

Plano amostral

Cadastros e fontes utilizadas

A fim de produzir um retrato representativo do uso das TICs no Brasil, o plano amostral da pesquisa TIC Domicílios 2010 utiliza informações do Censo Demográfico Brasileiro (2000) e da PNAD 2009, ambas realizadas pelo IBGE. A amostra considera as diferenças regionais e socioeconômicas dos domicílios e cidadãos, além de permitir a comparabilidade internacional.

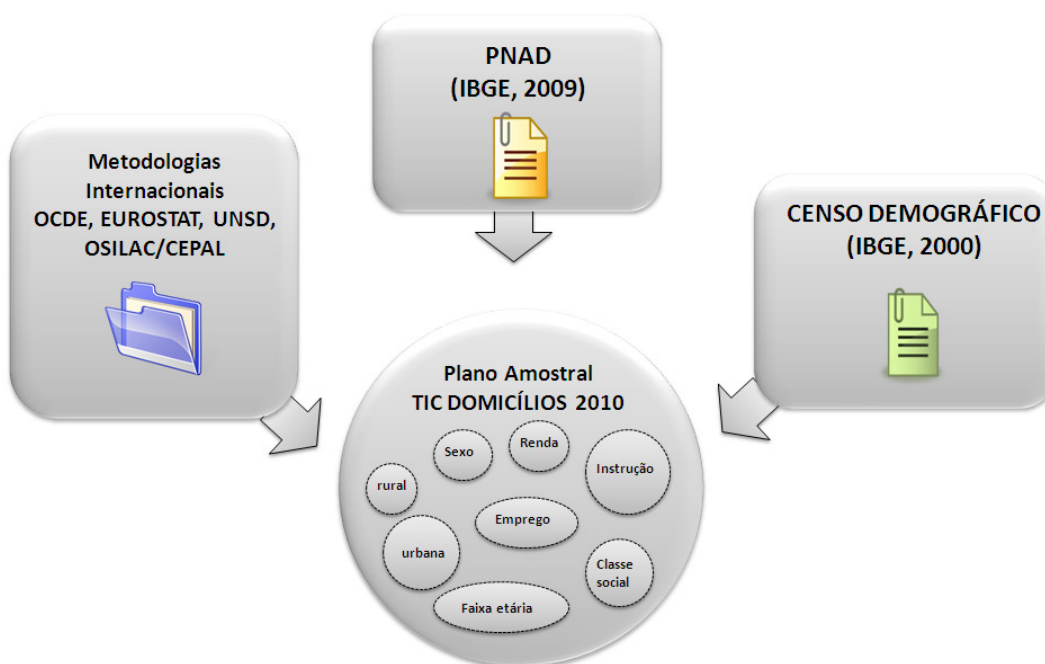


Figura 1. Processo do desenho amostral da pesquisa TIC Domicílios 2010.

Descrição do plano amostral

Introdução

Amostras probabilísticas foram usadas para estimar diversos parâmetros de interesse na população, sendo eles totais ou funções simples de totais populacionais. Considere N como o conjunto de rótulos distintos que são usados para identificar as N unidades da população da pesquisa; s é o conjunto de n rótulos das unidades na amostra selecionada e o total populacional da variável Y , definido como:

$$Y = y_1 + y_2 + \dots + y_n = \sum_{i=1}^N y_i = \sum_{i \in U} y_i$$

O estimador simples (ou de Horvitz-Thompson) do total populacional com base na amostra s é dado por:

$$\hat{Y} = \sum_{i \in s} \frac{y_i}{\pi_i} = \sum_{i \in s} d_i y_i$$

Onde $\pi_i = \Pr(i \in s)$ é a probabilidade de que a unidade i faça parte da amostra s (chamada de probabilidade de inclusão de i), e $d_i = 1/\pi_i$ é o peso amostral básico da unidade i . O peso d_i pode ser interpretado como o número de unidades da população representadas pela unidade i quando esta é selecionada para a amostra s . As probabilidades de inclusão das unidades da amostra foram calculadas a partir da forma de seleção da amostra.

Planejamento amostral

O planejamento amostral da pesquisa TIC Domicílios 2010 empregou duas abordagens distintas. Numa primeira fase do processo de amostragem, foi selecionada uma amostra probabilística de setores censitários. O plano amostral empregado nesta primeira fase para obtenção da amostra de setores pode ser descrito como amostragem estratificada de conglomerados em um ou dois estágios, dependendo do estrato (urbano ou rural).

O número de estágios do plano amostral depende essencialmente do papel conferido à seleção dos municípios. Vários municípios foram incluídos na amostra com probabilidade igual a 1 (municípios auto-representativos). Neste caso, os municípios funcionam como estratos para seleção da amostra de setores e, posteriormente, de moradores para as entrevistas, e por isso não contam como um primeiro estágio de seleção. Todos os demais municípios não representativos funcionam como unidades primárias de amostragem (UPAs) num primeiro estágio, e é nesta parte que a amostra probabilística tem dois estágios de seleção: sorteio de municípios e posterior sorteio de setores censitários nos municípios selecionados.

Em ambos os estágios (sorteio de municípios e sorteio de setores), o sorteio de unidades de amostragem foi feito com probabilidades proporcionais a medidas de tamanho e usando um método de amostragem denominado Amostragem Seqüencial de Poisson (Ohlsson, 1998). Para a pesquisa TIC Domicílios 2010, o último estágio de seleção da amostra probabilística foi sempre a seleção de setores censitários, que funcionaram como UPAs nos municípios autorrepresentativos e como unidades secundárias de amostragem (USAs) nos demais municípios da amostra.

Para assegurar a representatividade da população brasileira, procurou-se incorporar no desenho amostral a diversidade regional, econômica e social do país, tanto de pessoas como de domicílios, por meio de variáveis de estratificação e do estabelecendo de cotas para determinadas variáveis. Assim, a segunda abordagem da amostragem foi por cotas, empregada nos setores censitários selecionados para selecionar domicílios e, em seguida, nos moradores a entrevistar. A seleção de moradores empregou um procedimento de varredura sistemática dos domicílios, sendo visitado um a cada três domicílios encontrados no percurso do entrevistador.

O processo desenvolve-se com duas abordagens, em dois ou três estágios, em que os setores censitários são considerados UPA ou USA, respectivamente. A seguir, detalhes do processo em que os municípios são considerados UPA:

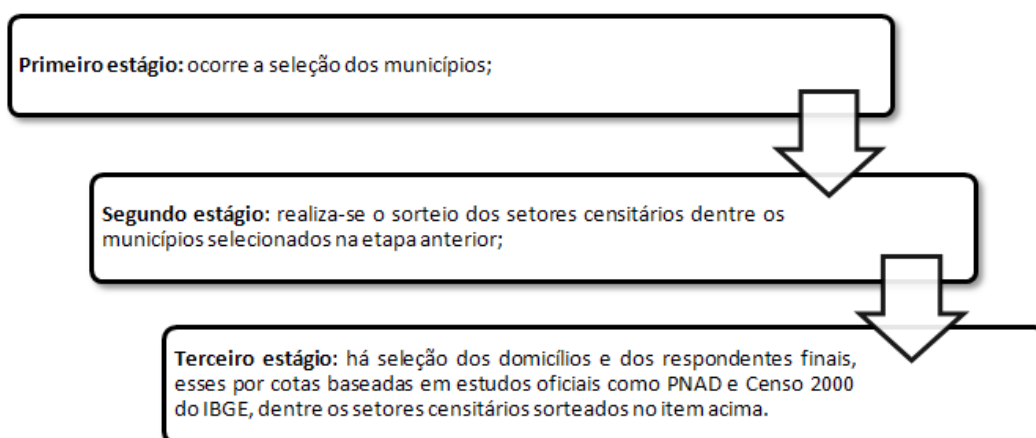


Figura 2. Processo para definição do município como UPA.

Primeiro estágio: seleção dos municípios.

Segundo estágio: sorteio dos setores censitários entre os municípios selecionados na etapa anterior.

Terceiro estágio: seleção dos domicílios e dos respondentes finais, estes por cotas baseadas em estudos oficiais como PNAD e Censo 2000, ambos do IBGE, entre os setores censitários sorteados na etapa anterior.

As figuras abaixo ilustram detalhadamente cada etapa do processo de coleta de dados.



Etapa A. Estratificação por regiões geográficas: para cada Unidade da Federação (UF), foram definidos até três estratos “geográficos”: capitais, regiões metropolitanas e interior. Todas as capitais foram incluídas na amostra (27 estratos). Em 9 UF's (Pará, Ceará, Pernambuco, Bahia, Minas Gerais, Rio de Janeiro, São Paulo, Paraná e Rio Grande do Sul), um segundo estrato foi formado pelos municípios que compõem a região metropolitana (RM) formada em torno da capital, exclusive o município da capital. Além disso, nessas 9 UF's todos os demais municípios, excluindo-se a capital e aqueles classificados na região metropolitana, foram incluídos num estrato chamado “interior”, totalizando os três estratos acima mencionados. Nas outras 17 UF's que não possuem RM (todas as demais, exceto o Distrito Federal), havia somente dois estratos de municípios: interior e capital.

Representação dos

5.507

municípios

Etapa B. Representação dos municípios: Nos estratos de regiões metropolitanas e interior, havia municípios que, por seu tamanho populacional, acabariam tendo probabilidade igual a 1 de inclusão na amostra. Todos esses foram incluídos na amostra (70 municípios autorrepresentativos), e passaram também a funcionar como estratos para fins de seleção da amostra de setores. Nos estratos de municípios não autorrepresentativos, foi aplicado um método de divisão dos municípios em zonas de amostragem para posterior seleção, conforme seu tamanho: a Amostragem Sequencial de Poisson (ASP). Essas zonas foram desenhadas de modo a ficarem com tamanho populacional total aproximadamente igual. Para a seleção de municípios, as medidas de tamanho usadas foram as estimativas do IBGE, em 2009, da população por municípios. Para o sorteio de setores, foram usadas as informações sobre a população dos setores conforme a base cadastral disponível (Censo 2000).

Amostragem
Sequencial de Poisson

$$\pi_i = n \times p_i$$



Etapa C. Amostragem Seqüencial de Poisson (ASP): a seleção dos municípios para compor a amostra é realizada a partir de um número aleatório modificado para cada estrato, de acordo com seu tamanho populacional relativo (p_i). As probabilidades aproximadas de inclusão na amostra são: $\pi_i = n \times p_i$, em que foram selecionadas para a amostra até 2 unidades ($n = 2$).

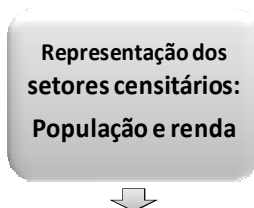
307

Municípios
selecionados

Resultado final. O resultado final é uma amostra de 307 municípios, sendo 70 incluídos com certeza e outros 237 selecionados pelo método ASP aplicado dentro dos estratos formados.



Etapa A.1. Estratificação dos setores censitários: Para os municípios selecionados no estágio anterior, há uma estratificação dos setores conforme a situação (urbana ou rural) antes da aplicação da estratificação por tamanho, configurada na formação das zonas de amostragem. Para amostragem de setores urbanos, todos os 307 municípios selecionados foram considerados, e para a amostragem de setores rurais foi considerada uma subamostra desses 307 municípios. Tanto o sorteio dos setores urbanos como dos setores rurais é feito por ASP. Entretanto serão diferentes as probabilidades de seleção e os estratos formados para aplicação deste método.



Etapa B.1. Representação dos setores: Os setores de cada município também foram estratificados em zonas de amostragem, com os setores ordenados de forma ascendente segundo a população do Censo 2000. Dessa forma, as zonas de amostragem de setores conferem um efeito de estratificação explícita conforme o tamanho do setor (população em 2000). Mais uma vez, as zonas de amostragem foram desenhadas de modo a ficarem com tamanho populacional total aproximadamente igual. Em cada uma dessas zonas de amostragem, dois setores foram selecionados (exceto quando a zona continha apenas um setor, situação em que este era incluído com certeza na amostra) usando o método ASP.

Amostragem
Sequencial de Poisson

$$\pi_{j|i} = m \times \frac{S_{ij}}{\sum_{k \in U_i} S_{ik}}$$

Etapa C.1. Amostragem Sequencial de Poisson (ASP):

Este processo foi seguido pela seleção da amostra de 1.932 setores urbanos dentro dos 307 municípios da amostra do primeiro estágio. A probabilidade de seleção de um setor urbano j dentro do município i é dada por:

$$\pi_{j|i} = m \times \frac{S_{ij}}{\sum_{k \in U_i} S_{ik}}$$

Em que m é a quantidade de setores selecionados para a amostra e S_{ij} é a população do setor j do município i .

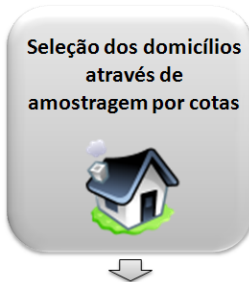
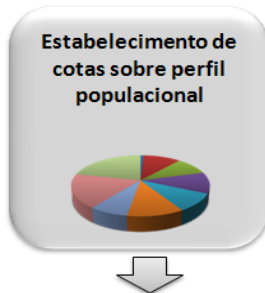
O sorteio de municípios para amostragem de setores rurais constitui amostragem em duas fases. A subamostragem de municípios para compor a amostra da zona rural incluiu 59 municípios da amostra original com certeza (autorrepresentativos). Dos demais 248, somente 216 tinham setores rurais. Destes, 88 foram selecionados por ASP, totalizando 147 municípios na subamostra para sorteio de setores rurais. Ao todo, foram incluídos na amostra 356 setores rurais nestes 147 municípios, sendo que 9 foram incluídos na amostra com certeza.

Para os setores urbanos, a probabilidade de inclusão na amostra é obtida pelo produto das probabilidades de inclusão do município e do setor:

$$\pi_{ij} = \pi_i \times \pi_{i|j}$$

Para os setores rurais, há um terceiro fator que é a probabilidade π_i de sorteio do município para a subamostra de municípios, portanto a probabilidade de inclusão do setor rural j do município i é dada por:

$$\pi_{ij} = \pi_i \times \tau_i \times \pi_{j|i}$$



Seleção dos respondentes satisfazendo às cotas

Resultado final. É uma amostra de 2.288 setores, sendo 1.932 urbanos e 356 rurais.

Etapa A.2. Estabelecimento de cotas: as cotas, estabelecidas segundo o perfil populacional extraído da PNAD 2009 e do Censo Demográfico 2000, são um meio de assegurar a representatividade da amostra em relação à população-alvo da pesquisa e garantem que indivíduos de todas as faixas etárias, gêneros, grau de instrução, regiões geográficas e faixas de renda estejam presentes na amostra. Foram utilizadas cotas de idade, sexo, instrução e PEA, estabelecidas de acordo com os dados oficiais da PNAD 2009. Definem-se essas cotas na fase de planejamento da pesquisa, e elas são controladas por folhas de cotas para cada setor censitário.

Etapa B.2. Seleção dos domicílios: dentro do setor censitário, o domicílio é escolhido com base em cotas de perfil populacional, baseadas no Censo 2000 e PNAD 2009. O entrevistador não tem liberdade total para procurar as cotas e deve seguir um conjunto de procedimentos pré-estabelecidos. O entrevistador deve numerar os quarteirões do setor censitário, começando do norte, e percorrê-los na ordem da numeração no sentido horário, abordando um a cada três domicílios. Isso permite que se possa fazer uma verificação do trabalho e obriga o entrevistador a realizar as entrevistas apenas dentro do setor censitário selecionado.

Seleção dos
respondentes
satisfazendo as cotas



Etapa C.2. Seleção do respondente: dentro do domicílio, escolhe-se o respondente com base em cotas de perfil populacional, de acordo com o Censo 2000 e PNAD 2009. Assim como no estágio anterior, em que ocorreu a seleção do domicílio, o entrevistador não tem liberdade para procurar respondentes, uma vez que estes devem satisfazer aos critérios estabelecidos em uma folha de cota.



Resultado Final. O resultado final de todas as etapas acima é a obtenção da unidade amostral: o entrevistado. Foram feitas 23.107 entrevistas para a amostra principal.

Procedimentos de ponderação

A estratégia de ponderação leva em conta as diversas etapas de amostragem probabilística e o estabelecimento de cotas relativas ao perfil populacional. Uma hipótese para permitir o cálculo de pesos segundo esta abordagem é que o sorteio de domicílios e entrevistados num setor se comportaria de maneira similar ao de uma amostra aleatória simples. Sob esta hipótese, a ponderação básica para entrevistas no setor j do município i foi feita calculando-se o peso da entrevista no setor como:

$$d_{k|ij} = \frac{N_{ij}}{n_{ij}}$$

Em que n_{ij} é o número de entrevistas realizadas no setor e N_{ij} é o número de domicílios existentes no setor segundo o Censo 2000. O peso básico de uma pessoa entrevistada no setor j do município i é dado então por:

$$d_{ijk} = \pi_{ij} \times d_{k|ij}$$

Após a obtenção dos pesos básicos, aplicou-se uma calibração nas respectivas distribuições marginais das variáveis: Região, Sexo, Situação de emprego, Faixa etária, Grau de instrução e Indicador de participação PEA. Essa estratégia assegura que os efeitos de seleção de municípios e setores com maior população serão adequadamente compensados. Conforme Särndal, Swensson e Wretman (1992), um estimador de calibração do total populacional de uma variável de pesquisa y é dado por:

$$\hat{Y}_C = \sum_{i \in S} w_i y_i$$

Em que os pesos amostrais calibrados w_i são escolhidos de modo a minimizar a distância dos pesos do desenho d_i , dada por:

$$G(w; d) = \sum_{i \in S} g(w_i, d_i)$$

E satisfazem à restrição:

$$\hat{X}_C = \sum_{i \in S} w_i x_i = X = \sum_{i \in U} X_i$$

Sendo x_i um vetor com as variáveis de calibração observadas para a unidade i , e X o vetor com os totais populacionais dessas variáveis de calibração. Os pesos w_i são ditos calibrados porque, conforme a restrição, quando aplicados para estimar os totais das variáveis auxiliares x , reproduzem exatamente seus totais populacionais conhecidos. A escolha da função de distância G leva a diferentes tipos de pesos amostrais ou estimadores de calibração, e foi utilizada a função distância abaixo, que dá origem aos chamados estimadores de regressão.

$$G(w; d) = \sum_{i \in S} (w_i - d_i)^2 / d_i$$

Amostra suplementar

Foi realizada uma amostra suplementar de domicílios urbanos visando firmar os resultados obtidos em determinados estratos. Para isso, foi usado um plano amostral idêntico ao descrito, mas aplicado somente a quatro dos estratos geográficos (Bahia – Interior, São Paulo – Interior, Santa Catarina – Total e Rio Grande do Sul – Interior). As amostras tiveram tamanhos idênticos aos da amostra principal nos mesmos estratos geográficos.

Os dados das duas amostras na produção dos resultados da pesquisa foram combinados com vistas à teoria estatística sobre a combinação ótima de estimadores (não viciados). Consideraram-se T um parâmetro de interesse, e \hat{T}_1 e \hat{T}_2 dois estimadores alternativos de T , ambos não viciados. Então um estimador não viciado de T , que sempre tem maior eficiência que \hat{T}_1 e \hat{T}_2 , isoladamente foi dado por:

$$\hat{T}_c = \alpha \hat{T}_1 + (1-\alpha)\hat{T}_2$$

Em que $0 < \alpha < 1$ é um peso a ser determinado. O valor ótimo de α depende da precisão relativa dos estimadores \hat{T}_1 e \hat{T}_2 . O valor de α aplicado foi de 0,3 (amostra principal), e conseqüentemente os dados da amostra suplementar tiveram peso igual a 0,7. Esses pesos foram aplicados para combinar os dados (e as estimativas) das duas amostras apenas nas regiões em que foi feita a amostra suplementar e como multiplicadores dos pesos calibrados calculados para cada uma das amostras separadamente, conforme o método descrito nas seções anteriores.

Assim o peso de uma pessoa entrevistada numa das quatro regiões onde foi feita a amostra suplementar foi obtido como:

$$w_{ijk} = \begin{cases} 0,3 w_{ijk}^P & \text{para pessoa da amostra principal} \\ 0,7 w_{ijk}^S & \text{para pessoa da amostra suplementar} \end{cases}$$

Isso assegurou que as estimativas das duas amostras fossem combinadas para cada uma das regiões com os pesos especificados.

Oversample de usuários de Internet

A amostra adicional de 1.500 entrevistas, chamada *oversample*, teve a finalidade de assegurar a leitura dos indicadores de uso da Internet cuja população respondente ainda era muito pequena para que se obtivesse a precisão desejada. Como a *oversample* dirigiu-se às pessoas que utilizaram a Internet nos últimos três meses, e por ser esta uma subpopulação com características bem específicas, optou-se por não utilizar os mesmos procedimentos de ponderação empregados na amostra principal, baseados nas informações do Censo 2000 e da PNAD 2009.

Os procedimentos de ponderação da *oversample* de usuários de Internet foram baseados nas informações extraídas da amostra principal, sendo esta uma referência mais atualizada. Ou seja, de acordo com as características dos estratos da população de usuários de Internet obtidas da amostra principal, desenhou-se um plano amostral específico para as entrevistas da *oversample*. Das 23.107 entrevistas da amostra principal, 8.432 eram usuários de Internet, ou seja, pessoas que informaram ter acessado a rede nos últimos três meses. Adicionaram-se a este grupo a *oversample* de usuários de Internet, com 1.500 novas entrevistas. A soma das duas amostras totalizou 9.932 respondentes, todos usuários da rede, para os quais foram aplicadas as questões específicas sobre o uso da Internet.

Tratamento de não-resposta

Na pesquisa TIC Domicílios 2010, o tratamento de não-resposta ocorre em dois níveis, descritos a seguir.

Não-resposta à unidade

A não-resposta à unidade ocorre quando não é possível entrevistar uma pessoa em um domicílio previamente sorteado, quando um domicílio não possui respondente satisfazendo aos critérios relativos às cotas estabelecidas, ou ainda se ocorrer divergência excessiva entre as respostas de diferentes questões. Para resolver esse problema, um novo sorteio de domicílio é

realizado, ou mesmo de setor censitário, até que seja atingido o número de entrevistas necessário para satisfazer à precisão da pesquisa.

No quadro abaixo estão listados todos os contatos realizados com a descrição do tipo de classificação quanto à efetivação da entrevista:

Descrição das ocorrências	Quantidade de contatos
1 - Ninguém no domicílio (NEC)	48.262
2 - Construção/casa vazia/comércio, etc.	9.311
3 - Prédio/condomínio não permite pesquisa	984
4 - Respondente ausente	4.394
5 - Informante não qualificado	13.407
6 - Recusa em responder	11.826
7 - Cota/filtro – Encerre	18.909
8 - Substituída	12
9 - Efetiva	27.123
10 - Pulo/arrolamento	46.038
TOTAL DE CONTATOS	180.266

Não-resposta ao item

A não-resposta ao item ocorre quando o entrevistado deixa de responder uma determinada questão, o que acontece, geralmente, porque este não tem o conhecimento sobre o assunto ou mesmo por outros motivos particulares.

Na maioria das questões, é disponibilizada uma opção para o respondente que se enquadra nesta classe. Isso possibilita mensurar a taxa de não-resposta para um dado indicador e, quando necessário, aplicar procedimentos de ponderação específicos utilizando-se as taxas de não-resposta para o indicador. Os dados foram descritos em cada indicador no campo “Não sabe / Não respondeu”.

Precisão da pesquisa

As medidas de erro amostral relativas a mensuração de indicadores e proporções da pesquisa TIC Domicílios 2010 foram calculadas utilizando-se um nível de confiança de 95% e supondo cenários de maior variabilidade possível para um dado indicador. A título de ilustração, a maior variabilidade ocorre quando uma dada proporção populacional assume o valor de 50% (P=0,5).

No quadro abaixo destacamos os erros amostrais, calculados com as ressalvas citadas acima, para as cinco grandes regiões brasileiras e também para o total da população.

Erro amostral – Variáveis de cruzamento (Total Brasil)

Variáveis de Cruzamento		Amostra Principal	Erro Amostral*	Oversample Internet	Total Usuários Internet	Erro amostral* (Amostra internet)
ÁREA	URBANA	19510	0,7%	1280	9028	1,0%
	RURAL	3597	1,6%	220	904	3,3%
REGIÕES DO PAÍS	SUDESTE	5301	1,3%	272	2424	2,0%
	NORDESTE	7009	1,2%	592	2822	1,8%
	SUL	4588	1,4%	132	1622	2,4%
	NORTE	3878	1,6%	332	1748	2,3%
	CENTRO-OESTE	2331	2,0%	172	1316	2,7%
SEXO	Masculino	11208	0,9%	702	4947	1,4%
	Feminino	11899	0,9%	798	4985	1,4%
GRAU DE INSTRUÇÃO	Analfabeto/Educação infantil	2025	2,2%	14	113	9,2%
	Fundamental	11973	0,9%	389	3544	1,6%
	Médio	6543	1,2%	592	3882	1,6%
	Superior	2566	1,9%	505	2393	2,0%
FAIXA ETÁRIA	De 10 a 15 anos	4440	1,5%	174	2760	1,9%
	De 16 a 24 anos	3621	1,6%	310	2431	2,0%
	De 25 a 34 anos	4140	1,5%	295	2212	2,1%
	De 35 a 44 anos	3903	1,6%	273	1395	2,6%
	De 45 a 59 anos	3857	1,6%	311	885	3,3%
	De 60 anos ou mais	3146	1,7%	137	249	6,2%
RENDA FAMILIAR	Até 1SM	4150	1,5%	76	793	3,5%
	1SM - 2SM	8274	1,1%	410	2781	1,9%
	2SM - 3SM	4027	1,5%	291	1998	2,2%
	3SM - 5SM	2831	1,8%	311	1781	2,3%
	5SM - 10SM	1408	2,6%	210	1158	2,9%
	10SM OU +	332	5,4%	74	312	5,5%
CLASSE ECONÔMICA ⁶	A	235	6,4%	49	234	6,4%
	B	4074	1,5%	557	3185	1,7%
	C	12514	0,9%	777	5589	1,3%
	DE	6284	1,2%	117	924	3,2%
SITUAÇÃO DE EMPREGO	Trabalhador	14610	0,8%	976	6272	1,2%
	Desempregado	463	4,6%	61	235	6,4%
	Não integra a população ativa ⁵	8034	1,1%	463	3425	1,7%
TOTAL		23107	0,6%	1500	9932	1,0%

Referências bibliográficas

LUMLEY, T. *Complex Surveys: a Guide to Analysis Using R*. Hoboken. New Jersey: John Wiley & Sons, 2010.

OHLSSON, E. "Sequential Poisson Sampling". *Journal of Official Statistics*, 14(2), 1998, p. 149-162.

SÄRNDAL, C. E.; SWENSSON, B.; WRETMAN, J. *Model Assisted Survey Sampling*. Nova York: Springer, 1992.

THOMPSON, S. K. *Sampling*. Hoboken: Wiley, 1999.