IBGE
Instituto Brasileiro de Geografia e Estatística

Pesquisa Nacional por Amostra de Domicílios Contínua – PNAD Contínua

Nota técnica 04/2021

Sobre a alteração do método de calibração dos fatores de expansão da PNAD Contínua

I) Introdução

Desde abril de 2020, o IBGE tem realizado estudos sobre a queda da taxa total de aproveitamento da amostra da pesquisa, observada com o início da pandemia. Após diversas avaliações[1], o IBGE entendeu que a reponderação da pesquisa, considerando adicionalmente a calibração por sexo e grupo etário, que estava incialmente planejada para ser implementada após o Censo que ocorreria em 2020, deveria ser feita imediatamente, mesmo com o uso de projeções populacionais baseadas em dados do Censo de 2010.

Dessa forma, a partir da divulgação do 3º trimestre de 2021[2], o processo de calibração dos fatores de expansão da Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD Contínua) sofrerá alteração metodológica.

Em pesquisas por amostragem, como a PNAD Contínua, é possível utilizar informações auxiliares, vindas de fontes externas, no intuito de melhorar a qualidade de suas estimativas. Essas informações são utilizadas para ajustar os fatores de expansão, processo conhecido como calibração.

Até a divulgação do trimestre móvel de junho-julho-agosto de 2021, a PNAD Contínua utilizava pesos ajustados de maneira que o total populacional da pesquisa, para alguns recortes geográficos, coincidisse com o das estimativas populacionais produzidas pela Coordenação de População e Indicadores Sociais do IBGE. Esse tipo de calibração é conhecido por pós-estratificação. Além de permitir a comparabilidade dos totais populacionais estimados com as

[1] Para maiores detalhamentos, consultar Nota técnica 03/2021 (out./2021) e Nota Técnica 02/2021 (abr./2021), disponíveis em: https://www.ibge.gov.br/estatisticas/sociais/trabalho/9171-pesquisa-nacional-por-amostra-de-domicilios-continua-mensal.html?=&t=notas-tecnicas

[2] Equivalente ao trimestre móvel de julho-agosto-setembro/2021.

estimativas populacionais, o processo de calibração proporciona uma melhoria na precisão das estimativas relacionadas com as variáveis utilizadas no processo, além de mitigar possíveis problemas de cobertura e não-resposta da pesquisa.

No novo processo de calibração implementado, além dos totais populacionais por recortes geográficos, os fatores de expansão da PNAD Contínua também serão ajustados para coincidir com estimativas de sexo e classes de idade para Brasil.

A alteração na metodologia de calibração motivou o IBGE a realizar também uma alteração no método de cálculo das estimativas de precisão da pesquisa. Para o cálculo das estimativas de variância, utilizava-se, até o trimestre móvel de junho-julho-agosto de 2021, o método do Conglomerado Primário e, por se tratar de estimadores de razão, aplicava-se, também, linearização de Taylor. A partir da divulgação do 3º trimestre de 2021, as estimativas de variância e os coeficientes de variação da PNAD Contínua serão calculados pelo método de replicação *bootstrap*.

A seguir, detalham-se alguns aspectos metodológicos dessas mudanças.

II) Calibração dos fatores de expansão

De maneira geral, o ajuste nos fatores de expansão utilizando calibração pode ser feito com base em outras informações auxiliares, além das informações de totais populacionais por recortes geográficos. Para isso, pode-se utilizar pesos calibrados, que são aqueles com maior proximidade do peso original do desenho, de acordo com uma determinada medida de distância, respeitando um conjunto de restrições.

Assim, considerando w_k o peso original do desenho, tem-se como peso calibrado w_k^* aquele que satisfaz a seguinte equação de calibração:

$$\sum_{k \in s} w_k^* \boldsymbol{x}_k = \sum_{k \in U} \boldsymbol{x}_k \qquad (1)$$

Onde w_k^* é o mais próximo possível de w_k, $\boldsymbol{x}_k$ é um vetor de variáveis auxiliares disponível para cada unidade k, s é o conjunto da amostra disponível, e U o conjunto de unidades da população de pesquisa**.** O vetor $\sum_{k \in U} \boldsymbol{x}_k$ com os totais populacionais das variáveis $\boldsymbol{x}$ é suposto conhecido.

Uma aplicação particular dessa técnica consiste na calibração com totais conhecidos na forma de uma tabela de frequências de múltiplas dimensões, que podem ser separadas em dois tipos possíveis:

a) Calibração em relação aos totais de cada célula

Neste caso, suponha-se, por simplicidade, que há 2 variáveis auxiliares: x_1 e x_2 com L e C categorias, respectivamente. Essas duas variáveis dão origem a uma tabela de frequências com L linhas e C colunas. Todos os elementos T_{lc} da tabela de frequências onde l = 1, ..., L e c = 1, ..., C, são supostos conhecidos e utilizados na calibração. O total populacional T pode ser calculado como: $T = \sum_l \sum_c T_{lc}$. Esse caso é chamado de pós-estratificação completa.

b) Calibração em relação aos totais marginais de cada variável auxiliar

Neste caso, a tabela bidimensional formada pelos elementos T_{lc} não é conhecida, mas os totais marginais de linhas T_{l+} e de colunas T_{+c} são conhecidos. Esse caso é chamado de pós-estratificação incompleta ou *raking* generalizado.

Processo de calibração na PNAD Contínua

Após estudos realizados, o novo modelo de calibração escolhido para ser adotado na PNAD Contínua foi o *raking* com a utilização de dois conjuntos de totais populacionais marginais.

Os primeiros totais marginais utilizados foram as estimativas populacionais para os 77 recortes geográficos[3] utilizados atualmente no processo de expansão da pesquisa. Cada Unidade da Federação é dividida em capital, resto da Região Metropolitana da capital (nas Unidades da Federação em que há Região Metropolitana na capital) e resto da Unidade da Federação, além das Regiões Integradas de Desenvolvimento (RIDEs).

A segunda marginal foi composta pelas estimativas dos contingentes populacionais por sexo e idade para o Brasil, fornecidas pela Coordenação de População e Indicadores Sociais da Diretoria de Pesquisas do IBGE, e foram divididas em 34 classes, com 2 categorias da variável sexo e 17 faixas etárias quinquenais, a saber: 0 a 4 anos , 5 a 9 anos, 10 a 13 anos, 14 a 19 anos, 20 a 24 anos, 25 a 29 anos, 30 a 34 anos, 35 a 39 anos, 40 a 44 anos, 45 a 49 anos, 50 a 54 anos, 55 a 59 anos, 60 a 64 anos, 65 a 69 anos, 70 a 74 anos, 75 a 79 anos e 80 anos ou mais. A faixa de 14 a 19 anos foi utilizada para que seu limite inferior coincida com a definição do limite da população em idade de trabalhar da pesquisa.

Para o cálculo dos pesos de calibração w_k^* foi utilizada a biblioteca *survey* do pacote estatístico R. Além das equações de calibração (1), outras duas restrições foram adicionadas. A primeira restrição refere-se à razão entre o peso calibrado w_k^* sobre o peso original do desenho w_k, que deve ficar entre 0,2 e 5,0. A segunda restrição consiste em que todos os moradores de um mesmo domicílio devam ter o mesmo peso calibrado. Esta última restrição corresponde à ideia de 'ponderação integrada por domicílios', que busca simplificar a obtenção de estimativas da pesquisa, mediante disponibilização de um único peso calibrado para os domicílios e seus moradores.

III) Cálculo das estimativas de variância da PNAD Contínua

A PNAD Contínua é uma pesquisa por amostragem probabilística de domicílios; portanto, o IBGE tem como prática divulgar, além das estimativas da pesquisa, uma medida de precisão associada. No caso da PNAD Contínua, juntamente com seus indicadores, são divulgados os coeficientes de variação.

Como mencionado anteriormente, com a mudança de metodologia da calibração da PNAD Contínua, o IBGE alterou a forma de cálculo das estimativas de variância da pesquisa. O método do Conglomerado Primário com Linearização de Taylor foi substituído pelo método de replicação *bootstrap*.

[3] O detalhamento dos 77 recortes geográficos encontra-se no Anexo 1.

Método *bootstrap*

O *bootstrap* é um método de replicação, proposto incialmente por Efron (1979) para variáveis aleatórias independentes identicamente distribuídas. Posteriormente, esta técnica passou a ser utilizada para estimação de variância em diversas pesquisas.

O método foi adaptado para amostras estratificadas e conglomeradas por Rao e Wu (1988). Em 1992, Rao, Wu e Yue fizeram uma modificação, na qual utilizaram a replicação nos pesos amostrais ao invés de aplicar nos próprios dados da pesquisa.

O método consiste em selecionar B amostras, com reposição, da própria amostra disponível. Para cada uma dessas amostras, calcula-se as estimativas de interesse e, a partir destas, calcula-se uma estimativa da variância.

Considerando uma amostra complexa, estratificada e conglomerada em mais de 1 estágio, o método pode ser descrito da seguinte forma:

Considere o interesse em estimar a variância do estimador $\hat{\theta}$ de um parâmetro de interesse θ:

a) Seleciona-se uma amostra aleatória simples, com reposição, de $n_h - 1$ UPAs das n_h UPAS existentes no estrato h; cada UPA selecionada leva para a amostra *bootstrap* todos os domicílios e pessoas ali pesquisados;

b) Repita o passo (a) por um número grande de vezes, digamos B, e seja $m_{hi}(b)$ o número de vezes que a UPA i do estrato h, aparece na b-ésima amostra *bootstrap*;

c) O peso *bootstrap* para uma unidade k (domicílio ou pessoa) pertencente à UPA i do estrato h na b-ésima amostra é calculado, então, da seguinte maneira:

$$w_k^*(b) = \frac{w_k^* n_h m_{hi}(b)}{n_h - 1} \qquad (2)$$

d) Para cada amostra *bootstrap* b (= 1, ..., B) calcula-se uma estimativa $\hat{\theta}(b)$ usando os pesos *bootstrap* dados em (2);

e) A variância *bootstrap* é estimada por:

$$v_B = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}(b) - \bar{\theta}\right)^2 \qquad (3)$$

Onde $\bar{\theta} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}(b)$.

A quantidade B de pesos replicados depende das características de cada pesquisa e deve ser escolhida de acordo com a convergência. Após estudos, definiu-se o número de 200 pesos replicados para a PNAD Contínua.

Quando ajustes de não-resposta e calibração são feitos nos pesos originais da pesquisa, esses devem ser feitos em cada uma das amostras *bootstrap*. Como na PNAD Contínua, o ajuste de não resposta é uniforme dentro da mesma UPA, utilizou-se a replicação dos pesos já com o ajuste de não resposta e, para cada réplica, realizou-se o processo de calibração como descrito na seção II desta nota.

Todo o processo de geração dos pesos replicados também foi realizado utilizando a biblioteca *survey* do pacote estatístico R.

Cálculo da variância do estimador de diferença entre dois trimestres

Para ajudar na análise de mudanças na conjuntura econômica, calcula-se, na PNAD Contínua, o intervalo de confiança da diferença entre dois indicadores em dois trimestres. Esse intervalo de confiança permite avaliar se a diferença foi significativa (quando o intervalo é estritamente positivo ou negativo) ou não (quando o valor zero está incluído no intervalo).

Desta maneira, seja $\hat{\theta}_t$ o estimador para o trimestre t e $\hat{\theta}_{t-1}$ o estimador para o trimestre anterior. O intervalo de confiança da diferença seria dado por:

$$IC(\theta_t - \theta_{t-1}; 95\%) = \left(\hat{\theta}_t - \hat{\theta}_{t-1}\right) \mp 1{,}96 \times \sqrt{v\left(\hat{\theta}_t - \hat{\theta}_{t-1}\right)} \quad (4)$$

Onde:

$$v\left(\hat{\theta}_t - \hat{\theta}_{t-1}\right) = v\left(\hat{\theta}_t\right) + v\left(\hat{\theta}_{t-1}\right) - 2 \times cov\left(\hat{\theta}_t - \hat{\theta}_{t-1}\right) \quad (5)$$

Até a divulgação do trimestre móvel de junho-julho-agosto de 2021, as estimativas da variância e da covariância nas fórmulas anteriores eram calculadas utilizando linearização de Taylor, sempre considerando o método do conglomerado primário. A partir da divulgação do 3º Trimestre de 2021, essas estimativas de variâncias em (5) também serão calculadas pelo método de replicação *bootstrap*.

IV) Liberação de microdados atualizados com os novos fatores de expansão

Além dos resultados produzidos para as suas divulgações, o IBGE disponibiliza microdados da PNAD Contínua para uso público, dentro das suas possibilidades de utilização e devidamente tratados para garantir o sigilo das informações individualizadas, conforme estabelecido em Lei. Portanto, sem qualquer interferência desta instituição, os usuários podem utilizar os microdados que são disponibilizados para calcular os indicadores de interesse para seus estudos, da forma que julguem melhor para atender os seus objetivos, produzi-los utilizando a linguagem de programação que lhes for conveniente (SAS, R, Stata, Python, SPSS etc.) e analisá-los.

No dia 30 de novembro de 2021, serão disponibilizados os seguintes microdados da PNAD Contínua atualizados com os novos fatores de expansão (representados pelo peso ajustado por calibração além do conjunto de outros 200 pesos amostrais obtidos pela técnica de *bootstrap*, que permitem obter estimativas dos coeficientes de variação dos indicadores[4]), a saber:

[4] Os coeficientes de variação calculados pelo IBGE são obtidos a partir dos pesos replicados. Caso o usuário utilize o Método do Conglomerado Primário com linearização de Taylor, os valores obtidos dos coeficientes de variação podem divergir.

- Trimestrais: 1º trimestre de 2012 até 3º trimestre de 2021 (bancos utilizados nas divulgações conjunturais trimestrais);
- Anuais acumulados em determinada visita: 1ª visita de 2012 até 2019; 5ª visita de 2020 (bancos utilizados na divulgação temática de rendimento de todas as fontes);
- Anuais concentrados em determinado trimestre: 4º trimestre de 2016 até 2019 (bancos utilizados na divulgação temática de TIC). A partir de 01 de dezembro de 2021, a cada divulgação temática anual, os referidos microdados públicos, inclusive sua série histórica, com os novos fatores de expansão serão disponibilizados nos meios usuais de disseminação oficial.

Adicionalmente, o repositório[5] com as projeções populacionais anteriormente utilizadas pela pesquisa conterá todos os pesos anteriormente utilizados, do início da série até o último período divulgado com o método de ponderação anterior.

19 de novembro de 2021

Diretoria de Pesquisas

Referências:

Canty AJ, Davison AC. (1999) Resampling-based variance estimation for labour force surveys. The Statistician 48.

Deville, J.-C., and Särndal, C. E. (1992), "Calibration Estimators in Survey Sampling," Journal of the American Statistical Association, 87, pp 376-382.

Deville, J. C., Särndal, C. E., and Sautory, O. (1993), “Generalized raking procedures in survey sampling”, Journal of the American Statistical Association 88, pp. 1013-1020.

Office for National Statistics. (2011) Labour Force Survey – User Guide Volume 1 – LFS Background and Methodology 2011.

Preston J. (2009) Rescaled bootstrap for stratified multistage sampling. Survey Methodology 35(2) 227-234

Rao JNK, Wu CFJ. (1993) Bootstrap inference for sample surveys. Proc Section on Survey Research Methodology.

Rao JNK, Wu CFJ. (1988) Resampling Inference With Complex Survey Data. Journal of the American Statistical Association. Vol 83. N 401.

Rao JNK, Wu CFJ. and Yue K. (1992) Some recent work on resampling methods for complex surveys. Survey Methodology, 18, 209-217.

T. Lumley (2014) "survey: analysis of complex survey samples". R package version 3.30.

[5] Os pesos anteriormente usados nas pesquisas suplementares anuais estão disponíveis em: https://ftp.ibge.gov.br/Trabalho_e_Rendimento/Pesquisa_Nacional_por_Amostra_de_Domicilios_continua/Anual/Microdados/Projecoes_Anteriores/. Já os pesos anteriormente usados nos indicadores conjunturais trimestrais podem ser acessados em: https://ftp.ibge.gov.br/Trabalho_e_Rendimento/Pesquisa_Nacional_por_Amostra_de_Domicilios_continua/Trimestral/Microdados/Projecoes_Anteriores/.

T. Lumley (2004) “Analysis of complex survey samples”, Journal of Statistical Software 9(1), pp. 1-19.

Anexo 1: Descrição dos 77 pós-estratos geográficos utilizados nos ajustes dos fatores de expansão da PNAD Contínua

Pós-estrato	Descrição
1	Porto Velho
2	Resto de Rondônia (exceto capital)
3	Rio Branco
4	Resto do Acre (exceto capital)
5	Manaus
6	RM de Manaus (exceto Manaus)
7	Resto do Amazonas (exceto capital e RM)
8	Boa Vista
9	Resto de Roraima (exceto capital)
10	Belém
11	RM de Belém (exceto Belém)
12	Resto do Pará (exceto capital e RM)
13	Macapá
14	RM de Macapá (exceto Macapá)
15	Resto do Amapá (exceto capital e RM)
16	Palmas
17	Resto de Tocantins (exceto capital)
18	São Luís
19	RM da Grande São Luís (exceto São Luís)
20	RIDE de Teresina (Timon)
21	Resto do Maranhão (exceto capital, RM e Timon).
22	Teresina
23	RIDE de Teresina (exceto capital)
24	Resto do Piauí (exceto capital e RIDE)
25	Fortaleza
26	RM de Fortaleza (exceto Fortaleza)
27	Resto do Ceará (exceto capital e RM)
28	Natal
29	RM de Natal (exceto Natal)
30	Resto do Rio Grande do Norte (exceto capital e RM)

31	João Pessoa
32	RM de João Pessoa (exceto João Pessoa)
33	Resto da Paraíba (exceto capital e RM)
34	Recife
35	RM de Recife (exceto Recife)
36	Resto de Pernambuco (exceto capital e RM)
37	Maceió
38	RM de Maceió (exceto Maceió)
39	Resto de Alagoas (exceto capital e RM)
40	Aracaju
41	RM de Aracaju (exceto Aracaju)
42	Resto de Sergipe (exceto capital e RM)
43	Salvador
44	RM de Salvador (exceto Salvador)
45	Resto da Bahia (exceto capital e RM)
46	Belo Horizonte
47	RM de Belo Horizonte (exceto Belo Horizonte)
48	RIDE do Distrito Federal e Entorno
49	Resto de Minas Gerais (exceto capital e RM)
50	Vitória
51	RM da Grande Vitória (exceto Vitória)
52	Resto do Espírito Santo (exceto capital e RM)
53	Rio de Janeiro - capital
54	RM do Rio de Janeiro (exceto capital)
55	Resto do Rio de Janeiro (exceto capital e RM)
56	São Paulo - capital
57	RM de São Paulo (exceto capital)
58	Resto de São Paulo (exceto capital e RM)
59	Curitiba
60	RM de Curitiba (exceto Curitiba)
61	Resto do Paraná (exceto capital e RM)
62	Florianópolis
63	RM de Florianópolis (exceto Florianópolis)

64	Resto de Santa Catarina (exceto capital e RM)
65	Porto Alegre
66	RM de Porto Alegre (exceto Porto Alegre)
67	Resto do Rio Grande do Sul (exceto capital e RM)
68	Campo Grande
69	Resto do Mato Grosso do Sul (exceto capital)
70	Cuiabá
71	RM do Vale do Rio Cuiabá (exceto Cuiabá)
72	Resto do Mato Grosso (exceto capital e RM)
73	Goiânia
74	RM de Goiânia (exceto Goiânia)
75	RIDE do Distrito Federal e Entorno
76	Resto de Goiás (exceto capital e RM)
77	Brasília