Métodos estatísticos de proteção de dados confidenciais sob a condição de *differential privacy*

Augusto Felix Marcolin

Departamento de Estatística - ICEx - UFMG

Fevereiro de 2018

Métodos estatísticos de proteção de dados confidenciais sob a condição de *differential privacy*

Augusto Felix Marcolin

Orientador: Thais Paiva Galletti

Departamento de Estatística
Instituto de Ciências Exatas
Universidade Federal de Minas Gerais

Belo Horizonte, MG - Brasil
Fevereiro de 2018

Agradecimentos

Primeiramente, agradeço aos meus pais, Milton Marcolin e Adriani Felix, por todo amor e confiança, e acima de tudo, pelo ensinamento de que a educação é a melhor herança.

Agradeço meu irmão pelas discussões(sim, eu sinto falta).

Á minha namorada, pela paciência ao longo desta etapa, aguentando todas minhas reclamações com o coração cheio de carinho e amor. Com certeza, você foi essencial.

Aos meus colegas de pós-graduação, Magno Tairone, Matheus Guerrero, Milton Pifano, Larissa Sayuri e Caio Oliveira pelas boas risadas e troca de conhecimento.

Aos meus amigos irmãos, Luis Gustavo Silva, Douglas Mesquita, Rodrigo Reis e Lucas Godoy por proporcionarem a minha estadia em Minas Gerais mais divertida, além é claro, de compartilharem o gigantesco conhecimento estatístico comigo, sem vocês o caminho seria muito mais difícil.

Á minha orientadora, Thaís Paiva Galletti, pela confiança, compreensão, incentivo e ensinamentos passados.

Ao grupo Stats4Good por mostrar que devemos pensar "fora da caixa" e que podemos melhorar a sociedade usando estatística.

Agradeço aos meus professores da graduação e pós-graduação por todo o conhecimento passado.

Resumo

A quantidade de dados produzidos no mundo digital tem crescido exponencialmente nas últimas décadas. Atentas a este fato, empresas e organizações não tem medido esforços para analisar toda essa gama de informação. Contudo, há um crescimento na preocupação acerca da privacidade da informação das pessoas. Nesse contexto, surge a àrea de *data privacy*, cujo objetivo é garantir anonimização das informações em bases de dados. Tendo em vista o problema exposto, este trabalho apresenta métodos para anonimização de variáveis binárias e categóricas, através de geração de bases sintéticas sob garantia de ϵ-*differential privacy*. Também apresentamos técnicas de inferência para lidar com esse tipo de dado. Inicialmente recriamos e complementamos o estudo de Charest (2011) no âmbito de variáveis binárias anonimizadas. Posteriormente, estendemos o modelo para variáveis de múltiplas categorias. Por fim, aplicamos as técnicas de anonimização e inferenciais em uma base de dados da SUSEP(Superintendência de Seguros Privados) a respeito de roubos de carros e indenizações de seguradoras, para o ano de 2016 na região metropolitana de Belo Horizonte e Zona da Mata. Quanto aos resultados, observamos que há uma perda de informação quando utilizamos a metodologia de bases sintéticas sob garantia de ϵ-*differential privacy*. Porém, utilizando as técnicas apropriadas para fazer inferência podemos obter estimativas precisas.

Palavras-chave: *data privacy*, ϵ-*differential privacy*, dados sintéticos

Abstract

The amount of data produced in digital era has increased in the last decades. Aware of this, companies and organizations have been making all necessary efforts to analyze this amount of information. However, the attention concerning privacy of individuals records is increasing. In this sense, the data privacy area emerges with the goal to guarantee users anonymity in researches. Given that, this work shows anonymization methods for binary and categorical data, using the concept of ϵ-differential privacy synthetic data. We also present inferential techniques to analyze this kind of data. First, we recreate and complement the scenarios proposed by Charest (2011) to binary anonymized data. We then extend the model to categorical variables. Lastly, we apply the anonymization and inferential techniques to a real dataset of car insurance claims in Brazil in 2016 for the metropolian region of Belo Horizonte and Zona da Mata. On the results, we noticed that there is some information loss when the methodology of ϵ-differential privacy synthetic data is applied. However, using the appropriate techniques to make inference can provide accurate estimates.

Keywords: *data privacy*, *ϵ-differential privacy*, synthetic data

Sumário

1 Introdução **8**

2 Metodologia **10**
2.1 *Data Privacy* 10
2.2 Statistical Disclosure Limitation 10
2.3 Dados Sintéticos 12
2.3.1 Geração de Bases Sintéticas 12
2.3.2 Análise via *Combining Rules* 13
2.4 Differential Privacy 14
2.5 Dados sintéticos sob Differential Privacy 16
2.5.1 Sintetizador Multinomial - Dirichlet 18
2.5.2 Sintetizador Binomial - Beta 19
2.6 Análise via Modelo Bayesiano 19
2.6.1 Modelo Beta - Binomial 20
2.6.2 Modelo Dirichlet - Multinomial 24

3 Resultados **28**
3.1 Análises Beta-Binomial 28
3.1.1 Sensibilidade L 28
3.1.2 Comparação entre Modelos 29
3.1.3 Tempo Computacional 30
3.1.4 Análise da distribuição de p 31
3.1.5 Análise de sensibilidade ϵ 33
3.1.6 Análise de sensibilidade n e $\tilde{n}$ 35
3.2 Análises Dirichlet-Multinomial 36
3.2.1 Efeito da variação de n 36
3.2.2 Efeito do Número de Categorias 37
3.2.3 Efeito de Categorias Desbalanceadas 40
3.2.4 Efeito do parâmetro de privacidade 42

4 Aplicação **44**
4.1 Caso 44
4.2 Definições 45
4.3 Inferência Sobre Dados sintéticos 45

5 Discussão **48**

Capítulo 1

Introdução

Atualmente é comum agências de pesquisa divulgarem bancos de dados à pesquisadores, analistas e para o público em geral. Porém, por questões éticas e jurídicas, cabe a estas agências garantir a segurança da informação dos indivíduos presentes na pesquisa. Existem alguns usuários destas bases de dados, chamados *intruders* ou intrusos, que podem por ventura tentar usar estas informações com intenções ilícitas.

Devido aos *intruders*, há uma gama de estudos sendo feitos com intuito de contornar o problema de identificação de indivíduos em bases de dados públicas. Estes estudos vem elaborando técnicas estatísticas de limitação das informações divulgadas, bem como analisando a probabilidade de identificação dos indivíduos.

Essa área de pesquisa é comumente conhecida por *data privacy* e vem sendo cada vez mais difundida na era do *Big Data*. Porém, cabe ressaltar que a preocupação sobre segurança da informação começou bem antes, quando em 1910 o então presidente dos Estados Unidos, Willian Taft, sancionou uma lei de proteção dos dados do censo do país. Tal feito desencadeou muitos outros projetos de regulamentação do censo objetivando a segurança da informação da população. Infelizmente no Brasil, tais medidas parecem distante da realidade.

Para evidenciar a dimensão do problema, temos como exemplo um caso que ficou famoso em 2007, no qual a empresa de filmes e séries via *streaming*, Netflix, ofereceu US$1 milhão por uma melhora em 10% no seu algoritmo de recomendação. Na época, a empresa divulgou uma base de dados contendo avaliações de filmes de milhares de usuários para que os competidores elaborassem seus algoritmos. Foram removidas apenas algumas identificações pessoais dos usuários, como nome, endereço e email. Tais precauções tomadas quanto à proteção da privacidade de seus clientes não foram suficientes. O Netflix não é o único portal de avaliação de filmes na internet, existem outros como o IMDb, um portal online onde usuários podem dar notas aos filmes que assistem. Narayanan e Shmatikov (2006) associaram a base de dados divulgada pelo Netflix com a do IMDb, e então conseguiram identificar alguns indivíduos na base, comprometendo assim a privacidade dos usuários.

Casos como o do Netflix são comuns atualmente. A revista Bloomberg publicou um artigo em 2013, "*States' hospital data for sale puts privacy in jeopardy*", descrevendo alguns casos em que a divulgação de bases de dados médicos estaduais nos EUA prejudicou certos indivíduos em contratações e acesso a planos de saúde. Por conta da má reputação

que isso pode gerar a uma empresa ou estado, muitas corporações estão atentas ao assunto de privacidade. Outro exemplo é a gigante da tecnologia Apple, que no lançamento de seu novo sistema operacional iOS 10, em setembro de 2016, deu enfâse à nova política de segurança da informação dos usuários com o *slogan*: *"Apple will not see your data"*. Obviamente a empresa não desprezará seus preciosos dados na elaboração de sistemas de recomendação, algoritmos de aprendizado de máquina e etc. Porém, a mensagem que a empresa passou para seus usuários é: "Seus dados estarão mascarados, poderemos apenas inferir comportamentos coletivos, não individuais".

Além dos casos citados, existem também situações onde as agências desejam liberar dados com informações sensitíveis aos analistas na forma de tabelas de contingência. Suponha então que o Ministério da Saúde está disposto a liberar dados para a população sobre casos de HIV nas cidades brasileiras, divididos por sexo e idade. Imagine que nessa situação hipotética a cidade seja pequena e hajam baixas contagens em algumas categorias, tal fato torna a informação usuários sensitíveis, impossibilitando assim a divulgação da base de dados.

Considerando tal situação, apresentamos nesse trabalho métodos de anonimização de usuários a liberação da informação seja de forma segura. /nosso objetivo é propor métodos inferências para dados dicotômicos e multicategóricos anonimizados. Ao longo do trabalho apresentaremos métodos de anonimização de bases de dados nas Seções 2.2 e 2.5, garantias ao usuário quanto à segurança de suas informações na Seção 2.4, bem como propostas inferenciais para analisarmos tais dados anonimizados na Seção 2.6. Avaliaremos os resultados no Capítulo 3 através de estudo de simulação. Por fim, no Capítulo 4, aplicamos nossa técnica aos dados da SUSEP.

Capítulo 2

Metodologia

2.1 *Data Privacy*

A área de *data privacy* ganhou notoriedade no inicio dos anos 90, quando os estudos a respeito do risco de identificação de usuários desenvolvidos por Duncan e Lambert (1986) e Lambert (1993) foram publicados. Os autores propuseram um cenário em que as agências deveriam modelar o comportamento dos *intruders*, com objetivo de estimar a probabilidade de identificação das unidades amostrais e o potencial prejuízo causado por tal, utilizando uma abordagem Bayesiana. De lá pra cá, houveram diversos trabalhos abordando o problema de identificação de usuário, tais como, Reiter (2005) propondo novas maneiras de estimar o risco de identificação e a aplicação de El Emam et al. (2009) em que avalia o risco de re-identificação de pacientes de um hospital através das prescrições médicas "vendidas" pelo hospital à empresas privadas.

Além de medir o risco de identificação, outro fator importante é elaborar técnicas capazes de diminuir esse risco, conhecidas como *Statistical Disclosure Limitation*(SDL). Tais técnicas tratam do processo de anonimização dos indivíduos de uma base de dados. Atualmente, existem inúmeros artifícios para mascarar os dados dos usuários; abordaremos o assunto na Seção 2.2, a seguir.

2.2 Statistical Disclosure Limitation

Considerando que as agências estão dispostas à divulgar suas bases de dados ao público, Karr e Reiter (2014) dizem que existem cerca de três procedimentos para avaliar o risco de identificação na base de dados antes dessa se tornar pública. Primeiramente, após a remoção dos identificadores diretos dos indivíduos, tais como, endereço e nome, as agências avaliam o risco de *linkage* da base de dados, que é a conexão entre as informações disponíveis com o usuário detentor de tais características. Comumente, o risco de tal *linkage* é elevado, tendo as agências que restringir o acesso aos dados ou utilizar alguma técnica de limitação de informação, em inglês *Statistical Disclosure Limitation*(SDL).

O segundo procedimento é justamente a aplicação de alguma técnica SDL com o objetivo de minimizar os riscos de identificação, como iremos descrever a seguir. O terceiro e último procedimento é a avaliação do risco na liberação da base e sua utilidade

inferencial. Segundo Reiter (2012), nestas avaliações as agências objetivam determinar se os riscos são suficientemente pequenos e se a utilidade é adequadamente alta, para justificar liberar os dados ao público.

O termo *Statistical Disclosure Limitation*(SDL) se refere a técnicas importantes na proteção da base de dados contra ataques de *intruders*, sendo responsáveis pela preservação da confidencialidade dos indivíduos. Existem inúmeras técnicas SDL, a seguir listamos as mais conhecidas e utilizadas, bem como suas aplicações e limitações.

Agregação

Esse método torna indivíduos atípicos, aqueles com alto risco de identificação, em indivíduos que possuem características comuns a outros membros na base de dados, através de uma agregação em classes maiores. Por exemplo, há uma pessoa com características incomuns em determinado bairro, porém na cidade podem haver muitas outras pessoas com tais características. Logo, liberar a informação ao nível de bairro pode trazer um alto risco de identificação, porém ao nível de município não.

Em contrapartida, esta agregação da informação faz com que percamos o poder de uma análise mais profunda, prejudicando a inferência.

Supressão

Nesta técnica, as agências simplesmente deletam valores de indivíduos com alto risco de identificação, ou até mesmo variáveis inteiras da base a ser divulgada (Cox, 1980). Ao suprimir informações, temos o problema de trazer dados faltantes não aleatórios para nossa base, o que acaba por viesar a amostra.

Data Swapping

Proposta por Dalenius e Reiss (1982), a técnica de *data swapping* trata de transformar dados através da troca de valores entre pares de indivíduos. Admita que em nossa base de dados tenhamos as variáveis bairro e número de residentes na moradia, para cada indivíduo. Agora suponha que há dois sujeitos (I e II) morando em bairros distintos, porém em suas residências vivem o mesmo número de pessoas. Se o indivíduo I está sob risco de *linkage*, troca-se o valor da variável "bairro" do sujeito I pela do sujeito II e vice-versa.

Esta técnica desencoraja os *intruders* de tentar identificar os indivíduos, porém danifica a estrutura de correlação presente entre as variáveis na base de dados.

Adicionar Ruído Aleatório

Autores como Tendick (1991) e Sullivan (1989) propõem métodos com intuito de mascarar os dados, porém não desconstruir a aleatoriedade na amostra. A proposta é justamente adicionar um ruído aos dados seguindo uma distribuição de probabilidade.

Por exemplo, adicionar um ruído com distribuição $Normal(0, \sigma^2)$ a uma variável contínua. A limitação do método é que, geralmente, a proteção da base aumenta conforme há um aumento na variância do ruído, o que causa uma distorção na distribuição dos dados.

2.3 Dados Sintéticos

As técnicas SDL apresentadas até aqui tratam de manipular a base original com intuito de garantir a anonimização dos indivíduos. Porém, causam distorção nas relações entre variáveis, além de viesar as observações. Desta maneira, há necessidade de um método que gere uma base de dados capaz de conter características e propriedades similares à base original, sem que seja necessário manipular esta. Nesse contexto, surgem os chamados dados sintéticos, método alternativo para lidar com tais problemas.

O conceito do método está vinculado à suposição de que há indivíduos na base de dados que tenham características muito específicas suscetíveis à *linkage*. Deste modo, é de interesse liberar ao público uma base que traga informações do comportamento geral da amostra, porém não comprometa nenhuma informação específica dos indivíduos. Neste caminho, os dados sintéticos podem ser vistos como uma "caricatura" da nossa base original, ou seja, nossa amostra é mascarada de tal forma que temos as características gerais dela, porém não podemos inferir nada ao nível do indivíduo.

Introduzido primeiramente por Rubin (1993), os dados sintéticos se popularizaram nos últimos anos devido à maior preocupação acerca da privacidade dos indivíduos na divulgação de microdados. Raghunathan et al. (2003) definem como criar e fazer inferência em bases de dados sintéticas. Tais definições serão explicitadas a seguir.

2.3.1 Geração de Bases Sintéticas

O conceito de dados sintéticos surgiu no contexto de imputação. O processo de imputação consiste em preencher valores para dados faltantes com objetivo de obter uma base de dados completa. Este preenchimento pode ser feito de diversas formas, desde algum método determinístico ou até mesmo via regressão. Contudo, estas técnicas se mostram falhas em casos multivariados, pois quando analisamos a base completa, não há variabilidade na parte imputada. Portanto, uma alternativa é utilizar os modelos de imputação múltipla proposto por Rubin (1987).

A geração dos valores imputados é feita da seguinte forma: considere a matriz inteira de dados $\boldsymbol{Y} = (\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis})$, onde $\boldsymbol{Y}_{obs}$ são os indivíduos observados e $\boldsymbol{Y}_{mis}$ os não observados. Após as especificações da verossimilhança $f(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}|\boldsymbol{\theta})$ e de uma distribuição a priori para os parâmetros $\boldsymbol{\theta}$, onde $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)$, a imputação é efetuada através da preditiva a posteriori via *Gibbs sampling*.

Na iteração i, amostramos:

$$\boldsymbol{y}_{mis}^{(i)} \sim f(\boldsymbol{y}_{mis}|\boldsymbol{y}_{obs}, \boldsymbol{\theta}^{(i-1)})$$
$$\boldsymbol{\theta}^{(i)} \sim f(\boldsymbol{\theta}|\boldsymbol{y}_{mis}^{(i)}, \boldsymbol{y}_{obs})$$

Após a convergência, seleciona-se aleatoriamente m amostras espalhadas para garantir a independência, obtendo assim uma base de dados com imputação múltipla.

Observe que podemos extender a semântica de imputação de dados para a ótica de dados sintéticos. A ideia geral do método é exatamente igual, o diferencial é o objetivo final. No âmbito de imputação, queremos gerar valores faltantes, enquanto ao gerar dados sintéticos queremos gerar valores "mascarados" de nossos dados. Para isso, basta utilizar a teoria de imputação, onde os dados faltantes são vistos como a base sintética a ser gerada.

Na Figura 2.1 temos um diagrama dando a intuição do método para geração de bases sintéticas. A caracterização do modelo Bayesiano é dada pelo conhecimento a priori representado por $f(\boldsymbol{\theta})$ e a verossimilhança por $\boldsymbol{X}$. A junção dessas informações resulta na distribuição a posteriori $f(\boldsymbol{\theta}|\boldsymbol{X})$. Nossa base sintética será então gerada através da distribuição preditiva a posteriori $f(x|\boldsymbol{\theta X})$ do modelo.

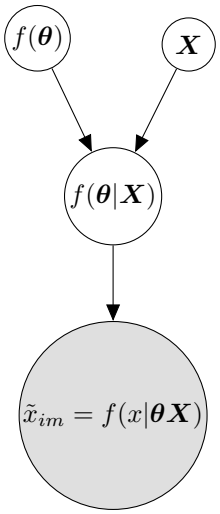

Figura 2.1: Geração de uma base sintética.

2.3.2 Análise via *Combining Rules*

Definido o conceito e estrutura de geração de dados sintéticos, é importante determinar como fazer inferência sobre esses dados. Nesse contexto, Rubin (1987) publicou as chamadas *combining rules*, técnicas para fazer inferência em bases de dados geradas via imputação múltipla.

Suponha que geremos M bases de dados sintéticas D_m, $m = 1, 2, ..., M$ e queremos estimar um parâmetro da população denotado por Q. Seja q_i o estimador pontual de Q e v_i o estimador da variância de q_i. Agora definimos:

$$\bar{q}_m = \sum_{i=1}^{m} \frac{q_i}{m}$$
$$b_m = \sum_{i=1}^{m} \frac{(q_i - \bar{q}_m)^2}{m-1}$$
$$\bar{v}_m = \sum_{i=1}^{m} \frac{v_i}{m}.$$

O estimador de Q é $\bar{q}_m$ e a variância deste é dada por:

$$T_m = \Big(1 + \frac{1}{m}\Big) b_m + \bar{v}_m.$$

Repare que este estimador leva em conta tanto a variância inerente aos dados, quanto a variância vinda da geração de múltiplas bases sintéticas. Extensões dessas regras para métodos multivariados podem ser encontradas em Li et al. (1991) e Raghunathan et al. (2001).

2.4 Differential Privacy

Foram apresentadas até aqui diversas formas de anonimização de uma base de dados via técnicas SDL, mostramos suas características e limitações. Porém, nenhuma dessas técnicas garante ao usuário, de forma clara e específica, uma medida avaliando quanto seus dados estão protegidos. Nesse contexto, Dwork (2008) introduz o *differential privacy* (DP), uma condição probabilística com objetivo de garantir a confidencialidade de um método de anonimização. De acordo com Dwork (2008), o DP garante a um indivíduo que ele não será afetado caso seus dados sejam utilizados em qualquer pesquisa ou estudo, independente da informação extra ou conjunto de dados que um *intruder* possuir.

A definição formal do método pode ser escrita como:

Definição 1. *Uma função aleatória de anonimização r com imagem $\boldsymbol{S}$ satisfaz a condição ϵ-differential privacy, se para todos os bancos de dados D e $D^{'}$ e todo subconjunto $S \in \boldsymbol{S}$:*

$$p(r(D) \in S) \leq \exp(\epsilon) \times p(r(D^{'}) \in S), \tag{2.1}$$

ou equivalente:

$$\frac{p(r(D) \in S)}{p(r(D^{'}) \in S)} \leq \exp(\epsilon), \tag{2.2}$$

onde D e $D^{'}$ se diferem em no máximo uma observação e $\epsilon > 0$.

Note que ϵ atua como um parâmetro que mede o nível de risco. Quanto menor o ϵ, maior será a confidencialidade dos indivíduos. Porém, não há uma relação direta entre ϵ e o risco de identificação, apenas uma associação forte. Também, não há um ϵ ótimo, este depende do problema que está sendo estudado.

A condição DP representa uma forte garantia de confidencialidade, pois mesmo que um *intruder* tenha informação sobre todos os indivíduos da base com exceção de um, ele não será capaz de aprender a respeito do elemento desconhecido.

Existem diversas maneiras de alcançar a condição de ϵ-*differential privacy*, muitas delas tratam de adicionar um ruído aleatório aos dados. Tal quantidade de ruído é calibrada através da *sensibilidade* de uma função.

Definição 2. *(Sensibilidade) Para qualquer função $f : D \to \mathbb{R}^d$, onde d é a dimensão de D, a sensibilidade de f é:*

$$\Delta f = \max_{D,D'} ||f(D) - f(D')|| \tag{2.3}$$

$\forall$ *D e D' que se diferem em no máximo uma observação.*

A sensibilidade mede o quanto a base anonimada se difere da real. Para ficar um pouco mais claro, vemos um exemplo abaixo:

Exemplo 1. *Suponha uma simples base de dados, onde temos o emprego dos indivíduos, idade, e um variável binária indicando se a pessoa tem alguma doença.*

Tabela 2.1: Base de dados

Emprego	**Idade**	**Classe**
Estatístico	34	S
Manicure	50	N
Economista	43	S
Barbeiro	26	N
Advogado	29	S
Manicure	31	N
Cientista da Computação	30	S
Estatístico	27	S

Considere a função f sendo a contagem do número de pessoas com menos de 45 anos. A sensibilidade Δf é 1, pois $f(D)$ só pode se diferenciar no máximo em uma contagem, caso haja um acréscimo ou remoção de um indivíduo.

Mecanismo de Laplace: Dwork et al. (2006) propõem um mecanismo para gerar dados sob a condição de ϵ-DP, baseado no conceito de sensibilidade. O método conhecido por Mecanismo de Laplace, trata de adicionar um ruído aleatório aos dados de uma distribuição Laplace. Uma variável aleatória segue uma distribuição Laplace(μ, λ) se sua função densidade de probabilidade é definida como:

$$P(x|\lambda) = \frac{1}{2\lambda} e^{(\frac{-|x-\mu|}{\lambda})} \tag{2.4}$$

com esperança igual a 0 e variância $2\lambda^2$.

Considere uma base de dados D e uma função f a ser aplicada a essa base. A versão anonimizada da base D' é obtida: primeiro, aplicando-se a função à base original $f(D)$, e depois adicionando o ruído da distribuição Laplace. Ou seja,

$$f(D') = f(D) + Lap(\lambda) \tag{2.5}$$

A ligação do método com a definição de sensibilidade é em relação à garantia de que os dados perturbados satisfaçam a condição de ϵ-*differential privacy*. Tal feito é sustentado pelo Teorema 2.4.1, encontrado no trabalho de Dwork et al. (2006).

Teorema 2.4.1. *Para qualquer função $f : D \to \mathbb{R}^d$, o algoritmo que adiciona, independentemente, um ruído aleatório de uma distribuição $Lap(\Delta f/\epsilon)$ para cada saída de D, satisfaz a condição de ϵ-differential privacy.*

Seguindo com o caso do Exemplo 1, para liberar uma base garantindo o ϵ-*differential privacy*, adicionando um ruído utilizando o mechanismo de Laplace, primeiramente computamos as verdadeiras contagens $f(D)$ e a sensibilidade ($\Delta f = 1$). O *output* gerado para o público será uma base de dados da forma: $f(D) + Lap(\frac{1}{\epsilon})$.

O mecanismo de Laplace é o algoritmo mais conhecido para alcançar a condição de ϵ-*differential privacy*. Porém pertubar os dados de forma direta, adicionando um ruído, causa distorções na distribuição da variável que podem levar pesquisadores a inferências viesadas. Outro fator negativo é quando temos a situação do exemplo, onde há uma variável discreta(contagens) e adicionamos um ruído com distribuição contínua, descaracterizando a natureza da variável.

Contudo, devido à condição fortíssima de privacidade e à garantia que pode ser dada aos usuários/clientes, o *differential privacy* tem ganhado notoriedade no campo de *data privacy*. As abordagens e métodos propostos para gerar bases de dados sob tal condição vem de diversas vertentes, como por exemplo em *machine learning*(ML) com o trabalho de (Mohammed et al., 2011), onde o autor propõe um algoritmo de anonimização, que além de atingir a condição de ϵ-DP, possui uma boa acurácia para classificadores. No mesmo contexto de ML, temos trabalhos voltados para sistemas de recomendação (Friedman et al., 2016), *deep learning* (Abadi et al., 2016) e PCA (Chaudhuri et al., 2012).

Este trabalho apresenta técnicas de criação de dados sintéticos sob a condição de ϵ-*differential privacy* para respostas binárias e multicategóricas. Também propomos métodos inferenciais para lidar com tais bases. A seguir, no Capítulo 2, apresentaremos a metodologia do trabalho, posteriormente resultados e avaliações das técnicas propostas, e por último uma aplicação.

2.5 Dados sintéticos sob Differential Privacy

O processo de geração de dados sintéticos sob a condição de differential privacy é feito através dos chamados sintetizadores ou DIPS(**DI**fferential **P**rivacy Data **S**ynthesis). Um sintetizador é uma função ou algoritmo que transforma a base original, sob risco de segurança, em uma base sintética capaz de ser divulgada ao público sem que haja riscos aos indivíduos nela contidos, garantido pela condição de *differential privacy*.

Segundo Bowen e Liu (2016), as DIPS podem ser dividas em duas classes: as técnicas paraméticas(P-DIPS) e não paramétricas(NP-DIPS). Os algoritmos da abordagem de

NP-DIPS baseam-se na distribuição empírica dos dados para a construção dos sintetizadores, quanto aos P-DIPS, os sintetizadores são implementados considerando distribuições paramétricas ou modelos paramétricos apropriados para os dados. Neste mesmo trabalho, o autor compara diferentes métodos de geração de DIPS, avaliando os prós e contras de cada proposta. No estudo comparativo de Bowen e Liu (2016) o autor cita diversas DIPS em diferentes contextos, como em dados contínuos, contagens e categóricos. Nessa conjuntura, iremos concentrar nosso estudo em DIPS para dados categóricos, no âmbito paramétrico.

Para esclarecer a problemática, vamos ilustrar o exemplo dado no Capítulo 1 de introdução. Suponha o caso em que o ministério da saúde esteja disposto a liberar dados para a população sobre os casos de HIV nas cidades brasileiras, divididas por sexo e idade. Hopotéticamente temos os dados expostos na Tabela 2.2:

Tabela 2.2: Número de casos de HIV para Pedro Osório

	Sexo		
Idade	Masc.	Fem.	Total
$18-24$	5	2	7
$25-40$	13	12	25
$41-55$	10	7	17
$55+$	21	26	47
Total	49	47	96

Observe que nas categorias onde existem poucas contagens há um risco alto em liberar a informação, devido a grande chance de *linkage*. Por exemplo, se um *intruder* desconfia que uma pessoa do sexo feminino que tem entre 18 e 24 anos, está com AIDS, e ainda, ele tem a informação extra de que já existe um caso, o usuário "mal-intencionado" pode levantar suas hipóteses. Portanto, antes do governo tornar público os dados é aconselhável utilizar alguma DIPS para garantiar a segurança dos usuários.

Neste cenário, existem alguns trabalhos abordando soluções para o problema, tais como: o estudo de Abowd e Vilhuber (2008) que propõe uma DIPS paramétrica onde a ideia é amostrar dados de tabelas de frequência a partir da distribuição preditiva a posteriori das contagens. Posteriormente, Charest (2011) avaliou a qualidade inferencial em bases de dados sintéticas binárias. Há ainda o trabalho de McClure e Reiter (2012) apresentando uma técnica similar para dados de contagens com diferença na especificação da priori.

Nosso trabalho propõe-se a revisar e analisar com maior profundidade os resultados obtidos por Charest (2011), bem como extender o modelo proposto para dados onde há múltiplas categorias. A seguir, apresentaremos os sintezadores abordados em Abowd e Vilhuber (2008), o método inferencial proposto por Charest (2011) e ainda nossas propostas.

2.5.1 Sintetizador Multinomial - Dirichlet

O sintetizador Multinomial - Dirichlet, proposto por Abowd e Vilhuber (2008), é desenvolvido para gerar bases sintéticas sob a condição de ϵ-*differential privacy* para dados de contagens em duas ou mais categorias.

O algoritmo foi elaborado para gerar bases sintéticas baseada na distribuição preditiva a posteriori, como visto na Seção 2.3, porém com uma adaptação nos parâmetros para garantir a condição de ϵ-DP.

Dado uma base de dados $X = (x_1, ..., x_k)$, onde cada x_i é não negativo e inteiro, e k é o número de categorias, tal que $\sum_1^k x_k = n$, assumimos que X segue uma distribuição multinomial com probabilidades $(p_1, ..., p_k)$, onde $p_i \in (0, 1) \forall i$ e $\sum p_i = 1$. A função densidade de probabilidade da multinomial pode ser definada como:

$$f(x, k) = \frac{n!}{x_1!, ..., x_k!} \prod_{i=1}^{k} p_i^{x_i}. \tag{2.6}$$

Dada tal conjuntura, temos uma base sintética $\tilde{X}$ sob ϵ-differential privacy para dados com duas ou mais categorias de tamanho $\tilde{n}$ seguindo o algoritmo proposto por Abowd e Vilhuber (2008):

Algoritmo 1: Sintetizador Multinomial

---

1 Seja $\alpha = (\alpha_1, ..., \alpha_k)$, a ser definido de acordo com a Definição 3 a seguir.

2 **Amostre** os parâmetros para a distribuição a posteriori seguindo:

$$\tilde{p} \propto Dirichlet(\alpha + X)$$

3 **Amostre** uma base sintética seguindo:

$$\tilde{X} \propto Multinomial(\tilde{n}, \tilde{p})$$

A adaptação necessária nos parâmetros para que a base sintética atinja a condição de ϵ-DP é dada através da Definição 3:

Definição 3. *O Sintetizador Multinomial - Dirichlet atinge ϵ-DP se e somente se*

$$\alpha_i \leq \frac{\tilde{n}}{\exp(\epsilon - 1)}, i = 1, \ldots, k.$$

Comumente, o parâmetro α é escolhido como o máximo valor que satisfaz a condição, sendo a razão mostrada um *upperbound* para o parâmetro. Além disso, é usual utilizar todos valores de α_i iguais.

A prova pode ser encontrada em Abowd e Vilhuber (2008). Observe que o parâmetro α depende do nível de confidencialidade ϵ e o tamanho da amostra sintética. O α é o parâmetro responsável por distorcer nossa base sintética a ponto de atingir a condição de ϵ-DP.

2.5.2 Sintetizador Binomial - Beta

Um caso específico do sintetizador Multinomial-Dirichlet é quando temos apenas duas categorias nos dados. O DIPS para esse caso é chamado de Binomial-Beta.

Além do modelo para múltiplas categorias, Abowd e Vilhuber (2008) introduziram um algoritmo para geração de dados sintéticos sob ϵ-*differential privacy* quando a variável em questão é dicôtomica.

Consideremos $X = (x_1, ..., x_n)$ onde $x_i \in 0, 1$ para $i = 1, ..., n$ variáveis dicotômicas. Assumindo verossimilhança binomial dos nossos dados, podemos resumir a informação da nossa amostra através da estatística suficiente minimal $x = \sum_{i=1}^{n} x_i$. Para garantir a proteção de nossa base, devemos publicar uma base de dados sintética $\tilde{x}$ que satisfaça a condição de ϵ-*differential privacy*. O algoritmo proposto é exposto a seguir:

Algoritmo 2: Sintetizador Binomial

---

1 Seja $\alpha = (\alpha_1, \alpha_2)$, de acordo com a Definição 3

2 **Amostre** os parâmetros para a distribuição a posteriori seguindo:

$$\tilde{p} \propto Beta(\alpha_1 + x, \alpha_2 + n - x)$$

3 **Amostre** uma base sintética seguindo:

$$\tilde{X} \propto Binomial(\tilde{n}, \tilde{p})$$

α_1 e α_2 são parâmetros de controle da privacidade da base de dados, $\tilde{p}$ e $\tilde{n}$ são o parâmetro de probabilidade de sucesso e o tamanho amostral, respectivamente, da base de dados sintética $\tilde{x}$.

Observe que o método permite gerar base de dados sintéticos de tamanho $\tilde{n}$ diferente da base original. Caso queiramos liberar múltiplas bases sintéticas, basta replicar o processo M vezes. É importante ressaltar que quando estamos considerando ϵ-*differential privacy* aliada aos dados sintéticos, ao liberar M bases de dados sintéticos, juntas estas tem de garantir o ϵ-DP. Para alcançar tal garantia, Charest (2011) propõe gerar cada base sintética, sob condição de ϵ/M *differential privacy*, onde M é o número de bases sintéticas, e constituir $\tilde{p}_m$ e $\tilde{x}_m$ para $m = 1, ..., M$.

2.6 Análise via Modelo Bayesiano

Naturalmente, a inferência em bases sintéticas se daria pelas *combining rules*, apresentadas na Seção 2.3.2. Porém, Charest (2011) mostrou que quando os dados sintéticos estão sob condição de ϵ-*differential privacy*, o estimador $\bar{q}_m$ para o parâmetro p, é viesado independente da quantidade de bases sintéticas divulgadas. Além disso, a variância deste estimador é superestimada. Portanto, as *combining rules* não se mostram propícias quando utilizadas para analisar bases de dados sintéticas sob ϵ-*differential privacy*.

2.6.1 Modelo Beta - Binomial

Nesta conjuntura, Charest (2011) introduz um método alternativo às *combining rules*. A ideia é usar o mecanismo de geração dos dados sintéticos dentro de um modelo Bayesiano e fazer inferência a partir da distribuição à posteriori de p.

O modelo completo é dado por:

$$
\begin{aligned}
&p \sim Beta(\gamma_1, \gamma_2) \\
&x \sim Binomial(n, p) \\
&\tilde{p}_m \sim Beta(\alpha_1 + x, \alpha_2 + n - x) \\
&\tilde{x}_m \sim Binomial(\tilde{n}, \tilde{p}_m),
\end{aligned}
$$

onde p e n são os parâmetros de probabilidade de sucesso e tamanho amostral, respectivamente, da base original, resumida pela estatística suficiente x. γ_1 e γ_2 são hiperparâmetros do modelo determinados pelo pesquisador. Na Figura 2.2 temos o modelo gráfico probabilístico exposto:

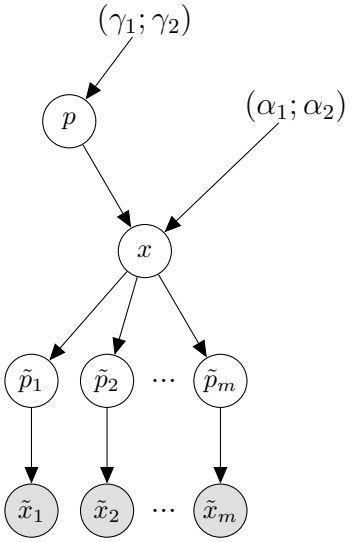

Figura 2.2: Modelo Gráfico Binomial.

Note que, a única informação disponível ao usuário da base são os dados sintéticos $\tilde{x}_m$ e o parâmetro de confidenciabilidade α_1 e α_2 do *differential privacy*. A distribuição conjunta do modelo é caracterizada por:

$$
\begin{aligned}
f(p, x, \tilde{p}, \tilde{x}) \;\propto\; & p^{\gamma_1 - 1}(1-p)^{\gamma_2} \binom{n}{x} p^x (1-p)^{n-x} \\
& \prod_{m=1}^{M} \tilde{p}_m^{\alpha_1 + x - 1}(1 - \tilde{p}_m)^{\alpha_2 + n - x - 1} \frac{\Gamma(\alpha_1 + \alpha_2 + n)}{\Gamma(\alpha_1 + x) + \Gamma(\alpha_2 + n - x)} \\
& \prod_{m=1}^{M} \binom{\tilde{n}}{\tilde{x}_m} \tilde{p}_m^{\tilde{x}_m}(1 - \tilde{p}_m)^{\tilde{n} - \tilde{x}_m}.
\end{aligned}
$$

Como não há forma fechada da distribuição para nenhum dos parâmetros, usaremos métodos MCMC para fazer a inferência. Para os parâmetros p e $\tilde{p}_m$, temos um modelo conjugado e podemos encontrar as distribuições condicionais:

$$p|x, \tilde{p}, \{\tilde{x}_m\}_{m=1}^{M} \sim Beta(\gamma_1 + x, \gamma_2 + n - x)$$
$$\tilde{p}_m|x, p, \tilde{x}_m \sim Beta(\alpha_1 + \tilde{x}_m + x, \alpha_2 + \tilde{n} - \tilde{x}_m + n - x).$$

Para x, é possível chegar na condicional completa, dada por:

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x} \prod_{m=1}^{M} \tilde{p}_m^{\alpha_1+x-1} (1-\tilde{p}_m)^{\alpha_2+n-x-1} \frac{\Gamma(\alpha_1+\alpha_2+n)}{\Gamma(\alpha_1+x)+\Gamma(\alpha_2+n-x)}. \quad (2.7)$$

Neste trabalho, propomos utilizar três métodos para amostrar destas distribuições: amostrando da própria condicional completa via *Gibbs sampling*, outro modo usando um *Gibbs sampling* com passo de *Metropolis-Hastings* para o parâmetro x, e por último utilizamos o software JAGS para o processo de amostragem. Os detalhes destas metodologias serão apresentados a seguir.

Gibbs Sampling

O *Gibbs sampling* é o algoritmo de *Markov chain Monte Carlo*(MCMC) mais antigo e conhecido. Proposto inicialmente por Geman e Geman (1984) no contexto de reconstrução de imagem, propunha um esquema de amostragem da distribuição a posteriori a partir das condicionais completas por meio de um algoritmo iterativo. O algoritmo ganhou notoriedade na área estatística com o trabalho de Gelfand e Smith (1990).

O algoritmo é aplicável quando a distribuição conjunta a posteriori não é conhecida, porém conseguimos amostrar, facilmente, da distribuição condicional de cada variável.

Assuma que $\boldsymbol{\theta} = (\boldsymbol{p}, \boldsymbol{p_m}, \boldsymbol{x})$ é nosso vetor de parâmetros. O *Gibbs sampling* é configurado da seguinte forma:

Algoritmo 3: Gibbs Sampling

Entrada: $\boldsymbol{\theta^{(0)}} = (p^{(0)}, p_m^{(0)}, x^{(0)})$, $t = 0$
Saída: Estimativas $p, \tilde{p}, x$

para *cada m cópia* **faça**
  enquanto $t \leq n_{iter}$ **faça**
    $p \leftarrow Beta(\gamma_1 + x, \gamma_2 + n - x)$
    $\tilde{p} \leftarrow Beta(\alpha_1 + \tilde{x}_m + x, \alpha_2 + \tilde{n} - \tilde{x}_m + n - x)$
    $\forall x = 1, ..., n$, calcula-se:

$$P(x) = \binom{n}{x} (p^{(t-1)})^x (1 - p^{(t-1)})^{n-x} \prod_{m=1}^{M} (\tilde{p}_m^{(t-1)})^{\alpha_1 + x - 1} (1 - \tilde{p}_m^{(t-1)})^{\alpha_2 + n - x - 1} B^{-1}(\alpha_1 + x, \alpha_2 + n - x),$$

    Sorteia-se $x^{(t)}$ de acordo com as probabilidades $P(x)$ calculadas.
    $t = t + 1$
  fim
fim

onde $B^{-1}(\alpha_1 + x, \alpha_2 + n - x) = \frac{\Gamma(\alpha_1 + \alpha_2 + n)}{\Gamma(\alpha_1 + x) + \Gamma(\alpha_2 + n - x)}$.

A medida em que t cresce, a cadeia se aproxima da estabilidade. Descartamos os primeiros t_0 valores simulados, período conhecido como *burn-in*, no qual a cadeia de Markov passa por um "aquecimento". E então, a partir da observação $t_0 + 1$, os valores de $\boldsymbol{\theta^{(t)}}$ são tratados como amostra da distribuição a posteriori do parâmetro.

Note que conforme n cresce, há mais possibilidades para o valor de x e consequentemente um esforço computacional maior em obter a probabilidade de cada possível x. Nesse contexto, trouxemos a proposta de utilizar o algoritmo de *Metropolis-Hastings*, apresentada a seguir, com intuito de diminuir o custo computacional.

Metropolis-Hastings L

O algoritmo Metropolis Hastings, proposto por Metropolis et al. (1953), e posteriormente extendido por Hastings (1970), é mais um membro da família MCMC que tem por objetivo amostrar valores da distribuição a posteriori, quando esta não é fácil de simular e, além disso, não temos uma forma fechada para a distribuição.

A ideia geral do algoritmo é simular amostras de uma distribuição através da distribuição conjunta completa e de uma outra distribuição auxiliar e independente, proposta pelo usuário.

Para o problema em questão, apenas x não tem forma fechada para a distribuição a posteriori, portanto, utilizaremos o MH para obter sua amostra. Logo, teremos uma mistura de dois algoritmos: *Gibbs sampling* para simular da posteriori de p e $\tilde{p}_m$, e um passo de MH para simular da posteriori de x.

Algoritmo 4: Gibbs com passo MH

Entrada: $\theta^{(0)} = (p^{(0)}, p_m^{(0)}, x^{(0)})$, $t = 0$
Saída: Estimativas $p, \tilde{p}, x$

enquanto $t \leq n_{iter}$ **faça**

- **Gibbs Sampling**
- $p \leftarrow Beta(\gamma_1 + x, \gamma_2 + n - x)$
- $\tilde{p} \leftarrow Beta(\alpha_1 + \tilde{x}_m + x, \alpha_2 + \tilde{n} - \tilde{x}_m + n - x)$
- **Metropolis Hastings**
- $x^{cand} \sim U(x^{(t-1)} - L, x^{(t-1)} + L)$,
- $\tau(x^{cand}|x^{t-1}) = \min\left[1, \frac{q(x^{(t-1)}|x^{cand})\pi(x^{cand})}{q(x^{cand}|x^{(t-1)})\pi(x^{(t-1)})}\right]$,
- Aceita-se x^{cand} se:
- **if** $\tau(x^{cand}|x^{t-1}) \leq u \propto Unif(0,1)$ **then**
  - $x = x^{cand}$
- **else**
  - $x = x^{t-1}$
- **end**
- $t = t + 1$

fim

De acordo com a metodologia do Metropolis-Hastings, precisamos definir uma distribuição proposta para gerar um valor candidato para o parâmetro. Neste caso, utilizamos uma distribuição proposta uniforme discreta de tamanho 2L ao redor do valor do parâmetro da iteração anterior. Note que, a distribuição proposta é simétrica em grande parte do espaço paramétrico, porém há a restrição em que, $[x^{(t-1)} - L > 0$; $x^{(t-1)} + L \leq n]$, tornando a distribuição assimétrica e truncada nas bordas. Outro fato interessante é a atuação do parâmetro L como controlador da taxa de aceitação do algoritmo de *Metropolis-Hastings*.

JAGS

Just Another Gibbs Sampler(JAGS) é um *software* gratuito desenvolvido para fazer análises em modelos Bayesianos usando *Markov Chain Monte Carlo*(MCMC), baseado na linguagem de programação do BUGS Plummer (2015), a partir da especificação do modelo na linguagem para, então, gerar uma amostra da distribuição a posteriori dos parâmetros. Isso pode ser feito através do R utilizando o pacote `rjags`, implementado por Plummer (2016).

Segundo Plummer et al. (2003), as maiores vantagens da família JAGS é ser uma plataforma independente e editável, além de sua programação ser executada inteiramente em C++, enquanto a família BUGS é *Component Pascal*, o que torna o JAGS mais eficiente do ponto de vista computacional.

A ideia geral dos desenvolvedores do JAGS é tornar a vida do usuário a mais simples

possível. "Rodar" um modelo refere-se a gerar amostras da distribuição a posteriori dos parâmetros. O usuário é encarregado de executar cinco passos:

1. Definir o modelo
2. Compilar
3. Inicialização
4. Adaptação e *burn-in*
5. Monitoramento

Feito isso, o JAGS utiliza-se de algoritmos de geração de amostras aleatórias da distribuição a posteriori do modelo. É importante ressaltar que a análise de convergência, análise descritiva e inferencial do modelo deve ser feita em outro *software*.

Uma das limitações do JAGS é quanto à capacidade em lidar com problemas mais complexos. Quando os modelos hierárquicos não tem conjugação, a forma da posteriori é desconhecida ou a quantidade de parâmetros cresce, não é possível obter uma boa aproximação da amostra para o modelo.

Para a implementação do modelo no *software* JAGS seguimos a proposta de Charest (2011), e utilizamos também o pacote `rjags` para análise das amostras a posteriori.

2.6.2 Modelo Dirichlet - Multinomial

Vamos agora considerar a extensão do modelo descrito na seção anterior para dados categóricos. Suponhamos que uma agência tem a intenção de divulgar aos analistas uma variável confidencial categórica $\mathbf{X} = X_1, ..., X_k$, onde k é o número de categorias. Vamos assumir que esta variável passou pelo processo de sintetização multinomial e satisfaz a condição de ϵ-*differential privacy*. Como foi mostrado anteriormente, as *combining rules* não são efetivas quando temos dados sintéticos sob tal condição. Portanto, precisamos de um modelo análogo ao Beta-Binomial para fazer inferência nesta variável sintética.

Uma extensão do modelo Beta-Binomial introduzido por Charest (2011) é o modelo Dirichlet-Multinomial. Esta é a alternativa inferencial quando a base de dados sintéticos divulgada possui dados com múltiplas categorias. O processo de inferência é uma generalização do caso binomial.

O processo inferencial nesta variável é dado pelo seguinte modelo:

$$
\begin{aligned}
\mathbf{p} &\sim Dirichlet(\boldsymbol{\gamma}) \\
\mathbf{x} &\sim Multinomial(n, \mathbf{p}) \\
\tilde{\mathbf{p}}_\mathbf{m} &\sim Dirichlet(\boldsymbol{\alpha} + \mathbf{x}) \\
\tilde{\mathbf{x}}_\mathbf{m} &\sim Multinomial(\tilde{n}, \tilde{\mathbf{p}}_\mathbf{m}),
\end{aligned}
$$

onde, $\boldsymbol{\gamma}$ é um vetor de hiperparâmetros para $\mathbf{p}$, determinados pelo pesquisador, n é o tamanho amostral, $\tilde{\mathbf{p}}_\mathbf{m}$ e $\tilde{\mathbf{x}}_\mathbf{m}$, são o vetor de probabilidades e contagens das categorias na

base sintética, respectivamente. Ainda temos o parâmetro de privacidade $\boldsymbol{\alpha}$ determinado pelo grau de de confidencialidade ϵ. O modelo gráfico probabilístico pode ser visto na Figura 2.3.

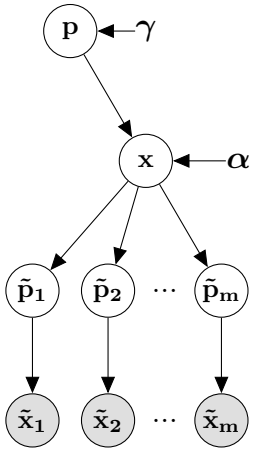

Figura 2.3: Modelo Gráfico Multinomial

Observe que a única parte observável no nosso modelo são os dados sintéticos $\tilde{\mathbf{x}}_\mathbf{m}$. O restante são variáveis latentes a serem estimadas.

Nesse modelo nossos parâmetros de interesse são as probabilidades originais das categorias $\mathbf{p} = (p_1, ..., p_k)$. A distribuição conjunta do nosso modelo é expressa por:

$$f(p, x, \tilde{p}, \tilde{x}) \propto \prod_{i=1}^{k} p_i^{\gamma_i - 1} \prod_{i=1}^{k} \Gamma(x_i + 1)^{-1} p_i^{x_i} \prod_{j=1}^{M} \Gamma\left(\sum_{i=1}^{k} \alpha_i + x_i\right)$$
$$\prod_{i=1}^{k} \tilde{p}_{ij}^{\alpha_i + x_i - 1} \Gamma(\alpha_i + x_i)^{-1} \prod_{j=1}^{M} \prod_{i=1}^{k} \tilde{p}_{ij}^{\tilde{x}_{ij}} \Gamma(\tilde{x}_{ij} + 1)^{-1}$$

Como a distribuição a posteriori não é de nenhuma família conhecida de distribuições, a solução para estimar os parâmetros do modelo é utilizar algoritmos MCMC. Da mesma forma que o caso Beta-Binomial, utilizaremos o Gibbs com passo de Metropolis-Hastings.

Para utilizar o algoritmo, primeiramente precisamos das distribuições condicionais dos parâmetros. Abaixo é explícito as condicionais para cada um dos parâmetros:

$$f(p|x, \tilde{p}, \tilde{x}) \propto \prod_{i=1}^{k} p_i^{\gamma_i - 1 + x_i} \propto Dir(\gamma_i + x_i)$$

$$f(\tilde{p}|p, x, \tilde{x}) \propto \prod_{i=1}^{k} \tilde{p}_{ij}^{\alpha_i + x_i - 1 + \tilde{x}_{ij}} \propto Dir(\gamma_i + x_i + \tilde{x}_{ij})$$

$$q(x) = f(x|p, \tilde{p}, \tilde{x}) \propto \Gamma\left(\sum_{i=1}^{k} \alpha_i + x_i\right) \prod_{i=1}^{k} \Gamma(x_i + 1)^{-1} p_i^{x_i} \Gamma(\alpha_i + x_i)^{-1} \tilde{p}_{ij}^{\alpha_i + x_i - 1}$$

Tendo em mãos o cálculo das distribuições condicionais, podemos aplicar o algoritmo de *Gibbs sampling* com passo de *Metropolis-Hastings*.

Algoritmo 5: Gibbs Sampling com Passo de Metropolis

Entrada: $\boldsymbol{\beta} = Unif(n_{cat})$,
Saída: Estimativas $\boldsymbol{p}, \tilde{\boldsymbol{p}}, \boldsymbol{x}$

enquanto $t \leq n_{iter}$ **faça**
- **Gibbs Sampling**
- $\boldsymbol{p} \leftarrow Dirichlet(\boldsymbol{\gamma} + \boldsymbol{x}^{(t-1)})$
- $\tilde{\boldsymbol{p}} \leftarrow Dirichlet(\boldsymbol{\gamma} + \boldsymbol{x}^{(t-1)} + \tilde{\boldsymbol{x}}^{(t-1)})$
- **Metropolis Hastings**
- Sorteia-se um candidato para $\boldsymbol{x}$ com distribuição:
- $\boldsymbol{x}^{cand} \leftarrow Multinomial(\boldsymbol{\beta})$
- A Probabilidade de aceitação τ é definida por:
- $\tau(\boldsymbol{x}^{cand}|\boldsymbol{x}^{(t-1)}) = \min\left[1, \frac{q(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{cand})\pi(\boldsymbol{x}^{cand})}{q(\boldsymbol{x}^{cand}|x^{(t-1)})\pi(x^{(t-1)})}\right],$
- **if** $\tau(\boldsymbol{x}^{cand}|\boldsymbol{x}^{(t-1)}) \leq u \propto Unif(0,1)$ **then**
  - $\boldsymbol{x} = \boldsymbol{x}^{cand}$
- **else**
  - $\boldsymbol{x} = \boldsymbol{x}^{(t-1)}$
- **end**

fim

Todavia, ao aplicar o algoritmo de Gibbs com passo de Metropolis-Hastings, notamos que os parâmetros demoravam a convegir ou até mesmo nem convergiam para o valor alvo, geralmente ficando em torno da distribuição proposta, a Multinomial(β). Para contornar o problema, é preciso que a distribuição proposta seja capaz de percorrer o espaço paramétrico de forma eficiente. É intuitivo pensar que após um número de passos do algoritmo tenhamos que fazer uma atualização nos parâmetros da distribuição proposta.

É justamente o que fazem os chamados algoritmos adaptativos. Proposto por Haario et al. (1999), como o próprio nome sugere, a medida em que o algoritmo "anda", após um certo número iterações há uma atualização dos parâmetros da distribuição proposta. Com isso, é possível percorrer de forma eficiente o espaço paramétrico e a convergência dos parâmetros acontece de forma mais rápida.

Abaixo é descrito o algoritmo de Gibbs com passo de Metropolis adaptativo para gerar amostras a posteriori:

Algoritmo 6: Gibbs Sampling com Passo de Metropolis

Entrada: ponto = 500, $\beta = Unif(n_{cat})$,
Saída: Estimativas $\boldsymbol{p}, \tilde{\boldsymbol{p}}, \boldsymbol{x}$

1 **enquanto** $t \leq n_{iter}$ **faça**
2 **Gibbs Sampling**
3 $\boldsymbol{p} \leftarrow Dirichlet(\boldsymbol{\gamma} + \boldsymbol{x}^{(t-1)})$
4 $\tilde{\boldsymbol{p}} \leftarrow Dirichlet(\boldsymbol{\gamma} + \boldsymbol{x}^{(t-1)} + \tilde{\boldsymbol{x}}^{(t-1)})$
5 **Metropolis Hastings**
6 **if** *t multiplo de ponto* **then**
7 $\boldsymbol{\beta} \leftarrow$ Média$(\boldsymbol{p}^{(t-500)}, ..., \boldsymbol{p}^{(t)})$
8 **end**
9 Sorteia-se um candidato para $\boldsymbol{x}$ com distribuição:
10 $\boldsymbol{x}^{cand} \leftarrow Multinomial(\boldsymbol{\beta})$
11 A Probabilidade de aceitação τ é definida por:
12 $\tau(\boldsymbol{x}^{cand}|\boldsymbol{x}^{(t-1)}) = \min\left[1, \frac{q(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{cand})\pi(\boldsymbol{x}^{cand})}{q(\boldsymbol{x}^{cand}|x^{(t-1)})\pi(x^{(t-1)})}\right]$,
13 **if** $\tau(\boldsymbol{x}^{cand}|\boldsymbol{x}^{(t-1)}) \leq u \propto Unif(0,1)$ **then**
14 $\boldsymbol{x} = \boldsymbol{x}^{cand}$
15 **else**
16 $\boldsymbol{x} = \boldsymbol{x}^{(t-1)}$
17 **end**
18 **fim**

Observe que na inicialização do algoritmo informamos em quantos passos devemos atualizar a distribuição proposta, no nosso caso 500 iterações. No passo 7 do algoritmo, é onde está descrita a atualização, ou seja, a cada 500 iterações, cada parâmetro da multinomial é atualizado pela média das 500 amostras anteriores.

Capítulo 3

Resultados

Nesta seção serão apresentados os resultados da pesquisa. Separamos o capítulo em duas sessões, uma referente ao estudo de simulação do modelo Beta-Binomial e outra do modelo Dirichlet-Multinomial.

Para o modelo dicotômico, seguimos os cenários de simulação propostos por Charest (2011) para avaliação dos métodos de análise de bases sintéticas sob condição de ϵ-*differential privacy*. Para tal, o modelo será implementado nas três abordagens apresentadas na Seção 2.6, para diferentes níveis de ϵ e M. A comparação de performance considera o poder inferencial e o tempo computacional dos métodos. Para o modelo cuja performance é superior, extenderemos as análises da distribuição a posteriori do parâmetro p sob diferentes níveis de ϵ e M.

Para o modelo multicategórico, consideramos apenas uma base sintética, $M = 1$, devido aos melhores resultados apresentados no caso binomial. Avaliaremos cenários análogos ao modelo dicotômico, sobre o efeito do parâmetro de privacidade ϵ, tamanho da amostra e etc, na inferência sobre a base de dados. No estudo de simulação, buscamos mostrar também o impacto quando estamos em um quadro de maior dimensionalidade nos parâmetros.

3.1 Análises Beta-Binomial

3.1.1 Sensibilidade L

Na proposta do Gibbs com passo *Metropolis-Hastings*, precisamos determinar o tamanho do passo L a ser dado na distribuição proposta. A avaliação para saber o quão bom é o tamanho do passo é baseada na taxa de aceitação do algoritmo, assim como apresentado por Charest (2011), nosso objetivo é algo em torno de 45%.

Para comparar as taxas de aceitação, determinamos um *grid* de $L = (6, ..., 12)$. Simulamos 1000 cadeias, cada uma com 2000 iterações e *burn-in* de 1000. E para cada um dos valores de L, calculamos a média da taxa de aceitação. O resultado pode ser visto na Figura 3.1.

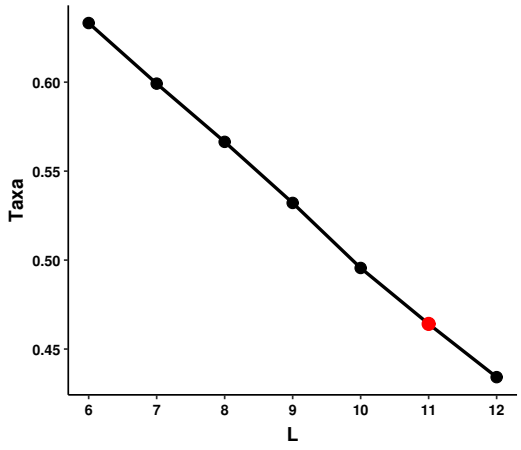

Figura 3.1: Taxa de aceitação para os valores de L

Note que há um decrescimo linear na taxa de aceitação conforme aumentamos o tamanho do passo no algoritmo *Metropolis-Hastings*, o que é intuitivo, pois o aumento do L faz com que a probabilidade de aceitação de um novo candidato diminua. O L ótimo para nossa condição é 11, tendo uma taxa de aceitação média de 0.464.

3.1.2 Comparação entre Modelos

Aqui serão apresentados os resultados de comparação entre os modelos propostos para fazer inferência sobre o parâmetro p nas bases sintéticas sob *differential privacy*. O objetivo é analisar em qual método há uma melhor aproximação da inferência feita na base de dados verdadeira.

Para via de comparações, replicamos o cenário proposto por Charest (2011). Consideramos que na base de dados real, a estatística suficiente é $x = 30$ com $n = 100$, e o parâmetro de privacidade $\epsilon = 2$. Utilizamos uma distribuição a priori para p sendo Uniforme[0,1] ou equivalentemente Beta(γ_1,γ_2), onde $\gamma_1 = \gamma_2 = 1$. Como distribuição a posteriori resultante, temos uma Beta(31,71), portanto o valor esperado a posteriori é de 0.3039 e a variância de 0.002053. Os resultados consideram os cenários com $M = 1, 2, 5, 10$ bases sintéticas.

Para cada cenário, foram geradas cadeias com 2000 iterações, *burn-in* de 1000 e *thinning* de 20. Este experimento foi replicado para 1000 amostras diferentes. Para cada algoritmo as comparações de estimativas dos métodos de *Gibbs Sampling*, *Gibbs + Metropolis-Hastings-L* e via *JAGS* são mostradas a seguir na Tabela 3.1 e Tabela 3.2.

Tabela 3.1: Esperança de p

Método	M = 1	M = 2	M = 5	M = 10
Metropolis	0.3104	0.3076	0.3186	0.3661
Gibbs	0.3105	0.3043	0.3116	0.3334
JAGS	0.3114	0.3130	0.3175	0.3395

Note que todos os métodos possuem boas estimativas pontuais para o parâmetro

p, quando temos $M = 1$ bases sintéticas. A medida que M cresce, há mais viés nas estimativas. Isto ocorre pois, quanto maior o número de bases sintéticas, maior será a variabilidade entre as bases.

Tabela 3.2: Variância de p

Método	M = 1	M = 2	M = 5	M = 10
Metropolis	0.0060	0.0069	0.0102	0.0119
Gibbs	0.0063	0.0068	0.0124	0.0156
JAGS	0.0060	0.0079	0.0107	0.0177

Quanto às estimativas da variância, os algoritmos superestimam um pouco a verdadeira variância a posteriori(0.002053). Porém esse resultado é esperado, pois a condição de ϵ-*differential privacy* induz variabilidade na bases sintéticas.

Quando estamos em cenários de simulação via MCMC, gostaríamos que as amostras fossem não correlacionadas ou pelo menos, perto disso. Um método de mensurar essa autocorrelação , é o *effective sample size*, que é capaz de nos fornecer o tamanho amostral efetivo, caso as amostras fossem independentes. Na Tabela 3.3 temos o *effective sample size* para os métodos MCMC considerando 1000 iterações.

Tabela 3.3: Effective Sample Size

Método	M = 1	M = 2	M = 5	M = 10
Metropolis	656	552	292	207
JAGS	1000	897	684	484
Gibbs	1000	1000	812	399

A interpretação do resultado é simples. Peguemos o caso do Gibbs com passo Metropolis: para uma cadeia com 1000 amostras e $M = 1$, equivale à uma amostra independentes de tamanho 656. Note que o Gibbs com passo Metropolis sempre possui tamanhos amostrais efetivos menores, o que é esperado devido à construção do algoritmo, mais especificamente na parte de aceitação dos candidatos. Quando não aceitamos um candidato, o valor de nossa amostra é igual ao do passo anterior.

A título de menção, testamos outras formas para a priori do modelo, como Beta($\gamma_1 = 0.1$,$\gamma_2 = 0.1$), Beta($\gamma_1 = 0.01$,$\gamma_2 = 0.01$), porém não houve ganho substancial nas estimativas do parâmetro p, inclusive trouxe uma piora inferencial.

3.1.3 Tempo Computacional

Para efeitos de comparação entre modelos, é interessante saber o gasto computacional efetuado por eles, com propósito de avaliar a viabilidade da implantação do método. Utilizamos uma máquina com 6MB de mémoria RAM e Processador intel(i5) 4º geração. Os resultados com o tempo para uma simulação e o tempo total são explicitados na Tabela 3.4.

Tabela 3.4: Tempo Computacional

Método	Tempo(1)	Tempo(total)
Metropolis	1.98s	2h59m
JAGS	0.75s	3h34m
Gibbs	2.22s	5h17m

Note que o software JAGS tem menor tempo computacional se comparado com o outros dois métodos, seguido do *Metropolis-Hastings* e o *Gibbs*. Esse fato era esperado já que o algoritmo é implementado em C++ e os outros dois em R. Contudo, se implementarmos o algoritmo de *Metropolis-Hastings* utilizando o pacote `Rcpp`, desenvolvido por Eddelbuettel e François (2011), temos uma redução de quase 80% do tempo.

Tendo em vista o tempo computacional e as estimativas mostradas pelos três métodos, decidimos continuar as análises utilizando o modelo do *Gibbs* com passo *Metropolis-Hastings*. Pois além de conseguirmos uma melhora computacional, este modelo também é mais flexível e será utilizada no caso multinomial, onde não há possibilidade de implementação no JAGS e do *Gibbs*. Portanto, todas as avaliações inferências, encontradas nas Seções 3.1.4 3.1.5 e 3.1.6, serão sob este método.

3.1.4 Análise da distribuição de p

Em estatística Bayesiana, a nossa informação sobre o parâmetro de interesse está na distribuição a posteriori. Portanto, apenas medidas resumo como estimações pontuais não são suficientes para tirar conclusões concretas sobre o parâmetro, é também necessário analisar a distribuição das estimativas.

Pela Figura 3.2, obtemos um *insight* do comportamento da distribuição das estimativas de p para os diferentes tamanhos de base sintética. O primeiro fato que destaca-se é o aumento da variância conforme aumentamos o número de bases sintéticas divulgadas.

Na Figura 3.3 temos as densidades aproximadas da distribuição dos parâmetros. Lembrando que M é o número de bases divulgadas ao público.

Um fato que fica claro é quanto a influência do número de bases divulgadas na distribuição das estimativas. Quanto maior o número de bases sintéticas divulgadas para análise, há mais ruído na inferência do parâmetro. A ocorrência dessa situação está ligada com o fato de que para garantir ϵ-*differential privacy* para M bases sintéticas, cada uma das bases atinge um nível de confidencialidade de $\frac{\epsilon}{M}$, tornando assim a condição muito mais rigorosa.

Taxa de cobertura

Um importante medida em análise Bayesiana é a taxa de cobertura das estimativas. Tal medida nada mais é que: a proporção de vezes que o valor alvo caiu dentro do intervalo HPD para as diferentes amostras.

Consideramos um intervalo HPD com 95% de credibilidade para nossas amostras. Obtivemos que, em 98% dos casos o valor verdadeiro do parâmetro estava contido no intervalo HPD. Taxa acima da recomendada, em torno de 95%.

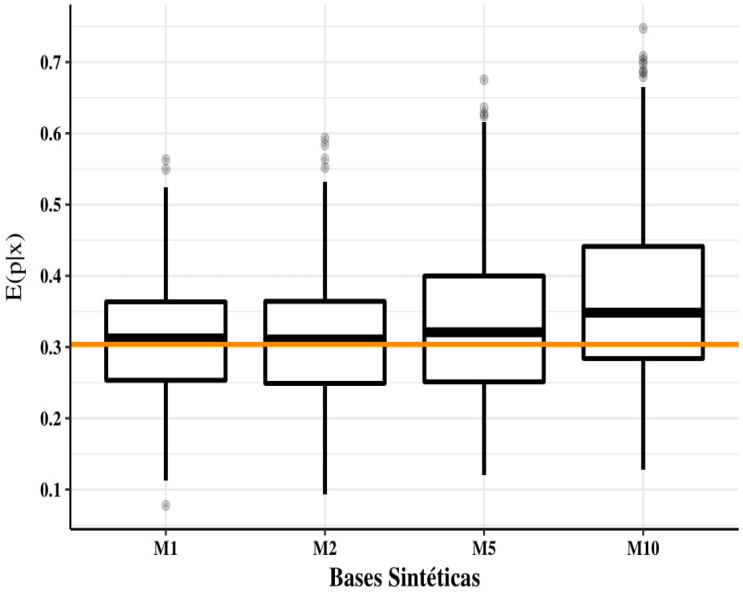

Figura 3.2: Boxplot Bases Sintéticas

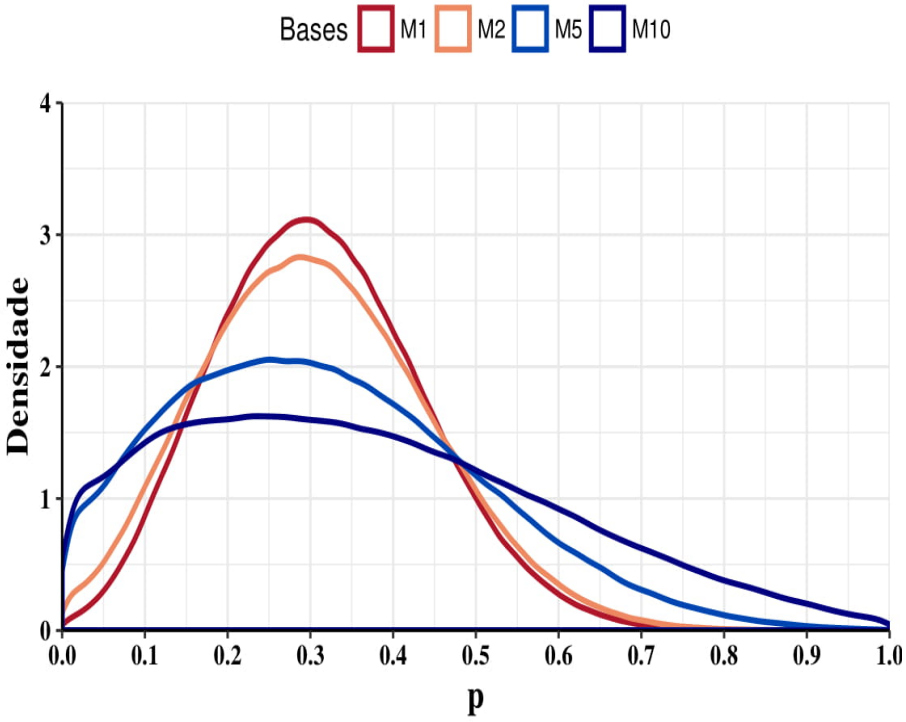

Figura 3.3: Densidades Bases Sintéticas

3.1.5 Análise de sensibilidade ϵ

Segundo Dwork e Roth (2013), quando aplicamos a metodologia de ϵ-*differential privacy* espera-se que, à medida que variamos o parâmetro ϵ de confidencialidade, temos uma mudança na base de dados e consequentemente, sobre as estimativas dos parâmetros do modelo também. Portanto é importante estudar a sensibilidade das estimativas ao parâmetro de privacidade.

Considerando o cenário de simulação proposto por Charest (2011), as simulações a seguir levam em conta os mesmos moldes da comparação entre modelos na Seção 3.1.2, para apenas uma base sintética, $M = 1$ e para os seguintes níveis de ϵ =(0.1,0.5,1,2,3,250).

Observe na Tabela 3.5 que, conforme aumentamos nossa confidencialidade, ou seja, diminuimos ϵ, perdemos poder inferencial sobre o modelo, levando-nos a resultados viesados e com aumento da variância do estimador.

Tabela 3.5: Análise de Sensibilidade ϵ

ϵ	$E(p\|x)$	$Var(p\|x)$
0.1	0.4891	0.0360
0.5	0.3766	0.0184
1	0.3236	0.0118
2	0.3106	0.0062
3	0.3110	0.0048
250	0.3077	0.0040

Considerando as distribuições a posteriori para ϵ mostradas na Figura 3.4, podemos concluir que o aumento da confidencialidade(ϵ menor) torna a inferência mais viesada. Novamente, nota-se a distância das distribuições com os dados sintéticos e os dados originais. Essa ocorrência deve-se ao fato de o ϵ-*differential privacy* ser uma garantia extremamente forte de anonimato dos indivíduos na base.

Quando $\epsilon < 0.5$, até mesmo as estimativas pontuais são muito viesadas. Conforme diminuimos a força da anonimização, temos uma melhora na inferência sobre o parâmetro p. Note que, quando temos $\epsilon = 250$ temos praticamente nossa base original, porém o ruído gerado pelos dados sintéticos tornam a inferência ainda um pouco distante da original.

Através da Figura 3.5 podemos observar claramente a melhora das estimativas quando diminuimos a confidencialidade da base. Também é evidente que, para $\epsilon = 250$, obtemos resultados próximos aos gerados por $\epsilon = 3$.

Os dados sintéticos sob a condição de ϵ-*differential privacy* se mostram eficientes quanto à anonimização dos indivíduos na base, porém há uma grande perda inferencial quando estamos sob tais condições.

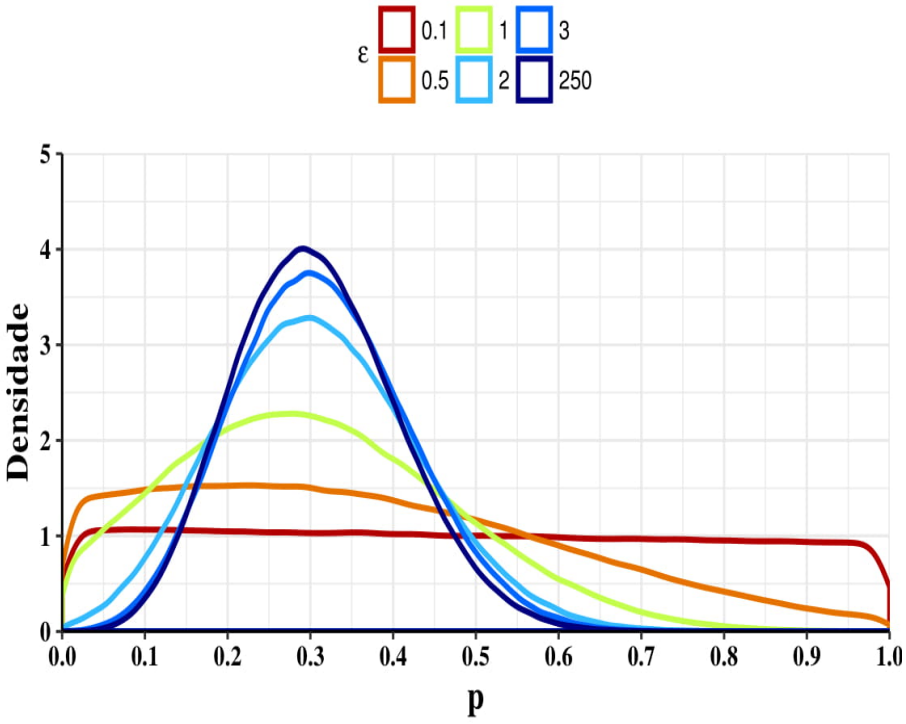

Figura 3.4: Densidades Bases Sintéticas

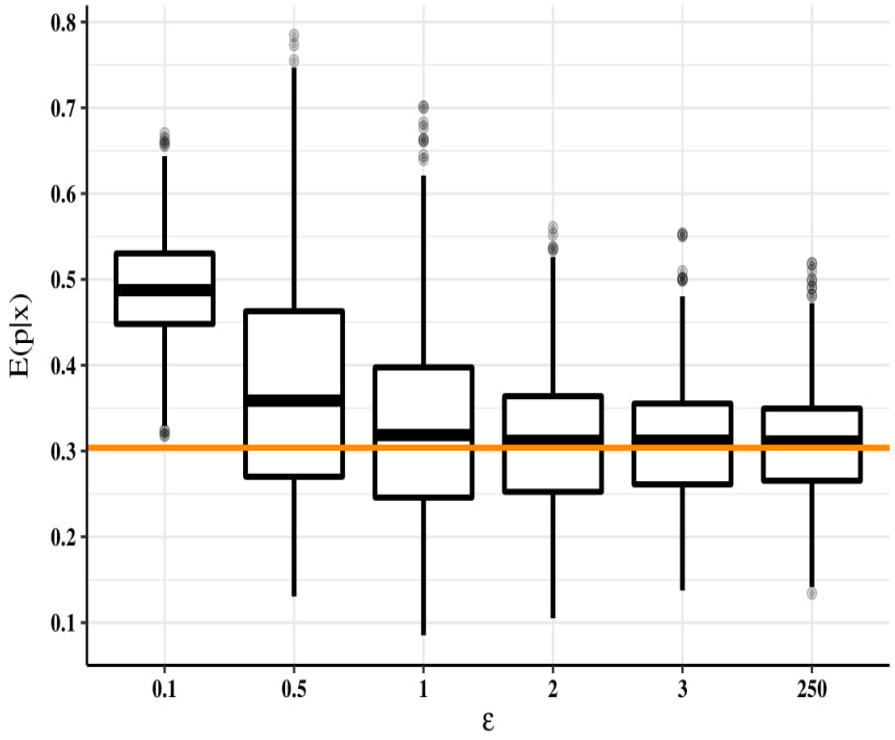

Figura 3.5: Boxplot Bases Sintéticas

3.1.6 Análise de sensibilidade n e $\tilde{n}$

Uma questão pertinente é a influência do tamanho da base original(n) e o da base sintética($\tilde{n}$) nas estimatimas do parâmetro p. Além disso, outro fator relevante de se avaliar é o comportamento do método quando temos p original próximo das bordas, ou seja, perto de 0 e de 1. Para investigar este contexto, montamos o seguinte cenário de simulação: fixamos o parâmetro $p = (0.01, 0.5, 0.99)$ e analisamos as estimativas para os seguintes pares de configuração: $(n = 100; \tilde{n} = 100)$, $(n = 1000; \tilde{n} = 1000)$ e $(n = 1000; \tilde{n} = 100)$. Também consideramos $M = (1; 5)$ bases sintéticas.

Tabela 3.6: Estimativas para p

$n = 100$, $\tilde{n} = 100$			
	p=0.01	p=0.5	p=0.99
m=1	0.076	0.497	0.925
m=5	0.097	0.498	0.903
$n = 1000$, $\tilde{n} = 1000$			
	p=0.01	p=0.5	p=0.99
m=1	0.015	0.501	0.986
m=5	0.050	0.501	0.949
$n = 1000$, $\tilde{n} = 100$			
	p=0.01	p=0.5	p=0.99
m=1	0.029	0.501	0.971
m=5	0.033	0.502	0.967

Observe, ao passo que aumentamos n, obtemos melhores estimativas para o parâmetro p. Quanto ao tamanho da base sintética $\tilde{n}$, quando este difere do tamanho real da amostra n, há uma leve piora nas estimativas.

3.2 Análises Dirichlet-Multinomial

Para avaliar o comportamento do sintetizador Multinomial-Dirichlet, montamos um cenário amplo de simulação. O objetivo é avaliar o comportamento das estimativas obtidas pelo modelo Dirichlet-Multinomial levando em consideração o efeito do tamanho amostral, número de categorias, proporções desbalanceadas e o parâmetro de confidencialidade.

Todos os cenários serão avaliados através de simulações de Monte Carlo, considerando 1000 amostras sintetizadas. Para cada amostra, o processo de modelagem Dirichlet-Multinomial é efetuado utilizando cadeias de tamanho 30000 e burn-in = 10000. O valor do hiperparâmetro da distribuição Dirichlet a priori do parâmetro p foi escolhido como $\gamma_i = 0.5$.

3.2.1 Efeito da variação de n

Um item importante na avaliação da qualidade das estimativas é a robustez perante à variação do tamanho amostral. Portanto, definimos o seguinte cenário de Monte Carlo:

- 1000 amostras sintetizadas.
- 4 categorias com proporções de (0.7, 0.1, 0.15, 0.05).
- n = (100, 500, 1000)
- $\epsilon = 2$.

Tabela 3.7: Estimativas para p

$\boldsymbol{p}$	0.70	0.10	0.15	0.05
n = 100				
$E(p\|x)$	0.7047	0.0944	0.1435	0.0573
$Var(p\|x)$	0.0101	0.0052	0.0069	0.0034
n = 500				
$E(p\|x)$	0.6925	0.0997	0.1401	0.0676
$Var(p\|x)$	0.0068	0.0026	0.0043	0.0016
n = 1000				
$E(p\|x)$	0.6826	0.1022	0.1426	0.0724
$Var(p\|x)$	0.0102	0.0030	0.0053	0.0029

A Figura 3.6 mostra a distribuição média dos estimadores para cada uma das quatro categorias para diferentes tamanhos amostrais, o ponto vermelho indica a proporção real da população. Note que, quanto maior o tamanho amostral, menor a variabilidade dos estimadores. Contudo, o efeito do tamanho amostral nas estimativas não se mostra relevante.

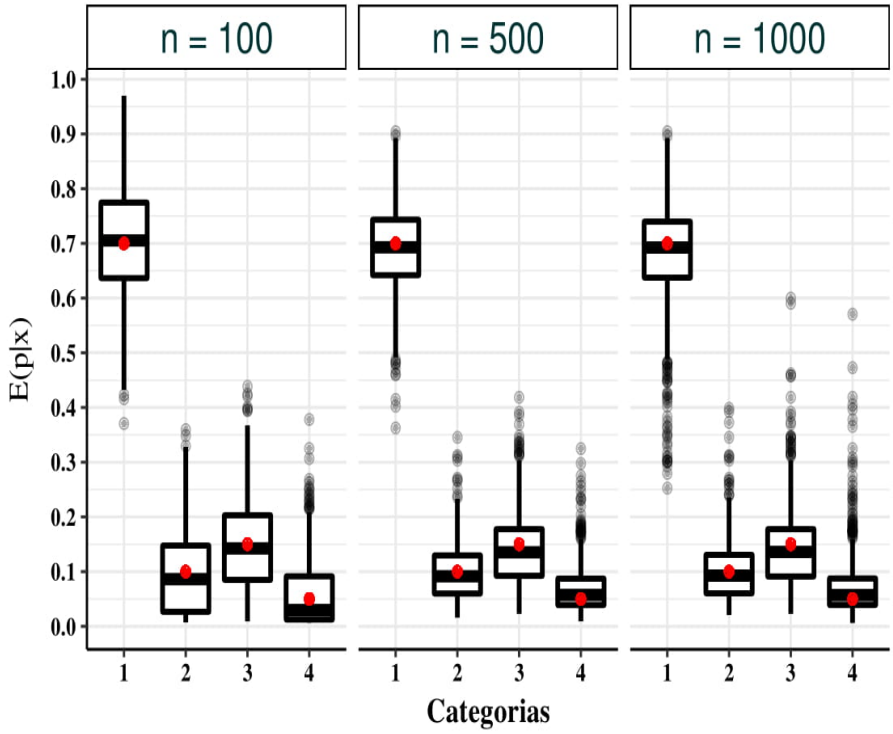

Figura 3.6: Boxplot: Influência do tamanho amostral

3.2.2 Efeito do Número de Categorias

A dimensionalidade dos problemas no processo inferencial está diretamente ligada à qualidade das estimativas, isto é, quanto mais parâmetros para estimar, possivelmente haverá uma perda na qualidade inferencial. Tendo em vista esse fato, montamos o seguinte cenário de simulação para avaliação do efeito do aumento de dimensão, ou seja, um maior número de categorias:

- n = 500.
- Parâmetro de confidencialidade $\epsilon = 2$.
- Número de categorias: (4, 8, 16).

As proporções para cada categoria nos cenários propostos foram escolhidas ao acaso e são apresentadas a seguir:

4 *Categorias* : $(0.7, 0.1, 0.15, 0.05)$

8 *Categorias* : $(0.30, 0.24, 0.16, 0.10, 0.08, 0.06, 0.04, 0.02)$

16 *Categorias* : $(0.20, 0.17, 0.12, 0.10, 0.06, 0.06, 0.04, 0.04, 0.04, 0.036, 0.03, 0.03, 0.024, 0.02, 0.02, 0.01)$

Nas Figuras 3.7, 3.8 e 3.9 temos a densidade das estimativas para os cenários de simulação propostos. Observe que a dimensionalidade do problema afeta consideravelmente

as estimativas e a variância do estimador. Por exemplo, na Figura 3.7 a densidade das estimativas estão em torno do valor real da proporção, porém quando há um aumento no número de parâmetros, as estimativas começam a possuir grande viés e alta variância.

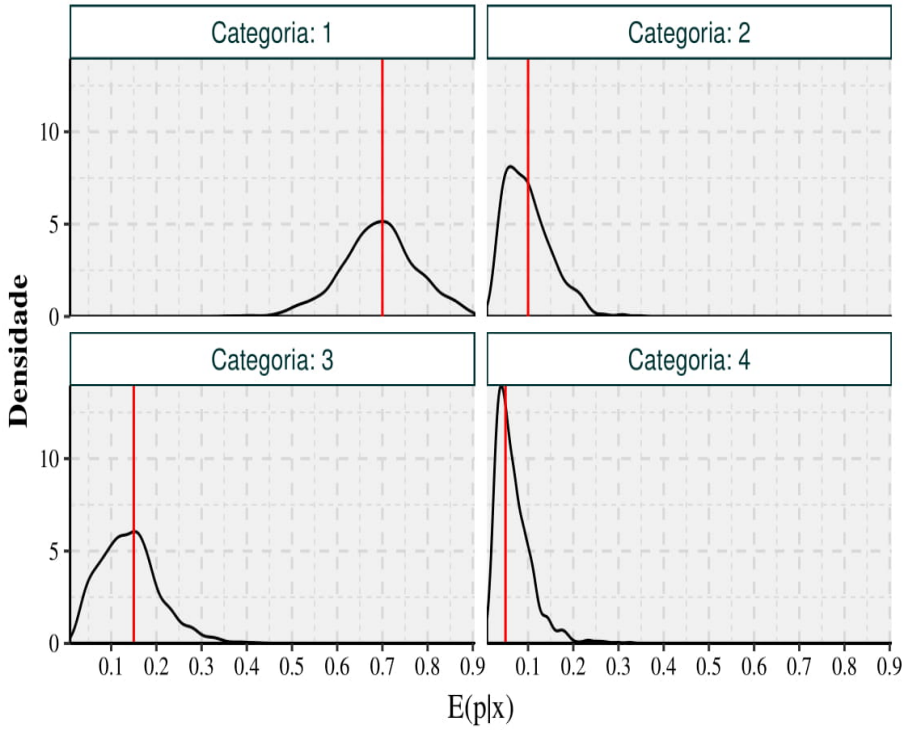

Figura 3.7: 4 categorias

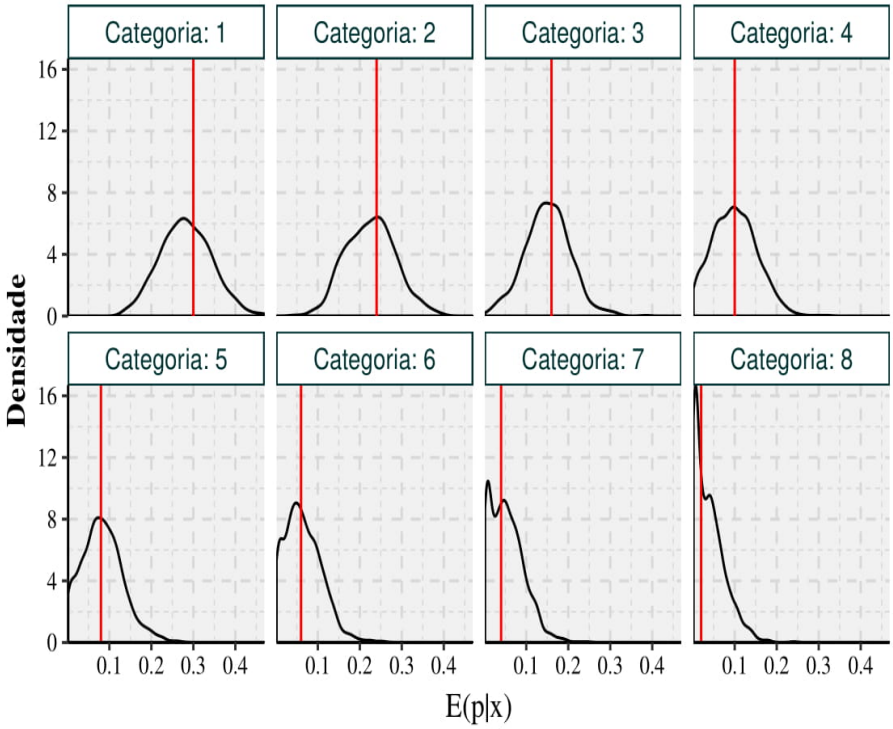

Figura 3.8: 8 categorias

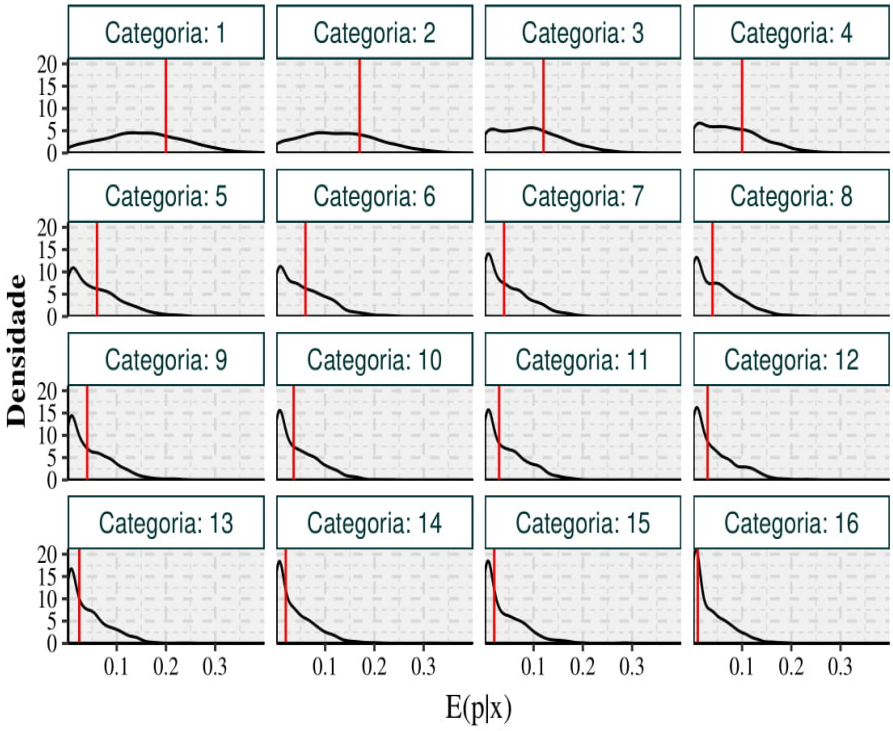

Figura 3.9: 16 categorias

É importante ressaltar que a taxa de cobertura para cada um dos cenários, considerando um intervalo HPD de 95% de credibilidade, não atingiu o valor padrão de 95% para nenhuma das categorias. Contudo, a taxa de cobertura apresentou um padrão similar para todos os cenários. Onde há uma proporção de contagens superior a 0.10, a taxa gira em torno de 80% e onde há uma proporção menor, a taxa de cobertura fica em volta de 60%.

3.2.3 Efeito de Categorias Desbalanceadas

O desbalanceamento das contagens em categorias é um fator que pode causar problemas na qualidade inferencial. Nesse sentido montamos um cenário em que temos 4 categorias com proporções balanceadas, visto na Tabela 3.8 e outro em que temos proporçõe desbalanceadas, como pode ser visto na Tabela 3.9:

Tabela 3.8: Estimativas p/ Categorias Balanceados

p	0.25	0.25	0.25	0.25
$E(p\|x)$	0.2481	0.2490	0.2534	0.2495
$Var(p\|x)$	0.0032	0.0029	0.0029	0.0030

Tabela 3.9: Estimativas p/ Categorias Desbalanceados

p	0.96	0.02	0.018	0.002
$E(p\|x)$	0.8264	0.0615	0.0588	0.0534
$Var(p\|x)$	0.0197	0.0050	0.0048	0.0053

Observe que o efeito do desbalanceamento de contagens nas categorias parece afetar mais o modelo Dirichlet-Multinomial do que o Beta- Binomial mostrado na Seção 3.1.6. Note que as categorias com baixas contagens acabam sendo superestimadas, em contrapartida a categoria que possui uma proporção maior nas contagens sofre com subestimação do parâmetro. A qualidade na inferência sobre as proporções de cada classe acaba por ser totalmente afetada quando há o desblanceamento, tendo como características alto viés e variância nas estimativas.

Observe a distribuição das estimativas para ambos cenários nas Figuras 3.10 e 3.11, em vermelho está marcado o verdadeiro valor do parâmetro. É evidente que quanto maior o desbalanceamento, maior será o viés nas estimativas. Além disso há também maior variabilidade nos valores estimados.

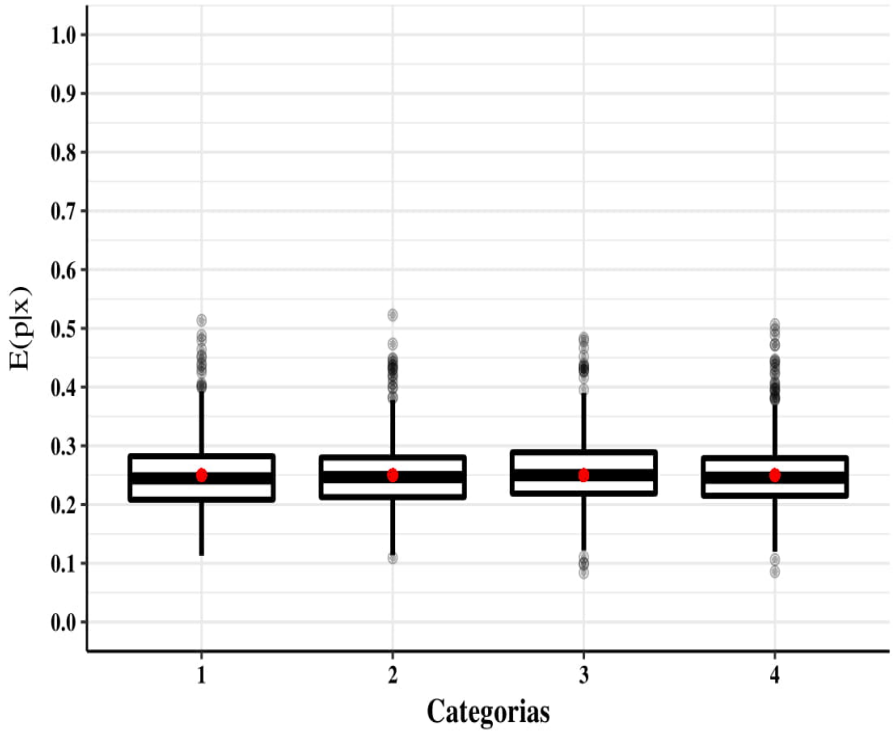

Figura 3.10: Proporções Balanceadas

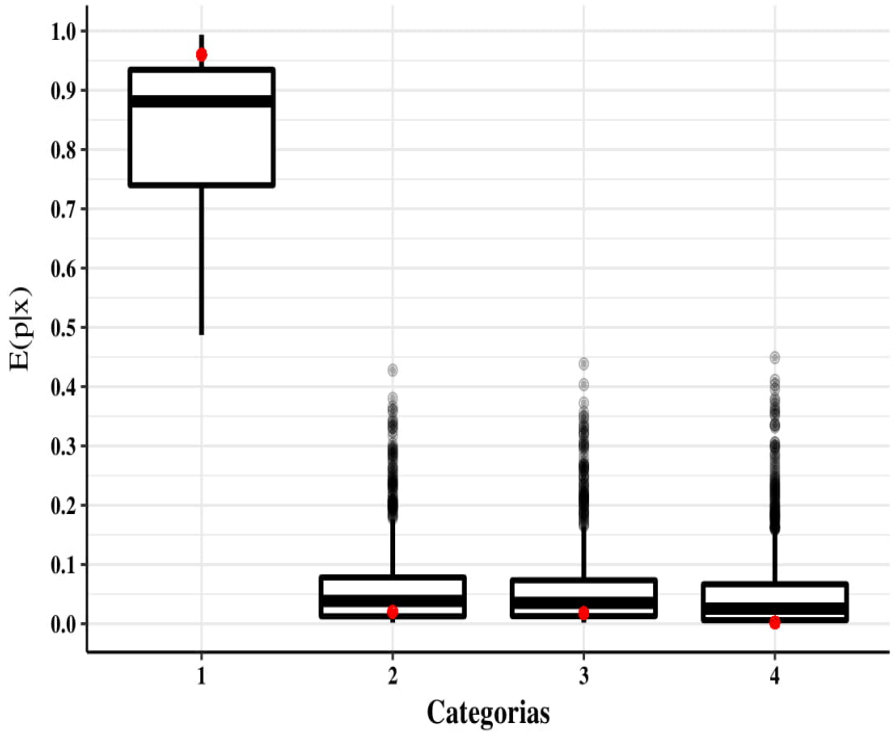

Figura 3.11: Proporções desbalanceadas

Quanto à taxa de cobertura, notamos que no caso onde há balanceamento entre as categorias, considerando um intervalo HPD com 95% de credibilidade, a taxa de cobertura foi superior à 85% em todas categorias. Porém, analisando os resultados no cenários onde há desbalanceamente, a taxa de cobertura foi de 80% para a categoria de maior contagem, e para as de menor contagem o resultado foi de uma cobertura de pouco menos que 50%.

Logo, acreditamos que um desbalanceamento combinado com baixas contagens causa um maior viés e também maior variabilidade nas estimativas, tornando-as imprecisas.

3.2.4 Efeito do parâmetro de privacidade

Na Seção 3.1 vimos a grande distorção nas estimativas que o parâmetro ϵ é capaz de causar. Com isso, também analisamos tal efeito para o modelo Dirichlet-Multinomial.

O cenário de simulação proposto é :

- n = 500.
- ϵ = (0.5, 1, 2, 250).
- k = (4, 8)

Na Figura 3.12 podemos ver a distribuição do estimador para cada categoria e com o respectivo nível de confidencialidade. Note que, novamente um aumento na anonimização da nossa base de dados (ϵ menor), causa uma distorção nas estimativas e inflacionando sua variância. Conforme flexibilizamos a confidenciabilidade (aumento de ϵ) as estimativas possuem menor variância e viés.

Para o caso de $k = 8$, temos resultados semelhantes, como podem ser vistos na Figura 3.13. Com isso, podemos concluir que o parâmetro de privacidade tem de ser escolhido com cautela pelo pesquisador, quanto maior anonimização requerida, pior será a qualidade inferencial. Outro fato interessante é que não há grande perda inferencial se temos um nível superior à 2 de ϵ, observe que as estimativas possuem bom desempenho se comparado com $\epsilon = 250$.

Quanto a taxa de cobertura, notamos que quanto maior a privacidade na base de dados, ou seja menor o ϵ, pior é o desempenho. No caso onde $\epsilon = 0.5$, tivemos taxas de cobertura menores que 30%. Tal resultado está de acordo com o encontrado para as estimativas, o aumento da privacidade traz muito viés à base de dados sintéticos.

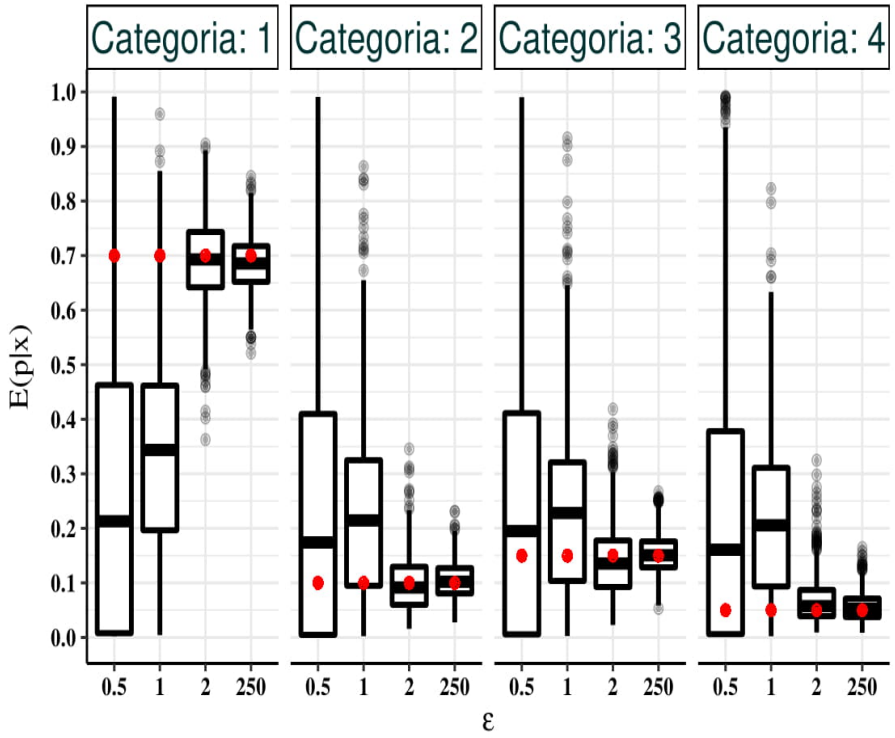

Figura 3.12: Efeito do ϵ para k = 4

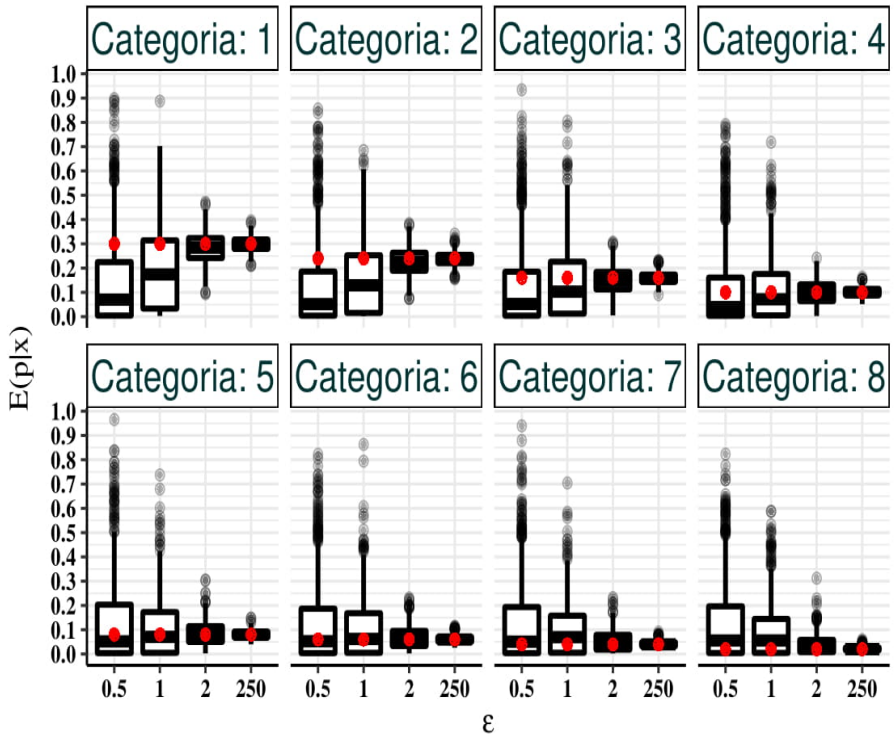

Figura 3.13: Efeito do ϵ para k = 8

Capítulo 4

Aplicação

4.1 Caso

O objetivo de desenvolver métodos de privacidade em bases de dados é para que as agências possam divulgar seus dados sem comprometer a anonimidade dos indivíduos que estão na base. Além disso, os analistas de dados devem ser capazes de fazer inferências próximas às reais.

O orgão governamental intitulado SUSEP, SUperintendência de SEguros Privados (http://www.susep.gov.br/), responsável pelo controle e fiscalização dos mercados de seguro, previdência privada aberta, capitalização e resseguro, possui diversas informações divulgadas ao público em geral, como estatísticas de resgastes pagos a empresas, índice de veículos roubados, estatísticas sobre automóveis, entre outras.

Dentro do site existe um sistema chamado Autoseg, que traz várias informações sobre roubos de carros, colisões, indenizações e etc. Para a consulta, o usuário pode escolher regiões brasileiras (Ex: MG-Met BH e ZN Mata), o modelo de carro que deseja (Ex: Fiat-Palio), o sexo do condutor, faixas etárias e o período em que quer analisar. O resultado é uma tabela contendo informações da frequência de roubos, colisões e a indenização total para cada categoria.

Na Figura 4.1 pode ser visto um exemplo do retorno de uma consulta para o modelo de carro Ford Fusion, na região metropolitana de BH e arredores, no período de 01/01/2015 à 31/12/2015.

Note que um usuário mal intencionado pode inferir algumas informações sobre os indivíduos, como por exemplo a indenização média para cada categoria. Há ainda indivíduos na base com alto risco de *linkage*, como por exemplo a linha 2 da Figura 4.1, onde apenas uma mulher entre 26 e 35 anos foi roubada e consta o valor da indenização.

Visando diminuir o risco associado à esta base de dados, utilizaremos a metodologia de geração de bases sintéticas Multinomial- Dirichlet e utilizaremos o nossa proposta de método inferencial. Consideramos como categorias a combinação do sexo do condutor e a faixa etária.

Sexo Condutor	Faixa Etária	Incêncio e Roubo	Incêncio e Roubo (R$)	Freq. Colisão	Indeniz. Colisão (R$)
Feminino	Entre 18 e 25 anos	0	0	0	0
Feminino	Entre 26 e 35 anos	1	59.561	6	85.574
Feminino	Entre 36 e 45 anos	2	24.985	2	24.105
Feminino	Entre 46 e 55 anos	0	0	10	69.266
Feminino	Maior que 55 anos	0	0	4	35.867
Masculino	Entre 18 e 25 anos	1	54.145	0	0
Masculino	Entre 26 e 35 anos	1	49.630	21	234.763
Masculino	Entre 36 e 45 anos	3	84.078	24	380.592
Masculino	Entre 46 e 55 anos	2	104.490	19	162.538
Masculino	Maior que 55 anos	3	128.177	21	133.617
Totais		**13**	**505.066**	**107**	**1.126.322**

Figura 4.1: Ilustração da tabela fornecida pela SUSEP

4.2 Definições

Utilizamos a mesma base apresentada na ilustração para nossa aplicação. Optamos pela frequência de colisões na zona metropolitana de BH e zona da mata, contendo ao total 107 casos para o modelo de carro Ford Fusion, no período de 01/01/2015 à 31/12/2015. Os dados a serem analisados estão resumidos na Tabela 4.1.

Tabela 4.1: Número de colisões por sexo e idade

	Sexo		
Idade	Masc.	Fem.	Total
$26-35$	21	6	27
$36-45$	24	2	26
$46-55$	19	10	29
55+	21	4	25
Total	85	22	107

Consideramos cada célula da tabela como uma categoria, resultando assim, em uma distribuição $Multinomial(n, \boldsymbol{p})$ para as contagens de colisões, onde $n = 107$ e $\boldsymbol{p}$ possui 8 parâmetros. Para priori, definimos uma $Dirichlet(\boldsymbol{\gamma})$, com $\boldsymbol{\gamma} = 0.5$. Para o nível de privacidade escolhemos $\epsilon = 2$, resultando em $\alpha = 39.37$.

4.3 Inferência Sobre Dados sintéticos

Nesta seção apresentamos os resultados da inferência sobre a base de dados da SUSEP anonimizada, utilizando nosso método. O modelo hierárquico é apresentado novamente abaixo:

$$\mathbf{p} \sim Dirichlet(\boldsymbol{\gamma})$$
$$\mathbf{x} \sim Multinomial(n, \mathbf{p})$$
$$\tilde{\mathbf{p}}_{\mathbf{m}} \sim Dirichlet(\boldsymbol{\alpha} + \mathbf{x})$$
$$\tilde{\mathbf{x}}_{\mathbf{m}} \sim Multinomial(\tilde{n}, \tilde{\mathbf{p}}_{\mathbf{m}}),$$

Para avaliar a inferência sobre os dados sintéticos, replicamos as estimativas em um cenário de 1000 bases sintéticas diferentes.

Para cargo de comparação nas estimativas, apresentamos as Tabelas 4.2 e 4.3 com os valores a posteriori verdadeiros de p para cada uma das categorias e também as estimativas obtidas pelo nosso algoritmo proposto.

Tabela 4.2: Prop de colisões à posteriori

	Sexo	
Idade	Masc.	Fem.
$26-35$	0.1937	0.0585
$36-45$	0.2207	0.0225
$46-55$	0.1757	0.0946
55+	0.1936	0.0405

Tabela 4.3: Estimativas para p

	Sexo	
Idade	Masc.	Fem.
$26-35$	0.1885	0.0573
$36-45$	0.2203	0.0375
$46-55$	0.1679	0.0951
55+	0.1853	0.0480

Note que as estimativas pontuais são realmente boas, é evidente que nosso método aproxima bem os resultados.A seguir mostramos as estimativas do viés e viés relativo, nas Tabelas 4.4 e 4.5:

Tabela 4.4: Estimativas do viés

	Sexo	
Idade	Masc.	Fem.
$26-35$	-0.0052	-0.0012
$36-45$	-0.0004	0.0150
$46-55$	-0.0077	0.0005
55+	-0.0083	0.0075

Tabela 4.5: Estimativas do viés Relativo

	Sexo	
Idade	Masc.	Fem.
$26-35$	-0.0268	-0.0205
$36-45$	-0.0018	0.6667
$46-55$	-0.0443	0.0053
55+	-0.0429	0.1852

Note que, as estimativas do viés apresentadas na Tabela 4.4 são ínfimas, contudo para fazer comparações relevantes, temos que considera o fato de que um erro em categorias com menores contagens é mais impactante. Ou seja, um viés de 0.01 em uma categoria com alto número de contagens é menos impactante do que nas categorias com proporções menores. Portanto, levamos consideração o viés relativo mostrado na Tabela 4.5, e assim podemos notar que o viés relativoé muita maior onde as proporções de contagens são menores.

A Figura 4.2 mostra a densidade das estimativas para cada um dos parâmetros. Observe que nas categorias onde as contagens são mais altas, temos uma distribuição simétrica em torno do valor real do parâmetro a posteriori. Porém, quando as contagens são mais baixas, onde há maior sensibilidade da informação, a distribuição das estimativas

é bem assimétrica. Esse fato ocorre devido à forte garantia de ϵ-DP mascarar muito as informações suscetíveis dos usuários.

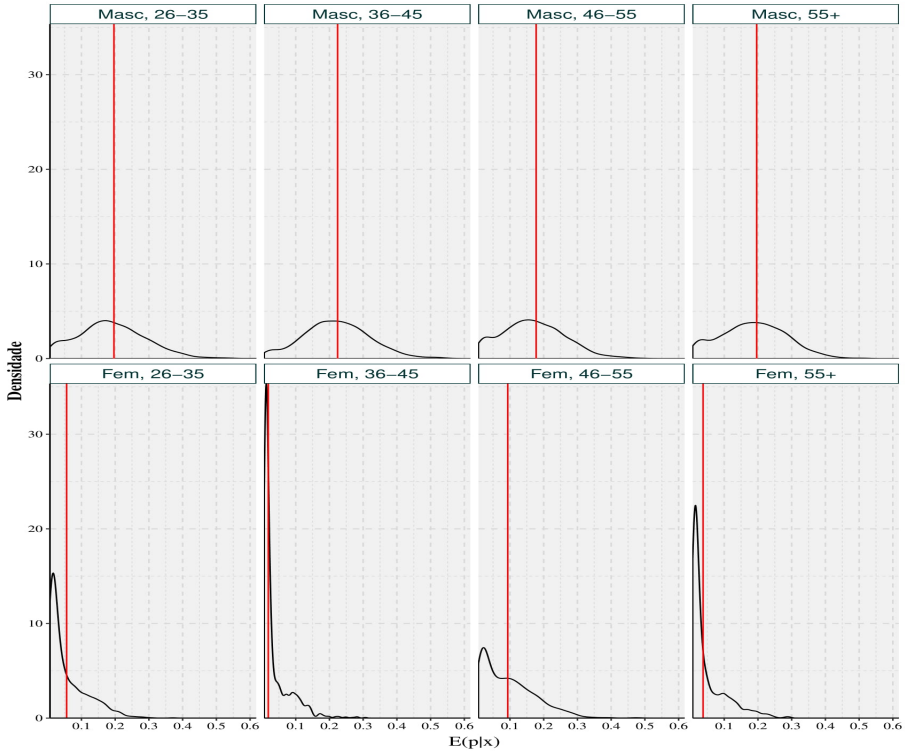

Figura 4.2: Densidade das estimativas para cada categoria

Apesar de algumas distribuições dos estimadores serem assimétricas, o modelo Dirichlet-Multinomial apresenta-se como uma boa técnica para obter estimativas pontuais da amostra sintética sob ϵ-DP da base de dados da SUSEP. Porém sempre temos de avaliar o *tradeoff* entre a qualidade inferencial e a segurança dos usuários.

Quanto à taxa de cobertura das estimativas, considerando um intervalo HPD de 95%, para categorias com maior número de contagens a taxa de cobertura ficou em torno de 70% a 85%. Contudo, para as categorias com baixa contagem, como idade entre 26-35 e 36-45 do sexo feminino, a taxa de cobertura foi superior a 60%.

Capítulo 5

Discussão

A área de *data privacy* tem ganhado muita atenção nos últimos anos, havendo interesse tanto de orgãos governamentais quanto de empresas em manter o sigilo dos indivíduos em suas bases de dados. Nosso objetivo neste trabalho era mostrar técnicas que além de garantir a anonimização destes indivíduos, fornecessem o quão seguro estão. Propomos apresentar métodos de anonimização para dados dicotômicos e categóricos, propostos por Abowd e Vilhuber (2008). Além disso, replicar e complementar o estudo de Charest (2011), sobre inferência em bases sintéticas binárias sob ϵ-*DP*. E por fim, extender o modelo binário para múltiplas categorias e fazer uma aplicação em dados de seguros de carros, fornecidos pela SUSEP.

Quanto à parte dicotômica, Charest (2011) apresenta um método inferencial, via modelo Bayesiano hierárquico, para as bases sintéticas sob ϵ-DP. A autora utiliza o *software* JAGS para obter os resultados de simulações via MCMC, cujo algoritmo se difere do apresentado por ela. Tendo em vista o problema exposto, propomos dois novos métodos de análise inferencial dos dados sintéticos sob ϵ-DP, via *Gibbs Sampling* e *Gibbs* com passo *Metropolis-Hastings*, e comparamos com os resultados gerados pelo JAGS. A *performance* apresentada pelo JAGS foi inferior quanto ao tempo computacional, em relação ao Gibbs com passo Metropolis; em relação as estimativas, os três métodos foram próximos. Dentre os comparados, o algoritmo *Gibbs* com passo *Metropolis-Hastings* foi escolhido para complementar o estudo de Charest (2011).

Outro adendo à publicação de Charest (2011) é que em nenhum momento é feita alguma análise quanto à distribuição do estimador. Tendo em vista esta questão, concentramos esforços em recriar suas simulações e complementá-las com este tipo de análise. Como resultado, observamos que as estimativas pontuais realmente são boas, porém a distribuição das estimativas possui alta variabilidade. Este fato pode estar atrelado à condição de ϵ-DP ser extremamente forte. Lembrando que o método garante aos usuários que mesmo que um *intruder* tenha informação sobre todos os indivíduos da base exceto a dele, ele não será capaz de aprender a respeito das suas informações. Para cumprir tal promessa, é introduzido um alto viés à amostra.

Notamos também que as estimativas são afetadas por várias questões. O número de bases sintéticas liberadas, quanto mais bases liberadas, maior viés. Quanto ao nível de privacidade, quando mais queremos proteger a informação dos usuários, pior serão as estimativas. E também outro fator é o desbalanceamento entre as classes e o tamanho

amostral, caso haja um grande desbalanceamente entre as classes, e ainda a amostra seja pequena, notamos que há um maior viés nas estimativas.

Em relação à geração e análise de bases sintéticas sob ϵ-DP com múltiplas categorias, além das conclusões gerais quanto ao nível de privacidade, desbalanceamento entre classes e qualidade da inferência, semelhantes ao modelo dicotômico, notamos também que a dimensionalidade do problema causa grande impacto nas estimativas, sendo esse um grande porém no método, pois na era do *Big data*, é comum trabalharmos com dimensionalidade alta.

Quanto à aplicação dos métodos nos dados sobre colisões de carros separados por região, categorizados em sexo e idade, fornecidos pela SUSEP, concluimos que as técnicas de anonimização e inferência via modelo Bayesiano hierárquico se apresentam como boa alternativa para garantir a segurança do usuário e fornecer estimativas precisas aos analistas.

Em geral, os métodos de DIPS apresentados se mostraram eficientes para estimativas pontuais, porém os estimadores possuem grande variabilidade e em alguns casos viés alto. Portanto, é ainda necessário desenvolver alternativas a estes métodos. Atualmente existem pesquisas voltadas para sintetizadores que garantam um relaxamento ao método de ϵ-DP. Outra esfera trabalha com outros tipos de sintetizadores e tem mostrado resultados promissores, como o trabalho de Hardt et al. (2012). Porém, métodos de anonimização em bases de dados seguem sendo um grande desafio.

Para nossos trabalhos futuros, temos interesse em estudar métodos de anonimização no âmbito multivariado, pois até aqui apresentamos técnicas que lidam com o problema de anonimização apenas para uma variável. Além disso, avaliar a qualidade da anonimização e inferência em problemas que relaxem a garantia de ϵ-DP.

Referências Bibliográficas

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., e Zhang, L. (2016), "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 308–318, ACM.

Abowd, J. M. e Vilhuber, L. (2008), "How protective are synthetic data?" in *International Conference on Privacy in Statistical Databases*, pp. 239–246, Springer.

Bowen, C. e Liu, F. (2016), "Comparative study of differentially private data synthesis methods." .

Charest, A.-S. (2011), "How can we analyze differentially-private synthetic datasets?" *Journal of Privacy and Confidentiality*, 2, 3.

Chaudhuri, K., Sarwate, A., e Sinha, K. (2012), "Near-optimal differentially private principal components," in *Advances in Neural Information Processing Systems*, pp. 989–997.

Cox, L. H. (1980), "Suppression methodology and statistical disclosure control," *Journal of the American Statistical Association*, 75, 377–385.

Dalenius, T. e Reiss, S. P. (1982), "Data-swapping: A technique for disclosure control," *Journal of statistical planning and inference*, 6, 73–85.

Duncan, G. T. e Lambert, D. (1986), "Disclosure-limited data dissemination," *Journal of the American statistical association*, 81, 10–18.

Dwork, C. (2008), "Differential privacy: A survey of results," in *International Conference on Theory and Applications of Models of Computation*, pp. 1–19, Springer.

Dwork, C. e Roth, A. (2013), "The algorithmic foundations of differential privacy," *Theoretical Computer Science*, 9, 1–227.

Dwork, C., McSherry, F., Nissim, K., e Smith, A. (2006), "Calibrating noise to sensitivity in private data analysis," Springer.

Eddelbuettel, D. e François, R. (2011), "Rcpp: Seamless R and C++ Integration," *Journal of Statistical Software*, 40, 1–18.

El Emam, K., Dankar, F. K., Vaillancourt, R., Roffey, T., e Lysyk, M. (2009), "Evaluating the risk of re-identification of patients from hospital prescription records," *The Canadian journal of hospital pharmacy*, 62, 307.

Friedman, A., Berkovsky, S., e Kaafar, M. A. (2016), "A differential privacy framework for matrix factorization recommender systems," *User Modeling and User-Adapted Interaction*, 26, 425–458.

Gelfand, A. E. e Smith, A. F. (1990), "Sampling-based approaches to calculating marginal densities," *Journal of the American statistical association*, 85, 398–409.

Geman, S. e Geman, D. (1984), "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on pattern analysis and machine intelligence*, pp. 721–741.

Haario, H., Saksman, E., e Tamminen, J. (1999), "Adaptive proposal distribution for random walk Metropolis algorithm," *Computational Statistics*, 14, 375–396.

Hardt, M., Ligett, K., e McSherry, F. (2012), "A simple and practical algorithm for differentially private data release," pp. 2339–2347.

Hastings, W. K. (1970), "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, 57, 97–109.

Karr, A. e Reiter, J. (2014), "Using statistics to protect privacy," *Lane J.et al. (eds.) Privacy, Big Data, and the Public Good Frameworks for Engagement*.

Lambert, D. (1993), "Measures of disclosure risk and harm," *Journal of Official Statistics*, 9, 313.

Li, K.-H., Raghunathan, T. E., e Rubin, D. B. (1991), "Large-sample significance levels from multiply imputed data using moment-based statistics and an F reference distribution," *Journal of the American Statistical Association*, 86, 1065–1073.

McClure, D. e Reiter, J. P. (2012), "Differential Privacy and Statistical Disclosure Risk Measures: An Investigation with Binary Synthetic Data." .

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., e Teller, E. (1953), "Equation of state calculations by fast computing machines," *The journal of chemical physics*, 21, 1087–1092.

Mohammed, N., Chen, R., Fung, B., e Yu, P. S. (2011), "Differentially private data release for data mining," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 493–501, ACM.

Narayanan, A. e Shmatikov, V. (2006), "How to break anonymity of the netflix prize dataset," *arXiv preprint cs/0610105*.

Plummer, M. (2015), "JAGS Version 4.0. 0 user manual," .

Plummer, M. (2016), *rjags: Bayesian Graphical Models using MCMC*, R package version 4-6.

Plummer, M. et al. (2003), "JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling," .

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., e Solenberger, P. (2001), "A multivariate technique for multiply imputing missing values using a sequence of regression models," *Survey methodology*, 27, 85–96.

Raghunathan, T. E., Reiter, J. P., e Rubin, D. B. (2003), "Multiple imputation for statistical disclosure limitation," *Journal of official statistics*, 19, 1.

Reiter, J. P. (2005), "Estimating risks of identification disclosure in microdata," *Journal of the American Statistical Association*, 100, 1103–1112.

Reiter, J. P. (2012), "Statistical approaches to protecting confidentiality for microdata and their effects on the quality of statistical inferences," *Public opinion quarterly*, 76, 163–181.

Rubin, D. B. (1987), "Multiple Imputation for Nonresponse in Surveys (Wiley Series in Probability and Statistics)," .

Rubin, D. B. (1993), "Statistical disclosure limitation," *Journal of official Statistics*, 9, 461–468.

Sullivan, G. R. (1989), "The use of added error to avoid disclosure in microdata releases," *Digital Repository Iowa State University*.

Tendick, P. (1991), "Optimal noise addition for preserving confidentiality in multivariate data," *Journal of Statistical Planning and Inference*, 27, 341–353.