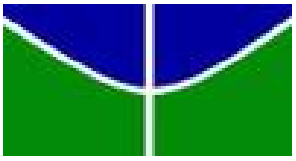

Universidade de Brasília - UnB
Instituto de Ciências Exatas - IE
Departamento de Estatística - EST

Pesquisas por telefone: Análise dos pedidos de Auxílio Emergencial.

Luana Pereira Ramos da Silva - 16/0132860

Brasília

2021

Luana Pereira Ramos da Silva **16/0132860**

Pesquisas por telefone: Análise dos pedidos de Auxílio Emergencial.

Relatório apresentado à disciplina Trabalho de Conclusão de Curso 2 do Bacharelado em Estatística, Departamento de Estatística, Instituto de Exatas, Universidade de Brasília, como parte dos requisitos necessários para o grau de Bacharel em Estatística.

Orientador: Prof. Dr. Alan Ricardo da Silva

Brasília

2021

Dedico este trabalho à minha família, amigos e professores, pilares da minha formação como ser humano. Dedico também àqueles que sempre me apoiaram, mas já se foram. Onde quer que estejam, espero que estejam orgulhosos.

Agradecimentos

Agradeço primeiramente à minha família pelo suporte e incentivo que me forneceram durante a minha trajetória acadêmica. Especialmente, minha mãe Rosilene Pereira da Silva, que sempre foi um exemplo de mulher e de ser humano, que sempre foi e sempre será um dos alicerces da minha vida.

Agradeço ao meu orientador Alan Ricardo da Silva, pela paciência, disposição e suporte que forneceu durante a realização deste trabalho e que foram essenciais. Agradeço ao professor Irmar Pereira Nunes, meu professor de matemática durante todo o Ensino Médio, que fez despertar em mim a paixão pela Estatística.

Agradeço à todos os meus amigos e colegas que me ajudaram durante a graduação e que fizeram essa experiência ainda mais marcante. Agradeço à todos os professores do Departamento de Estatística da Universidade de Brasília, pela paciência e pelos ensinamentos que proporcionaram e ainda proporcionam a todos que os procuram.

Agradeço à toda equipe do Instituto de Pesquisa DataSenado, pela disponibilização das informações que estão contidas neste trabalho e pela contribuição na minha vida profissional. Em especial, ao Marcos Ruben de Oliveira, que sempre esteve disponível para prestar auxílio durante a realização deste trabalho, e a minha grande amiga Isabella Cristine Figueiredo Vieira, que me incentivou a realizar este trabalho e que sempre foi uma inspiração para mim, tanto no âmbito profissional quanto pessoal.

Por fim, agradeço à todos aqueles que contribuíram, de alguma forma, para a realização deste trabalho. Muito obrigada!

"*Remember to look up at the stars and not down at your feet.*"

— Stephen Hawking

Resumo

A pesquisa por telefone é um método de coleta de dados bastante popular, mas apresenta algumas desvantagens, como a alta taxa de não-resposta. No Brasil, um exemplo de instituto que utiliza esse método é o Instituto de Pesquisa DataSenado. Assim, este trabalho pretende analisar a metodologia adotada pelo DataSenado para estimar a quantidade de pessoas que receberam a primeira parcela do Auxílio Emergencial, até maio de 2020, e analisar os efeitos do ajuste de não-resposta (e outros ajustes adotados pelo DataSenado) nas estimativas e em seus intervalos de confiança. A partir dos resultados obtidos, verificou-se que a metodologia adotada pelo DataSenado conseguiu gerar intervalos de confiança que capturaram os parâmetros populacionais referentes ao total de pessoas que receberam a primeira parcela do Auxílio Emergencial até maio de 2020, mas isso não aconteceu ao estimar o percentual que essas pessoas representavam entre aquelas que solicitaram o benefício. Também verificou-se que o ajuste de não-resposta, da forma como é aplicado pelo DataSenado, tem pouco impacto nas estimativas e nos intervalos de confiança após a calibração dos pesos amostrais, e que, dependendo dos ajustes aplicados aos pesos amostrais, os resultados obtidos podem ser diferentes.

Palavras-chaves: Pesquisa por telefone. Amostragem. Auxílio Emergencial. Instituto de Pesquisa DataSenado.

Lista de Tabelas

Lista de Figuras

Sumário

Capítulo 1

INTRODUÇÃO

O Instituto de Pesquisa DataSenado é um órgão vinculado à Secretaria de Transparência do Senado Federal e foi criado em 2005 com o objetivo de acompanhar, em resumo, a opinião pública a respeito do Senado Federal e sobre temas em debate no Congresso Nacional (DataSenado, 2020f). Segundo o DataSenado (2020f), até 2016 o instituto já ouviu mais de 4 milhões de brasileiros por meio de 147 pesquisas de opinião e pesquisas internas e 130 enquetes e pesquisas *online*. Os dados levantados servem como auxílio para parlamentares entenderem como a população brasileira pensa a respeito de um determinado assunto.

O DataSenado possui algumas pesquisas bastante relevantes que são realizadas periodicamente. A pesquisa Violência Doméstica e Familiar contra a Mulher, realizada em parceria com o Observatório da Mulher contra a Violência, já está em sua 8ª edição (DataSenado, 2019). Essa pesquisa investiga, principalmente, a percepção das mulheres a respeito da violência doméstica no Brasil ao longo do tempo e serve como base para diversas discussões sobre o tema. Além dela, existe também a pesquisa O Cidadão e o Senado Federal (DataSenado, 2020d) que busca investigar aspectos da população brasileira como posicionamento político e atitudes sociais.

As pesquisas de opinião do DataSenado são realizadas por meio de amostras representativas da população brasileira, por meio telefônico, onde todas as unidades da federação são representadas, incluindo capitais e cidades do interior. As enquetes do DataSenado são realizadas mensalmente, mas não são representativas e as perguntas ficam disponíveis no Portal do Senado (DataSenado, 2020f).

A pesquisa por telefone é um método de coleta de dados bastante popular, principalmente quando se trata de pesquisas de opinião. Nela, qualquer indivíduo da população é considerado como um elemento da amostra em potencial. Isso significa que uma pessoa só é incluída na amostra ao atender a ligação e concordar em participar da pesquisa (Sincero, 2012).

Pesquisas por telefone possuem algumas vantagens práticas e administrativas sobre pesquisas presenciais, como tempo e custo, principalmente quando os respondentes estão espalhados sobre uma área ampla (Colombotos, 1969). No entanto, sua grande desvantagem é uma maior taxa de não-resposta (Groves et al., 2001). Essa taxa de não-resposta pode ser um problema quando certos grupos da população de interesse deixam de responder a uma pesquisa por telefone e isso pode causar um viés nas estimativas. Por exemplo, considere uma pesquisa que tenha como objetivo conhecer a opinião de uma certa população a respeito de um tema, mas apenas as mulheres atendem o telefone e aceitam participar dessa pesquisa, como indica a Figura 1.1. Nesse caso, a opinião dos homens dessa população não estará representada nas estimativas obtidas por essa pesquisa, o que faz com que essas estimativas sejam viesadas.

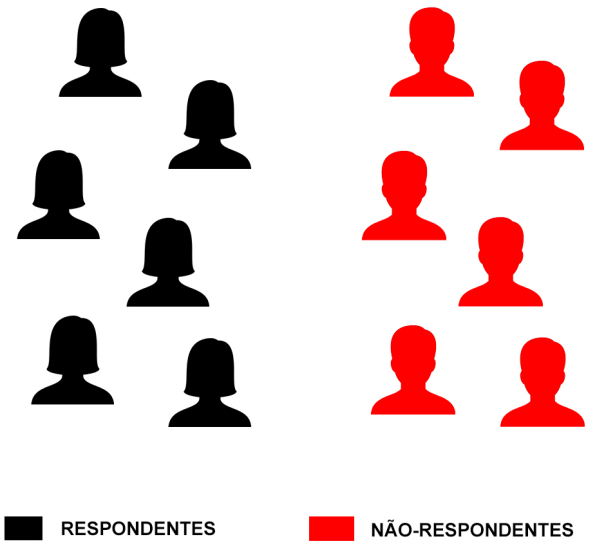

Figura 1.1: Exemplificação do viés de não-resposta.

No Brasil, alguns institutos de pesquisa de opinião utilizam pesquisas por telefone como meio para coletar dados, como é o caso do Instituto de Pesquisa DataSenado. Dessa forma, este trabalho pretende analisar uma pesquisa realizada pelo DataSenado, por telefone, a respeito da quantidade de pessoas que solicitaram e receberam (até 20 de maio de 2020) a primeira parcela do Auxílio Emergencial, um benefício concedido pelo Governo Federal durante a crise causada pela pandemia do coronavírus (Caixa, 2020a).

1.1 Objetivos

O objetivo geral do trabalho é analisar a metodologia utilizada pelo DataSenado para estimar a quantidade de pessoas que receberam a primeira parcela do Auxílio Emergencial.

Os objetivos específicos são:

- Entender o mecanismo da pesquisa por telefone;
- Analisar os efeitos da taxa de não-resposta;
- Analisar os métodos de ponderação para correção da não-resposta e probabilidade de seleção.

Capítulo 2

TÉCNICAS DE AMOSTRAGEM

2.1 Introdução

Geralmente, pesquisadores não conseguem coletar informações a respeito de todos os indivíduos da população que pretendem estudar, principalmente quando essa população é muito grande. Para contornar essa situação, são coletadas informações apenas de um subconjunto dessa população. Esse subconjunto é chamado de **amostra** e é definido como um subconjunto de uma população por meio do qual se estabelecem ou estimam as propriedades e características dessa população (Bolfarine e Bussab, 2005) e o processo de construção (ou seleção) dessa amostra é chamado de **amostragem**.

Quando uma amostra é selecionada, é esperado que ela possua características semelhantes às da população estudada, ou seja, espera-se que ela seja uma amostra representativa da população. Para isso, existem técnicas de amostragem que podem ser adotadas e cabe ao pesquisador decidir qual melhor se adequa ao seu estudo ou aos seus objetivos.

2.2 Amostragem Aleatória Simples

A Amostragem Aleatória Simples (AAS) consiste no método de amostragem onde todos os elementos da população possuem uma probabilidade igual e conhecida de pertencer à amostra. Sendo assim, enumera-se os elementos da população e, através de um procedimento aleatório, sorteia-se n elementos com igual probabilidade, sendo n prefixado anteriormente (Bolfarine e Bussab, 2005). Nesse caso, tem-se o método AAS **sem** reposição (AAS_s). Caso seja permitido que um mesmo elemento apareça mais de uma vez na amostra, têm-se o método AAS **com** reposição (AAS_c). Daqui em diante, será utilizado como referência apenas o método AAS_s.

2.2.1 Estimação de totais e médias

Total populacional

$$T = \sum_{i=1}^{N} Y_i = Y_1 + Y_2 + \cdots + Y_N. \tag{2.1}$$

Um estimador não viesado para T é dado por

$$\widehat{T} = N\frac{\sum_{i=1}^{n} y_i}{n} = N\overline{y}, \tag{2.2}$$

com variância

$$Var(\widehat{T}) = N^2\left(1 - \frac{n}{N}\right)\frac{S^2}{n}, \tag{2.3}$$

e

$$\widehat{Var}(\widehat{T}) = N^2\left(1 - \frac{n}{N}\right)\frac{s^2}{n}, \tag{2.4}$$

onde $\widehat{Var}(\widehat{T})$ é um estimador não viesado para a variância de $\hat{T}$.

Média populacional

$$\overline{Y} = \frac{\sum_{i=1}^{N} Y_i}{N} = \mu. \tag{2.5}$$

Define-se o estimador não viesado de $\overline{Y}$, ou seja, a média amostral, como

$$\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}. \tag{2.6}$$

com variância

$$Var(\overline{y}) = \frac{S^2}{n}\left(1 - \frac{n}{N}\right), \tag{2.7}$$

tendo um estimador não viesado dado por

$$\widehat{Var}(\overline{y}) = \frac{s^2}{n}\left(1 - \frac{n}{N}\right), \tag{2.8}$$

onde $\frac{n}{N}$ é conhecido como fator de correção para população finita ou, do inglês, *finite population correction* (fpc) (Cochran, 1977).

Variância de Y para uma população finita

Cochran (1977) define a variância de Y para população finita como

$$S^2 = \frac{\sum_{i=1}^{N}(Y_i - \overline{Y})^2}{N-1}. \tag{2.9}$$

Assim,

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{n-1}. \tag{2.10}$$

é um estimador não viesado para S^2.

Tamanho da amostra

O tamanho da amostra é dado por

$$n = \frac{\frac{z^2_{\alpha/2}S^2}{d^2}}{1 + \frac{1}{N}\left(\frac{z^2_{\alpha/2}S^2}{d^2}\right)} = \frac{n_0}{1 + \frac{n_0}{N}}, \tag{2.11}$$

onde

- $n_0 = \frac{z^2_{\alpha/2}S^2}{d^2}$;
- $1 - \alpha$ é o nível de confiança fixado anteriormente;
- d é o erro máximo admitido para as estimativas fixado anteriormente e;
- $z_{\alpha/2}$ é o quantil da distribuição normal $N(0, 1)$ tal que a área na densidade da $N(0, 1)$ no intervalo $(-z_{\alpha/2}; z_{\alpha/2})$ seja igual a $1 - \alpha$ (Cochran, 1977).

Para determinar n a partir da Equação (2.11) é necessário possuir um conhecimento prévio de S^2 (Bolfarine e Bussab, 2005). Ainda segundo Bolfarine e Bussab (2005), em muitos casos, uma amostra piloto pode ser utilizada para obter um estimador inicial de S^2, em outros casos, podem ser utilizadas estimativas de pesquisas anteriores sobre a população.

Intervalos de confiança

Costuma-se assumir que os estimadores $\overline{y}$ e $N\overline{y}$ são normalmente distribuídos em torno dos valores populacionais (Cochran, 1977). Nesse caso, os intervalos com $100(1 - \alpha)\%$ de confiança para a média e para o total populacional são dados, respectivamente, por

$$\left(\overline{y} \pm z_{\alpha/2}\sqrt{\frac{s^2}{n}\left(1 - \frac{n}{N}\right)}\right) \tag{2.12}$$

e

$$\left(N\overline{y} \pm z_{\alpha/2} N \sqrt{\frac{s^2}{n}\left(1 - \frac{n}{N}\right)} \right). \tag{2.13}$$

Caso o tamanho da amostra seja inferior à 50, os valores referentes ao nível de confiança podem ser obtidos a partir da distribuição t-Student com $(n-1)$ graus de liberdade, mas apenas se as obervações Y_i forem normalmente distribuídas e N tender ao infinito (Cochran, 1977).

2.2.2 Estimação de proporções

Em algumas situações, o interesse pode ser estudar a proporção de elementos de uma determinada população que possuem uma certa característica. Assim, a cada elemento da população é associada uma variável aleatória (Bolfarine e Bussab, 2005)

$$Y_i = \begin{cases} 1, & \text{se o elemento } i \text{ possui a característica} \\ 0, & \text{caso contrário.} \end{cases}$$

Dessa forma, a proporção de elementos da população que possuem a característica de interesse é dada por

$$P = \frac{\sum_{i=1}^{N} Y_i}{N} = \mu. \tag{2.14}$$

Como Y_i assume apenas os valores 0 e 1, a Equação (2.9) pode ser escrita como

$$S^2 = \frac{\sum_{i=1}^{N} (Y_i - P)^2}{N-1} = \left(\frac{N}{N-1}\right) P(1-P). \tag{2.15}$$

Segundo Bolfarine e Bussab (2005), seja n o tamanho da amostra e m a quantidade de elementos dessa amostra com uma determinada características, um estimador não viesado para P é dado por

$$p = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{m}{n}, \tag{2.16}$$

e com

$$Var(p) = \left(1 - \frac{n}{N}\right) \frac{S^2}{n} = \left(\frac{N-n}{N-1}\right) \frac{P(1-P)}{n}. \tag{2.17}$$

Um estimador não viesado para S^2, nesse caso, é dado por

$$s^2 = \left(\frac{n}{n-1}\right) p(1-p). \tag{2.18}$$

Assim, consequentemente,

$$\widehat{Var}(p) = \left(1 - \frac{n}{N}\right) \frac{p(1-p)}{n-1}. \tag{2.19}$$

é um estimador não viesado para $Var(p)$.

Tamanho da amostra

No caso da AAS_s, quando se deseja saber a proporção de indivíduos com uma determinada característica em uma população, (Cochran, 1977) define o tamanho da amostra como

$$n = \frac{\frac{z^2_{\alpha/2} P(1-P)}{d^2}}{1 + \frac{1}{N}\left(\frac{z^2_{\alpha/2} P(1-P)}{d^2}\right)} = \frac{n_0}{1 + \frac{n_0}{N}}, \tag{2.20}$$

onde $n_0 = \frac{z^2_{\alpha/2} P(1-P)}{d^2}$ e d é o erro máximo desejado fixado previamente.

Para utilizar (2.20), é preciso conhecer P, para se ter conhecimento à respeito da variabilidade da população. Segundo Bolfarine e Bussab (2005), assim como no caso de (2.11), uma forma de obter um estimador para P é utilizando uma amostra piloto ou a partir de pesquisas anteriores. Entretanto, uma forma alternativa para definir n é assumindo $P(1-P) = 1/4$, ou seja, tendo variância máxima. Assim, (2.20) se reduz à

$$n = \frac{z_{\alpha/2}/4d^2}{1 + \frac{1}{N}\left(z_{\alpha/2}/4d^2\right)} = \frac{n_0}{1 + \frac{n_0}{N}}, \tag{2.21}$$

onde $n_0 = z^2_{\alpha/2}/4d^2$.

Como P varia entre 0 e 1, nota-se que

$$P(1-P) \leq 0,5 \times 0,5 = 0,25 = 1/4.$$

Assim, (2.21) é considerada uma maneira conservadora para determinar o tamanho da amostra (n) quando não se conhece P (Bolfarine e Bussab, 2005).

Intervalos de confiança

Como o estimador p também varia entre 0 e 1, $p(1-p)$ será sempre menor que 1/4. Assim, Bolfarine e Bussab (2005) definem um intervalo com $100(1-\alpha)\%$ de confiança conservador para P como

$$\left(p \pm z_{\alpha/2} \sqrt{\frac{\left(1 - \frac{n}{N}\right)}{4(n-1)}} \right). \tag{2.22}$$

No entanto, a Equação (2.22) é definida de maneira diferente quando há conhecimento prévio à respeito de P. Nessa situação, basta substituir 1/4 pelo produto $P(1-P)$.

2.3 Amostragem Aleatória Estratificada

A Amostragem Aleatória Estratificada (AAE), ou Amostragem Estratificada (AE), é um método que consiste na divisão da população em H grupos (estratos) de acordo com alguma característica conhecida dos indivíduos de uma população e é usada principalmente para melhorar a precisão das estimativas (Bolfarine e Bussab, 2005). Na AAE, uma população de tamanho N é dividida em H grupos com $N_1, N_2, \ldots, N_H$ unidades, respectivamente, tal que $N_1 + N_2 + \cdots + N_H = N$ (Cochran, 1977).

A maneira mais simples de aplicar uma AAE é sorteando, de maneira independente, uma AAS de cada estrato para que, assim, n_h obervações sejam selecionadas aleatoriamente a partir da população do estrato h, $h = 1, 2, \ldots, H$ (Lohr, 1999).

2.3.1 Estimação de totais e médias

Cochran (1977) fornece algumas definições importantes que serão apresentadas a seguir:

N_h = número de unidades do estrato h.

n_h = número de unidades na amostra do estrato h.

y_{hi} = valor da i-ésima unidade do estrato h.

$W_h = \frac{N_h}{N}$ = peso do estrato h.

$f_h = \frac{n_h}{N_h}$ = fração amostral do estrato h.

$\overline{Y}_h = \frac{\sum_{i=1}^{N_h} Y_{hi}}{N_h}$ = média populacional do estrato h.

$\overline{y}_h = \frac{\sum_{i=1}^{n_h} y_{hi}}{n_h}$ = média amostral do estrato h.

$S_h^2 = \sum_{i=1}^{N_h} \frac{(Y_{hi} - \overline{Y}_h)^2}{N_h - 1}$ = variância populacional no estrato h.

Supondo que em cada estrato h foi aplicada, de maneira independente, uma AAS_s e amostradas n_h unidades, Cochran (1977) apresenta as seguintes definições.

Total populacional

Um estimador não viesado para o total populacional é dado por

$$\widehat{T}_{st} = \sum_{h=1}^{H} N_h \overline{y}_h, \tag{2.23}$$

com variância

$$Var(\widehat{T}_{st}) = \sum_{h=1}^{N}(1-f_h)N_h^2\frac{S_h^2}{n_h}. \tag{2.24}$$

Definindo

$$s_h^2 = \sum_{i=1}^{n_h}\frac{(y_{hi}-\overline{y}_h)^2}{n_h-1}, \tag{2.25}$$

para obter o estimador $\widehat{Var}(\widehat{T}_{st})$, basta substituir S_h^2 por s_h^2 em (2.24), como acontece na amostragem aleatória simples.

Média populacional

No caso da média populacional, um estimador não viesado é dado por

$$\overline{y}_{st} = \sum_{h=1}^{H}\frac{N_h}{N}\overline{y}_h = \sum_{h=1}^{H}W_h\overline{y}_h, \tag{2.26}$$

com variância

$$Var(\overline{y}_{st}) = \sum_{h=1}^{H}(1-f_h)W_h^2\frac{S_h^2}{n_h}. \tag{2.27}$$

Para obter $\widehat{Var}(\overline{y}_{st})$, basta substituir S_h^2 pelo seu estimador s_h^2.

Intervalos de confiança

Os intervalos com $100(1-\alpha)\%$ de confiança para a média e total populacional são dados, respectivamente, por

$$\left(\overline{y}_{st} \pm z_{\alpha/2}\sqrt{\sum_{h=1}^{H}(1-f_h)W_h^2\frac{s_h^2}{n_h}}\right) \tag{2.28}$$

e

$$\left(N\overline{y}_{st} \pm z_{\alpha/2}N\sqrt{\sum_{h=1}^{H}(1-f_h)\frac{s_h^2}{n_h}}\right), \tag{2.29}$$

assumindo que $\overline{y}_{st}$ e $N\overline{y}_{st}$ sejam normalmente distribuídos em torno dos valores populacionais.

2.3.2 Alocação da amostra nos estratos

Segundo Bolfarine e Bussab (2005), a distribuição das unidades da amostra pelos estratos é conhecida como alocação da amostra. Esse procedimento é importante para garantir a precisão do procedimento amostral (Bolfarine e Bussab, 2005).

Alocação ótima

Nesse procedimento, a alocação é feita de maneira a atender certas condições de custo fixo e custo linear. Sendo C o custo total, c_0 o custo fixo e c_h o custo por unidade do estrato h, Cochran (1977) define a função de custo linear como

$$C = c_0 + \sum_{h=1}^{H} c_h n_h \tag{2.30}$$

e $Var(\overline{y}_{st})$ é dada como em (2.27).

Segundo Cochran (1977), o objetivo pode ser minimizar $Var(\overline{y}_{st})$ para C fixo ou minimizar C para $Var(\overline{y}_{st})$ fixa. No primeiro caso, n é definido como

$$n = \frac{(C - c_0)\sum_{h=1}^{H}(N_h S_h/\sqrt{c_h})}{\sum_{h=1}^{H} N_h S_h \sqrt{c_h}}. \tag{2.31}$$

Já no segundo caso, n é definido como

$$n = \frac{\left(\sum_{h=1}^{H} W_h S_h \sqrt{c_h}\right)\sum_{h=1}^{H} W_h S_h/\sqrt{c_h}}{Var(\overline{y}_{st}) + \frac{\sum_{h=1}^{H} W_h S_h^2}{N}}. \tag{2.32}$$

Assim, Cochran (1977) define o tamanho amostral no estrato h (n_h) como

$$n_h \equiv n\frac{N_h S_h/\sqrt{c_h}}{\sum_{h=1}^{H}\left(N_h S_h/\sqrt{c_h}\right)}, \tag{2.33}$$

ou então, para n fixado,

$$n_h \equiv n\frac{N_h S_h}{\sum_{h=1}^{H}(N_h S_h)} \tag{2.34}$$

quando $c_h = c$, ou seja, o custo por unidade é o mesmo em todos os estratos.

Alocação proporcional

Segundo Cochran (1977), na alocação proporcional a fração amostral (f_h) é a mesma em todos os estratos, ou seja, a amostra é alocada de forma proporcional entre os estratos, isto é,

$$n_h = nW_h = n\frac{N_h}{N}. \tag{2.35}$$

Com isso, têm-se que o estimador $\overline{y}_{st}$ é dado como em (2.26), com variância

$$Var(\overline{y}_{st}) = \frac{(1 - \frac{n}{N})}{n} \sum_{h=1}^{H} W_h S_h^2. \tag{2.36}$$

Para obter $\widehat{Var}(\overline{y}_{st})$, basta substituir S_h^2 por s_h^2, como no caso da Equação (2.27).

Alocação uniforme

Na alocação uniforme, a amostra é distribuída igualmente entre todos os estratos. Assim, segundo Bolfarine e Bussab (2005), têm-se

$$n_h = \frac{n}{H} = k, \tag{2.37}$$

para cada um dos H estratos.

2.3.3 Estimação para proporções

Assim como na AAS, existem casos onde o interesse é estudar a proporção de elementos com uma determinada característica na população. Assim, segundo Bolfarine e Bussab (2005), a caraceterística de interesse associada ao i-ésimo elemento

do h-ésimo estrato pode ser representada como

$$Y_{hi} = \begin{cases} 1, & \text{se o elemento } (h,i) \text{ possui a característica} \\ 0, & \text{caso contrário.} \end{cases}$$

Dessa forma, a proporção de elementos que possuem uma certa característica na população do estrato h, $h = 1, 2, \ldots, H$, e a proporção desses elementos da amostra são, respectivamente, definidas por Cochran (1977) como

$$P_h = \frac{A_h}{N_h} = \frac{\sum_{i=1}^{N_h} Y_{hi}}{N_h} = \mu_h \tag{2.38}$$

e

$$p_h = \frac{a_h}{n_h} = \frac{\sum_{i=1}^{n_h} Y_{hi}}{n_h}. \tag{2.39}$$

Assim, ainda segundo Cochran (1977), a proporção estimada de elementos que possuem uma certa característica na população geral é definida como

$$p_{st} = \sum_{h=1}^{H} W_h p_h. \tag{2.40}$$

com variância dada por

$$Var(p_{st}) = \frac{1}{N^2} \sum_{h=1}^{H} \frac{N_h^2 (N_h - n_h)}{N_h - 1} \frac{P_h(1 - P_h)}{n_h}. \tag{2.41}$$

No caso da alocação proporcional, essa variância é dada por

$$Var(p_{st}) = \left(\frac{1 - f_h}{n} \right) \sum_{h=1}^{H} W_h P_h (1 - P_h). \tag{2.42}$$

Para obter $\widehat{Var}(p_{st})$ basta substituir $\frac{P_h(1-P_h)}{n_h}$ por $\frac{p_h(1-p_h)}{n_h - 1}$ em (2.41) ou (2.42).

Intervalos de confiança

O intervalo com $100(1-\alpha)\%$ de confiança para p_{st} pode ser definido como

$$\left(p_{st} \pm z_{\alpha/2} \sqrt{\left(\frac{1-f_h}{n} \right) \sum_{h=1}^{H} W_h P_h (1-P_h)} \right), \tag{2.43}$$

assumindo aproximação normal para a distribuição de p_{st}.

Tamanho da amostra

No caso da alocação proporcional, Cochran (1977) define n tal que

$$n = \frac{\frac{\sum_{h=1}^{H} W_h p_h (1-p_h)}{d^2}}{1 + \frac{1}{N}\left(\frac{\sum_{h=1}^{H} W_h p_h (1-p_h)}{d^2} \right)} = \frac{n_0}{1 + \frac{n_0}{N}}, \tag{2.44}$$

onde $n_0 = \frac{\sum_{h=1}^{H} W_h p_h (1-p_h)}{d^2}$ e d é o erro máximo admitido para as estimativas.

2.4 Peso amostral

O peso amostral representa a quantidade de indivíduos da população representada por um determinado elemento da amostra (Lohr, 1999). Isso significa que cada indivíduo da amostra representa ele mesmo e outros indivíduos com características semelhantes à ele, mas que não foram selecionados na amostra. Nesta seção, o peso amostral será abordado considerando apenas os métodos de amostragem aleatória simples e estratificada.

Amostragem Aleatória Simples

Segundo Bolfarine e Bussab (2005), a probabilidade do i-ésimo indivíduo, $i = 1, \ldots, n$, pertencer à amostra é dada por

$$p_i = \frac{n}{N}. \tag{2.45}$$

Assim, o peso amostral referente ao i-ésimo indivíduo da amostra (w_i) é sempre o inverso da sua probabilidade de seleção (Lohr, 1999). No caso da amostragem aleatória simples, o peso é definido como

$$w_i = \frac{1}{p_i} = \frac{N}{n}. \tag{2.46}$$

Considere uma população com $N = 4$. Se uma amostra aleatória simples com $n = 2$ é selecionada a partir dessa população, a probabilidade do indivíduo i pertencer à amostra, com $i = \{1, 2\}$, é dada por

$$p_i = \frac{2}{4} = 0,5, \quad \forall i.$$

Com isso, o peso do i-ésimo indivíduo selecionado na amostra é dado por

$$w_i = \frac{1}{0,5} = 2, \quad \forall i.$$

Isso significa que cada indivíduo representa ele mesmo e mais 1 indivíduo da população.

De acordo com Lohr (1999), a soma dos pesos amostrais é igual ao tamanho da população, uma vez que toda a amostra representa toda a população. Assim,

$$N = \sum_{i=1}^{n} w_i. \tag{2.47}$$

Aplicando ao exemplo,

$$N = \sum_{i=1}^{2} w_i = 2 + 2 = 4.$$

Por fim, o estimador não viesado para o total populacional definido em (2.2) pode ser escrito como

$$\widehat{T} = \sum_{i=1}^{n} w_i y_i \tag{2.48}$$

e o estimador não viesado para a média populacional definido em (2.6) pode ser escrito como

$$\overline{y} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} \tag{2.49}$$

utilizando os pesos amostrais.

Amostragem Estratificada

Segundo Lohr (1999), a probabilidade de selecionar o i-ésimo, $i = 1, \ldots, n_h$, indivíduo no h-ésimo estrato, $h = 1, \ldots, H$, na amostra é dada por

$$p_{hi} = \frac{n_h}{N_h}. \tag{2.50}$$

Assim, o peso amostral referente ao i-ésimo indivíduo no h-ésimo estrato é definido como

$$w_{hi} = \frac{1}{p_{hi}} = \frac{N_h}{n_h}. \tag{2.51}$$

Considere uma população com $N = 8$ dividida igualmente em dois estratos ($H = 2$) tal que $N_h = 4$, para $h \in \{1, 2\}$. Se uma amostra estratificada contento dois indivíduos de cada estrato ($n_h = 2$) é selecionada a partir dessa população, a probabilidade do indivíduo i do estrato h pertencer à amostra, com $i \in \{1, 2\}$, é dada por

$$p_{hi} = \frac{2}{4} = 0,5, \quad \forall h, i.$$

Assim, os pesos amostrais são dados por

$$w_{hi} = \frac{1}{0,5} = 2, \quad \forall h, i.$$

Isso significa que que cada indivíduo representa ele mesmo e mais 1 indivíduo do mesmo estrato.

Como no caso da amostragem aleatória simples, a soma dos pesos é igual ao tamanho da população, isto é,

$$N = \sum_{h=1}^{H} \sum_{i=1}^{n_h} w_{hi}.$$

Aplicando ao exemplo,

$$N = \sum_{h=1}^{2} \sum_{i=1}^{2} w_i = 2 + 2 = 4.$$

Por fim, o estimador não viesado para o total populacional definido em (2.23) pode ser escrito como

$$\widehat{T}_{st} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} w_{hi} y_{hi}, \tag{2.52}$$

e o estimador não viesado para a média populacional definido em (2.26) pode ser escrito como

$$\overline{y}_{st} = \frac{\sum_{h=1}^{H} \sum_{i=1}^{n_h} w_{hi} y_{hi}}{\sum_{h=1}^{H} \sum_{i=1}^{n_h} w_{hi}} \tag{2.53}$$

utilizando os pesos amostrais, assim como na amostragem aleatória simples.

2.5 Outras técnicas

Existem outras técnicas clássicas de amostragem que não foram citadas neste capítulo. Algumas delas são:

- Amostragem por Conglomerados (AC), onde a população é dividida em subpopulações (conglomerados) e apenas alguns desses conglomerados são selecionados

via AAS e são coletadas informações de todos os indivíduos pertencentes à esses conglomerados (Bolfarine e Bussab, 2005).

- Amostragem em Dois Estágios (A2E), onde a população também é dividida em subpopulações. No primeiro estágio, assim como na AC, apenas algumas subpopulações são selecionadas via AAS, e no segundo estágio são selecionados na amostra apenas alguns indivíduos de cada subpopulação (Bolfarine e Bussab, 2005).

- Amostragem Sistemática (AS), onde há uma listagem de todos os indivíduos de uma população e pode-se sortear um indivíduo entre os k primeiros e, a partir dele, observar todo k-ésimo indivíduo da lista (Bolfarine e Bussab, 2005).

Além dessas técnicas de amostragem clássicas, existem outras técnicas de amostragem que são utilizadas apenas em situações específicas, como no caso de pesquisa por telefone. O próximo capítulo abordará a técnica e problemas envolvendo amostragem por telefone.

Capítulo 3

AMOSTRAGEM POR TELEFONE

3.1 Introdução

No Capítulo 2 foram abordados os conceitos básicos de algumas técnicas clássicas de amostragem e seus métodos de estimação. Neste capítulo, serão abordadas as técnicas e problemas referentes à amostragem por telefone e os mecanismos utilizados em pesquisas por telefone.

A pesquisa por telefone é uma das muitas técnicas de pesquisa utilizadas e podem ser empregadas em vários tipos de pesquisas, como pesquisas de opinião. Essa técnica envolve, entre outros fatores, o estabelecimento da estratégia de amostragem correta e a coleta de dados por meio de entrevistas telefônicas (National Public Research, 2019).

A coleta de dados em pesquisas por telefone por ser feita por meio de um sistema CATI (do inglês, *Computer-Assisted Telephone Interview*), no qual os entrevistadores lêem as perguntas que aparecem no computador e as respostas dos participantes são inseridas diretamente no computador (National Public Research, 2019). No entanto, antes da coleta de dados, é necessário estabelecer a estratégia de amostragem adequada

que será empregada na pesquisa.

3.2 Listagem de telefones

Talvez uma das abordagens mais antigas de amostragem em pesquisas por telefone seja a partir de listas existentes de indivíduos, endereços ou organizações que possuem números de telefone, ou seja, bases ou quadros de telefone (do inglês, *telephone frames*) (Lepkowski et al., 2008). Essas listas podem ser obtidas através de empresas de telefonia ou registros governamentais, embora possa não ser um processo fácil.

No entanto, como discutido em Lepkowski et al. (2008), as listas telefônicas não incluem os domicílios e os indivíduos cujos números de telefone não estão registrados, aumentando o viés de cobertura. Quando todos os indivíduos da população-alvo possuem acesso a um numero de telefone e há o registro de todos esses números, o que ocorre é um processo de amostragem clássico, onde a única diferença está na forma de acessar os elementos da amostra e coletar as informações de interesse.

A partir dos quadros de telefone é possível realizar qualquer método de amostragem, desde que esses quadros, ou bases, possuam as informações necessárias para a aplicação de cada método. Caso o interesse seja realizar uma pesquisa por meio de uma amostragem aleatória simples, é necessário apenas informação à respeito dos números de telefone dos indivíduos da população-alvo, a partir disso basta enumerá-los (de 1 até N) e sortear os n elementos da amostra. Por exemplo, caso se deseje realizar uma pesquisa por telefone com os professores da Universidade de Brasília (UnB), uma forma de conseguir acesso aos números de telefone é através da própria UnB, assumindo que ela possui esse cadastro para uso interno. Assim, basta

enumerar esses números e sortear aleatoriamente os elementos da amostra.

Apesar de ser o método mais simples, nem sempre a amostragem aleatória simples é o método mais interessante para o pesquisador. Assim, caso o interesse seja realizar uma amostragem estratificada, é necessário ter, além da informação dos números de telefone, a informação necessária para definir o estrato (por exemplo, o sexo do indivíduo associado ao número de telefone). Assim, basta enumerar os indivíduos dentro de cada estrato h (de 1 até N_h) e sortear os n_h elementros da amostra em cada estrato. Seguindo o exemplo dos professores da UnB, para selecionar uma amostra estratificada, por exemplo, pelo cargo do professor (titular, associado, adjunto, assistente e auxiliar) é necessário que a lista disponibilizada tenha essa informação associada à cada número de telefone.

O Instituto Brasileiro de Geografia e Estatística (IBGE) realizou, em 2020, a Pesquisa Nacional por Amostra de Domicílios (PNAD COVID19) cuja coleta de dados foi feita por meio de ligações telefônicas (IBGE, 2020b). Para a realização da pesquisa, o IBGE utilizou os números de telefone cadastrados dos domicílios que participaram da PNAD Contínua no primeiro trimestre de 2019 (IBGE, 2020b). Este é um exemplo de pesquisa por telefone utilizando quadro (ou lista) de telefones aplicada à nível nacional.

Apesar de suas vantagens na realização de pesquisas por telefone, nem sempre se tem acesso a uma lista de telefones da população-alvo. Um exemplo disso é estimar, por meio de uma pesquisa por telefone, a quantidade de brasileiros que receberam a primeira parcela do Auxílio Emergencial em 2020. O processo para o pagamento do

Auxílio Emergencial revelou 46 milhões de brasileiros "inexistentes" para o governo, ou seja, brasileiros que não estavam em nenhuma lista do governo (G1, 2020). Neste caso, conseguir os números de telefone de todos os brasileiros com acesso à telefonia pode ser um processo quase impossível.

Assim, uma forma de resolver esse problema é construir uma lista de telefones totalmente aleatória onde não há nenhuma informação prévia à respeito do indivíduo associado ao número ou se há um indivíduo associado ao número gerado.

3.3 Plano de Numeração Brasileiro

Segundo Anatel (2020b), o Plano de Numeração Brasileiro segue as recomendações da União Internacional de Telecomunicações (UTI), que definiu para o Brasil o Código de País no formato [55], o Código de Área (ou DDD) no formato de dois dígitos e o número do assinante no formato de oito dígitos (quatro como prefixo e quatro como sufixo), no caso da telefonia fixa, e nove dígitos (cinco como prefixo e quatro como sufixo), no caso da telefonia móvel celular. O formato dos números de telefones definidos para o Brasil, para telefonia fixa e móvel, está ilustrado na Figura 3.1.

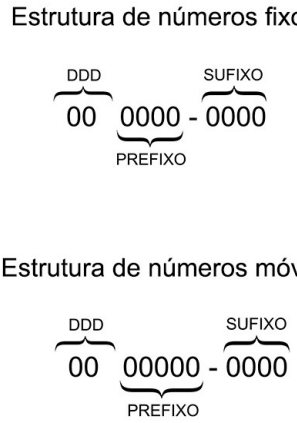

Figura 3.1: Estrutura básica de números telefônicos adotada no Brasil.

Em casos de pesquisas por telefone realizadas em território nacional, os números de telefone devem seguir essa estrutura e a não inclusão de alguma de suas partes estruturais pode acabar invalidando os mesmos. A única parte que pode ser desconsiderada, sem prejuízo à pesquisa, é o Código do País, exceto quando as ligações são realizadas de outro país.

Exemplo 3.1: Assumindo que a Anatel disponibiliza uma listagem dos números ativos no Distrito Federal, juntamente com o tipo de telefone (F - Fixo, M - Móvel), para a realização de uma pesquisa por telefone e ela contenha os seguintes dados:

Tabela 3.1: Exemplo de lista de números de telefone ativos no Distrito Federal.

TIPO	DDD	PREFIXO	SUFIXO
F	61	0000	2000
F	61	0000	2001
F	61	0000	2003
⋮	⋮	⋮	⋮
F	61	0000	2199
M	61	00001	1999
⋮	⋮	⋮	⋮
M	61	00001	4599

O primeiro número presente na Tabela 3.1 é o número 61 0000-2000, referente à telefonia fixa, no entanto, ele está dividido entre o Código de Área, prefixo e sufixo, assim como todos os outros. Assim, para que seja possível utilizar esses números, assumindo que a estrutura da base de telefones seja a mesma que a da Tabela 3.1, é necessário manipulá-los computacionalmente.

Ainda em relação à Tabela 3.1, é possível notar que o número 61 0000-2002 não se encontra na lista, indicando que este número não está ativo. Assim, não será necessário consumir tempo e esforço discando para ele, aumentando a eficácia na coleta de dados da pequisa.

3.4 Discagem de Dígitos Aleatórios

A Discagem de Dígitos Aleatórios ou, do inglês, *Random Digit Dialing* (RDD), é uma técnica de amostragem probabilística utilizada em pesquisas telefônicas como forma de selecionar indivíduos de uma população através da discagem para números de telefones aleatórios (IBPAD, 2020). Quando existe uma porcentagem muito baixa da população sem acesso a telefonia, seja fixa ou móvel, esse método pode ser mais vantajoso, economicamente falando, para obter cobertura de uma determinada área.

Segundo o IBPAD (2020), no Brasil, a discagem de dígitos aleatórios permite o controle geográfico da amostra por meio da inclusão do Código de Área (DDD) nos números sorteados. Com base nisso, é possível realizar uma pesquisa por telefone aplicando o método de Amostragem Estratificada, considerando cada Unidade da Federação (UF) como uma subpopulação (estrato) diferente. Caso o DDD associado ao número seja gerado de forma aleatória, assim como o número de telefone, têm-se

uma AAS aplicada ao país inteiro. Entretanto, em casos onde o DDD é desconsiderado na geração de um número, a amostra se limita apenas à UF da qual as ligações estão sendo realizadas, gerando uma AAS local.

Apesar de ser possível a aplicação de algumas técnicas clássicas de amostragem por meio do método de discagem de dígitos aleatórios, uma amostragem por conglomerados em dois ou mais estágios acaba se tornando inviável, uma vez as UFs são as menores subpopulações possíveis em um plano amostral baseado neste método, já que só é possível obter mais informações à respeito do indivíduo selecionado na amostra após o mesmo atender o telefone e concordar em participar da pesquisa.

O método de discagem de dígitos aleatórios é uma alternativa quando não se tem uma lista com os números de telefone da população-alvo da pesquisa. Por meio dele, qualquer número de telefone pode ser sorteado, estando habilitado ou não e não há garantia de que pertença à um indivíduo da população-alvo.

Em pesquisas por telefone realizadas no Brasil por meio do método em questão, os números podem ser sorteados a partir de uma lista disponibilizada pela Anatel, onde existem todos os números habilitáveis do Brasil. Um número habilitável é um número que está disponível para uso, mas não significa que esse número esteja ativo (habilitado) e, devido a isso, este sorteio pode conter uma grande quantidade de números inexistentes. Uma consulta realizada no dia 18 de agosto de 2020 mostrou que, no Brasil, existiam cerca de 223 milhões de números fixos habilitáveis e quase 555 milhões de números móveis (Anatel, 2020c). Em resumo, as bases disponibilizadas pela Anatel contêm todas as faixas de números habilitáveis no Brasil para um

determinado DDD e um determinado prefixo.

A estrutura básica das bases contendo todos os números habilitáveis do país está ilustratada nas Figuras 3.2 e 3.3.

DDD	PREFIXO	FAIXA INICIAL	FAIXA FINAL
61	0000	0	999
61	0000	2000	2999
12	0000	5000	5999
12	0000	0	999
43	0000	9000	9999

61 0000-2000
61 0000-2001
61 0000-2002
.
.
.
61 0000-2999

Figura 3.2: Ilustração da estrutura básica da lista de números fixos habilitáveis disponibilizada pela Anatel.

DDD	PREFIXO	FAIXA INICIAL	FAIXA FINAL
44	00000	1000	1999
44	00000	4000	4999
51	00000	0	999
51	00000	2000	2999
71	00000	0	999

44 00000-4000
44 00000-4001
44 00000-4002
.
.
.
44 00000-4999

Figura 3.3: Ilustração da estrutura básica da lista de números móveis habilitáveis disponibilizada pela Anatel.

Considerando o Exemplo 3.1, caso não seja possível ter acesso à listagem de números habilitados da população-alvo e o método de discagem de dígitos aleatórios seja adotado, números que não apareceram antes podem ser incluídos agora na listagem de telefones. Observando a Figura 3.2, é possível notar que o número 61 0000-2002 (que não aparece na Tabela 3.1) está presente em uma das faixas de números e pode ser sorteado para fazer parte da lista de telefones que será utilizada na pesquisa. A inclusão de números inexistentes pode acabar demandando um esforço maior para a realização da pesquisa, entretanto, nesse método não é possível saber se os números selecionados existem ou não.

3.5 Estimação e peso amostral

Em pesquisas por telefone, a inferência dos dados da amostra para a população-alvo é feita pela aplicação de fórmulas que levam em consideração as características do plano amostral, entretanto, não existe um protocolo único para o cálculo do peso amostral (Lepkowski et al., 2008).

Um problema comum em pesquisas por telefone é a não cobertura, ou o viés de cobertura. Em 2018, 14,3% dos brasileiros com 10 anos ou mais não possuíam nenhum tipo de acesso à telefonia, seja fixa ou móvel (IBGE, 2020c). Essa parte da população acaba se tornando inacessível em uma pesquisa telefônica realizada no Brasil.

Em pesquisas por telefone que utilizam o método de dígitos aleatórios, ao contrário do que acontece na amostragem aleatória simples, os indivíduos da população possuem uma probabilidade desigual e desconhecida de pertencer à amostra. Isso se deve ao fato de que não há como garantir que cada indivíduo da população-alvo possua acesso a apenas uma linha telefônica, ou seja, não há como garantir que a ligação entre o indivíduo e a lista de telefones seja "um-para-um" (do inglês, *one-to-one*), como define Lepkowski et al. (2008). Além dela, existem mais cinco ligações entre indivíduo e lista apresentadas por Lepkowski et al. (2008), sendo elas:

- ligação "um-para-nenhum" (do inglês, *one-to-none*), onde há números de telefone nas listas que não estão associados à nenhum indivíduo da população-alvo;

- ligação "nenhum-para-um" (do inglês, *none-to-one*), onde há indivíduos da

população-alvo sem acesso à nenhum número de telefone, o que os torna inacessíveis em uma pesquisa por telefone;

- ligação "muitos-para-um" (do inglês, *many-to-one*), onde um único indivíduo da população-alvo pode ser acessado por vários números de telefone;
- ligação "um-para-muitos" (do inglês, *one-to-many*), onde vários indivíduos da população-alvo podem ser acessados por um único número de telefone;
- ligação "muitos-para-muitos" (do inglês, *many-to-many*), quando vários indivíduos da população-alvo podem ser acessados por vários números de telefone.

Os problemas amostrais citados devem ser corrigidos caso deseje obter estimativas mais confiáveis e esta correção é feita por meio do peso amostral. Assim, como explicam Lepkowski et al. (2008), algumas combinações possíveis para o cálculo do peso do i-ésimo respondente de uma amostra por telefone podem ser feitas a partir dos seguintes componentes:

1. Peso básico (B_i).
2. Ajuste de não cobertura ($A_i^{(cob)}$).
3. Ajuste de não-resposta ($A_i^{(nr)}$).
4. Ajuste de calibração dos pesos ($A_i^{(cal)}$).

O DataSenado adotou uma estrutura diferente, considerando, dentre os ajustes enumerados, apenas o ajuste de não-resposta e a calibração dos pesos e adicionou um ajuste de seleção para corrigir a probabilidade dos respondentes serem selecionados

por meio de alguma linha telefônica que ele tenha acesso. Essa estrutura será discutida com mais detalhes no próximo Capítulo.

Segundo Lepkowski et al. (2008), quando todos os ajustes enumerados são feitos, o peso final associado ao respondente i (w_i) é o produto dos pesos obtidos em cada etapa, ou seja,

$$w_i = B_i \times A_i^{(cob)} \times A_i^{(nr)} \times A_i^{(cal)}. \tag{3.1}$$

Quando a amostragem é estratificada, essas etapas são repetidas e os pesos são aplicados para cada estrato.

O *peso base* (B_i) é geralmente calculado pelo inverso da probabilidade de seleção do indivíduo i na amostra (Lepkowski et al., 2008). Em resumo, B_i é o peso amostral associado ao indivíduo i e pode ser definido como em (2.46) ou (2.51).

O *ajuste de não cobertura* ($A_i^{(cob)}$) é calculado para subdivisões da população definidas com base em características que estão, ou possam estar, correlacionadas com a probabilidade de possuir ou não serviço telefônico (Lepkowski et al., 2008). Assim, para o i-ésimo indivíduo na h-ésima subpopulação $A_{hi}^{(cob)} = 1/\widehat{c}_h$, onde $\widehat{c}_h$ é a taxa de cobertura telefônica estimada.

O *ajuste de não-resposta* ($A_i^{(nr)}$) pode ser definido com o recíproco da propensão de resposta estimada para o i-ésimo indivíduo ($\widehat{p}_i$), ou seja, $A_i^{(nr)} = 1/\widehat{p}_i$, onde $\widehat{p}_i$ pode ser obtida com base em experiências anteriores de não-resposta na amostra (Lepkowski et al., 2008).

O *ajuste de calibração dos pesos* ($A_i^{(cal)}$) é o passo final onde há o ajuste da amostra ponderada para a distribuição da população com base em um conjunto de

variáveis categóricas (ou categorizadas) (Lepkowski et al., 2008). A calibração pode ser feita de duas maneiras: **pós-estratificação** (ou ponderação) para a distribuição conjunta da população (que pode ser obtida por meio de uma fonte externa) com base em certas variáveis ou **ajuste proporcional iterativo** (do inglês, *raking*) das distribuições marginais conjuntas da população com base nas variáveis de calibração (Lepkowski et al., 2008).

Segundo Lepkowski et al. (2008), a pós-estratificação pode ser implementada definindo uma classificação cruzada (ou tabela cruzada) das variáveis de calibração categóricas. O ajuste de pós-estratificação é calculado para a h-ésima célula de ajuste, $h = 1, \ldots, H$, como $A_h^{(cal)} = N_h / \sum_{i=1}^{r_h} w_{hi}$, onde N_h é a contagem da população na h-ésima célula de calibração, tal que $N = \sum_{h=1}^{H} N_h$ é o tamanho total da população, w_{hi} é o peso final obtido pelo i-ésimo indivíduo da amostra e r_h é o tamanho amostral na h-ésima célula de ajuste.

O método *raking*, essencialmente, força os totais amostrais a coincidirem (separadamente) com os totais conhecidos da população e, para isso, seu algoritmo real calcula, repetidamente, a estimativa de pesos em cada conjunto das variáveis de calibração até que eles convirjam (Fricker e Anderson, 2015). Este trabalho dará mais ênfase no método *raking*, uma vez que foi o método de calibração adotado pelo DataSenado na pesquisa em questão.

3.5.1 Método *raking*

O algoritmo básico do *raking* é descrito em termos dos pesos atribuídos à n respondentes de uma pesquisa, ou seja, $w_i, i = 1, 2, \ldots, n$, mas quando a amostra não

está ponderada, pode-se assumir os pesos iniciais como $w_i = 1, \forall i$ (Battaglia et al., 2009). Em uma classificação cruzada com L linhas e C colunas, w_{lc} denota a soma de w_i na célula (l, c) e para indicar um somatório, basta substituir um subscrito pelo sinal + (Izrael et al., 2000).

Assim, os totais iniciais dos pesos amostrais das linhas e os totais das colunas são definidos por w_{l+} e w_{+c}, respectivamente, e define-se os totais de controle correspondentes por T_{l+} e T_{+c} (Izrael et al., 2000). O algoritmo do *raking* produz pesos ajustados, cujas somas são denotadas por m com um índice entre parênteses que denota a iteração em que o algoritmo se encontra (Izrael et al., 2000).

Segundo Battaglia et al. (2009), combinando os totais de controle para as linhas, T_{l+}, o passo inicial do algoritmo *raking* é dado por

$$m_{lc}^{(0)} = w_{lc} \quad (l = 1, \ldots, L; c = 1, \ldots, C), \tag{3.2}$$

e os passos seguintes, considerando 2 iterações, são dados por

$$m_{lc}^{(1)} = m_{lc}^{(0)} \left(\frac{T_{l+}}{m_{l+}^{(0)}} \right) \quad (\forall c \text{ em cada } l),$$

$$m_{lc}^{(2)} = m_{lc}^{(1)} \left(\frac{T_{+c}}{m_{+c}^{(1)}} \right) \quad (\forall l \text{ em cada } c).$$

Aplicando a um exemplo numérico, considere uma amostra onde se conheça a distribuição conjunta das variáveis de calibração (Tabela 3.2), mas se conheça apenas as distribuições marginais dessas variáveis para a população.

Tabela 3.2: Exemplo de distribuição conjunta de duas variáveis (faixa etária e sexo).

Amostra	Feminino	Masculino	Total
16-39	55	36	91
40-60	43	21	64
60+	22	23	45
Total	120	80	200

Tabela 3.3: Exemplo de distribuição marginal de duas variáveis (faixa etária e sexo) para a população.

População	Feminino	Masculino	Total
16-39	?	?	85
40-60	?	?	69
60+	?	?	46
Total	105	95	200

Assim, iniciando pela faixa etária (linhas), o *raking* segue como indicado abaixo.

Raking para o sexo feminino:

$$m_{11}^{(1)} = m_{11}^{(0)} \left(\frac{T_{1+}}{m_{1+}^{(0)}} \right) = 55 \times \frac{85}{91} = 55 \times 0,934065 = 51,374,$$

$$m_{21}^{(1)} = m_{21}^{(0)} \left(\frac{T_{2+}}{m_{2+}^{(0)}} \right) = 43 \times \frac{69}{64} = 43 \times 1,078125 = 46,359,$$

$$m_{31}^{(1)} = m_{31}^{(0)} \left(\frac{T_{3+}}{m_{3+}^{(0)}} \right) = 22 \times \frac{46}{45} = 22 \times 1,02222 = 22,489.$$

Raking para o sexo masculino:

$$m_{12}^{(1)} = m_{12}^{(0)} \left(\frac{T_{1+}}{m_{1+}^{(0)}} \right) = 36 \times \frac{85}{91} = 36 \times 0,934065 = 33,626,$$

$$m_{22}^{(1)} = m_{22}^{(0)} \left(\frac{T_{2+}}{m_{2+}^{(0)}} \right) = 21 \times \frac{69}{64} = 21 \times 1,078125 = 22,640,$$

$$m_{32}^{(1)} = m_{32}^{(0)} \left(\frac{T_{3+}}{m_{3+}^{(0)}} \right) = 23 \times \frac{46}{45} = 23 \times 1,02222 = 23,511.$$

Os valores 0,934065, 1,078125 e 1,02222 são os pesos obtidos na primeira rodada para as faixas etárias 16-39, 40-60 e 60+, respectivamente. Assim, os totais marginais para faixa etária e a distribuição conjunta estimada na primeira iteração do *raking* é dada por:

Tabela 3.4: Aplicação do *raking* para a faixa etária (iteração 1).

Iteração 1	Feminino	Masculino	Total
16-39	51,374	33,626	**85**
40-60	46,360	22,640	**69**
60+	22,489	23,511	**46**
Total	120,223	79,777	200

Com os totais de controle das linhas ajustados na iteração 1, percebe-se que os totais das colunas estão desajustados. Agora, é necessário aplicar o *raking* para a variável sexo (colunas). Repetindo o procedimento para as colunas (iteração 2), têm-se o seguinte resultado:

Tabela 3.5: Aplicação do *raking* para o sexo (iteração 2).

Iteração 2	Feminino	Masculino	Total
16-39	44,870	40,043	84,913
40-60	40,490	26,960	67,450
60+	19,640	27,997	47,637
Total	**105**	**95**	200

O procedimento deve seguir até que a diferença $m_{lc}^{(k)} - m_{lc}^{(k-1)}$, para todo (l, c) no passo k, seja menor ou igual a um certo valor, denominado tolerância, ou seja, até que os totais convirjam, ou então, até que um número k de iterações seja atingido. Assim, definindo uma tolerância de 0,0001 (isto é, $m_{lc}^{(k)} - m_{lc}^{(k-1)} \leq 0,0001$) e um número máximo de iterações igual a 50, o resultado do método *raking* é representado na Tabela 3.6.

Tabela 3.6: Aplicação do *raking* para as variáveis sexo e faixa etária (iteração final).

Iteração final	Feminino	Masculino	Total
16-39	44,785	40,215	85
40-60	41,318	27,682	69
60+	18,897	27,103	46
Total	105	95	200

O resultado final é obtido pelo produto dos pesos obtidos na rodada de parada k com os valores da distribuição conjunta obtidos na rodada $k-1$. Note que os totais marginais obtidos ao final do procedimento (Tabela 3.6) é igual aos totais marginais da população de interesse (Tabela 3.3). O *raking* também pode ser utilizado quando se deseja ajustar um conjunto de dados utilizando três variáveis ou mais, mas os totais de controle, geralmente, envolvem variáveis únicas (Izrael et al., 2000).

3.5.2 Viés de não-resposta

Em pesquisas por telefone realizadas utilizando o método de discagem de dígitos aleatórios, um problema comum e que preocupa os pesquisadores é o viés de não-resposta (Groves et al., 2001). Esse viés acontece pelo alto número de recusas (ou alta taxa de não-resposta) que acaba gerando uma grande dificuldade para conseguir que um indivíduo que foi selecionado na amostra participe da pesquisa. Isso acaba resultando, na grande maioria das vezes, na substituição desse indivíduo por outro de mais fácil acesso. A questão é que esses dois grupos, respondentes e não respondentes, podem diferir muito dependendo do tema da pesquisa.

A taxa de resposta em pesquisas telefônicas pode ser influenciada por diversos fatores, como o tema da pesquisa, o órgão responsável pela pesquisa, o horário e o dia que as ligações são feitas. Na tentativa de minimizar esse viés, são adotados vários métodos que visam um melhor gerenciamento da amostra e o incentivo à colaboração dos indivíduos na pesquisa. Exemplos disso são o agendamento das pesquisas, a definição de um número de tentativas de chamadas, a definição do tempo entre essas tentativas, a elaboração de uma introdução que incentive a colaboração do indivíduo,

incentivos monetários, entre outros (Sangster, 2003).

Para o cálculo da taxa de resposta é preciso entender o conceito de elegibidade de uma linha telefônica. Uma linha é considerada elegível caso ela pertença a um indivíduo que faz parte da população alvo da pesquisa (AAPOR, 2016), que, por sua vez, é definida por Bolfarine e Bussab (2005) como a população que se pretende atingir. Uma linha telefônica é considerada com elegibilidade desconhecida quando não é possível saber se ela pertence a um indivíduo da população alvo, por exemplo, quando um número é selecionado na amostra, porém resulta em ligação não atendida ou linha ocupada (AAPOR, 2016). Por fim, uma linha telefônica é considerada inelegível quando seu usuário não pertence à população alvo da pesquisa, quando a linha não está ativa (ou habilitada) ou quando a linha telefônica pertence à uma empresa (AAPOR, 2016).

A taxa de resposta pode ser calculada de várias maneiras, que irá depender do pesquisador e do propósito do seu cálculo (Groves, 1989). Algumas definições importantes para definir seu cálculo são dadas por AAPOR (2016):

RR = Taxa de Resposta ou, do inglês, *Response Rate*;

I = Entrevista completa;

P = Entrevista parcial;

R = Recusa ou ligação interrompida;

NC = Sem contato;

O = Outro;

UH = Elegibilidade desconhecida, se é domicílio;

UO = Elegibilidade desconhecida, outro;

e = Proporção estimada de casos de elegibilidade desconhecida mas que são elegíveis.

Assim, uma das formas que AAPOR (2016) apresenta para o cálculo da taxa de resposta é dada por

$$RR1 = \frac{I}{(I+P)+(R+NC+O)+1\times(UH+UO)}. \tag{3.3}$$

A Taxa de Resposta 1 (RR1), ou taxa de resposta mínima, considera todos os casos de elegibilidade desconhecida no seu cálculo ($e = 1$), segundo a AAPOR (2016), e pode ser utilizada como estimativa de $\widehat{p}_i$ na etapa de ajuste de não-resposta na Equação (3.1) e, dessa forma, não é preciso se basear em experiências anteriores estimar a propensão de resposta na amostra.

Capítulo 4

MATERIAIS E MÉTODOS

4.1 Introdução

Neste capítulo são apresentados os materiais e métodos utilizados neste estudo. Os materiais principais são o banco de dados da pesquisa realizada pelo Instituto de Pesquisa DataSenado e o banco de dados do Portal da Transparência do Governo Federal, que contém informações sobre todos os pagamentos realizados de todas as parcelas do Auxílio Emergencial. O método utilizado será a incorporação dos pesos de amostragem por telefone para a geração de intervalos de confiança.

Os métodos utilizados neste estudo vizam avaliar se a pesquisa por telefone é capaz de estimar o parâmetro real de interesse por meio de suas estimativas intervalares.

4.2 Materiais

Os materiais utilizados foram disponibilizados pelo Instituto de Pesquisa DataSenado, pelo Portal da Transparência do Governo Federal e pelo Instituto Brasileiro de Geografia e Estatística.

4.2.1 Portal da Transparência

Os dados referentes ao Auxílio Emergencial foram disponibilizados pelo Portal da Transparência (2020). A base de dados contém informações sobre todas as parcelas pagas do benefício, como mês e ano do pagamento, nome do beneficiário, município que reside o beneficiário, número da parcela (primeira, segunda, etc.), entre outras informações, segundo o Portal da Transparência (2020), como mostra a Tabela 4.1.

Tabela 4.1: Dicionário do conjunto de dados dos pagamentos do Auxílio Emergencial.

COLUNA	**DESCRIÇÃO**
MÊS DISPONIBILIZAÇÃO	Ano/Mês a que se refere a parcela, no formato AAAAMM.
UF	Sigla da Unidade Federativa do beneficiário do Auxílio Emergencial.
CÓDIGO MUNICÍPIO IBGE	Código, do IBGE (Instituto Brasileiro de Geografia e Estatística), do município do beneficiário do Auxílio Emergencial.
NOME MUNICÍPIO	Nome do município do beneficiário do Auxílio Emergencial.
NIS BENEFICIÁRIO	Número de Identificação Social (NIS) do beneficiário do Auxílio Emergencial, caso possua.
CPF BENEFICIÁRIO	Número do Cadastro de Pessoas Físicas (CPF) do beneficiário do Auxílio Emergencial, caso possua.
NOME BENEFICIÁRIO	Nome do beneficiário do Auxílio Emergencial.
NIS RESPONSÁVEL	Número de Identificação Social (NIS) do responsável pelo beneficiário do Auxílio Emergencial, caso possua.
CPF RESPONSÁVEL	Número do Cadastro de Pessoas Físicas (CPF) do responsável pelo beneficiário do Auxílio Emergencial, caso possua.
NOME RESPONSÁVEL	Nome do responsável pelo beneficiário do Auxílio Emergencial, caso possua.
ENQUADRAMENTO	Identifica se o beneficiário é do grupo Bolsa Família, Inscrito no Cadastro Único (CadÚnico) ou Não Inscrito no Cadastro Único (ExtraCad).
PARCELA	Número sequencial da parcela disponibilizada.
OBSERVAÇÃO	Indica alterações na parcela disponibilizada como, por exemplo, se foi devolvida ou está retida.
VALOR BENEFÍCIO	Valor disponibilizado na parcela.

Os conjuntos de dados foram disponibilizados de acordo com o mês. Para este estudo, foram selecionadas apenas os dados referentes aos meses de abril e maio de 2020 e filtrados apenas os pagamentos referentes à primeira parcela do benefício e que não foram retidos ou devolvidos.

4.2.2 Pesquisa DataSenado

O Instituto de Pesquisa DataSenado disponibilizou dados referentes à pesquisa Coronavírus, realizada entre os dias 18 e 20 de maio de 2020, segundo o DataSenado (2020a), juntamente com um documento que contém informações técnicas sobre a metodologia adotada na realização da pesquisa.

Segundo DataSenado (2020b), a amostra da pesquisa é composta por 1.200 respondentes e população-alvo é formada por cidadãos brasileiros com 16 anos ou mais. Os participantes foram selecionados por meio de uma amostragem estratificada com alocação proporcional ao tamanho da população-alvo em cada UF, segundo as estimativas mais recentes do IBGE em relação ao período de realização da pesquisa (DataSenado, 2020c).

A coleta de dados foi feita por meio de um sistema CATI e os números de telefone discados foram gerados por meio do método de discagem de dígitos aleatórios e a partir de dados disponibilizados pela Anatel sobre os números habilitáveis existentes no Brasil (DataSenado, 2020c). As ligações foram feitas até que as 1.200 entrevistas fossem concluídas, entretanto, a entrevista só era realizada se o respondente confirmasse que pertencia à população-alvo e concordasse em participar da pesquisa (DataSenado, 2020b).

Os respondentes foram perguntados, entre outras coisas, acerca da solicitação do Auxílio Emergencial e, aqueles que afirmaram ter solicitado, foram perguntados sobre o recebimento ou não da primeira parcela do benefício (DataSenado, 2020a). Cada estimativa obtida e divulgada pelo DataSenado é acompanhada pela sua respectiva margem de erro calculada com 95% de confiança (DataSenado, 2020c).

Além dos dados sobre os respondentes, o DataSenado também disponibilizou dados que possibilitam a realização de alguns cálculos para a correção de problemas amostrais, tais como:

- taxa de resposta por região, para a correção do viés de não-resposta;
- quantidade de linhas telefônicas que os respondentes têm acesso e quantas pessoas têm acesso a cada linha, para correção da probabilidade de seleção;
- probabilidade de selecionar um número válido (habilitado) na UF a qual pertence o DDD do número utilizado para responder a pesquisa, para correção da probabilidade de seleção;
- informações sociodemográficas dos respondentes, para calibração dos pesos amostrais.

A nota técnica da pesquisa, disponível em DataSenado (2020c), descreve o delineamento amostral e os métodos ponderação e calibração adotados pelo DataSenado. Cada etapa da ponderação dispõe das fórmulas utilizadas para a realização dos cálculos necessários e serão descritas resumidamente a seguir.

Ponderação

Os dados coletados foram ponderados considerando os seguintes aspectos: taxas de resposta, probabilidade de seleção e calibração dos pesos amostrais por meio da distribuição da população-alvo (DataSenado, 2020c).

As taxas de resposta foram calculadas por meio da Equação (3.3) para cada região do Brasil e tipo de telefonia (fixa e móvel), a partir de dados coletados das discagens telefônicas realizadas durante a pesquisa (DataSenado, 2020c). O ajuste de não-resposta foi calculado conforme definido na Equação (3.1) para cada região e tipo de telefonia. Assim, o ajuste de não-resposta para cada região brasileira e tipo de telefonia é dado por

$$A_{reg,tipo}^{(nr)} = \frac{1}{RR1_{reg,tipo}} \tag{4.1}$$

A probabilidade de seleção do i-ésimo respondente foi calculada com base na quantidade de linhas telefônicas que ele tinha acesso (N_{di}), na quantidade de pessoas que tinham acesso a cada linha telefônica j (δ_{ij}) e na probabilidade de selecionar um número válido na UF a qual o número do respondente pertencia (π_h), segundo dados da Anatel (DataSenado, 2020c).

Considere $t_{disc,h}$ o total de números habilitados discados na pesquisa para a h-ésima UF ($h = 1, \ldots, 27$) e t_h o total de números habilitados existentes na h-ésima UF no mês de abril segundo Anatel (2020a). DataSenado (2020c) define π_h como

$$\pi_h = \frac{t_{disc,h}}{t_h}. \tag{4.2}$$

O tipo de telefonia não foi considerado em (4.2) pelo fato da lista de números de

telefone ter sido gerada de forma que π_h não dependesse do tipo de telefonia, isto é, a probabilidade de selecionar um número fixo e móvel na lista de telefones é a mesma (DataSenado, 2020c).

Assim, a probabilidade de selecionar o i-ésimo participante da h-ésima Unidade da Federação na amostra, é definida por DataSenado (2020c) como

$$f_{hi} = \pi_h \sum_{i=1}^{N_{di}} \frac{1}{\delta_{ij}}, \tag{4.3}$$

onde $i = 1, \ldots, 1.200$ e $j = 1, \ldots, N_{di}$ e o ajuste da probabilidade de seleção ($A_{hi}^{(sel)}$) é dado pelo inverso de f_{hi}.

Considerando a Equação (3.1) e segundo DataSenado (2020c), o peso não calibrado associado ao i-ésimo respondente é dado pelo produto dos pesos definidos anteriormente. Assim,

$$w^*_{reg,tipo,h,i} = A_{hi}^{(sel)} \times A_{reg,tipo}^{(nr)}. \tag{4.4}$$

A calibração de w^* foi feita por meio do método *raking* e, para que sua aplicação fosse possível, a coleta dos dados sociodemográficos dos respondentes foi feita de forma similar a da PNAD Contínua (DataSenado, 2020c). O conjunto de variáveis utilizado para a calibração inclui sexo, idade, escolaridade e cor/raça e o *raking* foi aplicado de forma que os totais amostrais que refletissem a distribuição da população por região do Brasil (DataSenado, 2020c).

Em relação ao conjunto de variáveis de calibração, os dados referentes à idade foram agrupados em: 16 a 29 anos, 30 a 39 anos, 40 a 49 anos, 50 a 59 anos e 60 anos ou mais; os dados de escolaridade foram agrupados em: até ensino fundamental incompleto, ensino fundamental completo, ensino médio completo e ensino superior

completo ou mais; e, por fim, os dados referentes à cor/raça foram agrupados em: branca e negra/indígena/amarela (DataSenado, 2020c). Assim, o método *raking* deve ajustar os totais amostrais referentes, por exemplo, a cor/raça branca de forma que eles reflitam a quantidade de pessoas (com 16 anos ou mais) autodeclaradas brancas em cada região brasileira.

4.2.3 PNAD Contínua

A Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD Contínua), relizada pelo IBGE, foi implantada em janeiro de 2012 em todo o Território Nacional (IBGE, 2020c). Os indicadores da pesquisa são divulgados, principalmente, de forma mensal, trimestral e anual e são relacionados à temas de habitação, características gerais dos moradores, educação, informações de força de trabalho, acesso à telefonia, entre outros (IBGE, 2020c).

Como foi visto na Equação (3.1), os totais populacionais a respeito da população-alvo são necessários para a calibração dos pesos amostrais, entretanto, a quantidade de pessoas que compõem a população brasileira depende do período de referência, como mostra a Tabela 4.2. Entre o primeiro trimestre de 2019 e o primeiro trimestre de 2020, a população brasileira cresceu cerca de 0,767%.

Tabela 4.2: População brasileira ao longo do tempo.

Período	Tamanho da população	Crescimento*
1^{o} trimestre de 2019	208.873.066	-
2^{o} trimestre de 2019	209.276.497	$0,193\%$
3^{o} trimestre de 2019	209.677.912	$0,192\%$
4^{o} trimestre de 2019	210.077.236	$0,19\%$
1^{o} trimestre de 2020	210.474.420	$0,189\%$

*O crescimento foi calculado com base no trimestre anterior.

Em 1º de agosto de 2010, período de referência do Censo 2010, também realizado pelo IBGE, a população brasileira era composta por 190.732.694 de pessoas (IBGE, 2020a). Entretanto, de 2010 até o primeiro trimestre de 2020, a população brasileira cresceu $10,35\%$ e, por esse motivo, os dados censitários não serão utilizados para calibrar os pesos amostrais.

Assim, devido às mudanças sofridas pela população, os dados acerca da população-alvo da pesquisa Coronavírus, realizada pelo DataSenado, foram obtidos por meio da PNAD Contínua referente ao 1º trimestre de 2020, dados mais recentes disponíveis no período de realização da pesquisa.

A população-alvo (16 anos ou mais com cor/raça declarada) no 1º trimestre de 2020 foi estimada em 166.298.359 (IBGE, 2020c). Essa será a população considerada para a calibração dos pesos amostrais da pesquisa.

4.3 Métodos

Para a realização deste estudo, serão estimados, de forma intervalar, o percentual e o total de pessoas que receberam a primeira parcela do Auxílio Emergencial, utilizando como base a metodologia adotada pelo DataSenado na realização da pesquisa Coronavírus.

Apesar do delineamento amostral da pesquisa considerar uma amostragem estratificada, em alguns casos será assumida uma amostragem aleatória simples com o objetivo de analisar o efeito de não adotar o plano amostral adequado nas estimativas obtidas, mesmo não sendo um meio adequado de se analisar esse efeito. Por fim, as estimativas serão analisadas das seguintes maneiras:

1. considerando uma AAS_s e sem a aplicação de pesos amostrais (isto é, $w_i = 1, \forall i$);

2. considerando uma AAS_s e aplicando o peso amostral simples, como definido em (2.46);

3. considerando uma AAS_s e aplicando o ajuste da probabilidade de seleção por UF, conforme a Equação (4.3);

4. considerando uma AAS_s e aplicando o ajuste de não-resposta, como definido em (4.1), além do ajuste do passo 3;

5. considerando uma AAS_s e aplicando o método *raking* para calibração dos pesos, além dos ajustes dos passos 3 e 4;

6. considerando uma AAS_s e aplicando o método *raking* para calibração dos pesos ajustados no passo 3, isto é, excluindo o efeito da não-resposta;

7. considerando uma AAE com alocação proporcional por Unidade da Federação[1] e sem a aplicação de pesos amostrais (isto é, $w_{hi} = 1, \forall h, i$);

8. considerando uma AAE com alocação proporcional por Unidade da Federação[1] e aplicando o peso amostral simples de cada estrato, como definido em (2.51);

9. considerando uma AAE com alocação proporcional por Unidade da Federação[1] e aplicando o ajuste da probabilidade de seleção por UF, como definido em (4.3);

[1]Plano amostral adotado na realização da pesquisa Coronavírus, segundo DataSenado (2020b).

10. considerando uma AAE com alocação proporcional por Unidade da Federação[1] e aplicando o ajuste de não-resposta, conforme a Equação (4.1), além do ajuste do passo 9;

11. considerando uma AAE com alocação proporcional por Unidade da Federação[1] e aplicando o método *raking* para calibração dos pesos, além dos ajustes dos passos 9 e 10;

12. considerando uma AAE com alocação proporcional por Unidade da Federação[1] e aplicando o método *raking* para calibração dos pesos ajustados no passo 9, isto é, excluindo o efeito da não-resposta.

As estimativas intervalares obtidas em cada passo serão comparadas a fim de verificar qual é o efeito de cada método de correção amostral adotado nos respectivos intervalos de confiança, além do efeito do plano amostral. Para isso, serão utilizados como base os dados populacionais disponibilizados pelo Portal da Transparência (2020). Os intervalos de confiança serão calculados considerando que os estimadores são normalmente distribuídos em torno dos valores populacionais, uma vez que a amostra possui um tamanho razoável para isso, e a variância de cada estimador será calculada por meio da linearização de Taylor (SAS Institute, 2020), onde, para uma amostra estratificada e por conglomerados e tomando como exemplo o caso da média

[1]Plano amostral adotado na realização da pesquisa Coronavírus, segundo DataSenado (2020b).

amostral, têm-se:

$$\hat{V}_h(\hat{\bar{Y}}) = \frac{n_h(1-f_h)}{n_h-1} \sum_{i=1}^{n_h} (e_{hi\cdot} - \bar{e}_{h\cdot\cdot})^2 \tag{4.5}$$

$$\hat{\bar{Y}} = \left(\sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij}\ y_{hij} \right) / \ w_{\cdots} \tag{4.6}$$

onde

$$e_{hi\cdot} = \left(\sum_{j=1}^{m_{hi}} w_{hij}\ (y_{hij} - \hat{\bar{Y}}) \right) / \ w_{\cdots} \tag{4.7}$$

$$w_{\cdots} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} \tag{4.8}$$

$$\bar{e}_{h\cdot\cdot} = \left(\sum_{i=1}^{n_h} e_{hi\cdot} \right) / \ n_h \tag{4.9}$$

tal que

- $h = 1, 2, \ldots, H$ representa o estrato,
- $i = 1, 2, \ldots, n_h$ representa o conglomerado dentro do estrato h e
- $j = 1, 2, \ldots, m_{hi}$ representa a unidade dentro do conglomerado i do estrato h.

Capítulo 5

ANÁLISE DOS RESULTADOS

5.1 Introdução

Este capítulo apresenta a análise dos resultados obtidos por meio dos dados descritos na Seção 4.2 e seguindo a metodologia descrita na Seção 4.3. Os resultados foram gerados por meio do *software* SAS 9.4.

5.2 Análise descritiva dos dados

A pesquisa do DataSenado estimou, entre outras coisas, quantas pessoas solicitaram o Auxílio Emergencial até o dia 20 de maio de 2020 e quantas, entre as que solicitaram, receberam a primeira parcela do benefício (DataSenado, 2020a). As Tabelas 5.1 e 5.2 foram geradas com base no banco de dados da pesquisa Coronavírus, disponibilizado pelo DataSenado, e mostram a distribuição dos dados para as características referentes à solicitação e ao recebimento da primeira parcela do Auxílio Emergencial.

Tabela 5.1: Distribuição dos respondentes segundo a solicitação do Auxílio Emergencial.

Solicitou	Total de respondentes
Sim	**472**
Não	727
Não sabe ou não respondeu	1
Total	1.200

Segundo o DataSenado (2020e), o texto da pergunta foi "Você solicitou o Auxílio Emergencial?" e as respostas possíveis foram "Sim", "Não" ou "Não sei ou prefiro não responder".

Tabela 5.2: Distribuição dos respondentes segundo o recebimento da primeira parcela do Auxílio Emergencial.

Recebeu	Total de respondentes
Sim	283
Não	188
Total	**471***

Segundo o DataSenado (2020e), o texto da pergunta foi "Você recebeu a primeira parcela do Auxílio Emergencial?" e as respostas possíveis foram "Sim", "Não" ou "Não sei ou prefiro não responder".
*Inconsistência nos dados.

Pela Tabela 5.1, têm-se que a amostra total da pesquisa é composta por 1.200 entrevistas (ver também DataSenado (2020b)), das quais, em 472, os entrevistados afirmaram ter solicitado o Auxílio Emergencial e em 727, afirmaram não ter solicitado. Dentre os que afirmaram ter solicitado o benefício, a Tabela 5.2 mostra que 283 afirmaram ter recebido a primeira parcela e 188 afirmaram não ter recebido. A Tabela 5.2 também mostra uma leve inconsistência nos dados, causada pela ausência de dados para um dos respondentes que afirmou ter solicitado o Auxílio Emergencial. Segundo o banco de dados disponibilizado pelo DataSenado, não há informação se esse indivíduo recebeu a primeira parcela ou não.

5.3 Ponderação dos dados

O plano amostral adotado na pesquisa, segundo o DataSenado (2020b), considerou uma amostragem estratificada e distribuída proporcionalmente entre as Unidades da Federação (estratos), como mostra a Tabela 5.3. Considerando uma amostragem aleatória simples (apesar de ser um erro metodológico), o peso simples é dado pela razão entre o tamanho da população de interesse e o tamanho da amostra total. Como mostra a Tabela 5.3, o tamanho da população-alvo foi estimado em 166.298.359 e o tamanho da amostra para todo o Brasil foi 1.200. Assim, o peso amostral simples para AAS_s é dado por

$$W_i = \frac{166.298.359}{1.200} \approx 138.581,96 \quad \forall i \in \{1, \ldots, 1.200\}$$

e a soma dos pesos resulta no total populacional de interesse.

Considerando uma amostragem estratificada e a população brasileira com 16 anos ou mais e cor/raça declarada, têm-se a distribuição da amostra da pesquisa e da população em cada estrato representada na Tabela 5.3. De acordo com (2.51), o peso atribuído ao estrato h é calculado pela razão entre o tamanho da população de interesse e o tamanho da amostra alocada no estrato h. No caso de Rondônia (estrato 1), o peso indicado na Tabela 5.3 foi calculado por

$$W_1 = \frac{1.359.832}{10} = 135.983,2.$$

Tabela 5.3: Distribuição da população e amostra por Unidade da Federação.

Unidade da Federação	Estrato (h)	População (N_h)	Amostra (n_h)	Peso (W_h)	$n_h \times W_h$
Rondônia	1	1.359.832	10	135.983,20000	1.359.832
Acre	2	626.468	5	104.411,33333	626.468
Amazonas	3	2.895.084	21	137.861,14286	2.895.084
Roraima	4	386.196	3	128.732,00000	386.196
Pará	5	6.351.100	46	141.135,55556	6.351.100
Amapá	6	601.742	5	120.348,40000	601.742
Tocantins	7	1.192.631	9	132.514,55556	1.192.631
Maranhão	8	5.165.607	37	139.611,00000	5.165.607
Piauí	9	2.506.699	18	131.931,52632	2.506.699
Ceará	10	7.164.553	52	137.779,86538	7.164.553
Rio Grande do Norte	11	2.732.835	20	136.641,75000	2.732.835
Paraíba	12	3.096.694	23	134.638,86957	3.096.694
Pernambuco	13	7.562.646	55	140.049,00000	7.562.646
Alagoas	14	2.511.863	18	132.203,31579	2.511.863
Sergipe	15	1.771.421	13	136.263,15385	1.771.421
Bahia	16	11.599.485	83	139.752,83133	11.599.485
Minas Gerais	17	17.094.351	123	138.978,46341	17.094.351
Espírito Santo	18	3.198.819	23	139.079,08696	3.198.819
Rio de Janeiro	19	14.363.058	103	139.447,16505	14.363.058
São Paulo	20	37.250.189	266	140.038,30451	37.250.189
Paraná	21	90.69.349	65	139.528,44615	9.069.349
Santa Catarina	22	5.807.978	42	138.285,19048	5.807.978
Rio Grande do Sul	23	9.304.789	67	138.877,44776	9.304.789
Mato Grosso do Sul	24	2.100.326	16	140.021,73333	2.100.326
Mato Grosso	25	2.621.190	19	137.957,36842	2.621.190
Goiás	26	5.545.319	40	138.632,97500	5.545.319
Distrito Federal	27	2.418.135	18	134.340,83333	2.418.135
Total		166.298.359	1.200		166.298.359

A Tabela 5.3 mostra os pesos atribuídos a cada estrato (W_h), como definido em (2.51), que variam entre 104.411,3 e 141.135,6, além de mostrar que a soma desses pesos atribuídos aos respondentes ($n_h \times W_h$) soma o total da população-alvo. Nota-se também que o produto entre o tamanho da amostra em cada estrato e W_h é igual ao tamanho da população-alvo em cada estrato.

5.3.1 Ajuste de seleção

O ajuste de seleção dos respondentes da pesquisa, como descrito na Subseção 4.2.2, necessita, além da probabilidade de selecionar um número habilitado em cada UF (disponibilizada na base por meio da variável nomeada PS), da quantidade de linhas telefônicas que cada respondente tinha acesso e quantas pessoas compartilhavam cada uma dessas linhas na época da pesquisa.

Ao longo da pesquisa, o DataSenado coletou esses dados e os disponibilizou por meio das variáveis nomeadas na base de dados como V01, V04_A, V04_B, ..., V04_K. Pela descrição das variáveis, foi assumido que cada respondente teria acesso, no máximo, a mais 10 linhas telefônicas, além da utilizada para realizar a pesquisa, entretanto, nenhum respondente da pesquisa tinha acesso a mais de 8 linhas telefônicas na época da pesquisa.

As variáveis V01, V04_A, V04_B, ..., V04_K correspondem à δ_{ij}, definido na Equação (4.3), que representa a quantidade de pessoas que tinham acesso a cada linha telefônica j pertencente ao respondente i. Assim, o ajuste de seleção é obtido pelo produto da probabilidade de selecionar um número habilitado em cada UF pela soma do inverso dessas variáveis. Por exemplo, a probabilidade de selecionar um número habilitado no Paraná (estrato 12) é $0,00995038\%$ e um respondente i selecionado nessa mesma UF possui acesso a mais 1 linha telefônica, além da utilizada para responder a pesquisa (totalizando 2 linhas telefônicas), e que ele divida essas duas linhas com mais 1 pessoa. Então, a probabilidade de selecionar esse respondente

é dada, aproximadamente, por

$$f_{12,i} = 0,0000995038 \times \left(\frac{1}{2} + \frac{1}{2}\right) = 0,0000995038$$

e o ajuste de seleção desse respondente é dado por

$$A_{12,i}^{(sel)} = \frac{1}{f_{12,i}} = \frac{1}{0,0000995038} = 10.049,872061.$$

É importante notar que, embora a probabilidade de selecionar um número habilitado é calculada para cada UF, a probabilidade de selecionar cada respondente dentro de cada UF difere.

5.3.2 Ajuste de não-resposta

Para o ajuste de não-resposta é necessário apenas o cálculo da taxa de resposta. Nessa caso, foram disponibilizadas, pelo DataSenado, as taxas de resposta obtidas pela pesquisa em cada região do Brasil e cada tipo de telefonia, obtida pela pesquisa. Esses dados foram divulgados por meio da variável TR, presente na base de dados.

Assim, considerando a taxa de resposta para telefonia móvel na região Sul de 4,88798371%, o ajuste de não-resposta para os respondentes residentes na região Sul que responderam a pesquisa por meio de um telefone móvel é dado por

$$A_{Sul,movel}^{(nr)} = \frac{1}{0,0488798371} = 20,45833.$$

Nesse caso, quando respondentes que residiam na mesma região e utilizaram o mesmo tipo de telefonia para responder a pesquisa, recebem o mesmo ajuste de não-resposta.

Como o ajuste de não-resposta é aplicado em conjunto com o ajuste de seleção, considerando que o respondente i dos dois exemplos (ajuste de seleção e ajuste de

não-resposta) seja o mesmo, isto é, um respondente que residia no Paraná (região Sul, na época da pesquisa, que possuía acesso à duas linhas telefônicas e compartilhava cada uma delas com outra pessoa e respondeu a pesquisa utilizando um telefone móvel. O peso associado a ele, segundo a Equação (4.4), é dado por

$$w^*_{Sul,movel,12,i} = 10.049,872061 \times 20,45833 = 205.603,63258.$$

Como a aplicação dos pesos se tornou mais complexa, não há garantia de que a soma dos pesos resulte no total da população-alvo. Logo, é necessário calibrar os pesos amostrais para garantir que isso aconteça.

5.3.3 Calibração dos pesos pelo método *raking*

A calibração dos pesos amostrais foi feita utilizando como variáveis de calibração: o sexo, a faixa etária, a cor ou raça e a escolaridade da população-alvo, além de considerar os totais dessas variáveis para cada região do Brasil. Nesse procedimento, foi considerada apenas a população com 16 anos ou mais e com cor ou raça declarada, assim como a população-alvo da pesquisa. A programação utilizada para realizar a calibração dos pesos está detalhada no Apêndice A.

As diferenças entre os pesos obtidos neste trabalho e os pesos gerados e divulgados pelo DataSenado apresentaram uma média muito próxima a 0, assim como os quartis de sua distribuição, como mostra a Tabela 5.4.

Tabela 5.4: Medidas descritivas das diferenças entre os pesos gerados neste trabalho e os pesos obtidos pelo DataSenado.

Média	$1^{\underline{o}}$ quartil	Mediana	$3^{\underline{o}}$ quartil
$0,0000000000175$	$-0,0133526$	$0,0067681$	$0,0225067$

Os resultados apresentados na Tabela 5.4 indicam que foi possível reproduzir

resultados bem semelhantes aos obtidos pelo DataSenado. As diferenças encontradas podem ter sido causadas por divergências nos critérios de parada do algoritmo, já que os resultados foram gerados em *softwares* diferentes. O DataSenado utilizou o *software* R para a calibração.

Entretanto, antes de aplicar o *raking* é necessário que os pesos amostrais estejam na mesma escala da população-alvo, já que a amostra é sorteada considerando linhas telefônicas. Assim, é necessário reescalonar os pesos aplicados aos respondentes de forma que eles somem o total da população-alvo da pesquisa. Essa transformação, considerando o peso w_i, é feita da seguinte forma:

$$w_{i(reesc)} = N \times \frac{w_i}{\sum_{i=1}^{1.200} w_i}, \quad \forall i. \tag{5.1}$$

Considerando o respondente i do estado do Paraná, já citado anteriormente, cujo o ajuste de seleção aplicado a ele foi de $10.049,872061$ e o ajuste de seleção e não-resposta foi de $205.603,63258$, considerando a população-alvo e a Equação (5.1), os ajustes aplicados ao respondente i passam a ser os seguintes:

$$A^{(sel)}_{12,i(reesc)} = 166.298.359 \times \frac{A^{(sel)}_{12,i}}{\sum_{h=1}^{27}\sum_{i=1}^{n_h} A^{(sel)}_{h,i}}$$

$$= 166.298.359 \times \frac{10.049,872061}{12.417.943,87} = 134.585,6649$$

e

$$w^*_{Sul,movel,12,i(reesc)} = 166.298.359 \times \frac{w^*_{Sul,movel,12,i}}{\sum_{h=1}^{27}\sum_{i=1}^{n_h} w^*_{Sul,movel,h,i}}$$

$$= 166.298.359 \times \frac{205.603,63258}{352.817.243} = 96.910,07846.$$

Sendo, no exemplo em questão, $A^{(sel)}_{12,i(reesc)}$ e $w^*_{Sul,movel,12,i(reesc)}$ os pesos $A^{(sel)}_{12,i}$ e

$w^*_{Sul,movel,12,i}$ reescalonados para o tamanho da população-alvo para o respondente i, respectivamente. Por fim, no geral, o peso utilizado para aplicar calibração foi o peso $w^*_{Reg,tipo,h,i(reesc)}$, uma vez que ele foi ajustado para corresponder à quantidade de pessoas.

5.3.4 Análise dos pesos

A soma total dos pesos amostrais pode variar dependendo dos ajustes aplicados em cada etapa de ponderação. Analisando o ajuste de não-resposta individualmente, os pesos amostrais dos respondentes somaram $35.412,92$. Esse valor representa a quantidade de linhas telefônicas utilizadas durante a pesquisa para conseguir as 1.200 entrevistas. Caso o ajuste de não-resposta fosse ser utilizado individualmente, esse valor também deveria ser reescalonado.

A Tabela 5.5 mostra as somas dos pesos aplicados aos respondentes da pesquisa Coronavírus, conforme os passos especificados na Seção 4.3.

Tabela 5.5: Soma dos pesos amostrais aplicados.

Ponderação	Soma total
Sem peso	1.200,00
Peso simples (AAS sem reposição)	166.298.359,00
Peso simples (AAE com alocação proporcional)	166.298.359,00
Ajuste de seleção	12.417.943,87
Ajuste de seleção (reescalonado)	166.298.359,00
Ajuste de seleção e não-resposta	352.817.243,00
Ajuste de seleção e não-resposta (reescalonado)	166.298.359,00
Ajuste de seleção e não-resposta calibrados	166.298.359,00
Ajuste de seleção calibrado	166.298.359,00

A Tabela 5.5 mostra como cada peso apresenta diferentes totais. Quando não há a aplicação de pesos amostrais, o total obtido é o total amostral. Quando a amostragem é aleatória simples ou estratificada com alocação proporcional ou quando há calibração,

os pesos somam o total da população-alvo. Entretanto, quando há aplicação de pesos mais complexos, a soma difere bastante da tamanho da população-alvo.

A soma dos pesos considerando o ajuste de seleção e a soma dos pesos considerando o ajuste de seleção e o ajuste de não-resposta (Tabela 5.5) diferem do total da população devido à pesquisa ser feita com base na quantidade de linhas telefônicas e não na quantidade de pessoas. Como a quantidade de linhas telefônicas disponíveis é maior que a população total, os pesos aplicados para corrigir não-resposta e probabilidade de seleção acabam somando um valor maior que o total da população-alvo.

5.4 Resultados

Após a aplicação de cada peso amostral, foram calculados os intervalos de confiança para as estimativas do total e percentual de pessoas (entre as que solicitaram) que receberam a primeira parcela do Auxílio Emergencial. Para o total de pessoas que receberam a primeira parcela do benefício, a Figura 5.1 mostra os intervalos obtidos para cada um das estimativas geradas em nos passos descritos na Seção 4.3. O plano amostral descrito, na Figura 5.1, como "Amostragem Estratificada" refere-se ao adotado na realização da pesquisa Coronavírus, isto é, uma amostragem estratificada com alocação proporcional por Unidade da Federação (DataSenado, 2020b).

As linhas verticais presentes da Figura 5.1 indicam os valores do parâmetro populacional. A linha contínua indica a quantidade de pessoas, com 16 anos ou mais, que receberam a primeira parcela do Auxílio Emergencial até 21 de maio de 2020, isto é, aproximadamente 52,3 milhões, segundo a Caixa (2020b). A linha tracejada indica

a quantidade de pessoas que receberam a primeira parcela do Auxílio Emergencial até o fim de maio de 2020, isto é, aproximadamente 58 milhões, segundo o Portal da Transparência (2020), desconsiderando os pagamentos retidos ou devolvidos. Este último parâmetro foi considerado apenas para verificar se os intervalos de confiança gerados por meio da metodologia adotada pelo DataSenado são capazes de capturar ambos os valores. Apesar disso, o valor mais adequado para o parâmetro populacional é o que considera a data mais próxima à data de término de coleta de dados da pesquisa – 20 de maio de 2020, segundo o DataSenado (2020a) – ou seja, o parâmetro dado pela Caixa (2020b).

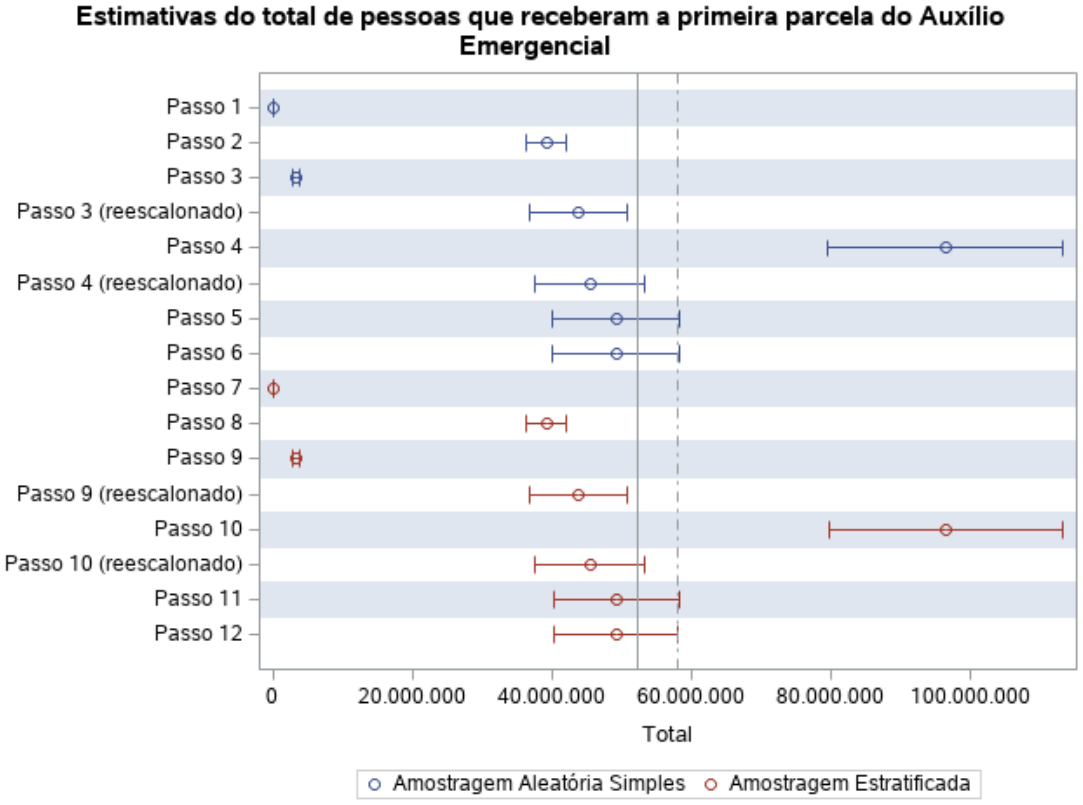

Figura 5.1: Estimativas obtidas em cada um dos passos descritos na Seção 4.3.

Linha contínua: ≈**52.300.000** (Caixa, 2020b).
Linha tracejada: 58.023.204 (Portal da Transparência, 2020).

Analisando a Figura 5.1, é possível notar que não houve diferença entre os planos amostrais adotados, indicando que os estratos (no caso, as Unidades da Federação) podem não ter influenciado no fato dos indivíduos terem recebido ou não a primeira parcela do Auxílio Emergencial. Nota-se também que a aplicação do método *raking* nos métodos de ponderação adotados pelo DataSenado (passos 5, 6, 11 e 12) foi capaz de gerar intervalos de confiança que capturaram ambos os parâmetros. Além disso, quando não houve a aplicação do ajuste de não-resposta (passos 6 e 12), os intervalos se mostraram bem semelhantes ao intervalos gerados com a aplicação do ajuste de não-resposta (passos 5 e 11), embora as estimativas sejam levemente menores. Com isso, pode-se afirmar que a aplicação do ajuste de não-resposta não teve um impacto significativo nas estimativas quando o método *raking* foi aplicado. Uma das possíveis hipóteses que podem explicar esse "não impacto" está relacionada com a fonte do viés de não-resposta, ou seja, pode ser que o fato do indivíduo atender o telefone e responder a pesquisa ou não, não esteja relacionado com o região do Brasil onde ele mora (já que o ajuste de não-resposta é aplicado de acordo com a região onde os respondentes moram).

Ainda em relação à Figura 5.1, nota-se que, quando há o reescalonamento dos pesos de ajuste de seleção e não-resposta para o tamanho da população-alvo (passos 4 e 10), os intervalos obtidos se aproximam dos valores reais e até mesmo capturam o parâmetro principal. No entanto, sem esse reescalonamento, esses intervalos são maiores e superestimam os valores reais. Já quando os pesos amostrais foram desconsiderados (passos 1 e 7) e quando foi aplicado apenas o ajuste de seleção (passos

3 e 9), em ambos os planos amostrais (AAS_s e AAE), os intervalos obtidos foram mais curtos e se distanciaram muito dos parâmetros. Em relação às estimativas obtidas sem a aplicação de pesos amostrais (passos 1 e 7), não se pode dizer que elas estejam incorretas, já que elas se referem aos totais na amostra e não na população-alvo. Por fim, vale ressaltar que os intervalos muito próximos aos parâmetros, mas que não os capturaram, não estão necessariamente incorretos, uma vez que a confiança dos intervalos é de 95%.

Analisando numericamente, a Tabela 5.6 traz as estimativas e as margens de erro obtidas em cada procedimento de ponderação. As estimativas pontuais nos passos 1 e 7 foram obtidas através da contagem da quantidade de respondentes que afirmou ter recebido a primeira parcela do Auxílio Emergencial na amostra ($\widehat{T} = 283$). Agora usando como exemplo o passo 2 (AAS_s), a estimativa pontual foi obtida por meio da Equação (2.2), sendo assim,

$$\widehat{T} = N \times \overline{y} = 166.298.359 \times \frac{283}{1200} = 166.298.359 \times 0,23583 = 39.218.696.$$

Nesse caso, $\overline{y}$ é equivalente à p, definido em (2.16), já que o interesse é saber se o indivíduo recebeu ou não a primeira parcela do benefício (no geral, 1-"Sim"ou 0-"Não"). Para o passo 8 (AAE), a estimativa pontual foi obtida por meio da Equação (2.23), considerando como N_h os valores apresentados na Tabela 5.3. As estimativas pontuais nos demais passos foram obtidas através da aplicação dos pesos amostrais e as variâncias foram calculadas por meio da linearização de Taylor.

A Tabela 5.6 confirma a discussão a respeito dos resultados apresentados na Figura 5.1. Também é possível notar que o impacto de não considerar o plano

amostral correto (passos 1 ao 6) foi quase nulo, uma vez que as estimativas geradas foram muito parecidas. No mais, apesar do efeito do plano amostral ser quase nulo, as margens de erro obtidas considerando o plano amostral adotado na pesquisa Coronavírus (indicado por "Amostragem Estratificada") foram levemente menores, mas essa diferença é muito pequena, a ponto de poder ser desconsiderada. Os resultados presentes na Tabela 5.6 foram arredondados.

Tabela 5.6: Estimativas obtidas em cada um dos passos descritos na Seção 4.3.

Ponderação	Estimativa	Limite inferior	Limite superior	Margem de erro
Passo 1*	283	262	304	21
Passo 2*	39.218.696	36.316.784	42.120.609	2.901.913
Passo 3*	3.264.640	2.738.294	3.790.986	526.346
Passo 3* (reescalonado)	43.719.331	36.670.622	50.768.041	7.048.709
Passo 4*	96.396.448	79.576.128	113.216.768	16.820.320
Passo 4* (reescalonado)	45.435.906	37.507.746	53.364.066	7.928.160
Passo 5*	49.203.477	40.051.371	58.355.583	9.152.106
Passo 6*	49.167.855	40.027.164	58.308.546	9.140.691
Passo 7**	283	262	304	21
Passo 8**	39.236.820	36.377.610	42.096.031	2.859.211
Passo 9**	3.264.640	2.742.723	3.786.557	521.917
Passo 9** (reescalonado)	43.719.331	36.729.934	50.708.728	6.989.397
Passo 10**	96.396.448	79.616.267	113.176.629	16.780.181
Passo 10** (reescalonado)	45.435.906	37.526.665	53.345.147	7.909.241
Passo 11**	49.203.477	40.250.840	58.156.114	8.952.637
Passo 12**	49.167.855	40.224.145	58.111.565	8.943.710

* Amostragem Aleatória Simples.
** Amostragem Estratificada com alocação proporcional por Unidade da Federação. Parâmetros reais: ≈**52.300.000** (Caixa, 2020b) e 58.023.204 (Portal da Transparência, 2020).

Agora analisando as estimativas percentuais, a Figura 5.2 mostra os intervalos obtidos para o percentual de pessoas que receberam a primeira parcela do Auxílio Emergencial em relação à quantidade de pedidos processados até o dia 21 de maio de 2020, isto é, 101,2 milhões, segundo a Caixa (2020b). O plano amostral descrito, na

Figura 5.2, como “Amostragem Estratificada” refere-se ao plano amostral adotado na realização da pesquisa Coronavírus, assim como no caso dos totais.

A linha contínua presente na Tabela 5.2 indica o percentual de pessoas, com 16 anos ou mais, que receberam a primeira parcela do Auxílio Emergencial até o dia 21 de maio de 2020, isto é, aproximadamente 51,68% (52.300.000/101.200.000 = 0,5168). A linha tracejada indica o percentual de pessoas que receberam a primeira parcela do benefício até o fim de maio de 2020, mas considerando como denominador a quantidade de pedidos processados até o dia 21 de maio de 2020 pela Caixa (2020b), isto é, aproximadamente 57,34% (58.023.204/101.200.000 = 0,5734), uma vez que não foi possível ter acesso a quantidade de pedidos processados até o fim de maio de 2020. No entanto, ao considerar esse denominador, esse último percentual acabou sendo superestimado, já que o esperado é que mais pessoas tenham solicitado o benefício até o fim de maio de 2020.

Assim como no caso das estimativas para os totais, o último parâmetro descrito foi considerado apenas para verificar se os intervalos de confiança gerados por meio da metodologia adotada pelo DataSenado são capazes de capturar ambos os valores. Entretanto, o valor mais adequado para o parâmetro populacional é o que considera a data mais próxima à data de término de coleta de dados da pesquisa, ou seja, o percentual calculado com os dados fornecidos pela Caixa (2020b).

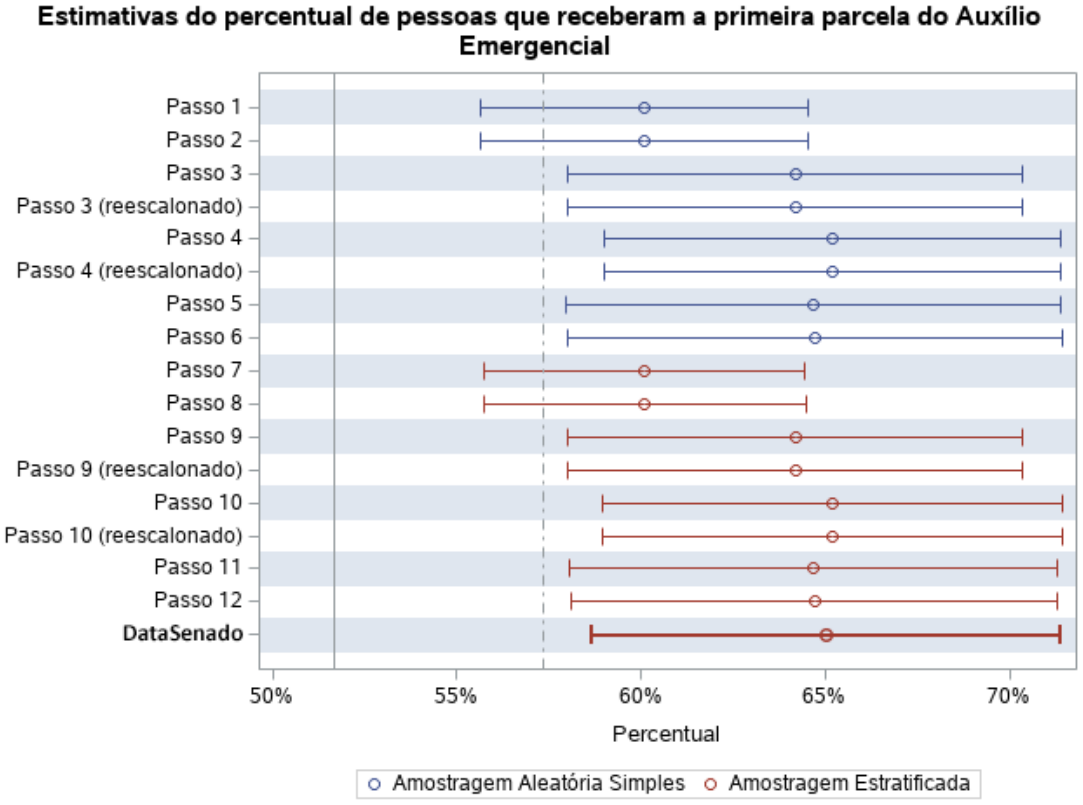

Figura 5.2: Estimativas percentuais obtidas em cada um dos passos descritos na Seção 4.3 e estimativas divulgadas pelo DataSenado (2020a).

Linha contínua: **51,68%** (Caixa, 2020b).
Linha tracejada: $57,34\%$ (Portal da Transparência (2020) e Caixa (2020b)).

A Figura 5.2 mostra que, para o percentual, os intervalos gerados foram bastante semelhantes e também não houve diferença entre os planos amostrais adotados, indicando também que os estratos (Unidades da Federação) podem não ter influenciado no fato dos indivíduos terem solicitado e recebido ou não a primeira parcela do Auxílio Emergencial. É possível perceber também que reescalonamento dos pesos não teve influência nos intervalos de confiança, diferente do que acontece com as estimativas para os totais. Além disso, a Figura 5.2 mostra que apenas quando não houve a aplicação de pesos ou quando houve a aplicação dos pesos simples (N/n e

N_h/n_h - passos 1, 2, 7 e 8) um dos parâmetros populacionais foi capturado pelos intervalos de confiança. Entretanto, isso não indica que os métodos mais simples sejam os mais eficazes em estimar os parâmetros populacionais. Neste caso, o fato dos métodos mais simples terem capturado um dos parâmetros, mesmo o parâmetro menos preciso, pode ter sido apenas fruto do acaso.

Um aspecto importante, no caso das estimativas percentuais, é a possibilidade de verificar a eficácia da metodologia adotada pelo DataSenado, por meio das estimativas que foram divulgadas oficialmente pelo instituto (ver DataSenado (2020a)). Assim, pela Figura 5.2, nota-se que a metodologia adotada pelo DataSenado não conseguiu capturar, de forma intervalar, o percentual populacional de pessoas que receberam a primeira parcela do Auxílio Emergencial, mas as estimativas ficaram bem próximas das demais. Isso não indica que a metodologia esteja incorreta, uma vez que a confiança do intervalo é de 95%, segundo o (DataSenado, 2020b). Em relação aos intervalos gerados para este trabalho, ao alterar o nível de confiança para 99%, todos eles passam a capturar o parâmetro representado pela linha tracejada (57,34%), logo, dependendo do nível de confiança, os resultados podem ser diferentes.

Analisando numericamente, a Tabela 5.7 mostra as estimativas percentuais e as margens de erro obtidas em cada procedimento de ponderação. As estimativas pontuais, nos passos 1 e 7, foram calculadas pela razão entre a quantidade de respondentes que afirmaram ter recebido a primeira parcela do Auxílio Emergencial (283) e a quantidade de respondentes que afirmaram ter solicitado o benefício (471), ou seja, $283/471 = 0,601$. Usando como exemplo o passo 2 (AAS_s), a estimativa

pontual foi feita com base na Equação (2.16) e o cálculo se dá da mesma forma, uma vez que o n não é considerado como denominador, e sim a quantidade de respondentes que afirmaram ter solicitado o auxílio. No passo 8 (AAE), a estimativa pontual foi calculada por meio da Equação (2.40) e, para o cálculo do termo p_h, o denominador considerado, ao invés de n_h, foi a quantidade de respondentes que afirmaram ter solicitado o auxílio dentro de cada estrato (UF). As estimativas pontuais nos demais passos foram calculadas utilizando os pesos amostrais obtidos e a variância foi calculada por meio da linearização de Taylor.

Tabela 5.7: Estimativas percentuais obtidas em cada um dos passos descritos na Seção 4.3 e estimativas divulgadas pelo DataSenado (2020a).

Ponderação	Estimativa	Limite inferior	Limite superior	Margem de erro
Passo 1*	60,1%	55,6%	64,5%	4,4%
Passo 2*	60,1%	55,6%	64,5%	4,4%
Passo 3*	64,2%	58,0%	70,3%	6,2%
Passo 3* (reescalonado)	64,2%	58,0%	70,3%	6,2%
Passo 4*	65,2%	59,0%	71,4%	6,2%
Passo 4* (reescalonado)	65,2%	59,0%	71,4%	6,2%
Passo 5*	64,7%	57,9%	71,4%	6,7%
Passo 6*	64,7%	58,0%	71,4%	6,7%
Passo 7**	60,1%	55,7%	64,4%	4,4%
Passo 8**	60,1%	55,7%	64,5%	4,4%
Passo 9**	64,2%	58,0%	70,3%	6,2%
Passo 9** (reescalonado)	64,2%	58,0%	70,3%	6,2%
Passo 10**	65,2%	58,9%	71,4%	6,2%
Passo 10** (reescalonado)	65,2%	58,9%	71,4%	6,2%
Passo 11**	64,7%	58,1%	71,3%	6,6%
Passo 12**	64,7%	58,1%	71,3%	6,6%
DataSenado	**65,0%**	**58,6%**	**71,4%**	**6,4%**

* Amostragem Aleatória Simples
** Amostragem Estratificada com alocação proporcional por Unidade da Federação. Parâmetros reais: **51,68%** (Caixa, 2020b) e $57,34\%$ (Portal da Transparência (2020) e Caixa (2020b)).

A partir da Tabela 5.7, é possível confirmar os resultados discutidos anteriormente,

referentes à Figura 5.2, e notar que impacto de não considerar o plano amostral real (passos 1 ao 6) nas estimativas é muito baixo, assim como no caso dos totais. Por fim, apesar da maioria dos intervalos não capturarem nenhum dos parâmetros considerados, principalmente o parâmetro principal (51,68%), as estimativas são bastante parecidas. Além disso, o fato de ter considerado a quantidade de solicitações do Auxílio Emergencial processadas até o dia 21 de maio de 2020 pode ter influenciado na geração dos parâmetros, uma vez que, devido a limitações de sistema ou outros fatores desconhecidos, esse número possivelmente é menor do que o real número de pessoas que solicitaram o benefício. Assim, uma possível causa para o comportamento apresentado pelos intervalos de confiança gerados para os percentuais pode ser o fato do percentual de pessoas, na amostra, que afirmaram ter solicitado o Auxílio possa ser menor do que o percentual de pessoas que solicitaram o benefício entre os indivíduos da população-alvo, fazendo com as estimativas percentuais fossem maiores.

Por fim, com o objetivo de analisar a semelhança entre as estimativas geradas para os dois planos amostrais adotados, foi calculado o efeito do planejamento (EPA) ou, do inglês, *design effect* (DEFF), para todos os métodos de ponderação e planos amostrais e os resultados se encontram na Tabela 5.8. O efeito do planejamento é utilizado, segundo Bolfarine e Bussab (2005), para comparar um plano amostral qualquer com um plano amostral padrão por meio das variâncias. Neste trabalho, o plano amostral padrão considera uma AAS e, por esse motivo, para os dois primeiros métodos de ponderação apresentados na Tabela 5.8, o efeito do planejamento na amostragem aleatória simples é igual a 1.

Tabela 5.8: Efeito do planejamento em cada método de ponderação e plano amostral.

Ponderação	Amostragem Aleatória Simples	Amostragem Estratificada
Sem peso	1,0000	0,9660
Peso simples	1,0000	0,9687
Ajuste de seleção	2,0141	2,0108
Ajustes de seleção e não-resposta	2,0469	2,0897
Ajustes de seleção e não-resposta calibrados	2,4192	2,3330
Ajuste de seleção calibrado	2,4101	2,3246

Os resultados presentes na Tabela 5.8 mostram que os efeitos do planejamento são bem próximos em ambos os planos amostrais, indicando que os estratos da amostra não exerceram quase nenhum efeito na variância dos estimadores, o que explica a semelhança entre os intervalos de confiança gerados para os dois planos amostrais. Por fim, por "Amostragem Estratificada" entende-se o plano amostral adotado na pesquisa Coronavírus.

Capítulo 6

CONCLUSÕES

O objetivo do trabalho foi verificar o efeito de diferentes métodos de ponderação para corrigir problemas existentes em pesquisas por telefone e como esses métodos impactam nos intervalos de confiança das estimativas. Para isso, houve a não aplicação de pesos amostrais, a aplicação de pesos amostrais considerando os ajustes simples, o ajuste de seleção, o ajuste de não-resposta e a aplicação do método *raking*. Além disso, houve a alteração no tipo de plano amostral, considerando a amostragem aleatória simples, ao invés da amostragem estratificada, em parte dos ajustes.

Para as estimativas obtidas para os totais, foi visto que a junção do ajuste de seleção, do ajuste de não-resposta e do método *raking* conseguiu capturar, por meio do intervalo de confiança, os parâmetros reais. Entretanto, isso também aconteceu ao desconsiderar o ajuste de não-resposta. Já para as estimativas obtidas para os percentuais, observou-se um comportamento diferente das estimativas para os totais. Nesse caso, a aplicação dos ajustes de seleção e não-resposta, assim como a aplicação do método *raking*, não obtiveram sucesso ao estimar os parâmetros reais. Apesar disso, os intervalos de confiança não precisam, obrigatoriamente, capturar o parâmetro de interesse, uma vez que seu nível de confiança é de 95%, tanto no caso

dos totais quanto dos percentuais. No entanto, apenas o fato dos intervalos estarem próximos do parâmetro pode indicar que o método é preciso.

Os resultados também mostraram que ignorar o plano amostral real da pesquisa não alterou, de forma significativa, nas estimativas obtidas. Isso pode indicar que os estratos considerados na amostra, ou seja, as Unidades da Federação, não tenham influenciado no fato dos indivíduos terem recebido ou não o pagamento da primeira parcela do Auxílio Emergencial. Além disso, os resultados mostraram que, quando houve a calibração dos pesos amostrais, o ajuste de não-resposta teve um impacto quase nulo nas estimativas, mas isso não significa que o ajuste de não-resposta não deva ser aplicado. Como o DataSenado aplica o ajuste em questão considerando as linhas telefônicas das quais não obtiveram retorno de acordo com a região que elas pertencem, esse resultado pode indicar que o viés de não-resposta não provém da região ou talvez outro fator cause esse viés e mereça ser investigado. Mas, dada as limitações da pesquisa por telefone, é compreensível a forma com que esse método de correção foi implementado.

Com base nos resultados apresentados, conclui-se que a junção dos métodos de ponderação adotados pelo DataSenado foi eficaz em capturar os parâmetros reais quando se tratava dos totais, mas isso não aconteceu quando se tratava dos percentuais. Entretanto, como os percentuais foram calculados com base na quantidade de solicitações do Auxílio Emergencial processadas até o dia 21 de maio de 2020[1], é possível que eles não sejam precisos. Além disso, durante a realização da pesquisa, é

[1]Data mais próxima à data do término da coleta de dados da pesquisa Coronavírus, isto é, 20 de maio de 2020, segundo o DataSenado (2020a)

possível que mais solicitações do benefício tenham sido realizadas, mas não processadas pelo governo. Com isso, não foi possível ter acesso ao número real de pessoas que haviam solicitado o benefício até o término da pesquisa, isto é, até o dia 20 de maio de 2020, ou até o fim de maio de 2020.

Por fim, este trabalho mostrou como a escolha dos métodos de ponderação pode influenciar os intervalos de confiança e o resultado final, podendo levar o pesquisador à conclusões incorretas, ao enviesar as estimativas, ou beneficiá-lo, ajustando as estimativas de forma que os intervalos de confiança capturem o parâmetro de interesse.

Referências Bibliográficas

AAPOR (2016). Standard definitions: Final dispositions of case codes and outcome rates for surveys. American Association for Public Opinion Research.

Anatel (2020a). Acessos. Agência Nacional de Telecomunicações. Disponível em: URL `https://www.anatel.gov.br/paineis/acessos`. Acesso em: 21 de set. de 2020.

Anatel (2020b). Plano de Numeração Brasileiro. Agência Nacional de Telecomunicações. Disponível em: URL `https://www.anatel.gov.br/setorregulado/plano-de-numeracao-brasileiro#:~:text=O%20n%C3%BAmero%20do%20assinante%20tem,Especializado%20(telefonia%20m%C3%B3vel%20r%C3%A1dio)..` Acesso em: 13 de set. de 2020.

Anatel (2020c). SAPN - Administração do Plano de Numeração. Agência Nacional de Telecomunicações. Disponível em: URL `https://sistemas.anatel.gov.br/sapn/`. Acesso em: 18 de ago. de 2020.

Battaglia, M. P., Izrael, D., Hoaglin, D. C., & Frankel, M. R. (2009). Practical considerations in raking survey data. *Survey Practice*, 2:5.

Bolfarine, H. & Bussab, W. O. (2005). *Elementos de Amostragem*. ABE - Projeto Fisher.

Caixa (2020a). Auxílio Emergencial. Caixa. Disponível em: URL `http://www.caixa.gov.br/auxilio/PAGINAS/DEFAULT2.ASPX`. Acesso em: 28 de ago. de 2020.

Caixa (2020b). Download de Arquivos. A Caixa - Demonstrativo financeiro. Disponível em: URL `https://www.caixa.gov.br/site/paginas/downloads.aspx`. Acesso em: 28 de nov. de 2020.

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

Colombotos, J. (1969). Personal versus telephone interviews: Effect on responses. *Public Health Reports (1896-1970)*, 84(9):773–782.

DataSenado (2019). Violência contra a mulher: agressões cometidas por "ex" aumentam quase 3 vezes em 8 anos. Portal Institucional do Senado Federal. Disponível em: URL https://www12.senado.leg.br/institucional/datasenado/materias/enquetes/publicacaodatasenado?id=violencia-contra-a-mulher-agressoes-cometidas-por-2018ex2019-aumentam-quase-3-vezes-em-8-anos-1. Acesso em: 23 de ago. de 2020.

DataSenado (2020a). Brasileiros acreditam que número de contaminados é maior que o noticiado. Portal Institucional do Senado Federal. Disponível em: URL https://www12.senado.leg.br/institucional/datasenado/publicacaodatasenado?id=brasileiros-acreditam-que-numero-de-contaminados-e-maior-que-o-noticiado. Acesso em: 26 de out. de 2020.

DataSenado (2020b). Metodologia Coronavírus (Covid-19). Instituto de Pesquisa DataSenado, Secretaria de Transparência, Senado Federal.

DataSenado (2020c). Nota Técnica - Delineamento amostral. Instituto de Pesquisa DataSenado, Secretaria de Transparência, Senado Federal.

DataSenado (2020d). Para a maioria dos brasileiros, a democracia é a melhor forma de governo. Portal Institucional do Senado Federal. Disponível em: URL https://www12.senado.leg.br/institucional/datasenado/publicacaodatasenado?id=para-a-maioria-dos-brasileiros-a-democracia-e-a-melhor-forma-de-governo. Acesso em: 23 de ago. de 2020.

DataSenado (2020e). Questionário Coronavírus (Covid-19). Instituto de Pesquisa DataSenado, Secretaria de Transparência, Senado Federal.

DataSenado (2020f). Sobre o DataSenado. Portal Institucional do Senado Federal. Disponível em: URL https://www12.senado.leg.br/institucional/datasenado/sobre. Acesso em: 23 de ago. de 2020.

Fricker, R. & Anderson, L. (2015). Raking: An important often overlooked survey analysis tool. *Phalanx*, pages 36–42.

G1 (2020). Auxílio emergencial de R$ 600 revela 46 milhões de brasileiros invisíveis aos olhos do governo. Disponível em: URL https://g1.globo.com/fantastico/noticia/2020/04/26/auxilio-emergencial-de-r-600-revela-42-milhoes-de-brasileiros-invisiveis-aos-olhos-do-governo.ghtml. Acesso em: 27 de set. de 2020.

Groves, R. M. (1989). *Survey Errors and Survey Costs*. John Wiley & Sons.

Groves, R. M., Biemer, P. P., Lyberg, L. E., Massey, J. T., Nicholls, W. L., & Waksberg, J. (2001). *Telephone Survey Methodology*. John Wiley & Sons.

IBGE (2020a). Censo 2010: população do Brasil é de 190.732.694 pessoas. Instituto Brasileiro de Geografia e Estatística. Disponível em: URL https://censo2010.ibge.gov.br/noticias-censo.html?view=noticia&id=3&idnoticia=1766&busca=1&t=censo-2010-populacao-brasil-190-732-694-pessoas. Acesso em: 26 de out. de 2020.

IBGE (2020b). Pesquisa Nacional por Amostra de Domicílios - PNAD COVID19. Instituto Brasileiro de Geografia e Estatística. Disponível em: URL https://www.ibge.gov.br/estatisticas/investigacoes-experimentais/estatisticas-experimentais/27946-divulgacao-semanal-pnadcovid1?t=conceitos-e-metodos&utm_source=covid19&utm_medium=hotsite&utm_campaign=covid_19. Acesso em: 27 de set. de 2020.

IBGE (2020c). Pesquisa Nacional por Amostra de Domicílios Contínua - PNAD Contínua. Instituto Brasileiro de Geografia e Estatística. Disponível em: URL https://www.ibge.gov.br/estatisticas/multidominio/condicoes-de-vida-desigualdade-e-pobreza/9173-pesquisa-nacional-por-amostra-de-domicilios-continua-trimestral.html?edicao=28690&t=o-que-e. Acesso em: 26 de out. de 2020.

IBPAD (2020). O que é e como é feito random-digit dialing (RDD)?. Instituto Brasileiro de Pesquisa e Análise de Dados. Disponível em: URL https://www.ibpad.com.br/blog/o-que-e-random-digit-dialing-rdd/. Acesso em: 20 de set. de 2020.

Izrael, D., Hoaglin, D., & Battaglia, M. (2000). A SAS Macro for Balancing a Weighted Sample.

Lepkowski, J. M., Tucker, C., Brick, J. M., de Leeuw, E., Japec, L., Lavrakas, P. J., Link, M. W., & Sangster, R. L. (2008). *Advances in Telephone Survey Methodology*. John Wiley & Sons.

Lohr, S. L. (1999). *Sampling: Design and Analysis*. Duxbury Press.

National Public Research (2019). Telephone surveys. Disponível em: URL https://nationalpublicresearch.com/services/telephone-surveys/. Acesso em: 20 de set. de 2020.

Portal da Transparência (2020). Auxílio Emergencial. Portal da Transparência. Disponível em: URL http://www.portaltransparencia.gov.br/pagina-interna/603519-download-de-dados-auxilio-emergencial. Acesso em: 05 de nov. de 2020.

Sangster, R. L. (2003). Do current methods used to improve response to telephone surveys reduce nonresponse bias? U.S. Bureau of Labor Statistics. Disponível em: https://www.bls.gov/osmr/research-papers/2003/st030290.htm.

SAS Institute (2020). SAS Help Center: Taylor Series Variance Estimation. The SURVEYMEANS Procedure. Disponível em: URL https://documentation.sas.com/?docsetId=statug&docsetTarget=statug_surveymeans_details06.htm&docsetVersion=15.2&locale=en#statug.surveymeans.variancedetails. Acesso em: 12 de fev. de 2021.

Sincero, S. M. (2012). Telephone survey. **Explorable.com**. Disponível em: URL https://explorable.com/telephone-survey. Acesso em: 23 de ago. de 2020.

Apêndice A

SAS *macro* para *Raking*

Até o término deste trabalho, o *software* SAS não tinha o procedimento *raking* implementado em seus códigos-fonte, apesar de existirem *macros* desenvolvidos para esse fim, como é o caso da *macro* desenvolvida por Izrael et al. (2000). Com isso, foi necessário o desenvolvimento de uma *macro* para calibrar os pesos amostrais aplicando o *raking*. A *macro* utilizada para a calibração dos pesos amostrais segue abaixo.

```
%macro raking(sample=,wsamp=,population=,wpop=,var=,tol=0.0001,int=10);

proc means data=&sample noprint nway;
    class &var;
    var &wsamp;
    output out=&sample._ranking(drop=_type_ _freq_) sum=freq;
run;

data &sample._ranking;set &sample._ranking;
    totp=.; tots=.; weight=1;
run;

data _null_;
    var=compbl("&var");
    var2=compress(var);
    nvar=length(var)-length(var2)+1;
    drop var2;
    call symput('nvar',trim(left(nvar)));
run;
```

```
%put &nvar;
%let dpar=100;
%let int2=1;

%do %while (&dpar>&tol and &int2<=&int);
    %do i=1 %to &nvar;

        %let variable=%scan(&var,&i);
        %put &variable;

        proc means data=&sample._ranking noprint nway;
            class &variable;
            var freq;
            output out=_freq_(drop=_type_ _freq_) sum=tots;
        run;

        proc means data=&population noprint nway;
            class &variable;
            var &wpop;
            output out=_freqp_(drop=_type_ _freq_) sum=totp;
        run;

        proc sort data=&sample._ranking;by &variable;run;
        proc sort data=&population;by &variable;run;

        data &sample._ranking;
            merge &sample._ranking(drop=totp tots) _freq_ _freqp_;
            by &variable;
            w&int2=totp/tots;
            freqold=freq;
            freq=freqold*w&int2;
            dpar=abs(freq-freqold);
            weight=weight*w&int2;
        run;

        proc means data=&sample._ranking noprint;
            var dpar;
            output out=_maxdpar_ max=dpar;
        run;

        data _maxdpar_;set _maxdpar_;
            if dpar<&tol then dpar2=&tol/10;else dpar2=dpar;
            call symput('dpar',trim(left(dpar2)));
        run;

        %put dpar=&dpar;
        %let int2=%eval(&int2+1);
    %end;
```

```
%end;

proc sql;
    select sum(freq) as Total_Sample from &sample._ranking;
    select sum(totp) as Total_Population from _freqp_;
quit;

%mend raking;
```

A *macro* recebe o banco de dados com os totais da amostra (*sample*), o peso a ser calibrado (*wsamp*), o banco de dados com os totais da população-alvo (*population*), o peso aplicado aos indivíduos da população (*wpop*), as variáveis de calibração, que devem estar separadas por um espaço (*var*) e os critérios de parada: a tolerância mínima para a diferença entre o total obtido na iteração atual e na iteração anterior do *raking* (*tol*) e o número máximo de iterações que o *raking* deve fazer (*int*). Para a geração dos resultados, foi considerado $tol = 10$ e $int = 500$ e a *macro* foi aplicada considerando os totais em cada região do Brasil, seguindo os métodos adotados por DataSenado (2020c). Por fim, ela retorna os ajustes necessários para que os pesos somem o total da população-alvo ($A_i^{(cal)}$), como indicado na Equação (3.1).

O DataSenado realizou a calibração dos pesos amostrais por meio do *software* R, utilizando a função *rake*, contida no pacote *survey* (DataSenado, 2020c). Calculando a diferença entre os pesos obtidos no R pelo pacote *survey* e no SAS 9.4 pela *macro* descrita neste Apêndice, tanto o peso considerando o ajuste de seleção e não-resposta quanto o peso considerando apenas o ajuste de seleção apresentaram uma diferença média muito próxima de 0, indicando que os dois algoritmos geram pesos calibrados muito semelhantes.