Delineamento amostral das pesquisas DataSenado

Marcos R. de Oliveira, Marina B. de Oliveira, Roberto de S. M. Buffone, Aretha P. Cordeiro

2023-11-17

Instituto de Pesquisa DataSenado
Secretaria de Transparência
SENADO FEDERAL

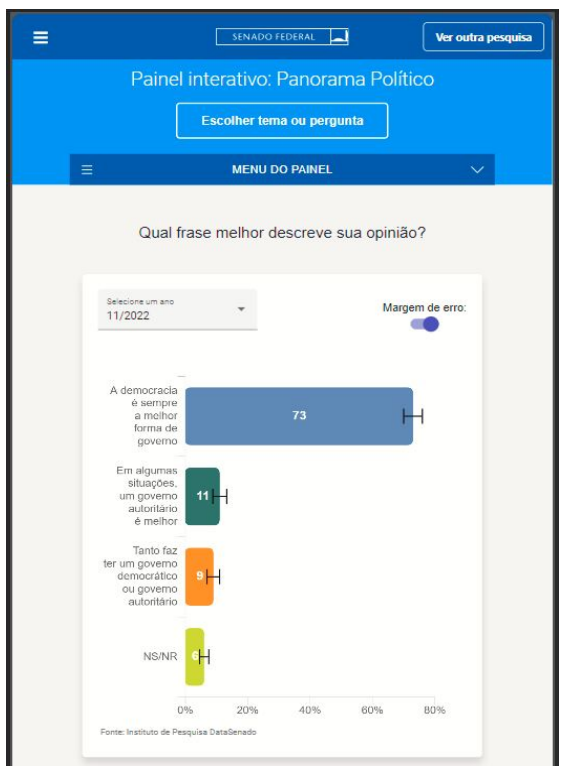
https://proceedings.science/p/174405?lang=pt-br

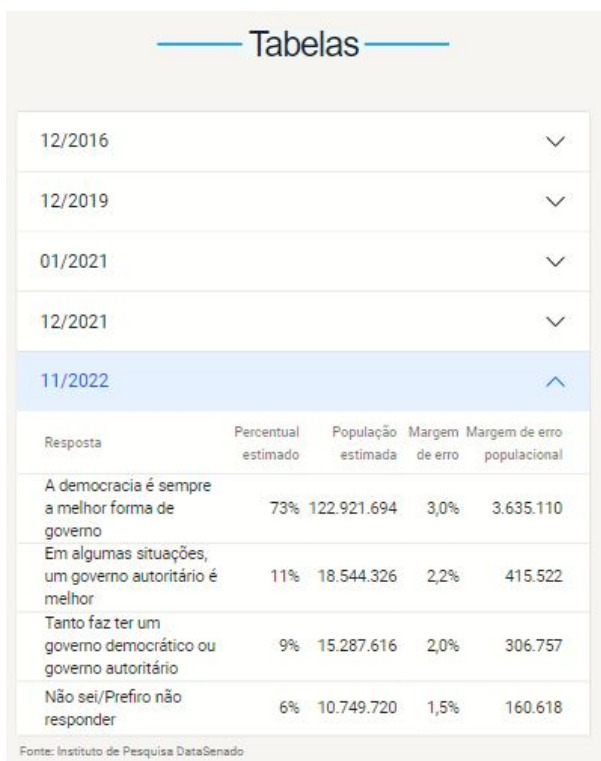
Instituto de Pesquisa DataSenado

- Criado em 2005 para auxiliar senadores e comissões nas duas principais atividades do Senado Federal: legislar e fiscalizar o Poder Executivo.
  - Mais de 4,7 milhões de entrevistas
  - 128 pesquisas nacionais probabilísticas
  - Levantamentos on-line (público interno e enquetes)
  - Pesquisas qualitativas - grupos focais e entrevistas em profundidade
  - Exemplo: Pesquisa Nacional sobre Violência contra a Mulher: primeira edição (2005) auxiliou a criação da Lei Maria da Penha. Em 2023, 10ª edição, maior e mais longa série histórica sobre o tema

Instituto de Pesquisa DataSenado

- O Senado Federal disponibiliza publicamente relatórios, dados e descrição detalhada do método: www.senado.leg.br/datasenado
- Painéis interativos para consulta, cruzamentos e download de microdados das principais pesquisas:





Delineamento amostral DataSenado

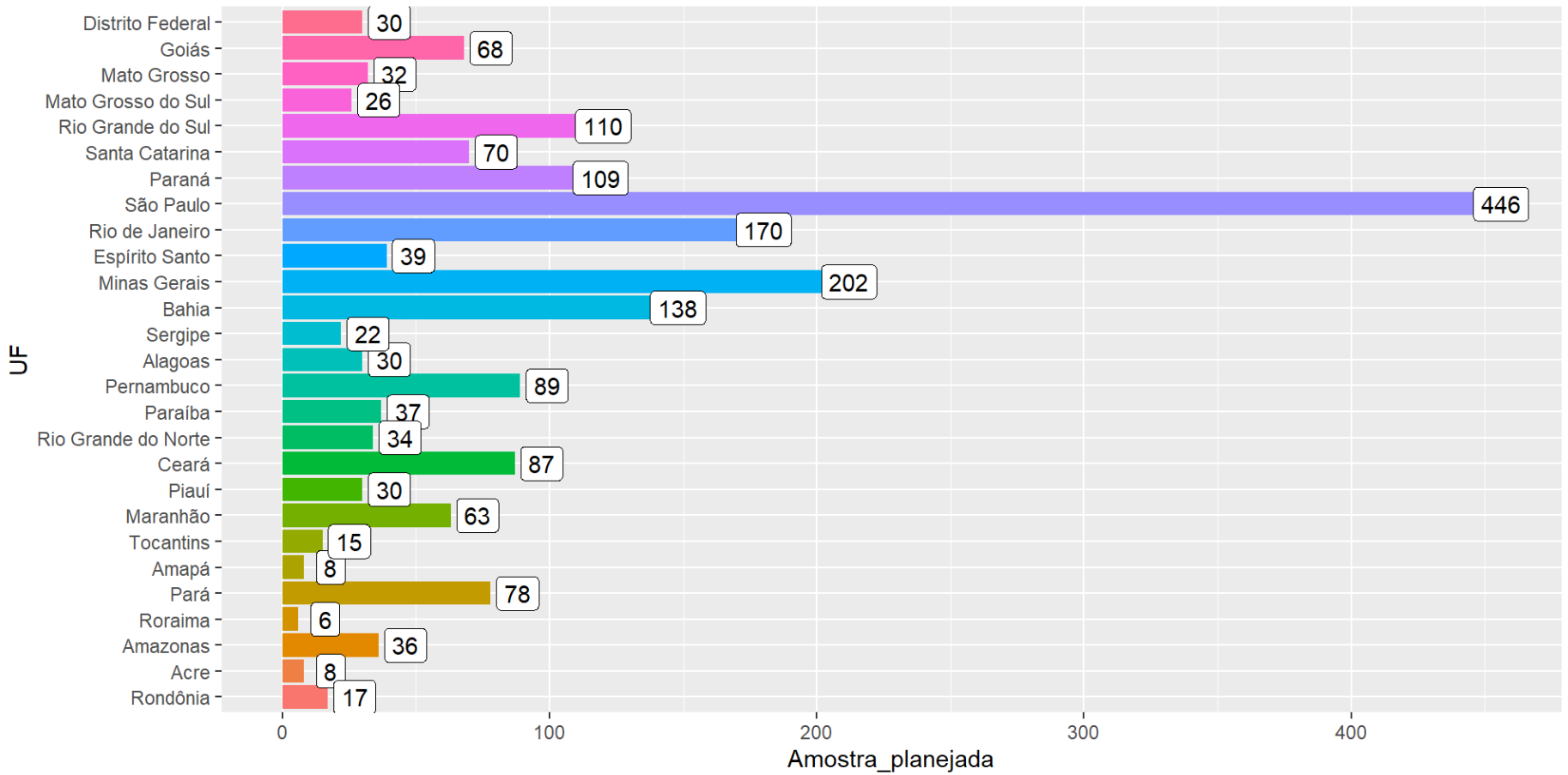
- Método original, criado pelos estatísticos do DataSenado
- Características **usuais** das pesquisas DataSenado:
  - População-alvo: residentes no Brasil com 16 anos ou mais de idade
  - Amostragem Aleatória Estratificada por estados e DF, com alocação proporcional à população-alvo
  - Coleta de dados via CATI (Computer-Assisted Telephone Interviewing)
  - Seleção aleatória de números discados - RDD (Random digit dialing)
  - Protocolos rígidos de qualidade e auditoria das entrevistas

https://proceedings.science/p/174405?lang=pt-br

Delineamento amostral DataSenado

Exemplo: pesquisa “Violência nas Escolas”

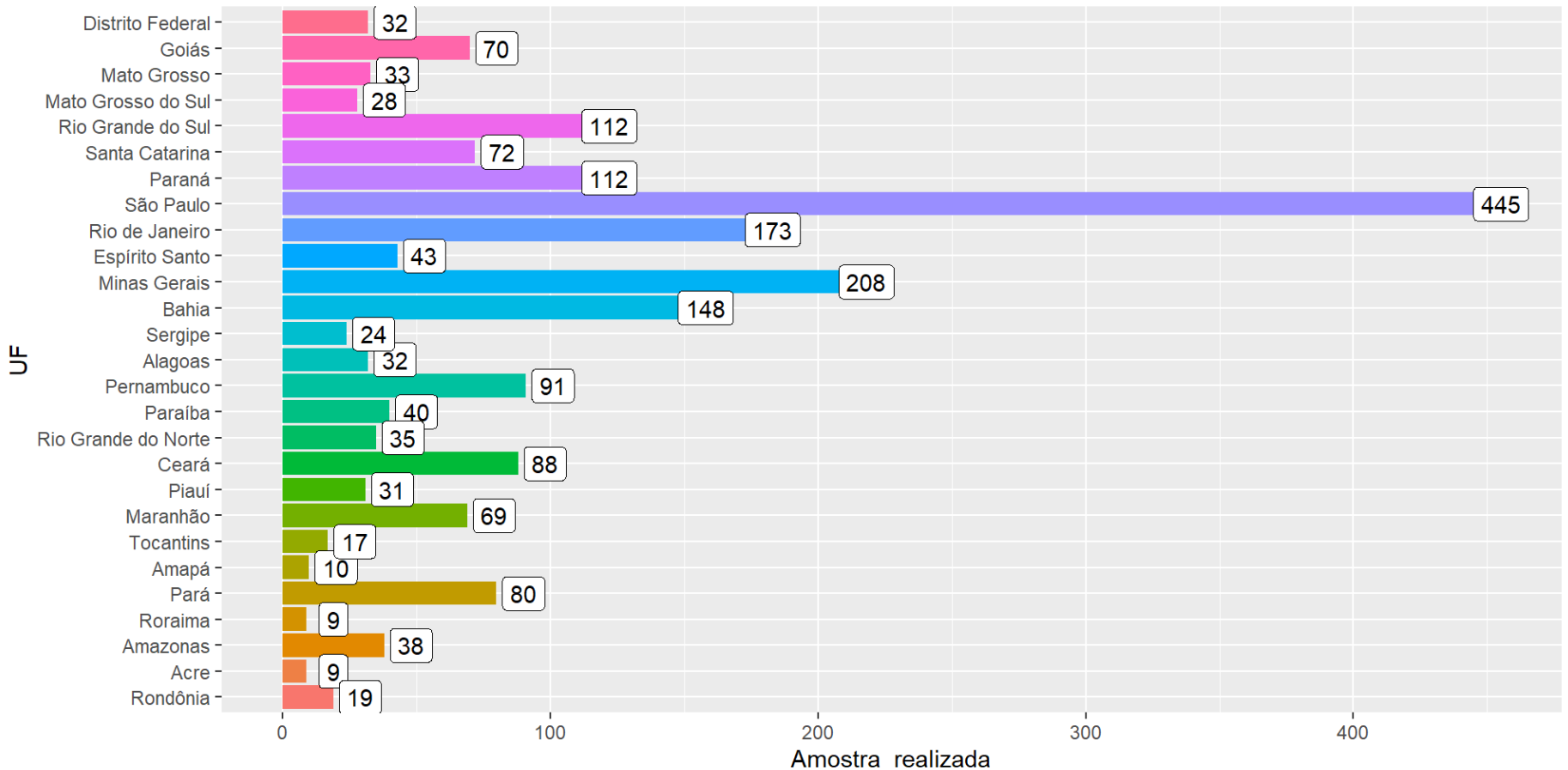
- Amostra planejada: 2.000 entrevistas



Delineamento amostral DataSenado

Exemplo: pesquisa “Violência nas Escolas”

- Amostra realizada: 2.068 entrevistas



Geração de números aleatórios - RDD

- ANATEL regulamenta a numeração de telefonia fixa e móvel no Brasil e disponibiliza os números liberados para uso pelas operadora. Consulta em 27/4/2023:
  - 585.333.050 números móveis possíveis
  - 229.151.319 números fixos possíveis
- ANATEL também informa quantidade de linhas ativas por tipo de telefonia. Consulta em abril, com dados referentes a fevereiro de 2023:
  - 250.644.275 linhas ativas de telefonia móvel
  - 26.644.230 linhas ativas de telefonia fixa
- Dados por estados e DF

Geração de números aleatórios - RDD

- Amostra divida em dois grupo: números de telefones fixos e móveis
- Probalidade de seleção de um número fixo válido igual à probabilidade de seleção de número móvel válido.
- Procedimento aplicado a cada um dos 26 estados e o Distrito Federal (UFs)
- Seja i a UF e j o tipo de telefonia, $j = \{fixo, móvel\}$. O tamanho de amostra n_{ij} é dada pela proporção entre a quantidade de linhas ativas:

$$n_{ij} = \frac{linhas.ativas_{ij}}{\sum_j (linhas.ativas_{ij})} \times \sum_j n_{ij}$$

Geração de números aleatórios - RDD

A quantidade de números selecionados a partir da lista de números autorizados pela ANATEL ($Lista_{ij}$) é calculada de maneira a preservar a igualdade de chances de seleção entre telefones fixos e móveis:

$$Lista_{ij} = \frac{Números.possíveis_{ij}}{\sum_j linhas.ativas_{ij}} \times \sum_j n_{ij}$$

Geração de números aleatórios - RDD

A taxa de sucessos no uso da lista é influenciada por: números inválidos, de empresas, usados apenas para internet, ….

- Grande maioria: números inválidos
- Últimos 3 anos:
  - a cada mil números, cerca de 6 “alôs”
  - após “alô”: em média 20% de entrevistas concluídas

A lista final efetivamente usada em campo é inflada por esses resultados segundo a UF:

$$Lista^{*}_{ij} = \frac{Números.possíveis_j}{\sum_j linhas.ativas_j} \times \sum_j n_{ij} \times \frac{histórico.discados_i}{histórico.concluídos_i}$$

Geração de números aleatórios - RDD

Exemplo: pesquisa “Violência nas Escolas”

- Lista de 2.557.007 números aleatórios, 73% móveis
- Coleta de dados - campo: 9 e 10 de maio de 2023
- 655.202 ligações para 386.023 números distintos (renitência média de 1,7)
- 11.673 “alôs”
- 2.068 entrevistas concluídas

Análise dos dados amostrais

O delineamento amostral DataSenado é incorporado às análises considerando:

1. Probabilidade de seleção dos entrevistados
2. Taxas de respostas
3. Ponderação via *raking*

Probabilidade de seleção

A chance de um número habilitado qualquer ser selecionado é conhecida e dada pela razão entre a quantidade de linhas ativas detectadas no processo de discagem e a quantidade de linhas ativas total na UF i, tipo j, informada pela Anatel:

$$\pi_{ij} = \frac{linhas.ativas.discadas_{ij}}{linhas.ativas.Anatel_{ij}}$$

Já a chance de uma pessoa k qualquer ser selecionada (f_{ijk}) depende, além de π_{ij}, da quantidade de pessoas que compartilham a mesma linha de telefone (δ_k)[1]:

$$f_{ijk} = \pi_{ij} \times \frac{1}{\delta_k}$$

1. *Nota: até recentemente considerávamos outras linhas a que a pessoa tem acesso, porém desconsiderar essa informação não impacta nas estimativas divulgadas (diferença ocorre após a segunda casa decimal), além de simplificar muito o questionário, o que aumenta a taxa de resposta.*

Peso em função da probabilidade de seleção

Computada a probabilidade de seleção, o peso associado a essa informação é dado por

$$w_{sel,ijk} = \frac{1}{f_{ijk}}$$

Exemplo: pesquisa “Violência nas Escolas”

$$\bar{f} = 0.000739$$

$$\overline{w_{sel}} = 1956.7$$

Taxas de respostas

Ligações para linhas ativas, com números válidos, podem resultar em: entrevistas concluídas, interrompidas, agendadas e não finalizadas, recusadas, ligações não atendidas, linhas ocupadas, caixas postais, telefones desligados, pessoas fora da amostra, linhas não particulares, dentre outros

A taxa de resposta é calculada considerando essas classificações e a *RR1* da AAPOR:

$$RR1 = \frac{Entrevistas.completas}{Linhas.ativas + Linhas.potencialmente.ativas - Inelegíveis}$$

Onde linhas potencialmente ativas são aquelas em que não é possível afirmar que a linha está inativa, e inelegíveis são empresas e pessoas fora da população-alvo.

Taxas de respostas

Problema: em amostras com alocação proporcional, algumas UFs tem poucas entrevistas alocadas e o estrato é finalizado apenas com telefonia móvel.

Solução: agrupar por Grande Região e tipo de telefonia:

$$RR1_{região,tipo} = \frac{Entrevistas.completas_{região,tipo}}{Linhas.potencialmente.ativas_{região,tipo} - Inelegíveis_{região,tipo}}$$

Peso em função da taxa de resposta

Dado por

$$w_{RR_{região,j}} = \frac{1}{RR1_{região,j}}$$

No exemplo da pesquisa ‘Violência nas Escolas’:

- RR1 varia de 0.001260 (fixo, Norte) a 0.013355 (móvel, Sul).
- Os pesos variam de 74.87832 a 793.6508.

Raking

- População-alvo: residentes no Brasil com 16 anos ou mais de idade.
- População amostrada: pessoas com 16 anos ou mais com acesso à telefonia. Em 2022 apenas 1,7% de moradores não tinham acesso a telefonia:

Tabela 7304 - Domicílios e Moradores, por existência de telefone no domicílio (inclui UF, RM e RD)						
Variável - Distribuição percentual dos moradores em domicílios (%)						
Brasil						
Ano	Existência de telefone					
	Total	Havia telefone	Havia telefone fixo convencional	Havia telefone móvel celular	Havia telefone fixo convencional e telefone móvel celular	Não havia telefone
2018	100,0	96,4	26,8	95,6	26,0	3,6
2019	100,0	96,7	22,7	96,1	22,1	3,3
2021	100,0	98,0	15,1	97,7	14,8	2,0
2022	100,0	98,3	11,8	98,0	11,5	1,7
Fonte: IBGE - Pesquisa Nacional por Amostra de Domicílios Contínua Anual - 4º trimestre						

Raking

- *Raking:* ajuste das **distribuições marginais** por **região** das características sociodemográficas da população-alvo: sexo, raça/cor (PPI e não PPI), faixa etária, escolaridade, situação de domicílio, porte do município
- Dados de referência para o *raking*: PNAD Contínua 1º/2023 e Estimativa Populacional 2021 para porte do município
- Peso por respondente pré-*raking*:

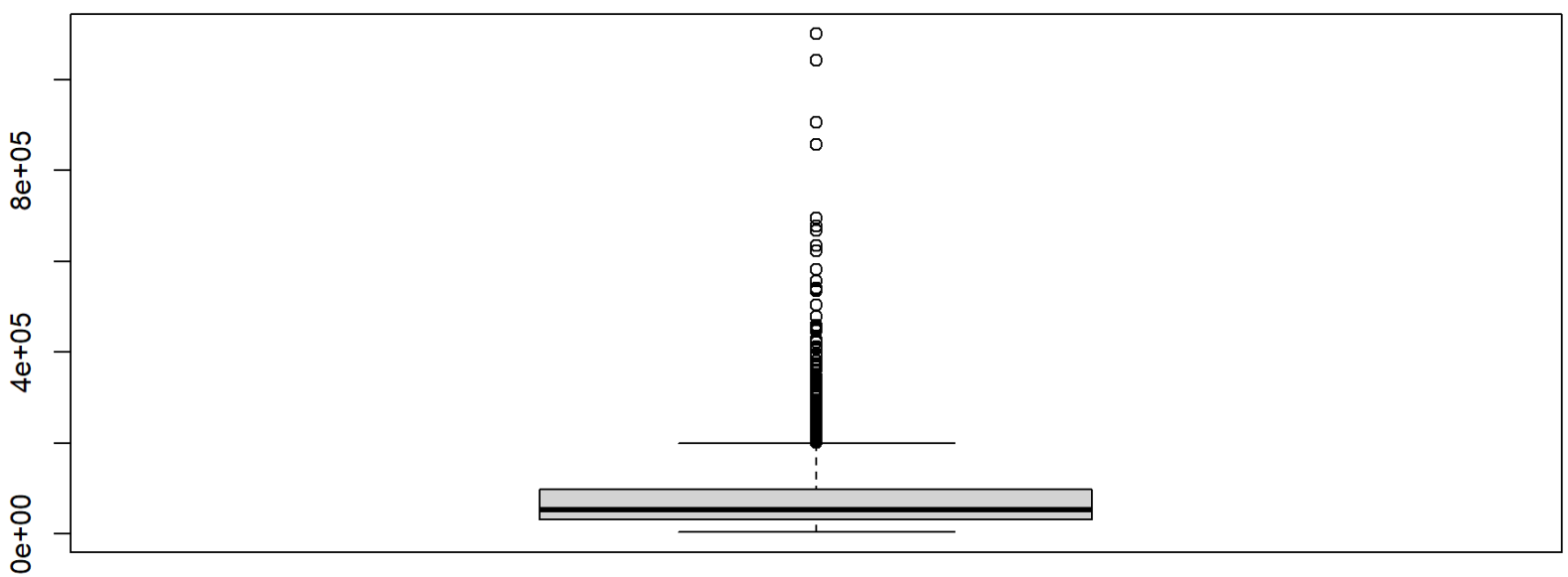
$$w_k^1 = w_{sel,ijk} \times w_{RR1_{região,j}}$$

- Peso por repondente pós-*raking*: w_k^2

Raking

Resultados - pesos após *raking*:

media	desvpad
81376.68	89960.68



Raking

Estimativa de pessoas com 16 anos ou mais que já sofreram violência na escola **em algum momento da vida**:

Você já sofreu algum tipo de violência no ambiente escolar?

	Estimativa	**E.P.**
Sim	0.22	0.0123
Não	0.78	0.0124

IC(95%) para a pergunta ‘Você já sofreu algum tipo de violência no ambiente escolar?’

	2.5 %	**97.5 %**
Sim	0.19	0.24
Não	0.75	0.80

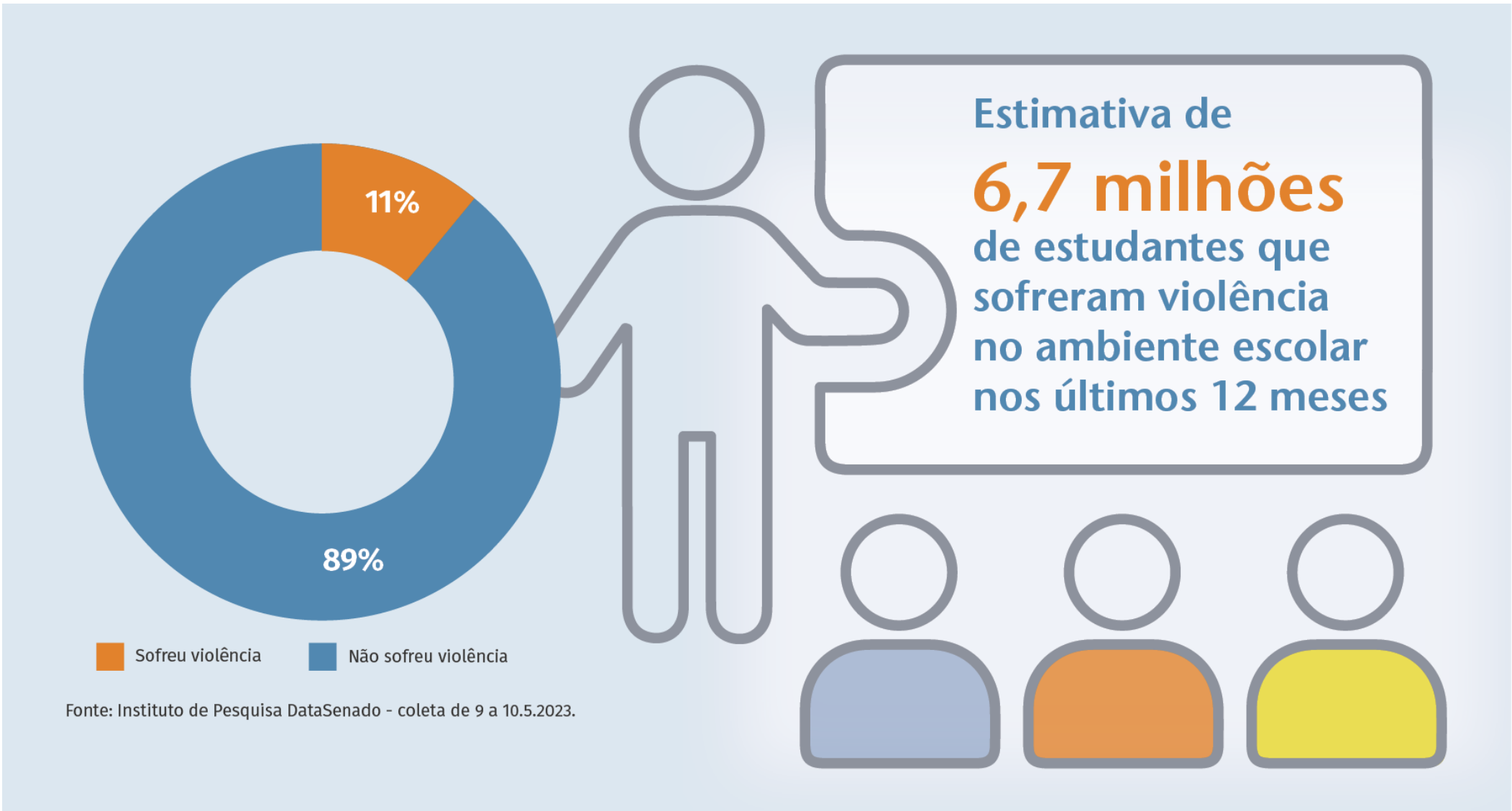
Estimativa de total - ‘Você já sofreu algum tipo de violência no ambiente escolar?’

	Total	E.P.
Sim	36496528	2056404
Não	129782744	2108468

Resultados adicionais

- Na pesquisa ‘Violência nas escolas’ foi feita, adicionalmente, a pergunta outros moradores do domicílio são estudante e sofreram alguma violência nos últimos 12 meses
- Obteve-se, assim, uma informação do DOMICÍLIO
- Com o objetivo de estimar a quantidade de pessoas que sofreram violência na escola nos últimos 12 meses no Brasil, independente da idade, aplicou-se o *raking* para domicílio por **região** usando as variáveis
  - 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).

Resultados adicionais



Resultados adicionais

"Estudantes que sofreram algum tipo de violência em ambiente escolar nos últimos 12 meses" - Brasil - 2023

	Estimativa	Margem de erro	População de estudantes estimada
Sim	11%	±2,3%	6.730.480
Não	89%	±2,6%	53.021.618
Total	100%	-	59.752.098

Fonte: Instituto de Pesquisa DataSenado - coleta de 9 a 10.5.2023.

"Estudantes que sofreram violência em ambiente escolar nos últimos 12 meses" por região - Brasil - 2023

	Estimativa (± margem de erro)				
	Norte/Nordeste	Sudeste	Sul	Centro-Oeste	Brasil
Sim	12% (±4,5%)	12% (±5,0%)	9% (±4,6%)	5% (±3,7%)	11% (±2,3%)
Não	88% (±4,5%)	88% (±5,0%)	91% (±4,6%)	95% (±3,7%)	89% (±2,6%)
Total	100%	100%	100%	100%	100%
População Estimada	23.330.328	24.065.528	8.126.595	4.229.648	59.752.098

Fonte: Instituto de Pesquisa DataSenado - coleta de 9 a 10.5.2023.
Nota: Soma dos percentuais difere de 100% devido ao arredondamento.

Conclusão

- Checagem das estimativas: comparativo do total de alunos matriculados em 2022 a partir da pesquisa DataSenado x dados PNAD Contínua x INEP 2022:

INEP 2022 (mil)	PnadC Educação (mil)	DataSenado mai/2023	DataSenado Lim.Inf. (95%)
56.826	58.246	59.752	57.753

- Resultados DataSenado acertam o parâmetro populacional, considerada a margem de erro
- Outras pesquisas confirmaram esse resultado (Auxílio Emergencial, pesquisa com candidatos, …)

Conclusão

Dúvidas, sugestões, críticas, contatos:

Marcos Ruben de Oliveira

mruben@senado.leg.br

Obrigado!