RELATÓRIO METODOLÓGICO
PAINEL TIC COVID-19 – Edição 2

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 do Painel TIC COVID-19: Pesquisa sobre o uso da Internet no Brasil durante a pandemia do novo coronavírus.

A pandemia COVID-19 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. A realização de pesquisas domiciliares presenciais tem sido diretamente impactada pelas medidas de distanciamento social e pela necessidade de preservar a saúde de entrevistadores e informantes, atendendo a recomendações da Organização Mundial de Saúde (OMS).

Diante das limitações no atual momento para a coleta de dados por métodos tradicionais que requeriam entrevistas presenciais, o Cetic.br/NIC.br implementou um estudo piloto com usuários de Internet por meio de questionários *on-line*, de forma a acelerar o conhecimento a respeito de estratégias alternativas de coleta e obtenção de informação de qualidade sobre o acesso e uso das TIC durante a pandemia.

2 - OBJETIVOS DA PESQUISA

O Painel TIC COVID-19 tem como objetivo coletar informações sobre o uso da Internet durante a pandemia causada pelo novo coronavírus.

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;

• **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).[1]

6 - INSTRUMENTOS DE COLETA

6.1 - 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). Foram utilizados dois instrumentos para coleta dos dados que continham as mesmas perguntas: questionário *web* e questionário pelo telefone – CATI (do inglês, *computer-assisted telephone interviewing*). O questionário *web* requeria autopreenchimento, sem mediação de entrevistador. O questionário CATI foi aplicado com mediação de entrevistadores devidamente treinados.

6.2 - TEMÁTICAS ABORDADAS

Planejada para ser realizada e divulgada em três edições, 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).

[1] A Abep utiliza para tal classificação a posse de alguns itens duráveis de consumo doméstico, mais o grau de instrução do chefe do domicílio declarado. A posse dos itens estabelece um sistema de pontuação em que a soma para cada domicílio resulta na classificação como classes econômicas A1, A2, B1, B2, C, D e E. O Critério Brasil foi atualizado em 2015, resultando em classificação não comparável à anteriormente vigente (Critério Brasil 2008). Para os resultados divulgados a partir de 2016, foi adotado o Critério Brasil de 2015.

Além disso, foram criados módulos temáticos para aprofundar e detalhar aspectos do uso da rede relacionados ao contexto de enfrentamento da pandemia da COVID-19 e seus efeitos na sociedade. Para tanto, o Painel TIC COVID-19 inclui indicadores referentes aos seguintes temas:

1ª EDIÇÃO	Cultura Comércio eletrônico
2ª EDIÇÃO	Serviços públicos on-line Privacidade Telessaúde
3ª EDIÇÃO	Ensino remoto Trabalho remoto

7 - PLANO AMOSTRAL

7.1 - CADASTROS E FONTES DE INFORMAÇÃO

Para o desenho amostral do Painel TIC COVID-19 foi utilizado como fonte primária o Painel Conectaí, mantido pelo IBOPE Inteligência, que conta com aproximadamente 95 mil painelistas de 16 anos ou mais de idade. Para complementar as entrevistas obtidas foram contactados painelistas de outras empresas parceiras do IBOPE Inteligência, como a Offerwise. O recrutamento dos participantes nos painéis se dá por uma série de canais e métodos, 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.

Para além dos painéis, foram realizadas entrevistas complementares por meio de abordagem telefônica, para contemplar segmentos populacionais mais raros nos painéis on-line e que compõem a população da pesquisa. O cadastro utilizado foi obtido junto a empresas especializadas, que seguem critérios de privacidade e orientações da Abep e da World Research Association (ESOMAR). Além disso, estão em conformidade com a norma internacional de qualidade em pesquisa de mercado e opinião (ISO 20.252) e a norma internacional de gestão de qualidade (ISO 9001).

7.2 - DIMENSIONAMENTO DA AMOSTRA

A amostra foi dimensionada em 2.600 entrevistas considerando a otimização de recursos e capacidade de coleta em curto período.

7.3 - MÉTODOS PARA OBTENÇÃO DA AMOSTRA

O plano amostral empregado para a obtenção da amostra de respondentes foi do tipo amostragem de cotas. As cotas foram estabelecidas considerando sexo, faixa etária, escolaridade, macrorregião e classe social, e foram aplicadas para indicar os indivíduos a serem abordados para coleta pela *web* e para os contatados por meio telefônico. A alocação da amostra segundo os critérios estabelecidos foi 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 COVID-19.

8 - COLETA DE DADOS EM CAMPO

8.1 - MÉTODO DE COLETA

Os dados foram coletados por meio de questionários estruturados. Foram utilizados dois modos de coleta: CATI (do inglês, *computer-assisted telephone interviewing*), que consiste na aplicação de questionário estruturado e programado em computador por meio telefônico, e 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 COVID-19 foi utilizada como referência primária a pesquisa TIC Domicílios 2019 (CGI.br, 2020). Adicionalmente, os resultados da TIC Domicílios foram recalibrados para a população da Pesquisa Nacional por Amostra de Domicílios (Pnad) Contínua, do Instituto Brasileiro de Geografia e Estatística (IBGE), referente ao primeiro trimestre de 2020. O processo de ponderação dos respondentes do Painel TIC COVID-19 foi 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;
2. Estimação de pseudo-probabilidades de seleção desses respondentes para ponderação do Painel TIC COVID-19.

9.1.1 - ETAPA 1 - Estimação do contingente de usuários de Internet representados no Painel TIC COVID-19

A pesquisa TIC Domicílios 2019, 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[2]. Já o Painel TIC COVID-19 conta com respondentes de 16 anos ou mais que são usuários de Internet, segundo parâmetros adotados internacionalmente (UIT, 2014). Para que as duas amostras fossem comparáveis, foram filtrados os resultados da TIC Domicílios 2019 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 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 disponível (TIC Domicílios 2019). A seguir, esse mesmo modelo é então utilizado para estimar os escores de propensão para os respondentes do Painel TIC COVID-19.

Comparando a distribuição dos escores de propensão do Painel TIC COVID-19 com a verificada na pesquisa TIC Domicílios 2019 é possível determinar qual parte (ou se toda) a população da pesquisa em 2019 poderia ser considerada representada pelos respondentes do painel. Isso equivale a estimar o erro de cobertura do Painel TIC COVID-19 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 pesquisa TIC Domicílios 2019, o conjunto de unidades investigadas cujos escores de propensão parecem bem representados pelos respondentes do Painel TIC COVID-19.

[2] 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.

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

O processo de determinação dessa população seguiu quatro passos:

I. Atualização de totais de população da pesquisa TIC Domicílios 2019 para o primeiro trimestre de 2020, mediante recalibração dos pesos amostrais dessa pesquisa com base em informações da Pnad Contínua referentes ao primeiro trimestre de 2020 disponibilizadas 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 como variáveis explicativas. Esse modelo foi então usado para estimar os escores de propensão a ser usuário de Internet para os respondentes da pesquisa TIC Domicílios 2019;
III. Estimação dos escores de propensão para os respondentes da Painel TIC COVID-19 com base no modelo ajustado com os dados da pesquisa TIC Domicílios 2019;
IV. Determinação do ponto de corte que separou tanto na amostra da pesquisa TIC Domicílios 2019 como no Painel TIC COVID-19 a parcela da população que estaria representada.

Passo I. Atualização dos totais populacionais da pesquisa TIC Domicílios 2019 para o primeiro trimestre de 2020

O objetivo desse passo foi atualizar as estimativas populacionais para a população de 10 anos ou mais de idade da pesquisa TID Domicílios 2019, tendo como base dados divulgados pelo IBGE no primeiro trimestre de 2020. Os cálculos atualizaram o total da população de 10 anos ou mais de idade a partir das estimativas informadas nos microdados da Pnad Contínua do primeiro trimestre de 2020. Em seguida, e seguindo a mesma distribuição percentual dos calibradores utilizados na pesquisa TIC Domicílios 2019, foi refeita a calibração dos pesos da pesquisa segundo os novos totais das distribuições marginais das variáveis consideradas na calibração.[3]

[3] 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

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 2019

Essa etapa buscou estimar com qualidade a probabilidade de um indivíduo ser usuário de Internet a partir de variáveis socioeconômicas observadas na pesquisa TIC Domicílios 2019 e que também estão disponíveis no Painel TIC COVID-19. 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:

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

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

α e $\boldsymbol{\beta}$ são parâmetros do modelo, a serem estimados.

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

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

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

O modelo ajustado utilizou como opções de variáveis independentes (**X**) apenas informações que estivessem presentes em ambas as fontes: TIC Domicílios e Painel TIC COVID-19. O modelo final mais parcimonioso e com grande grau de acerto na previsão de quais indivíduos eram usuários da Internet incluiu as seguintes variáveis: sexo, faixa etária, grau de instrução, classe social e indicador de uso de computador[4]. A Tabela 1 apresenta os resultados de ajuste.

[4] Modelos livres da restrição de variáveis comuns entre as investigações foram ajustados. O melhor modelo considerava um conjunto maior de variáveis, mas o grau de qualidade da previsão da variável "usuários de Internet" não se mostrou significativamente diferente da apresentada pelo modelo com variáveis comuns às duas coletas.

TABELA 1
ESTATÍSTICAS DE AJUSTE DO MODELO

Variáveis independentes no modelo	TIC Domicílios 2019	
	R^2	Taxa de classificação correta (1)
Sexo, Idade, Grau de instrução, Classe social, Indicador de usuário de computador	0,431	83%

Fonte: CGI.br, TIC Domicílios 2019.
(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

A partir do modelo ajustado com os dados da pesquisa TIC Domicílios 2019, foram estimados os escores de propensão para o conjunto de respondentes do Painel TIC COVID-19. Em seguida, foi feita a comparação das distribuições dos escores de propensão na amostra da TIC Domicílios 2019 com os escores da amostra do Painel TIC COVID-19 para os usuários de Internet. Os resultados são apresentados na Tabela 2. É possível notar que a distribuição dos escores dos respondentes do Painel TIC COVID-19 em sua segunda edição tem um perfil distinto do observado para a população usuária de Internet de 16 anos ou mais segundo a TIC Domicílios 2019.

TABELA 2
ESTATÍSTICAS DESCRITIVAS DE ESCORES DE PROPENSÃO A SER USUÁRIO DE INTERNET

	Mínimo	Q1	Mediana	Média	Q3	Máximo
TIC Domicílios 2019	0,03	0,72	0,92	0,82	0,99	1,00
Painel TIC COVID-19	0,08	0,96	0,99	0,94	1,00	1,00

Fonte: CGI.br, TIC Domicílios 2019 (2020) e Painel TIC COVID-19 (2020).

Passo IV. Determinação de população de suporte comum das pesquisas TIC Domicílios 2019 e Painel TIC COVID-19

Dado que as distribuições dos escores obtidos em ambas as pesquisas eram diferentes, optou-se por buscar identificar um recorte da amostra de usuários de Internet da TIC Domicílios 2019 que fosse mais parecido com o conjunto de respondentes do Painel TIC COVID-19 em sua segunda edição. A escolha deste recorte levou em conta a observação das distribuições dos escores e a variabilidade em pesos que foram atribuídos aos respondentes do painel, em cada população recortada. Essa avaliação foi feita estimando-se os pesos dos respondentes segundo três opções de corte estudadas na primeira edição e procurando aproximar o corte na segunda edição do corte adotado na primeira edição da pesquisa Painel TIC COVID-19. Os cortes estudados foram:

I. Seleção de todos os respondentes de ambas as pesquisas, sem recorte;
II. Seleção dos respondentes de ambas as pesquisas que têm escores de propensão maior ou igual a dois terços;
III. Seleção dos respondentes de ambas as pesquisas que têm escores de propensão maior ou igual a três quartos;
IV. Seleção dos respondentes de ambas as pesquisas que têm escores de propensão entre 2/3 e 3/4.

Para cada uma das opções, foram estimados pseudo-pesos para os respondentes do Painel TIC COVID-19. A metodologia para a estimação dos pseudo-pesos é apresentada na próxima seção. As Tabelas 3 e 4 apresentam os resultados.

TABELA 3
COMPARAÇÃO DA DISTRIBUIÇÃO DOS PESOS DOS RESPONDENTES DO PAINEL TIC COVID-19, SEGUNDO ALTERNATIVAS DE RECORTE DOS ESCORES DE PROPENSÃO

Estatísticas dos pesos calibrados	Mínimo	Q1	Mediana	Média	Q3	Máximo
Sem recorte de escores	1 288	11 566	24 382	45 685	51 616	941 924
Recorte para escores maiores ou iguais a 2/3	962	12 352	22 355	40 102	44 673	1 069 091
Recorte para escores maiores ou iguais a 11/16	978	13 117	22 807	40 968	46 458	631 865
Recorte para escores maiores ou iguais a 72/100	1 641	14 052	22 913	40 347	46 059	758 985
Recorte para escores maiores ou iguais a 3/4	910	14 702	23 526	38 429	42 074	699 864

Fonte: CGI.br, TIC Domicílios 2019 (2020) e Painel TIC COVID-19 (2020).

TABELA 4
COMPARAÇÃO DA DISTRIBUIÇÃO DOS FATORES DE CALIBRAÇÃO DOS PESOS DOS RESPONDENTES DO PAINEL TIC COVID-19, SEGUNDO ALTERNATIVAS DE RECORTE DE ESCORES DE PROPENSÃO

Estatísticas de razão (pesos calibrados/pesos básicos)	Mínimo	Q1	Mediana	Média	Q3	Máximo
Sem recorte de escores	0,111	0,314	0,449	0,678	0,735	5,369
Recorte para escores maiores ou iguais a 2/3	0,045	0,397	0,549	0,846	0,932	12,100
Recorte para escores maiores ou iguais a 11/16	0,060	0,438	0,608	0,918	1,001	8,968
Recorte para escores maiores ou iguais a 72/100	0,111	0,494	0,656	0,953	1,044	8,573
Recorte para escores maiores ou iguais a 3/4	0,049	0,549	0,732	0,963	1,083	7,405

Fonte: CGI.br, TIC Domicílios 2019 (2020) e Painel TIC COVID-19 (2020).

Os recortes considerados indicam uma grande variabilidade nos pesos. Optou-se pelo recorte de escores maiores ou iguais a 0,72 (72/100), uma vez que os fatores de calibração (razão entre os pesos calibrados e os pesos básicos) têm média próxima a 1 e que este é um recorte que aproxima bastante as populações representadas na primeira e segunda edição do Painel TIC COVID-19. Como na primeira edição, as características dos pesos com esse corte têm as propriedades desejáveis, com os pesos calibrados mais próximos dos pesos inicialmente estabelecidos pela

metodologia de estimação de pseudo-pesos. O recorte de 0,72 estabelece uma população com os totais estimados apresentados na Tabela 5. Considerando esse recorte, 2.408 respondentes do Painel TIC COVID-19 foram utilizados na estimação dos indicadores de interesse.

TABELA 5
ESTIMATIVAS DAS POPULAÇÕES DE INDIVÍDUOS USUÁRIOS DE INTERNET DE 16 OU MAIS ANOS REPRESENTADOS PELO PAINEL TIC COVID-19

Características	População de usuários de Internet de 16 anos e mais de idade (TIC Domicílios 2019)	População representada pelo Painel TIC COVID-19	
		Edição 1	Edição 2
Sexo			
Masculino	57 529 132	48 969 763	46 868 635
Feminino	63 946 589	52 206 246	50 286 021
Classe social			
AB	28 021 597	30 107 156	29 537 283
C	60 187 722	51 599 711	49 696 700
DE	30 238 053	19 469 143	17 920 673
Usuário de computador			
Não usuário de computador	61 117 479	40.825 295	36 806 747
Usuário de computador (menos de 3 meses)	60 358 242	60 350 714	60 347 909
Faixa Etária			
De 16 a 24 anos	30 745 323	30 677 229	30 672 190
De 25 a 34 anos	29 151 958	28 134 975	27 051 411
De 35 a 44 anos	26 345 229	22 278 737	21 010 437
De 45 a 59 anos	25 515 249	15 993 718	14 579 877
60 anos ou mais	9 717 962	4 091 350	3 840 741
Escolaridade			
Até fundamental	35 572 623	20 223 133	17 814 922
Ensino Médio	53 760 372	51 334 055	49 897 095
Superior	29 846 696	29 618 821	29 442 639
Região			
Norte	9 144 295	7 729 775	7 519 332
Nordeste	30 695 701	25 406 258	24 284 472
Sudeste	53 292 747	44 364 582	45 222 481
Sul	18 495 682	15 312 771	12 068 385
Centro-Oeste	9 847 295	8 362 624	8 059 986
Total	121 475 721	101 176 009	97 154 656

Fonte: CGI.br, Pesquisa TIC Domicílios 2019 (2020).

9.1.2 - ETAPA 2 - *Estimação de pseudo-probabilidades de inclusão para determinação de pesos dos respondentes do Painel TIC COVID-19*

O processo de estimação de pseudo-pesos consiste na estimação de pseudo-probabilidades de inclusão dos respondentes do Painel TIC COVID-19 (amostra não probabilística) na pesquisa TIC Domicílios 2019 (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 pesquisa.

Para estimar as pseudo-probabilidades, 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 consideradas quatro diferentes possibilidades, conforme os recortes de população estabelecidos na seção anterior. Tais recortes visaram identificar a população suporte comum das duas pesquisas avaliando os pesos obtidos, como sugere Valliant (2019).

O processo de estimação de pseudo-probabilidades empregou 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 COVID-19 (amostra não probabilística) e o valor 0 para respondentes da TIC Domicílios 2019 (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;
III. Ajuste de um modelo de regressão logística tendo a variável **Z** como resposta, levando em consideração o desenho amostral da pesquisa, para estimar as probabilidades de inclusão dos respondentes do Painel TIC COVID-19 na amostra probabilística.

No ajuste do modelo, a amostra do Painel TIC COVID-19 foi considerada como um estrato à parte, e cada respondente dessa amostra foi considerado como sendo uma unidade primária de amostragem (UPA) distinta. Esse procedimento foi 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 contém as seguintes variáveis: idade, classe social, indicador de uso de computador, escolaridade e número de moradores do domicílio. A partir desse modelo, foram estimadas as pseudo-probabilidades de inclusão dos respondentes do Painel TIC COVID-19 na pesquisa TIC Domicílios 2019. Os recíprocos dessas pseudo-probabilidades são os pesos iniciais alocados para cada respondente do Painel TIC COVID-19.

Esses pesos iniciais foram calibrados para totais marginais estimados das variáveis estrato TIC; sexo; faixa etária; escolaridade; classe social e Indicador de uso de computador. Os pesos assim calibrados foram 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 COVID-19 um peso que o trata como se fosse um participante de pesquisa com plano amostral igual ao da pesquisa TIC Domicílios 2019, 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 pseudo-probabilidades 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 pseudo-probabilidades, foram 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, foi ajustado o modelo para estimação de pseudo-probabilidades de inclusão, e correspondentes pseudo-pesos; e
III. Os pseudo-pesos de cada réplica foram calibrados e guardados para estimação da variância.

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

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

Onde:

$\hat{y}$ é 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 COVID-19 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 comparação com a primeira edição do Painel TIC COVID-19, foram feitas novas tabulações dos indicadores comuns das duas edições para a população com recortes iguais a 0,72.

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 commom 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.