Delineamento amostral das pesquisas DataSenado

Marcos Ruben de Oliveira[1], Marina Barros de Oliveira[1], Roberto de S. Marques Buffone[1], Aretha Pessanha Cordeiro[1],

MRUBEN@senado.leg.br

[1] *Instituto de Pesquisa DataSenado – Brasília, Brasil.*

1 Introdução

O Instituto de Pesquisa DataSenado realiza pesquisas de opinião para subsidiar o Senado Federal em suas decisões. Para tanto, faz uso delineamento amostral e técnicas de análises implementados em novembro de 2019. Este texto apresenta os principais procedimentos adotados, bem como suas fontes de dados e um comparativo de resultados que permite verificar as estimativas DataSenado com dados do IBGE e do INEP.

A população-alvo das pesquisas DataSenado é, em geral, composta de cidadãos brasileiros com 16 anos ou mais. Os participantes são selecionados via amostragem aleatória 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.

A pesquisa utilizada no texto versou sobre a violência nas escolas e foi realizada pelo instituto em maio de 2023, quando foram entrevistadas 2068 pessoas.

A seguir são apresentadas as duas etapas dos estudos realizados pelo DataSenado. Na primeira etapa, abordar-se o delineamento amostral. Na segunda etapa, trata-se da análise dos dados, que leva em conta a amostragem complexa levada a efeito.

2 Metodologia

2.1 Delineamento amostral

O DataSenado faz, via de regra, amostragens aleatórias estratificadas com alocação proporcional à população com 16 anos ou mais nas unidades da

Federação, como foi o caso da pesquisa "Violência nas escolas 2023". Como referência de proporcionalidade, foram usados dados da Pesquisa Nacional por Amostra de Domicílios Contínua do 4º trimestre de 2022, último dado disponível quando da elaboração do delineamento. Durante a realização da coleta, algumas entrevistas excederam o delineado e a Tabela 1 apresenta os quantitativos planejados e realizados.

Tabela 1- Tamanho da amostra planejada e realizada

Unidade da Federação	Amostra planejada	Amostra realizada
Rondônia	17	19
Acre	8	9
Amazonas	36	38
Roraima	6	9
Pará	78	80
Amapá	8	10
Tocantins	15	17
Maranhão	63	69
Piauí	30	31
Ceará	87	88
Rio Grande do Norte	34	35
Paraíba	37	40
Pernambuco	89	91
Alagoas	30	32
Sergipe	22	24
Bahia	138	148
Minas Gerais	202	208
Espírito Santo	39	43
Rio de Janeiro	170	173
São Paulo	446	445
Paraná	109	112
Santa Catarina	70	72
Rio Grande do Sul	110	112
Mato Grosso do Sul	26	28
Mato Grosso	32	33
Goiás	68	70
Distrito Federal	30	32
Brasil	2000	2068

Na etapa de delineamento amostral são gerados os números a serem discados, de forma aleatória, a partir do cadastro de números habilitáveis da Agência Nacional de Telecomunicações – Anatel – e de tal forma que a probabilidade de ocorrência de números fixos e móveis válidos sejam iguais.

Para tanto, usando as estatísticas mais recentes de acesso a telefonia fornecidas pela Anatel (fevereiro de 2023), tem-se que o tamanho da lista de telefones móveis por UF é dado por (raciocínio análogo fornece o tamanho da lista para telefones fixos):

$$Tamanho.lista_{Móvel} = \frac{Números\_habilitáveis_{móvel}}{Qtde_{usuários.(fixo+móvel)}} \times (Qtde\_números\_desejados)$$

No cálculo acima, a quantidade de números desejados leva em conta as taxas de respostas de pesquisas passadas do DataSenado.

Por fim, a lista de discagem é totalmente aleatorizada de maneira que as entrevistas ocorram aproximadamente de forma síncrona por unidade da Federação, respeitada eventual aleatoriedade das discagens.

2.2 Análise de dados da amostra

Tratando-se de delineamento amostral complexo, as análises inferenciais requerem o cuidado de 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.

2.2.1 Probabilidade de seleção

A probabilidade de seleção dos entrevistados, representada por f_{hj}, onde h representa o estrato e j o participante, 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 estatísticas mais recentes da Anatel. No cálculo da probabilidade de seleção foram considerados todos os telefones habilitados da lista usada na amostra, e o peso amostral devido a esse fator é dado por $w_{sel,hj} = \frac{1}{f_{hj}}$.

2.2.2 Taxa de resposta

Como várias ligações resultam em recusa a participar da pesquisa, números não atendidos, dentre outras ocorrências que impedem que uma ligação para número habilitado resulte em pesquisa concluída, fez-se necessário considerar o peso amostral devido à Taxa de Resposta. A estimativa da taxa de resposta por região e tipo de telefonia foi obtida via *Response Rate 1* (RR1) da *American Association for Public Opinion Research* (AAPOR, 2023, p. 85), a partir de metadados das discagens telefônicas coletados no decorrer da pesquisa.

A taxa de resposta da pesquisa representa o número de entrevistas concluídas em relação ao número de ligações realizadas para números habilitados e válidos. 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 AAPOR. A fórmula abaixo representa o procedimento adotado:

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

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; e 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 UFs 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 por:

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

2.2.3 Rake

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 ($w_{sel,hj}$):

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

No Brasil, em 2021, segundo dados do IBGE, 10,8% da população com 16 anos ou mais não tinha acesso à telefonia. Como a população referenciada (população com acesso à telefonia) não abrange toda a população-alvo, há viés de não cobertura. Esse viés pode ser atenuado pelo uso da pós-estratificação.

A pós-estratificação ajusta os pesos amostrais de modo que reflitam o tamanho da população em cada estrato populacional, tornando possível a inferência para a população-alvo. Esse método auxilia a 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, o que nem sempre ocorre. 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).

Para utilizar esse método na pesquisa sobre Violência nas escolas, foram consideras as seguintes características sociodemográficas: sexo, idade, escolaridade, raça/cor, situação do domicílio e porte do município. A

distribuição estimada da população brasileira segundo tais características foram estimadas usando dados da Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD Contínua) do 1º trimestre de 2023. Já o porte do município foi categorizado segundo os dados da Estimativa Populacional de 2021 do IBGE, e os municípios foram divididos em três categorias: até 50.000 habitantes, de 50.001 a 500.000 habitantes e mais de 500.000 habitantes.

Para o *raking*, os dados referentes à idade foram agrupados em 5 categorias (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". Com relação à situação de domicílio dos respondentes, classificou-se em área "Urbana" e "Rural". E, por fim, os dados referentes à cor/raça foram agrupados em "Branca/Amarela" e "Preta/Parda /Indígena", dado o tamanho da amostra. Dessa forma, foi possível gerar estimativas diversas para a população-alvo, bem como informar as respectivas margens de erro para cada estimativa.

Com o objetivo estimar a quantidade de estudantes que sofreram algum tipo de violência em ambiente escolar nos últimos 12 meses, o DataSenado perguntou se o entrevistado que estuda sofreu alguma violência e, adicionalmente, perguntou para todos os entrevistados se algum morador do seu domicílio é estudante e se sofreu violência na escola nos últimos 12 meses. Para os que responderam sim à essa última pergunta, foi perguntado quantos moradores passaram por essa experiência. Tais perguntas permitiram obter, na amostra, a quantidade de moradores por domicílio que sofreu violência em ambiente escolar nos últimos 12 meses. Trata-se de uma característica do domicílio, não mais do entrevistado, devendo ser computada como tal nos cálculos inferenciais.

A partir das informações domiciliares acima foi possível estimar a quantidade de brasileiros que passaram por algum tipo de violência escolar nos últimos 12 meses. Para tanto, o DataSenado ponderou as respostas, via *raking*, pelas seguintes características de domicílios: região, situação do domicílio (urbana ou rural), localização do município do domicílio (capital ou não capital), e quantidade de moradores por domicílio (até 2 pessoas, 3, 4, 5, 6 ou mais). Como

referência no *rake*, foram utilizados os dados de domicílio disponíveis na PNAD Contínua 1º de 2023.

Os pesos com pós-estratificação aproximados pelo *raking* foram gerados por meio do *software R,* utilizando o algoritmo contido na função *rake* do pacote *survey*. O delineamento amostral, o peso sem pós-estratificação 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. Cada estimativa divulgada pelo DataSenado é acompanhada das respectivas margens de erros, calculadas com nível de confiança de 95%.

No link da pesquisa[1] está disponível painel interativo que permite selecionar qualquer pergunta da pesquisa e conferir as estimativas pontuais e suas respectivas margens de erro, computadas a 95% de confiança. Resultados mais detalhados estão disponíveis no relatório, acessível pelo mesmo link.

2 Conclusão

Para checar os resultados obtidos a partir da aplicação do método descrito, comparou-se a estimativa de total de alunos matriculados em 2022 a partir da pesquisa DataSenado com as estatísticas disponibilizadas pelo IBGE e pelo INEP.

A Tabela 2 apresenta essa comparação:

Tabela 2- Dados de matrícula do INEP X Estimativa IBGE X Resultados DataSenado

Matrículas – INEP(1)	Nº de estudantes – IBGE (2) (Pnadc 2º Trimestre de 2022)	Estimativa do nº de estudantes – DataSenado (3)	Margem de erro da estimativa DataSenado (95%)
56.369.194	58.246.106	59.752.098	1.999.354

Fontes:

(1) Censo da Educação. Básica 2022 e do Ensino. Superior 202. As informações de matrículas no Ensino Superior pelo INEP não haviam sido divulgadas até o momento de redação do texto.
(2) IBGE PNAD Contínua 2º Trimestre de 2022.
(3) Instituto de Pesquisa DataSenado - coleta de 9 a 10.5.2023

Nota: intervalo calculado a 95% de confiança.

[1] Ver https://www12.senado.leg.br/institucional/datasenado/publicacaodatasenado?id=quase-7-milhoes-de-brasileiros-sofreram-violencia-no-ambiente-escolar-nos-ultimos-12-meses.

Verifica-se que a estimativa de estudantes do DataSenado engloba o resultado apresentado pelo IBGE, dentro da margem de erro calculada e do intervalo com 95% de confiança. Os resultados do INEP, por sua vez, apresentaram comparabilidade comprometida por não estarem disponíveis os dados de matrículas no ensino superior em 2022 até o momento.

Dentre os vários resultados da pesquisa, apresentados em audiência pública na Comissão de Educação no dia 4 de julho de 2023, destaca-se a inferência de que quase 7 milhões de brasileiros sofreram algum tipo de violência no ambiente escolar nos 12 meses anteriores à coleta de dados. Tais resultados são utilizados para desenvolvimento de políticas públicas e aprimoramento do aparato legal brasileiro.

Agradecimentos

Agradecemos ao Instituto de Pesquisa DataSenado, à Secretaria de Transparência e à Presidência do Senado Federal pelo apoio na execução e publicação dos resultados da pesquisa.

Referências bibliográficas

AMERICAN ASSOCIATION FOR PUBLIC OPINION RESEARCH. Standard definitions: Final dispositions of case codes and outcome rates for surveys, 2023. Disponível em https://aapor.org/standards-and-ethics/standard-definitions/. Consultado em 11/9/2023.

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.