Delineamento amostral das pesquisas DataSenado

Nota técnica

junho/2020

Instituto de Pesquisa DataSenado

Secretaria de Transparência

SENADO FEDERAL

Introdução

Esta nota técnica apresenta a nova metodologia amostral usada pelo Instituto de Pesquisa DataSenado nas pesquisas de opinião realizadas por meio de telefone, implementada a partir de novembro de 2019. O documento fornece as referências necessárias para a correta análise desses dados, o que implica em considerar o delineamento amostral complexo no cálculo das estimativas, a fim de obter resultados não viesados e margens de erro para cada uma das perguntas feitas nos levantamentos. O rigor científico adotado pelo DataSenado visa romper a abordagem simplista em pesquisas de opinião que, normalmente, reportam sua variabilidade numa única margem de erro, ignorando o delineamento amostral efetivamente levado a efeito e supondo cenários irreais. Ao considerar o delineamento amostral complexo no cálculo das estimativas e das margens de erro individuais, o DataSenado fornece aos seus leitores instrumentos precisos para tomadas de decisões.

Para exemplificar o método, foram utilizados dados da pesquisa sobre Fake News realizada pelo instituto em junho de 2020.

A população-alvo das pesquisas DataSenado é, em geral, composta de cidadãos brasileiros com 16 anos ou mais. Os participantes são selecionados via amostragem estratificada por Unidade da Federação (UF) com alocação proporcional à população-alvo segundo os dados mais recentes da Pesquisa Nacional por Amostra de Domicílios Contínua - PNAD Contínua, do Instituto Brasileiro de Geografia e Estatística - IBGE. Os questionários são do tipo estruturado, com questões objetivas. Na pesquisa sobre Fake News, por exemplo, a amostra foi composta por 1.200 entrevistas, com a seguinte alocação por UF:

Tabela 1 - Tamanho da amostra por UF

UF	Tamanho da amostra
Rondônia	10
Acre	5
Amazonas	21
Roraima	3
Pará	46
Amapá	5
Tocantins	9
Maranhão	37
Piauí	18
Ceará	52
Rio Grande do Norte	20

Tabela 1 - Tamanho da amostra por UF	
Paraíba	23
Pernambuco	55
Alagoas	18
Sergipe	13
Bahia	83
Minas Gerais	123
Espírito Santo	23
Rio de Janeiro	103
São Paulo	266
Paraná	65
Santa Catarina	42
Rio Grande do Sul	67
Mato Grosso do Sul	16
Mato Grosso	19
Goiás	40
Distrito Federal	18
Brasil	1.200

Coleta de dados

A coleta de dados foi feita por meio de entrevistas telefônicas via CATI (*Computer Assisted Telephone Interviewing*). Nesse método, o entrevistador segue roteiro disponibilizado em computador e composto por questionário estruturado, com questões objetivas e orientações para a condução da entrevista. Essa estrutura visa eliminar possíveis vieses, bem como maximizar a aderência dos cidadãos contatados à pesquisa.

Os números de telefone usados nas discagens foram selecionados aleatoriamente, respeitando o delineamento amostral a partir de dados públicos da Anatel sobre os números habilitáveis do país. A quantidade de números fixos e móveis sorteados na amostra foi estabelecida de forma a garantir que, por UF, a probabilidade de sorteio de qualquer número fosse a mesma, independente de se tratar de telefone fixo ou móvel.

Para compor a amostra, foram realizadas ligações telefônicas para todo o país. Atendido o telefone, e após verificar se o(a) entrevistado(a) pertencia à população-alvo, o entrevistador solicitava autorização para realizar a pesquisa. As entrevistas foram realizadas até que os 1.200 questionários estivessem preenchidos, respeitando a alocação por UF do plano amostral.

Foram auditadas 25% das entrevistas, verificando itens como cordialidade, leitura fluente, marcação correta das respostas, não direcionamento das respostas, dentre outros aspectos de qualidade e imparcialidade durante a aplicação da pesquisa.

Ponderação

Tratando-se de delineamento amostral complexo, faz-se necessário ponderar os dados considerando três aspectos: probabilidades de seleção dos entrevistados, taxas de respostas e calibração da distribuição demográfica da amostra em relação à população-alvo. Busca-se, assim, obter estimativas não viesadas para a população-alvo da pesquisa, com cálculo ajustado dos erros padrões.

A probabilidade de seleção dos entrevistados foi calculada com base na quantidade de linhas telefônicas a que cada indivíduo tinha acesso, na quantidade de pessoas que compartilhavam cada uma dessas linhas e no total de linhas habilitadas usadas na pesquisa em relação ao total de linhas habilitadas no Brasil por UF, segundo as estatísticas mais recentes da Anatel.

A estimativa da taxa de resposta por região e tipo de telefonia foi obtida de forma equivalente à *Response Rate 1* (RR1) da American Association for Public Opinion Research (AAPOR, 2016, p. 61), a partir de metadados das discagens telefônicas coletados no decorrer da pesquisa.

Por fim, os pesos foram ajustados para refletirem a proporção da população por região, segundo características demográficas. Na pesquisa sobre Fake News essas variáveis foram: sexo, idade, escolaridade e raça/cor. Para tanto, foi utilizado o método *rake*, considerando a distribuição estimada da população brasileira segundo dados da Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD Contínua) do 1º trimestre de 2020.

A seguir estão detalhadas etapas do processo de ponderação da amostra.

Probabilidade de seleção

A alocação da amostra por estrato é proporcional à população-alvo, o que, em situações mais simples, faria com que a probabilidade f_h de seleção fosse a mesma para todos os elementos, independente da UF h de residência ($f_h = n_h/N_h = n/N$, com $h = 1, \dots, 27$ representando as UF do Brasil). Mas, como a pesquisa foi feita por meio de ligações telefônicas, a probabilidade de seleção dos indivíduos em um mesmo estrato h é desigual, influenciada por dois fatores, (1) a quantidade de telefones fixos e móveis a que cada um

tem acesso e (2) a quantidade de pessoas que compartilham esses acessos. Considerando essa característica do método de coleta, a fórmula abaixo foi usada para estimar f_{hj}, a probabilidade de seleção por participante j do estrato h da pesquisa, de acordo com sua UF de residência:

$$f_{hj} = \pi_h \sum_{i=1}^{N_{dj}} \frac{1}{\delta_{ij}}$$

onde

- π_h é a probabilidade de seleção de um número de telefone qualquer da UF h, dada por $\pi_h = \frac{Total\ de\ números\ habilitados\ usados\ na\ pesquisa\ para\ a\ UF}{Total\ de\ números\ habilitados\ na\ UF,\ \ segundo\ Anatel\ (abril/2020)}$, com $h = 1, \dots, 27$. As listas telefônicas por UF foram geradas de tal forma que π_h independesse do tipo de telefonia (fixa ou móvel).
- j representa o j-ésimo indivíduo selecionado na amostra pertencente ao estrato h,
- $i = 1, \dots, N_{dj}$ representa cada linha de telefone fixo ou móvel a que o j-ésimo indivíduo do estrato h tem acesso, num total de N_{dj} linhas, e
- δ_{ij} é a quantidade de pessoas que podem atender ligações na linha i do j-ésimo indivíduo do estrato h.

O peso associado à probabilidade de seleção é dado pelo inverso de f_{hj}:

$$w_{sel,hj} = \frac{1}{f_{hj}}.$$

Os valores de δ_{ij} para cada $i = 1, \dots, N_{dj}$ são identificados pelas variáveis V01, V04A, ..., V04K do arquivo de dados da pesquisa.

Taxa de resposta

No cálculo da probabilidade de seleção são considerados todos os telefones habilitados da lista usada na amostra. Ocorre que várias ligações resultam em recusa a participar da pesquisa, números ocupados, números não atendidos, dentre outras ocorrências que impedem que uma ligação para número habilitado resulte em pesquisa concluída. Para atenuar o viés devido aos diferentes padrões de não resposta na amostra,

faz-se necessário considerar esse fator no cálculo do peso amostral, por meio da Taxa de Resposta. O preço a ser pago por essa correção será o incremento no erro padrão das estimativas ponderadas. A taxa de resposta da pesquisa representa o número de entrevistas concluídas em relação ao número de ligações realizadas para números habilitados. Adotando postura conservadora, considera-se números habilitados todas as linhas telefônicas usadas na pesquisa que se mostraram elegíveis ou de elegibilidade desconhecida, mesmo critério da *Response Rate 1 (RR1*) da American Association for Public Opinion Research (AAPOR, 2016, p. 61):

$$TR_{Região,Tipo} = RR1 = \frac{EC}{EC+NR+INT+OC+NA+RE+OUTRO},$$

onde

- Região = região do Brasil: Norte, Nordeste, Sul, Sudeste ou Centro-Oeste;
- Tipo = tipo de telefonia: Móvel ou Fixa;
- EC = Entrevista completa;
- NR = Não quis responder a entrevista;
- INT = Entrevista Interrompida;
- OC = Ligação Ocupada;
- NA = Não atendimento;
- RE = Chamada recusada;
- OUTRO = Outras classificações de elegibilidade desconhecida.

Foram consideradas linhas telefônicas elegíveis aquelas pertencentes a usuários da população-alvo, com discagens classificadas como EC, NR ou INT. Já as linhas telefônicas de elegibilidade desconhecida foram aquelas nas quais não foi possível definir se o usuário da linha pertencia ou não à população-alvo por apresentarem as classificações OC, NA, RE e OUTRO.

As taxas de respostas foram calculadas separadamente para telefonia móvel e fixa, pois cada tipo de acesso apresenta comportamento peculiar nesse quesito. Os cálculos foram feitos, além disso, por região, uma vez que algumas UF contam com poucas unidades amostrais, o que torna seus resultados extremamente voláteis.

Para cada indivíduo da amostra, o peso para ajuste de não resposta é dado pela fórmula abaixo:

$$w_{NR,Região,tipo} = \frac{1}{TR_{Região,Tipo}}$$

Peso sem pós-estratificação

O peso **sem pós-estratificação** do respondente k (w_k), com $k = 1, \dots, n,$ é obtido pela multiplicação entre o peso referente ao ajuste de não resposta ($w_{NR,Região,Tipo}$) e o peso obtido para o ajuste da probabilidade de seleção ($w_{sel,hj}$):

$$w_k = w_{sel,hj} \times w_{NR,Região,Tipo}$$

Por fim, os pesos são ajustados de forma que somem o total da amostra:

$$w'_k = n \times \frac{w_k}{\sum_{k=1}^{n} w_k}.$$

Pós-estratificação

No Brasil, em 2018, segundo dados do IBGE, 14% da população com 16 anos ou mais (população-alvo da pesquisa) não tinha acesso à telefonia. Temos, então, que a população referenciada (população com acesso à telefonia) não abrange toda a população-alvo. Há, portanto, viés de não cobertura. Esse viés pode ser atenuado pelo uso da pós-estratificação.

A pós-estratificação é um método para ajustar os pesos amostrais de modo que eles reflitam o tamanho da população em cada estrato populacional, tornando possível a inferência para a população-alvo. Esse método é utilizado para atenuar o viés de não cobertura, como o existente em pesquisas telefônicas. O processo de pós-estratificação pressupõe a existência, na amostra, de elementos de todos os perfis possíveis considerados na ponderação, que no caso das pesquisas DataSenado são, em geral: região, sexo, idade, escolaridade e raça/cor, o que resulta em centenas de combinações possíveis. No entanto, dado o tamanho da amostra, esse pré-requisito não pode ser atendido. Como alternativa à pós-estratificação, foi utilizada uma metodologia de aproximação conhecida como método *rake* ou *raking*.

O *raking* utiliza os totais populacionais conhecidos para ajustar os pesos amostrais, de forma que os valores marginais (soma das linhas e/ou colunas) de uma tabela na amostra ponderada somem os totais conhecidos da população. O algoritmo envolve a estimativa de pesos em cada par de variáveis repetidamente, até que os pesos convirjam. Essencialmente, o *raking* força que os totais ponderados da pesquisa correspondam aos totais da população, atribuindo um peso adequado a cada respondente (Fricker e Anderson, 2015, p. 38).

Na pesquisa sobre Fake News, o *raking* foi aplicado visando refletir a proporção da população por região segundo as seguintes características demográficas: sexo, idade, escolaridade e raça/cor. Para utilizar esse método, os totais populacionais foram estimados a partir da Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD Contínua) do 1º trimestre de 2020 para a população com 16 anos ou mais e raça/cor declarada. As perguntas da pesquisa usadas para ponderação foram feitas da forma mais similar possível às perguntas da PNAD Contínua.

Para o *raking*, os dados referentes à idade foram agrupados em 5 grupos (16 a 29 anos, 30 a 39 anos, 40 a 49 anos, 50 a 59 anos e 60 anos ou mais), os dados referentes à 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”, dado o tamanho da amostra.

Os pesos com pós-estratificação (w_{POS}) aproximados pelo *raking* foram gerados por meio do *software R,* utilizando o algoritmo contido na função *rake* do pacote *survey*.

Por fim, os pesos são ajustados novamente de forma que somem o total da amostra:

$$w'_{POS,k} = n \times \frac{w_{POS,k}}{\sum_{k=1}^{n} w_{POS,k}}.$$

Apresentação de resultados

O delineamento amostral, o peso sem pós-estratificação (w'_k) e a aplicação do método *rake* foram considerados para gerar as estimativas pontuais e respectivas margens de erro da pesquisa. Também é possível obter somente as estimativas pontuais, sem as respectivas margens de erro, por meio da aplicação direta do peso com pós-estratificação

($w'_{POS,k}$). Cada estimativa divulgada pelo DataSenado é acompanhada das respectivas margens de erros, calculadas com nível de confiança de 95%.

Os percentuais apresentados nas divulgações do DataSenado foram arredondados de maneira que, para números com decimal menor que 0,5, foi mantida a parte inteira; e para números com decimal maior ou igual a 0,5, adicionou-se uma unidade à parte inteira do número. O uso dessa metodologia de arredondamento faz com que, em alguns casos, a soma dos percentuais de gráficos e de algumas colunas das tabelas seja diferente de 100%, para mais ou para menos, sem que isso implique em erro de cálculo.

Referências bibliográficas

AMERICAN ASSOCIATION FOR PUBLIC OPINION RESEARCH. **Standard definitions: Final dispositions of case codes and outcome rates for surveys**, 2011.

BOLFARINE, H.; BUSSAB, W. de O. **Elementos de amostragem**. Edgard Blucher. São Paulo, 2005.

FRICKER, R; ANDERSON, L. **Raking: An Important Often Overlooked Survey Analysis Tool.** Phalanx, 2015. p. 36-42.

Realização

Senado Federal

Presidente: Senador Davi Alcolumbre

Secretaria de Transparência

Diretora: Elga Lopes

Instituto de Pesquisa DataSenado

Coordenadora: Laura do Nascimento

Elaboração e responsáveis técnicos

Estatístico responsável: Marcos Ruben de Oliveira
Estatística: Isabella Cristine Figueiredo Vieira- estatística
Estagiária de estatística: Luana Pereira Ramos da Silva
Revisora: Eleonora Stanziona Viggiano