Tamanho de amostra para a pesquisa especial sobre tabagismo - PETab

Marcos Paulo Soares de Freitas – marcos.freitas@ibge.gov.br
Maurício Franca Lila – mauricio.lila@ibge.gov.br
Giuseppe de Abreu Antonaci – giuseppe.antonaci@ibge.com.br
Sonia Albieri – sonia.albieri@ibge.gov.br

Introdução

Este trabalho descreve os estudos realizados para dimensionamento da amostra para a pesquisa que investigou o tema tabagismo e o plano amostral escolhido. A Pesquisa Especial de Tabagismo – PETab foi uma pesquisa suplementar que foi a campo junto com a PNAD 2008, utilizando toda a estrutura amostral dessa pesquisa.

A seguir são descritos o plano amostral, o tamanho e a alocação da amostra, os estimadores utilizados e, finalmente, uma avaliação da precisão das estimativas obtidas e divulgadas com os dados da pesquisa.

Plano amostral

Um das principais finalidades do plano amostral para esta investigação, Pesquisa Especial sobre Tabagismo – PETab, foi permitir a obtenção de estimativas da proporção de pessoas com determinadas características relacionadas com o consumo de tabaco em nível nacional e em cada uma das Grandes Regiões. A investigação foi domiciliar e o plano amostral empregado foi o mesmo da PNAD, com mais um estágio de seleção: morador com 15 anos ou mais de idade.

O plano de amostragem da PNAD consistiu de uma amostra probabilística de domicílios com três estágios de seleção: unidades primárias - municípios; unidades secundárias - setores censitários; e unidades terciárias - unidades domiciliares (domicílios particulares). No primeiro estágio, as unidades (municípios) foram classificadas em duas categorias: auto-representativas (probabilidade 1 de pertencer à amostra) e não auto-representativas. Os municípios pertencentes à segunda categoria passaram por um processo de estratificação e, em cada estrato, foram selecionados 2 municípios sem reposição e com probabilidade proporcional à população residente obtida no último censo demográfico (no caso, o Censo Demográfico 2000). No segundo estágio, as unidades (setores censitários) foram selecionadas em cada município da amostra, também com probabilidade proporcional e sem reposição, sendo utilizado o número de unidades domiciliares existentes por ocasião do Censo Demográfico 2000 como medida de tamanho.

No terceiro estágio foram selecionados, com eqüiprobabilidade os domicílios particulares e as unidades de habitação em domicílios coletivos para investigação das características dos moradores e da habitação. O questionário da PETab foi respondido pelo próprio morador selecionado, o que implicou em uma modificação na metodologia padrão da PNAD que utiliza informante proxy para obter informações sobre os moradores que não estão presentes no momento da visita do entrevistador ao domicílio. Sendo assim, a PETab foi aplicada em uma sub-amostra de domicílios da PNAD em todos os setores pertencentes à amostra desta pesquisa.

No último estágio de seleção, um morador com 15 anos ou mais de idade em cada domicílio foi selecionado, por amostragem aleatória simples e somente ele respondeu o questionário completo sobre tabagismo. O processo de seleção foi realizado por meio de uma tabela de aleatorização.

Cadastros de Seleção

1º estágio: base territorial contendo a divisão administrativa vigente em 2000, com informações sobre tamanho de população obtidas no Censo Demográfico 2000.

2º estágio: base operacional geográfica contendo a malha setorial vigente em 2000, com informações sobre tamanho de população e quantitativos de unidades domiciliares obtidas no Censo Demográfico 2000.

3º estágio: cadastro de unidades domiciliares construído na operação de listagem, que consiste em relacionar, ordenadamente, todas as unidades residenciais e não residenciais existentes na área dos setores censitários selecionados para a pesquisa. Anualmente, com a finalidade de manter atualizado o cadastro e, desta forma, preservar as frações de amostragem prefixadas, essa operação é realizada. Além desta atualização, com a finalidade de acompanhar o crescimento dos municípios pertencentes à amostra, criou-se um cadastro complementar constituído pelas unidades domiciliares existentes em conjuntos residenciais, edifícios e aglomerados subnormais com 30 ou mais unidades residenciais, que tenham surgido nestes municípios após a realização do Censo Demográfico 2000.

4º estágio: lista de moradores com 15 anos ou mais de idade construída no momento da entrevista na unidade domiciliar selecionada.

Tamanho da Amostra

O tamanho de amostra foi sugerido através de cálculo baseado em amostragem aleatória simples (AAS), a fim de definir os níveis de precisão, medida pelo coeficiente de variação (cv), que seriam obtidos para as estimativas de proporções em diversos níveis geográficos.

O plano amostral que foi adotado na investigação não foi amostragem aleatória simples de pessoas, por isso foi feita uma correção nos valores dos coeficientes de variação considerando o efeito de plano amostral (EPA) (design effects). Esta medida, EPA, indica o quanto o plano amostral por conglomerados é menos eficiente (maior cv) que a amostragem aleatória simples. Os fatores que interferem no valor do EPA são o número de domicílios/pessoas selecionados em cada setor e o coeficiente de correlação intraclasse, que mede o grau de homogeneidade dentro dos setores em relação à variável de interesse.

Como não há informação sobre o comportamento da variável de interesse para calcular o EPA, seu valor foi determinado considerando estudos sobre outras variáveis que provavelmente possuem efeito de conglomeração maior. Para determinação do valor de EPA foram levados em conta a seleção da subamostra de domicílios e o tamanho médio da amostra de pessoas por setor.

A fórmula utilizada foi a seguinte: $CV\left(\hat{P}\right)=\frac{\sqrt{EPA\cdot\frac{N-n}{N-1}\cdot\frac{P\cdot Q}{n}}}{P}$, onde:

CV é o coeficiente de variação da estimativa de proporção;
n é o tamanho de amostra de pessoas;
N é o número total de pessoas;
P é a proporção de interesse;
$Q=1-P$;
$EPA=1+\left(\bar{n}-1\right)\cdot\rho$ é o efeito de plano amostral;
$\bar{n}$ é o tamanho médio da amostra de pessoas por setor e
ρ é o coeficiente de correlação intraclasse.

O tamanho de amostra esperado considerado para a PETab foi de 40.000 pessoas, o que representa aproximadamente a seleção de um domicílio a cada três domicílios da amostra da PNAD, considerando uma taxa de não-resposta de 20% (que inclui domicílios desocupados, domicílios destruídos, recusas e entrevistas incompletas). Importante ressaltar que, antes de uma entrevista ser classificada como incompleta, foram realizadas três tentativas (3 visitas) para completar o questionário.

O tamanho médio esperado da amostra de domicílios por setor foi de 5 domicílios, considerando a seleção em todos os setores da amostra da PNAD. Com este tamanho de amostra, o valor de EPA obtido foi de 1,5.

Após a análise dos coeficientes de variação, esperava-se ser possível estimar, com qualidade, proporções de pessoas que possuem características raras (maiores que 0,01) em nível nacional tanto na área urbana, quanto na área rural e, em nível das Grandes Regiões, proporções maiores que 0,10 na área urbana e na área rural. Além desses níveis geográficos, também esperava-se ser possível estimar com precisão aceitável (cv até 15%) algumas proporções pequenas, maiores que 0,15, em nível de Unidade da Federação (UF), total e área urbana.

O tamanho total da amostra planejado foi de 50.000 pessoas, com a expectativa de realização de 40.000 entrevistas. Com a aplicação da seleção de 1 a cada 3 domicílios em cada setor, ao final da coleta foram selecionadas 51.011 pessoas, das quais 39.425 foram entrevistadas.

Fatores de expansão

Os fatores de expansão ou pesos amostrais, para a PETab foram calculados em três etapas, considerando: a probabilidade de seleção; os ajustes pela não-resposta e os ajustes para calibração dos totais estimados pelas estimativas provenientes da PNAD.

As etapas de cálculo dos pesos amostrais foram:

Peso amostral básico

O peso amostral básico é definido como o inverso da probabilidade de seleção e foi calculado da seguinte forma:

- probabilidade de seleção do domicílio para a PETab, dado que o domicílio foi selecionado para a PNAD

$$p^{d}_{hijk} = \frac{n^{T}_{hij}}{n_{hij}}$$

onde,

h é indicador de estrato de seleção da PNAD;
i é indicador de município;
j é indicador de setor censitário;
k é o indicador de domicílios;

n^{T}_{hij} é o número de domicílios selecionados para a amostra da PETab no setor j, do município i, do estrato h e
n_{hij} é o número de domicílios selecionados para a amostra da PNAD no setor j, do município i, do estrato h.

- probabilidade de seleção da pessoa para a PETab, dado que o domicílio foi selecionado para a PNAD

$$p^{p}_{hijk} = \frac{1}{O_{hijk}} \cdot \frac{n^{T}_{hij}}{n_{hij}}$$

onde,
O_{hijk} é o número e pessoas com 15 anos ou mais de idade no domicílio k, do setor j, do município i, do estrato h.

- probabilidade de seleção do domicílio para a PNAD

$$p_{hijk}$$

esta probabilidade é constante em cada pós-estrato g da PNAD (Região Metropolitana e Resto da Unidade da Federação; Rural e Urbano e combinação destas duas subdivisões no Pará).

- probabilidade de seleção da pessoa para a PETab

$$p^{p^*}_{hijk} = \frac{1}{O_{hijk}} \cdot \frac{n^{T}_{hij}}{n_{hij}} \cdot p_{hijk}$$

- peso amostral básico da pessoa selecionada para a PETab

$$w^{p}_{hijk} = \frac{1}{p^{p^*}_{hijk}} = O_{hijk} \cdot \frac{n_{hij}}{n^{T}_{hij}} \cdot \frac{1}{p_{hijk}}$$

Peso amostral com ajuste de não-resposta

Para compensar a ocorrência de perda de entrevista por não-resposta na PETab, ou seja, por domicílio fechado, por recusa dos moradores em atender o entrevistador e por recusa da pessoa selecionada em responder o questionário, o peso amostral básico foi ajustado como se segue:

- peso amostral com ajuste por não-resposta da pessoa selecionada para a PETab

$$w^{p*}_{hijk} = O_{hijk} \cdot \frac{n_{hij}}{n^T_{hij}} \cdot \frac{r^T_{hij} + na^T_{hij}}{r^T_{hij}} \cdot \frac{1}{p_{hijk}}$$

onde,

r^T_{hij} é o número de domicílios selecionados para a amostra da PETab no setor j, do município i, do estrato h com entrevista realizada e

na^T_{hij} é o número de domicílios selecionados para a amostra da PETab no setor j, do município i, do estrato h sem entrevista realizada por não-resposta (domicílio fechado, recusa dos moradores em atender o entrevistador, outro motivo em domicílios ocupados e recusa da pessoa selecionada).

Peso amostral final

Observou-se que na PETab, a não-resposta foi ligeiramente diferenciada por sexo. Devido a isso optou-se por calibrar as estimativas de pessoas de 15 anos ou mais de idade por sexo provenientes da PETab com as estimativas obtidas com a PNAD.

As estimativas da PNAD consideram o peso final desta pesquisa, que é calibrado para que os totais estimados de pessoas nos pós-estratos sejam iguais as estimativas de população feitas pela Coordenação de População de Indicadores Sociais do IBGE. Os pós-estratos são os citados anteriormente.

O ajuste no peso da PETab foi feito também em cada um destes pós-estratos da PNAD, e a expressão final do peso amostral para as pessoas selecionadas é dada por

$$w^{gs}_{hijk} = O_{hijk} \cdot \frac{n_{hij}}{n^T_{hij}} \cdot \frac{r^T_{hij} + na^T_{hij}}{r^T_{hij}} \cdot \frac{1}{p_{hijk}} \cdot \frac{\hat{Y}_{gs}}{\hat{Y}^T_{gs}}$$

onde,

w^{gs}_{hijk} é o peso amostral final da pessoa selecionada no domicílio k, do setor j, do município i, do estrato h, do sexo s e do pós-estrato g;

$\frac{\hat{Y}_{gs}}{\hat{Y}^T_{gs}}$ é o fator de ajuste do peso das pessoas selecionadas do sexo s do pós estrato h;

$\hat{Y}_{gs} = \sum_{hijk} \frac{1}{p_{hijk}} \cdot f_g \cdot y^s_{hijk} \cdot I^g_{hijk}$ é o total estimado de pessoas de 15 anos ou mais de idade do sexo s do pós-estrato g proveniente da PNAD;

$f_g = \frac{T_g}{\hat{T}_g}$ é o fator de calibração do peso amostral da PNAD no pós-estrato g;

T_g é a estimativa de população no pós-estrato g para o ano de 2008[1];

$\hat{T}_g = \sum_{hijk} \frac{1}{p_{hijk}} \cdot y_{hijk} \cdot I^g_{hijk}$ é o total estimado de pessoas no pós-estrato g proveniente da PNAD, utilizando como peso o inverso da probabilidade de seleção para esta pesquisa;

y_{hijk} é o total de pessoas no domicílio k, do setor j, do município i, do estrato h;

y^s_{hijk} é o total de pessoas do sexo s no domicílio k, do setor j, do município i, do estrato h;

$$I^g_{hijk} = \begin{cases} 1 & \text{se domicílio k, do setor j, do município i, do estrato h é do pós – estrato g} \\ 0 & \text{caso contrário} \end{cases}$$

[1] PROJEÇÃO da população do Brasil por sexo e idade 1980-2050: revisão 2008. Rio de Janeiro: IBGE, 2008. Disponível em:<http://www.ibge.gov.br/home/estatistica/populacao/projecao_da_populacao/2008/default.shtm>

$\hat{Y}_{gs}^{T} = \sum_{hijk} w_{hijk}^{p^*} \cdot I_{hijk}^{s} \cdot I_{hijk}^{g}$ é o total estimado de pessoas de 15 anos ou mais de idade do sexo s do pós-estrato g proveniente da PETab, utilizando o peso amostral com ajuste de não-resposta;

$$I_{hijk}^{s} = \begin{cases} 1 & \text{se pessoa selecionada no domicílio k, do setor j, do município i, do estrato h é do sexo s} \\ 0 & \text{caso contrário} \end{cases}$$

Estimativas de Erro Amostral

A precisão das estimativas produzidas com os dados da PETab foi expressa em termos relativos, utilizando o coeficiente de variação (CV). Esses coeficientes de variação (CV) foram estimados utilizando-se o Método do Conglomerado Primário[2], por meio do software SUDAAN (*Survey Data Analysis*). Apresenta-se, a seguir, o estimador da variância do estimador de total de uma variável x

$$\hat{V}(\hat{X}) = \sum_{h} \frac{m_h}{m_h - 1} \sum_{u} \left(\hat{Z}_{hu} - \frac{\hat{Z}_h}{m_h} \right)^2$$

onde,

u é indicador da unidade primária de amostragem (UPA). Nos estratos em que os municípios são autorrepresentativos a UPA é o setor censitário, nos demais estratos a UPA é o município;

m_h é o número de UPAs selecionadas no estrato h;

x_{hijk} é o valor da variável x para a pessoa selecionada no domicílio k, do setor j, do município i, do estrato h;

$$\hat{Z}_{hu} = \sum_{ijkgs} w_{hijk}^{gs} \cdot \left(x_{hijk} - \frac{\hat{X}_h^{gs}}{\hat{Y}_{gs}^{T}} \right);$$

$$\hat{Z}_h = \sum_{u} \hat{Z}_{hu} \text{ e}$$

$$\hat{X}_h^{gs} = \sum_{ijk} w_{hijk}^{gs} \cdot x_{hijk} \cdot I_{hijk}^{g} \cdot I_{hijk}^{s}$$

E o estimador do coeficiente de variação para as estimativas de total é dado por:

$$cv(\hat{X}) = \frac{\sqrt{\hat{V}(\hat{X})}}{\hat{X}}$$

Para cada célula das tabelas com as estimativas da PETab foi estimado o cv, e a média desses cvs ficou em 7,8%, indicando boa precisão para o geral das estimativas. A tabela C1 contém uma distribuição dos coeficientes de variação segundo classes de valor.

Tabela C1 – Distribuição e média dos coeficientes de variação estimados segundo classes de valores

Classe de valor do CV	Número de estimativas divulgadas	Percentual de estimativas divulgadas	Média dos CVs (%)
Total	8.575	100,0	7,8
Até 5%	4.173	48,7	2,2
Mais de 5% até 15%	3.217	37,5	9,1
Mais de 15% até 30%	9.50	11,1	20,4
Mais de 30% até 50%	204	2,4	37,0
Mais de 50%	31	0,4	58,8

[2] *Ultimate Cluster*; detalhes em: Hansen at alli (1953). *Sample Survey Methods and Theory* ou em Pessoa, D.G.C e Silva, P.L.N.(1988) *Análise de dados amostrais complexos.* São Paulo: Associação Brasileira de Estatística

Pela tabela anterior, observa-se que mais de 85% das estimativas foram obtidas com precisão dentro do que foi previsto, ou seja com coeficiente de variação de até 15%. E menos de 3% possuem precisão baixa, por serem estimativas referentes a características mais raras, portanto mais difíceis de serem obtidas na amostra.

Para esta pesquisa, foi efetuada uma avaliação da qualidade do plano tabular. Esta avaliação foi realizada com o auxílio do programa Índice de Qualidade de Tabelas[3] – ou IQT – que avalia a qualidade de cada tabela e também a qualidade global do conjunto de tabelas, mediante fatores calculados a partir da precisão de suas estimativas. O resultado desta avaliação é uma nota que varia de 0 a 10, quanto maior a nota, melhor a precisão das estimativas.

A nota final para o plano tabular da PETab foi de 9,6, que é mais um indicador a confirmar a boa precisão geral das estimativas da pesquisa. A seguir, a tabela C2 mostra o resultado da avaliação para as tabelas agrupadas por temas e por níveis geográficos de divulgação .

Tabela C2 – Nota final proveniente da avaliação no IQT, segundo temas e níveis geográficos

Tema e nível geográfico	Nota final
Total	9,6
Uso de tabaco – Brasil	9,8
Uso de tabaco – Grandes Regiões	9,5
Uso de tabaco – Unidades da Federação	9,7
Cessação – Brasil	9,6
Cessação – Grandes Regiões	9,8
Exposição à fumaça – Grandes Regiões	9,3
Exposição à fumaça – Unidades da Federação	7,2
Economia – Grandes Regiões	9,1
Mídia – Brasil	9,7
Mídia – Grandes Regiões	9,9
Mídia – Unidades da Federação	9,8
Conhecimento, atitudes e percepções – Grandes Regiões	10,0
Conhecimento, atitudes e percepções – Unidades da Federação	10,0

[3] ALBIERI, S. e SILVA, A. N.. *Índice de Qualidade de Tabelas: Avaliação de um plano tabular de pesquisas por amostragem em função da precisão das estimativas.* .[documento interno] Rio de Janeiro: IBGE, Coordenação de Métodos e Qualidade, 2001.