

# RELATÓRIO METODOLÓGICO

## 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 metodologia da *Pesquisa on-line com usuários de Internet no Brasil – Painel TIC COVID-19 (4ª Edição)*.

A partir da pandemia COVID-19, que afetou de forma substancial o trabalho dos institutos nacionais de estatística e demais produtores de dados em todo o mundo e, particularmente, entre os países da América Latina, o Cetic.br inovou na produção de indicadores a partir de pesquisa em painéis web.

Essa experiência se mostrou bastante relevante, e a partir de 2021 esse painel passa a ser uma nova forma de investigação utilizada pelo Cetic.br na produção de estatísticas TIC para usuários de Internet. Essa nova ferramenta de investigação vai permitir um acompanhamento mais regular de indicadores de conjuntura, bem como avaliar e coletar informação sobre novos temas e aspectos do uso de Internet no Brasil.

### 2. OBJETIVOS DA PESQUISA

O Painel TIC tem como objetivo coletar informações sobre o uso da Internet por usuários de Internet em todo o Brasil.

### 3. POPULAÇÃO-ALVO

A população-alvo da pesquisa é composta por indivíduos usuários de Internet de 16 anos ou mais de idade no Brasil. São considerados usuários de Internet os indivíduos que fizeram uso da rede nos três meses que antecedem a entrevista, segundo recomendação metodológica da União Internacional de Telecomunicações (UIT, 2014).

### 4. UNIDADE DE ANÁLISE E REFERÊNCIA

Indivíduos usuários de Internet com 16 anos ou mais de idade.

## 5. DOMÍNIOS DE INTERESSE PARA ANÁLISE E DIVULGAÇÃO

Para as unidades de análise e referência, os resultados são divulgados para domínios definidos com base nas variáveis e níveis descritos a seguir.

- **Sexo:** Corresponde à divisão em Masculino e Feminino;
- **Grau de instrução:** Corresponde à divisão em Ensino Fundamental, Ensino Médio e Ensino Superior;<sup>1</sup>
- **Faixa etária:** Corresponde à divisão das faixas de 16 a 24 anos, de 25 a 34 anos, de 35 a 44 anos, de 45 a 59 anos e de 60 anos ou mais;
- **Região:** Corresponde à divisão regional do Brasil, segundo critérios do IBGE, nas macrorregiões Norte, Nordeste, Sudeste, Sul e Centro-Oeste;
- **Classe social:** Corresponde à divisão em AB, C e DE, conforme o Critério de Classificação Econômica Brasil (CCEB), da Associação Brasileira de Empresas de Pesquisa (Abep).

## 6. INSTRUMENTOS DE COLETA

### 6.1. INFORMAÇÕES SOBRE OS INSTRUMENTOS DE COLETA

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

### 6.2. TEMÁTICAS ABORDADAS

A pesquisa investiga 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 a temas relevantes à época da pesquisa.

## 7. PLANO AMOSTRAL

### 7.1. CADASTROS E FONTES DE INFORMAÇÃO

Para o desenho amostral do Painel TIC é utilizado como fonte primária um painel de respondentes de empresa de pesquisa de mercado, que conta com aproximadamente 167 mil painelistas de 16 anos ou mais de idade. O recrutamento dos participantes no painel se dá por uma série de canais e métodos, entre eles: pesquisas

---

<sup>1</sup> Os graus de instrução apresentados referem-se à agregação dos níveis de escolaridade declarados que são iguais ou inferiores às categorias apresentadas, isto é, no grau Ensino Fundamental estão incluídas as escolaridades: sem instrução, pré-escola e qualquer nível de escolaridade do Ensino Fundamental completo ou incompleto. Isso também se aplica ao Ensino Médio e Superior.

probabilísticas, escolha criteriosa de parceiros de recrutamento e parcerias com veículos de comunicação e mídia, avaliação contínua da taxa de resposta dos painelistas, foco em ações de recrutamento para públicos específicos conforme as necessidades dos clientes, processo de recrutamento em conformidade com os mais altos padrões de mercado. Além disso, é importante mencionar que os participantes dos painéis recebem incentivos para responderem às pesquisas.

## 7.2. DIMENSIONAMENTO DA AMOSTRA

A amostra é dimensionada de acordo com as necessidades de informação, recursos disponíveis e o prazo em que se deseja ter a informação para análise. O total de entrevistas de cada onda do Painel TIC é apresentado no Relatório de Coleta de Dados do Painel TIC.

## 7.3. MÉTODOS PARA OBTENÇÃO DA AMOSTRA

O plano amostral empregado para a obtenção da amostra de respondentes é do tipo amostragem de cotas. As cotas são estabelecidas considerando sexo, faixa etária, escolaridade, macrorregião e classe social, e são utilizadas para indicar os indivíduos a serem abordados para coleta pela *web*. A alocação da amostra segundo os critérios estabelecidos é desproporcional às informações constantes no cadastro, dada a necessidade de atender à demanda por informações para todos os domínios de interesse. A amostra resultante deste esforço de coleta é daqui por diante denominada Painel TIC.

## 8. COLETA DE DADOS EM CAMPO

### 8.1. MÉTODO DE COLETA

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

## 9. PROCESSAMENTO DE DADOS

### 9.1. PROCEDIMENTOS DE PONDERAÇÃO

Pesquisas amostrais com utilização de cotas para seleção de respondentes são classificadas como não probabilísticas. Tipicamente, tais estratégias não permitem o cálculo de erros amostrais e podem carregar alguns vieses de seleção, na medida em que as probabilidades de seleção de cada unidade não são conhecidas. Abordagens não probabilísticas são usuais em pesquisas de opinião, de intenção de voto, de avaliação de produtos e de satisfação de clientes. Tais pesquisas contam, em geral, com períodos de coleta mais curtos e com menores orçamentos, mas não seguem o rigor habitual dos métodos de amostragem probabilística para obtenção das amostras.

Recentemente, a crescente demanda por informações mais frequentes e desagregadas, além da emergência de novas fontes de informação (tipo *Big Data*), tem impulsionado inúmeros estudos que tentam atribuir estruturas de pesos que permitam amenizar os vieses de bases de dados coletadas por métodos não tradicionais. Em geral, tais estudos utilizam uma pesquisa amostral ou censo tradicional como referência para o cálculo de pesos para as observações da amostra não probabilística, que então servem de base para a obtenção de estimativas da precisão, intervalos de confiança etc. Como exemplos de estudos nessa linha podem ser citados Elliott e Valliant (2017) e Valliant (2019).

Para o Painel TIC é utilizada como referência primária a última pesquisa probabilística – TIC Domicílios (CGI.br, 2020). Adicionalmente, os resultados da TIC Domicílios são atualizados para a população da Pesquisa Nacional por Amostra de Domicílios (Pnad) Contínua, do Instituto Brasileiro de Geografia e Estatística (IBGE), referente ao último trimestre divulgado. O processo de ponderação dos respondentes do Painel TIC 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;
2. Estimação de pseudoprobabilidades de seleção desses respondentes para ponderação do Painel TIC.

#### 9.1.1. ETAPA 1 - Estimação do contingente de usuários de Internet representados no Painel TIC

A pesquisa TIC Domicílios (último resultado disponível), a partir de uma abordagem probabilística tradicional, permite estimar o total de brasileiros de 10 anos ou mais de idade que são usuários de Internet<sup>2</sup>. Já o Painel TIC 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 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 última pesquisa probabilística TIC Domicílios disponível. A seguir, esse mesmo modelo é então utilizado para estimar os escores de propensão para os respondentes do Painel TIC.

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

---

<sup>2</sup> 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](http://cetic.br/media/microdados/256/ticdom_2019_relatorio_metodologico_v1.0.pdf).

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.

Esse procedimento tem por objetivo determinar a população que é representada pelo Painel TIC e considerar, para efeitos de comparação de resultados, essa mesma população entre os usuários de Internet na última pesquisa TIC Domicílios.

O processo de determinação dessa população segue quatro passos:

- I. Atualização de totais de população da última pesquisa TIC Domicílios para totais do último trimestre divulgado pela 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 como variáveis explicativas. Esse modelo é então usado para estimar os escores de propensão a ser usuário de Internet para os respondentes da última pesquisa TIC Domicílios;
- III. Estimação dos escores de propensão para os respondentes da Painel TIC com base no modelo ajustado com os dados da última pesquisa TIC Domicílios;
- IV. Determinação do ponto de corte que separa tanto na amostra da última pesquisa TIC Domicílios como no Painel TIC a parcela da população que estaria representada.

*Passo I. Atualização dos totais populacionais da última pesquisa TIC Domicílios para o trimestre mais recentemente divulgado pela Pnad Contínua*

O objetivo desse passo foi atualizar as estimativas populacionais para a população de 10 anos ou mais de idade da última pesquisa TIC Domicílios, 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, é refeita a atualização dos pesos da pesquisa segundo os novos totais das distribuições marginais das variáveis consideradas na calibração.<sup>3</sup>

*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. 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:

---

<sup>3</sup> Maiores detalhes sobre as variáveis de calibração de usuários podem ser obtidos no capítulo “Relatório Metodológico” da pesquisa TIC Domicílios. Recuperado em 1 agosto, 2020, de [http://cetic.br/media/microdados/256/ticdom\\_2019\\_relatorio\\_metodologico\\_v1.0.pdf](http://cetic.br/media/microdados/256/ticdom_2019_relatorio_metodologico_v1.0.pdf)

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

Onde:

$Y_i$  é uma variável indicadora, tomando valor 1 se o indivíduo  $i$  é usuário de Internet, e valor 0, caso contrário;

$\mathbf{X}_i$  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

$\alpha$  e  $\beta$  são parâmetros do modelo, a serem estimados.

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

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

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 ( $\mathbf{X}$ ) apenas informações que estejam presentes em ambas as fontes: TIC Domicílios e Painel TIC. O modelo final é apresentado no Relatório de Coleta de dados do Painel TIC.

### *Passo III. Estimação dos escores de propensão para os respondentes do Painel*

A partir do modelo ajustado com os dados da última pesquisa TIC Domicílios, são estimados os escores de propensão para o conjunto de respondentes do Painel TIC. Em seguida, é feita a comparação das distribuições dos escores de propensão na amostra da TIC Domicílios com os escores da amostra do Painel TIC para os usuários de Internet. Os resultados são apresentados no Relatório de Coleta de Dados do Painel TIC.

### *Passo IV. Determinação de população de suporte comum das pesquisas TIC Domicílios e Painel TIC*

Se as distribuições dos escores obtidos em ambas as pesquisas forem diferentes, busca-se identificar um recorte da amostra de usuários de Internet da TIC Domicílios que seja mais parecido com o conjunto de respondentes do Painel TIC. A escolha deste recorte leva em conta a observação das distribuições dos escores e a variabilidade em pesos que são atribuídos aos respondentes do painel, para um conjunto de possíveis recortes de escores de propensão a ser usuário de Internet. Essa avaliação é feita estimando-se os pesos dos respondentes do Painel TIC segundo situações alternativas:

- I. Seleção de todos os respondentes de ambas as pesquisas, sem recorte; e

- II. Seleção dos respondentes de ambas as pesquisas que têm escores de propensão maior ou igual a uma fração específica.

A fração é escolhida de forma a utilizar “pedaços” do conjunto de respondentes de ambas as pesquisas que sejam comparáveis, o que determina uma população suporte comum em ambas as pesquisas.

Para cada opção de fração – determinadora de população suporte comum, são estimados pseudopesos para os respondentes do Painel TIC<sup>4</sup>. Os recortes considerados são avaliados segundo a variabilidade nos pesos. Opta-se pelo recorte em que os pesos resultantes têm a menor amplitude na distribuição dos valores absolutos, assim como os fatores de calibração (razão entre os pesos calibrados e os pesos básicos) possuem média mais próxima a 1. Isso é desejável, já que nessa situação a calibração mantém os pesos calibrados mais próximos dos pesos inicialmente estabelecidos pela metodologia de estimação de pseudopesos. Os resultados dessa etapa são apresentados no Relatório de Coleta de Dados do Painel TIC.

#### 9.1.2. ETAPA 2 - Estimação de pseudoprobabilidades de inclusão para determinação de pesos dos respondentes do Painel TIC

O processo de estimação de pseudopesos consiste na estimação de pseudoprobabilidades de inclusão dos respondentes do Painel TIC (amostra não probabilística) na última pesquisa TIC Domicílios (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 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 pesquisa.

Para estimar as pseudoprobabilidades, os dados de ambas as amostras (probabilística e não probabilística) são empilhados em uma única base de dados, e as probabilidades de inclusão são estimadas por meio de um modelo de regressão logística que leva em consideração o plano amostral da pesquisa probabilística de referência.

Para esse estudo foram são consideradas diferentes possibilidades, conforme os recortes de população estabelecidos na seção anterior. Tais recortes visam identificar a população suporte comum das duas pesquisas avaliando os pesos obtidos, como sugere Valliant (2019).

O processo de estimação de pseudoprobabilidades emprega os seguintes passos:

- I. União dos casos em uma mesma base de dados (empilhamento), garantindo a presença de variáveis independentes comuns (**X**), coletadas segundo os mesmos critérios e conceitos. Nesta base, foi criada uma variável indicadora **Z**, que assume o valor 1 para respondentes do Painel TIC (amostra não probabilística) e o valor 0 para respondentes da TIC Domicílios (amostra probabilística);
- II. Criação de uma coluna de pesos neste arquivo, a qual considera os pesos provenientes da amostra probabilística (para os seus casos) e peso igual a 1 para os casos da amostra não probabilística;

---

<sup>4</sup> A metodologia para a estimação dos pseudopesos é apresentada na próxima seção.

- III. Ajuste de um modelo de regressão logística tendo a variável **Z** como resposta, levando em consideração o desenho amostral da pesquisa TIC Domicílios, para estimar as probabilidades de inclusão dos respondentes do Painel TIC na amostra probabilística.

No ajuste do modelo, a amostra do Painel TIC é considerada como um estrato à parte, e cada respondente dessa amostra é considerado como sendo uma unidade primária de amostragem (UPA) distinta. Esse procedimento é necessário na declaração das variáveis de estrutura do plano amostral para o arquivo de dados empilhados das duas pesquisas.

O modelo mais parcimonioso considerando as variáveis independentes (**X**) disponíveis e comuns às duas bases de dados é apresentado no Relatório de Coleta de Dados da pesquisa Painel TIC. A partir desse modelo, são estimadas as pseudoprobabilidades de inclusão dos respondentes do Painel TIC na última pesquisa TIC Domicílios. Os recíprocos dessas pseudoprobabilidades são os pesos iniciais alocados para cada respondente do Painel TIC.

Esses pesos iniciais são calibrados para totais marginais estimados das variáveis da pesquisa TIC Domicílios. Os pesos assim calibrados são considerados para a estimação de todos os indicadores de resultados de interesse e das medidas de precisão associadas.

## 9.2. ESTIMAÇÃO DE VARIÂNCIA

O processo de estimação atribuiu a cada respondente do Painel TIC um peso que o trata como se fosse um participante de pesquisa com plano amostral igual ao da última pesquisa TIC Domicílios, mas com tamanho total da amostra menor. Dessa forma, é possível estimar variâncias e margens de erro. Segundo Valliant (2019), são duas as possibilidades para a estimação de variância: estimação considerando a amostra como aleatória simples com reposição ou estimação com base em método de replicação.

O segundo método (estimação com base em método de replicação) tem a vantagem de considerar a estimação do modelo para estimação das pseudoprobabilidades de inclusão para subamostras retiradas da amostra principal. Isso permite incluir na estimação da variância a variabilidade associada à estimação desse modelo. Por esta razão esse foi o método utilizado para estimação de variâncias. O procedimento seguiu os seguintes passos:

- I. A partir da base comum (empilhada) utilizada para a estimação do modelo de pseudoprobabilidades, são selecionadas 200 amostras *bootstrap* com uso da função *as.svrepdesign* do pacote *survey* do programa R, considerando o plano amostral;
- II. Para cada uma destas 200 réplicas, é ajustado o modelo para estimação de pseudoprobabilidades de inclusão, e correspondentes pseudopesos; e
- III. Os pseudopesos de cada réplica são calibrados e guardados para estimação da variância.

A variância de estimativas de indicadores de interesse é estimada usando:

$$\hat{V}(\hat{y}) = \frac{R-1}{R} \sum_{r=1}^R (\hat{y}_r - \hat{y})^2,$$

Onde:

$\hat{y}$  é a estimativa do indicador  $y$  obtida usando a amostra do Painel TIC COVID-19 (com 2.511 respondentes);

$\hat{y}_r$  é a estimativa do indicador  $y$  na réplica  $r$ ;

$R = 200$  é o total de réplicas *bootstrap* formadas.

## 10. DISSEMINAÇÃO DOS DADOS

Os resultados do Painel TIC 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 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. Para efeito de comparação com edições anteriores da pesquisa TIC Domicílios, são disponibilizadas as tabelas da pesquisa considerando o mesmo recorte utilizado no Painel TIC separando na pesquisa TIC Domicílios a população de suporte comum – quando necessário.

## 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](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.