ASPECTOS DE AMOSTRAGEM - Pesquisa Padrões de Vida - 1996-1997

1. O PLANEJAMENTO DA AMOSTRA

O desenho amostral da PPV - Pesquisa sobre Padrões de Vida - foi discutido com os técnicos do Banco Mundial e a dimensão da amostra foi fixada em função do orçamento disponível para a realização da pesquisa.

Como pesquisa piloto optou-se por sua realização apenas nas Regiões Nordeste e Sudeste do País, considerando 10 estratos geográficos, a saber: Região Metropolitana de Fortaleza, Região Metropolitana de Recife, Região Metropolitana de Salvador, restante da área urbana do Nordeste, restante da área rural do Nordeste, Região Metropolitana de Belo Horizonte, Região Metropolitana do Rio de Janeiro, Região Metropolitana de São Paulo, restante da área urbana do Sudeste e restante da área rural do Sudeste.

Tal como em outras pesquisas domiciliares realizadas pelo IBGE, optou-se por um desenho com dois estágios de seleção, com estratificação das unidades primárias e seleção proporcional a uma medida de tamanho e seleção aleatória das unidades de segundo estágio. A unidade primária é o setor da base geográfica do Censo Demográfico de 1991 e a unidade de segundo estágio é o domicílio.

O tamanho da amostra para cada estrato geográfico foi fixado em 480 domicílios. Em cada estrato geográfico foi fixado em 60 o número de setores a serem selecionados e 8 domicílios em cada setor, com exceção para os estratos que correspondem ao restante da área rural de cada Região onde fixou-se em 30 o número de setores e em 16 o número de domicílios a serem selecionados por setor, em função da dificuldade de acesso a esses setores, o que implicaria em aumento de custo.

O tamanho da amostra fixado foi defendido pelos técnicos do Banco Mundial em função da experiência nos outros países onde a pesquisa foi ou está sendo conduzida, pela necessidade de produzir informações com a maior rapidez possível e por julgar que o objetivo da pesquisa não é produzir tabulações com cruzamentos de variáveis, tal como ocorre com as informações da Pesquisa Nacional por Amostra de Domicílios - PNAD, mas o de fornecer indicadores de tendência ou de variação em níveis bastante agregados.

1.1. A definição dos estratos estatísticos

Conforme descrito anteriormente, o setor é a unidade primária de amostragem, o domicílio é a unidade secundária e unidade de investigação.

A estratificação das unidades primárias de amostragem foi definida em duas etapas: a primeira, considerando a divisão geográfica de interesse, que resultou na definição de 10 estratos geográficos; para cada um dos estratos geográficos, a segunda estratificação foi definida por critérios estatísticos, considerando as informações sobre a renda média mensal do chefe do domicílio, variável que foi investigada no Censo Demográfico de 1991 para todos os domicílios.

1.2. A alocação da amostra nos estratos de renda

Vale lembrar que o tamanho final da amostra de domicílios foi fixada em função do custo, mais especificamente dos recursos financeiros disponíveis. Em conseqüência, o tamanho da amostra de setores e o número de domicílios a serem selecionados por setor também foram fixados, a saber:

- 60 setores e 8 domicílios por setor, nos estratos geográficos urbanos e regiões metropolitanas (estratos geográficos 1,2,3,4,6,7,8 e 9);
- 30 setores e 16 domicílios por setor, nos estratos geográficos rurais (estratos geográficos 5 e 10).

Antes da alocação nos estratos de renda, a amostra total nos 10 estratos geográficos ficou com 540 setores e 4.800 domicílios.

Foi utilizada a alocação proporcional, com base no número de domicílios particulares permanentes ocupados, obtidos pelo Censo 91.

A Tabela 1, a seguir, apresenta os tamanhos da amostra de setores e de domicílios após a alocação nos estratos de renda, bem como a fração esperada de domicílios. Vale lembrar quem, durante o procedimento de alocação, os valores resultantes foram arredondados para o maior inteiro e em um único estrato, após o arredondamento, o valor resultante 1 foi alterado para 2 a fim de permitir o cálculo de variâncias. Como pode ser observado, em função da variabilidade da fração amostral, a amostra resultante não é a autoponderada.

Tabela 1 - Número de domicílios no universo, número de domicílios esperados na amostra e fração amostral de domicílios, por estrato geográfico e estrato de renda

Estrato Geográfico	Estrato de Renda	Número de setores na amostra	Número de domicílios esperados na amostra	Fração amostral domicílios (por 10.000)
1 – RM Fortaleza	I	46	368	9,85
	II	11	88	10,22
	III	5	40	10,83
2 – RM Recife	I	45	360	7,32
	II	9	72	7,41
	III	7	56	8,00
3 – RM Salvador	I	46	368	8,49
	II	9	72	8,61
	III	6	48	9,38
4 – Restante Nordeste Urbano	I	47	376	1,22
	II	11	88	1,22
	III	3	24	1,30
5 – Restante Nordeste Rural	I	26	416	1,48
	II	5	80	1,82
	III	2	32	7,42
6 – RM Belo Horizonte	I	45	360	5,94
	II	11	88	6,26
	III	6	48	6,33
7 – RM Rio de Janeiro	I	44	352	1,80
	II	10	80	1,81
	III	7	56	1,91
8 – RM São Paulo	I	44	352	1,23
	II	12	96	1,23
	III	5	40	1,21
9 – Restante Sudeste Urbano	I	33	264	0,71
	II	20	160	0,72
	III	8	64	0,75
10 – Restante Sudeste Rural	I	19	304	3,10
	II	9	144	3,20
	III	3	48	3,96
Total	I	395	3.520	2,03
	II	107	968	1,77
	III	52	456	2,21
	---	---	---	---
	Total	554	4.944	1,99

2. A SELEÇÃO DA AMOSTRA

2.1. A seleção da amostra de setores

Para a seleção da amostra de setores, segundo o desenho adotado, qual seja, amostra estratificada com probabilidade proporcional ao tamanho, foi utilizado um programa em linguagem SAS, utilizando a macro de seleção PPTCOM (ver Silva (1989), que foi adaptada para considerar automaticamente os estratos geográficos e estratos de renda definidos. A medida de tamanho adotada foi o número de domicílios em cada setor, conforme definição de P_{hi} mais adiante.

Após a seleção dos setores, foi realizada uma comparação desses setores com os setores pertencentes às amostras da PNAD - Pesquisa Nacional por Amostra de Domicílios, da PME - Pesquisa Mensal de Emprego e da amostra selecionada para a POF 96/96 - Pesquisa de Orçamentos Familiares. Como o esquema de seleção das amostras dessas pesquisas é o mesmo, qual seja, seleção com probabilidade proporcional ao tamanho, era de se esperar que houvesse coincidências de setores selecionados para duas ou mais pesquisas. Foram avaliados os procedimentos adotados nessas outras pesquisas para contornar o problema de setores (ou domicílios) serem investigados em mais de um pesquisa no mesmo período. Nenhuma das soluções adotadas em outras pesquisas foi considerada satisfatória, tendo sido decidido substituir os setores coincidentes com os de outras pesquisas, além daqueles que foram selecionados mais de uma vez na própria PPV, uma vez que a seleção foi com reposição.

Em função dessa decisão, foi selecionada uma segunda amostra, usando os mesmos procedimentos adotados quando da seleção da primeira amostra. Dessa segunda amostra foram extraídos todos os setores coincidentes com os das outras três pesquisas, todos os setores coincidentes com os selecionados na primeira amostra e, também, aqueles selecionados mais de uma vez nessa segunda amostra da PPV. Os setores restantes foram analisados comparativamente àqueles a serem substituídos e, para a substituição propriamente dita, foram consideradas algumas variáveis de controle, a saber: estrato geográfico, estrato de renda, situação (urbana ou rural) e tipo de setor (normal ou de favela). Além disso, foi considerado o valor da probabilidade de seleção. Isto significa que um setor substituto tem as mesmas características nas variáveis de controle e tem uma probabilidade de seleção aproximadamente igual à de um setor qualquer dentre os que foram substituídos. Ao todo, foram substituídos 78 setores.

2.2. A operação de listagem e a seleção de domicílios

A operação de listagem de setores tem por objetivo construir um cadastro, o mais atualizado possível, dos domicílios existentes nos setores selecionados para a amostra, a fim de permitir a seleção dos domicílios a serem investigados. Em função disso, a operação de listagem foi realizada em quatro etapas, cada uma abrangendo os setores de um trimestre da pesquisa.

Uma vez terminada a listagem dos setores, foi realizada a seleção dos domicílios, que, como definido anteriormente, foi feita com eqüiprobabilidade, considerando os tamanhos de amostra fixados, quais sejam 8 domicílios em cada setor dos estratos metropolitanos e urbanos, e 16 nos dois estratos rurais. Para contornar as possíveis recusas, domicílios vagos ou fechados na hora da realização da entrevista, foi definido um procedimento para substituição de domicílios, que consistiu na seleção de uma amostra reserva de domicílios em cada setor da amostra. Essa amostra foi selecionada previamente, utilizando o mesmo método utilizado na seleção da amostra principal. Ao todo foram

realizadas 4940 entrevistas entre as 4944 esperadas. Por problemas operacionais durante a coleta, em dois setores da amostra só foram realizadas 6 entrevistas.

3. A EXPANSÃO DA AMOSTRA

Cada domicílio da amostra representa um certo número de domicílios na população objetivo da pesquisa. Este número é o fator de expansão ou peso do domicílio que, associado às características investigadas na pesquisa, permite a obtenção de estimativas para o universo do qual a amostra foi selecionada. A etapa de expansão da amostra consiste basicamente na determinação do peso a ser associado a cada unidade da amostra.

3.1. A obtenção de pesos

Para a obtenção dos pesos ou fatores de expansão, foi utilizado o estimador natural, obtido pela fórmula diretamente associada ao plano amostral utilizado na PPV. Neste caso, o peso de cada domicílio é obtido pelo inverso da probabilidade de inclusão do domicílio na amostra.

O estimador natural para estimar o total da característica *y* é dado por:

$$\hat{Y} = \sum_{h=1}^{L} \sum_{i=1}^{m_h} \frac{1}{m_h \; P_{hi}} \sum_{j=1}^{n_{hi}} \frac{1}{n_{hi} \; P_{hij}} \; y_{hij}$$

onde:

L é o número de estratos;

m_h é o número de setores na amostra no estrato h;

P_{hi} é a probabilidade de seleção, num sorteio, do i-ésimo setor do estrato h;

$$P_{hi} = \frac{N_{hi}}{N_h}$$

N_{hi} é o número de domicílios particulares permanentes ocupados do i-ésimo setor do estrato h, obtidos pelo Censo 91;

N_h é o número de domicílios particulares permanentes ocupados do estrato h, obtidos pelo Censo 91;

n_{hi} é o número de domicílios com entrevista realizada no i-ésimo setor do estrato h;

P_{hij} é a probabilidade de seleção, num sorteio, do j-ésimo domicílio do i-ésimo setor do estrato h;

$$P_{hij} = \frac{1}{N^*_{hi}}$$

N^*_{hi} é o número de domicílios particulares ocupados do i-ésimo setor do estrato h, obtido pela listagem;

y_{hij} é o valor da característica y associado ao j-ésimo domicílio do i-ésimo setor do estrato h;

Seja:

p_{hij} o peso do j-ésimo domicílio do i-ésimo setor do estrato h;

$$p_{hij} = \frac{1}{m_h P_{hi}} \frac{1}{n_{hi} P_{hij}}$$

A expressão do estimador pode ser reescrita como:

$$\hat{Y} = \sum_{h=1}^{L} \sum_{i=1}^{m_h} \sum_{j=1}^{n_{hi}} p_{hij} y_{hij}$$

Cabe observar que ao considerar, no cálculo dos pesos, o número de domicílios com entrevista realizada em cada setor, ao invés do número de domicílios selecionados para a entrevista, o estimador resultante já considera o tratamento para a não resposta, através de um procedimento de reponderação da amostra.

Além disso, vale ressaltar que os pesos dos domicílios de um mesmo setor são constantes, variando apenas de setor para setor. Sendo assim, tem-se:

$p_{hi} = p_{hij}$ ($\forall j$ do i-ésimo setor do estrato h)

Portanto, têm-se pesos distintos para domicílios de um mesmo estrato, o que confirma a não autoponderação da amostra.

É importante analisar a variabilidade dos pesos, através de sua distribuição em cada estrato. A Tabela 2 a seguir apresenta essas distribuições.

Tabela 2 – Tamanho da amostra e estatísticas da distribuição dos pesos associados às unidades amostrais, por região da pesquisa

Região da Pesquisa	Tamanho da amostra		Média dos pesos	Distribuição dos pesos				
	Setores	Domicílios		Mínimo	1º Quartil	Mediana	3º Quartil	Máximo
Nordeste + Sudeste	554	4940	5505	724	1364	4034	8481	29234
Região Nordeste	278	2484	3675	724	1159	1407	6752	15348
Região Sudeste	276	2456	7349	991	2940	5892	10496	29234

3.2. A estimação de totais, proporções e razões

A estimação de totais é feita utilizando-se, para cada unidade (pessoa ou domicílio), o peso correspondente, que foi determinado para cada domicílio da amostra e atribuído a cada pessoa desse domicílio. Assim, para estimar o total de uma característica *y* utiliza-se o estimador definido na seção anterior, que pode ser expresso de forma resumida como segue:

$$\hat{Y}=\sum_{k=1}^{n} w_k \, y_k$$

onde:

w_k é o peso associado à k-ésima unidade da amostra (pessoa ou domicílio);

y_k é o valor da característica *y* associado à k-ésima unidade da amostra;

n é o número de unidades na amostra da área em questão.

Dessa forma, é possível calcular estimativas de totais para quaisquer variáveis investigadas na PPV, seja para características relativas a pessoas ou a domicílios.

As proporções são estimadas dividindo-se o número total estimado de unidades com uma determinada característica pelo número total estimado de unidades na população.

As razões da forma $R= Y / X$ são estimadas por $\hat{r}= \hat{Y} / \hat{X}$, onde $\hat{Y}$ e $\hat{X}$ são estimadores de total para as características consideradas no numerador e no denominador, respectivamente.

3.3. A precisão das estimativas

A avaliação da precisão das estimativas produzidas por pesquisas amostrais é um ponto fundamental do processo de produção de informações mediante amostragem. Dela depende o grau de confiança das conclusões analíticas advindas dos resultados da pesquisa.

Para cada estimativa derivada da pesquisa é possível obter uma medida de precisão, que é estimada com os próprios dados da pesquisa. Essa medida é estimada pelo coeficiente de variação (CV) do estimador obtido através da estimativa da variância do estimador. A expressão da variância é função da forma do estimador, do desenho amostral e do procedimento de expansão da amostra utilizado.

Dessa forma, aplicando o método clássico do *ultimate cluster*, (ver Hansen, Hurwitz e Madow, 1953 ou Wolter, 1985), em cada estrato da amostra, o estimador da variância do estimador de total de uma característica *y* é dado por:

$$\text{var}(\hat{Y}) = \sum_{h=1}^{L} \frac{m_h}{m_h - 1} \sum_{i=1}^{m_h} \left(\hat{Y}_{hi} - \hat{Y}_h \right)^2$$

onde:

$\hat{Y}_{hi}$ é o total expandido ao nível do *ultimate cluster*, ou seja:

$$\hat{Y}_{hi} = \sum_{j=1}^{n_{hi}} p_{hi} \; y_{hij}$$

$\hat{Y}_h$ é o total expandido da característica *y* no estrato h;

$$\hat{Y}_h = \sum_{i=1}^{m_h} \hat{Y}_{hi}$$

Portanto, o estimador do coeficiente de variação do estimador de total $\hat{Y}$ para a característica *y* é dado por:

$$cv(\hat{Y}) = \frac{\sqrt{\text{var}(\hat{Y})}}{\hat{Y}}$$

O estimador da variância de uma razão $\hat{r} = \hat{Y} / \hat{X}$, de acordo com Cochran (1977) é aproximado por:

$$\text{var}(\hat{r}) = \frac{1}{\hat{X}^2}\left(\text{var}(\hat{Y}) + \hat{r}^2 \ \text{var}(\hat{X}) - 2\,\hat{r}\ \text{cov}(\hat{Y}, \hat{X})\right)$$

onde:

$$\text{cov}(\hat{Y}, \hat{X}) = \sum_{h=1}^{L} \frac{m_h}{m_h - 1} \ \sum_{i=1}^{m_h} \left(\hat{Y}_{hi} - \hat{Y}_h\right)\left(\hat{X}_{hi} - \hat{X}_h\right)$$

Analogamente, o estimador para o coeficiente de variação da razão $\hat{r} = \hat{Y} / \hat{X}$ é dado por:

$$cv(\hat{r}) = \frac{\sqrt{\text{var}(\hat{r})}}{\hat{r}}$$

O coeficiente de variação é utilizado para construir intervalos de confiança que conterão o valor do universo com uma certa probabilidade decorrente do nível de confiança desejado na tomada de decisão.

Na prática, um intervalo de confiança de 95%, por exemplo, indica que, em cada 100 amostras selecionadas com o mesmo desenho e dimensão, 95 produzirão estimativas cujo intervalo de confiança conterá o valor do universo e em apenas 5 amostras este valor estará fora do intervalo de confiança.

4. REFERÊNCIAS

ALBIERI, S.; Bianchini, Z.M. e Cardoso, R.L. *Pesquisa domiciliar sobre padrões de vida: planejamento da amostra*. Rio de Janeiro: IBGE, Divisão de Metodologia e Departamento de Indicadores Sociais, 1995, 24p.

ALBIERI, S. e Bianchini, Z.M. *Aspectos de amostragem relativos à pesquisa domiciliar sobre padrões de vida.* Rio de Janeiro: IBGE, Departamento de Metodologia. 1997, 15p.

COCHRAN, W.G. *Sampling Techniques*. New York: John Wiley, 1977, 3ª Edição, 428p.

HANSEN, M.H., Hurwitz, W.N.; Madow, W.G. *Sample surveys methods and theory*. New York: John Wiley, 1953, v. 2.

PESQUISA de Orçamentos Familiares, volume 3 - Aspectos de amostragem. Rio de Janeiro: IBGE, 1992, 218p. [Série Relatórios Metodológicos, v.10].

SILVA, P.L.N. *Um algoritmo para seleção de amostras aleatórias simples sem reposição - aplicações na seleção de amostras do IBGE*. Rio de Janeiro: IBGE, Departamento de Coordenação de Métodos, 1983, 7p.

SILVA, P.L.N. *Macros para seleção de amostras*. Rio de Janeiro: IBGE, Núcleo de Metodologia, 1989, 64p.

WOLTER, K.M. *Introduction to variance estimation*. New York: Springer-Verlag, 1985.