

Amostragem na coleta de dados do Censo Demográfico 2010

Sonia Albieri - sonia.albieri@ibge.gov.br

Marcos Paulo Soares de Freitas - marcos.freitas@ibge.gov.br

IBGE, Coordenação de Métodos e Qualidade

Introdução

O Censo Demográfico 2010 vai adotar o mesmo modelo de investigação de características usado nos censos anteriores, qual seja, a utilização de dois questionários, aqui entendidos como dois conjuntos diferentes de quesitos para a mensuração de características de interesse: um questionário grande para ser aplicado em uma amostra de domicílios (e cada um de seus moradores) e outro, pequeno, para ser aplicado aos domicílios (e cada um de seus moradores) não selecionados para a amostra.

Esse modelo permite formar o que chamamos de conjunto universo, junção de todos os domicílios com as variáveis comuns aos dois questionários, que permite a obtenção de dados para 100% dos domicílios e da população brasileira. Adicionalmente, os domicílios e pessoas que respondem ao questionário da amostra formam o denominado conjunto amostra, a partir do qual são feitas as estimativas das características de interesse, nos níveis geográficos definidos pelo método de estimação, as áreas de ponderação, que não devem ser maiores do que um município.

Este documento apresenta as definições relacionadas com o plano amostral definido para a seleção das unidades a serem pesquisadas por meio do questionário grande, abordando as questões de tamanho da amostra, refletidas nas frações amostrais a serem aplicadas, e o algoritmo de seleção definido como alternativa à seleção sistemática simples em cada setor censitário, que foi usado até então nos censos brasileiros.

1 - Amostragem na coleta

A discussão sobre o uso de amostragem na coleta de informações em censos demográficos passa sempre por duas grandes definições: o plano amostral e a fração (ou frações) de amostragem. A pesquisa por amostragem probabilística na coleta dos censos demográficos no Brasil teve início em 1960 e foi realizada desde então em todos os censos decenais. Apenas as duas operações de Contagem da População, em 1996 e em 2007, não fizeram uso de amostragem na coleta das informações, devido ao número restrito de variáveis que investigaram, em função do próprio objetivo desse tipo de operação, apenas a contagem da população.

Nos Censos de 1960, 1970 e 1980, foi usada uma fração amostral de 25% dos domicílios¹ para a aplicação do questionário da amostra. Nos Censos de 1991 e 2000, foram usadas duas frações amostrais de acordo com o tamanho do município, medido em população. Os municípios com até 15.000² habitantes tiveram 20% de seus domicílios investigados por meio do questionário da amostra e nos municípios com mais de 15.000 habitantes, a fração amostral foi de 10%.

1.1 - Plano amostral

O plano amostral a ser adotado em cada censo foi discutido no momento em que estava sendo planejado e nunca foi encontrada nenhuma justificativa para alterar o desenho definido em 1960, a saber: seleção sistemática de domicílios particulares e de unidades domiciliares em domicílios coletivos, independentemente em cada setor censitário, o que equivale a um plano de amostragem estratificada, onde os estratos são os setores, e a seleção de domicílios é sistemática com equiprobabilidade em cada estrato. Os estudos realizados sobre esse ponto estão descritos em diversos documentos citados nas referências, principalmente em Albieri e Bianchini, 1998 e em METODOLOGIA...2003.

É importante destacar que uma vantagem desse plano amostral, ao considerar cada setor censitário como um estrato, permite atender a demanda por informações em níveis geográficos

¹ Domicílios particulares e famílias e pessoas sós moradoras em domicílios coletivos.

² População estimada pelo IBGE para a data de referência do Censo, em 1991 e em 2000.

pequenos (agregados de poucos setores) e a necessidade de informações da amostra em todos os setores, para o planejamento de amostras das pesquisas domiciliares no período intercensitário.

Uma característica desse plano amostral, a seleção sistemática de domicílios para a aplicação do questionário da amostra justifica-se pela facilidade de aplicação do processo de seleção das unidades amostrais, sem a necessidade de se dispor de um cadastro prévio dessas unidades; ou seja, o cadastro vai sendo construído ou atualizado durante a varredura do território, pelo recenseador, e a definição de qual questionário aplicar no domicílio é feita simultaneamente a isso. Ou seja, a ausência de um cadastro de endereços de domicílios atualizado é a motivação principal para a adoção do processo de seleção sistemática de unidades para comporem a amostra, que sabidamente não é o melhor método de seleção probabilística.

1.2 - Fração amostral

A fração amostral a ser adotada na coleta de dados em cada censo também foi objeto de estudos, principalmente para os censos de 1991 e 2000.

Durante o planejamento do Censo Demográfico de 1991, o assunto foi tratado no âmbito de uma comissão interna criada especialmente para esse fim, que encomendou estudos específicos visando avaliar o efeito da redução da fração amostral que vinha sendo adotada nos censos anteriores (25%). A conclusão dos estudos apontou alternativas que levaram à definição das frações amostrais usadas em 1991, citadas anteriormente.

Posteriormente, durante o planejamento do Censo Demográfico de 2000, novos estudos foram realizados, visando subsidiar a decisão sobre manter o mesmo esquema adotado em 1991 ou introduzir uma terceira ou mesmo uma quarta fração amostral, a ser utilizada nos casos extremos de municípios muito pequenos ou muito grandes, em população residente.

Tanto em um caso como no outro, os estudos estão relatados em documentos técnicos internos citados nas referências, destaque para Silva e Bianchini, 1990 e para a publicação que descreve os aspectos metodológicos do Censo Demográfico de 2000 (METODOLOGIA...2003).

Em outubro de 2004, o IBGE, em parceria com o INEGI, do México, realizou no Brasil, o Seminário Sobre Métodos Alternativos para Censos Demográficos, que contou com a participação de especialistas da França, que apresentaram a experiência com censo contínuo, e dos Estados Unidos da América, que apresentaram os aspectos metodológicos da *American Community Survey*, pesquisa que faz uso de amostras espalhadas no tempo para dar estimativas de características tradicionalmente investigadas no questionário da amostra do censo de população dos EUA.

Em 2005, para estudar essas questões no IBGE, foi constituído um grande grupo de trabalho (cerca de 60 pessoas), coordenado pela Diretoria de Pesquisas – DPE e pela Coordenação Operacional de Censos – COC, para desenvolver o projeto que foi denominado EMACD – Estudos de Modalidades Alternativas de Censos Demográficos. Dada a complexidade do projeto e para realizar os diversos estudos, o grande grupo foi dividido em 5 grupos de trabalho (GT), numa tentativa de refletir as diferentes dimensões do projeto. Um desses grupos ficou encarregado dos estudos relacionados com Amostragem, Estimção e Acumulação de Informações.

Como os métodos alternativos para censos demográficos fazem uso intensivo de amostragem probabilística, o GT específico para tratar dos aspectos de amostragem iniciaram os estudos visando a determinação de tamanhos de amostras, para posteriormente avaliarem os aspectos de estimção e acumulação temporal e espacial de informações.

Da mesma forma que nos censos demográficos de 1991 e 2000, as modalidades alternativas de censo também fazem uso de frações amostrais diferentes em grupos de municípios definidos por seu tamanho, medido em população ou em número de domicílios. Assim, os estudos realizados no âmbito daquele GT tiveram como objetivo definir esses grupos de município e as correspondentes frações amostrais, de forma a garantir uma precisão estipulada para as estimativas em todos os municípios. Ou seja, o menor nível geográfico de interesse, controlado no plano amostral, é o município.

Como ressalta o documento interno, preparado pelo GT, que descreve os estudos realizados (MODALIDADES...2005),

“Os resultados encontrados, ou seja, os grupos e as frações, servem para serem utilizados em qualquer modalidade de censo demográfico, pois para estimar com a qualidade requerida é necessário um tamanho total de amostra que tanto pode ser obtido num mesmo instante, da forma como é feito em um censo realizado nos moldes tradicionais, como pode ser acumulado, se a amostra for coletada continuamente ao longo de um período de tempo, que é a maneira proposta nas modalidades alternativas.”

Durante a realização desses estudos, características de interesse, precisão das estimativas e frações amostrais ou limites de corte foram os fatores considerados na definição de propostas alternativas. Os dados usados foram os do Censo Demográfico 2000.

Ao todo, o documento apresenta 10 propostas resultantes da combinação de diferentes definições de grupos de municípios por tamanho com diferentes frações amostrais. Em cada proposta, para cada grupo de municípios, são apresentados o número de municípios no universo, a fração amostral definida, o número de domicílios e de pessoas no universo e o número de domicílios e de pessoas na amostra resultante da aplicação da correspondente fração.

É importante lembrar, como dito anteriormente, que os cálculos de tamanho de amostra efetuados para as diferentes propostas no âmbito do projeto EMACD são **exatamente** da mesma natureza que os que teriam que ser feitos para subsidiar as discussões para a definição da amostragem na coleta dos dados do Censo 2010. Assim, os resultados encontrados foram usados como subsídios para a definição das frações amostrais a serem usadas no Censo 2010, que estão apresentadas na tabela 1, a seguir, considerando ³ os 5.565 municípios existentes em 2009:

Tabela 1 - Fração amostral e número de municípios, segundo as classes de municípios em número de pessoas

Classes de número de pessoas	Fração amostral	Número de municípios
Até 2.500	50%	260
Mais de 2.500 até 8.000	33%	1.912
Mais de 8.000 até 20.000	20%	1.749
Mais de 20.000 até 500.000	10%	1.604
Mais de 500.000	5%	40
Total	11%	5.565

Para os 40 municípios com mais de 500.000 habitantes, que terão fração amostral menor do que a aplicada nos censos anteriores, decidiu-se avaliar a possibilidade de aplicar frações amostrais diferentes em cada uma de suas divisões administrativas intramunicipais, de forma a permitir a divulgação de estimativas e de microdados nesses níveis geográficos. Esses municípios foram, então, separados em 3 grupos:

- Grupo 1 - composto por 7 municípios sem subdivisão administrativa na base territorial 2010 - a fração de 5% ficou mantida;

³ De fato, são 5.563 municípios. As outras duas unidades referem-se uma ao distrito de Fernando de Noronha, subordinado ao estado de Pernambuco, e outra a Brasília, Distrito Federal, que são consideradas unidades de planejamento de atividades de pesquisa em nível municipal.

- Grupo 2 - composto por 15 municípios onde não é necessário o aumento da fração em nenhuma de suas subdivisões, pois o tamanho da amostra resultante já contempla o mínimo estabelecido. - a fração de 5% ficou mantida;
- Grupo 3 - composto por 18 municípios onde há necessidade de aumento na fração amostral em pelo menos uma subdivisão.

Um ponto que foi muito discutido durante as etapas de planejamento dos censos de 1991 e de 2000, que influenciou fortemente a definição das frações amostrais usadas, refere-se às questões operacionais relacionadas com a utilização de mais de uma fração amostral, dependendo do tamanho do município. Havia uma preocupação muito grande com os controles para garantir a aplicação correta da fração definida, para cada município. Isso determinou que o número de frações diferentes ficasse reduzido a dois, mesmo havendo evidências das vantagens de se ter mais uma ou mais duas frações distintas. Nesse sentido, o uso de computadores de mão é um facilitador, uma vez que os controles podem ser feitos de forma automática e centralizada. E a adoção de mais de duas frações amostrais passou a ser uma alternativa de fato.

2 – A seleção de unidades para a amostra

Para completar as definições relacionadas com o uso de amostragem na operação de campo, falta ainda definir o esquema de seleção das unidades da amostra.

Vale registrar que o uso de computadores de mão para a realização das entrevistas pessoais nos domicílios foi considerada na definição dos pontos acima e também estão norteando as discussões e as definições sobre os demais aspectos ainda por definir.

Por exemplo, a definição de mais de duas frações amostrais só se deu em função da facilidade operacional, de aplicação e de controle, permitida com o uso dos computadores de mão na operação de coleta de dados.

Assim, a idéia é contar com a construção das listas de domicílios (Folha de Coleta, com a relação de todas as unidades, inclusive as não residenciais, e Folha de Domicílio Coletivo, específica para listar as famílias e pessoas sós moradoras em domicílio coletivo) simultaneamente com a operação de coleta.

2.1 – Os procedimentos de seleção discutidos

Vale lembrar que, até o momento, não foi encontrada nenhuma justificativa contundente para alterar o esquema de seleção que vem sendo usado desde 1960, a saber: seleção sistemática de domicílios particulares e de unidades em domicílios coletivos, independentemente em cada setor censitário, o que equivale a um plano de amostragem estratificada, onde os estratos são os setores, e a seleção de domicílios é sistemática com equiprobabilidade em cada estrato. Os estudos realizados sobre esse ponto estão descritos em diversos documentos citados nas referências, principalmente em Albieri e Bianchini, 1998 e em METODOLOGIA...2003.

A seleção sistemática de domicílios para a aplicação do questionário da amostra justifica-se pela facilidade de aplicação do processo de seleção das unidades amostrais, sem a necessidade de se dispor de um cadastro prévio dessas unidades; ou seja, o cadastro vai sendo construído ou atualizado durante a varredura do território, pelo recenseador, e a definição de qual questionário aplicar no domicílio é feita simultaneamente a isso.

Uma questão ou crítica que se coloca com relação à seleção sistemática está relacionada com a facilidade de introdução de vício na seleção, por parte do recenseador. Isto porque, após a definição do ponto de partida, que é feita a partir de um processo aleatório, o intervalo de seleção a ser aplicado é conhecido, dependendo apenas da fração amostral de cada setor censitário. Assim, em um setor com fração amostral de 20%, 1 em cada 5 domicílios deve pertencer à amostra, ou seja, deve ser entrevistado por meio do questionário grande. No caso de fração amostral de 33%, 1 em cada 3 deve pertencer à amostra. Conhecendo este intervalo, o recenseador poderia, então, propositadamente manipular a ordem de introdução do domicílio na folha de coleta, que relaciona as unidades domiciliares e controla a seleção para a amostra, de forma a obter na amostra a melhor combinação de domicílios grandes ou pequenos, em número de moradores, de acordo com as taxas de pagamento, que

são estipuladas para cada tipo de questionário, para cada domicílio entrevistado e para cada pessoa moradora do domicílio.

Esse risco sempre existiu e, após a coleta dos dados, o vício de seleção foi de fato detectado no Censo Demográfico de 1980, o que acarretou na definição de procedimentos de correção durante o processo de expansão da amostra, cuja descrição pode ser encontrada em METODOLOGIA...1983.

Com a finalidade de prevenir contra a introdução de vícios por parte do recenseador, as discussões na DPE sobre formas alternativas de seleção de unidades chegaram a gerar propostas de seleção que não garantem que todas as unidades tenham a mesma probabilidade de seleção, fato já devidamente comprovado com o esquema de seleção sistemática e seus estimadores associados.

Algumas dessas propostas foram:

- I. introduzir um ruído aleatório no intervalo de seleção, após a partida aleatória, para que o intervalo não seja constante, não permitindo ao recenseador deduzir qual tipo de questionário deve ser aplicado a um domicílio em particular. Esse procedimento não garante probabilidades de seleção iguais para todas as unidades do setor;
- II. dividir o setor em partes e reiniciar a partida aleatória em cada parte, causando uma interrupção na série dentro do setor, sem o conhecimento por parte do recenseador. Na prática essa proposta equivale a ter subsetores fictícios e pode ser de difícil implementação, em função dos limites da partição;
- III. fazer um sorteio aleatório associado a cada domicílio, com probabilidade igual à fração amostral do município/setor a que pertence, para determinar o tipo de questionário a ser aplicado. Se o tamanho do setor em número de domicílios é conhecido, ou seja, se já estiver finalizada a etapa de varredura e preenchimento da folha de coleta do setor, este procedimento pode ser aplicado e garante a equi-probabilidade de seleção dos domicílios. Caso o tamanho do setor não seja conhecido, este procedimento implica em ter o tamanho final da amostra como uma variável aleatória, fato indesejável em uma operação em larga escala como é a pesquisa amostral do censo, além de haver o risco de setores sem nenhum domicílio selecionado para a amostra.

2.2 – O procedimento de seleção proposto

A prevenção contra vícios de coleta, de qualquer natureza, só se faz com supervisão de campo e ações decorrentes, nos casos de detecção de sua ocorrência.

Acredita-se que o uso do computador de mão é um auxiliar importante no controle da seleção da unidade para a amostra, porque a operação de listar o domicílio na folha de coleta também é feita no computador, simultaneamente com a determinação da espécie da unidade (particular ocupado, fechado, vago, uso ocasional, domicílio coletivo, não residencial). Isso permite que a indicação de qual tipo de questionário deve ser aplicado na unidade registrada seja feita antes mesmo que o recenseador saiba o seu tamanho em número de moradores. E claro, não deve existir a possibilidade de cancelamento do início de entrevista, com posterior alteração do tipo de questionário!

Posto isso, em um primeiro momento, propôs-se que a seleção dos domicílios particulares para a amostra do Censo 2010 fosse feita independentemente em cada setor censitário, através da Folha de Coleta, de forma sistemática, com início aleatório e intervalo de seleção definido em função da fração amostral do setor. O início aleatório tanto poderia ser um atributo do setor, definido antecipadamente de forma centralizada, como poderia ser determinado aleatoriamente no computador de mão no momento da abertura do setor. Caso fosse selecionado para a amostra um domicílio particular fechado, vago ou de uso ocasional, para os quais não é possível obter nenhuma informação, deveria ser mantida a série de seleção de unidades para a amostra.

O mesmo procedimento seria aplicado para a seleção das famílias e pessoas sós moradoras em domicílios coletivos, considerando a construção de um cadastro separado para essas unidades (Folha de Domicílio Coletivo).

Posteriormente, decidiu-se avaliar a possibilidade de implementar, no computador de mão, um procedimento tal como descrito na proposta II acima. Assim, foram realizados estudos para avaliar a

seleção sistemática considerando um intervalo de salto duas vezes o tamanho do salto tradicional que seria obtido pela aplicação da seleção sistemática simples.

Assim, por exemplo, em um setor com fração amostral de 20%, em vez de selecionar 1 em cada 5 domicílios, o algoritmo prevê a seleção de 2 em cada 10, aleatoriamente no grupo formado. Dessa forma, não fica evidenciado o salto constante entre os números de ordem de unidades selecionadas em sequência, dificultando a introdução de vícios operacionais por parte dos entrevistadores.

O teste de aplicabilidade foi bem sucedido e os estudos para avaliar probabilidades de seleção e impactos nas fórmulas de variância do estimador de total concluíram pela viabilidade de utilização desse procedimento, que foi implementado de forma definitiva no sistema que controla toda a operação de coleta de dados do Censo 2010.

Considerando as cinco frações amostrais definidas, a seleção dos domicílios para a amostra, que significa dizer que serão entrevistados por meio do questionário grande, se dará, então, conforme apresentado abaixo:

- para a fração de 50%, 2 a cada 4 domicílios;
- para a fração de 33%, 2 a cada 6 domicílios;
- para a fração de 20%, 2 a cada 10 domicílios;
- para a fração de 10%, 2 a cada 20 domicílios;
- para a fração de 5%, 2 a cada 40 domicílios.

Com isso, para cada setor censitário, a amostra será composta por domicílios particulares selecionados e todos os seus moradores e as famílias e pessoas sós moradoras em domicílios coletivos selecionadas, de acordo com o procedimento descrito.

3 - Considerações finais

A definição dos aspectos de amostragem para a realização de uma pesquisa por amostragem probabilística, como é o caso da investigação do questionário da amostra do Censo Demográfico 2010, engloba a definição do plano amostral, das frações amostrais e do esquema de seleção das unidades da amostra, para permitir a operação de campo, qual seja a coleta das informações. Posteriormente, outras definições serão necessárias, tais como o método de expansão da amostra (obtenção dos fatores de expansão e das estimativas), o tratamento da não resposta, a definição das áreas de ponderação e o método de estimação das medidas de precisão associadas às estimativas. Para tanto serão necessários novos estudos, experimentos, análise de resultados e avaliações relacionadas com esses aspectos de forma a permitir a discussão técnica e posterior definições metodológicas.

4 - Referências

ALBIERI, S. e Bianchini, Z.M. *Sobre o uso de amostragem de setores na coleta dos dados dos censos demográficos brasileiros*. Rio de Janeiro: IBGE, Diretoria de Pesquisas, Departamento de Metodologia, 1998. 12p.

ALBIERI, S. e Freitas, M.P.S de. *Censo Demográfico de 2010: plano amostral para a coleta de dados*. Rio de Janeiro: IBGE, Diretoria de Pesquisas, agosto de 2008, 25p.

METODOLOGIA do censo demográfico de 1980. Rio de Janeiro: IBGE, 1983. 478p. (Série relatórios metodológicos, volume 4).

METODOLOGIA do censo demográfico 2000. Rio de Janeiro: IBGE, 2003. 574p. (Série relatórios metodológicos, volume 25).

MODALIDADES alternativas de censos demográficos: aspectos de amostragem. 1ª versão. Rio de Janeiro: IBGE, junho de 2005, 77p. [preparado pelo Grupo de Trabalho de Amostragem, Estimação e Acumulação de Informações].

SILVA, L.F. e Bianchini, Z.M. *A redução da amostra e a utilização de duas frações amostrais no censo demográfico de 1990*. Rio de Janeiro: IBGE, 1990. 49p. (Textos para Discussão, n.33).