UNIVERSIDADE FEDERAL DE JUIZ DE FORA
INSTITUTO DE CIÊNCIAS EXATAS
DEPARTAMENTO DE ESTATÍSTICA

Filipe Oliveira Fernandes

AVALIAÇÃO DA EFICIÊNCIA DE PLANOS AMOSTRAIS EM BIG DATA DE REGISTROS ADMINISTRATIVOS

Juiz de Fora
2018

Filipe Oliveira Fernandes

AVALIAÇÃO DA EFICIÊNCIA DE PLANOS AMOSTRAIS EM BIG DATA DE REGISTROS ADMINISTRATIVOS

Monografia apresentada ao Curso de Estatística, da Universidade Federal de Juiz de Fora como requisito parcial a obtenção do título de Bacharel em Estatística.

Orientador: Prof. Marcel de Toledo Vieira, Ph.D.

Juiz de Fora
2018

Ficha catalográfica elaborada através do programa de geração
automática da Biblioteca Universitária da UFJF,
com os dados fornecidos pelo(a) autor(a)

Fernandes, Filipe Oliveira.
Avaliação de eficiência de planos amostrais em Big Data de registros administrativos / Filipe Oliveira Fernandes. -- 2018.
46 f. : il.

Orientador: Marcel de Toledo Vieira
Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Juiz de Fora, Instituto de Ciências Exatas, 2018.

1. Amostragem. 2. CadÚnico. 3. Big Data. I. Vieira, Marcel de Toledo , orient. II. Título.

Filipe Oliveira Fernandes

AVALIAÇÃO DA EFICIÊNCIA DE PLANOS AMOSTRAIS EM BIG DATA DE REGISTROS ADMINISTRATIVOS

Monografia apresentada ao Curso de Estatística, da Universidade Federal de Juiz de Fora como requisito parcial a obtenção do título de Bacharel em Estatística.

Aprovada em 3 de dezembro de 2018

BANCA EXAMINADORA

_____

Prof. Marcel de Toledo Vieira, Ph.D. - Orientador

Universidade Federal de Juiz de Fora

_____

Prof. Augusto Carvalho Souza, D.Sc.

Universidade Federal de Juiz de Fora

_____

Prof. Ricardo Freguglia, D.Sc.

Universidade Federal de Juiz de Fora

Dedicado a São Filipe Apóstolo e a todos aqueles que encontrei na Universidade.

AGRADECIMENTOS

Agradeço primeiro a Deus e pela intercessão de Nossa Senhora, que me afastou da angústia e renovou as minhas esperanças.

Ao meu orientador Marcel pela paciência e pela confiança que teve em mim durante todo esse tempo.

À Faculdade de Economia, por ter cedido o servidor e ter tornado esse trabalho possível.

Aos meus colegas estatísticos que me fizeram companhia nessa aventura durante todos esses anos.

Aos integrantes do GOU, pela acolhida e por ser essa corrente de graça.

Ao povo do forró, obrigado pela dança e por ter realizado o meu sonho de fazer uma faculdade muito mais alegre.

E com carinho especial a todos que encontrei, seja numa disciplina eletiva ou numa conversa rápida, mas que compartilharam comigo o sonho de estar na Universidade.

“No final, o que importa são as pessoas”

(Autor desconhecido)

“O Senhor é meu pastor, nada me faltará”

(Salmo 23)

RESUMO

As novas tecnologias do século XXI propiciaram um grande aumento na produção de dados, o que ocasionou no fenômeno conhecido como *Big Data*. Surgem bancos de dados cada vez mais complexos e difíceis de serem analisados, que requerem uma nova infraestrutura e capacidade maior de processamento computacional. Uma solução para esse problema seria o uso de técnicas de amostragem probabilística. A proposta deste estudo foi a realização de um estudo de simulação considerando diferentes desenhos amostrais através do *software* R e com o auxílio de um servidor. Planos amostrais simples e complexos foram comparados quanto ao erro padrão e nível de cobertura, na finalidade de obter estimativas com as melhores propriedades a partir de tamanhos de amostra reduzidos, a fim de possibilitar a análise dos dados também ao usuário de um *desktop*. Nesse estudo foram considerados dados do *CadUnico* (Cadastro Único para Programas Sociais do Governo), que possui mais de 20 milhões de registros e distingue-se como a plataforma de acesso ao Bolsa Família. A Amostragem Aleatória Simples destacou-se como o método mais rápido e também o mais preciso inclusive nos menores tamanhos de amostra.

Palavras-chave: Amostragem. *CadÚnico*. *Big Data*.

ABSTRACT

New technologies of the XXI century provided a great increase in the production of data, which caused the phenomenon known as Big Data. Increasingly complex and difficult-to-analyse databases are emerging that require new infrastructure and increased computational processing power. A solution to this problem would be the use of probabilistic sampling techniques. The purpose of this study was to carry out a simulation study considering different sample designs through the software R and with the assistance of a server. Simple and complex sample designs were compared based on the standard error and coverage level in order to obtain estimates with the best properties from reduced sample sizes in order to allow data analysis also to the user of a desktop. In this study, data from the *CadÚnico* (Single Register for Social Programs of the Government), which has more than 20 million records and is distinguished as the platform for access to Bolsa Família, were considered. Simple Random Sampling has stood out as the fastest and most accurate method even in the smallest sample sizes.

Keywords: Sampling. *CadÚnico*. *Big Data.*

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

$IBGE$	Instituto Brasileiro de Geografia e Estatística
ISI	*International Statistical Institute*
$Cad\acute{U}nico$	Cadastro Único para Programas Sociais do Governo Federal
AAS	Amostragem Aleatória Simples
TLC	Teorema do Limite Central
AES	Amostragem Estratificada Simples
AES_u	Amostragem Estratificada Simples com Alocação Uniforme
AES_p	Amostragem Estratificada Simples com Alocação Proporcional
Var	Variância
EP	Erro Padrão
CV	Coeficiente de Variação
NIS	Número de Identificação Social
$CRAS$	Centro de Referência de Assistência Social

SUMÁRIO

1. INTRODUÇÃO

Vivemos em uma era em que nos está disponível uma quantidade cada vez maior de informações, dados são gerados a todo momento devido principalmente a fatores como a grande evolução que a Internet e que outros sistemas digitais tiveram nos últimos anos. O grande volume de dados atualmente é um desafio para as técnicas de análise e armazenamento convencionais, caracterizando o chamado fenômeno “Big Data”, um dos assuntos mais debatidos na ciência nos últimos anos.

Os desafios do “Big Data” não consistem somente em trabalhar com grandes bases de dados. Existem também questões também acerca da velocidade com que são processadas todas essas informações, colocando em cheque o poder dos computadores que estão disponíveis hoje em dia. Outro aspecto muito importante que também é colocado em pauta é a questão da qualidade desses dados, um grande volume de dados não implica necessariamente que todas eles sejam uteis e bons para aproveitamento.

As informações que formam os bancos de dados do “Big Data” são majoritariamente coletadas através de dispositivos eletrônicos e de maneira automática, dando ampla margem a inclusão de dados incorretos nos registros (HAND, 2016). A limpeza e tratamento desses dados se fazem então necessários e são importantes nesse estudo.

Os registros administrativos são então uma das fontes dos grandes bancos de dados atuais. Define-se como registros administrativos os dados que não foram coletados especificamente para estudos acadêmicos. Geralmente, são registros que são usados no cotidiano de empresas e outras instituições públicas.

Por não passar por um crivo de uma pesquisa científica, os registros administrativos não seguem uma regularidade metodológica na forma de serem coletados, ficando a critério das necessidades da organização detentora dos dados. As mudanças repentinas na forma dos dados podem gerar várias ambiguidades para quem queira fazer uma análise estatística dos registros.

O pesquisador ao fazer inferência sobre esse tipo de dados deve se ater a uma série de precauções, como definir bem quem será a população a ser estudada e observar se não há registros duplos ou inválidos, o que pode enviesar as suas estimativas.

Nesse contexto as técnicas de amostragem ganham relevância, se afastando do senso comum que é necessário usar as informações de todos os indivíduos presentes nos arquivos para se chegar a bons resultados.

Uma das principais vantagens do uso de métodos de amostragem com relação ao uso de censo em Big Data é a economia de recursos computacionais. O tamanho da amostra é menor que o da população, podendo gerar resultados precisos de maneira consideravelmente mais rápida.

Existem vários métodos de amostragem na literatura, mas nem todos se fundamentam em princípios estatísticos. A validade dos resultados obtidos fica a cargo da escolha de métodos probabilísticos. Os métodos se diferenciam também em sua complexidade, amostras com mesmo tamanho podem gerar estimativas mais precisas que outras devido ao seu plano amostral.

O objetivo desta monografia é avaliar a eficiência de dois diferentes planos amostrais probabilísticos, envolvendo dados amostrais simples e complexos, a partir da utilização de um grande banco de dados de registros administrativos, o CadÚnico do ano de 2015.

O CadÚnico é um cadastro do governo federal criado pelo Ministério do Desenvolvimento Social em 2001. Possui a finalidade de registrar toda a população de baixa renda do país e pode ser considerado o principal meio de acesso a vários programas sociais. O banco de dados do CadÚnico domiciliar do ano de 2015 possui mais de 20 milhões de famílias cadastradas.

Os métodos de amostragem considerados foram a Amostragem Aleatória Simples e a Amostragem Estratificada Simples, com dois tipos diferentes de locação. Em cada método foram considerados também quatro tamanhos de amostras diferentes (frações amostrais de 0,1%, 0,5%, 5% e 10% do total da população). Para avaliarmos os métodos foi realizado um estudo de simulação, com um auxílio de um servidor e implementado em *software* R, em que foram selecionadas 1.000 amostras segundo cada plano amostral, totalizando 12.000 amostras distintas.

A avaliação dos métodos consistiu em estimar as médias para algumas das principais variáveis do CadÚnico: renda, número de pessoas presentes em cada domicilio e em cada família, despesas com energia, água e esgoto, gás e alimentação das famílias relacionadas no cadastro. Para analisarmos a exatidão e a precisão das estimativas também foram calculados o Erro Padrão, o Coeficiente de Variação e o Nível de Cobertura para essas médias. Com a finalidade ser avaliado também o custo computacional e a complexidade de cada método também fora levado em consideração o tempo necessário para a obtenção de cada amostra.

A monografia é estruturada em 5 capítulos. Após o primeiro em que é feita apresentação do estudo, o segundo contém três seções, cada uma destinada a fazer uma breve revisão literária dos principais temas relacionados a este trabalho, que são a amostragem, a

questão do Big Data e dos registros administrativos; neste capítulo também há uma seção específica sobre o CadÚnico. No terceiro capítulo é apresentada a metodologia empregada na pesquisa, onde são detalhados os planos amostrais utilizados nos estudos de simulação. No quarto capítulo são abordadas as características do banco de dados e as técnicas de limpeza aplicadas, em seguida são apresentados os resultados das simulações. No quinto capítulo é exibida a conclusão do estudo.

2. BIG DATA DE REGISTROS ADMINISTRATIVOS

2.1 Amostragem

A amostragem é a ação de se selecionar uma amostra, ato de escolher unidades de uma população e através dos resultados dessa amostra conseguir obter informações sobre o todo (BUSSAB, 2005), ou seja, fazer inferência estatística.

No contexto em que estamos vivendo, em que as novas tecnologias impulsam a geração cada vez maior de banco de dados, a amostragem mostra-se relevante, com condições de gerar resultados com qualidade. Utilizando-se de uma amostra não se faz necessária a investigação de uma população inteira, diminuindo os custos, tempo de pesquisa, assim como o número de indivíduos pesquisados. Muitas pesquisas não seriam viáveis economicamente sem a utilização de técnicas de amostragem.

As técnicas de amostragem fundamentam-se como teoria estatística principalmente com Kiaer (1895), a partir do século XIX, que demostrou empiricamente que poderiam ser estimadas médias e totais de populações finitas a partir de pequenas amostras. Em 1925, o ISI (Instituto Internacional de Estatística) reconhece oficialmente a amostragem como método científico (VIEIRA, 2017).

A capacidade de gerar resultados precisos em amostragem fica a cargo do método utilizado. Existem os métodos probabilísticos e os não-probabilísticos (NEYMAN, 1934). Os métodos probabilísticos são aqueles em que cada indivíduo possui probabilidade positiva de ser selecionado para a amostra, métodos que possuem aporte teórico, sendo capaz de se realizar inferência sobre a população; já os métodos não-probabilísticos são aqueles que não possuem esse aporte, sendo incapaz de apresentar resultados confiáveis, estando presente na literatura pelo hábito de sua utilização e por possuir os menores custos, sendo chamados de “métodos artísticos” por alguns autores (NEYMAN, 1934).

2.2 *Big Data*

"*Big data*" é um termo originário do inglês e significa "grandes bases de dados" ou "grandes bancos de dados", há analogias também que fazem referência a expressão "tempestade de dados"; pode ser considerado o termo mais discutido em Ciência da Computação e Estatística na última década. Surge junto a termos como "*data cloud*" (computação em nuvem), "*data warehouse*" (armazém de dados), "*business analytics*" (análise de negócios) entre outros, que fazem referência principalmente à 3ª época da era da informação (MINELI, CHAMBERS E DHIRAJ, 2013), momento em que há uma grande quantidade de dados disponíveis devido ao avanço dos sistemas digitais, fenômeno que pode ser mais notado na evolução da Internet, da computação móvel (smartphones e tablets) e popularidade das redes sociais.

A expressão "*Big Data*" foi citada pela primeira vez no relatório "*Data, data, everywhere: a special report on managing information*", do periódico britânico *The Economist*. Nasce nas áreas de Astronomia e Genética, ciências que primeiro perceberam os desafios em trabalhar com gigantescas quantidades de dados, principalmente por causa da gama de informações fornecidas pelo telescópio Hubble e no estudo do genoma humano (CUKIER, 2010).

Uma questão importante em torno desse excesso de dados é o seu armazenamento e sua análise; se faz cada vez mais necessários computadores mais robustos para trabalhar com todas essas informações. Hoje estima-se que haja no mundo uma quantidade em torno de 10.000 *exabytes* de dados digitais (LETOUZÉ, 2012). O que nos instiga a recordarmos a Lei de Moore, que prevê que o poder de processamento dos computadores seguiria uma linearidade e duplicaria a cada 18 meses, lei que pode ser superada com a utilização de novos materiais como o grafeno.

Bernard Marr (2014) elaborou um conceito que foi amplamente difundido na mídia e no ambiente empresarial que são os "5 v's do *Big Data"*:

a) Volume, que diz respeito a grande quantidade de dados, em que os arquivos manipulados estão na ordem dos *gigabytes* de tamanho;
b) Velocidade, o tempo que se demora para processar esses dados pode ser muito grande sendo um fator fundamental;
c) Variedade, os dados a serem manipulados podem ter diversas origens (vídeo clips, gravações de voz, documentos de repositórios, logs na web, dados sociais, dados abertos do governo entre outros) e conter todo tipo de informação;

d) Veracidade, as decisões que são tomadas a partir da análise desses dados necessitam que sejam condizentes com a realidade e;
e) Valor, todo o investimento em estrutura de *Big Data* deve ser compensado.

Grandes empresas como Google e Facebook já reconhecem o valor do Big Data e destinam setores relevantes de suas companhias a inteligência corporativa (NESELLO, P.; FACHINELLI, A.C.,2014), constituindo as suas receitas baseadas na análise dos seus bancos de dados, gerando riquezas com o uso das próprias informações que os seus clientes fornecem.

Fica cada vez mais nítido para a sociedade o valor agregado desses grandes bancos de dados. Cresce a estima não só dos empresários, mas do meio acadêmico pelo uso desse tipo de informação. Já pode ser notado um interesse em *Big Data*, apesar de ainda pequeno, em novas áreas como Saúde e Ciências Agrárias (NESELLO P.; FACHINELLI A.C., 2014), caracterizando esse fenômeno como próprio do ser humano na sua contínua busca por querer medir, registrar e analisar o mundo (SCHÖNBERGERMAYER V.; CUKIER K., 2013)

2.3 Registros Administrativos

Em Estatística entende-se registros administrativos como dados pertencentes a uma organização, empresa ou órgão governamental, que não foram colhidos para fins de pesquisa, mas para processos internos da instituição. Uma definição mais abrangente do termo foi dada pela Organização para Cooperação e Desenvolvimento Econômico que considera registros administrativos como dados que foram coletados para um propósito não estatístico, em que o agente que fornece os dados e a unidade à qual os dados se referem são distintos, e os métodos empregados nessa coleta são determinados pelo próprio órgão administrativo e visa abranger toda a população alvo (OCDE, 2016).

A discussão em torno desse tipo de registro tem ganhado força pois eles refletem todo o potencial de análise em Big Data. A grande quantidade de dados disponíveis devido a nova era da informação se dá pela descoberta dessas novas fontes alternativas de dados. Fontes essas muito mais baratas que os usuais censos e levantamentos encomendados, fazendo com que o próprio Instituto Brasileiro de Geografia e Estatística (IBGE) estudasse extinguir o censo como conhecemos hoje para calcular estimativas para a população brasileira através da análise de Big Data (SENRA, 2016).

Vários benefícios podem ser abordados quando se opta por trabalhar com esse tipo de dados, dentre outros pode-se levantar (HAND, 2016):

a) a própria questão da diminuição de custos da pesquisa, em que se economiza por não precisar mais pagar pela coleta dos dados sendo que já estão disponíveis previamente;
b) a maioria dos registros possuem informações de todos os indivíduos relativos aquele órgão ou empresa, se assemelhando portando a um censo populacional;
c) a qualidade desses dados, aspecto que gera dúvidas, mas se tratando de informações de uma empresa por exemplo, a qualidade dessa informações implicaria diretamente na gestão e nos lucros da instituição;
d) a veracidade dos registros administrativos; a obtenção dos mesmos dados através de questionários tradicionais, por exemplo, daria margem a dissimulação do entrevistado, levando a mentir ou evitando responder certas perguntas, problema que não há se tratando de registros administrativos, onde a coleta costuma se dá de maneira indireta ou automática.

Em contraposição a esses benefícios podem ser levantados também aspectos negativos quanto ao uso desse tipo de dados (HAND, 2016):

a) registros administrativos, especialmente aqueles relacionados ao governo e à política pública constantemente estão sujeitos a alguma espécie de regulamentação e legislação;
b) a dificuldade que os pesquisadores encontram para fazer análises longitudinais; como as empresas não possuem esses dados para fins científicos, há recorrentes mudanças no método de coleta, havendo a interrupção da série histórica;
c) a última questão relaciona-se estritamente com outra característica negativa que é a fusão errônea de banco de dados. Bases que são geradas da união de outras menores podem conter graves erros de incompatibilidade, um exemplo clássico é a tentativa de formar um banco de dados nacional da união de bancos estaduais que foram coletados de maneiras distintas;
d) a questão da confidencialidade, privacidade e anonimato de registros administrativos; que geralmente possuem variáveis que podem identificar o indivíduo (ao nível do CPF), e revelar outros informações sigilosas, como renda, saldo da conta corrente e gastos pessoais; mais comum em banco de dados de financeiras e grandes empresas;
e) a questão da limpeza do banco de dados pode ser considerado o principal aspecto negativo com relação a esse tipo de dados; como visto anteriormente,

> registros administrativos são coletados na maioria das vezes de maneira automática, informações erradas podem ser acrescentadas aos registros sem critério e julgamento de um analista. Problema que pode se tornar grave na era do Big Data, dependendo do método de limpeza empregado pelo pesquisado, pode haver a retirada de dados bons e a subestimação da variância dependendo da imputação de dados.

Os registros administrativos diferem-se daqueles coletados para uma pesquisa por serem um banco de dados estritamente operacional, são usados fundamentalmente para análises exploratórias nas empresas. A tentativa de se fazer inferência com esses dados deve ser vista com cautela, com a definição precisa da população a ser estudada.

2.4 *CadÚnico*

O Cadastro Único para Programas Sociais do Governo Federal, mais conhecido como CadÚnico, é um banco de dados nacional voltado para as famílias de baixa renda do país; foi instituído através do decreto 3.877 de julho de 2001 durante o governo de Fernando Henrique (BRASIL, 2001) e posteriormente sofreu alterações e foi regulamentado através do decreto 6.135 de 27 de junho de 2007 (BRASIL, 2007), já no governo Lula. Desde de 2003 tem sido considerado o principal instrumento que o governo utiliza para reunir informações de todas as famílias que se encontram em estado de pobreza e extrema pobreza no Brasil (MINISTÉRIO DO DESENVOLVIMENTO SOCIAL, 2015).

> Art. 2º O Cadastro Único para Programas Sociais - CadÚnico é instrumento de identificação e caracterização sócio-econômica das famílias brasileiras de baixa renda, a ser obrigatoriamente utilizado para seleção de beneficiários e integração de programas sociais do Governo Federal voltados ao atendimento desse público (BRASIL, 2007).

Estão registradas no CadÚnico as famílias que ganham até 3 salários mínimos de renda mensal total ou até meio salário mínimo de renda familiar mensal *per capita*. O cadastro das famílias é de responsabilidade dos governos municipais e são processados pelo Agente Operador do Cadastro Único (Caixa Econômica Federal) que fica incumbida de atribuir a cada pessoa da família cadastrada um número de identificação social (NIS) de caráter único, pessoal e intransferível (BRASIL, 2010). A inscrição das famílias é feita nos Centros de Referência de Assistência Social (CRAS). O Ministério de Desenvolvimento Social utiliza desde cadastro para a seleção dos beneficiários de programas sociais, como o

Bolsa Família, Minha Casa Minha Vida, Cisternas, Passe Livre, dentre outros (BRASIL, 2007). É proibida a utilização do CadÚnico para outros fins que não seja a concepção de políticas públicas e o uso de seus dados para pesquisas no meio acadêmico.

> Art. 8º Os dados de identificação das famílias do CadÚnico são sigilosos e somente poderão ser utilizados para as seguintes finalidades:
> I - formulação e gestão de políticas públicas; e
> II - realização de estudos e pesquisas. (BRASIL, 2007)

A utilização do cadastro não ocorre de maneira exclusiva pelo governo federal, o CadÚnico foi concebido para uso descentralizado, tendo participação tanto das esferas municipal, estadual e pela União. Cabe a cada governo a gerencia de seus projetos sociais e análise socioeconômica das famílias cadastradas.

> Art 8º § 2º A União, os Estados, os Municípios e o Distrito Federal poderão utilizar suas respectivas bases para formulação e gestão de políticas públicas no âmbito de sua jurisdição (BRASIL, 2007).

A validade dos registros constantes no CadÚnico será de 2 anos, sendo que as principais informações relacionadas em cada cadastro são características do domicilio, renda de cada integrante da família, qualificação escolar e profissional, e despesas familiares. (BRASIL, 2010)

3 MÉTODOS DE AMOSTRAGEM

3.1 Amostragem Aleatória Simples

A Amostragem Aleatória Simples (AAS) é considerado o método mais intuitivo e o mais elementar. Como o próprio nome já revela, consiste em selecionar aleatoriamente indivíduos de uma população em que cada um possui uma probabilidade igual de ser escolhido (VIEIRA, 2017). Esse método de amostragem subdivide-se em Amostragem Aleatória Simples com Reposição (AAS_C) e Amostragem Aleatória Simples sem Reposição (AAS_s), a diferença é a possibilidade ou não de um indivíduo ser selecionado mais de uma vez para amostra; no caso de AAS_C, após ser sorteado ele é reposto na população. A AAS_C é costumeiramente mais considerada no momento da análise dos dados, pois por ocasionar independência entre as observações, facilita os cálculos estatísticos e matemáticos.

Esse tipo de amostragem distingue-se por ser o mais simples, servindo de referência para o cálculo do efeito do plano amostral e podendo ser utilizado combinado a outros métodos.

Em seguida são apresentados os estimadores da média ($\bar{y}$) e variância populacional (S^2), ambos não enviesados:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

$$Var(\bar{y}) = \frac{\sigma^2}{n}$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{y})^2$$

Em que define-se Y como a variável aleatória de interesse; n o tamanho da amostra e σ^2 a variância populacional.

3.2 Amostragem Estratificada

No caso da Amostragem Estratificada Simples (AES) a população é dividida em subgrupos, mutuamente exclusivos e exaustivos (VIEIRA, 2017), antes da seleção aleatória dos indivíduos para amostra. A definição desses grupos (estratos) é feita de acordo com a variável de interesse, com intuito sempre de formar grupos internamente mais homogêneos, o que proporciona a diminuição do erro amostral.

O método por estratificação pode ser considerado, em muitas ocasiões, mais vantajoso que a AAS. A separação em grupos homogêneos reduz o tamanho da amostra tanto quanto permite a inferência para cada um desses grupos (BUSSAB, 2005). Esses fatores fazem com que a Amostragem Estratificada seja mais empregada por inclusive conseguir reduzir os gastos de grandes pesquisas.

A seguir são apresentados estimadores não enviesados para a média ($\bar{y}_{es}$) e variância populacional (S^2) segundo o método da Amostragem Estratificada:

$$\bar{y}_{es} = \sum_{h=1}^{H} W_h \bar{Y}_h$$

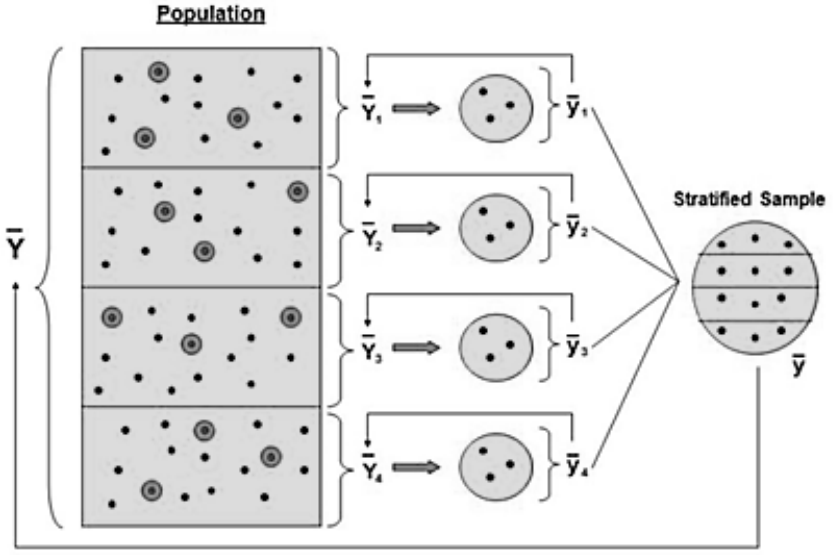
$$Var(\bar{y}_{es}) = \sum_{h=1}^{H} W_h^2 \frac{\sigma_h^2}{n_h^2}$$

$$S^2 = \sum_{h=1}^{H} \frac{N_h - 1}{N - 1} S_h^2 + \sum_{h=1}^{H} \frac{N_h}{N - 1} (\bar{Y}_h - \mu)^2$$

Sendo que H corresponde ao número de estratos e $W_h = \frac{N_h}{N}$, em que N_h é o tamanho do estrato h.

A Figura 1 abaixo ilustra o método por estratificação:

Figura 1 – Amostragem Estratificada



Fonte: VIEIRA,2017

A Amostragem Estratificada distingue-se por possuir mais de uma forma de distribuir o número de indivíduos amostrados por cada estrato, o que é chamado de alocação.

Nesse estudo foram considerados dois tipos de alocação, a Alocação Uniforme e a Alocação Proporcional:

- Alocação Uniforme: neste método a amostra é distribuída igualmente entre os estratos

$$n_h = \frac{n}{H}$$

- Alocação Proporcional: a amostra é dividida entre os estratos de maneira proporcional ao seu tamanho

$$n_h = n\frac{N_h}{N}$$

3.3 Tamanho da Amostra

Umas das primeiras questões a serem levantadas ao se trabalhar com amostragem é com relação ao tamanho da amostra. O tamanho que a amostra irá ter relaciona-se não somente com o tamanho da população, mas também com a variabilidade presente nos dados e o erro amostral que se está disposto a assumir.

Define-se B (erro) como a semi-amplitude do intervalo de confiança para a média populacional (VIEIRA, 2017).

$$P(|\bar{y} - \bar{Y}| \leq B) \cong 1 - \alpha$$

Em que α é o nível de significância adotado.

De acordo com o Teorema do Limite Central (TLC), podemos assumir que a média amostral possua distribuição normal (z).

$$P(|\bar{y} - \bar{Y}| \leq z_{\alpha} \sqrt{\frac{S^2}{n_0}})$$

Então, o tamanho da amostra para Amostragem Aleatória Simples com Reposição pode ser definido como:

$$n_0 = \frac{z_{\alpha}^2 S^2}{B^2}$$

Neste trabalho foram considerados tamanhos de amostras definidos a partir das frações de 0,5%, 0,1%, 5% e 10% do total da população. Através dos estudos de simulação com 1.000 amostras foram estimados os valores de B para os tamanhos de amostra considerados. Estes resultados confirmaram o TLC.

4 ESTUDOS DE SIMULAÇÃO

4.1 Banco de Dados

O banco de dados utilizado nesse estudo foi o CadÚnico domiciliar do ano de 2015. O CadÚnico em questão possui no total 27.192.314 de registros e 52 variáveis que medem diversas características socioeconômicas e identificam cada domicilio. Das diversas variáveis presentes no cadastro foram selecionadas oito consideradas principais e de maior relevância para o estudo. A tabela 1 apresenta as variáveis utilizadas no CadÚnico 2015.

Tabela 1 – Variáveis utilizadas no CadÚnico 2015

Variavel	**Descrição**
codIBGE	Código de identificação do IBGE
Renda	Renda média familiar
Pessoas	Número de pessoas no domicílio
Famílias	Número de famílias no domicílio
Energia	Valor das despesas com energia elétrica
Água/Esgoto	Valor das despesas com água e esgoto
Gás	Valor das despesas com gás
Alimentação	Valor das despesas com alimentação

Fonte: ELABORADO PELO PRÓPRIO AUTOR

Entre as variáveis que não foram utilizadas neste estudo, muitas foram descartadas por não possuir aplicação neste estudo ou apresentar inconsistências nos dados, o que foi percebido pelo excesso de valores nulos ou atípicos.

Em banco de dados muitos grandes como o CadÚnico, é comum que estejam presentes muitos valores discrepantes, *outliers*, e valores faltantes, *missing*, como descrito no Capítulo 2. A estratégia empregada para limpar o banco de dados neste estudo foi a eliminação dos registros que apresentavam valores acima de determinados quantis. A eliminação era feita de maneira progressiva, começando pela variável de maior coeficiente de variação até a de menor valor. Os quantis considerados foram de 99.9%, 99,8% e 99,7%. Através da observação do valor da variância e do número de *missing* restantes, optou-se pela limpeza do banco de dados pelo percentil 99,8%. A tabela 2 apresenta o banco de dados antes da limpeza, em que $\bar{Y}$ e σ^2 representam a média e variância populacionais, respectivamente.

Tabela 2 – Valores do CadÚnico 2015 completo (parâmetros)

Variável	**Observações válidas**	***Missing***	$\overline{Y}$	σ^2	σ
Renda	26.556.332	635.982	184,20	150.039,76	387,35
Pessoas	24.429.813	2.762.501	3,22	3,34	1,83
Famílias	24.432.085	2.760.229	1,05	0,57	0,76
Energia	26.780.521	411.793	39,79	79.284,75	281,58
Água/Esgoto	26.316.616	875.698	19,52	37.718,00	194,21
Gás	26.872.648	319.666	32,36	13.888,38	117,85
Alimentação	27.126.754	65.560	219,55	67.231,30	259,29

Fonte: ELABORADO PELO PRÓPRIO AUTOR

A exclusão de valores acima do quantil 99,8% foi o critério adotado para eliminação de todos os valores atípicos presentes nestas variáveis. A tabela 3 mostra como ficou o banco de dados após a ‘limpeza’.

Tabela 3 – Valores do CadÚnico 2015 após o tratamento dos dados (parâmetros)

Variável	**Observações válidas**	$\overline{Y}$	σ^2	σ
Renda	26.501.235	178,08	52.429,00	228,97
Pessoas	24.375.314	3,17	2,44	1,56
Famílias	24.306.209	1,03	0,03	0,18
Energia	26.721.569	37,14	1.248,39	35,33
Água/Esgoto	26.263.893	18,05	492,38	22,19
Gás	26.802.862	31,67	272,52	16,51
Alimentação	27.063.429	220,58	20.751,36	144,05

Fonte: ELABORADO PELO PRÓPRIO AUTOR

Para a manipulação do arquivo com os dados do CadÚnico foi necessária a conversão do arquivo originalmente em *.css* para *.sav*, o formato utilizado pelo software SPSS. Em seguida o arquivo foi particionado em cinco para poder ser implementado no programa R. O tamanho do arquivo original do CadÚnico e das frações amostrais utilizadas no estudo estão expostas na tabela 4.

Tabela 4 – Tamanho dos arquivos do CadÚnico

Nº de observações	**Tamanho**
$N = 27.192.314$	6,96 GB
$n_{10\%} = 2.719.231$	696 MB
$n_{5\%} = 1.359.613$	348 MB
$n_{0,5\%} = 118.848$	35 MB
$n_{0,1\%} = 27.192$	7 MB

Fonte: ELABORADO PELO PRÓPRIO AUTOR

Os estudos de simulação demandam de muito recurso computacional, tanto pelo tamanho dos arquivos quanto pelo número de amostras geradas. Fez-se imprescindível a utilização de um servidor fornecido pela Faculdade de Economia para a conclusão deste trabalho. O servido operado possuía quatro servidores Intel ® Xeon ® E7-4870 de 2,40 GHz e 128 GB de memória RAM.

4.2 Cenários Considerados

Neste estudo foram selecionadas 1.000 amostras do CadÚnico por meio de cada método de amostragem relacionado no Capítulo 3 e considerando os quatro diferentes tamanhos de amostra escolhidos, frações amostrais de 10%, 5%, 0,5% e 0,1% da população.

A média e a variância de cada variável selecionada no cadastro foram calculadas, segundo cada plano amostral, e para comparamos os resultados foram calculados também medidas de precisão e exatidão para essas estimativas:

- Erro Padrão (EP):

$$\widehat{EP} = \sqrt{\left(1 - \frac{n}{N}\right)\frac{s^2}{n}}$$

- Coeficiente de Variação (CV):

$$\widehat{CV}(\%) = \frac{\widehat{EP}}{\bar{y}} \times 100$$

- Nível de cobertura (NC): número de amostras em que o intervalo de confiança inclui o verdadeiro valor da média.

Nesse estudo também foi analisado o tempo médio necessário para a geração de cada amostra, que revela a complexidade e o gasto computacional de cada método.

4.3 Amostragem Aleatória Simples (AAS)

Na tabela 5 são exibidos os resultados para a Amostragem Aleatória Simples de tamanho 2.378.380, que corresponde a fração de 10% da população.

Tabela 5 – AAS (10%)

y	$\bar{Y}$	$\mathbf{E}(\bar{y}_{AAS})$	$\mathbf{E}(\widehat{EP}(\bar{y}_{AAS}))$	$\mathbf{E}(\widehat{CV}(\bar{y}_{AAS}))$	$\mathbf{NC}_{AAS}$
Renda	178,08	178,0798	0,140299	0,078784	94,0%
Pessoas	3,17	3,1678	0,000936	0,029533	94,9%
Famílias	1,03	1,0307	0,000108	0,010468	96,2%
Energia	37,14	37,1380	0,021573	0,058088	94,6%
Água/Esgoto	18,05	18,0511	0,013603	0,075359	94,8%
Gás	31,67	31,6655	0,010190	0,032179	95,3%
Alimentação	220,58	220,5789	0,087638	0,039731	95,9%

Fonte: ELABORADO PELO PRÓPRIO AUTOR

A fração amostral de 10% é a maior utilizada neste estudo. No caso da Amostragem Aleatória Simples, apresentou melhores resultados para a variável Famílias, em que obtivemos o menor erro padrão, já a maior erro padrão foi referente a variável Renda.

Na tabela 6 são exibidos os resultados para a Amostragem Aleatória Simples de tamanho 1.118.518, que corresponde a fração de 5% da população.

Tabela 6 – AAS (5%)

y	$\bar{Y}$	$\mathbf{E}(\bar{y}_{AAS})$	$\mathbf{E}(\widehat{EP}(\bar{y}_{AAS}))$	$\mathbf{E}(\widehat{CV}(\bar{y}_{AAS}))$	$\mathbf{NC}_{AAS}$
Renda	178,08	178,0852	0,210533	0,118220	96,9%
Pessoas	3,17	3,1678	0,001404	0,044317	95,6%
Famílias	1,03	1,0307	0,000162	0,015708	96,0%
Energia	37,14	37,1374	0,032370	0,087164	94,4%
Água/Esgoto	18,05	18,0519	0,020413	0,113080	94,7%
Gás	31,67	31,6643	0,015290	0,048288	95,1%
Alimentação	220,58	220,5796	0,131505	0,059618	94,3%

Fonte: ELABORADO PELO PRÓPRIO AUTOR

Diminuindo a fração amostral para 5% os resultados continuam bons, mas houve um aumento do erro padrão para as variáveis Renda e Alimentação, o que acabou acarretando na diminuição no nível de cobertura.

Na tabela 7 são exibidos os resultados para a Amostragem Aleatória Simples de tamanho 118.848, que corresponde a fração de 0,5% da população.

Tabela 7 – AAS (0,5%)

y	$\bar{Y}$	$\mathbf{E}(\bar{y}_{\mathbf{AAS}})$	$\mathbf{E}(\widehat{\mathbf{EP}}(\bar{y}_{\mathbf{AAS}}))$	$\mathbf{E}(\widehat{\mathbf{CV}}(\bar{y}_{\mathbf{AAS}}))$	$\mathbf{NC}_{\mathbf{AAS}}$
Renda	178,08	178,0688	0,659852	0,370561	95,4%
Pessoas	3,17	3,1677	0,004400	0,138912	95,0%
Famílias	1,03	1,0307	0,000508	0,049248	95,2%
Energia	37,14	37,1381	0,101476	0,273239	96,0%
Água/Esgoto	18,05	18,0537	0,063995	0,354474	95,1%
Gás	31,67	31,6646	0,047932	0,151374	95,1%
Alimentação	220,58	220,5908	0,412339	0,186925	94,2%

Fonte: ELABORADO PELO PRÓPRIO AUTOR

Para o caso da fração amostral de 5%, podemos notar um grande aumento no erro padrão para a variável Renda, se mostrando esta a que possui maior variabilidade nos dados.

Na tabela 8 são exibidos os resultados para a Amostragem Aleatória Simples de tamanho 23.770, que corresponde a fração de 0,1% da população.

Tabela 8 – AAS (0,1%)

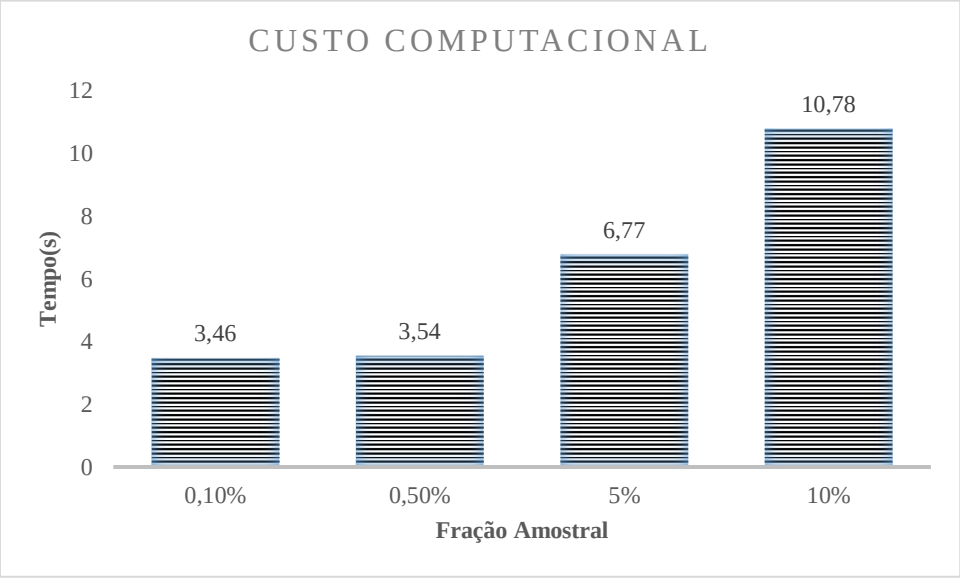
y	$\bar{Y}$	$\mathbf{E}(\bar{y}_{\mathbf{AAS}})$	$\mathbf{E}(\widehat{\mathbf{EP}}(\bar{y}_{\mathbf{AAS}}))$	$\mathbf{E}(\widehat{\mathbf{CV}}(\bar{y}_{\mathbf{AAS}}))$	$\mathbf{NC}_{\mathbf{AAS}}$
Renda	178,08	178,0598	1,478416	0,830303	95,3%
Pessoas	3,17	3,1680	0,009861	0,311265	94,8%
Famílias	1,03	1,0306	0,001136	0,110255	94,9%
Energia	37,14	37,1305	0,227334	0,612263	94,7%
Água/Esgoto	18,05	18,0431	0,143321	0,794340	95,4%
Gás	31,67	31,6630	0,107411	0,339238	93,1%
Alimentação	220,58	220,5603	0,923762	0,418827	94,2%

Fonte: ELABORADO PELO PRÓPRIO AUTOR

No caso da menor fração amostral utilizada neste estudo, os resultados são satisfatórios, com o nível de cobertura próximo ao estimado em todas as variáveis. Há ressalvas que podem ser feitas quanto a Renda e Alimentação, que por possuírem maior desvio padrão exercem maior influência em amostras pequenas.

Na figura 2 é apresentado o tempo médio necessário para se obter cada amostra no método da Amostragem Aleatória Simples.

Figura 2 – Tempo médio para se obter cada amostra AAS no CadÚnico 2015



Fonte: ELABORADO PELO AUTOR

O gráfico mostra que não há uma relação linear entre o tamanho da amostra e o tempo gasto. O tempo médio das amostras de 0,1% e 0,5% são muitos próximos, já a de fração amostral 10% se sobressai com um tempo médio superior a 10s.

4.4 Amostragem Estratificada Simples (AES)

No método da Amostragem Estratificada, os estratos foram definidos de acordo com a unidade federativa e a qualidade de ser da zona rural ou urbana de cada domicilio, contabilizando 54 estratos.

4.4.1 Alocação Uniforme (AES_u)

Na tabela 9 são exibidos os resultados para a Amostragem Estratificada Simples com Alocação Uniforme de tamanho 2.282.494, que corresponde a fração de 10% da população.

Tabela 9 – AES_u (10%)

$\boldsymbol{y}$	$\overline{\boldsymbol{Y}}$	$\mathbf{E}(\bar{\mathbf{y}}_{\boldsymbol{AES_u}})$	$\mathbf{E}(\widehat{\mathbf{EP}}(\bar{\mathbf{y}}_{\boldsymbol{AES_u}}))$	$\mathbf{E}(\widehat{\mathbf{CV}}(\bar{\mathbf{y}}_{\boldsymbol{AES_u}}))$	$\mathbf{NC}_{\boldsymbol{AES_u}}$
Renda	178,08	178,0689	0,226289	0,127080	96,3%
Pessoas	3,17	3,1679	0,001450	0,045783	95,0%
Famílias	1,03	1,0307	0,000165	0,016034	94,4%
Energia	37,14	37,1360	0,034201	0,092096	95,2%
Água/Esgoto	18,05	18,0502	0,020999	0,116336	95,3%
Gás	31,67	31,6651	0,015198	0,047997	94,7%
Alimentação	220,58	220,5735	0,138623	0,062847	94,5%

Fonte: ELABORADO PELO PRÓPRIO AUTOR

Há um aumento no erro padrão em todas as variáveis com relação ao mesmo tamanho de amostra da Amostragem Aleatória Simples, outro ponto a se notar é o aumento expressivo no nível de cobertura da variável Renda.

Na tabela 10 são exibidos os resultados para a Amostragem Estratificada Simples com Alocação Uniforme de tamanho 1.100.557, que corresponde a fração de 5% da população.

Tabela 10 – AES_u (5%)

y	$\bar{Y}$	$\mathbf{E}(\bar{y}_{AES_u})$	$\mathbf{E}(\widehat{\mathbf{EP}}(\bar{y}_{AES_u}))$	$\mathbf{E}(\widehat{\mathbf{CV}}(\bar{y}_{AES_u}))$	$\mathbf{NC}_{AES_u}$
Renda	178,08	178,0557	0,333424	0,187258	94,1%
Pessoas	3,17	3,1679	0,002140	0,067547	94,9%
Famílias	1,03	1,0307	0,000244	0,023668	95,8%
Energia	37,14	37,1349	0,050366	0,135629	94,4%
Água/Esgoto	18,05	18,0495	0,030896	0,171172	95,9%
Gás	31,67	31,6642	0,022429	0,070835	95,2%
Alimentação	220,58	220,5761	0,204254	0,092600	95,7%

Fonte: ELABORADO PELO PRÓPRIO AUTOR

No caso da Amostragem Estratificada com Alocação Uniforme de tamanho 5% já pode ser visto um afastamento da média da variável Renda, refletindo também no aumento do erro padrão. Os valores do nível de cobertura que eram baixos para a fração de 10% sofreram um aumento neste caso.

Na tabela 11 são exibidos os resultados para a Amostragem Estratificada Simples com Alocação Uniforme de tamanho 118.854, que corresponde a fração de 0,5% da população.

Tabela 11 – AES_u (0,5%)

y	$\bar{Y}$	$\mathbf{E}(\bar{y}_{AES_u})$	$\mathbf{E}(\widehat{\mathbf{EP}}(\bar{y}_{AES_u}))$	$\mathbf{E}(\widehat{\mathbf{CV}}(\bar{y}_{AES_u}))$	$\mathbf{NC}_{AES_u}$
Renda	178,08	178,1006	1,031799	0,579328	95,6%
Pessoas	3,17	3,1677	0,006624	0,209092	96,4%
Famílias	1,03	1,0307	0,000755	0,073258	94,7%
Energia	37,14	37,1363	0,155704	0,419276	94,4%
Água/Esgoto	18,05	18,0533	0,095461	0,528768	95,8%
Gás	31,67	31,6635	0,069450	0,219341	95,1%
Alimentação	220,58	220,5648	0,631733	0,286416	95,2%

Fonte: ELABORADO PELO PRÓPRIO AUTOR

Para a fração amostral de 0,5% o erro padrão supera uma unidade em Renda, seguido de aumento nas outras variáveis. O nível de cobertura da variável Pessoas se destaca superando os 95%.

Na tabela 12 são exibidos os resultados para a Amostragem Estratificada Simples com Alocação Uniforme de tamanho 23.814, que corresponde a fração de 0,1% da população.

Tabela 12 – AES_u (0,1%)

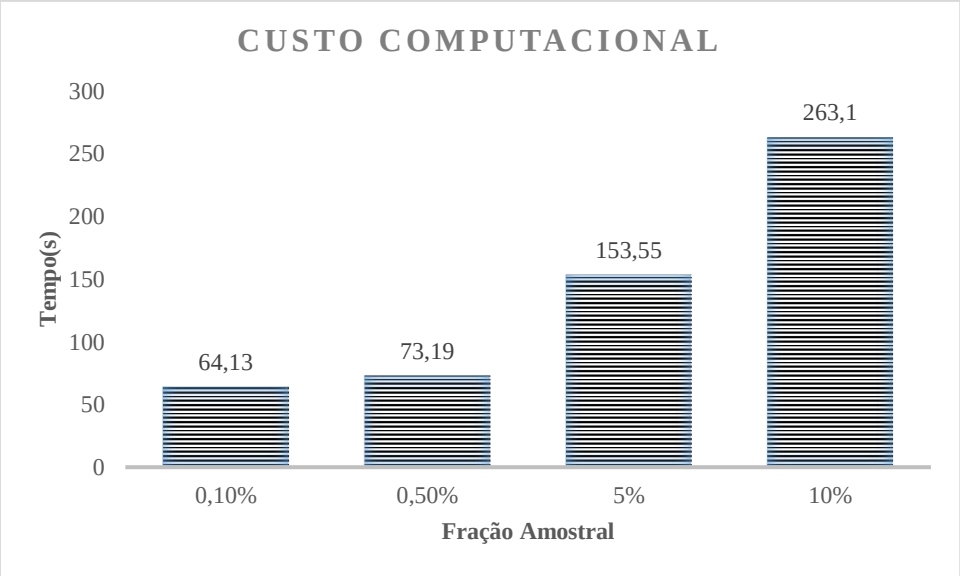
y	$\bar{Y}$	$\mathbf{E}(\bar{y}_{AES_u})$	$\mathbf{E}(\widehat{\mathbf{EP}}(\bar{y}_{AES_u}))$	$\mathbf{E}(\widehat{\mathbf{CV}}(\bar{y}_{AES_u}))$	$\mathbf{NC}_{AES_u}$
Renda	178,08	178,2000	2,307342	1,294800	94,8%
Pessoas	3,17	3,1700	0,014804	0,467400	95,1%
Famílias	1,03	1,0300	0,001687	0,163700	94,4%
Energia	37,14	37,1500	0,348336	0,937600	94,3%
Água/Esgoto	18,05	18,0500	0,213230	1,181400	93,8%
Gás	31,67	31,6700	0,155325	0,490500	94,7%
Alimentação	220,58	220,6100	1,410802	0,639500	95,7%

Fonte: ELABORADO PELO PRÓPRIO AUTOR

Para a menor fração amostral da Amostragem Estratificada com Alocação Uniforme os resultados sofrem um grande aumento no erro padrão, o que se exemplifica com o coeficiente de variação superando uma unidade nos casos das varáveis Água/Esgoto e Renda. O nível de cobertura ficou abaixo do esperado na maioria dos casos.

Na figura 3 é apresentado o tempo médio necessário para se obter cada amostra no método da Amostragem Estratificada Simples com Alocação Uniforme.

Figura 3 – Tempo médio para se obter cada amostra AES_u no CadÚnico 2015



Fonte: ELABORADO PELO AUTOR

O nível de complexidade para o método da Amostragem Estratificada é maior que o da Amostragem Aleatória Simples, o que se reflete no tempo da amostra de 10% superando quatro minutos.

4.4.2 Alocação Proporcional (AES_p)

Na tabela 13 são exibidos os resultados para a Amostragem Estratificada Simples com Alocação Proporcional de tamanho 2.378.407, que corresponde a fração de 10% da população.

Tabela 13 – AES_p (10%)

y	$\overline{Y}$	$\mathbf{E}(\bar{y}_{AES_p})$	$\mathbf{E}(\widehat{\mathbf{EP}}(\bar{y}_{AES_p}))$	$\mathbf{E}(\widehat{\mathbf{CV}}(\bar{y}_{AES_p}))$	$\mathbf{NC}_{AES_p}$
Renda	178,08	178,08	0,158330	0,088908	95,0%
Pessoas	3,17	3,168	0,001050	0,033068	94,7%
Famílias	1,03	1,031	0,000120	0,011709	94,6%
Energia	37,14	37,137	0,023600	0,063561	95,0%
Água/Esgoto	18,05	18,051	0,014130	0,078288	94,2%
Gás	31,67	31,665	0,011000	0,034729	93,8%
Alimentação	220,58	220,579	0,095830	0,043444	93,9%

Fonte: ELABORADO PELO PRÓPRIO AUTOR

A Amostragem Estratificada com Alocação Proporcional por muitas vezes apresenta melhores resultados que com Alocação Uniforme. Neste caso podemos ver uma queda nas estimativas de erro padrão, mas também níveis de cobertura que ficam abaixo dos 95% previsto.

Na tabela 14 são exibidos os resultados para a Amostragem Estratificada Simples com Alocação Proporcional de tamanho 1.118.546, que corresponde a fração de 5% da população.

Tabela 14 – AES_p (5%)

y	$\bar{Y}$	$\mathbf{E}(\bar{y}_{AES_p})$	$\mathbf{E}(\widehat{\mathbf{EP}}(\bar{y}_{AES_p}))$	$\mathbf{E}(\widehat{\mathbf{CV}}(\bar{y}_{AES_p}))$	$\mathbf{NC}_{AES_p}$
Renda	178,08	178,0835	0,237600	0,133419	94,3%
Pessoas	3,17	3,1678	0,001570	0,049619	94,9%
Famílias	1,03	1,0307	0,000180	0,017573	94,7%
Energia	37,14	37,1385	0,035410	0,095358	93,9%
Água/Esgoto	18,05	18,0507	0,021200	0,117471	93,9%
Gás	31,67	31,6655	0,016500	0,052107	95,0%
Alimentação	220,58	220,5805	0,143790	0,065188	94,0%

Fonte: ELABORADO PELO PRÓPRIO AUTOR

O método da Alocação Proporcional gera melhores resultados que a Alocação Uniforme, com estimativas da média mais próximas do valor real.

Na tabela 15 são exibidos os resultados para a Amostragem Estratificada Simples com Alocação Proporcional de tamanho 118.848, que corresponde a fração de 0,5% da população.

Tabela 15 – AES_p (0,5%)

y	$\bar{Y}$	$\mathbf{E}(\bar{y}_{AES_p})$	$\mathbf{E}(\widehat{\mathbf{EP}}(\bar{y}_{AES_p}))$	$\mathbf{E}(\widehat{\mathbf{CV}}(\bar{y}_{AES_p}))$	$\mathbf{NC}_{AES_p}$
Renda	178,08	178,0917	0,744932	0,418285	96,3%
Pessoas	3,17	3,1677	0,004926	0,155518	94,5%
Famílias	1,03	1,0307	0,000568	0,055062	95,9%
Energia	37,14	37,1410	0,111026	0,298930	95,4%
Água/Esgoto	18,05	18,0509	0,066474	0,368254	93,9%
Gás	31,67	31,6648	0,051727	0,163360	93,9%
Alimentação	220,58	220,5772	0,450774	0,204361	95,8%

Fonte: ELABORADO PELO PRÓPRIO AUTOR

No caso da fração amostral de 0,5%, a variabilidade da Renda ganha destaque com um aumento no erro padrão, o que é acompanhada também da variável Alimentação.

Na tabela 16 são exibidos os resultados para a Amostragem Estratificada Simples com Alocação Proporcional de tamanho 23.795, que corresponde a fração de 0,1% da população.

Tabela 16 – AES_p (0,1%)

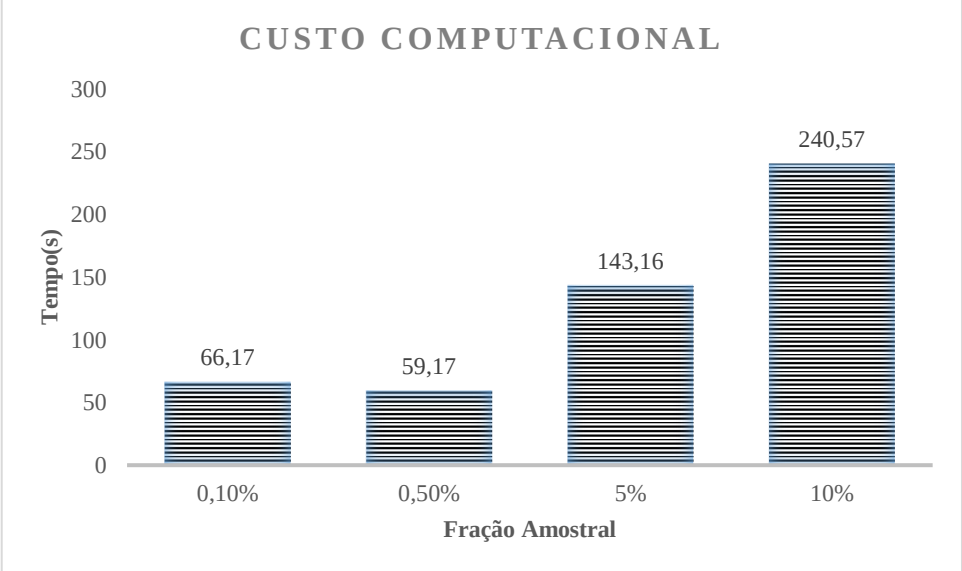
y	$\bar{Y}$	$\mathbf{E}(\bar{y}_{AES_p})$	$\mathbf{E}(\widehat{\mathbf{EP}}(\bar{y}_{AES_p}))$	$\mathbf{E}(\widehat{\mathbf{CV}}(\bar{y}_{AES_p}))$	$\mathbf{NC}_{AES_p}$
Renda	178,08	178,0386	1,666425	0,935974	94,4%
Pessoas	3,17	3,16777	0,011024	0,348018	95,1%
Famílias	1,03	1,03068	0,001270	0,123256	95,3%
Energia	37,14	37,1271	0,248296	0,668767	94,7%
Água/Esgoto	18,05	18,0412	0,148754	0,824508	96,4%
Gás	31,67	31,6638	0,115808	0,365754	94,9%
Alimentação	220,58	220,5371	1,009057	0,457544	95,4%

Fonte: ELABORADO PELO PRÓPRIO AUTOR

De maneira geral o método da Alocação Proporcional se mostra um bom método, com a capacidade de gerar estimativas com menor erro padrão em todos os casos.

Na figura 4 é apresentado o tempo médio necessário para se obter cada amostra no método da Amostragem Estratificada Simples com Alocação Proporcional.

Figura 4 – Tempo médio para se obter cada amostra AES_p no CadÚnico 2015



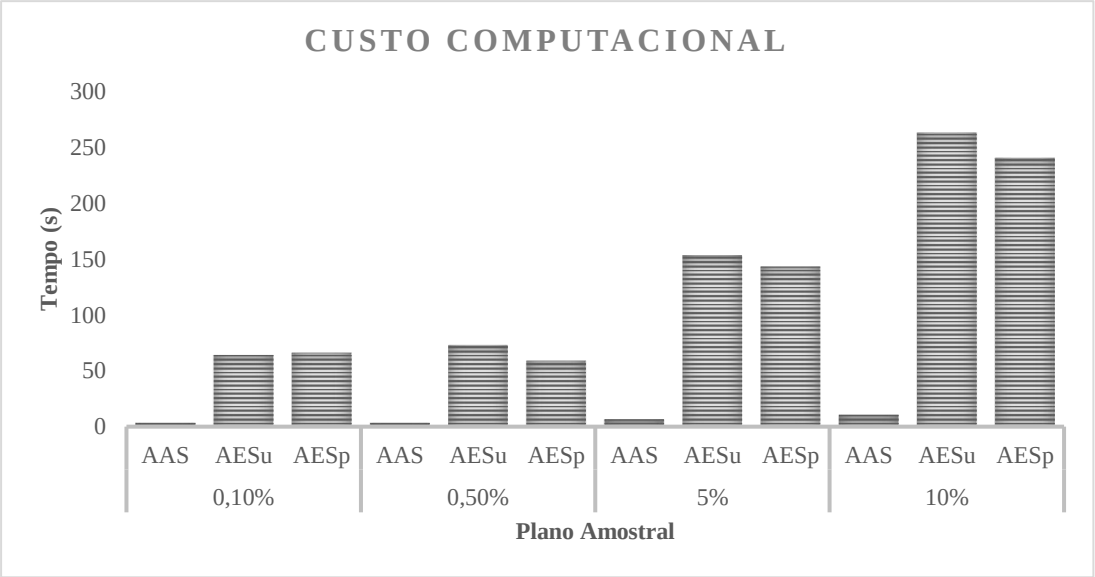
Fonte: ELABORADO PELO PRÓPRIO AUTOR

A Amostragem Estratificada com Alocação Proporcional se mostrou mais rápida que a com Alocação Uniforme. O tempo da amostra de 5% surpreende por ser menor que de 1% mas pode ser explicado por disformidades no funcionamento do servidor.

4.4.3 Custo Computacional

A figura 5 apresenta a comparação dos tempos médios necessários para se selecionar cada amostra segundo os diferentes métodos empregados neste estudo.

Figura 4 – Tempo médio para se obter cada amostra no CadÚnico 2015



Fonte: ELABORADO PELO AUTOR

O método da Amostragem Aleatória Simples, por ser o único a não envolver amostragem complexa, fica evidente como o método com os menores tempos, seguido da Amostragem Estratificada com Alocação Proporcional e com Alocação Uniforme.

O tempo de obtenção de cada amostra não segue uma tendência linear, sendo o tempo das amostras de 0,1% e 0,5% muito parecidos em comparação com as amostras maiores, de 5% e 10%.

Avaliando os resultados das seções anteriores fica nítido que o aumento do tempo nos planos amostrais complexos não resultou em ganho de precisão neste estudo. A Amostragem Aleatória Simples se apresenta como a mais vantajosa, sendo a mais precisa e também a mais rápida.

CONCLUSÃO

O resultado das amostras são bons de maneira geral e todas apresentaram estimativas próximas ao valor real do parâmetro. O método da Amostragem Aleatória Simples se sobressaiu com os menores valores de erro padrão, coeficiente de variação e como o método mais rápido; seguida da Amostragem Estratificada com Alocação Proporcional e com Alocação Uniforme.

O método da Amostragem Estratificada, com ambos os tipos de alocação, apesar de ser um método de amostragem complexa eficiente não apresentou as vantagens que se esperaria. Podemos questionar a escolha das variáveis para a definição dos estratos e a opção da região do domicílio como fator explicativo para os índices calculados.

Em relação ao tamanho, até as amostras menores apresentaram boas estimativas, apesar de serem maior influenciadas pela variabilidade dos dados. Este efeito fica mais nítido nos casos das variáveis Renda e Alimentação.

Se mostrou imprescindível o uso de um servidor, os estudos de simulação e a manipulação de grandes arquivos fica limitado a usuários de computadores potentes e mais caros. Há a necessidade do desenvolvimento de novas técnicas para garantir ao usuário de um computador comum a realização de outras pesquisas como essas. Novos pacotes no R surgem prometendo sanar essa deficiência.

Nosso estudo garantiu que até a menor amostra do método mais simples consegue gerar boas estimativas. A vantagem de se utilizar amostras menores fica evidente na contraposição em se fazer uma análise tipo censo, do banco de dados inteiro, o que acarretaria em carregar um arquivo de 6 GB, ao invés de usar um arquivo de 7 MB (como mostra a Tabela 4), que seria um processo muito mais simples.

Ao término deste estudo podemos definir o método da Amostragem Aleatória Simples, com tamanho de 0,1%, como o que mais se destacou, sendo o mais rápido e também com uma precisão aceitável. A utilização deste método pode ser um diferencial em situações reais que envolvam a tomada de decisões diárias, em que a análise por censo é computacionalmente custosa e não é uma opção. A análise dos dados do CadÚnico sob diferentes métodos e tamanhos de amostras nos leva a perceber o valor da amostragem para os dias atuais, se mostrando uma solução para a análise de dados em um mundo sob o fenômeno do Big Data.

REFERÊNCIAS

BOLFARINE, Heleno; BUSSAB, Wilton. **Elementos de amostragem**. São Paulo: Editora Edgard Blücher, 2005

BRASIL. Decreto nº 3.877, de 24 de junho de 2001**. Institui o Cadastramento Único para Programas Sociais do Governo Federal.**, Brasília, DF, jun 2001. Disponivel em: < http://www.planalto.gov.br/ccivil_03/decreto/2001/D3877.htm>. Acesso em: 15 nov. 2018.

BRASIL. Decreto nº 6.135, de 26 de junho de 2007**. Dispõe sobre o Cadastro Único para Programas Sociais do Governo Federal e dá outras providências**, Brasília, DF, jun 2007. Disponivel em: <http://www.mds. gov.br/webarquivos/legislacao/cadastro_unico/decretos/2007/decreto_6135.pdf>. Acesso em: 15 nov. 2018.

CUKIER, K. Data, data, everywhere: a special report on managing information. **The Economist**, v. 394, Issue 867, 2010.

HAND, D. J. Statistical challenges of administrative and transaction data. Imperial College London and Winton Capital Management, Londres, 2017.

IBM Corp. Released 2012. IBM SPSS Statistics for Windows, Version 21.0. Armonk, NY: IBM Corp.

LETOUZÉ, Emmanuel. **Big data for development**: challenges & opportunities. UN Global Pulse, 2012.

MARR, B. **Big Data**: The 5 Vs Everyone Must Know, Linkedin. Disponível em: < https://www.linkedin.com/pulse/20140306073407-64875646-big-data-the-5-vs-everyone-must-know >. Acesso em: 4 jul. 2018

MINELI, Michael; CHAMBERS, Michele; DHIRAJ Ambiga. **Big data, big analytics**: emerging business intelligence and analytic trends for today's businesses. New Jersey: John Wiley & Sons, Inc., 2013.

MINISTÉRIO DO DESENVOLVIMENTO SOCIAL, **CadÚnico**. 2015 Disponivel em: < http://www.brasil.gov.br/economia-e-emprego/2010/03/cadunico>. Acesso em: 14 nov. 2018.

NESELLO, P.; FACHINELLI, A.C. Big Data: O novo desafio para a gestão. **Revista Inteligência Competitiva**, São Paulo, v. 4, n. 1, p. 18-38, jan./mar. 2014.

NEYMAN, J. On the two different aspects of the representative method, **Journal of the Royal Statistical Society**, 1934.

OCDE. Short-Term Economic Statistics (STES) Administrative Data: Two Frameworks of Papers. Disponível em <http://www.oecd.org/std/short-termeconomicstatisticsstesadministra> Acesso em: 15 de nov. 2018.

R Development Core Team.R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2017

SCHÖNBERGER-MAYER, Viktor; CUKIER, Kenneth. Tradução Paulo Palzonoff Junior. **Big data: como extrair volume, variedade, velocidade e valor da avalanche de informação cotidiana**. Rio de Janeiro: Elsevier, 2013.

SENRA, Nelson; FONSECA, Silvia; MILLIONS, Teresa. **O desafio de retratar o país**: Entrevistas com os presidentes do IBGE no período de 1985 a 2015. Rio de Janeiro: IBGE, 2016

VIEIRA, M. D. T. Notas de aula de Amostragem. Juiz de Fora, Universidade Federal de Juiz de Fora, 2017.

APÊNDICE A – SINTAXE SPSS

A sintaxe utilizada no software SPSS para o tratamento do banco de dados é apresentada a seguir. Devido a sua extensão algumas partes não foram exibidas.

```
GET
  FILE='E:\...\CADÚNICO_2015 LABEL.sav'.
DATASET NAME Conjunto_de_dados1 WINDOW=FRONT.
FREQUENCIES VARIABLES=vlr_renda_media_fam qtd_pessoas_domic_fam
qtd_familias_domic_fam val_desp_energia_fam val_desp_agua_esgoto_fam
val_desp_gas_fam val_desp_alimentacao_fam val_desp_transpor_fam
val_desp_aluguel_fam val_desp_medicamentos_fam
  /FORMAT=NOTABLE
  /PERCENTILES=99.8
  /ORDER=ANALYSIS.
Frequencies
Notes
Comments
Input Data  E:\...\CADÚNICO_2015 LABEL.sav
      Active Dataset    Conjunto_de_dados1
      Filter      <none>
      Weight      <none>
      Split File  <none>
      N of Rows in Working Data File     27192314
Missing Value Handling  Definition of Missing  User-defined missing values
are treated as missing.
      Cases Used  Statistics are based on all cases with valid data.
Syntax

FILTER OFF.
USE ALL.
SELECT IF (ANY(val_desp_transpor_fam<300,1)).
EXECUTE.
FREQUENCIES VARIABLES=val_desp_transpor_fam val_desp_agua_esgoto_fam
  /FORMAT=NOTABLE
  /PERCENTILES=99.8
  /ORDER=ANALYSIS.
Frequencies
Notes
Comments
```

```
Input Data  E:\...\CADÚNICO_2015 LABEL.sav
      Active Dataset    Conjunto_de_dados1
      Filter      <none>
      Weight      <none>
      Split File  <none>
      N of Rows in Working Data File     25072938
Missing Value Handling  Definition of Missing   User-defined missing values
are treated as missing.
      Cases Used  Statistics are based on all cases with valid data.
Syntax            FREQUENCIES VARIABLES=val_desp_transpor_fam
val_desp_agua_esgoto_fam
  /FORMAT=NOTABLE
  /PERCENTILES=99.8
  /ORDER=ANALYSIS.
Resources   Processor Time    00:28:11,39
      Elapsed Time      00:30:21,71



[Conjunto_de_dados1] E:\...\CADÚNICO_2015 LABEL.sav
Statistics
            Valor de despesas com transporte. Campo no formato NNNNN (não
existem casas decimais) Valor de despesas com água. Campo no formato NNNNN
(não existem casas decimais)
N     Valid 25072938    24914145
      Missing     0     158793
Percentiles 99,8  200,00      162,00

DATASET ACTIVATE Conjunto_de_dados1.

SAVE OUTFILE='E:\...\CADÚNICO_2015 LABEL - Copia (3).sav'
  /COMPRESSED.
```

APÊNDICE B – CÓDIGO R

Uma parte do script utilizado para a Amostragem Estratificada com Alocação Uniforme de tamanho 10% é apresentado abaixo. O código não é descrito por completo devido a sua extensão.

```
require("foreign")
require("survey")
require("sampling")
require("tictoc")

CADUNICO1 <- as.data.frame(read.spss("Corte1.sav"))
CADUNICO2 <- as.data.frame(read.spss("Corte2.sav"))
CADUNICO3 <- as.data.frame(read.spss("Corte3.sav"))
CADUNICO4 <- as.data.frame(read.spss("Corte4.sav"))
CADUNICO <-rbind(CADUNICO1,CADUNICO2,CADUNICO3,CADUNICO4)

CADUNICO <- subset(CADUNICO,CADU$Estrato!= '0')
CADUNICO <- CADUNICO[order(CADUNICO$Estrato),]

tab <- table(CADUNICO$Estrato)
c <-rep(NA, length=54)
t <-rep(NA, length=54)

for(i in 1:54){

  c[i] <- ifelse(tab[[i]]<44045,tab[[i]],44045)
}
sum(c)

for(i in 1:54){
  t[i] <- tab[[i]]
}

attach(CADUNICO)

YA=mean(vlr_renda_media_fam)
tic()
yA<-rep(NA, length=1000)
EPA<-rep(NA, length=1000)
CVA<-rep(NA, length=1000)
ICIA<-rep(NA, length=1000)
ICSA<-rep(NA, length=1000)
NCA<-rep(NA, length=1000)
```

```
for(i in 1:1000){

  fpc=rep(t,c)
  IAESs=sampling::strata(CADUNICO, stratanames=c("Estrato"), c,
method=c("srswor"))
  AESs1=getdata(CADUNICO,IAESs)
  Plano=svydesign(~1, strata=~Estrato, data = AESs1, probs=~IAESs$Prob,
fpc=~fpc)

  yA[i]<-svymean(~vlr_renda_media_fam,Plano)[[1]]
  EPA[i]<- SE(svymean(~vlr_renda_media_fam,Plano))[[1]]
  ICIA[i]<-yA[i]-1.96*EPA[i]
  ICSA[i]<-yA[i]+1.96*EPA[i]
  if(YA>ICIA[i] & YA<ICSA[i]){
    NCA[i]<-1
  }
  if(YA<ICIA[i] || YA>ICSA[i]){
    NCA[i]<-0
  }

########Renda######
CVA<-(EPA/yA)*100
myA <- mean(yA)
mEPA <- mean(EPA)
mCVA <- mean(CVA)
sNCA <- sum(NCA)

dat1 <- data.frame(Var="Renda",Media=myA,EP=mEPA,CV=mCVA,NC=sNCA)
dat1

dat <- rbind(dat1,dat2,dat3,dat4,dat5,dat6,dat7)
dat

ttol <- file("TAmostr10EstrUn.txt", open = "wt")
sink(ttol)
sink(ttol, type = "message")
toc()
sink(type = "message")
sink()

write.table(dat, file="Amostr10EstrUn.csv",sep =';')
```