3.1), a estimação será feita em duas etapas: estima-se os graus dos indivíduos e, a seguir, condicional às estimativas encontradas para os graus, estima-se o tamanho da população de interesse. Para a etapa 1, utiliza-se as populações conhecidas para estimar δ_i. Após utilizar a aproximação do

modelo Binomial pelo modelo Poisson (conforme discutido na Seção 3.3), assumindo uma distribuição *a priori* $Gama(a_\delta, b_\delta)$ para δ_i segue da conjugação que

$$\delta_i | \underset{\sim}{X}_{i.} = \underset{\sim}{x}_{i.} \sim Gama\left(\sum_{k=1}^{m} x_{ik} + a_\delta, \frac{\sum_{k=1}^{m} N_k}{N} + b_\delta\right).$$

De fato, apesar do grau do indivíduo representar uma quantidade discreta, foi considerada uma aproximação na qual uma distribuição *a priori* contínua gama foi utilizada, com o intuito de explorar a conjugação com a distribuição de Poisson. Neste caso, uma estimativa pontual para δ_i será dada por:

$$\hat{\delta}_{iB} = \mathbb{E}[\delta_i | \underset{\sim}{X}_{i.} = \underset{\sim}{x}_{i.}] = \frac{\sum_{k=1}^{m} x_{ik} + a_\delta}{\frac{\sum_{k=1}^{m} N_k}{N} + b_\delta}. \tag{3.11}$$

A etapa 2 consiste em estimar θ dada a estimativa 3.11. Atribuída uma distribuição *a priori* conjugada Beta (a_θ, b_θ) para θ e considerando $\underset{\sim}{\delta}$ como sendo o vetor $(\delta_1, ..., \delta_n)$, a distribuição *a posteriori* de θ condicional à estimativa pontual $\hat{\underset{\sim}{\delta}}$ será dada por:

$$\theta | (\underset{\sim}{Y}, \underset{\sim}{\delta}) = (\underset{\sim}{y}, \hat{\underset{\sim}{\delta}}) \sim Beta\left(\sum_{i=1}^{n} y_i + a_\theta, \sum_{i=1}^{n} \hat{\delta}_{iB} - \sum_{i=1}^{n} y_i + b_\theta\right).$$

Então, uma estimativa pontual para θ é da forma:

$$\hat{\theta}_B = \mathbb{E}[\theta | \underset{\sim}{Y} = \underset{\sim}{y}, \underset{\sim}{\delta} = \hat{\underset{\sim}{\delta}}] = \frac{\sum_{i=1}^{n} y_i + a_\theta}{\sum_{i=1}^{n} \hat{\delta}_{iB} + a_\theta + b_\theta}. \tag{3.12}$$

Neste caso, um intervalo de credibilidade de $100(1-\alpha)\%$ de credibilidade para θ será dado pelos quantis de $\frac{\alpha}{2}$ e $1 - \frac{\alpha}{2}$ da distribuição a posteriori de θ, isto é, da distribuição Beta $\left(\sum_{i=1}^{n} y_i + a_\theta, \sum_{i=1}^{n} \hat{\delta}_{iB} - \sum_{i=1}^{n} y_i + b_\theta\right)$.

3.3.3 Estimação via Monte Carlo via Cadeias de Markov

Conforme visto nesta seção, tanto sob o ponto de vista frequentista quanto sob o ponto de vista bayesiano, o método NSUM baseia-se na estimação em duas etapas. Na etapa 1, baseados apenas nas informações sobre as subpopulações conhecidas (X_{ik}, $k = 1, ..., m$),

estima-se os graus dos indivíduos. Vale notar, porém, que a população de interesse também traz informações sobre δ_i. De fato, supõe-se $Y_i \sim Bin(\delta_i, \theta)$, indicando que as observações $y_1, \ldots, y_n$ podem conter informações relevantes sobre os graus; informações estas que não são bem aproveitadas pelo método NSUM. Visando utilizar um método alternativo que aproveite as informações de Y_i na estimação dos graus, propõe-se neste trabalho a utilização de um método iterativo de estimação, baseado em simulações de Monte Carlo.

Os métodos de Monte Carlo via cadeias de Markov (MCMC) são uma alternativa aos métodos não iterativos em problemas complexos. A ideia central nos métodos MCMC é utilizar técnicas de simulação iterativa, baseadas em cadeias de Markov, para obter uma amostra da distribuição *a posteriori* conjunta de todos os parâmetros e calcular estimativas amostrais de características desta distribuição. A simulação é feita com base em distribuições auxiliares que percorrem o espaço paramétrico, convergindo para uma distribuição estacionária, que é a de interesse no problema (EHLERS, 2007).

Nesta seção, um particular algoritmo MCMC, denominado Amostrador de Gibbs (Anexo B), será utilizado para gerar amostras da distribuição *a posteriori* conjunta de

$$(\delta_1, ..., \delta_n, \theta)|(\underset{\sim}{X} = \underset{\sim}{x}, \underset{\sim}{Y} = \underset{\sim}{y}),$$

onde $\underset{\sim}{X} = \{X_{ik}, i = 1, \ldots, n \text{ e } k = 1, \ldots, m\}$.

Mais especificamente, após observar $\underset{\sim}{x}$ e $\underset{\sim}{y}$, amostras da distribuição conjunta

$$p(\underset{\sim}{\delta}, \theta|\underset{\sim}{x}, \underset{\sim}{y}) \propto p(\underset{\sim}{\delta}, \theta, \underset{\sim}{x}, \underset{\sim}{y}) = p(\underset{\sim}{y}|\underset{\sim}{\delta}, \theta, \underset{\sim}{x})p(\underset{\sim}{x}|\underset{\sim}{\delta}, \theta)p(\underset{\sim}{\delta}|\theta)p(\theta) = p(\underset{\sim}{y}|\underset{\sim}{\delta}, \theta)p(\underset{\sim}{x}|\underset{\sim}{\delta})p(\underset{\sim}{\delta})p(\theta)$$

serão obtidas gerando alternadamente valores das distribuições condicionais completas:

$$p(\delta_i|\underset{\sim}{y}, \underset{\sim}{x}, \theta) \ \alpha \ p(y_i|\theta, \delta_i)p(x_i|\delta_i)p(\delta_i), \ i = 1, \ldots, n$$

$$p(\theta|\underset{\sim}{\delta}, \underset{\sim}{y}, \underset{\sim}{x}) \ \alpha \ \prod_{i=1}^{n} p(y_i|\theta, \delta_i)p(\theta).$$

Aproximando as distribuições de Y_i e X_{ik} pelas respectivas distribuições de Poisson, e tomando as distribuições *a priori* conjugadas $\delta_i \sim Gama(a_\delta, b_\delta)$, $i = 1, ..., n$, e $\theta \sim$

$Beta(a_\theta, b_\theta)$ (seguindo as mesmas ideias do que foi feito na Seção 3.3.2), as condicionais completas neste caso tomam a forma:

$$\delta_i|(\underset{\sim}{Y}, \underset{\sim}{X}, \theta) \sim Gama\left(a_\delta + \sum_{k=1}^{m} X_{ik} + Y_i, b_\delta + \frac{\sum_{k=1}^{m} N_k}{N} + \theta\right), \ i = 1, \ldots, n, \tag{3.13}$$

$$\theta|(\underset{\sim}{\delta}, \underset{\sim}{Y}, \underset{\sim}{X}) \sim Beta\left(a_\theta + \sum_{i=1}^{n} Y_i, b_\theta + \sum_{i=1}^{n} \delta_i - \sum_{i=1}^{n} Y_i\right). \tag{3.14}$$

Com base nas amostras geradas das distribuições condicionais obtidas em 3.13 e 3.14, serão calculadas estimativas pontuais (médias *a posteriori*) e intervalares (MDP) relacionadas aos parâmetros $\underset{\sim}{\delta}$ e θ.

4 Análise dos Resultados

O objetivo deste capítulo é apresentar e discutir os resultados provenientes das estimativas dos métodos apresentados no Capítulo 3. Os resultados serão obtidos em duas diferentes situações: um estudo simulado e uma aplicação em dados reais. Uma das dificuldades da verificação dos resultados de estimação do NSUM é que não se sabe o tamanho certo das subpopulações de difícil-acesso. Dessa forma, a Seção 4.1 trata de um estudo simulado para obter estimativas de subpopulações conhecidas a serem comparadas com o verdadeiro tamanho, para avaliar a eficácia dos métodos diretos e indiretos tratados neste trabalho. Além disso, nesta seção também será abordado o conceito de rede aleatória, uma vez que esta tem papel essencial para a elaboração das redes de contatos dos indivíduos que fundamentalizam os métodos indiretos. Por último, na Seção 4.2, será realizada uma aplicação em dados reais no qual espera-se o mesmo comportamento apresentado pelos métodos no estudo simulado.

Tanto no estudo simulado quanto na aplicação a dados reais, sob abordagem bayesiana, como o objetivo não é priorizar nenhum conhecimento *a priori* para o parâmetro de interesse, será atribuída uma distribuição *a priori* uniforme no intervalo (0,1) para θ, isto é, uma distribuição *a priori* não-informativa. De fato, isto equivale a tomar $a_\theta = b_\theta = 1$ na distribuição *a priori* conjugada $\theta \sim Beta(a_\theta, b_\theta)$. Para a rede de contatos individual, δ_i, também chamada grau da rede, será atribuída uma distribuição *a priori* não-informativa baseada no número de Dunbar, que define um número limitado de pessoas com o qual se poderia fazer uma relação social estável (KUDO; DUNBAR, 2001). Baseada na obra de Dunbar, que limita o número de conhecidos de um indivíduo em 150, pode-se elicitar uma distribuição *a priori* gama para δ_i tal que a média é 150 e a variabilidade é alta (para gerar uma distribuição *a priori* não-informativa). Sendo assim, será considerada uma distribuição *a priori* para o grau da forma $\delta_i \sim Gama(a_\delta = 0,00015; b_\delta = 0,000001)$. Todos os resultados a serem apresentados neste Capítulo foram obtidos utilizando o *software* estatístico R (R Core Team, 2016).

4.1 Estudo Simulado

4.1.1 Rede Aleatória

Conforme discutido no capítulo anterior, métodos indiretos para estimação em populações de difícil acesso baseiam-se no uso de informações da rede de contatos dos indivíduos de uma população geral. Assim, um passo importante para um estudo simulado neste trabalho é a construção de uma rede aleatória que simule o relacionamento entre indivíduos, bem como suas características, representativas das subpopulações em estudo.

Uma rede aleatória, também denominada por grafo aleatório, é um conjunto de v vértices (ou nós) aos quais arestas são adicionadas aleatoriamente com probabilidade p (ERDOS, 1959). Diferentes modelos de grafos aleatórios produzem diferentes distribuições de probabilidade nos grafos. Existem diversos tipos de redes aleatórias, tais como, rede de pequeno mundo (conhecidas pelo inglês *small-word networks*), proposta por Watts e Strogatz (1998), na qual são gerados grafos matemáticos com um certo número de nós com poucas interligações diretas onde qualquer par de nós pode ser conectado por um número relativamente pequeno de passos através das ligações existentes. Entretando, neste estudo simulado, como o intuito era criar redes nas quais não existe nenhum critério que privilegie ligações em relação a outras e que pudesse ser definida uma quantidade v de indivíduos e probabilidade p de eles estarem conectados aleatoriamente, foi utilizada a rede Erdos- Rényi, simulada em R usando a biblioteca igraph (CSARDI; NEPUSZ, 2006). Portanto, para gerar um modelo de Erdos-Rényi, dois parâmetros devem ser especificados: o número de nós no gráfico gerado e a probabilidade de que uma ligação seja formada entre dois nós (ERDOS; RÉNYI, 1960). Para um par de valores de (v, p), diversos grafos podem ser gerados.

A Figura 1 mostra uma realização de um grafo aleatório, baseado na escolha de $v = 30$ vértices e uma probabilidade $p = 2log(v)/v$ de um vértice estar conectado a qualquer outro. A probabilidade p foi escolhida baseada no fato de que a maior probabilidade para o qual um grafo aleatório está conectado é dada por $2log(v)/v$ (ERDOS, 1959). Além disso, para cada nó foram atribuídas algumas características binárias que caracterizam determinados grupos. Assim, pode-se pensar que a Figura 1 ilustra uma população de 30 indivíduos representados pelos vértices no qual cada indivíduo está conectado a outros com características similares e/ou distintas e, dessa forma, é possível determinar o grau de um indivíduo, isto é, número de conhecidos através da quantidade de arestas conectadas a ele. Com isso, pode-se utilizar informações a respeito da rede de contatos dos indivíduos

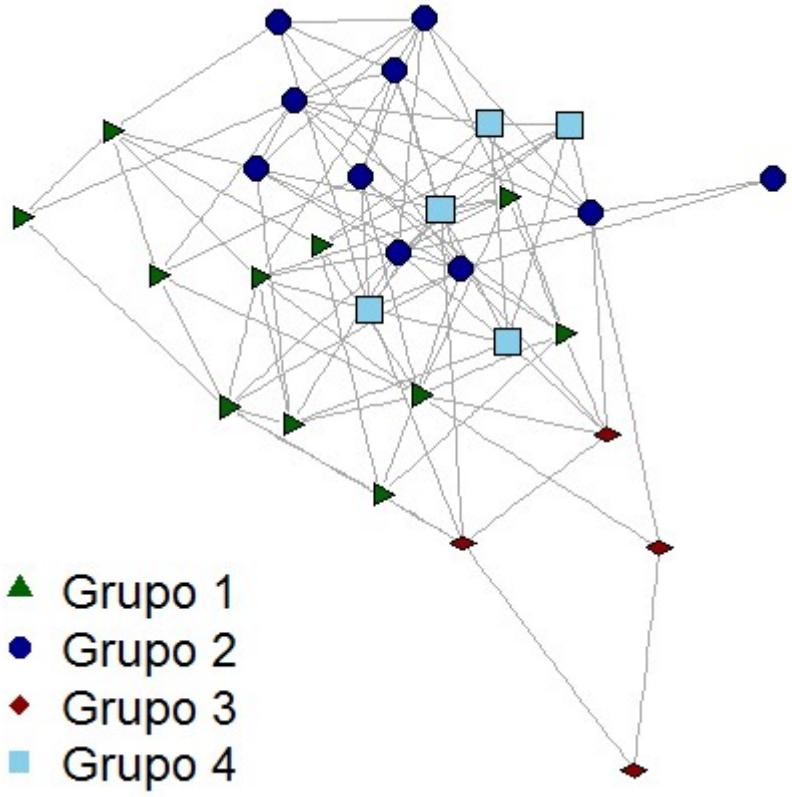


Figura 1: Rede aleatória simulada pelo modelo de Erdos-Rényi.

para estimar a proporção de uma característica de interesse como utilizado no método *Network Scale-up*.

4.1.2 *Toy Example*

A fim de ilustrar o comportamento do método NSUM, assim como comparar suas estimativas com as estimativas obtidas pelo método direto e pelo método indireto com grau conhecido, 300 populações de $N = 1000$ indivíduos, usando as redes de Erdos-Rényi, foram geradas. Para cada indivíduo dessas populações, foram atribuíduas 8 características binárias com uma mesma probabilidade θ (0,02; 0,2; 0,5). Uma dada subpopulação foi considerada desconhecida (população de interesse) e o objetivo desse estudo simulado era obter estimativas dessa população-alvo considerando as outras 7 subpopulações conhecidas. Para todas as simulações, as estimativas obtidas foram baseadas em amostras de tamanho 50, 100 e 500.

Através dessas simulações, mediu-se a raiz do erro quadrático médio (RMSE) para verificar o quanto de erro ocorre nas estimativas dos tamanhos populacionais obtidas por cada método. Neste caso, para as 300 populações o RMSE será definido por:

$$\sqrt{\frac{\sum_{t=1}^{300}(\hat{\theta}_t - \theta)^2}{300}} \times N,$$

onde $\hat{\theta}_t$ é a estimativa da característica de interesse na população t, $t = 1, 2, \ldots, 300$, θ é o valor verdadeiro da probabilidade da característica de interesse comum em todas as populações (0,02; 0,2; 0,5) e N é o total da população geral ($N = 1000$).

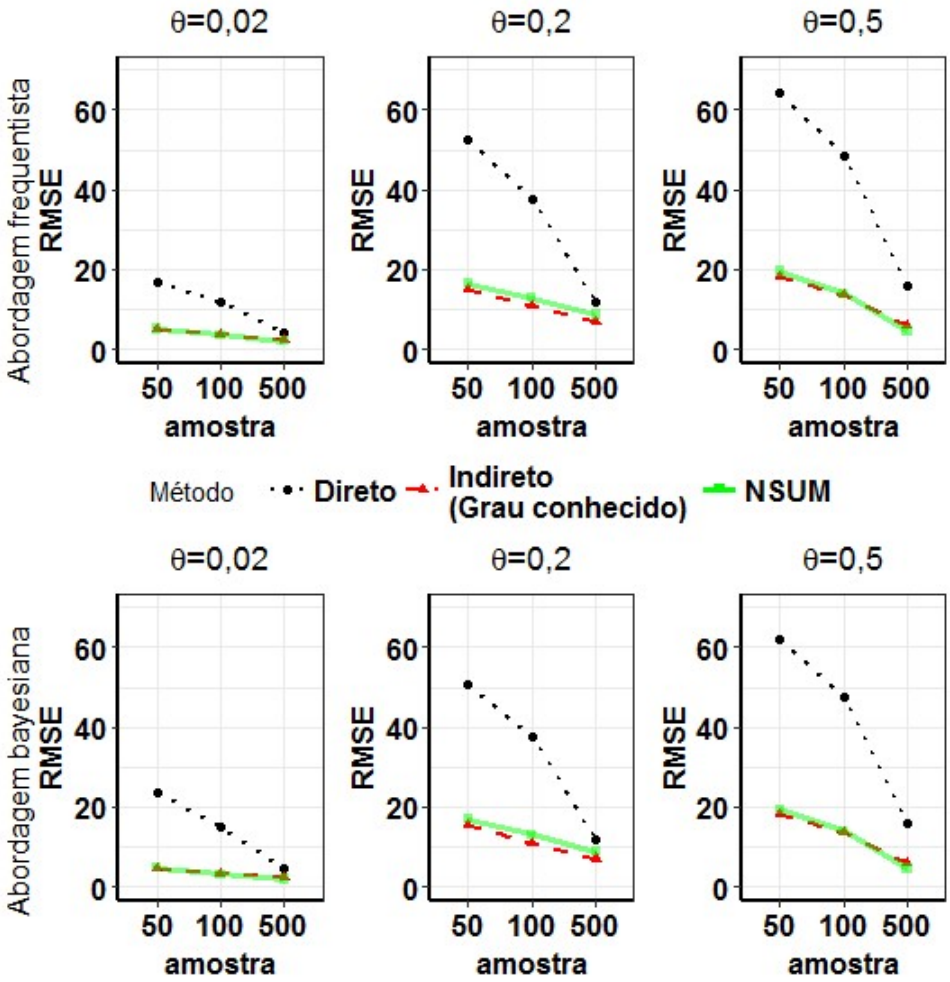


Figura 2: RMSE para amostras de tamanho 50, 100 e 500 de populações de 1000 indivíduos com 2%, 20% e 50% dos indivíduos com a característica de interesse, respectivamente.

Pode-se perceber através da Figura 2 que, em todas as situações (para todo tamanho de amostra fixado e para todo valor de θ), o método direto apresentou o RMSE mais alto, de modo que há evidências de que o método indireto é mais preciso. No entanto, quando comparada a abordagem frequentista com a abordagem bayesiana não são encontradas diferenças significativas entre elas. Além disso, quanto menor o tamanho da amostra e maior o θ, maior será o RMSE. O método direto é influenciado fortemente pelo tamanho da amostra, enquanto os métodos indiretos mantém-se quase que invariantes. Portanto, é importante destacar que mesmo com tamanho de amostra pequeno, as estimativas geradas pelos métodos indiretos são precisas, o que é, na prática, interessante por reduzir custos de pesquisa.

Após verificar o bom desempenho dos métodos indiretos quando comparados ao método direto com base nas 300 populações, todo o processo foi aplicado para uma única população com a finalidade de estimar o tamanho de uma determinada subpopulação de interesse. Sendo assim, foi gerada uma rede de 10000 indivíduos, no qual foram atribuídas 20 características binárias com probabilidade $\theta = 0,2$. Desta população, foi selecionada uma amostra aleatória de tamanho 500. Suponha que a característica de interesse é o gênero do indivíduo e é preciso estimar o número de mulheres dessa população. Com este objetivo, os métodos direto e indiretos serão utilizados sob as abordagens frequentista e bayesiana.

Para todos os métodos e abordagens serão apresentadas estimativas pontuais (EMV no caso frequentista e médias *a posteriori* no caso bayesiano) e intervalares. No caso das estimativas intervalares, intervalos de confiança/credibilidade de 95% serão obtidos. Vale ressaltar aqui que todos os intervalos de credibilidade (bayesianos) serão tomados de forma a representarem intervalos de máxima densidade *a posteriori* (MDP), obtidos com o auxílio da função `HPDinterval`, disponível no pacote coda (PLUMMER et al., 2006), do R.

A Tabela 1 apresenta as estimativas referentes aos indivíduos do sexo feminino, utilizando o método direto sob as duas abordagens estatísticas. Verifica-se a superposição

Tabela 1: Estimativas para a proporção de indivíduos do sexo feminino obtidas via método direto sob as abordagens frequentista e bayesiana.

Abordagem	$\hat{\theta}$	$IC_{95\%}(\theta)$
Frequentista	0,2120	[0,1761 , 0,2478]
Bayesiana	0,2131	[0,1811 , 0,2508]

dos intervalos de confiança e credibilidade associados ao parâmetro θ, isto indica que, estatisticamente, as estimativas referentes a proporção de mulheres são similares entre as abordagens frequentista e bayesiana. Note que, tal conclusão já era esperada visto que foi utilizada uma distribuição *a priori* não-informativa para θ. Com base nas estimativas apresentadas na Tabela 1 para a proporção de mulheres, obter estimativas para o tamanho desta subpopulação de interesse é simples. Basta lembrar que $\hat{\theta} \times N$ é o estimador do total populacional e, dessa forma, como $N = 10000$, as abordagens frequentista e bayesiana apresentam, respectivamente, estimativas de 2120 e 2131 para o total de mulheres na população geral.

Sabendo o grau de cada indivíduo, pode-se encontrar as estimativas para a proporção de mulheres utilizando o método indireto com grau conhecido como mostra a Tabela 2.

Tabela 2: Estimativas para a proporção de indivíduos do sexo feminino obtidas via método indireto com grau conhecido sob as abordagens frequentista e bayesiana.

Abordagem	$\hat{\theta}$	$IC_{95\%}(\theta)$
Frequentista	0,1989	[0,1907 , 0,2071]
Bayesiana	0,1990	[0,1912 , 0,2071]

Novamente, não houve diferenças significativas entre as duas abordagens, devido a distribuicão *a priori* não-informativa atribuída a θ. Entretanto, ao comparar as estimativas intervalares deste método com o método direto (apresentado na Tabela 1), segue-se que, para este último, os intervalos de confiança/credibilidade de 95% tem maior amplitude, indicando estimativas menos precisas. Neste caso, a população de mulheres é estimada em 1989 e 1990, segundo as abordagens frequentista e bayesiana, respectivamente.

No método indireto com grau desconhecido, *Network Scale-up*, primeiro estimou-se o grau dos indivíduos. Lembrando que, na abordagem bayesiana, para a rede de contatos individual, δ_i, foi considerado uma distribuição *a priori* da forma $\delta_i \sim Gama(0,00015; 0,000001)$ baseada no número de Dunbar. As estimativas pontuais e intervalares para o grau individual podem ser observadas na Figura 3.

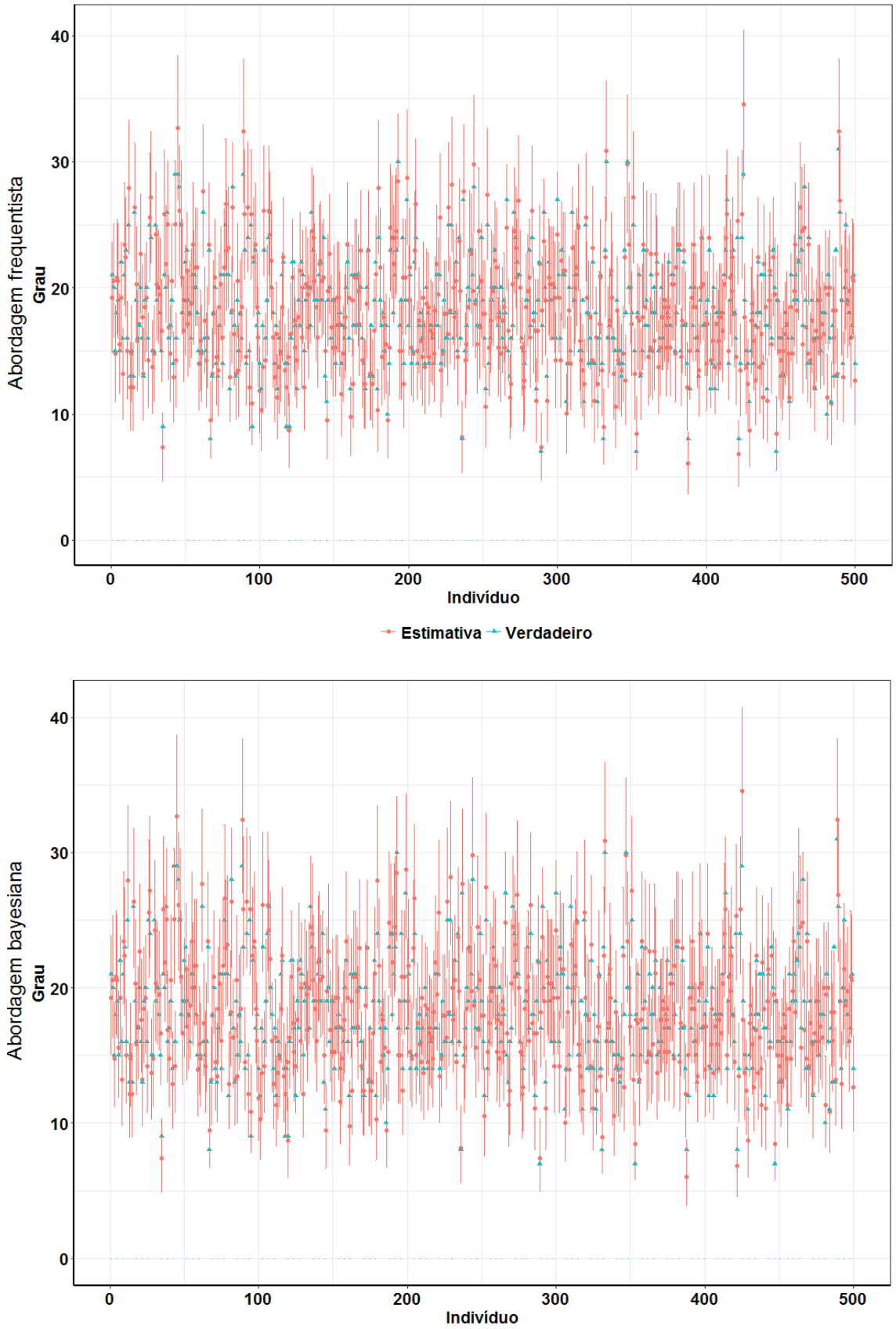


Figura 3: Comparacão entre o valor verdadeiro e as estimativas (pontuais e intervalares) obtidas para os graus dos 500 indivíduos da amostra via NSUM.

Na Figura 3 pode-se observar as estimativas pontuais, representadas pelos pontos, e os respectivos intervalos de confiança/credibilidade de 95% (representados pelos segmentos verticais) para os graus dos 500 indivíduos da amostra, além dos valores verdadeiros destes parâmetros (representados na figura pelos triângulos). Pode-se perceber que o método de populações conhecidas utilizado para estimar o grau, de fato, é eficiente, pois, as estimativas pontuais estão próximas do valor verdadeiro e este último está no intervalo de confiança/credibilidade de 95% estimado. De forma mais precisa, as abordagens frequentista e bayesiana apresentaram proporção de cobertura de $96,6\%$ e 97%, respectivamente.

Dessa forma, condicionado à estimativa dos graus dos indivíduos, e atribuindo uma distribuição *a priori* Beta $(1, 1)$ para θ, a proporção de mulheres na população tanto na versão frequentista quanto na bayesiana é dada pela Tabela 3, a seguir.

Tabela 3: Estimativas para a proporção de indivíduos do sexo feminino obtidas via método NSUM sob as abordagens frequentista e bayesiana.

Abordagem	$\hat{\theta}$	$IC_{95\%}(\theta)$
Frequentista	0,1990	[0,1908 , 0,2072]
Bayesiana	0,1991	[0,1912 , 0,2068]

Pela Tabela 3 chega-se a mesma conclusão obtida pela Tabela 2, isto é, os métodos indiretos apresentaram estimativas próximas entre elas e mais precisas que o método direto. Mais uma vez, para a população geral de $N = 10000$ indivíduos considerada no contexto do problema, as estimativas do número de mulheres nessa população é dada por 1990 e 1991, sob as abordagens frequentista e bayesiana, respectivamente.

As Figuras 4 e 5 sumarizam as informações obtidas pelos métodos (direto e indiretos) a fim de facilitar a visualização e interpretação dos resultados.

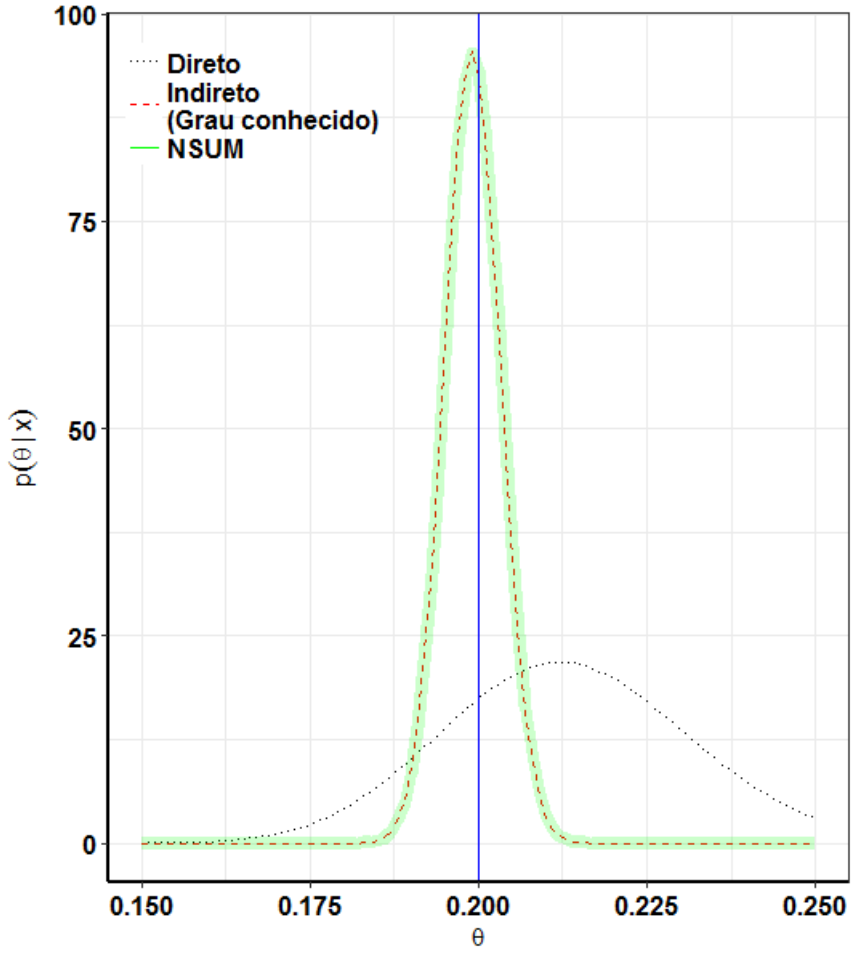


Figura 4: Distribuição *a posteriori* de θ considerando os três métodos sob abordagem bayesiana (linha vertical: valor verdadeiro).

Através das Figuras 4 e 5, pode-se perceber que, independente da abordagem utilizada, a estimativa direta apresenta maior variabilidade quando comparada aos métodos indiretos. Também, verifica-se que o método indireto com grau desconhecido (NSUM) tem um comportamento bastante similar ao método indireto com grau conhecido, o que é interessante pois, na prática, não é conhecido o grau de um indivíduo e então pode-se concluir que utilizar estimativas para este funciona tão bem quanto o valor verdadeiro. Além disso, pela Figura 4, nota-se que as distribuições *a posteriori* (abordagem bayesiana) associadas aos métodos indiretos estão mais concentradas em torno da proporção verdadeira ($\theta = 0,2$). A Figura 5 mostra que, ao usar abordagem frequentista, também verifica-se a estimativa do método NSUM próxima do valor verdadeiro de 0,2.

Por fim, para a abordagem bayesiana, utilizou-se técnicas de simulação iterativa baseadas em cadeias de Markov, conforme discutido na Seção 3.3.3, para gerar amostras da distribuição *a posteriori* dos parâmetros relacionados ao modelo. Antes de examinar os resultados obtidos via MCMC, é necessário examinar alguns diagnósticos para

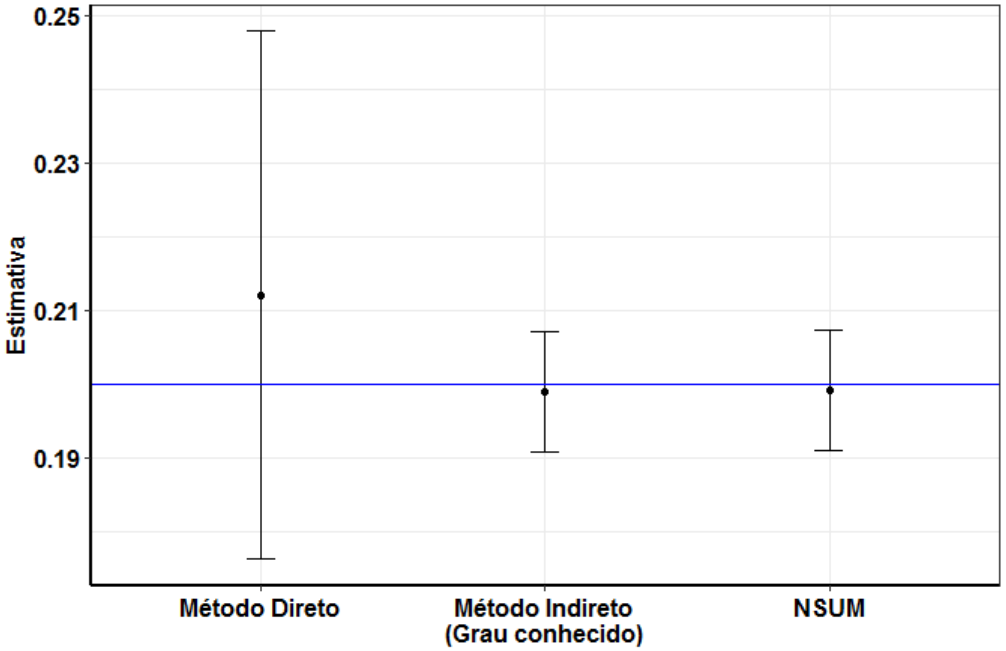
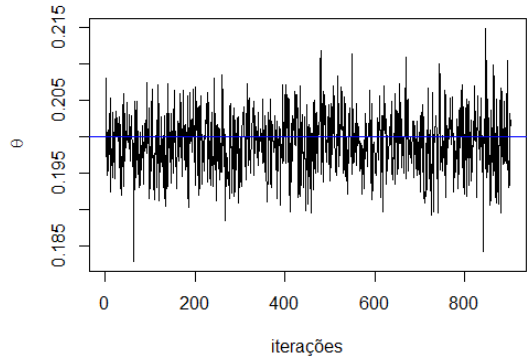


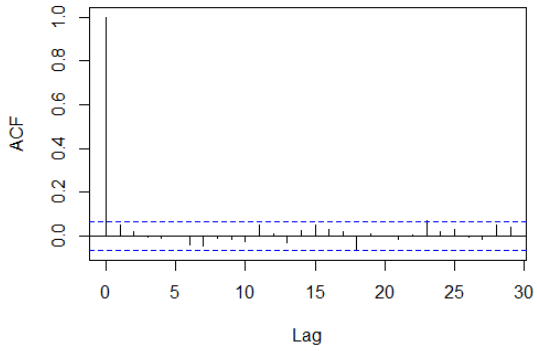
Figura 5: Comparação entre o valor verdadeiro e as estimativas (pontuais e intervalares) obtidas pelos métodos sob abordagem frequentista (linha horizontal: valor verdadeiro).

avaliar se a cadeia de Markov convergiu para sua distribuição estacionária. Usando o algoritmo Amostrador de Gibbs, uma seqüência de 1000 amostras foi gerada para a distribuição *a posteriori* de cada parâmetro, retirando os 100 primeiros valores com *burn-in* (aquecimento). A convergência das cadeias foi verificada através de critérios visuais, inspecionando os traços das cadeias geradas, além do pacote coda do *software* R.

As Figuras 6 e 7 a seguir ilustram as cadeias simuladas e as autocorrelações amostrais obtidas via MCMC para os parâmetros θ e δ_{100}, isto é, o grau do centésimo indivíduo, respectivamente. Pode-se perceber, através dos gráficos de autocorrelação que, em ambos os casos, a cadeia não é altamente correlacionada ao longo das iterações. Além disso, a cadeia de θ convergiu para uma distribuição *a posteriori* tal que a média *a posteriori* resultou em 0,209 e, tal valor, é próximo do verdadeio valor do parâmetro (0,2), representado pela linha contínua na Figura 6(a). No caso do δ específico, aparentemente, a cadeia também convergiu (Figura 7(a)). Além disso, pode-se observar que o grau verdadeiro do centésimo indivíduo foi contemplado na simulação da cadeia deste parâmetro.

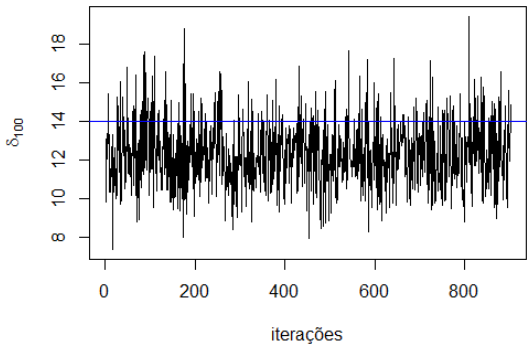


(a) Cadeia simulada via MCMC de θ (linha contínua: valor verdadeiro).

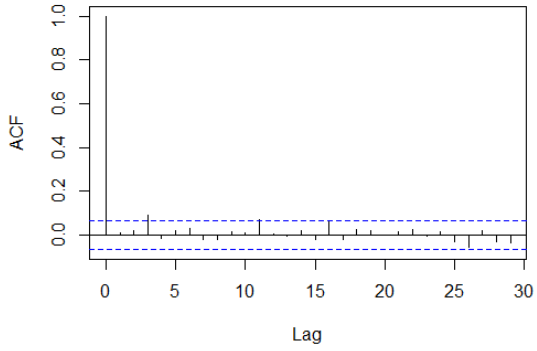


(b) Autocorrelação amostral de θ.

Figura 6: Dados relacionados ao parâmetro θ.



(a) Cadeia simulada via MCMC de δ_{100} (linha contínua: valor verdadeiro).



(b) Autocorrelação amostral de δ_{100}.

Figura 7: Dados relacionados ao grau do centésimo indivíduo.

Na Figura 8 é possível observar o histograma da amostra gerada para δ_{100}, assim como sua média *a posteriori* (linha vertical contínua) e seu intervalo de credibilidade MDP de 95% (linhas verticais tracejadas).

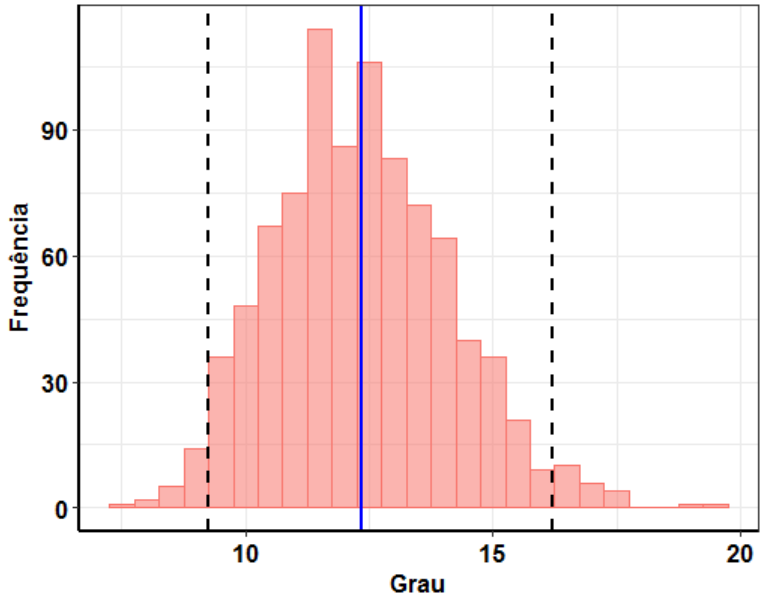


Figura 8: Histograma da amostra final obtida via MCMC para o grau.

Apesar da Figura 8 trazer informações a respeito de um único δ, é possível sumarizar as informações com relação aos graus estimados via MCMC dos 500 indivíduos e compará-los com seus respectivos graus verdadeiros. Com este objetivo, todos os graus foram estimados e para cada cadeia, a média *a posteriori* foi calculada. Como os graus verdadeiros são valores inteiros, justamente por representarem o tamanho da rede de contato de um indivíduo, uma solução encontrada para fazer essa comparação foi utilizar o inteiro mais próximo de cada média *a posteriori* calculada. Veja tal comparação na Figura 9 a seguir. Note que, apesar da Figura 9 mostrar que as frequências dos valores estimados

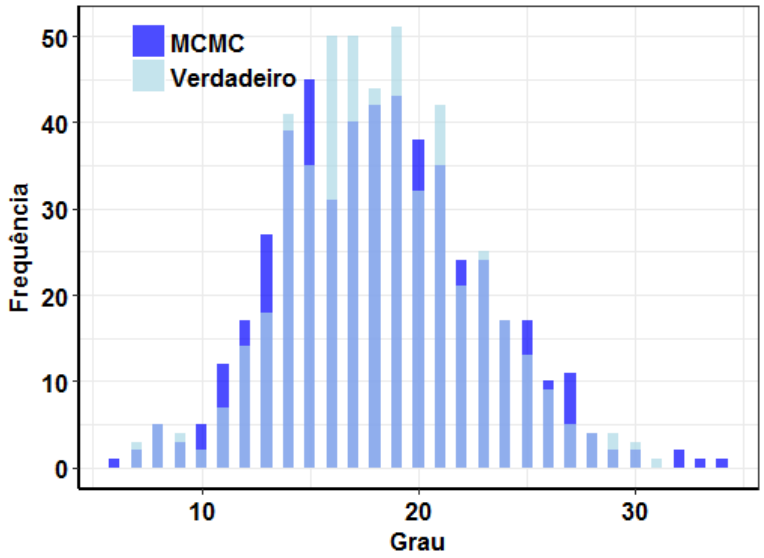


Figura 9: Gráfico de barras das médias *a posteriori* obtidas para os graus via MCMC, sobreposto ao gráfico de barras associado aos graus verdadeiros.

e verdadeiros nem sempre coincidem, a Figura 10 mostra que 96, 6% dos intervalos de credibilidade MDP de 95% estimado para cada um dos graus contém o valor verdadeiro.

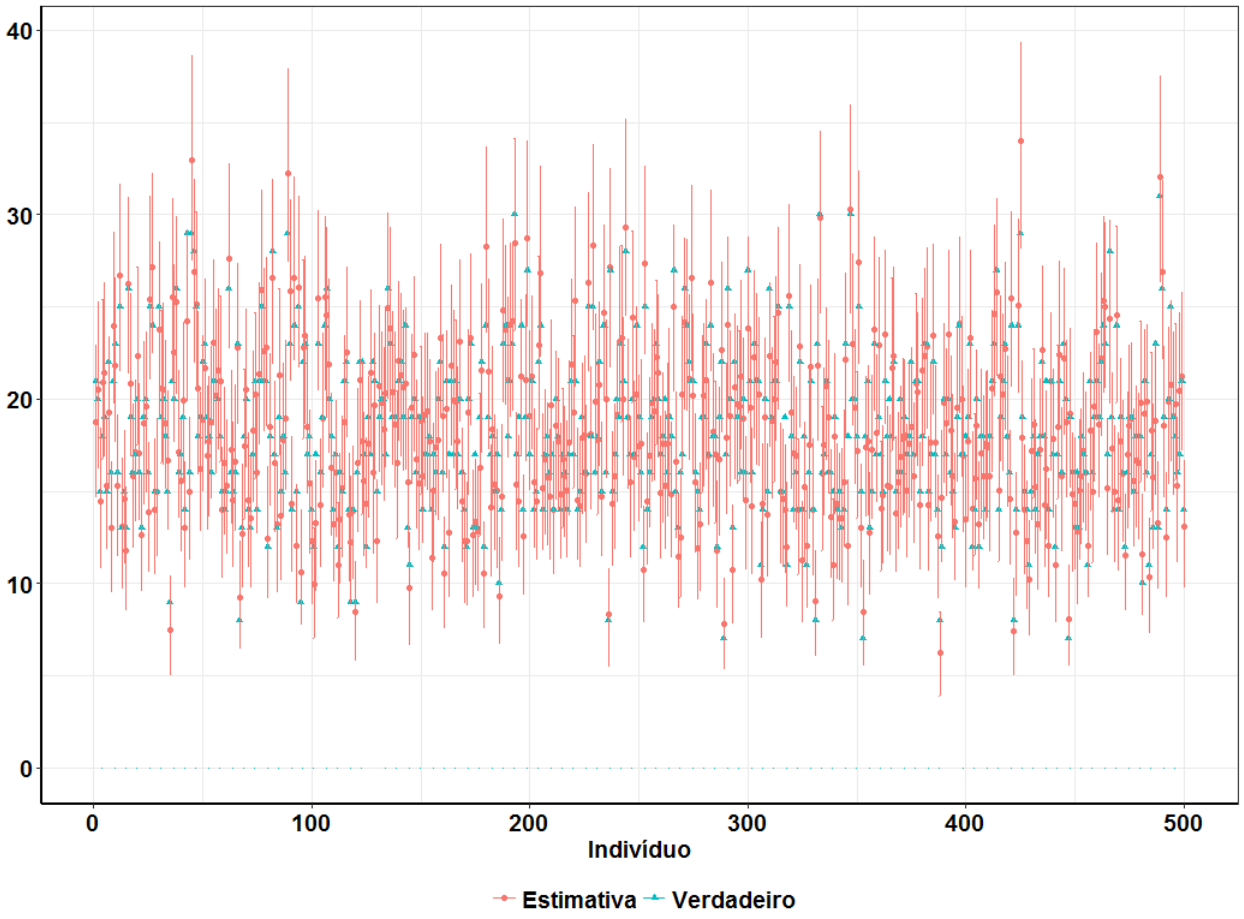


Figura 10: Comparação entre o valor verdadeiro e as estimativas (pontuais e intervalares) obtidas para os graus dos 500 indivíduos da amostra via MCMC.

A Tabela 4 fornece informações sobre a distribuição *a posteriori* de θ via MCMC. Veja

Tabela 4: Resumo da distribuição *a posteriori* de θ obtido via MCMC.

$\hat{\theta}$	$IC_{95\%}(\theta)$
0,1992	[0,1911 , 0,2067]

que, neste caso, as estimativas pontuais são bastante precisas e que o valor verdadeiro de θ está contido no intervalo de credibilidade MDP de 95% estimado. Com isso, pode-se concluir que, apesar das estimativas pontuais dos graus não serem tão precisas (de fato, é possível notar uma certa diferença entre os valores estimados e verdadeiros através da Figura 9), note que as estimativas de θ parecem não ser afetadas, indicando o bom desempenho do método em estimar o tamanho das populações de interesse.

Como a proposta de utilizar a estimação via MCMC é verificar se as informações provenientes dos $Y's$ na estimação dos graus melhora as estimativas do θ (conforme discutido na Seção 3.3.3), a Figura 11 a seguir apresenta os *boxplots* associados às distribuições *a posteriori* via método NSUM e via MCMC. Ressalta-se que, para construir o *boxplot* associado ao método NSUM, foram geradas 901 amostras aleatórias da distribuição *a posteriori* beta associada a θ (conforme definida na Seção 3.3.2). Através

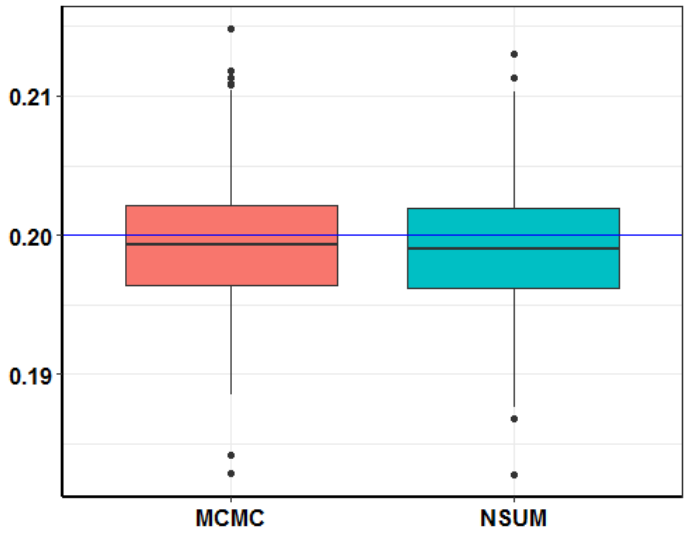


Figura 11: *Boxplots* associados às distribuições *a posteriori* via método MCMC e via NSUM (linha horizontal: valor verdadeiro).

da Figura 11, percebe-se que, o método MCMC teve uma leve melhora na estimativa de θ, no sentido de que sua média *a posteriori* está mais próxima do valor verdadeiro de θ que a estimativa obtida via NSUM. Conclui-se assim que, utilizar também as informações do número de conhecidos do indivíduo na população-alvo para estimar o grau é um fator importante para o problema de estimar uma população de interesse. Apesar da similaridade entre os resultados de θ via NSUM e via MCMC, houve um pequeno ganho em não fazer a estimação bayesiana apenas em duas etapas.

4.2 Dados reais

Uma vez que o estudo simulado demostrou um bom desempenho dos métodos indiretos no que diz respeito à estimação do tamanho de populações de difícil acesso, nesta Seção será apresentada uma aplicação da metodologia aos dados reais. Os dados utilizados nesta aplicação referem-se a um estudo de 2011 realizado em Curitiba, no qual informações de 500 indivíduos residentes neste município foram utilizadas para

estimar o número de usuários de drogas ilícitas nesta população (SALGANIK et al., 2011a). O conjunto de publicações, banco de dados e rotinas de análise referente à pesquisa anterior, realizada em Curitiba, está disponível para *download* gratuito em: http://opr.princeton.edu/archive/NSUM/.

O banco de dados em questão contém informações referentes aos respondentes do inquérito (representadas na Tabela 5), assim como informações relacionadas ao número de conhecidos dos respondentes em cada uma de 20 subpopulações conhecidas. Todas as 20 subpopulações conhecidas utilizadas nesta aplicação estão descritas na Tabela 6. Neste estudo, as variáveis listadas na Tabela 5 foram utilizadas de forma a definir 20 subpopulações, conforme a descrição na Tabela 6.

Tabela 5: Informações referentes aos respondentes do inquérito.

Variável	Categorias/Valores
Idade	Em anos
Estado Civil	Solteiro
	Casado
	Separado
	Divorciado
	União Consensual
Raça/cor	Branca
	Preta
	Amarela
	Parda
Local de Nascimento em Curitiba	Código
Local de Emprego	Código
Gênero	Feminino
	Masculino

Tabela 6: As 20 subpopulações de tamanho conhecido que foram usadas para estimar os tamanhos de rede pessoal dos entrevistados.

Subpopulações	Tamanho absoluto	Tamanho percentual	Fonte
Estudantes do ensino fundamental de escolas públicas	100527	5,5%	MEC[1]
Meninos abaixo de 5 anos	54129	3,0%	DATASUS[2]
Meninas abaixo de 5 anos	51948	2,9%	DATASUS[2]
Mulheres acima de 70 anos	50159	2,8%	DATASUS[2]
Empregados na cidade de Curitiba	37372	2,1%	IBGE/MUNIC[3]
Trabalhador de construção	35056	1,9%	RAIS/CAGED[4]
Homens acima de 70 anos	29768	1,6%	DATASUS[2]
Estudantes de universidade pública	26282	1,4%	MEC[1]
Aposentados por deficiência	26029	1,4%	MPS[5]
Mulheres com pelo menos 20 anos que tiveram filho nos últimos 12 meses	23344	1,3%	IBGE/SIDRA[6]
Estudantes do ensino médio de escolas particulares	17627	1,0%	MEC[1]
Caixa de banco	17056	0,9%	RAIS/CAGED[4]
Estudantes do ensino fundamental de escolas particulares	16461	0,9%	MEC[1]
Morreu nos últimos 12 meses	10310	0,6%	DATASUS[2]
Mulheres casadas nos últimos 12 meses	9960	0,5%	IBGE/MUNIC[3]
Homens casados nos últimos 12 meses	9960	0,5%	IBGE/MUNIC[3]
Motoristas de ônibus	4309	0,2%	RAIS/CAGED[4]
Mulheres com menos que 20 anos que tiveram filho nos últimos 12 meses	3593	0,2%	IBGE/SIDRA[6]
Motoristas de taxi	3252	0,2%	RAIS/CAGED[4]
Hospitalizado por acidente no trânsito nos últimos 12 meses	568	0,03%	DATASUS[2]
Total	527710	29,0%	

[1]Ministério da Educação, Censo Educacional (2009). [2]Departamento de Informática do SUS (2009). [3]Instituto Brasileiro de Geografia e Estatística, Pesquisas de Informações Básicas Municipais (2009). [4]Relação Anual de Informações Sociais, Cadastro Geral de Empregados e Desempregados (2009). [5]Ministério da Previdência Social (2009). [6]Instituto Brasileiro de Geografia e Estatística, Sistema IBGE de Recuperação Automática (2009).

Considerando que o estudo simulado apresentado na Seção 4.1 apontou uma grande similaridade entre os resultados obtidos pelas abordagens frequentista e bayesiana na estimação dos métodos direto e indiretos, justificado pelo uso das distribuições *a priori* não-informativas e a grande quantidade de dados, para os dados reais será adotada apenas a metodologia bayesiana.

Com o intuito de verificar a qualidade dos métodos, estimou-se uma das subpopulações conhecidas, a saber, o número de homens acima de 70 anos, pelos métodos direto e indireto com grau desconhecido e comparou-se o resultado com a proporção verdadeira na população de Curitiba (1,6%, de acordo com a Tabela 6). Lembre-se que para os dados reais não é possível utilizar o método indireto com grau conhecido.

Para ambos os métodos, utilizou-se distribuições *a priori* não-informativas para os parâmetros, conforme discutido no estudo simulado da Seção 4.1.2. A Figura 12 apresenta as estimativas pontuais e intervalares obtidas pelo método direto e indireto (NSUM), assim como o valor verdadeiro para a população em questão.

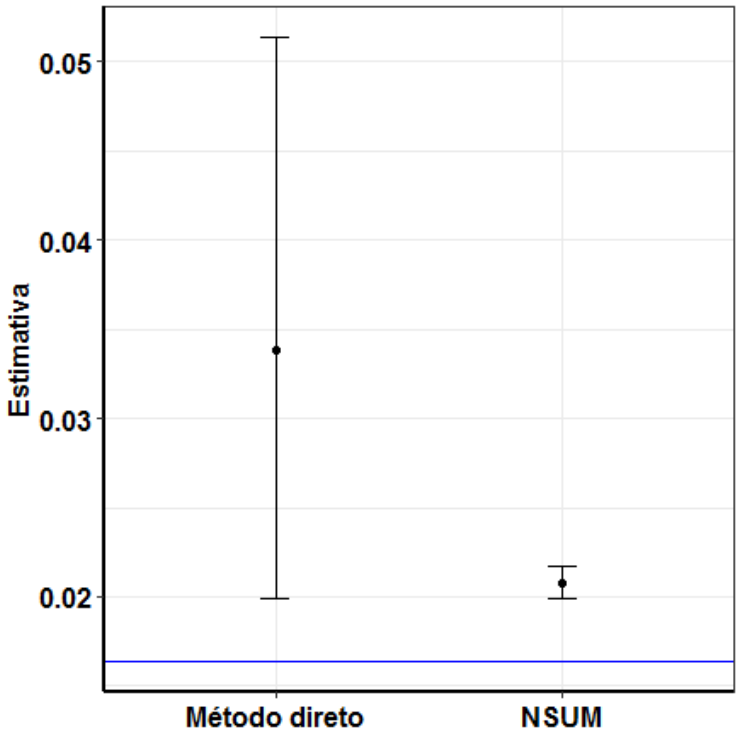


Figura 12: Comparação entre o valor verdadeiro da proporção de homens acima de 70 anos, residentes na cidade de Curitiba e as estimativas (pontuais e intervalares) obtidas pelos métodos (linha horizontal: valor verdadeiro).

A estimativa pontual (média *a posteriori*) obtida via NSUM foi de 2,07%, enquanto a estimativa obtida via método direto foi de 3,38%, ou seja, a última se distancia

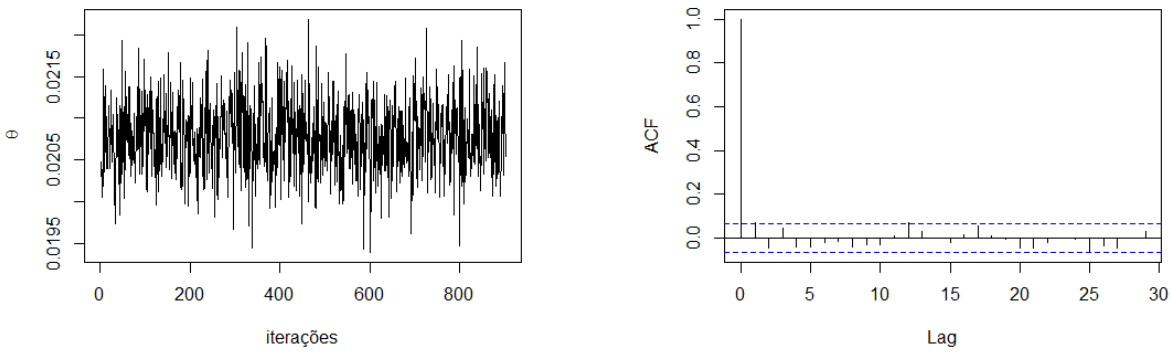
significativamente do valor verdadeiro (de fato, apresenta um valor, aproximadamente, duas vezes maior que o verdadeiro), além de apresentar uma amplitude elevada e consequentemente menor precisão que o método NSUM.

A Tabela 7 sumariza as estimativas referentes aos indivíduos, residentes em Curitiba, que são homens acima de 70 anos. Lembre-se que esta subpopulação corresponde a uma proporção de, aproximadamente, 1,6% da população de referência (que é a população geral de Curitiba), o que representaria cerca de 29768 mil homens acima de 70 anos neste município.

Tabela 7: Estimativas referentes aos homens acima de 70 anos residentes na cidade de Curitiba obtidas via método direto e NSUM.

Método	$\hat{\theta}$	$IC_{95\%}(\theta)$	$\hat{\theta} \times N$	$IC_{95\%}(\theta \times N)$
Direto	0,0338	[0,0183 , 0,0477]	61546,5700	[33286,6400 , 86867,1000]
NSUM	0,0207	[0,0199 , 0,0217]	37784,0400	[36256,4800 , 39510,8400]

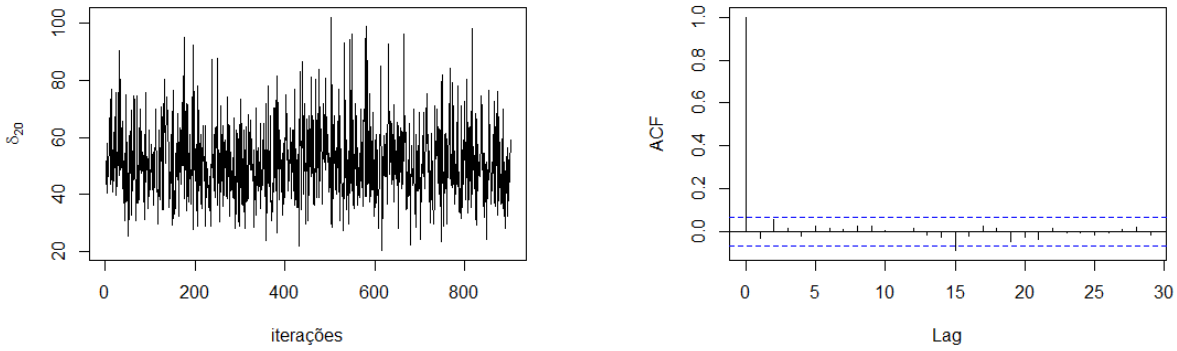
O método via MCMC foi desenvolvido conforme visto no *Toy Example*: usando o Amostrador de Gibbs, uma sequência de 1000 amostras foi gerada da distribuição *a posteriori* dos parâmtros do modelo, retirando os 100 primeiros valores com *burn-in*. Como, novamente, o critério para convergência será visual, as Figuras 13 e 14 apresentam dados relacionados aos parâmetro δ_{20} e θ. Em ambos os casos, as cadeias não são altamente correlacionadas, vide Figuras 13(b) e 14(b), entretanto, apesar delas terem convergido, a cadeia associada ao parâmetro θ não contém a proporção verdadeira do número de homens acima de 70 anos na população de Curitiba (1,6%).



(a) Cadeia simulada via MCMC de θ.

(b) Autocorrelação amostral de θ.

Figura 13: Dados relacionados ao parâmetro θ.



(a) Cadeia simulada via MCMC de δ_{20}.

(b) Autocorrelação amostral de δ_{20}.

Figura 14: Dados relacionados ao grau do vigésimo indivíduo.

Com base nas amostras geradas via MCMC, a Tabela 8 fornece estimativas pontuais (médias *a posteriori*) e intervalares (MDP) relacionadas aos parâmetros do modelo.

Tabela 8: Resumo da distribuição *a posteriori* dos parâmetros.

Parâmetro	Média	$IC_{95\%}$
δ_{20}	51,4779	[29,1000 , 80,9500]
θ	0,0208	[0,0198 , 0,0217]

Comparando as estimativas obtidas para θ na Tabela 8 com as obtidas via método NSUM (Tabela 7), é possível notar que, assim como no estudo simulado, ambos os métodos apresentam comportamentos similares. Entretanto, é nítido que, apesar do método indireto com grau desconhecido, seja via NSUM ou via MCMC, apresentar intervalo de credibilidade tão estreito a ponto de não "alcançar"o valor verdadeiro do parâmetro, suas estimativas pontuais são mais próximas deste quando comparado ao método direto. Para retratar tal informação, a Figura 15 apresenta os *boxplots* associados às distribuições *a posteriori* de θ via métodos direto e indiretos com grau desconhecido (NSUM e MCMC).

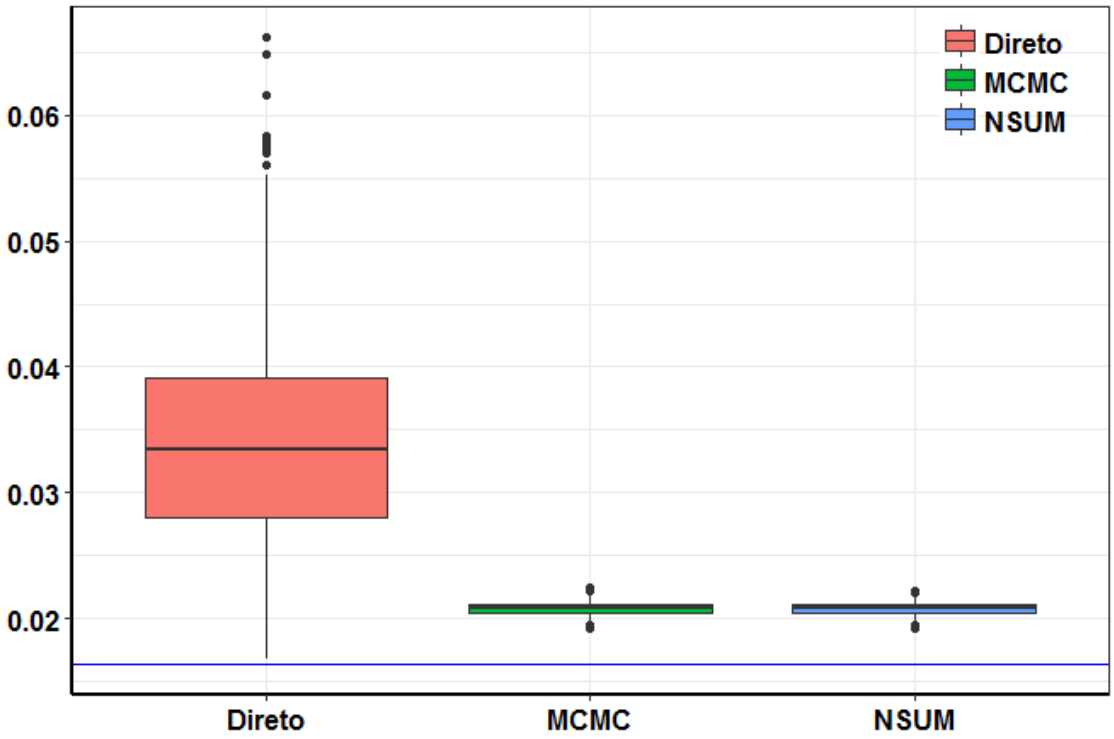


Figura 15: *Boxplots* associados às distribuições *a posteriori* de θ via métodos direto e indiretos com grau desconhecido (NSUM e MCMC) (linha horizontal: valor verdadeiro).

Por meio da Figura 15, fica evidente que os métodos NSUM e MCMC apresentam comportamentos parecidos e embora não contenham o verdadeiro valor da proporção de homens acima de 70 anos residentes na cidade de Curitiba (representado pela linha horizontal), possuem intervalos de credibilidade bastante precisos, diferentemente do método direto (conforme discutido na Figura 12). Destaca-se também que, as estimativas pontuais desses métodos indiretos estão mais próximas do valor verdadeiro quando comparadas ao método direto.

5 Conclusão

Este trabalho avaliou o quanto o método direto e o método indireto, seja com grau conhecido ou com grau desconhecido (NSUM), diferem com relação às estimativas da população-alvo. A metodologia *Network Scale-up* apresentou excelentes estimativas quando comparada ao método tradicional e similaridade em relação ao método MCMC. Além disso, o fato de não haver diferenças significativas entre os métodos indiretos sugere que utilizar o método NSUM é tão eficiente quanto ao método com grau conhecido que só pode ser utilizado em estudos simulados. Também ficou claro que as abordagens frequentista e bayesiana apresentaram resultados similares justificado pelo uso de distribuições *a priori* gama com grande variabilidade para o tamanho da rede contato individual e de uma distribuição *a priori* própria mas fracamente informativa para o parâmetro θ.

Apesar de, em muitas situações, no estudo simulado, o intervalo de confiança ou o de máxima densidade *a posteriori* obtido pelo método direto conter o valor verdadeiro do tamanho da população de interesse, os métodos de estimação indireta são preferíveis pois não é necessário o contato direto com a população estimada, não expondo, desta forma, os comportamentos desses indivíduos, frequentemente estigmatizado, e eventualmente criminalizado. Além disso, é notório que existem situações onde não é possível utilizar a abordagem direta. Neste sentido, pode-se citar, por exemplo, a estimação da quantidade de mortos no terremoto do México (BERNARD et al., 1991). Ainda nesse contexto, vale ressaltar que os intervalos gerados pelos métodos indiretos foram mais precisos em toda a análise.

Este estudo permitiu validar as estimativas produzidas pelo método *Network Scale-up* ao fazer um estudo simulado totalmente controlado, no qual sabia-se exatamente a prevalência de todas as características da população, os graus dos indivíduos e as características das suas redes de contatos e, posteriormente, através de uma aplicação em dado de representatividade nacional para estimar uma subpopulação conhecida através de cadastros disponíveis.

Acredita-se que, no contexto dos métodos indiretos discutidos neste trabalho, a quantidade de populações conhecidas pode influenciar a estimação do tamanho de populações desconhecidas. Supõe-se que não existe um número mínimo nem máximo de populações conhecidas para esta estimação, embora alguns estudos apontem o uso de 20 subpopulações (REIS, 2014). Sendo assim, para estudos futuros seria interessante buscar o modelo mais parcimonioso, implicando em um menor o custo computacional, que estime o tamanho de populações desconhecidas. Acredita-se também que a informação da rede de contatos dos participantes provenientes da população geral quando mal quantificada; por exemplo, quando, influenciada pelos efeito de barreira e viés de visibilidade (MCCORMICK; SALGANIK; ZHENG, 2010), pode afetar a estimação do tamanho da população desconhecida. Dessa forma, seria interessante desenvolver critérios de comparação de modelos hierárquicos bayesianos usando o método *Network Scale-up* para diferentes números de populações conhecidas para estimação do tamanho de populações de difícil acesso. Então, um número ótimo de populações conhecidas necessárias para se fazer a estimação poderia ser determinado.

Apesar de todas as dificuldades em se estimar o grau dos indivíduos nos métodos indiretos, acarretados por exemplo por problemas na falta de exatidão das respostas dos respondentes dos inquéritos, entre outros problemas, é notório que as estimativas dos tamanhos das populações de interesse parecem não ser tão afetadas. Este fato justifica a similaridade entre as estimativas obtidas via método indireto com grau conhecido e desconhecido, além da boa performance do algoritmo MCMC em estimar θ no estudo simulado, apesar das divergências observadas nas estimativas de alguns graus (conforme representado na Figura 9).

A grande vantagem do método NSUM é que sua pesquisa é realizada através da implementação de um inquérito com a população geral (REIS, 2014). Com isso, as perguntas sobre as populações a estimar e as demais questões referentes a este método podem ser inseridos em inquéritos regulares já realizados ou a serem realizados no local, diminuindo custos de pesquisa e permitindo que um mesmo questionário possa ser utilizado para estimar diferentes populações de difícil acesso.

Referências

BERNARD, H. R.; JOHNSEN, E. C.; KILLWORTH, P. D.; ROBINSON, S. Estimating the size of an average personal network and of an event subpopulation: Some empirical results. *Social science research*, Elsevier, v. 20, n. 2, p. 109–121, 1991.

CASELLA, G.; BERGER, R. L. *Statistical inference*. 2th. ed. [S.l.]: Duxbury Pacific Grove, CA, 2002.

COELI, C. M.; VERAS, R. P.; COUTINHO, E. d. S. F. Capture-recapture methodology: an option for surveillance of non-communicable diseases in the elderly. *Cadernos de Saúde Pública*, SciELO Public Health, v. 16, n. 4, p. 1071–1082, 2000.

CSARDI, G.; NEPUSZ, T. The igraph software package for complex network research. *InterJournal*, Complex Systems, p. 1695, 2006. Disponível em: ⟨http://igraph.org⟩.

EHLERS, R. S. Inferência bayesiana. *Departamento de Matemática Aplicada e Estatıstica, ICMC-USP*, 2007.

ERDOS, P. Graph theory and probability. *Canad. J. Math*, v. 11, p. 34G38, 1959.

ERDOS, P.; RÉNYI, A. On the evolution of random graphs. *Mathematical Institute of the Hungarian Academy of Sciences*, v. 5, p. 290–297, 1960.

EZOE, S.; MOROOKA, T.; NODA, T.; SABIN, M. L.; KOIKE, S. Population size estimation of men who have sex with men through the network scale-up method in Japan. *PloS one*, Public Library of Science, v. 7, n. 1, p. e31184, 2012.

JOHNSEN, E. C.; BERNARD, H. R.; KILLWORTH, P. D.; SHELLEY, G. A.; MCCARTY, C. A social network approach to corroborating the number of AIDS/HIV+ victims in the US. *Social Networks*, Elsevier, v. 17, n. 3, p. 167–187, 1995.

JOHNSEN, E. C.; KILLWORTH, P. D.; ROBINSON, S. Estimating the size of an average personal network and of an event subpopulation. In: *The small world*. [S.l.]: Ablex Press, 1989. p. 159–175.

KADUSHIN, C.; KILLWORTH, P. D.; BERNARD, H. R.; BEVERIDGE, A. A. Scale-up methods as applied to estimates of heroin use. *Journal of Drug Issues*, SAGE Publications, v. 36, n. 2, p. 417–440, 2006.

KILLWORTH, P. D.; JOHNSEN, E. C.; MCCARTY, C.; SHELLEY, G. A.; BERNARD, H. R. A social network approach to estimating seroprevalence in the United States. *Social Networks*, Elsevier, v. 20, n. 1, p. 23–50, 1998a.

KILLWORTH, P. D.; MCCARTY, C.; BERNARD, H. R.; SHELLEY, G. A.; JOHNSEN, E. C. Estimation of seroprevalence, rape, and homelessness in the United States using a social network approach. *Evaluation Review*, Sage Publications, v. 22, n. 2, p. 289–308, 1998b.

KUDO, H.; DUNBAR, R. Neocortex size and social network size in primates. *Animal Behaviour*, Elsevier, v. 62, n. 4, p. 711–722, 2001.

MCCARTY, C.; KILLWORTH, P. D.; BERNARD, H. R.; JOHNSEN, E. C.; SHELLEY, G. A. Comparing two methods for estimating network size. *Human organization*, Society for Applied Anthropology, v. 60, n. 1, p. 28–39, 2001.

MCCORMICK, T. H.; SALGANIK, M. J.; ZHENG, T. How many people do you know?: Efficiently estimating personal network size. *Journal of the American Statistical Association*, Taylor & Francis, v. 105, n. 489, p. 59–70, 2010.

MCCORMICK, T. H.; ZHENG, T. Adjusting for recall bias in "How many x's do you know?" surveys. In: *Proceedings of the joint statistical meetings*. [S.l.: s.n.], 2007.

MIGUEL, M. I. R. *Ensino e aprendizagem do modelo Poisson: uma experiência com modelagem*. Tese (Doutorado) — Pontifícia Universidade Católica de São Paulo, 2005.

PLUMMER, M.; BEST, N.; COWLES, K.; VINES, K. Coda: Convergence diagnosis and output analysis for mcmc. *R News*, v. 6, n. 1, p. 7–11, 2006. Disponível em: ⟨http://CRAN.R-project.org/doc/Rnews/⟩.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2016. Disponível em: ⟨https://www.R-project.org/⟩.

REIS, N. B. d. *Quantos usuários de crack e/ou similares existem nas capitais brasileiras? Resultados de um inquérito nacional com a utilização da metodologia Network Scale-Up*. Tese (Doutorado) — Escola Nacional de Saúde Pública, Fundação Oswaldo Cruz, 2014.

SALGANIK, M. J.; FAZITO, D.; BERTONI, N.; ABDO, A. H.; MELLO, M. B.; BASTOS, F. I. Assessing network scale-up estimates for groups most at risk of HIV/AIDS: evidence from a multiple-method study of heavy drug users in curitiba, brazil. *American journal of epidemiology*, Oxford Univ Press, v. 174, n. 10, p. 1190–1196, 2011a.

SNIDERO, S.; MORRA, B.; CORRADETTI, R.; GREGORI, D. Use of the scale-up methods in injury prevention research: An empirical assessment to the case of choking in children. *Social Networks*, Elsevier, v. 29, n. 4, p. 527–538, 2007.

UNAIDS. Estimating the size of populations at risk for HIV. number. *Geneva*, n. 03.36E, 2003.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of 'small-world'networks. *nature*, Nature Publishing Group, v. 393, n. 6684, p. 440–442, 1998.

WHITE, G. C. *Capture-recapture and removal methods for sampling closed populations*. [S.l.]: Los Alamos National Laboratory, 1982.

ANEXO A – Lei dos pequenos números

Teorema A.0.1 ***Lei dos pequenos números***

Sejam $(p_n)_{n\in\mathbb{N}}$ *sequência real com valores em* $(0,1)$ *e* $\lambda > 0$*, tais que* $np_n \to \lambda$

- $p_n \to 0$
- *Dado qualquer* $k \in \{0,1,2,3,\ldots\}$, $\binom{n}{k} p_n^k (1-p_n)^{n-k} I_{\{0,1,2,3,\ldots\}}(k) \to \frac{e^{-\lambda}\lambda^k}{k!}$

Interpretação: Se $X \sim Bin(n,p)$ com n "grande"e p "pequeno", então:

$$X \approx Poisson(np).$$

Concluindo, as probabilidades em um Modelo Binomial coincidem exatamente com aquelas de um Modelo de Poisson em que $np = \lambda$, quando n tende a infinito. Nos dois modelos, as probabilidades são próximas, quando n for grande e p pequeno; em geral, $n \geq 20$ fornece uma aproximação aceitável, desde que $np < 7$ (para o caso em que $n = 20$, equivale a $p < 0,35$; quando $n = 25$, equival e a $p < 0,28$, etc.) (MIGUEL, 2005)

ANEXO B – Amostrador de Gibbs

A amostragem de Gibbs ou amostrador de Gibbs é um algoritmo para gerar uma sequência de amostras da distribuição conjunta de probabilidades de duas ou mais variáveis aleatórias. O propósito de tal sequência é aproximar a distribuição conjunta, ou computar uma integral (tal como um valor esperado)(EHLERS, 2007).

Suponha que $\theta = (\theta_1, \theta_2, \ldots, , \theta_{i-1}, \theta_i, \theta_{i+1}, \ldots \theta_k)'$ é um vetor aleatório com k variáveis aleatórias. A ideia do amostrador de Gibbs é dividir o vetor θ de forma que cada pedaço de θ (θ_i uni ou multidimensional) possa ser gerado de sua distribuição condicional completa $f(\theta_i|\theta_{-i})$, onde $\theta_{-i} = (\theta_1, \theta_2, \ldots, , \theta_{i-1}, \theta_{i+1}, \ldots \theta_k)'$.

O **algoritmo** é dado da seguinte forma:

- Inicialize a cadeia com um valor inicial $\theta^{(0)}$ e inicialize um contador $t = 1$;
- Gere um novo valor $\theta_i^{(t)} \sim f(\theta_i|\theta_{-i}^{t-1})$ onde, $\theta_{-i}^{(t)} = (\theta_1^{(t)}, \theta_2^{(t)}, \ldots, , \theta_{i-1}^{(t)}, \theta_{i+1}^{(t-1)}, \ldots \theta_k^{(t-1)})'$, $i = 1, \ldots, k$.
- Faça $t = t + 1$ e volte ao passo 2 até atingir a convergência.

O ponto crucial para implementar o algoritmo de Gibbs num problema particular é ser capaz de obter as distribuições condicionais completas $f(\theta_i|\theta_{-i})$.