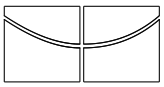

Universidade de Brasília

Instituto de Ciências Exatas

Departamento de Estatística

Dissertação de Mestrado

Metodologia para Subamostragem em Grandes Bancos de Dados Amostrais Complexos para Realização de Testes de Hipóteses

por

Gilberto Rezende de Almeida Junior

Orientador: Prof. Dr. Alan Ricardo da Silva

Outubro de 2017

Gilberto Rezende de Almeida Júnior

Metodologia para Subamostragem em Grandes Bancos de Dados Amostrais Complexos para Realização de Testes de Hipóteses

Dissertação apresentada ao Departamento de Estatística do Instituto de Ciências Exatas da Universidade de Brasília como requisito parcial à obtenção do título de Mestre em Estatística.

Universidade de Brasília

Brasília, Outubro de 2017

"There were 5 Exabytes of information
created between the dawn of civilization through 2003,
but that much information is now created every 2 days."
Google's CEO, Eric Schmidt, 2010.

Agradecimentos

A minha família, que me apoiou sempre que necessário.

Aos meus amigos, que souberam entender a ausência nesta fase de estudos.

Aos meus professores, que ensinaram a teoria e aplicação da estatística atiçando a curiosidade do questionamento de antigas e imaginação de novas técnicas.

Sumário

Agradecimentos **ii**

Lista de Figuras **4**

Lista de Tabelas **7**

Resumo **8**

Abstract **9**

Introdução **10**

1 Testes de Hipóteses **13**

1.1 Introdução 13

1.2 Testes de Hipóteses 13

1.2.1 Teste para média em uma população 16

1.2.1.1 Teste para média com variância conhecida 16

1.2.1.2 Teste para média com variância desconhecida 16

1.2.2 Comparação de variâncias entre duas populações 17

1.2.3 Comparação de médias entre duas populações 18

1.2.3.1 Teste t para amostras pareadas 18

1.2.3.2 Teste t para amostras independentes com variância conhecida 18

1.2.3.3 Teste t para amostras independentes com variâncias desconhecidas iguais 19

1.2.3.4 Teste t para amostras independentes com variâncias desconhecidas diferentes . . . 19
1.3 Consistência de estimadores . . . 20

2 Amostragem Complexa **22**
2.1 Introdução . . . 22
2.2 Peso amostral . . . 23
2.3 Teste para média com variância desconhecida . . . 24
2.4 Estimação de variância em pesquisas amostrais complexas . . . 26
2.5 A PNAD . . . 30

3 Material e métodos **32**
3.1 Introdução . . . 32
3.2 Material . . . 32
3.2.1 Tamanho amostral . . . 36
3.3 Métodos . . . 38
3.3.1 Primeira amostra . . . 38
3.3.2 Subamostragem . . . 40
3.3.3 Técnica de Subamostragem mista em Grandes Amostras . . . 42
3.3.4 Teste de Efeito do Tamanho da Amostra na Significância de Testes de Hipóteses . . . 43

4 Análise dos Resultados **45**
4.1 Amostra Aleatória Simples . . . 45
4.1.1 Primeira amostra . . . 45
4.1.2 Subamostragem . . . 46
4.2 Amostra Estratificada . . . 50
4.2.1 Primeira amostra . . . 50
4.2.2 Subamostragem . . . 52
4.3 Amostra Complexa . . . 54
4.3.1 Primeira amostra . . . 54
4.3.2 Subamostragem . . . 56

4.4 Pesos Utilizados 57
4.5 Técnica de subamostragem mista na PNAD 58

5 Conclusões **62**

5.1 Conclusões 62
5.2 Limitações do trabalho 63
5.3 Recomendações para trabalhos futuros 64

Referências Bibliográficas **65**

A Apêndice **67**

A.1 Amostra Aleatória Simples 67
A.2 Amostra Estratificada 70
A.3 Amostra Complexa 71

Lista de Figuras

1 Média e variância estimada da média por tamanho da amostra - Bernoulli 11

2 Média e variância estimada da média por tamanho da amostra - Normal(0 ; 25) 12

3 Média e variância estimada da média por tamanho da amostra - Normal(0 ; 625) 12

3.1 Histograma do rendimento mensal de todas as fontes para pessoas de 10 anos ou mais de idade - PNAD 2014 35

4.1 Distribuições normais simuladas 47

4.2 Amplitude do intervalo de confiança ao aumentar o tamanho mínimo da amostra - Normal 47

4.3 Amplitude do intervalo de confiança ao aumentar o tamanho mínimo da amostra - Log-normal 48

4.4 Amplitude Média - Intervalo de Confiança - Amostragem Estratificada - Normal e Log-normal 51

4.5 Valor da estatística t referentes a diferentes μ_D - amostragem principal e mista 61

Lista de Tabelas

1.1 Tipos de erro em testes de hipóteses 15

2.1 Dados populacionais fictícios para exemplificação 24
2.2 Pesos atribuídos ao exemplo 24
2.3 Estimação de médias em amostragem 24
2.4 Ultimate cluster de tamanho $n_a = 4$ 28

3.1 Algorítmo de geração das variáveis do estudo 33
3.2 Dados utilizados para simulação da renda - PNAD 2014 33
3.3 Variáveis PNAD 2014 - Medidas Resumo 40
3.4 Graus de liberdade para o teste t por tipo de amostragem 43

4.1 Tamanho amostral - Amostragem aleatória simples 46
4.2 Percentual de acertos da estimativa da média nas simulações para AAS com desvio padrão e tamanho de amostra variáveis 49
4.3 Percentual de acertos da estimativa da média nas simulações para AAS com desvio padrão e tamanho de amostra variáveis - Bases "principal", "subamostragem" e "caso misto" 50
4.4 Tamanho amostral - Amostragem estratificada 51
4.5 Percentual de acertos dos intervalos de confiança gerados pelas amostras estratificadas por tamanho de amostra e nível de confiança do teste - Normal e Log-normal 52

4.6 Percentual de acertos da estimativa da média nas simulações para amostragem estratificada com desvio padrão e tamanho de amostra variáveis - Bases “principal”, “subamostragem” e “caso misto” - Subamostragem com plano amostral estratificado 53
4.7 Percentual de acertos da estimativa da média nas simulações para amostragem estratificada com desvio padrão e tamanho de amostra variáveis - Bases “principal”, “subamostragem” e “caso misto” - Subamostragem com plano amostral aleatório simples 53
4.8 Percentual de acertos dos intervalos de confiança gerados pelas amostras complexas e amplitude dos IC por tamanho de amostra 55
4.9 Estimação equivocada de parâmetros em amostragem complexa - Amplitude do IC e percentual de acerto da média dentro do IC por tamanho de amostra - Amostragem aleatória simples e estratificada . . 55
4.10 Amplitude dos intervalos de confiança da média nas simulações para amostragem inicial complexa e diferentes tipos de técnicas de subamostragem - Bases “principal”, “subamostragem” e “caso misto” 57
4.11 Percentual de acertos da estimativa da média nas simulações para amostragem inicial complexa e diferentes tipos de técnicas de subamostragem - Bases “principal”, “subamostragem” e “caso misto” 57
4.12 Média da soma dos pesos obtidos nas simulações por tipo de amostragem, subamostragem e utilização de pesos 58
4.13 Tamanho da amostra - PNAD 2012, 2014 e subamostragem mista . . 59
4.14 Estimativas de rendimento médio - PNAD 2012 e 2014 - Amostragem “principal” e subamostragem mista 59
4.15 Tamanho da amostra - PNAD 2012, 2014 e subamostragem mista . . 60

A.1 Simulações para AAS com desvio padrão e tamanho de amostra variáveis 68
A.2 Amplitude média do IC gerado nas simulações para Amostragem Aleatória Simples (AAS) - Casos "principal", "subamostragem"e "misto" 69
A.3 Simulações para amostragem estratificada segundo o nível de confiança e o tamanho de amostra variáveis 70

A.4 Amplitude média do IC gerado nas simulações para amostragem estratificada - Casos "principal", "subamostragem"e "misto- Subamostragem estratificada e aleatória simples 70
A.5 Tamanho amostral no segundo estágio - Amostragem complexa 71
A.6 Dados para amostragem complexa - estratos 72
A.7 Dados para amostragem complexa - conglomerados 73

Resumo

A amostragem é uma metodologia utilizada para auxiliar a seleção de amostras e estimação de parâmetros com base nessas amostras. Usualmente é discutido o tamanho mínimo que deve se tomar em uma amostra. No entanto, ao utilizar amostras grandes, podem surgir problemas na realização de testes de hipóteses pois, segundo a propriedade da consistência dos estimadores, ao aumentar o tamanho amostral a variância do estimador diminui, podendo influenciar no valor da estatística do teste de hipótese. O problema se agrava em amostras complexas.

Neste trabalho é proposto uma técnica de subamostragem para ser aplicada nessas grandes amostras, assim como o algoritmo para fazer uma subamostragem de maneira correta. Um teste para o efeito do tamanho amostral na significância de teste de hipóteses também é apresentado. Foram simulados dados em que os resultados mostraram a importância dessa verificação. Também foi feita uma aplicação utilizando os dados da Pesquisa Nacional por Amostra de Domicílios - PNAD, e os resultados mostraram uma mudança na inferência quando o tamanho da amostra foi reduzido.

Palavras-Chave: Amostragem; Grandes Bancos de Dados; Teste de Hipóteses; Subamostragem .

Abstract

Sampling is a statistical methodology used to aid the sample selection and the parameters estimation based on this sample. It is common to discuss the minimum size to be taken in a sample, however, when using large samples, problems may arise in performing hypothesis tests because, according to the consistency property of the estimators, by increasing the sample size the variance of the estimator decreases. This may influence the value of the hypothesis test statistic. The problem is exacerbated in complex samples.

In this work we propose a subsampling technique to be applied in large samples, as well as an algorithm to conduct resampling. A test for the effect of sample size on the significance of the hypothesis test is also presented. Simulated data have been used and the results showed the importance of this method. In addition, in application to the Brazilian National Household Sample Survey (PNAD) showed that the inference was changed when the sample size was reduced.

Keywords: Sampling; Big Data; Hypothesis Test; Subsampling .

Introdução

Segundo Casella e Berger (2002), a inferência estatística busca generalizar os resultados obtidos a partir de uma amostra com objetivo de estimar parâmetros populacionais. Os diferentes tipos de técnicas de amostragem, que variam desde a forma de coleta à seleção das observações, tratam da especificidade de cada plano amostral. Heeringa et al. (2010) ressaltam que quando há mistura de técnicas de amostragem a complexidade inerente ao plano requer tratamento especial na estimação pontual e na variância das estimativas.

Com a evolução da tecnologia de informação e a facilidade de obtenção de dados, o pesquisador se depara, cada vez mais, com grandes quantidades de informações. O fenômeno *Big Data* é um exemplo dessa grande quantidade de informações à disposição do pesquisador.

Nesse contexto, ao utilizar estimativas intervalares inferenciais em grandes amostras, pode acontecer de que algumas análises estatísticas indiquem uma diferença significativa entre o que está sendo testado, mas não necessariamente o fenômeno apresenta tal diferença. Como exemplo podemos citar o caso do teste para média em uma população normal em que a estatística do teste é calculada como:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \tag{1}$$

onde $\bar{x}$ é a média amostral, μ_0 é a média populacional a ser testada, σ é o desvio-padrão populacional e n é o tamanho da amostra.

Caso o tamanho da amostra n cresça o valor da estatística do teste z também tende a crescer, conforme pode ser visto na Equação (2). Como o valor da estatística do teste é grande e o valor crítico não muda, o tamanho da amostra pode levar à

rejeição da hipótese nula, puramente por questões matemáticas.

$$\lim_{n\to\infty} \frac{\bar{x} - \mu_0}{\sigma}\sqrt{n} = \infty \tag{2}$$

No caso em que o tamanho da amostra é pequeno, como apresentado em (3), é tratado, comumente, ao considerar um tamanho mínimo de amostra, calculado para o plano aleatório simples como $n = \frac{z^2_{\gamma/2}\sigma^2}{E^2}$, onde γ e o nível de confiança e E é erro amostral máximo admitido.

$$\lim_{n\to 0} \frac{\bar{x} - \mu_0}{\sigma}\sqrt{n} = 0 \tag{3}$$

Para exemplificar o exposto, foi gerada via simulação uma variável de distribuição Bernoulli com média 0,6 e duas Normais com média 20 e variâncias 25 e 625, todas com tamanho populacional de 100 mil. Foram retiradas amostras aleatórias simples com tamanhos variando de 100 a 40 mil. O comportamento da média e do erro padrão da média pode ser observado nas Figuras 1, 2 e 3. A fórmula do tamanho amostral para esse caso pode ser encontrada na Seção 3.2.1.

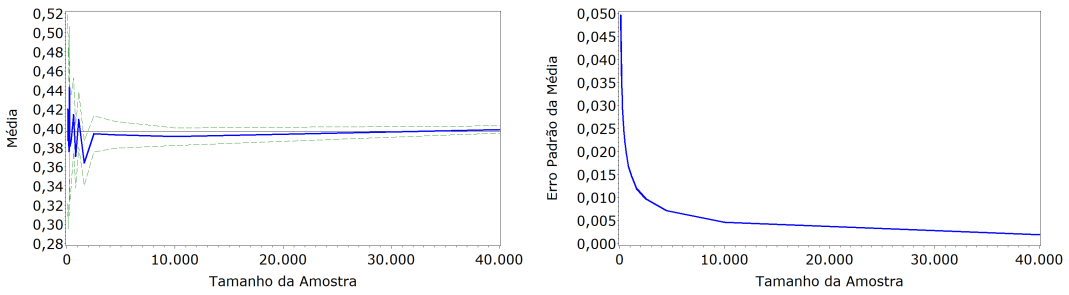

Figura 1: Média e variância estimada da média por tamanho da amostra - Bernoulli

É evidente que o tamanho da amostra influencia diretamente o valor do erro padrão da média, influência principalmente da propriedade de consistência dos estimadores, discutida na Seção 1.3. Percebemos também que a variância começa a se estabilizar a partir de amostras maiores do que 10 mil.

Este trabalho visa discutir a influência do tamanho amostral e seus impactos em testes de hipóteses aplicados a planos amostrais complexos. Uma solução para reduzir a quantidade de dados amostrais, por questões de tempo de processamento, foi dada por Zhu et al. (2015), em que foi retirada uma amostra aleatória simples dos

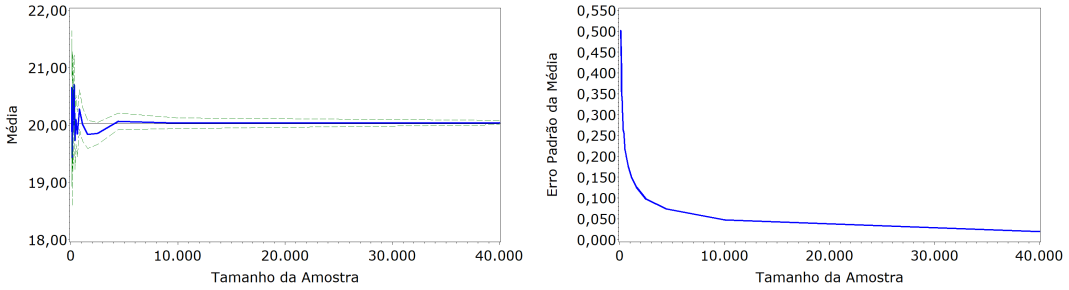

Figura 2: Média e variância estimada da média por tamanho da amostra - Normal(0 ; 25)

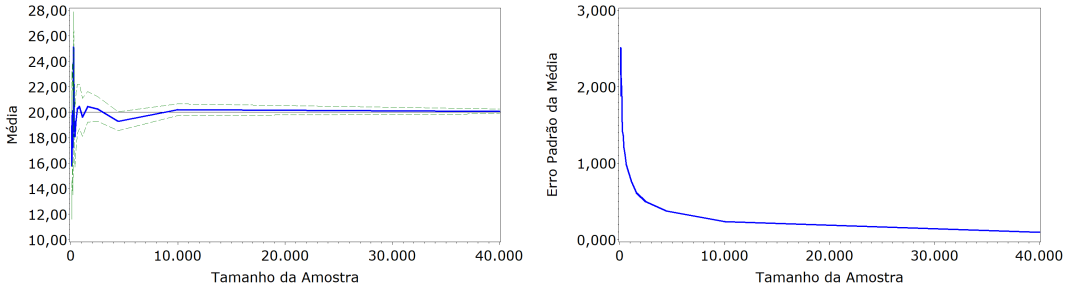

Figura 3: Média e variância estimada da média por tamanho da amostra - Normal(0 ; 625)

dados. No caso de dados advindos de pesquisas amostrais complexas, deve-se tomar um cuidado especial quanto à retirada de subamostras, uma vez que a estrutura inicial do plano de amostragem deve ser preservada.

O objetivo deste trabalho é apresentar uma alternativa para tratamento de testes de hipóteses em grandes bancos de dados provenientes de pesquisa amostrais complexas. Para isso, será apresentada uma forma de subamostragem que mantém a estrutura complexa da amostragem, produzindo assim um teste mais robusto.

O trabalho está organizado da seguinte forma: o Capítulo 1 trata de testes de hipóteses. O Capítulo 2 introduz o conceito de testes de hipóteses em planos amostrais complexos. O Capítulo 3 apresenta a metodologia do trabalho e no Capítulo 4 consta a análise dos resultados. Por fim o Capítulo 5 apresenta as conclusões, limitações do trabalho e recomendações para trabalhos futuros.

Capítulo 1

Testes de Hipóteses

1.1 Introdução

Os testes de hipóteses são métodos inferenciais que auxiliam a quantificação da tomada de decisão. O objetivo do teste de hipóteses é tomar decisões, baseado em uma amostra da população.

Walpole et al. (1993) comentam que nunca sabemos com absoluta certeza se uma hipótese estatística é verdadeira ou falsa. A não ser que examinemos a população inteira, o que seria impraticável na maioria das situações. Podemos retirar uma amostra aleatória, grande o suficiente, da população e usarmos os dados contidos nela para fornecer evidências que apoie ou rejeite a hipótese. Segundo Bussab e Morettin (2010), o objetivo do teste estatístico de hipóteses é fornecer uma metodologia que nos permita verificar se os dados amostrais trazem evidências que apoiem ou não uma hipótese formulada.

Este capítulo tem por objetivo mostrar testes de hipóteses clássicos aplicados em amostras aleatórias simples com reposição e verificar, para cada um, sua relação com o tamanho da amostra.

1.2 Testes de Hipóteses

Em um teste de hipóteses existem duas hipóteses complementares: a hipótese nula, H_0, e a alternativa, H_A. É definido aqui θ como parâmetro populacional a ser

testado. São exemplos de hipóteses comumente testadas $\theta_0 = 0$ e $\theta_0 \neq 0$ ou $\theta_0 \geq 0$ e $\theta_0 < 0$. Após definir as hipóteses e verificar o p-valor do teste, o pesquisador pode tomar sua decisão em rejeitar ou não a hipótese nula (Casella e Berger, 2002).

Na metodologia de teste de hipóteses é calculada a região crítica do teste, definida pelo valor amostral mínimo, ou máximo, que o pesquisador pode observar para rejeitar a hipótese nula, fixando um determinado nível de significância. Essa probabilidade é definida em dois tipos de erros. O erro do tipo I e o erro do tipo II (Casella e Berger, 2002).

O erro do tipo I é definido quando o parâmetro populacional θ está contido no espaço paramétrico da hipótese nula, porém o teste de hipótese, incorretamente, decide rejeitar a hipótese nula. Por outro lado, o erro do tipo II acontece quando o parâmetro populacional θ está contido na hipótese alternativa mas o teste indica a não rejeição da hipótese nula (Mood et al., 1974). Essas definições são melhor visualizadas na Tabela 1.1. O erro do tipo I também é conhecido como nível de significância α e o erro do tipo II como β.

Um exemplo clássico de erro do tipo I pode ser verificado na aplicação da legislação brasileira, em que um juiz evita ao máximo a condenação de um inocente, uma vez que a Constituição Federal de 1988 estabelece como premissa básica (hipótese nula) a presunção de inocência descrita em seu artigo 5, inciso LVII, "*ninguém será considerado culpado até o trânsito em julgado de sentença penal condenatória*" (BRASIL, 1988). Nesse mesmo contexto, o erro do tipo II seria a não condenação de um cidadão culpado.

Bussab e Morettin (2010) descrevem a função poder do teste (π) como a probabilidade de se rejeitar a hipótese nula, dado um valor qualquer de μ, especificado ou não pela hipótese alternativa, ou $\pi = 1 - \beta$. Em outras palavras é o complementar da probabilidade de erro do tipo II.

Segundo Walpole et al. (1993), o tamanho de amostra escolhido para se ter um bom poder de teste para um dado α é obtido, em testes unilaterais, por:

$$n = \frac{(z_\alpha + z_\beta)^2 \sigma^2}{E^2} \tag{1.1}$$

Tabela 1.1: Tipos de erro em testes de hipóteses

Decisão	Situação	
	H_0 é verdadeira	H_0 é falsa
Não rejeitar H_0	Decisão correta	Erro do tipo II
Rejeitar H_0	Erro do tipo I	Decisão correta

onde z_α é o quantil da distribuição normal para um nível de significância α e z_β é o quantil da distribuição normal para um erro do tipo II β e E é o erro máximo de estimativa.

Em teste bilaterais é dada por:

$$n = \frac{(z_{\alpha/2} + z_{\beta/2})^2 \sigma^2}{E^2} \tag{1.2}$$

Casella e Berger (2002) comentam que o teste de hipóteses pode ser expresso pela estatística do teste $W(\boldsymbol{X_1}, ..., \boldsymbol{X_n}) = W(X)$ que é função da amostra coletada. Como descrito em Magalhães e de Lima (2008), após a mensuração do teste, as decisões podem ser tomadas ao interpretar o nível descritivo do teste, ou p-valor. O p-valor, $p(X)$ é uma estatística de teste que satisfaz $0 \leq p(X) \leq 1$ e quando apresenta valor pequeno indica que há poucas evidências de que H_0 seja verdadeira (Casella e Berger, 2002).

Cabe ao pesquisador definir quão pequeno deve ser o p-valor do teste para rejeitar a hipótese nula. Dependendo do fenômeno estudado este valor de corte pode variar, considerando o custo de ter um falso positivo ou um falso negativo.

Como exemplo, Arizola e Teixeira (2015) utilizaram um teste de hipótese para verificar se o impacto do zumbido (percepção do som sem a presença de estímulo sonoro externo) é diferente entre idosos praticantes ou não de atividades físicas. Foi verificado que o impacto do zumbido é menor entre os idosos praticantes de atividade física, com p-valor igual a 0,004, sendo esse p-valor considerado aceitável para a pesquisa.

1.2.1 Teste para média em uma população

Os testes para média são utilizados para testar se a média populacional é igual, maior ou menor ao valor definido na hipótese nula. O teste para média paramétrico pode diferir quando conhecemos ou não a variância populacional. Pode-se citar como exemplo de hipóteses testadas $H_0 : \mu_X = \mu_0$ e $H_A : \mu_X \neq \mu_0$.

1.2.1.1 Teste para média com variância conhecida

Segundo Magalhães e de Lima (2008), o teste z pode ser utilizado para testar se a média populacional μ é igual a média amostral $\bar{x}$. Neste caso, é considerado que o modelo normal é adequado aos dados. A estatística do teste z é calculada como:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1) \tag{1.3}$$

onde $\bar{x}$ é a média amostral, μ_0 é o valor utilizado para comparação com a média na hipótese nula, σ é o desvio padrão populacional conhecido e n é o tamanho da amostra. A variância populacional σ^2 é calculada por (1.4), em que N é o tamanho da população.

$$\sigma^2 = \sum_{i=1}^{N} \frac{(X_i - \mu)^2}{N} \tag{1.4}$$

O comportamento em grandes amostras é verificado em (1.5)

$$\lim_{n\to\infty} \frac{\bar{x} - \mu_0}{\sigma}\sqrt{n} = \infty \tag{1.5}$$

Verifica-se que à medida que o tamanho da amostra cresce, a estatística do teste também cresce, concluindo, consequentemente, que a média amostral é diferente da média populacional que está sendo testada.

1.2.1.2 Teste para média com variância desconhecida

Segundo Bussab e Morettin (2010), o teste t de Student é adequado para testar variáveis que seguem uma seguem a distribuição normal com média μ e variância σ^2 desconhecida.

A estatística do teste é dada por:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{n-1} \tag{1.6}$$

onde $\bar{x}$ é a média amostral, μ_0 é o valor utilizado para comparação com a média na hipótese nula, s é o desvio padrão da amostra, calculado em (1.7), e n é o tamanho da amostra. A estatística do teste, segue uma distribuição t de Student com $n-1$ graus de liberdade.

$$s^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{x})^2}{n-1} \tag{1.7}$$

O comportamento em grandes amostras é verificado em (1.8)

$$\lim_{n\to\infty} \frac{\bar{x} - \mu_0}{s}\sqrt{n} = \infty \tag{1.8}$$

Verifica-se que à medida que o tamanho da amostra cresce, a estatística do teste também cresce, concluindo, consequentemente, que a média amostral é diferente da média populacional que está sendo testada.

1.2.2 Comparação de variâncias entre duas populações

Segundo Magalhães e de Lima (2008), os testes de comparação de variâncias são importantes como procedimentos preliminares para testes de comparações de médias entre duas populações pois conhecer se as populações testadas possuem variâncias iguais é determinante na escolha do teste utilizado, como verificado na próxima subseção.

Seja X uma normal com média μ_X e variância σ_X^2 e Y uma normal com média μ_Y e variância σ_Y^2. Desejamos testar as hipóteses $H_0 : \sigma_X^2 = \sigma_Y^2$ e $H_A : \sigma_X^2 \neq \sigma_Y^2$.

A estatística do teste é dada por:

$$F = \frac{s_X^2}{s_Y^2} \tag{1.9}$$

onde s_X^2 é a variância amostral de X e s_Y^2 a variância amostral de Y. A estatística F do teste segue o modelo de *Fisher-Snedecor* com $n_X - 1$ e $n_Y - 1$ graus de liberdade.

n_X e n_Y são os tamanhos amostrais de X e Y, respectivamente.

1.2.3 Comparação de médias entre duas populações

Segundo Magalhães e de Lima (2008), o teste t de Student também pode ser utilizado para comparar média entre duas populações. A independência ou não das amostras é um fator importante a ser considerado quando da comparação de média entre duas amostras, assim como o valor das variâncias. É utilizado aqui o teste de comparação de variâncias entre as duas populações.

Assim, serão citados os casos em que as amostras são pareadas, em que as amostras são independentes e com variância conhecida, independentes com variâncias desconhecidas mas iguais e independentes com variâncias desconhecidas e diferentes.

1.2.3.1 Teste t para amostras pareadas

Segundo Magalhães e de Lima (2008), o teste t para amostras pareadas é adequado quando há uma dependência entre os elementos de amostras testadas. Esse caso é utilizado para comparar duas médias populacionais sendo que para cada unidade amostral é realizada duas medições da característica de interesse. Geralmente essas medidas são tomadas antes e após uma dada intervenção. As medidas tomadas antes e depois serão representadas por X_i e Y_i respectivamente.

Seja $D_i = Y_i - X_i$, $\bar{D} = \frac{\sum_{i=1}^{n} D_i}{n}$ e $s_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2$. As hipóteses do teste são $H_0 : \mu_D = 0$ e $H_A : \mu_D \neq 0$.

A estatística do teste é é dada por:

$$\frac{\bar{D} - \mu_D}{\sqrt{s_D^2/n}} \sim t_{n-1} \tag{1.10}$$

e segue uma distribuição t de Student com $n - 1$ graus de liberdade.

1.2.3.2 Teste t para amostras independentes com variância conhecida

Segundo Magalhães e de Lima (2008), consideremos o caso em que queremos comparar as médias de duas populações independentes em que conhecemos as vari-

âncias. A informação dos valores das variâncias populacionais pode ser estimada via estudos anteriores ou experimentos similares. Considere X e Y variáveis aleatórias representando a característica de interesse em cada população com médias μ_X e μ_Y, respectivamente.

Seja $\bar{D} = \bar{X} - \bar{Y}$, e $Var(\bar{D}) = \sigma_X^2/n_1 + \sigma_Y^2/n_2$. As hipóteses do teste são $H_0 : \mu_D = 0$ e $H_A : \mu_D \neq 0$. O teste é dado por:

$$\frac{\bar{D} - \mu_D}{\sqrt{\sigma_X^2/n_1 + \sigma_Y^2/n_2}} \sim N(0,1) \tag{1.11}$$

e segue uma distribuição Normal com média 0 e variância 1.

1.2.3.3 Teste t para amostras independentes com variâncias desconhecidas iguais

Magalhães e de Lima (2008) citam o caso em que queremos comparar as médias de duas populações independentes em que não conhecemos suas variâncias. Neste caso, pode ser realizado um teste de comparação de variâncias entre as duas populações a fim de verificar se são iguais ou diferentes. Caso o teste resulte em que as duas populações testadas possuem variâncias iguais, este caso é adequado.

Seja $\bar{D} = \bar{X} - \bar{Y}$ e seja a variância:

$$s_c^2 = \frac{(n_1 - 1)s_X^2 + (n_2 - 1)s_Y^2}{(n_1 - 1) + (n_2 - 1)} \tag{1.12}$$

Considere as hipóteses do teste $H_0 : \mu_D = 0$ e $H_A : \mu_D \neq 0$. Assim, a estatística do teste é dada por:

$$\frac{\bar{D} - \mu_D}{\sqrt{s_c^2(1/n_1 + 1/n_2)}} \sim t_{n_1+n_2-2} \tag{1.13}$$

e segue uma distribuição t de Student com $n_1 + n_2 - 2$ graus de liberdade.

1.2.3.4 Teste t para amostras independentes com variâncias desconhecidas diferentes

Complementarmente à última subseção, o caso em que queremos comparar as médias de duas populações independentes em que não conhecemos as variâncias e

que, via testes adequados ou conhecimento prévio, as variâncias das populações são diferentes, este teste é adequado (Magalhães e de Lima, 2008).

Considere $\bar{D} = \bar{X} - \bar{Y}$ e a variância de D como:

$$\sigma^2_{\bar{D}} = \frac{s^2_X}{n_1} + \frac{s^2_Y}{n_2} \tag{1.14}$$

Considere as hipóteses do teste $H_0 : \mu_D = 0$ e $H_A : \mu_D \neq 0$. Assim, a estatística do teste é dada por:

$$\frac{\bar{D} - \mu_D}{\sqrt{\sigma^2_{\bar{D}}}} \sim t_\nu \tag{1.15}$$

que segue uma distribuição t de Student com ν graus de liberdade. Os graus de liberdade do teste são dados por:

$$\nu = \frac{\sigma^4_{\bar{D}}}{\frac{(s^2_X/n_1)^2}{n_1-1} + \frac{(s^2_Y/n_2)^2}{n_2-1}} \tag{1.16}$$

1.3 Consistência de estimadores

Segundo Casella e Berger (2002), devemos considerar as propriedades assintóticas dos estimadores à medida que o tamanho da amostra se torna grande ou tendendo ao infinito. A estatística inferencial assintótica usa de técnicas de maneira a simplificar alguns cálculos que, seriam difíceis em populações finitas. Das propriedades dos estimadores, a que apresenta relação direta a este trabalho é a consistência, que é uma das propriedades mais importantes de um estimador.

A definição de um estimador consistente é apresentada em Casella e Berger (2002). Um estimador é dito consistente se o estimador converge em probabilidade para o valor populacional à medida que o tamanho da amostra cresce. Sejam $X_1, X_2, ..., X_n$ variáveis independentes e identicamente distribuídas seguindo a distribuição $f(x|\theta)$. Podemos construir uma sequência de estimadores $W_n(X_1, ..., X_n)$ que pode ser definida como um estimador consistente para o parâmetro θ, pertencente ao espaço paramétrico Θ, se as equações (1.17) e (1.18) forem válidas.

$$\lim_{n\to\infty} Var_\theta W_n = 0 \tag{1.17}$$

$$\lim_{n\to\infty} Vies_\theta W_n = 0 \tag{1.18}$$

Em outras palavras, é desejável que o estimador seja consistente pois ao aumentar o tamanho da amostra, ele se aproximará do valor real do parâmetro estimado. Para que o estimador seja consistente, o estimador utilizado deve diminuir sua variância à medida que o tamanho da amostra cresce e apresentar um viés tendendo à zero.

A investigação deste trabalho parte da propriedade de consistência dos estimadores, investigando o impacto do tamanho da amostra na variância desses e no resultados de testes inferenciais influenciados pela pequena variância, estimada cada vez menor à medida que o tamanho da amostra cresce.

Capítulo 2

Amostragem Complexa

2.1 Introdução

Segundo Kish e Frankel (1974), os principais métodos estatísticos inferenciais foram criados com a suposição de Amostragem Aleatória Simples (AAS). Assumir que as observações são independentes é uma escolha que facilita os cálculos e simplifica as interpretações. Porém, nem sempre o estudo apresenta essas condições ideais.

Segundo Chambers e Skinner (2003), uma das primeiras perguntas que o pesquisador deve fazer quando está trabalhando com dados provenientes de um plano amostral complexo é como e onde o tipo de amostragem influencia na análise. A amostragem é considerada complexa quando a amostra é coletada utilizando conglomerados, estratificação e/ou probabilidade desigual de seleção, ou seja, quando duas ou mais técnicas de amostragem são combinadas, capturando assim a complexidade do fenômeno.

Heeringa et al. (2010) complementam que o pesquisador pode aproveitar o conhecimento da teoria amostral para gerar resultados mais precisos. Livros-Texto e programas estatísticos utilizados por pesquisadores, de maneira geral, não discutem a forma de trabalhar com os dados que foram coletados em um plano amostral que envolva amostragem por conglomerados, estratificação e probabilidade desigual de seleção.

A aplicação de métodos estatísticos sem levar em consideração as características do plano amostral gera tanto estimativas pontuais erradas quanto estimativas

erradas da variância do estimador, podendo levar a diagnósticos equivocados do problema estudado. Pessoa et al. (1997) mostram que o uso de técnicas convencionais, que baseiam-se em um plano amostral aleatório simples com variáveis independentes e identicamente distribuídas, podem influenciar, principalmente a estimação das variâncias das estimativas pontuais.

O artigo de Silva et al. (2006) exemplifica a necessidade de se incorporar o plano amostral com objetivo de melhorar a qualidade das estimativas. Nesse trabalho, é comparado o resultado com a aplicação e a não aplicação do plano amostral nas estimativas.

Este capítulo abordará o cálculo de pesos amostrais, a estimação de variância e realização de testes de hipóteses em amostragem complexa. Por fim, apresentará a PNAD, uma pesquisa amostral complexa de abrangência nacional realizada no Brasil.

2.2 Peso amostral

O peso amostral (W) é uma variável auxiliar utilizada para expandir estimativas amostrais para estimativas populacionais. Seu cálculo varia dependendo do tipo de técnica amostral utilizada. Sua fórmula é dada pelo inverso da probabilidade de seleção daquela unidade amostral. Segundo Lohr (2009), uma estimativa para o tamanho populacional é dada pela soma dos pesos amostrais selecionados, ou seja, uma forma usual de verificar se os pesos amostrais foram calculados corretamente é soma-los e esse resultado, $\sum_{i=1}^{n} W_i = \hat{N}$, deve ser igual ao tamanho populacional.

Em amostragem complexa, o cálculo do peso se mostra complicado pois devemos considerar a especificidade de seleção da amostra em cada estágio. Um exemplo seria uma amostra de 2 estágios como descrita a seguir: suponha que temos 3 localidades e nestas localidades o número de domicílios está descrito na Tabela 2.1.

O plano amostral consiste em selecionar 2 localidades e 4 domicílios, dentre as localidades selecionadas no primeiro estágio, sem reposição.

O peso calculado para esse exemplo é mostrado na Tabela 2.2. Pode-se notar que a soma dos pesos atribuídos as observações selecionadas na amostra final é de

Tabela 2.1: Dados populacionais fictícios para exemplificação

Localidade	Domicílios
1	10
2	20
3	30
Total	60

60, tamanho da população pesquisada.

Tabela 2.2: Pesos atribuídos ao exemplo

Localidade	Primeiro Estágio	Segundo Estágio	Peso Final
1	$W_{11} = 3$	$W_{21} = 10/4 = 2,5$	7,5
2	$W_{12} = 1,5$	$W_{22} = 20/4 = 5$	7,5
3	$W_{13} = 1$	$W_{23} = 30/4 = 7,5$	7,5

Assim, segundo Lohr (2009), para o caso geral para l estágios $w_{NEWj} = \prod_{k=1}^{l} w_{kj}$, em que k é o número de estágios da amostra e j identifica a unidade amostral observada.

2.3 Teste para média com variância desconhecida

Para realização de teste de hipótese para a média μ, considerando as características do plano amostral, primeiro devemos estimar a variância da média amostral. A variância é calculada com base no plano amostral utilizado. As variâncias podem ser calculadas segundo a Tabela 2.3, dependendo da técnica de amostragem utilizada para a coleta dos dados.

Tabela 2.3: Estimação de médias em amostragem

Técnica de amostragem	Média	Variância da média
AAS com reposição	$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$	$V(\bar{x}) = \frac{\sigma^2}{n}$
AAS sem reposição	$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$	$V(\bar{x}) = (1-f)\frac{S^2}{n}$
Estratificada	$\bar{x}_{st} = \sum_{h=1}^{H} W_h \bar{x}_h$	$V(\bar{x_{st}}) = \sum W_h^2 (1-f_h)\frac{s_h^2}{n_h}$
Sistemática	$\bar{x}_s = \sum_{i=1}^{n} \frac{x_{[i+(i-1)(k-1)]}}{n}$	$V(\bar{x_s}) = \frac{N-1}{N} S^2 - \frac{k(n-1)}{N} S_{ws}^2$
Conglomerado	$\bar{\bar{x}} = \frac{1}{nM} \sum_{i=1}^{n} \sum_{j=1}^{M} x_{ij}$	$V(\bar{\bar{x}}) = \frac{1-f}{nM} S^2[1 + (M-1)\rho]$

Para a AAS sem reposição, $S^2 = \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{N-1}$ é a variância populacional, escrita dessa forma para facilitar os cálculos, e $f = \frac{n}{N}$ (Cochran, 1977). Para a amostragem

estratificada, N_h é o tamanho do estrato, $W_h = \frac{N_h}{N}$ é o peso do estrato e $f_h = \frac{n_h}{N_h}$ é a fração amostral do h-ésimo estrato.

Para a amostragem sistemática, k é o inteiro mais próximo de $\frac{N}{n}$ e $S^2_{ws} = \frac{1}{k(n-1)} \sum_{i=1}^{k} \sum_{j=1}^{n} (X_{ij} - \bar{X}_i)^2$.

Para a amostragem por conglomerados de tamanhos iguais, M é o tamanho do conglomerado e ρ é o coeficiente de correlação intraclasse dado em (2.1).

$$\rho = \frac{E(X_{ij} - \bar{\bar{X}})(X_{il} - \bar{\bar{X}})}{E(X_{ij} - \bar{\bar{X}})^2} \tag{2.1}$$

Todas as equações do Capítulo 1 podem ser reformuladas, utilizando o novo tratamento para cálculo de estimativas pontuais e de variabilidade, incorporando o plano amostral. Um exemplo é a Equação 1.6 que pode ser reescrita como em (2.2).

$$t = \frac{\bar{x}_{complex} - \mu_0}{s_{complex}/\sqrt{n}} \tag{2.2}$$

onde $\bar{x}_{complex}$ é a média amostral, e $s_{complex}$ é o desvio padrão da amostra, ambos calculados incorporando o plano amostral utilizado na pesquisa. A estatística do teste t também segue uma distribuição t de Student, mas agora com o número de graus de liberdade dados por (Särndal et al., 2003; Heeringa et al., 2010):

- Se o plano possui conglomerados e estratificação, então o número de graus de liberdade é o número de conglomerados menos o número de estratos.
- Se o plano não possui conglomerados, então o número de graus de liberdade é o número de observações menos o número de estratos.
- Se o plano não possui estratos, então o número de graus de liberdade é o número de conglomerados menos um.

Importante notar que o comportamento para grandes amostras é mantido como em (2.3).

$$\lim_{n\to\infty} \frac{\bar{x}_{complex} - \mu_0}{s_{complex}} \sqrt{n} = \infty \tag{2.3}$$

2.4 Estimação de variância em pesquisas amostrais complexas

À medida que a complexidade inerente ao desenho amostral vai aumentando, a variância da média pode apresentar uma função não linear da amostra. Isso pode tornar o cálculo da variância algebricamente e computacionalmente difícil.

Para tratar esse caso pode-se chegar a valores aproximados da variância da média via Linearização de Taylor, que aproxima a não linearidade da variância como uma função linear dos totais, como visto em Särndal et al. (2003).

Em amostragem complexa, a linearização de Taylor (ou do inglês, *Taylor Series Linearization* - TSL) é uma técnica bastante utilizada para estimar a variância de qualquer total ponderado. Isso inclui estimadores de razão, coeficientes de regressão e coeficiente de correlação. Em geral, num plano amostral complexo, a média pode ser estimada por (Särndal et al., 2003):

$$\bar{y}_w = \frac{\sum_h \sum_\alpha \sum_i y_{h\alpha i} w_{h\alpha i}}{\sum_h \sum_\alpha \sum_i w_{h\alpha i}} = \frac{x}{z} \tag{2.4}$$

onde $y_{h\alpha i}$ é uma medida na unidade i, no conglomerado α no estrato h e $w_{h\alpha i}$ é o peso correspondente.

Como forma de calcular uma aproximação para a variância de funções não-lineares, seja $f(x, z)$ a função de variância não linear, com 2 ou mais variáveis, que queremos aproximar via TSL. Note que

$$f(x,z) \approx f(x_0,z_0) + (x-x_0)\left[\frac{df}{dx}\right]_{x=x_0,z=z_0} + (z-z_0)\left[\frac{df}{dz}\right]_{x=x_0,z=z_0} \tag{2.5}$$

onde x_0 e z_0 são os valores de x e z em que a função $f(x, z)$ será aproximada.

Seja $A = \left[\frac{df}{dx}\right]_{x=x_0,z=z_0}$ e $B = \left[\frac{df}{dz}\right]_{x=x_0,z=z_0}$, temos que:

$$var(f(x,z)) \approx A^2 var(x) + B^2 var(z) + 2AB cov(x,z) \tag{2.6}$$

No caso da média apresentada em (2.4), $f(x, z) = \frac{x}{z}$, $A = \frac{1}{z_0}$ e $B = -\frac{x_0}{z_0^2}$. Dessa

forma:

$$Var\left(\frac{x}{z}\right) \approx \frac{Var(x) + \left(\frac{x_0}{z_0}\right)^2 Var(z) + 2\left(\frac{x_0}{z_0}\right) Cov(x,z)}{z_0^2} \tag{2.7}$$

tendo como estimativa:

$$var(\bar{y}_w) \approx \frac{var(x) + \bar{y}_w^2 var(z) + 2\bar{y}_w cov(x,z)}{z^2} \tag{2.8}$$

A aplicação do TSL já é difundida nos softwares estatísticos. No SAS, o *PROC SURVEYMEANS* e *PROC SURVEYREG*. No R, o pacote *survey* pode ser utilizado. No SPSS, o módulo *Complex Samples*.

Heeringa et al. (2010), citam que, alternativamente, também pode-se calcular essa variância via técnicas de reamostragem como a *Balanced Repeated Replication* (BRR), *Jackknife Repeated Replication* (JRR) e o *Bootstrap*. Eles formam uma classe de métodos não paramétricos para calcular a variância de estimativas amostrais e utilizam replicações de subamostragem do banco de observações amostrais para para estimar variâncias para estatísticas lineares e não lineares.

Segundo Heeringa et al. (2010) cada um desses métodos de replicação seguem um algorítimo de 5 passos:

1) Replicação amostral ($r = 1, ..., R$) da amostra completa da pesquisa são definidas pelas regras do método (BRR, JRR ou *Bootstrap*), em que R é o número desejado de replicações;

2) Replicação de pesos - pesos amostrais são recalculados para cada replicação para criar $r = 1, ..., R$;

3) Estimativas ponderadas de estatísticas de interesse são calculadas para a amostra completa e separadamente para cada replicação subamostral (usando os pesos replicados);

4) As variâncias estimadas são calculadas pelas fórmulas específicas do BRR, JRR ou do *Bootstrap*;

5) São construídos os intervalos de confiança (ou testes de hipóteses) baseados nas estimativas pontuais, variância e graus de liberdades obtidos em cada um dos métodos.

Tabela 2.4: Ultimate cluster de tamanho $n_a = 4$

PSU	SSU	HU
1	1	2
1	1	5
1	2	1
1	2	4

Heeringa et al. (2010) explicam os pressupostos adotados pelos principais *softwares* para cálculo do TSL, JRR e BRR. São eles:

1) Em amostragem de mais de um estágio as unidades primárias de amostragem (do inglês *Primary Sample Unit* - PSU) são consideradas estimadas com reposição. A correção de populações finitas da amostra do primeiro estágio é ignorada. As estimativas das variâncias amostrais são levemente superestimadas;

2) A amostragem multiestágio da PSU selecionada resulta em um único *Ultimate Cluster* de observações para aquela PSU, como por exemplo na Tabela 2.4. Esse *Ultimate Cluster* contém todas unidades secundárias de amostragem (*Secondary Sample Unit* - SSU) e, consequentemente, todos os *Housing Units* - HU .

Métodos de estimação de variância baseado em *Ultimate Cluster* sorteiam os componentes de variância de múltiplos estágios para uma única fórmula de um estágio que requer conhecimento apenas do primeiro estágio e dos identificadores das PSU para cálculo do resultado final. Todas as fontes de variabilidade do PSU são capturadas na variância estimada composta.

A técnica *Balanced Repeated Replication* (BRR) é um método de *half sample* que foi desenvolvido especificamente para estimar variâncias amostrais em planos amostrais com dois PSUs por estrato. O termo *half sample* é utilizado para uma subamostra em que tenha metade dos elementos da amostra principal. O método surgiu do conceito de de formar réplicas ao escolher uma metade da amostra. Segundo Kish e Frankel (1974), a variância via BRR pode ser calculada pela Equação (2.9).

$$Var_{BRR}\{g(S)\} = \frac{1-f}{2k}\sum_{i=1}^{k}\{[g(H_i - g(S))]^2 + [g(C_i - g(S))]^2\} \qquad (2.9)$$

onde S denota a amostra completa, H_i a i-ésima *half sample* formada incluindo um dos PSU de cada estrato e C_i é o complementar de H_i. São formadas k *half samples*.

A técnica *Jacknife Repeated Replication* (JRR) é comumente utilizada para uma grande variedade de planos amostrais complexos incluindo desenhos amostrais em que existem duas ou mais Unidades Primárias de Amostragem (PSUs). Segundo Kish e Frankel (1974), a técnica utiliza a informação de variabilidade que cada estrato possui na variabilidade total. Para isso, a técnica consiste em retirar algumas observações, e posteriormente recalcular os pesos. A Equação (2.10) mostra como a variância é calculada.

$$Var_{JRR}\{g(S)\} = \frac{1-f}{2}\sum_{i=1}^{h}\{[g(J_i - g(S))]^2 + [g(CJ_i - g(S))]^2\} \tag{2.10}$$

onde J_i é a reamostragem obtida ao retirar da amostra completa uma seleção no i-ésimo estrato mas incluindo 2 vezes o complementar da seleção naquele estrato. CJ_i é o complementar da reamostragem produzida em J_i.

O *Bootstrap* é o terceiro e menos comumente utilizado para estimativas de variância em amostras complexas. Comparações simuladas e verificações epíricas mostraram que na maioria das aplicações amostrais, razoavelmente grandes, o *Bootstrap* não oferece vantagem perante aos métodos de TSL, BRR ou o JRR (Kovar et al., 1988).

Como Skinner et al. (1989) apontaram, o método *Bootstrap* permite a validação direta de distribuições amostrais das estimativas e não necessita da normalidade de muitas observações para formulação de intervalos de confiança. Assim, o método *Bootstrap* possui aplicações específicas em analisar planos amostrais complexos de tamanho pequeno.

Heeringa et al. (2010) compararam os métodos TSL, JRR e BRR e verificaram que eles são não viesados e produzem resultados idênticos no caso especial em que o estimador de interesse é uma estatística linear, como o peso amostral total. Para estimativas não-lineares, comumentes aplicados em amostragem complexa, ou coeficientes de regressão, o TSL e o JRR tendem a apresentar um viés menor (menor MSE) para estimativas de variância amostral. Por outro lado, os intervalos de confiança construídos utilizando as estimativas BRR ou *Bootstrap* produzem melhor cobertura nominal.

Neste trabalho, quando for necessário utilizar técnicas de estimação de variância para amostras complexas, como as citadas nesta Seção, utilizaremos a Linearização de Taylor, pois segundo Rao e Wu (1985) assintoticamente as técnicas JRR, BRR e Linearização de Taylor convergem para valores bem próximos.

2.5 A PNAD

A Pesquisa Nacional por Amostra de Domicílios (PNAD) ocorre anualmente, em anos em que não há o censo, e investiga características gerais da população. Esse formato foi mantido até meados de 2015 quando o IBGE decidiu mudar a metodologia, e passou a adotar exclusivamente a PNAD contínua, a qual é analisada trimestralmente (IBGE, 2015).

São medidas variáveis demográficas que subsidiam estudos sobre educação, trabalho, rendimento, habitação, dentre outros. O levantamento dessas informações é um instrumento importante para formulação de políticas públicas, validação e avaliação de pesquisas aplicadas em diversas áreas de estudo.

A Pesquisa Nacional por Amostra de Domicílios é realizada por meio de amostra probabilística de domicílios obtida em três estágios de seleção, ou seja, possui um plano amostral complexo visto que utiliza estratificação, probabilidade desigual de seleção e conglomeração. As unidades primárias são os municípios, as secundárias os setores censitários e as terciárias os domicílios.

No primeiro estágio, é utilizada a probabilidade desigual de seleção. As unidades primárias são os municípios, que são classificados em autorrepresentativos, com probabilidade 1 de pertencer à amostra, e não autorrepresentativos. Os municípios não autorrepresentativos passam por um processo de estratificação e, em cada estrato, são selecionados com reposição e com probabilidade proporcional à população residente obtida no Censo Demográfico.

Os setores censitários, unidades secundárias, são selecionados, em cada município da amostra, com probabilidade proporcional e com reposição, sendo utilizado o número de unidades domiciliares existentes por ocasião do Censo Demográfico como medida de tamanho.

No terceiro estágio são selecionados, com equiprobabilidade, em cada setor censitário da amostra, as unidades domiciliares para investigação das características dos moradores e da habitação (IBGE, 2012)

Nascimento Silva et al. (2002) citam a complexidade do plano amostral da PNAD e exemplifica como a estratificação, a conglomeração, probabilidades desiguais de seleção e ajustes dos pesos para calibração podem ser considerados na análise dos dados. Indicam, também, os cuidados que o pesquisador deve ter ao analisar os dados dessa pesquisa.

Capítulo 3

Material e métodos

3.1 Introdução

Este Capítulo apresenta a descrição dos materiais e métodos utilizados neste trabalho. Serão utilizados bancos de dados populacionais simulados, amostras provenientes de diferentes tipos de amostragem e a Pesquisa Nacional por Amostra de Domicílios (PNAD) de 2012 e 2014. O Software SAS 9.4 será utilizado em todo o trabalho.

3.2 Material

O material utilizado é composto por bancos de dados simulados para verificar o comportamento em planos amostrais aleatório simples, estratificado e complexo. Também será feito um estudo de caso a partir dos dados da Pesquisa Nacional por Amostra de Domicílios (PNAD), que possui amostragem complexa como descrita na Seção 2.5.

As variáveis do estudo foram geradas utilizado o algoritmo exposto na Tabela 3.1 como proposto em Wicklin (2013). As variáveis normal e log-normal foram geradas com média M e desvio padrão S, de tamanho populacional N. A função *rand('Normal', M, S)* do SAS gera números aleatórios de uma normal com média M e desvio padrão S. Os vetores com os valores finais simulados estão alocados nas variáveis *normal* e *lognormal*.

Tabela 3.1: Algorítmo de geração das variáveis do estudo

$\phi = \sqrt{S^2 + M^2}$
$\mu = log(M^2/\phi)$
$\sigma = \sqrt{log(\phi^2/M^2)}$
faça i de 1 até N
{
x = rand('Normal', μ, σ)
$lognormal = e^x$
$normal$ = rand('Normal', M, S)
}

Fonte: Wicklin (2013) com adaptações.

Para o primeiro ensaio, considerando o plano amostral aleatório simples, foi gerada uma população, normal e log-normal, com média 1.000 e coeficiente de variação de 10%, 50% e 100% de tamanho 200 milhões, similar ao tamanho da população brasileira, a fim de mostrar a influência da variabilidade dos dados.

Para o segundo ensaio, considerando o plano amostral estratificado, com objetivo de simular as regiões brasileiras, foi gerada uma população similar a do Brasil, utilizando os dados da PNAD 2014, com tamanho de 190.610.814 habitantes. A Tabela 3.2 apresenta os valores utilizados para gerar as informações de renda e população. A população foi simulada utilizando as distribuições log-normal e normal, com as médias e variâncias para os estratos descritas na Tabela 3.2. A média de renda salarial para o Brasil, segundo a PNAD 2014, foi de R$1.195,53.

Tabela 3.2: Dados utilizados para simulação da renda - PNAD 2014

Região	População	Amostra PNAD	Média	Desvio Padrão
Sudeste	80.364.410	91.026	1.409,66	2.970,9
Nordeste	53.081.950	88.418	766,01	1.581,6
Sul	27.300.000	48.133	1.436,67	2.281,9
Norte	15.864.454	47.023	851,27	1.641,5
Centro-Oeste	14.000.000	32.156	1.466,94	2.723,2
Total	190.610.814	306.756	1.195,53	xx

Para o terceiro ensaio, considerando o plano amostral complexo, foi gerada a população de uma variável log-normal, de tamanho 200 milhões, com média 1.000 e desvio padrão 1.000 (CV=100%). Essa população possui 4 estratos de tamanhos distintos. Com objetivo de obter tamanhos diferentes, o primeiro estrato foi classi-

ficado com as observações até o 10º percentil da população, o segundo estrato com as observações entre o 10º e o 40º percentis, o terceiro estrato entre o percentil 40º e o 80º, e o último estrato entre o 80º e o 100º percentis.

Dentro de cada estrato existem 10 conglomerados, também de tamanho diferentes. Para classificar os conglomerados dentro de cada estrato foi utilizada a técnica de classificação K-Médias. Essa técnica de classificação iterativa, segundo MacQueen et al. (1967), minimiza a distância entre as observações de um vetor ao centróide do conglomerado, construindo assim conglomerados com características semelhantes. Para a população gerada, os conglomerados obtidos, assim como os estratos, foram de tamanhos distintos.

Nesse ensaio, foi realizada uma amostragem complexa em 2 estágios. No primeiro estágio foram selecionados 4 conglomerados em cada estrato, definidos via alocação ótima utilizando a metodologia presente na Seção 3.2.1. O maior conglomerado de cada estrato está na amostra com probabilidade igual a 1.

Essa particularidade se justifica pois em pesquisas amostrais usualmente a presença de um município na amostra é importante pois ele representa grande parte das observações estudadas. Imagine, por exemplo, se o município de São Paulo não fosse escolhido em uma pesquisa amostral municipal a nível de Brasil, a amostra poderia não representar bem a realidade brasileira pois São Paulo possui um papel importante por ser o município mais populoso do país.

Como consequência da cauda pesada da distribuição log-normal, o quarto estrato apresentou 3 conglomerados muito grandes, assim, os três maiores conglomerados foram escolhidos com probabilidade 1. O conglomerado restante foi escolhido utilizando o método de probabilidade proporcional ao tamanho.

A amostragem utilizada, calculada pelo critério de alocação ótima para amostragem complexa em dois estágios, é descrita na Seção 3.2.1. Foi utilizado um custo total de R\$100.000 ($C = 100.000$), com custo fixo de R\$5.000 ($c_0 = 5.000$), custo de seleção no primeiro estágio de R\$25.000 ($c_1 = 25.000$) e custo de seleção no segundo estágio de R\$5 ($c_2 = 5$).

No segundo estágio, as unidades amostrais foram selecionadas via amostragem aleatória simples. Segundo Lohr (2009), esse tipo de amostragem é considerado

complexo pois envolve a mistura de diferentes técnicas como a estratificação e probabilidade desigual de seleção.

Nos dois primeiros ensaios (amostragem AAS e estratificada) foram simuladas duas variáveis. Uma normal, a fim de verificar a influência em distribuições bem comportadas e simétricas e uma log-normal, a fim de verificar o comportamento em distribuições assimétricas. O terceiro ensaio (amostragem complexa) foi simulado utilizando somente a distribuição log-normal.

A escolha dessas variáveis se deu pelo fato de que no mundo real pode-se encontrar distribuições das mais variadas formas e para que o teste fosse feito nesses diferentes comportamentos. Por exemplo, na Figura 3.1 é apresentado o histograma do rendimento mensal de todas as fontes para pessoas de 10 anos ou mais de idade, variável de estudo neste trabalho e disponível na PNAD 2014. Percebe-se que a maioria das observações estão concentradas em valores mais baixos, porém, poucos apresentam grandes rendimentos. Isso caracteriza como uma distribuição de cauda pesada.

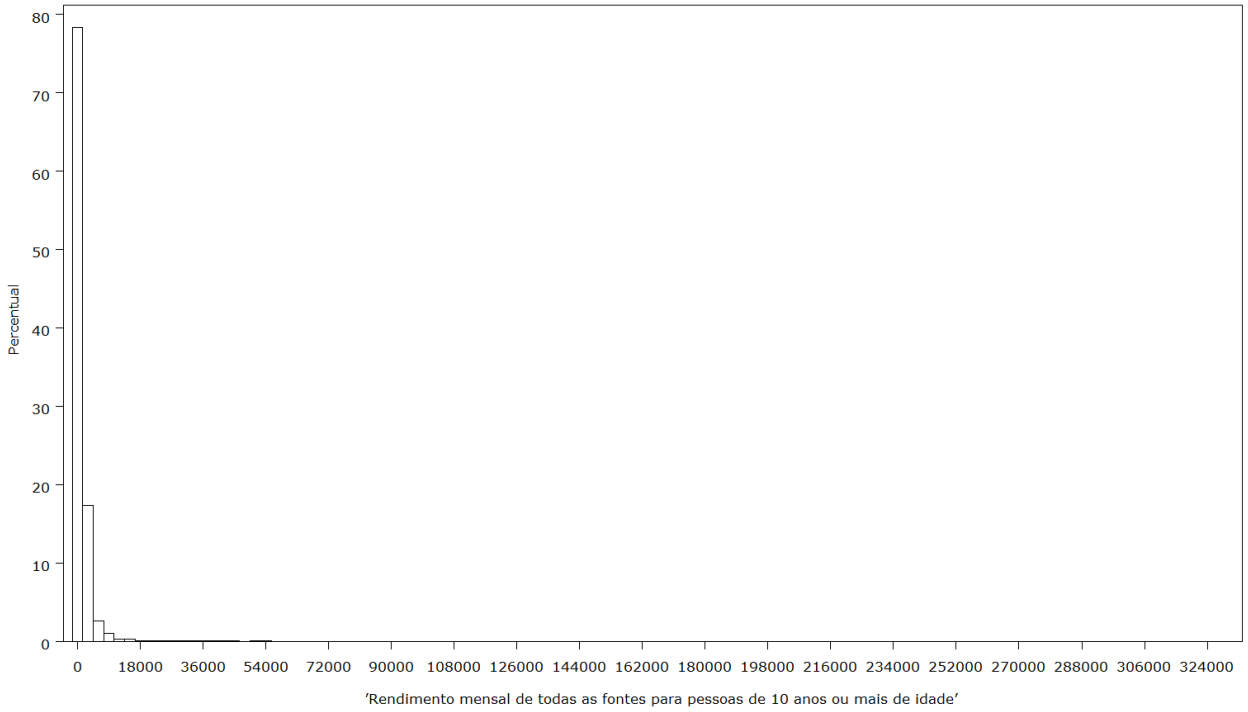

Figura 3.1: Histograma do rendimento mensal de todas as fontes para pessoas de 10 anos ou mais de idade - PNAD 2014

Para o quarto ensaio, foi utilizado o banco de dados da PNAD 2014.

3.2.1 Tamanho amostral

Segundo Cochran (1977), no planejamento de uma pesquisa amostral, uma decisão importante é a decisão do tamanho amostral. Se a amostra for muito grande, pode-se estar gastando recursos desnecessariamente, se muito pequena os resultados podem não ser muito úteis. Assim, a teoria amostral fornece parâmetros para considerar o problema de maneira inteligente.

Serão apresentados os tamanhos de amostras mínimos, como definidos em Cochran (1977), que foram utilizados nos três primeiros ensaios descritos na Seção 3.2, aleatório simples, estratificado e complexo em dois estágios.

Para a amostragem aleatória simples, o tamanho mínimo da amostra para estimativa da média, e considerando populações infinitas, é calculado pela Equação (3.1).

$$n = \frac{z^2_{\frac{\alpha}{2}}\sigma^2}{\epsilon^2} \tag{3.1}$$

onde $z_{\frac{\alpha}{2}}$ é o quantil $(1 - \frac{\alpha}{2})$ de uma distribuição $N(0, 1)$, ϵ é o erro máximo de estimação e σ^2 é a variância estimada da variável estudada. As populações deste trabalho são todas na casa dos milhões, justificando assim as fórmulas para populações infinitas.

Considerando o coeficiente de variação (CV) da variável, a Equação (3.1) pode ser reescrita como na Equação (3.2).

$$n = \frac{z^2_{\frac{\alpha}{2}}CV^2}{r^2} \tag{3.2}$$

em que r é o desvio percentual em relação à média.

Para amostra estratificada, considerando custos iguais e alocação ótima de Neyman, o tamanho de amostra para cada estrato é dado na Equação (3.3).

$$n_h = n\frac{N_h S_h}{\sum N_h S_h} \tag{3.3}$$

em que h é o estrato, N_h é o tamanho populacional do estrato, S_h é o desvio-padrão

populacional do estrato e n é o tamanho total da amostra dado na Equação (3.4).

$$n = \frac{\sum \frac{W_h^2 s_h^2}{w_h}}{\left(\frac{\epsilon}{z_{\frac{\alpha}{2}}}\right)^2 + \frac{1}{N}\sum W_h s_h^2} \tag{3.4}$$

Por fim, para o caso da amostragem complexa em dois estágios descrita na Seção 3.2, o método utilizado para alocação ótima da amostra dentre os estratos e os conglomerados de uma amostragem complexa em dois estágios, em que o primeiro estágio é estratificado e o segundo por conglomerado é descrito, para esse caso específico, em Khan et al. (2006). Esse trabalho faz referência para amostragem multivariada, porém, neste estudo foi feita a adaptação das fórmulas para o caso em que temos apenas uma variável.

A população em estudo possui N unidades primárias de seleção (PSU), de maneira que $N = \sum_{h=1}^{L} N_h$. O índice h identifica os L estratos presente no primeiro estágio da pesquisa. Considere, também, que existam M_{hi} unidades secundárias de seleção (SSU) no estrato h e conglomerado i, de maneira que $M_{h0} = \sum_{i=1}^{N_h} M_{hi}$.

Considere a Equação de custos totais em (3.5).

$$C = c_0 + \sum_{h=1}^{L}\left(c_{1h}n_h + c_{2h}\sum_{i=1}^{n_h} m_{hi}\right) \tag{3.5}$$

onde C é o custo total da amostragem, c_0 o custo fixo, c_1 e c_2 os custos unitários de cada unidade amostral no primeiro e segundo estágio respectivamente. A variável n_h identifica o número de unidades amostrais no primeiro estágio do estrato h e a variável m_{hi} identifica o número de unidades amostrais no estrato h e conglomerado i.

A alocação ótima do tamanho amostral, para o segundo estágio, no estrato h e conglomerado i (m_{hi}^*) é dada em (3.6) (Khan et al., 2006):

$$m_{hi}^* = \frac{M_{hi}S_{hiy}^2}{\overline{M}_h}\sqrt{\frac{c_{1h}}{A_h c_{2h}}} \tag{3.6}$$

Assim, partindo da Equação (3.5), a alocação ótima para o primeiro estágio (n_h^*) é

dada na Equação (3.7)

$$n_h^* = \frac{(C - c_0) W_h \sqrt{A_h} / \sqrt{c_{1h}}}{\sum_{h=1}^{L} \left(W_h \sqrt{A_h c_{1h}} + \frac{W_h \sqrt{c_{2h}}}{N_h} \sum_{i=1}^{N_h} \frac{M_{hi}}{\overline{M}_h} S_{hiy}^2 \right)} \tag{3.7}$$

onde $A_h = S_{hb}^2 - \frac{1}{N_h} \sum_{i=1}^{N_h} \frac{M_{hi}}{M_h^2} S_{hiy}^2$. S_{hb}^2 é a variância entre as médias das unidades primárias de amostragem e S_{hiy}^2 é a variância entre as subunidades dentre as unidades primárias de seleção, dadas pelas Equações (3.8) e (3.9).

$$S_{hb}^2 = \frac{\sum_{i=1}^{N} (\overline{Y}_i - \overline{\overline{Y}})^2}{N - 1} \tag{3.8}$$

$$S_{hiy}^2 = \frac{\sum_{i=1}^{N} \sum_{j=1}^{M} (Y_{ij} - \overline{Y})^2}{M(N - 1)} \tag{3.9}$$

3.3 Métodos

Foram realizados ensaios englobando 3 tipos de diferentes métodos de amostragem e um exemplo em um banco de dados real. O primeiro ensaio utilizou o caso de amostragem aleatória simples (sem reposição), o segundo para amostragem estratificada, o terceiro amostragem complexa e o exemplo será aplicado na PNAD 2014. Foram analisados resultados para o teste de média, segundo as técnicas descritas nos Capítulos 1 e 2.

Para cada ensaio será aplicada a técnica de subamostragem mista, descrita na Seção 3.3.3, a fim de verificar se as conclusões tomadas serão iguais utilizando os dados da amostra maior e da menor.

3.3.1 Primeira amostra

Com o intuito de verificar o impacto do tamanho da amostra no teste da média, este ensaio irá observar o comportamento da amplitude do intervalo de confiança da média e do número de vezes em que esse intervalo engloba a média populacional ao aumentar o tamanho mínimo da amostra.

Serão utilizados os seguintes passos:

Passo 1: Selecionar i subamostras utilizando o método descrito (amostragem aleatória simples, estratificada ou complexa) do ensaio e utilizando o tamanho de amostra mínimo como verificado em Cochran (1977). Nesse estudo utilizamos $i = 100$.

Passo 2: Para cada uma das i amostras retiradas no passo anterior, verificar o número de vezes em que o intervalo de confiança gerado englobou a média populacional.

Passo 3: Para cada uma das i amostras, verificar a amplitude do intervalo de confiança gerado. Este passo tem por objetivo verificar a influência do tamanho da amostra no resultado do teste de média, como visto na Seção 1.3, resultado da propriedade de consistência dos estimadores.

Walpole et al. (1993) mostra a relação entre o p-valor de um teste de hipótese e o intervalo de confiança. O valor de μ_0 é coberto por um intervalo de confiança bicaudal de tamanho $(1 - \alpha)$ dado pela Equação (3.10) se o p-valor de um teste de hipóteses bicaudal com as hipóteses $H_0 : \mu = \mu_0$ e $H_A : \mu \neq \mu_0$ é maior que α.

$$\left(\overline{X} - \frac{t_{\alpha/2,n-1}S}{\sqrt{n}}, \overline{X} + \frac{t_{\alpha/2,n-1}S}{\sqrt{n}}\right) \tag{3.10}$$

Portanto, se μ_0 está contido em um intervalo de confiança de tamanho $(1-\alpha)$, o teste de hipótese de tamanho α não rejeita a hipótese nula e se μ_0 não está contido nesse a hipótese nula é rejeitada.

Dessa maneira, a influência do tamanho da amostra no p-valor de testes de média será avaliada por essa relação com o intervalo de confiança, representada, aqui, pelo valor da amplitude desse intervalo. A amplitude é calculada pela diferença entre o limite superior e o limite inferior do intervalo de confiança gerado em cada amostra.

Passo 4: Repetir os passos anteriores utilizando o tamanho de amostra como o tamanho de amostra calculado no passo 1 multiplicado pelas constantes 10, 100 e 1.000. Esses valores multiplicados se justificam pela possibilidade de diferenças de variabilidade entre as variáveis de um banco de dados amostral.

Pode acontecer de uma pesquisa conter variáveis de estudo apresentando diferentes variâncias. Caso ocorra, por exemplo 3 variáveis, de mesma média, em um

banco de dados, porém variâncias diferentes o tamanho mínimo da amostra será o tamanho mínimo da variável com maior variância, fazendo com que a variável com menor variabilidade seja coletada mais vezes do que o necessário. Isso pode afetar conclusões inferenciais via teste de hipóteses, como demonstrado nos Capítulos 1 e 2.

Utilizando a PNAD 2014 como exemplo, podemos comparar o coeficiente de variação da variável binária "Nasceu no município de residência" (V501), com respostas possíveis "Sim" e "Não", com a variável quantitativa "Rendimento mensal de todas as fontes para pessoas de 10 anos ou mais de idade" (V4720). Os resultados podem ser avaliados na Tabela 3.2. O coeficiente de variação (CV) da variável V4720 é igual a 226% enquanto o da variável V501 é igual a 82%. Caso utilizemos o cálculo do tamanho de amostra como aleatório simples via coeficiente de variação, considerando 95% de confiança e 5% de erro, o tamanho de amostra mínimo para a segunda é 21 vezes maior que a primeira. Assim, em uma pesquisa que precisemos obter essas duas informações, a variável V501 será coletada um tamanho 21 vezes maior que o mínimo necessário. O tamanho amostral da PNAD 2014 é de 306.756.

Tabela 3.3: Variáveis PNAD 2014 - Medidas Resumo

Variável	Amostra PNAD	Média	Desvio-padrão	CV	Amostra AAS
V0501	425.627	0,6	0,49	82%	369
V4720	306.756	1195,52	2697,28	226%	7.852

Passo 5: Por fim, construir tabela comparativa dos resultados para análise da influência do tamanho da amostra no teste de média.

3.3.2 Subamostragem

Com objetivo de comparar os resultados do teste da média em situações distintas, serão avaliados três casos. São eles:

1) Base "principal": serão calculados i intervalos de confiança da média nas amostras retiradas da população descrita na Seção 3.2 com tamanho 1.000 vezes o tamanho mínimo apresentado na Seção 3.2.1. Nesse estudo utilizamos $i = 100$.

2) Base "subamostragem": serão calculados i intervalos de confiança da média

para a subamostragem retirada das i amostras da base "principal", definida no passo anterior, utilizando o tamanho de amostra mínimo, como apresentado na Seção 3.2.1, calculado com as informações definidas pelo pesquisador.

3) Caso "misto": serão calculados i intervalos de confiança da média utilizando as estimativas combinadas entre a base "principal" e a de "subamostragem". Será utilizada a estimativa pontual da base "principal" e a estimativa de variabilidade da base de "subamostragem".

A utilização do "caso misto" se justifica pela propriedade de consistência dos estimadores (Seção 1.3). Como neste caso o pesquisador possui informações de um grande banco de dados, onde a estimativa está mais próxima do valor populacional, porém com variabilidade da média pequena, esse caso é proposto. O objetivo é utilizar a informação da estimativa pontual mais acurada com a variabilidade desejada pelo pesquisador ao definir um nível de significância e erro, e consequentemente, um tamanho mínimo para sua amostra.

Para comparação de resultado serão avaliados o percentual de vezes em que o intervalo de confiança gerado englobou a média populacional.

Os três casos serão avaliados para planos amostrais aleatório simples, estratificado e complexo.

A técnica de subamostragem a seguir é utilizada para diminuir o tamanho de grandes amostras. Para aplicá-la, serão utilizados os seguintes passos:

Passo 1: Identificar o plano amostral da grande amostra e guardar o peso dessa grande amostra.

Passo 2: Utilizar o mesmo plano amostral da grande amostra, e utilizando a grande amostra como população, calcular o peso da subamostra. Guarde o peso gerado por essa subamostra.

Passo 3: A fim de incorporar os diferentes pesos, como mostrado na Seção 2.2, o peso da amostra final será dado pelo produto do peso da grande amostra com o peso da subamostra.

3.3.3 Técnica de Subamostragem mista em Grandes Amostras

A técnica de subamostragem mista é útil quando o pesquisador dispõe de um banco de dados amostral grande. O tamanho amostral pode interferir nas conclusões sobre o fenômeno testado, portanto essa técnica auxilia o pesquisador a utilizar as informações de estimativas pontuais do banco grande, definindo uma subamostra para utilização das estimativas de variabilidade dos estimadores, dado um erro e nível de confiança definido pelo pesquisador.

A utilização da subamostragem para cálculo da variabilidade do estimador é justificada pela possível influência do grande volume de dados obtido pelo pesquisador na variância estimada. Assim, o resultado do teste utiliza uma estimativa pontual mais próxima do parâmetro observado (consequência da amostra de tamanho grande) e uma estimativa de variabilidade do estimador em uma escala obtida pelo tamanho mínimo amostral definido pelo pesquisador. Ressalta-se que para utilização desta técnica o tamanho mínimo definido pelo pesquisador é menor que o banco de dados amostral disponível ao pesquisador.

A técnica proposta para subamostragem mista em grandes amostras segue os seguintes passos:

Passo 1: Utilize as informações do grande banco de dados para estimar a variância da variável de estudo (S_B^2).

Passo 2: Com base na informação da variabilidade estimada no passo 1 calcule o tamanho mínimo de amostra utilizando erro e nível de confiança desejado. O caso da amostragem aleatória simples está representado na Equação 3.11.

$$n_R = \frac{z_{\frac{\alpha}{2}_R} s_B^2}{\epsilon_R^2} \tag{3.11}$$

Passo 3: Utilizando o tamanho amostral obtido no passo 2 (n_R), faça uma subamostragem seguindo os passos apresentados na Seção 3.3.2, não esquecendo de reajustar os pesos, conforme o passo 3.

Passo 4: Para realização do teste de hipótese de subamostragem mista utilize a estatística pontual estimada no banco principal e a estatística de variabilidade

estimada no banco de subamostragem.

$$t = \frac{\bar{x}_B - \mu_0}{s_R/\sqrt{n_R}} \tag{3.12}$$

em que as estimativas identificadas com a letra B se referem ao banco principal e as estimativas identificadas pela letra R se referem ao banco de subamostragem.

Os graus de liberdade relativos aos três cenários apresentados neste trabalho estão resumidos na Tabela 3.4.

Tabela 3.4: Graus de liberdade para o teste t por tipo de amostragem

Amostragem	Graus de Liberdade do teste t
AAS	#observações - 1
Estratificada	#observações - #estratos
Complexa[1]	#conglomerados - #estratos

[1]Para amostragem complexa com estratos no primeiro estágio e conglomerados no segundo estágio.

Passo 5: Caso o teste de subamostragem mista continue rejeitando a hipótese nula tem-se indícios que o fenômeno é de fato verificado. Caso passe a não rejeitar verificar o comportamento adotando outros níveis de confiança.

3.3.4 Teste de Efeito do Tamanho da Amostra na Significância de Testes de Hipóteses

Esse teste é proposto com o objetivo de verificar se o tamanho amostral está influenciando no nível de significância de testes de hipóteses. É aplicado quando o pesquisador acredita que o tamanho amostral pode ser muito grande e possivelmente estaria influenciando no resultado do p-valor de seu teste. Esse fenômeno pode acontecer como consequência da possível redução da variância do estimador que pode ocorrer ao aumentar o tamanho da amostra, como visto no Capítulo 2 .

Para realização do Teste de Efeito do Tamanho da Amostra na Significância de Testes de Hipóteses, são realizados os seguintes passos:

Passo 1: Realize o teste de hipótese, definido pelo pesquisador, no banco "principal". Guarde o valor da variância do estimador e o resultado da conclusão do teste, com base no p-valor obtido.

Passo 2: Aplique a técnica de subamostragem mista, apresentada na Seção 3.3.3, no banco "principal". Guarde a variância do estimador e o resultado da conclusão do teste de hipóteses realizado, com base no p-valor obtido.

Passo 3: Compare as variâncias do banco "principal" e do banco de "subamostragem", utilizando um teste para igualdade de variâncias, como o teste visto na Seção 1.2.2. Caso as variâncias sejam diferentes, continue para o próximo passo. Caso as variâncias forem iguais, o tamanho amostral não está influenciando na variabilidade do estimador.

Passo 4: Compare os resultados das conclusões dos testes de hipóteses aplicados nos passos 1 e 2. Caso os resultados sejam divergentes, existe uma evidência de que o tamanho amostral está influenciando para a conclusão do teste de hipóteses.

Capítulo 4

Análise dos Resultados

4.1 Amostra Aleatória Simples

4.1.1 Primeira amostra

Este ensaio tem objetivo de verificar a influência do tamanho da amostra na conclusão via teste de hipóteses do pesquisador utilizando amostragem aleatória simples, como descrita na Seção 3.2. O teste foi feito para as distribuições normal e log-normal como descrito na Seção 3.3. Foram selecionadas, via estudo simulado, amostras das distribuições estudadas com média R$1.000.

Com a finalidade de verificar o impacto do tamanho da amostra em diferentes condições de desvio padrão, foram utilizados valores de 100, 500 e 1.000. Esses valores foram escolhidos para simular distribuições normais platicúrticas, mesocúrticas e leptocúrticas, ao adotar um coeficiente de variação de 10%, 50% e 100%, como na Figura 4.1.

Foi calculado o tamanho mínimo das amostras, com metodologia descrita na Seção 3.2.1, considerando um erro de R$100,00 e níveis de confiança de 90%, 95% e 99%. Para verificar a influência do tamanho da amostra foram selecionadas amostras maiores que esse mínimo de 10 a 1.000 vezes, com quantitativos apresentados na Tabela 4.1.

Foram implementados os intervalos de confiança da média, com $\alpha = 0,05$, estimada para os desvios-padrão e os parâmetros citados acima. A Figura 4.2 mostra o

Tabela 4.1: Tamanho amostral - Amostragem aleatória simples

CV	Confiança	Tamanho amostral			
		Mínimo	10x	100x	1.000x
10%	90%	3	30	300	3.000
	95%	4	40	400	4.000
	99%	7	70	700	7.000
50%	90%	68	680	6.800	68.000
	95%	97	970	9.700	97.000
	99%	166	1.660	16.600	166.000
100%	90%	271	2.710	27.100	271.000
	95%	385	3.850	38.500	385.000
	99%	664	6.640	66.400	664.000

comportamento do intervalo de confiança para a média em amostras desenhadas via amostragem aleatória simples (AAS), para a variável normal e a Figura 4.3 mostra o comportamento para a variável log-normal.

Essas figuras mostram que a amplitude dos intervalos de confiança para a média diminui à medida em que o tamanho da amostra cresce, o que ratifica a propriedade de consistência do estimador da média (Seção 1.3). O resultado é similar para as variáveis normal e log-normal. As informações completas estão na Tabela A.1).

Percebe-se, pelas Figuras 4.2 e 4.3, que à medida em que a amostra fica maior que seu tamanho mínimo, a amplitude do intervalo de confiança vai diminuindo. Esse resultado impacta diretamente a rejeição dos testes de hipóteses para a média.

A Tabela 4.2 resume o percentual de vezes em que a média amostral esteve dentro do intervalo de confiança verificado nos resultados apresentados acima nas 100 amostras retiradas para cada combinação de nível de confiança e tamanho de amostra pelas variáveis normal e log-normal. O resultado mostra que o percentual de vezes que a amostra está dentro do intervalo de confiança é similar ao nível de confiança adotado.

4.1.2 Subamostragem

Com objetivo de verificar o comportamento comparativo dos testes de hipóteses da média nos três casos descritos na Seção 3.3.2 foram calculados intervalos de confiança para amostras aleatória simples retiradas da população, como apresentado

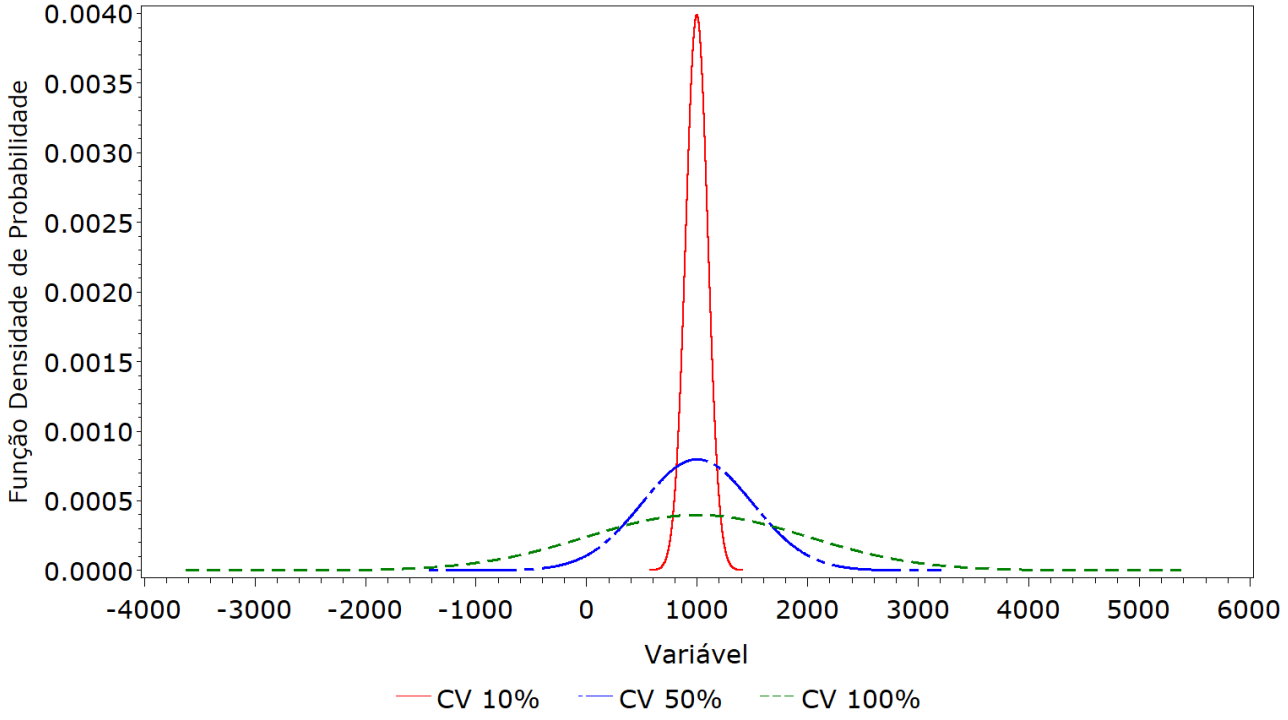

Figura 4.1: Distribuições normais simuladas

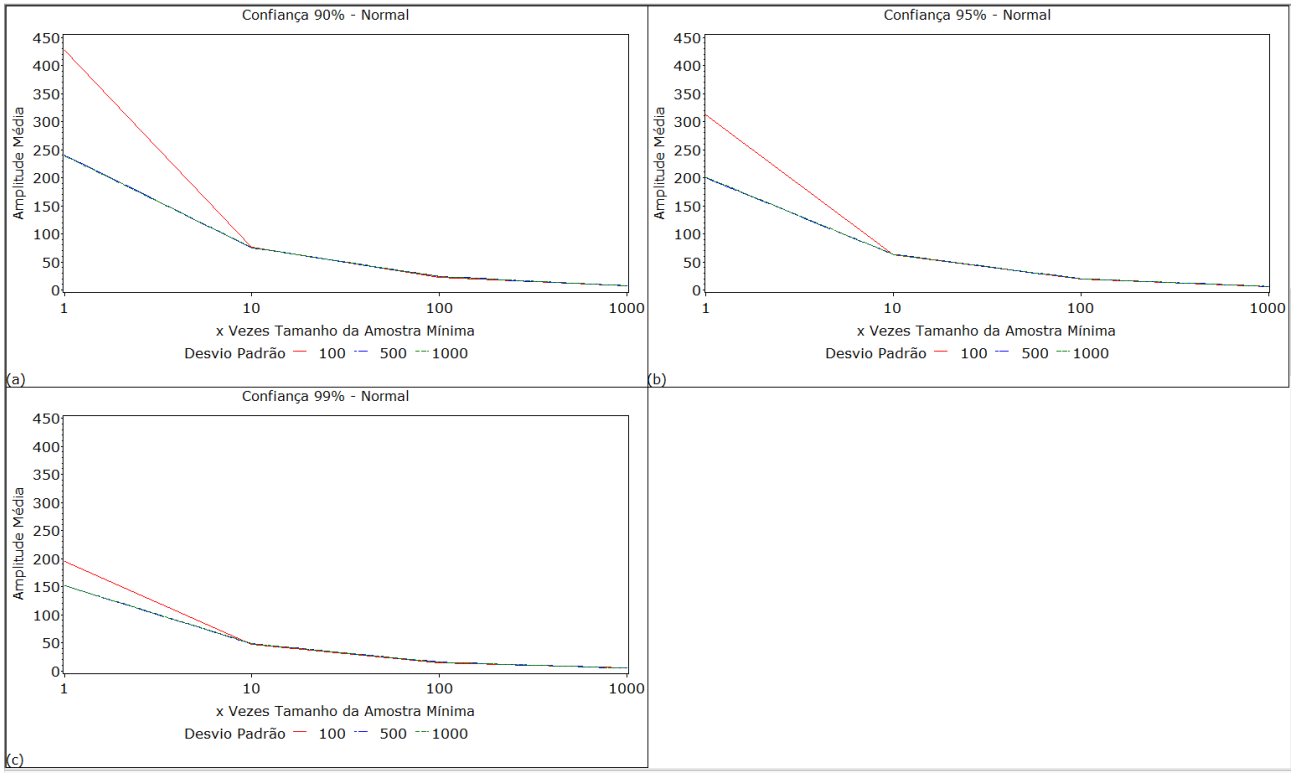

Figura 4.2: Amplitude do intervalo de confiança ao aumentar o tamanho mínimo da amostra - Normal

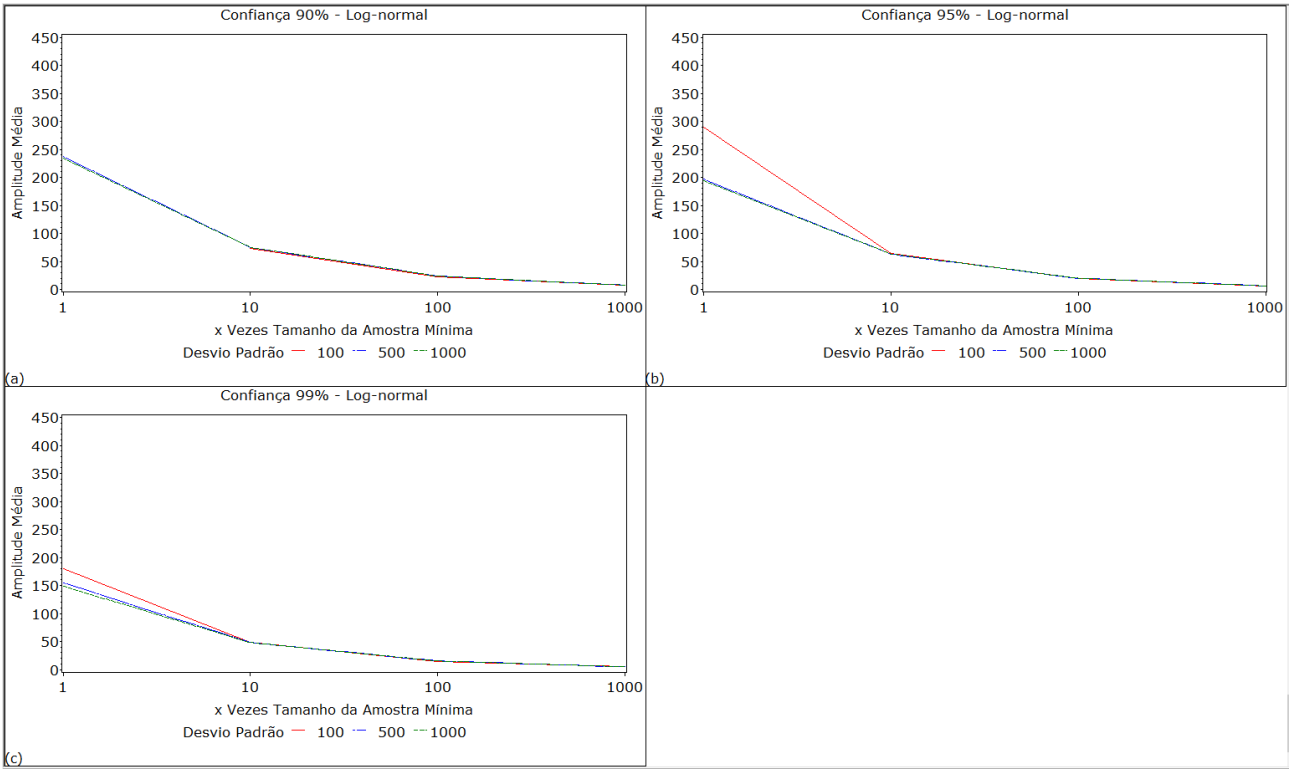

Figura 4.3: Amplitude do intervalo de confiança ao aumentar o tamanho mínimo da amostra - Log-normal

na Seção 4.1.1.

Para a base "principal" foram consideradas as amostras de tamanho 1.000 vezes o tamanho mínimo apresentadas na Seção 3.2.1. Para a base de "subamostragem" foram retiradas subamostras utilizando a técnica de subamostragem descrita em 3.3.2 e para o caso "misto" foram calculados intervalos de confiança ($\alpha = 0,05$) utilizando a estimativa pontual da base "principal" com a estimativa de variabilidade da base de "subamostragem".

O processo foi realizado 100 vezes para cada caso. A Tabela 4.3 mostra o resultado compilado para os caso de CV 10%, 50% e 100% e níveis de confiança amostrais de 90%, 95% e 99%. O erro utilizado para cálculo do tamanho mínimo amostral foi de R$100,00. A Tabela A.2 mostra o resultado da amplitude média nos 100 ensaios realizados.

Percebe-se pela Tabela 4.3 que o percentual de acertos para a base "principal" é similar à base de "subamostragem", porém o caso misto englobou a média populacional em 100% dos seus intervalos gerados. Esse fato ocorre pois a estimativa da

Tabela 4.2: Percentual de acertos da estimativa da média nas simulações para AAS com desvio padrão e tamanho de amostra variáveis

Variável	Coeficiente de Variação (CV)	Vezes	IC 90%	IC 95%	IC 99%
Log-normal	10%	1	95%	96%	93%
		10	95%	92%	93%
		100	93%	95%	96%
		1.000	97%	97%	96%
	50%	1	94%	94%	93%
		10	93%	98%	96%
		100	93%	95%	96%
		1.000	95%	97%	94%
	100%	1	94%	93%	91%
		10	93%	95%	91%
		100	95%	95%	96%
		1.000	95%	95%	96%
Normal	10%	1	94%	96%	99%
		10	93%	98%	94%
		100	97%	96%	93%
		1.000	96%	96%	97%
	50%	1	97%	97%	95%
		10	95%	95%	95%
		100	97%	92%	96%
		1.000	87%	97%	97%
	100%	1	92%	98%	97%
		10	95%	95%	95%
		100	95%	96%	93%
		1.000	98%	97%	95%

média da base "principal" é mais próxima da populacional e a variância utilizada, da base de "subamostragem", é maior, permitindo assim que o percentual de acertos seja maior.

Tabela 4.3: Percentual de acertos da estimativa da média nas simulações para AAS com desvio padrão e tamanho de amostra variáveis - Bases "principal", "subamostragem" e "caso misto"

Confiança	Variável	CV	Amostragem - AAS		
			Principal (1.000 vezes)	Subamostragem (AAS)	Misto (AAS)
90%	Log-normal	10%	97%	94%	100%
		50%	95%	91%	100%
		100%	95%	92%	100%
	Normal	10%	96%	93%	100%
		50%	87%	95%	100%
		100%	98%	94%	100%
95%	Log-normal	10%	97%	91%	100%
		50%	97%	94%	100%
		100%	95%	93%	100%
	Normal	10%	96%	96%	100%
		50%	97%	97%	100%
		100%	97%	99%	100%
99%	Log-normal	10%	96%	94%	100%
		50%	94%	95%	100%
		100%	96%	93%	100%
	Normal	10%	97%	95%	100%
		50%	97%	100%	100%
		100%	95%	97%	100%

4.2 Amostra Estratificada

4.2.1 Primeira amostra

Este ensaio tem objetivo de verificar a influência do tamanho da amostra na conclusão via teste de hipóteses do pesquisador utilizando amostragem aleatória estratificada. A população em análise foi gerada utilizando algoritmo descrito na Seção 3.2. O teste foi feito para as distribuições normal e log-normal, como apresentado na Seção 3.3. Foram selecionadas, via simulação, amostras das distribuições estudadas com médias e desvios-padrão expostos na Tabela 3.2.

Com o objetivo de verificar o impacto do tamanho da amostra em testes de média realizados em amostras estratificadas, foram selecionadas 100 amostras estratificadas por região. Este trabalho procurou gerar uma população de tamanho similar à população brasileira, com dados extraídos dos microdados da PNAD 2014.

O tamanho amostral base foi calculado via alocação de Neyman, descrito na Seção 3.2.1, com nível de confiança de 90%, 95% e 99% e erro de R$100.

Com o objetivo de verificar a influência do tamanho da amostra nas conclusões inferenciais via testes de hipóteses, além da amostra em tamanho mínimo, foram selecionadas amostras de tamanho 10, 100, e 1.000 vezes maiores que as definidas pelo tamanho mínimo, com quantitativos apresentados na Tabela 4.4. Os resultados estão representados na Figura 4.4. Os valores detalhados podem ser verificados em A.3.

Tabela 4.4: Tamanho amostral - Amostragem estratificada

Região	Tamanho amostral			
	Minimo	10x	100x	1.000x
Norte	63	630	6.300	63.000
Nordeste	202	2.020	20.200	202.000
Sudeste	574	5.740	57.400	574.000
Sul	150	1.500	15.000	150.000
Centro-Oeste	92	920	9.200	92.000
Total	1.081	10.810	108.100	1.081.000

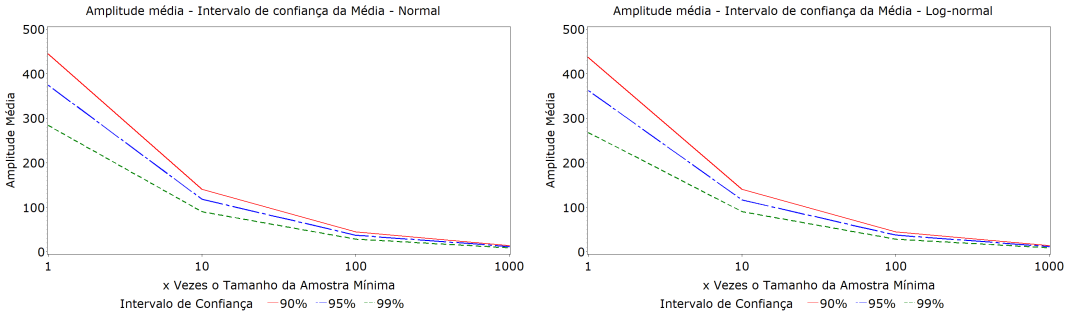

Figura 4.4: Amplitude Média - Intervalo de Confiança - Amostragem Estratificada - Normal e Log-normal

Percebe-se que, à medida que a amostra fica maior que seu tamanho mínimo, a amplitude do intervalo de confiança vai diminuindo. Como a relação de impacto nos resultados dos testes de hipóteses está sendo verificada pela amplitude, verificamos que, assim como o resultado obtido na Seção 4.1, esse resultado impacta diretamente a rejeição dos testes de hipóteses para a média em amostragem estratificada.

O percentual de acertos, verificado na Tabela 4.5, dos intervalos de confiança, com $\alpha = 0,05$, para a média nas amostras obtidas, neste ensaio de amostragem

estratificada, estiveram próximos de seus níveis de confiança definidos.

Tabela 4.5: Percentual de acertos dos intervalos de confiança gerados pelas amostras estratificadas por tamanho de amostra e nível de confiança do teste - Normal e Log-normal

Confiança	Variável	Vezes	Percentual de acertos IC
90%	Log-normal	1	91%
		10	95%
		100	95%
		1000	98%
	Normal	1	99%
		10	97%
		100	94%
		1000	94%
95%	Log-normal	1	88%
		10	94%
		100	94%
		1000	94%
	Normal	1	97%
		10	97%
		100	99%
		1000	96%
99%	Log-normal	1	93%
		10	89%
		100	96%
		1000	93%
	Normal	1	97%
		10	94%
		100	92%
		1000	97%

4.2.2 Subamostragem

A aplicação da subamostragem na amostra estratificada definida na Seção 3.2 resultou os valores apresentados nas Tabelas 4.6 e 4.7. A primeira tabela apresenta os resultados para o caso em que a primeira amostra, de tamanho grande (1.000 vezes o tamanho mínimo), foi retirada com plano de amostragem estratificada.

Com objetivo de diminuir o tamanho da grande amostra e readequar a variância dos estimadores ao nível escolhido, vamos imaginar que o pesquisador aplicou um plano de subamostragem estratificada e utilizou o peso combinado dos dois planos

de maneira correta. Neste caso a soma dos pesos da amostra final coincide com o tamanho populacional.

A Tabela 4.7 apresenta os resultados para o caso em que a primeira amostra, de tamanho grande (1.000 vezes o tamanho mínimo), foi retirada com plano de amostragem estratificada. Com objetivo de diminuir o amostra anterior, porém, vamos imaginar novamente que o pesquisador aplicou um plano de subamostragem aleatório simples. Foi utilizado o peso combinado dos dois planos. Neste caso a soma dos pesos da amostra final não coincide com o tamanho populacional, não sendo a maneira correta de realizar a subamostragem.

Os resultados do percentual de acertos da média nos intervalos de confiança foram similares nos dois casos pois a soma dos pesos no segundo caso, apesar de não coincidir com a população, apresentou valor similar a esta (ver Tabela 4.12).

Tabela 4.6: Percentual de acertos da estimativa da média nas simulações para amostragem estratificada com desvio padrão e tamanho de amostra variáveis - Bases "principal", "subamostragem" e "caso misto" - Subamostragem com plano amostral estratificado

Variável	Confiança	Amostragem - AE		
		Principal (1.000 vezes)	Subamostragem - AE	Misto - AE
Log-Normal	90%	97%	93%	100%
	95%	98%	96%	100%
	99%	94%	97%	100%
Normal	90%	94%	97%	100%
	95%	94%	95%	100%
	99%	94%	99%	100%

Tabela 4.7: Percentual de acertos da estimativa da média nas simulações para amostragem estratificada com desvio padrão e tamanho de amostra variáveis - Bases "principal", "subamostragem" e "caso misto" - Subamostragem com plano amostral aleatório simples

Variável	Confiança	Amostragem - AE		
		Principal (1.000 vezes)	Subamostragem - AAS	Misto - AAS
Log-Normal	90%	97%	93%	100%
	95%	98%	91%	100%
	99%	94%	95%	100%
Normal	90%	94%	96%	100%
	95%	94%	94%	100%
	99%	94%	95%	100%

Novamente, a base "principal" e a "subamostragem" apresentaram valores, de

acerto da média no intervalo de confiança, semelhantes nos dois casos, porém, ao utilizar a técnica "mista", o percentual de acerto sobe para 100% em todos os casos. Os valores das amplitudes médias dos intervalos de confiança obtidos nas simulações podem ser observados na Tabela A.4.

4.3 Amostra Complexa

4.3.1 Primeira amostra

Para este ensaio, o objetivo é verificar a influência do tamanho da amostra na conclusão via teste de hipóteses do pesquisador utilizando amostragem complexa. A amostragem é complexa em dois estágios. No primeiro estágio amostragem estratificada com probabilidades desiguais de seleção e no segundo estágio amostragem aleatória simples. A população em análise foi gerada utilizando algoritmo descrito na Seção 3.2.

O teste foi feito para a distribuição log-normal como apresentado na Seção 3.3. Foram selecionadas, via simulação, amostras da população gerada de uma log-normal com média 1.000 e desvio padrão 1.000 (CV=100%). Os estratos e conglomerados populacionais foram gerados como descrito na Seção 3.2, e a população possui tamanho de 200 milhões de observações divididas pelos conglomerados. As informações de cada estrato e conglomerado estão detalhadas nas Tabelas A.6 e A.7.

Com o objetivo de verificar o impacto do tamanho da amostra em testes de média realizados em amostras complexas, foram selecionadas 100 amostras. Além da amostra em tamanho mínimo, foram selecionadas amostras de tamanho 10, 100 e 1.000 vezes maiores que as definidas pelo tamanho mínimo (m^*_{hi}). Caso o tamanho mínimo multiplicado seja maior que o tamanho do conglomerado (M_{hi}) o tamanho amostral será igual ao tamanho do conglomerado. Esse fato ocorreu nos conglomerados 2, 3, 6 e 8 do estrato 4.

Como definido em 3.2, para o primeiro estágio foram selecionados 4 conglomerados por estrato. Na Tabela A.5 estão o quantitativo de observações sorteadas para cada conglomerado escolhido. Como o conglomerado selecionado é variável para

cada uma das 100 repetições, o tamanho amostral final não é fixo no caso complexo.

O valor da amplitude média e o percentual de vezes que o intervalo de confiança amostral gerado englobou a média populacional, considerando 95% de confiança, está representado na Tabela 4.8.

Tabela 4.8: Percentual de acertos dos intervalos de confiança gerados pelas amostras complexas e amplitude dos IC por tamanho de amostra

Amostragem	Vezes	Amplitude IC	Percentual de Acertos
Complexa	1	526,20	93%
	10	526,26	95%
	100	526,12	95%
	1.000	526,19	95%

Percebe-se que, para este caso, o percentual de acertos e a amplitude do Intervalo de confiança gerado permaneceram constantes com o aumento da amostra até 1.000 vezes.

Um resultado interessante aconteceu quando o plano amostral não foi incorporado no cálculo nas estimativas da variância. A Tabela 4.9 mostra o resultados da Tabela 4.8 caso o plano amostral fosse considerado uma amostragem aleatória simples ou uma amostragem estratificada. As estimativas pontuais nos três casos são iguais, porém, as estimativa de variâncias são bem diferentes.

Tabela 4.9: Estimação equivocada de parâmetros em amostragem complexa - Amplitude do IC e percentual de acerto da média dentro do IC por tamanho de amostra - Amostragem aleatória simples e estratificada

Amostragem	Vezes	Amplitude IC	Percentual de Acertos
AAS	1	70,98	20%
	10	22,44	8%
	100	7,09	6%
	1.000	2,35	1%
Estratificada	1	39,35	10%
	10	12,43	6%
	100	3,93	2%
	1.000	1,37	0%

Percebe-se que o percentual de acertos do intervalo de confiança contendo a média cai substancialmente para os dois casos. Pela propriedade de consistência dos estimadores, apresentada na Seção 1.3, ao aumentar o tamanho da amostra e

considerando os planos que, nesse caso, apresentam variância menor do que o plano complexo, o tamanho da amplitude do intervalo de confiança é bastante reduzido. Assim, ao aumentar o tamanho da amostra, o percentual de acertos do intervalo de confiança contendo a média cai pois a amplitude do intervalo vai diminuindo.

4.3.2 Subamostragem

Com objetivo de diminuir o tamanho da grande amostra e readequar a variância dos estimadores ao nível escolhido, imagine que o pesquisador aplicou um plano de subamostragem complexa e utilizou o peso combinado dos dois planos de maneira correta.

São apresentados os resultados para os casos em que a primeira amostra, de tamanho grande (1.000 vezes o tamanho mínimo), foi retirada com plano de amostragem complexo e, com objetivo de diminuir a amostra anterior, porém, o pesquisador aplicou um planos de subamostragem complexo, estratificado e aleatório simples. Foi utilizado o peso combinado dos dois planos.

Como será apresentado na Seção 4.4 o caso em que a soma dos pesos coincide com o tamanho populacional é no caso da amostragem complexa com subamostragem complexa, nos demais casos isso não ocorre.

A aplicação da subamostragem na amostra complexa definida na Seção 3.2 resultou os valores apresentados nas Tabelas 4.10 e 4.11. A primeira tabela apresenta os resultados da amplitude dos intervalos de confiança da média obtidos nas combinações de reamostragens possíveis. A segunda, mostra o percentual de acerto desse intervalo na média real.

No caso complexo em estudo, nenhuma subamostragem apresentou variância superior à em estudo, sinal de que, neste caso a subamostragem não precisa ser realizada pois a variabilidade estimada no banco "principal" não está influenciada pelo tamanho da amostra.

A diferença de valores na utilização ou não da correção de população finita (do inglês *Finite Population Correction* - FPC, dado por $1 - f = 1 - \frac{n}{N}$), foi identificada pois os conglomerados 2, 3, 6 e 8 (do estrato 4) apresentaram tamanhos amostrais iguais aos populacionais (veja na Tabela A.5) e, consequentemente, variância zero.

Como a variância é muito maior no estrato 4 (Tabela A.7) a utilização do FPC reduz o valor da variância verificada, influenciando assim os resultados.

Um resultado interessante é notado nos casos em que a subamostragem é feita de maneira aleatória simples ou estratificada em cima de uma amostragem complexa. Como o valor da variância estimada é significativamente diferente, o percentual de acerto do intervalo de confiança na média real cai muito. Esse resultado reitera a importância da aplicação de uma subamostragem de maneira correta. Nesse exemplo o correto seria a utilização de uma subamostragem complexa com a mesma estrutura do banco principal.

Tabela 4.10: Amplitude dos intervalos de confiança da média nas simulações para amostragem inicial complexa e diferentes tipos de técnicas de subamostragem - Bases "principal", "subamostragem" e "caso misto"

Tipo de Subamostragem	Principal* (complexa)	Subamostragem	Mista
Complexa Com fpc	526,2	429,6	429,6
Complexa Sem fpc	686,2	686,2	679,3
AAS	526,2	74,2	74,2
Estratificada	526,2	322,0	321,8

* Amostra de 1.000 vezes o tamanho mínimo.

Tabela 4.11: Percentual de acertos da estimativa da média nas simulações para amostragem inicial complexa e diferentes tipos de técnicas de subamostragem - Bases "principal", "subamostragem" e "caso misto"

Tipo de Subamostragem	Principal* (complexa)	Subamostragem	Mista
Complexa Com fpc	95%	89%	89%
Complexa Sem fpc	99%	99%	99%
AAS	95%	23%	20%
Estratificada	95%	32%	36%

* Amostra de 1.000 vezes o tamanho mínimo.

4.4 Pesos Utilizados

Uma observação importante é da utilização dos pesos amostrais de maneira correta. Das simulações geradas, foram verificados as somas dos pesos em cada amostra realizada. Na Tabela 4.12 pode-se perceber que no caso de realizar a subamostragem, e o recálculo dos pesos não seja efetuado, a soma dos pesos não coincide com o tamanho populacional, gerando assim estimativas erradas.

Considere P_1 o peso amostral obtido no desenho amostral da primeira amostra, P_2 o peso obtido no desenho amostra da subamostragem sem ser combinado ao obtido na primeira amostra e P_c o peso combinado obtido ao multiplicar o peso obtido na primeira amostra com o peso obtido na técnica de subamostragem utilizada.

Os únicos resultados de soma de peso que coincidiram com a população em estudo foram os casos em que não houve subamostragem e utilização do P_1 e os casos em que a técnica de subamostragem obedeceu a complexidade da técnica utilizada na primeira amostra e foi utilizado o peso combinado, P_c.

Tabela 4.12: Média da soma dos pesos obtidos nas simulações por tipo de amostragem, subamostragem e utilização de pesos

Amostragem	Subamostragem	Peso utilizado	Soma de pesos (média)	População
AAS	Nenhuma	P_1	200.000.000	200.000.000
	AAS	P_c	200.000.000	200.000.000
		P_1	200.000	200.000.000
		P_2	185.000	200.000.000
Estratificada	Nenhuma	P_1	190.610.814	190.610.814
	AAS	P_c	190.938.075	190.610.814
		P_1	190.938	190.610.814
		P_2	717.333	190.610.814
	Estratificada	P_c	190.610.814	190.610.814
		P_1	190.611	190.610.814
		P_2	717.333	190.610.814
Complexa	Nenhuma	P_1	200.000.000	200.000.000
	AAS	P_c	199.554.816	200.000.000
		P_1	199.589	200.000.000
		P_2	3.549.105	200.000.000
	Estratificada	P_c	200.367.945	200.000.000
		P_1	54.504	200.000.000
		P_2	3.549.106	200.000.000
	Complexa	P_c	200.000.000	200.000.000
		P_1	200.025	200.000.000
		P_2	3.549.106	200.000.000

4.5 Técnica de subamostragem mista na PNAD

Com o objetivo de comparar o rendimento *per capta* médio no Brasil nos anos de 2012 e 2014 foi realizado teste de hipótese, utilizando as PNADs respectivas de cada ano.

O valor do rendimento *per capta* (V4720 - Rendimento mensal de todas as fontes para pessoas de 10 anos ou mais de idade) de 2012 foi inflacionado a valores de 2014 utilizando o deflator IPCA-E do IBGE de julho de 2012 a julho de 2014, de valor 1,1226477. Os valores de renda zero e *missing* foram excluídos da base.

A Tabela 4.14 apresenta os valores das estimativas do rendimento médio na PNAD considerando o banco completo e utilizando a técnica de subamostragem mista descrita na Seção 3.3.3. O resultado mostra que não há intersecção entre os intervalos de confiança do rendimento médio na amostra PNAD "principal" nem quando utilizamos a técnica de subamostragem mista, apesar de aproximar um pouco os intervalos. Os tamanhos amostrais estão na Tabela 4.13.

Tabela 4.13: Tamanho da amostra - PNAD 2012, 2014 e subamostragem mista

Amostragem	Ano	Amostra
Principal	2014	219.288
	2012	212.520
Mista	2014	8.663
	2012	9.242

Tabela 4.14: Estimativas de rendimento médio - PNAD 2012 e 2014 - Amostragem "principal" e subamostragem mista

Amostragem	Ano	Média	Erro Padrão	Limite Inferior*	Limite Superior*
Principal	2014	1.665,2	14,43	1.636,9	1.693,5
	2012	1594,5	15,23	1.564,6	1.624,4
Mista	2014	1.665,2	23,75	1.618,6	1.711,7
	2012	1594,5	21,8	1551,8	1637,2

*Limites obtidos para intervalo de confiança de 95%.

Para realização do teste de hipótese para comparar o rendimento médio *per capta* nos dois anos, foi utilizada as fórmulas descritas na Seção 1.2.3.4. O resultado está apresentado na Tabela 4.15.

Ao realizar o Teste de Efeito do Tamanho da Amostra na Significância de Testes de Hipóteses, como descrito na Seção 3.3.4, primeiro foram comparadas as variâncias dos estimadores nos casos "principal" e "misto". O valor do teste F foi de 2,35 com p-valor menor que 0,00001. Com base nesse resultado, temos que as duas variâncias são diferentes. Depois de verificar que as variâncias dos bancos são diferentes o

próximo passo é verificar se a conclusão do teste de hipóteses ao utilizar as duas variâncias é igual.

Assim, considerando um nível de confiança de 99% e o $\mu_D = 0$ (para testar se as duas médias são iguais), a renda média do brasil nas PNADs 2012 e 2014 apresentou resultado diferente no caso "principal" enquanto o caso "misto" não apresentou evidências de que as rendas são diferentes.

Como a conclusão do teste de hipótese mudou, conclui-se que, no caso de comparação de média salarial para o brasil, o tamanho amostral da PNAD influenciou na decisão do teste de diferença de médias t.

Tabela 4.15: Tamanho da amostra - PNAD 2012, 2014 e subamostragem mista

Amostragem	Medida	Resultado
Principal	$\bar{D}$	70,68
	σ_D^2	440,43
	ν	94.985.432
	t_ν	3,37
	p-valor	0,0004
Mista	$\bar{D}$	70,68
	σ_D^2	1.038,43
	ν	9.259.863
	t_ν	2,19
	p-valor	0,0141

Por outro lado, a fim de verificar o valor da diferença entre as duas metodologias, foi utilizada a Equação (1.2.3.4) com valor de μ_D variando de R\$0,00 a R\$25,00 com incrementos de R\$1,00. A Figura 4.5 mostra que a diferença dos rendimentos médios entre as PNADS de 2012 e 2014 começa a apresentar p-valor menor que 0,01 (bicaudal) a partir de R\$17,00 para a base "principal". Para o caso "misto" verificamos na aplicação que já apresenta diferença com o $\mu_D = 0$.

É importante ressaltar que a Pesquisa Nacional Por Amostras de Domicílios (PNAD) possui objetivos que vão além da produção de estimativas gerais para o Brasil. O tamanho amostral grande deve-se pelo fato dessa pesquisa possibilitar estudos locais, possibilitando maiores desagregações regionais.

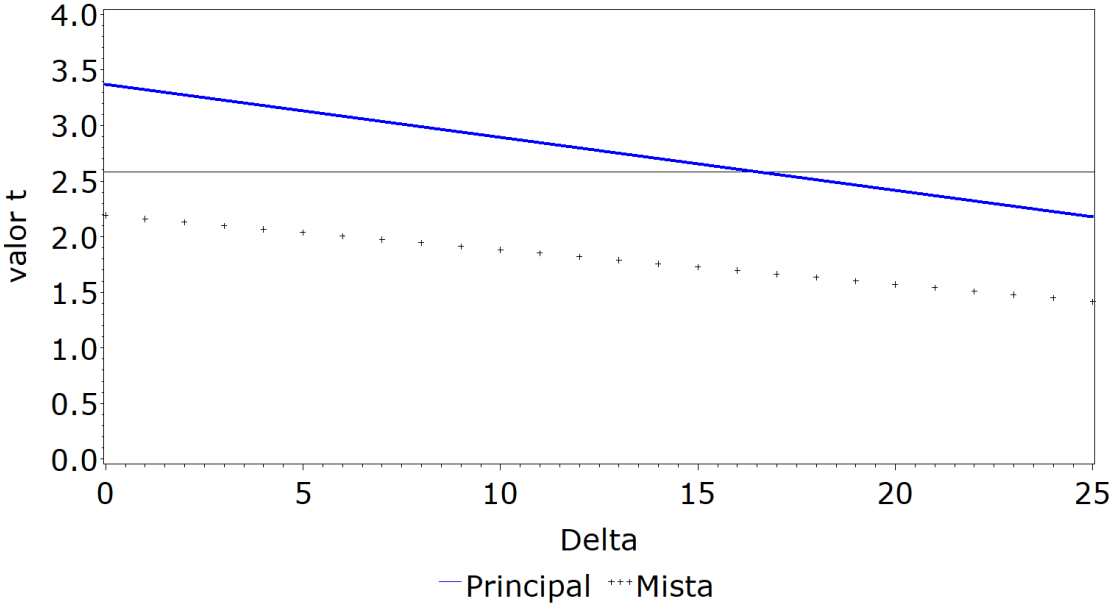

Figura 4.5: Valor da estatística t referentes a diferentes μ_D - amostragem principal e mista

Capítulo 5

Conclusões

5.1 Conclusões

Grandes amostras são ótimas para gerar estimativas pontuais porém, como consequência da consistência dos estimadores (Seção 1.3), podem gerar estimadores com variância próximas de zero. Assim, com objetivo de verificar a influência do tamanho da amostra no comportamento do p-valor do teste de hipótese de média, foram realizados estudos simulados em diferentes tipos de amostragem.

Essas simulações, foram realizadas para amostragem aleatória simples, amostragem estratificada e amostragem complexa (com estratos, clusters e probabilidade desigual de seleção). Os resultados mostraram que, seguindo a teoria de consistência dos estimadores, ao aumentar o tamanho amostral a variância do estimador diminui. Esse resultado mostra que amostras muito grandes podem influenciar na conclusão dos testes de média, como verificado no Capítulo 4, pois uma variância estimada pequena pode gerar estatísticas do teste t de valores elevados, como vista na Equação 1.5.

A Seção 3.3.3 apresentou uma alternativa para realização de testes de média para grandes amostras. Essa alternativa, chamada de teste de hipótese misto, consiste em realizar o teste de hipóteses utilizando informações das estatísticas pontuais do banco de dados considerado "grande" e as estimativas de variabilidade do banco de dados composto pela subamostragem do banco de dados "grande", dado o erro e o nível de confiança proposto pelo pesquisador.

A Seção 3.3.4 introduziu o Teste de Efeito do Tamanho da Amostra na Significância de Testes de Hipóteses que verifica se o tamanho amostral está influenciando na conclusão do teste de hipóteses, tomada com base no nível de significância do teste.

Este trabalho detalhou, também, como realizar a subamostragem de uma amostra complexa. O detalhamento, apresentado na Seção 3.3.2, mostra o passo a passo da técnica e explica o cuidado que o pesquisador deve ter ao recalcular o peso final, que é a combinação dos pesos do banco amostral inicial e o final.

O resultado apresentado na Tabela 4.9 mostrou a importância de se manter o mesmo esquema amostral ao realizar uma subamostragem quando a plano amostral é complexo. A estimativa de variância é diferente e o a quantidade de acertos do intervalo de confiança diminuiu.

A Seção 4.4 apresentou os resultados da utilização dos pesos nas variadas combinações de amostragem e subamostragem nos casos de estudo simulados deste trabalho. Foi verificado que apenas ao replicar a mesma técnica de amostragem utilizada no banco grande na subamostragem e utilizar o peso adequado e corrigido existe a coincidência de a soma dos pesos da subamostragem ser igual à população em estudo.

Por fim, na Seção 4.5, o teste de efeito do tamanho da amostra na significância de testes de hipóteses e a a técnica de subamostragem mista foram aplicados para um teste de comparação dos rendimentos médios entre as PNADs de 2012 e de 2014. Concluiu-se que para este caso o tamanho amostral da PNAD influenciou no resultado obtido, a um nível de significância de 99%, de que o rendimento médio é diferente entre 2012 e 2014.

5.2 Limitações do trabalho

Este trabalho focou em verificar a influência do tamanho amostral apenas no teste de média, no entanto o estudo pode ser replicado para testes dos parâmetros de um modelo de regressão e em outros testes de hipóteses, paramétricos ou não.

Inicialmente foi pensado em realizar simulações considerando os Erros do tipo I

e II, porém neste trabalho foi considerado apenas o primeiro.

5.3 Recomendações para trabalhos futuros

Como sugestão de estudos para trabalhos futuros, sugere-se que:

- seja verificada se há influência do tamanho da amostra no comportamento do p-valor nos testes para coeficientes de regressão;
- teste F;
- testes de correlação;
- testes de verificação de distribuição;
- testes de hipóteses não paramétricos.
- verificar a influência do tamanho amostral no valor do erro do tipo II de um teste de hipótese;
- Estudar o tamanho máximo de uma amostra, considerando o ganho na variabilidade, principalmente para amostras maiores que 10 mil.

Referências Bibliográficas

Arizola, H. G. A. e Teixeira, A. R. (2015). Impacto do zumbido em idosos praticantes e não praticantes de exercício físico. *ConScientiae Saúde*, 14(1):80–88.

BRASIL, C. (1988). *Constituição da República Federativa do Brasil*. Senado Federal: Centro Gráfico.

Bussab, W. e Morettin, P. (2010). *Estatistica basica*. Saraiva.

Casella, G. e Berger, R. (2002). *Statistical Inference*. Duxbury advanced series in statistics and decision sciences. Thomson Learning.

Chambers, R. L. e Skinner, C. J. (2003). *Analysis of survey data*. John Wiley & Sons.

Cochran, W. G. (1977). *Sampling Techniques*, (3rd ed.). John Wiley & Sons.

Heeringa, S. G., West, B. T., e Berglund, P. A. (2010). *Applied survey data analysis*. CRC Press.

IBGE (2012). Pesquisa nacional por amostra e domicílios - notas metodológicas. Technical report.

IBGE (2015). Pesquisa nacional por amostra de domicílios contínua - trimestral. Technical report. Acesso em 27 Out. 2017. Disponivel em: `https://ww2.ibge.gov.br/home/estatistica/indicadores/trabalhoerendimento/pnad_continua/default.shtm`.

Khan, M., Chand, M. A., e Ahmad, N. (2006). Optimum allocation in two-stage and stratified two-stage sampling for multivariate surveys. *A A*, 1(2):2.

Kish, L. e Frankel, M. R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–37.

Kovar, J., Rao, J., e Wu, C. (1988). Bootstrap and other methods to measure errors in survey estimates. *Canadian Journal of Statistics*, 16(S1):25–45.

Lohr, S. (2009). *Sampling: design and analysis*. Brooks/Cole.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

Magalhães, M. e de Lima, A. (2008). *Noções de Probabilidade e Estatística (6a Edição Revista e Ampliada) Vol. 40*. Edusp.

Mood, A. M., Boes, D. C., e Graybill, F. A. (1974). *Introduction to the Theory of Statistics*, (3th ed.). McGraw-Hill.

Nascimento Silva, P. L., Pessoa, D. G. C., e Lila, M. F. (2002). Análise estatística de dados da pnad: incorporando a estrutura do plano amostral. *Ciência & Saúde Coletiva*, 7(4):659–670.

Pessoa, D. G., SILVA, P. L., e Duarte, R. P. (1997). Análise estatística de dados de pesquisas por amostragem: Problemas no uso de pacotes-padrão. *Revista Brasileira de Estatística*, 58(210):53–75.

Rao, J. e Wu, C. J. (1985). Inference from stratified samples: second-order analysis of three methods for nonlinear statistics. *Journal of the American Statistical Association*, 80(391):620–630.

Särndal, C., Swensson, B., e Wretman, J. (2003). *Model Assisted Survey Sampling*. Springer Series in Statistics. Springer New York.

Silva, D. B., Carvalho, A., e Neri, M. C. (2006). Diferenciais de salários por raça e gênero: aplicação dos procedimentos de oaxaca e heckman em pesquisas amostrais complexas. *XV Encontro de Estudos Populacionais*.

Skinner, C., Holt, D., e Smith, T. (1989). *Analysis of complex surveys*. Wiley series in probability and mathematical statistics. Wiley.

Walpole, R. E., Myers, R. H., Myers, S. L., e Ye, K. (1993). *Probability and statistics for engineers and scientists*, volume 5. Macmillan New York.

Wicklin, R. (2013). *Simulating data with SAS*. SAS Institute.

Zhu, R., Ma, P., Mahoney, M. W., e Yu, B. (2015). Optimal subsampling approaches for large sample linear regression. *arXiv preprint arXiv:1509.05111*.

Apêndice A

Apêndice

A.1 Amostra Aleatória Simples

Tabela A.1: Simulações para AAS com desvio padrão e tamanho de amostra variáveis

Distribuição	CV	Vezes	Confiança 90%		Confiança 95%		Confiança 99%	
			Amplitude Média	Acertos IC	Amplitude Média	Acertos IC	Amplitude Média	Acertos IC
Log-normal	10%	1	463,3	95%	290,4	96%	179,9	93%
		10	72,5	95%	64,3	92%	48,4	93%
		100	22,7	93%	19,6	95%	14,8	96%
		1000	7,2	97%	6,2	97%	4,7	96%
	50%	1	236,8	94%	196,6	94%	154,6	93%
		10	75,0	93%	63,0	98%	48,1	96%
		100	23,8	93%	19,9	95%	15,2	96%
		1000	7,5	95%	6,3	97%	4,8	94%
	100%	1	233,8	94%	193,6	93%	148,5	91%
		10	75,1	93%	63,2	95%	48,0	91%
		100	23,8	95%	19,9	95%	15,2	96%
		1000	7,5	95%	6,3	95%	4,8	96%
Normal	10%	1	426,3	94%	312,5	96%	195,3	99%
		10	75,6	93%	63,3	98%	47,5	94%
		100	22,8	97%	19,7	96%	14,8	93%
		1000	7,2	96%	6,2	96%	4,7	97%
	50%	1	240,2	97%	200,1	97%	151,8	95%
		10	75,1	95%	63,1	95%	48,1	95%
		100	23,8	97%	19,9	92%	15,2	96%
		1000	7,5	87%	6,3	97%	4,8	97%
	100%	1	239,1	92%	200,5	98%	152,0	97%
		10	75,4	95%	63,1	95%	48,1	95%
		100	23,8	95%	20,0	96%	15,2	93%
		1000	7,5	98%	6,3	97%	4,8	95%

*Tamanho mínimo da amostra calculado com erro de R$100,00 e conforme o nível de confiança.

Tabela A.2: Amplitude média do IC gerado nas simulações para Amostragem Aleatória Simples (AAS) - Casos "principal", "subamostragem"e "misto"

Confiança	Variável	CV	Amostragem - AAS		
			Principal (1.000 vezes)	subamostragem - AAS	Mista
90%	Log-normal	10%	7,2	451,8	451,8
		50%	7,5	231,9	231,9
		100%	7,5	238,4	238,4
	normal	10%	7,2	456,6	456,6
		50%	7,5	240,3	240,3
		100%	7,5	238,2	238,2
95%	Log-normal	10%	6,2	300,7	300,7
		50%	6,3	199,5	199,5
		100%	6,3	200,9	200,9
	normal	10%	6,2	302,6	302,6
		50%	6,3	202,7	202,7
		100%	6,3	200,4	200,4
99%	Log-normal	10%	4,7	183,2	183,2
		50%	4,8	154,9	154,9
		100%	4,8	150,2	150,2
	normal	10%	4,7	177,4	177,4
		50%	4,8	153,3	153,3
		100%	4,8	152,7	152,7

A.2 Amostra Estratificada

Tabela A.3: Simulações para amostragem estratificada segundo o nível de confiança e o tamanho de amostra variáveis

Confiança	Variável	Vezes	Amplitude média	Acertos IC
90%	lognormal	1	405,5	91%
		10	141,4	95%
		100	43,6	95%
		1000	13,9	98%
	normal	1	439,3	99%
		10	138,8	97%
		100	43,9	94%
		1000	13,9	94%
95%	lognormal	1	361,5	88%
		10	115,5	94%
		100	36,6	94%
		1000	11,7	94%
	normal	1	368,7	97%
		10	116,6	97%
		100	36,8	99%
		1000	11,7	96%
99%	lognormal	1	259,7	93%
		10	87,7	89%
		100	28,1	96%
		1000	8,9	93%
	normal	1	281,9	97%
		10	88,8	94%
		100	28,1	92%
		1000	8,9	97%

Tabela A.4: Amplitude média do IC gerado nas simulações para amostragem estratificada - Casos "principal", "subamostragem"e "misto- Subamostragem estratificada e aleatória simples

Variável	Confiança	Principal	Estratificada		AAS	
			Subamostragem	Misto	Subamostragem	Misto
lognormal	0,95	13,8	422,4	422,4	414,8	414,8
	0,975	11,6	348,9	348,9	343,8	343,8
	0,995	8,8	263,9	263,9	277,5	277,5
normal	0,95	13,8	439,2	439,2	445,4	445,4
	0,975	11,6	371,2	371,2	374,0	374,0
	0,995	8,8	280,8	280,8	284,2	284,2

A.3 Amostra Complexa

Tabela A.5: Tamanho amostral no segundo estágio - Amostragem complexa

Estrato	Conglomerado	M_{hi}	m^*_{hi}	m^*_{hi} x 10	m^*_{hi} x 100	m^*_{hi} x 1.000
1	1	1.304.462	29	290	2.900	29.000
	2	2.336.003	42	420	4.200	42.000
	3	2.031.813	38	380	3.800	38.000
	4	3.027.590	48	480	4.800	48.000
	5	433.688	27	270	2.700	27.000
	6	2.585.740	45	450	4.500	45.000
	7	1.679.909	33	330	3.300	33.000
	8	895.140	25	250	2.500	25.000
	9	2.782.218	47	470	4.700	47.000
	10	2.923.436	48	480	4.800	48.000
2	1	6.734.717	94	940	9.400	94.000
	2	6.022.548	63	630	6.300	63.000
	3	6.862.672	87	870	8.700	87.000
	4	6.884.638	91	910	9.100	91.000
	5	4.896.096	40	400	4.000	40.000
	6	6.401.723	73	730	7.300	73.000
	7	4.672.124	37	370	3.700	37.000
	8	5.616.438	54	540	5.400	54.000
	9	6.690.810	81	810	8.100	81.000
	10	5.218.235	46	460	4.600	46.000
3	1	10.105.310	145	1.450	14.500	145.000
	2	9.383.196	181	1.810	18.100	181.000
	3	10.465.524	134	1.340	13.400	134.000
	4	7.099.869	174	1.740	17.400	174.000
	5	6.194.196	152	1.520	15.200	152.000
	6	4.778.856	120	1.200	12.000	120.000
	7	5.389.767	132	1.320	13.200	132.000
	8	8.000.942	189	1.890	18.900	189.000
	9	8.782.809	191	1.910	19.100	191.000
	10	9.799.530	163	1.630	16.300	163.000
4	1	3.023.401	641	6.410	64.100	641.000
	2	21.836	185	1.850	18.500	21.836
	3	3.122	229	2.290	3.122	3.122
	4	625.752	397	3.970	39.700	397.000
	5	1.435.092	514	5.140	51.400	514.000
	6	83.505	230	2.300	23.000	83.505
	7	17.998.436	911	9.110	91.100	911.000
	8	245.576	300	3.000	30.000	245.576
	9	10.672.295	857	8.570	85.700	857.000
	10	5.890.986	761	7.610	76.100	761.000

Tabela A.6: Dados para amostragem complexa - estratos

Estrato	Média	S^2_{hb}	$\bar{M}_h$	A_h	W_h	n_h^{*1}	n_h^*
1	148,6	3.616,0	2.000.000	3.616,0	0,25	3,77	4
2	417,1	9.968,9	6.000.000	9.968,9	0,25	3,75	4
3	971,9	68.825,3	8.000.000	68.825,3	0,25	3,68	4
4	8.942,6	67.954.500,5	4.000.000	67.954.500,5	0,25	3,45	4

[1] valor apresentado sem a aproximação para o próximo inteiro.

Note que o cálculo de m^*_{hi} é feito utilizando a Equação 3.6. Por exemplo, utilizando os dados das tabelas A.6 e A.7, representados na Equação A.1, o tamanho amostral do estrato 1 do conglomerado 1 é 29, o próximo valor inteiro obtido ao resolver essa equação.

$$m^*_{11} = \frac{1.304.462 \times 37,1}{2.000.000}\sqrt{\frac{25.000}{3.616 \times 5}} = 28,45 \tag{A.1}$$

Tabela A.7: Dados para amostragem complexa - conglomerados

Estrato	Conglomerado	M_{hi}	S^2_{hiy}	Média	m^*_{hi}
1	1	1.304.462	37,1	101,3	29
	2	2.336.003	30,5	160,6	42
	3	2.031.813	31,6	141,4	38
	4	3.027.590	26,6	234,5	48
	5	433.688	104,5	52,7	27
	6	2.585.740	29,5	179,5	45
	7	1.679.909	33,3	121,7	33
	8	895.140	47,3	79,3	25
	9	2.782.218	28,5	198,2	47
	10	2.923.436	27,5	216,5	48
2	1	6.734.717	117,1	262,3	94
	2	6.022.548	88,1	438,4	63
	3	6.862.672	107,3	335,5	87
	4	6.884.638	112,0	299,3	91
	5	4.896.096	68,8	529,8	40
	6	6.401.723	95,5	405,3	73
	7	4.672.124	66,9	558,4	37
	8	5.616.438	80,5	470,2	54
	9	6.690.810	101,9	370,9	81
	10	5.218.235	73,7	500,6	46
3	1	10.105.310	423,4	675,0	145
	2	9.383.196	569,5	828,6	181
	3	10.465.524	377,8	605,8	134
	4	7.099.869	725,3	1.095,8	174
	5	6.194.196	727,6	1.189,2	152
	6	4.778.856	743,6	1.376,4	120
	7	5.389.767	724,4	1.282,5	132
	8	8.000.942	698,2	1.003,4	189
	9	8.782.809	643,1	913,8	191
	10	9.799.530	491,2	749,0	163
4	1	3.023.401	98.795,9	3.974,1	641
	2	21.836	3.946.687,8	17.365,6	185
	3	3.122	34.147.452,7	27.753,6	229
	4	625.752	295.339,2	6.853,2	397
	5	1.435.092	166.961,7	5.218,1	514
	6	83.505	1.283.988,8	12.256,6	230
	7	17.998.436	23.584,0	1.661,8	911
	8	245.576	568.607,2	9.072,9	300
	9	10.672.295	37.430,2	2.257,7	857
	10	5.890.986	60.231,8	3.012,3	761