

RELATÓRIO DE COLETA DE DADOS

PAINEL TIC COVID-19 – 4ª EDIÇÃO

1. INTRODUÇÃO

O Comitê Gestor da Internet no Brasil (CGI.br), por meio do Centro Regional de Estudos para o Desenvolvimento da Sociedade da Informação (Cetic.br), departamento do Núcleo de Informação e Coordenação do Ponto BR (NIC.br), apresenta a avaliação da coleta de dados e processos associados aos resultados da *Pesquisa on-line com usuários de Internet no Brasil – Painel TIC COVID-19 (4ª Edição)*.

2. INSTRUMENTOS DE COLETA

2.1. INFORMAÇÕES SOBRE OS INSTRUMENTOS DE COLETA

Os dados foram coletados por meio de questionários estruturados, com perguntas fechadas e respostas predefinidas (respostas únicas ou múltiplas). Foi utilizado um questionário *web* que requeria autopreenchimento, sem mediação de entrevistador.

2.2. TEMÁTICAS ABORDADAS

A pesquisa investigou atividades realizadas na Internet e dispositivos utilizados para acesso à rede, tendo como referência os indicadores validados pela pesquisa TIC Domicílios (Comitê Gestor da Internet no Brasil [CGI.br], 2020) e indicadores de uso relacionados aos temas:

- Cultura
- Comércio eletrônico
- Serviços públicos on-line
- Telessaúde
- Ensino remoto
- Teletrabalho

3. COLETA DE DADOS EM CAMPO

3.1. MÉTODO DE COLETA

Os dados foram coletados por meio de questionários estruturados. Foi utilizado o modo de coleta CAWI (do inglês, *computer-assisted web interviewing*), que utiliza um questionário programado e autoaplicado via questionário *on-line*.

3.2. PERÍODO DE COLETA

Os dados foram coletados entre 15 e 30 de julho de 2021.

3.3. AMOSTRA COLETADA

Ao todo 5.552 painelistas responderam à pesquisa de um total de 16.427 painelistas contatados.

4. PROCESSAMENTO DE DADOS

4.1. PROCEDIMENTOS DE PONDERAÇÃO

Para o Painel TIC COVID-19 – 4ª Edição foi utilizada como referência primária a pesquisa TIC Domicílios 2020. Adicionalmente, os resultados da TIC Domicílios foram atualizados para a população da Pesquisa Nacional por Amostra de Domicílios (Pnad) Contínua, do Instituto Brasileiro de Geografia e Estatística (IBGE), referente ao segundo trimestre de 2021. O processo de ponderação dos respondentes do Painel TIC COVID-19 – 4ª Edição é dividido em duas etapas:

1. Estimação do contingente total de usuários de Internet de 16 anos ou mais de idade no Brasil na data de referência da pesquisa que são representados pelos respondentes do Painel TIC COVID-19 – 4ª Edição;
2. Estimação de pseudoprobabilidades de seleção desses respondentes para ponderação do Painel TIC COVID-19 – 4ª Edição.

4.1.1. ETAPA 1 - Estimação do contingente de usuários de Internet representados no Painel TIC – 4ª Edição

A pesquisa TIC Domicílios 2020, a partir de uma abordagem probabilística tradicional, permite estimar o total de brasileiros de 10 anos ou mais de idade que são usuários de Internet¹. Já o Painel TIC COVID-19 – 4ª Edição conta com respondentes de 16 anos ou mais que são usuários de Internet, segundo parâmetros adotados internacionalmente (UIT, 2014). Para que as duas amostras sejam comparáveis, são filtrados os resultados da TIC Domicílios referentes a mesma faixa etária, aqueles usuários de Internet de 16 anos ou mais.

Uma vez que a construção do conjunto de respondentes do Painel TIC COVID-19 – 4ª Edição não é feita de forma probabilística, não é possível considerá-lo *a priori* como representativo do conjunto da população de usuários de Internet de 16 ou mais anos de idade. Para estimar o contingente da população que é representada pelos respondentes do painel, adotou-se o procedimento de estimação baseado em escores de propensão (*propensity scores*). Nessa metodologia são calculados, inicialmente, os escores de propensão de ser usuário de Internet segundo variáveis socioeconômicas com base na pesquisa probabilística TIC Domicílios 2020. A seguir, esse mesmo modelo é então utilizado para estimar os escores de propensão para os respondentes do Painel TIC COVID-19 – 4ª Edição.

¹ Maiores detalhes em no *website* do Cetic.br. Recuperado em 1 agosto, 2020, de http://cetic.br/media/microdados/256/ticdom_2019_relatorio_metodologico_v1.0.pdf.

Comparando a distribuição dos escores de propensão do Painel TIC COVID-19 – 4ª Edição com a verificada na pesquisa TIC Domicílios 2020 foi possível determinar qual parte (ou se toda) a população da pesquisa TIC Domicílios 2020 pode ser considerada representada pelos respondentes do painel. Isso equivale a estimar o erro de cobertura do Painel TIC COVID-19 – 4ª Edição em relação à população-alvo inicialmente considerada para a pesquisa.

A partir dessa comparação é estabelecido um ponto de corte que determina, na base da última pesquisa TIC Domicílios, o conjunto de unidades investigadas cujos escores de propensão parecem bem representados pelos respondentes do Painel TIC COVID-19 – 4ª Edição.

Esse procedimento teve por objetivo determinar a população que é representada pelo Painel TIC COVID-19 – 4ª Edição e considerar, para efeitos de comparação de resultados, essa mesma população entre os usuários de Internet na última pesquisa TIC Domicílios 2020.

O processo de determinação dessa população seguiu quatro passos:

- I. Atualização de totais de população da pesquisa TIC Domicílios 2020 para totais do segundo trimestre de 2021 da pesquisa Pnad Contínua realizada pelo IBGE;
- II. Ajuste de modelo de regressão logística tendo a variável “usuário de Internet” como variável resposta e um conjunto de variáveis socioeconômicas comuns a essa pesquisa e ao Painel TIC COVID-19 – 4ª Edição como variáveis explicativas. Esse modelo é então usado para estimar os escores de propensão a ser usuário de Internet para os respondentes da pesquisa TIC Domicílios 2020;
- III. Estimação dos escores de propensão para os respondentes da Painel TIC COVID-19 – 4ª Edição com base no modelo ajustado com os dados da pesquisa TIC Domicílios 2020;
- IV. Determinação do ponto de corte que separa tanto na amostra da pesquisa TIC Domicílios 2020 como no Painel TIC COVID-19 – 4ª Edição a parcela da população que está representada.

Passo I. Atualização dos totais populacionais da pesquisa TIC Domicílios 2020 para o primeiro trimestre de 2021

O objetivo desse passo foi atualizar as estimativas populacionais para a população de 10 anos ou mais de idade da pesquisa TIC Domicílios 2020, tendo como base dados divulgados pelo IBGE na última PNAD Contínua. Os cálculos atualizam o total da população de 10 anos ou mais de idade a partir das estimativas informadas nos microdados da Pnad Contínua. Em seguida, e seguindo a mesma distribuição percentual dos calibradores utilizados na última pesquisa TIC Domicílios 2020, é refeita a atualização dos pesos da pesquisa segundo os novos totais das distribuições marginais das variáveis consideradas na calibração.²

² Maiores detalhes sobre as variáveis de calibração de usuários podem ser obtidos no capítulo “Relatório Metodológico” da pesquisa TIC Domicílios. Recuperado em 3 fevereiro, 2022, de https://cetic.br/media/microdados/604/tic_dom%C3%ADlios_2020_relatorio_metodologico_v1.0.pdf

Passo II. Ajuste do modelo de regressão logística para a variável “usuário de Internet” entre os respondentes de 16 ou mais anos de idade na TIC Domicílios

Essa etapa busca estimar com qualidade a probabilidade de um indivíduo ser usuário de Internet a partir de variáveis socioeconômicas observadas na última pesquisa TIC Domicílios e que também estão disponíveis no Painel TIC COVID-19 – 4ª Edição. Com o objetivo de obter um modelo parcimonioso e que desse bons resultados na estimação de usuários de Internet foram testados diversos modelos da forma:

$$\log \left(\frac{P(Y_i = 1)}{1 - P(Y_i = 1)} \right) = \alpha + \beta X_i$$

Onde:

Y_i é uma variável indicadora, tomando valor 1 se o indivíduo i é usuário de Internet, e valor 0, caso contrário;
 X_i é um vetor com os valores de variáveis explicativas (sexo, faixa etária, escolaridade etc.) do indivíduo i ,
 $P(Y_i = 1)$ representa a probabilidade do indivíduo ser usuário de Internet, e
 α e β são parâmetros do modelo, a serem estimados.

As estimativas para $P(Y_i = 1)$ fornecidas pela expressão

$$\hat{P}(Y_i = 1) = \frac{\exp(\hat{\alpha} + \hat{\beta} X_i)}{1 + \exp(\hat{\alpha} + \hat{\beta} X_i)}$$

são os chamados escores de propensão considerados na metodologia, sendo que $\hat{\alpha}$ e $\hat{\beta}$ são as estimativas dos parâmetros obtidas com base no modelo ajustado.

O modelo ajustado utiliza como opções de variáveis independentes (X) apenas informações que estejam presentes em ambas as fontes: TIC Domicílios e Painel TIC COVID-19 – 4ª Edição. O modelo final mais parcimonioso e com grande grau de acerto na previsão de quais indivíduos são usuários da Internet incluiu as seguintes variáveis: sexo, idade, grau de instrução, classe social e indicador de uso de Internet no computador³. A Tabela 1 apresenta os resultados de ajuste.

³ Utilizou-se como modelo de partida o modelo final utilizado nas pesquisas Painel TIC COVID-19 – 4ª Edição de 2020. A variável classe social não se mostrou significativa para a pesquisa TIC Domicílios 2020, mas optou-se por manter essa variável no modelo dada a relevância temática dessa informação.

TABELA 1
ESTATÍSTICAS DE AJUSTE DO MODELO

Variáveis independentes no modelo	TIC Domicílios 2020	
	R ²	Taxa de classificação correta ⁽¹⁾
Sexo, Idade, Grau de instrução, Classe social, Indicador de usuário de computador	0,389	87%

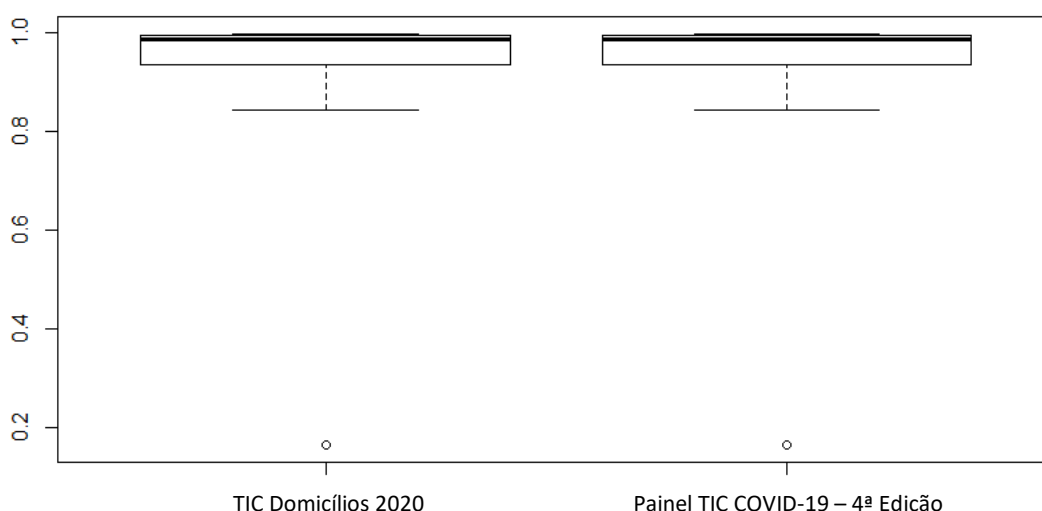
Fonte: CGI.br, TIC Domicílios 2020.

(1) = Percentual de indivíduos classificados corretamente com base no modelo ajustado.

Passo III. Estimação dos escores de propensão para os respondentes do Painel TIC COVID-19 – 4ª Edição

A partir do modelo ajustado com os dados da pesquisa TIC Domicílios 2020, foram estimados os escores de propensão para o conjunto de respondentes do Painel TIC COVID-19 – 4ª Edição. Em seguida, foi feita a comparação das distribuições dos escores de propensão na amostra da TIC Domicílios 2020 com os escores da amostra do Painel TIC COVID-19 – 4ª Edição para os usuários de Internet. Os resultados são apresentados na Figura 1. É possível notar que a distribuição dos escores dos respondentes do Painel TIC COVID-19 – 4ª Edição tem um perfil similar ao observado para a população usuária de Internet de 16 anos ou mais segundo a TIC Domicílios 2020.

FIGURA 1: DISTRIBUIÇÃO DOS ESCORES DE PROPENSÃO A SER USUÁRIO DE INTERNET



Fonte: CGI.br, TIC Domicílios 2020 e Painel TIC COVID-19 – 4ª Edição.

Passo IV. Determinação de população de suporte comum das pesquisas TIC Domicílios 2020 e Painel TIC COVID-19 – 4ª Edição

Dado que as distribuições dos escores obtidos em ambas as pesquisas são bastante similares, os respondentes da pesquisa Painel TIC COVID-19 – 4ª Edição foi considerada integralmente como representativa da população de usuários de Internet com 16 anos ou mais de idade no Brasil.

4.1.2. ETAPA 2 - Estimação de pseudoprobabilidades de inclusão para determinação de pesos dos respondentes do Painel TIC COVID-19 – 4ª Edição

O processo de estimação de pseudopesos consiste na estimação de pseudoprobabilidades de inclusão dos respondentes do Painel TIC COVID-19 – 4ª Edição (amostra não probabilística) na pesquisa TIC Domicílios 2020 (amostra probabilística), e usar seus recíprocos como pesos, tal como em uma pesquisa por amostragem probabilística tradicional. Com isso, estima-se a probabilidade de um indivíduo ser selecionado e responder à pesquisa TIC Domicílios 2019 com base em variáveis independentes (**X**) relacionadas ao perfil dos entrevistados, considerando que, dadas essas variáveis (**X**), as probabilidades de inclusão são independentes das variáveis de interesse da pesquisa⁴.

O modelo mais parcimonioso considerando as variáveis independentes (**X**) disponíveis e comuns às duas bases de dados contém as seguintes variáveis: macrorregião, classe social, indicador de uso de Internet no computador, escolaridade, faixa etária, sexo e número de moradores do domicílio. A partir desse modelo, foram estimadas as pseudoprobabilidades de inclusão dos respondentes do Painel TIC COVID-19 – 4ª Edição na pesquisa TIC Domicílios 2020. Os recíprocos dessas pseudoprobabilidades são os pesos iniciais alocados para cada respondente do Painel TIC COVID-19 – 4ª Edição.

Esses pesos iniciais foram calibrados para totais marginais estimados das variáveis macrorregião, sexo; faixa etária e escolaridade. Os pesos assim calibrados foram considerados para a estimação de todos os indicadores de resultados de interesse e das medidas de precisão associadas.

4.2. ESTIMAÇÃO DE VARIÂNCIA

O processo de estimação atribuiu a cada respondente do Painel TIC COVID-19 – 4ª Edição um peso que o trata como se fosse um participante de pesquisa com plano amostral igual ao da pesquisa TIC Domicílios 2020, mas com tamanho total da amostra menor. Dessa forma, é possível estimar variâncias e margens de erro. O método utilizado para a estimação de erros foi o método de replicação⁴.

5. DISSEMINAÇÃO DOS DADOS

Os resultados do Painel TIC COVID-19 – 4ª Edição são apresentados de acordo com as variáveis de classificação descritas no item “Domínios de Interesse para Análise e Divulgação”. Arredondamentos fazem com que, em alguns

⁴ Maiores detalhes no Relatório Metodológico Painel TIC - Edição 1.

resultados, a soma das categorias parciais difira de 100% em questões de resposta única. O somatório de frequências em questões de respostas múltiplas usualmente é diferente de 100%. Vale ressaltar que, nas tabelas de resultados, o hífen (–) é utilizado para representar a não resposta ao item. Por outro lado, como os resultados são apresentados sem casa decimal, as células com valor zero significam que houve resposta ao item, mas ele é explicitamente maior do que zero e menor do que um por cento.

Os resultados são publicados em relatório *on-line* e disponibilizados no *site* do Cetic.br (<http://www.cetic.br>). As tabelas com estimativas de totais e margens de erro calculadas para cada indicador estão disponíveis para *download* no *website* do Cetic.br.

REFERÊNCIAS

Comitê Gestor da Internet no Brasil – CGI.br. (no prelo). *Pesquisa sobre o uso das tecnologias de informação e comunicação nos domicílios brasileiros: TIC Domicílios 2019*. São Paulo: CGI.br.

Dever, J. A. (2018). Combining probability and nonprobability samples to form efficient hybrid estimates: An evaluation of the common support assumption. *Proceedings of the 2018 Federal Committee on Statistical Methodology (FCSM) Research Conference*, Washington, Estados Unidos, 15.

Elliott, M. R. (2009). Combining data from probability and non-probability samples using pseudo-weights. *Survey Practice*, 2(6), 1–7.

Elliott, M. R., & Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32(2), 249–64.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. *Wiley Series in Probability and Statistics*.

União Internacional de Telecomunicações – UIT. (2014). *Manual for measuring ICT access and use by households and individuals 2014*. Recuperado em 1 agosto, 2020, de http://www.itu.int/dms_pub/itu-d/opb/ind/D-IND-ITCMEAS-2014-PDF-E.pdf

Valliant, R. (2019). Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology*, 8(2), 231–263.

Valliant, R., & Dever, J. A. (2011). *Estimating propensity adjustments for volunteer web surveys*. *Sociological Methods and Research*, 40(1), 105–137.