Análise estatística de dados da PNAD: incorporando a estrutura do plano amostral

Statistical analysis of data from PNAD: incorporating the sample design

Pedro Luis do Nascimento Silva [1]
Djalma Galvão Carneiro Pessoa [1]
Maurício Franca Lila [2]

Abstract *This paper describes how the complex sample design aspects of stratification, clustering, unequal selection probabilities and calibration adjustments of the sample weights affecting the National Household Sample Survey (PNAD) of IBGE can be incorporated into the analysis of its data. For this purpose, it includes: a brief but comprehensive description of this survey's sample design; indication of how its data can be used for estimating totals; and also a description of the essential methods required to fit parametric models taking into account the complex sample survey design aspects mentioned. It also presents some estimates for household and personal characteristics obtained from PNAD/1998, for which standard errors and design effects are computed as an illustration. It concludes with an indication of the precautions users must take when analysing data from PNAD in their survey practice.*

Key words *Design effect, Estimation, Survey data analysis, Structured data, Household survey, Parametric models*

Resumo *Este artigo descreve como podem ser considerados na análise dos dados da Pesquisa Nacional por Amostra de Domicílios (PNAD) do IBGE os diversos aspectos de seu plano amostral complexo: estratificação, conglomeração, probabilidades desiguais de seleção e ajustes dos pesos para calibração. Para isso, inclui: uma descrição resumida porém completa do plano amostral dessa pesquisa; indicação de como seus dados podem ser usados para estimar totais; e também uma descrição resumida dos métodos essenciais para ajustar modelos paramétricos regulares com os dados da pesquisa levando em conta os aspectos de amostragem complexa. Apresenta ainda os resultados de algumas estimativas para características de pessoas e domicílios calculadas com base nos dados da PNAD/1998, para as quais são apresentadas estimativas dos respectivos desvios padrão e dos efeitos do plano amostral. Conclui com uma indicação dos cuidados que os usuários devem ter ao analisar tais dados em sua prática de pesquisa.*

Palavras-chave *Efeito do plano amostral, Estimação, Análise de dados amostrais, Dados estruturados, Pesquisa domiciliar, Modelos paramétricos*

[1] Departamento de Metodologia, Instituto Brasileiro de Geografia e Estatística. Av. Chile 500/10º andar 20031-170 Rio de Janeiro RJ. pedrosilva@ibge.gov.br
[2] Departamento de Emprego e Rendimento, Instituto Brasileiro de Geografia e Estatística.

Introdução

Este artigo trata de problema de grande importância para os usuários de dados obtidos através de pesquisas amostrais "complexas", como é o caso da PNAD (Pesquisa Nacional por Amostra de Domicílios, do IBGE – ver IBGE, 1981). Dados da PNAD são usados rotineiramente para análises descritivas que envolvem o cálculo, comparação e interpretação de estimativas para totais, médias, taxas, proporções e razões populacionais. Quando são considerados nos cálculos os pesos das unidades amostrais (fornecidos nos arquivos de microdados), as estimativas obtidas para os parâmetros populacionais correspondentes são não viciadas. A incorporação dos pesos na estimação de medidas descritivas, como as citadas acima, pode ser feita com simplicidade empregando-se as opções de ponderação disponíveis nos pacotes e sistemas estatísticos padrão, tais como SAS, SPSS, SPLUS, STATA e muitos outros.

Já se o interesse for a estimação de medidas de dispersão (variância, desvio padrão), concentração (índices de Gini e similares), função de distribuição empírica e quantis associados (quartis, decis, percentis, etc.), soluções adequadas muitas vezes ainda não estão disponíveis nos pacotes padrão. Isso ocorre porque a estimação de tais medidas requer considerar diversos aspectos adicionais do planejamento da amostra que foi usada para obter os dados além dos pesos usualmente disponíveis. Por esse motivo, a estimação de medidas de precisão das estimativas de medidas descritivas (desvios padrão e coeficientes de variação), que depende da estimação de variâncias e é essencial para análises qualificadas da significância dessas estimativas e de diferenças entre elas, também enfrenta as mesmas dificuldades.

É comum, também, a utilização de dados da PNAD para a construção e ajuste de modelos em análises secundárias usando, por exemplo, modelos de regressão (Reis *et al.*, 2001), modelos de regressão logística (Leote, 1996; Pessoa *et al.*, 1997), modelos de regressão multinomial logística (Leite, 2001), e outros. Tais análises, muitas vezes feitas por analistas que trabalham fora da agência produtora dos dados, freqüentemente utilizam para a modelagem pacotes estatísticos que se baseiam em hipóteses válidas somente quando os dados são obtidos através de amostras aleatórias simples com reposição (AASC). As exceções são os trabalhos de Pessoa *et al.* (1997) e Leite (2001).

Acontece que o plano (desenho) amostral da PNAD incorpora todos os aspectos que definem um "plano amostral complexo": estratificação das unidades de amostragem, conglomeração (seleção da amostra em vários estágios, com unidades compostas de amostragem), probabilidades desiguais de seleção em um ou mais estágios, e ajustes dos pesos amostrais para calibração com totais populacionais conhecidos. Por esse motivo, dados obtidos através das amostras das PNADs geralmente não podem ser tratados como se fossem observações independentes e identicamente distribuídas (isto é, como se tivessem sido gerados por amostras aleatórias simples com reposição), como fazem os procedimentos usuais de análise disponíveis nos pacotes estatísticos padrão.

As estimativas pontuais de medidas descritivas da população são influenciadas pelos pesos distintos das observações. Já as estimativas de variância e desvio padrão (medidas de precisão dos estimadores) e as estimativas de parâmetros para ajustes de alguns tipos de modelos são influenciadas conjuntamente pela estratificação, conglomeração e pesos. Ao ignorar esses aspectos, as técnicas e sistemas de análise tradicionais podem produzir resultados incorretos tanto para as estimativas pontuais como para os respectivos desvios padrão e níveis de significância, o que pode comprometer a qualidade do ajuste de modelos e a interpretação dos resultados obtidos.

O assunto tem recebido maior atenção nas últimas duas décadas, e já são muitos os recursos disponíveis para tornar mais fácil e prática a aplicação das técnicas de análise capazes de incorporar adequadamente os diversos aspectos de planos amostrais complexos, tanto na estimação de medidas descritivas e da precisão dessas estimativas, como no ajuste de modelos, no diagnóstico e avaliação de significância dos modelos ajustados, e na interpretação de resultados. Algumas referências úteis sobre o tema incluem: Pessoa e Nascimento Silva (1998), Skinner, Holt & Smith (1989), Korn e Graubard (1999), e Lehtonen e Pahkinen (1995), entre outras.

O objetivo deste artigo é apresentar uma descrição de como os métodos modernos de análise de dados incorporando os aspectos de complexidade do plano amostral podem ser aplicados para análise dos dados da PNAD, tomando como exemplo os dados coletados na edição de 1998 dessa pesquisa.

A seção 2 contém uma descrição do plano amostral utilizado na PNAD durante a década

Ciência & Saúde Coletiva, 7(4):659-670, 2002

de 1990, e indicações de como podem ser construídas as variáveis descritoras da estrutura do plano amostral a partir das informações existentes nos arquivos de microdados. A seção 3 apresenta estimadores para totais e suas variâncias, bem como o método de cálculo dos pesos que acompanham os microdados da PNAD. A seção 4 apresenta uma breve revisão dos métodos requeridos para ajuste de modelos paramétricos regulares com dados de pesquisas amostrais complexas, os quais formam a base para o desenvolvimento de pacotes estatísticos especializados tais como SUDAAN, entre outros. Na seção 5 são apresentadas estimativas de algumas medidas descritivas para variáveis de pessoas e domicílios com base na PNAD/1998, junto com uma avaliação do impacto de ignorar o plano amostral ao estimar a precisão destas estimativas. Finalmente, na seção 6 são discutidas as dificuldades encontradas pelos usuários dos dados da PNAD para incorporar adequadamente na modelagem aspectos importantes do plano amostral como os que aqui foram discutidos.

Plano amostral da PNAD

A PNAD é uma pesquisa anual por amostragem probabilística de domicílios, realizada em todo o território nacional exclusive a área rural da região Norte. A população alvo é composta pelos domicílios e pessoas residentes em domicílios na área de abrangência da pesquisa. A PNAD adota um plano amostral estratificado e conglomerado com um, dois ou três estágios de seleção, dependendo do estrato.

A estratificação da amostra básica da PNAD foi feita em duas etapas. Primeiro há uma estratificação geográfica que dividiu o país em 36 estratos "naturais". Nesta estratificação, 18 unidades da federação formaram cada uma um estrato independente para fins de amostragem. As outras nove unidades da federação (PA, CE, PE, BA, MG, RJ, SP, PR, RS) deram origem a outros 18 estratos, pois em cada uma delas foram definidos dois estratos naturais: um com todos os municípios da Região Metropolitana sediada na capital, e o outro com os demais municípios da unidade da federação.

Nos nove estratos naturais formados pelas regiões metropolitanas, o plano amostral da PNAD é estratificado adicionalmente por município e conglomerado em dois estágios. Nestes estratos (municípios), as unidades primárias de amostragem (UPAs) são os setores censitários. As unidades secundárias de amostragem (USAs) são os domicílios. Dentro de cada município, a seleção dos setores (UPAs) foi feita usando amostragem sistemática com probabilidades proporcionais ao tamanho (PPT), usando como medida de tamanho o número de domicílios conforme obtido do Censo Demográfico de 1991. Antes de efetuar a seleção dos setores em cada estrato (município), os setores foram ordenados segundo a situação (urbano, rural) e o código geográfico, conferindo um efeito de estratificação implícita por situação devido ao uso do sorteio sistemático.

Nos 27 estratos naturais formados com os municípios que não são situados em regiões metropolitanas ou ficam nas unidades da federação sem região metropolitana, o plano amostral da PNAD é conglomerado em três estágios. As unidades primárias de amostragem são os municípios, as unidades secundárias são os setores e as unidades terciárias de amostragem são os domicílios. Nesses estratos naturais, alguns municípios considerados grandes em termos populacionais foram incluídos na amostra com certeza. Tais municípios são chamados de auto-representativos. Os municípios auto-representativos são, portanto, estratos geográficos dentro dos quais o plano amostral é igual ao utilizado nos municípios das regiões metropolitanas, isto é, conglomerado em dois estágios, com os setores como unidades primárias de amostragem e os domicílios como unidades secundárias de amostragem.

Os demais municípios não situados nas regiões metropolitanas são chamados de não auto-representativos. Os municípios não auto-representativos foram estratificados por tamanho e proximidade geográfica, buscando formar estratos com população total aproximadamente igual, conforme os dados do último censo demográfico.

Em cada um dos estratos de municípios não auto-representativos, municípios foram selecionados através de sorteio sistemático, com probabilidades proporcionais à população existente na época do censo demográfico. No segundo estágio de seleção, o sorteio de setores foi feito dentro de cada município contido na amostra do primeiro estágio, usando o mesmo método já descrito para a seleção de setores nos estratos de regiões metropolitanas.

A cada ano, antes da última etapa de seleção da amostra (amostragem de domicílios), é feita uma Operação de Listagem dentro de ca-

da setor selecionado. Essa operação fornece o cadastro atualizado para a seleção de domicílios em cada setor, permitindo assim localizar, identificar e quantificar as unidades domiciliares ali existentes no ano de realização da pesquisa.

Usando a listagem atualizada de unidades domiciliares existentes nos setores da amostra, faz-se então a seleção das unidades domiciliares a serem pesquisadas a cada ano mediante amostragem sistemática simples. As unidades domiciliares são formadas pelos domicílios particulares e unidades de habitação em domicílios coletivos com moradores na ocasião da Operação de Listagem. Nos domicílios selecionados, as entrevistas são realizadas face a face, usando-se questionários em papel. Todos os moradores das unidades domiciliares selecionadas são incluídos na pesquisa.

A descrição acima indica como é selecionada a **amostra básica da PNAD**. Esta é complementada com unidades domiciliares selecionadas do Cadastro de Projetos de Novas Construções. Este cadastro inclui projetos habitacionais com mais de 30 domicílios que surgiram após o censo realizado na década. O universo das Novas Construções é estratificado por municípios, e nesses estratos o plano amostral é conglomerado em apenas um estágio, pois neste caso as unidades primárias de amostragem são os domicílios, cujo sorteio é efetuado mediante amostragem sistemática simples.

Em cada estrato natural, o plano amostral da PNAD é **autoponderado**, isto é, procura assegurar que todos os domicílios tenham igual probabilidade de seleção. Entretanto, as probabilidades de seleção podem variar bastante de um estrato natural para outro. No caso das 18 unidades da federação que formam cada uma um único estrato natural, a fração amostral é fixa e constante para todos os municípios. No caso das nove unidades da federação em que existem dois estratos naturais, os estratos naturais contendo os municípios das regiões metropolitanas podem ter frações amostrais diferentes dos estratos contendo os demais municípios que pertencem à mesma unidade da federação. O quadro 1 apresenta as frações amostrais usadas em cada um dos estratos naturais da pesquisa durante a década de 1990.

No momento em que foi feita a primeira seleção de setores no início da década, o número de domicílios a selecionar para a amostra por setor foi fixado em 13 e seria igual para todos os setores da amostra (Bianchini e Albieri, 1999). Quando as listagens de domicílios nos setores selecionados são atualizadas a cada ano, o número de domicílios a selecionar por setor pode variar, pois é mantido fixo o intervalo de seleção de domicílios calculado por ocasião da primeira seleção. Por exemplo, num setor onde o número de domicílios existente no ano da pesquisa tivesse dobrado em relação ao número existente no último censo demográfico, seria dobrado o número de domicílios a selecionar para a PNAD desse ano, passando de 13 para 26.

A figura 1 ilustra as partes do plano amostral da PNAD indicando, para cada parte, como devem ser construídas as variáveis que definem a estrutura do plano amostral. Vale notar que a primeira parte, referente à população residente em regiões metropolitanas, não existe em 18 das 27 unidades da federação.

Considerando este esquema geral do plano amostral da PNAD numa unidade da federação qualquer, pode-se empregar o algoritmo a se-

Quadro 1
Frações amostrais da PNAD por estratos naturais da pesquisa durante a década de 1990.

Áreas de abrangência	Fração amostral
Região metropolitana de Belém	1/150
Distrito Federal e regiões metropolitanas de Fortaleza, Recife, Salvador e Porto Alegre	1/200
Regiões metropolitanas de Belo Horizonte e Curitiba	1/250
Rondônia, Acre, Amazonas, Roraima, Amapá, Tocantins, Sergipe, Mato Grosso do Sul, Mato Grosso e Goiás	1/300
Pará	1/350
Piauí, Ceará, Rio Grande do Norte, Paraíba, Pernambuco, Alagoas, Bahia, Minas Gerais, Espírito Santo, Rio de Janeiro e região metropolitana do Rio de Janeiro	1/500
Paraná, Santa Catarina e Rio Grande do Sul	1/550
Maranhão, São Paulo e região metropolitana de São Paulo	1/750

guir para definir os valores das variáveis que descrevem a estrutura do plano amostral.

Algoritmo para criação das variáveis que definem a estrutura do plano amostral da PNAD (ESTRATO e UPA)

Este algoritmo é descrito como deve ser aplicado para os registros de domicílios nos arquivos de microdados da PNAD. Uma vez criadas as variáveis de estrutura do plano amostral para os domicílios, estas podem ser repassadas para os registros das pessoas moradoras correspondentes. Note que a variável "município" está contida na variável denominada UPA no arquivo de domicílios da PNAD. A nova variável UPA criada no algoritmo abaixo deve ser guardada em nome distinto.

Processa amostra básica

Domicílio de região metropolitana ou município auto-representativo
SE (1<=V4107<=2) ENTÃO FAÇA:
    ESTRATO = UF*100000000 + MUNICÍPIO.
    UPA = V0102*1000;
FIM1.

Domicílio na amostra de município não auto-representativo
SE (V4107=3) ENTÃO FAÇA:
    ESTRATO =UF*100000000 + 99*1000000 + V4602*10000;
    UPA= UF*1000000 + V4602*10000 + MUNICÍPIO;
FIM2.

Processa amostra de novas construções

SE Novas Construções ENTÃO FAÇA:
    ESTRATO = UF*100000000 + 98*1000000 + MUNICÍPIO;
    UPA=V0102*1000+V0103;
FIM3.

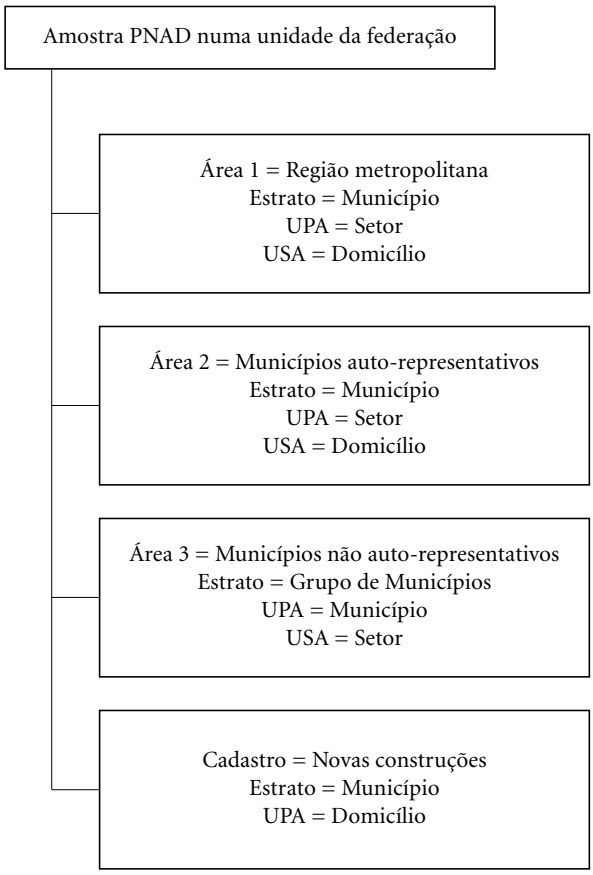
Estimação de totais na PNAD

Boa parte das estimativas publicadas pela PNAD corresponde à estimação de totais populacionais. Além disso, a estimação de totais é a base sobre a qual se assenta a estimação de médias, razões, taxas e proporções. Por esse motivo, apresenta-se aqui uma breve revisão de como são estimados totais usando os dados da amostra da PNAD para um ano qualquer (o ano de 1998 não foge à regra). O estimador simples para o total de uma característica y qualquer observada na amostra da PNAD para um estrato natural especificado é definido por:

$$\hat{Y} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} d_{hij}\, y_{hij} \qquad (1)$$

onde H é o número de estratos existentes no estrato natural, n_h é número de unidades primárias de amostragem (UPAs) selecionadas para a amostra no estrato h, n_{hi} é numero de unidades elementares de interesse (domicílios ou pessoas) pesquisadas na amostra da UPA i do estrato h, d_{hij} é o peso amostral básico da j-ési-

Figura 1
Ilustração do plano amostral da PNAD durante a década de 1990.



ma unidade elementar pesquisada na UPA *i* do estrato *h*, e y_{hij} é o valor observado da variável de interesse *y* para a *j*-ésima unidade elementar pesquisada na UPA *i* do estrato *h*, cujo total se deseja estimar.

Os pesos amostrais d_{hij} correspondem aos valores inversos das probabilidades de inclusão dos domicílios em cada estrato, isto é, aos denominadores das frações amostrais apresentadas no quadro 1. Variam, portanto, entre 150 e 750, dependendo do estrato natural a que pertence a unidade pesquisada. Como todas as pessoas residentes num domicílio selecionado são pesquisadas (não há sorteio de moradores), todas recebem o peso calculado para o domicílio. Vale destacar que não são estes os pesos usualmente gravados nos arquivos de dados da PNAD, como se verá mais adiante.

Um estimador da variância do estimador simples do total $\hat{Y}$ pode ser obtido usando:

$$\hat{V}(\hat{Y}) = \sum_{h=1}^{H} \frac{s_{hy}^2}{n_h} \qquad (2)$$

onde $s_{hy}^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} \left(\frac{\hat{Y}_{hi}}{p_{hi}} - \hat{Y}_h \right)^2$,

$\hat{Y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{\hat{Y}_{hi}}{p_{hi}}$,

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

e p_{hi} é o tamanho relativo da UPA *i*, no estrato *h*, conforme o último censo demográfico.

Este estimador em (2) corresponde à aproximação do plano amostral PPT sistemático adotado para seleção da amostra de unidades primárias na PNAD por um plano PPT com reposição no momento de estimar variâncias das estimativas, chamado de método do Conglomerado Primário por Pessoa e Nascimento Silva (1998). O método está descrito também em Cochran (1977) ou Korn e Graubard (1999). Essa é a opção usual em casos como esse, porque quando o plano amostral tem sorteio sistemático de UPAs não existem estimadores exatamente não viciados de variância das estimativas pontuais de totais. O estimador de variância adotado é considerado "conservador", no sentido de que seu valor esperado deve ser ligeiramente maior que a variância do estimador sob o plano efetivamente utilizado que tem sorteio sem reposição das UPAs. Na PNAD, o vício desse estimador de variância deve ser modesto, de vez que a fração amostral é no máximo igual a 1/150 (ver quadro 1), o que faz com que o efeito do fator de correção de população finita (*1 - f*) desprezado com a aproximação adotada seja muito próximo de 1 (*1 - 1/150 = 149/150 ≅ 0,993*).

O estimador (1) é não viciado para o total populacional de *y* no estrato natural, mas pode ser melhorado com a incorporação de ajustes de calibração que aproveitam informações populacionais auxiliares disponíveis. Na PNAD, o método efetivamente empregado no processo de expansão da amostra utiliza estimadores de razão (caso particular dos estimadores de calibração – ver Särndal, Swensson & Wretman, 1992), considerando informação auxiliar as projeções independentes da população total para cada um dos 36 estratos naturais.

O estimador de razão empregado em um estrato natural qualquer é definido como:

$$\hat{Y}^R = \hat{Y} \times \frac{P}{\hat{P}} = P \times \frac{\hat{Y}}{\hat{P}} = P \times \hat{R} \qquad (3)$$

onde *P* representa a população residente *projetada* para o estrato natural obtida através de um processo de projeção independente da amostra, e é o total estimado da população residente no estrato natural através da amostra com base no estimador simples do plano amostral, isto é,

$$\hat{P} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} d_{hij} \, x_{hij},$$

onde x_{hij} é o número de moradores do *j*-ésimo domicílio pesquisado na UPA *i* do estrato *h*.

Correspondendo ao estimador (3) para o total, cada unidade amostrada tem um peso ajustado, que é calculado e adicionado aos registros de dados da PNAD. Esse peso ajustado corresponde ao valor do peso básico d_{hij} referente ao estimador (1) multiplicado pela razão ou fator de ajuste *P/P*, e é dado por

$$w_{hij} = d_{hij} \, \frac{P}{\hat{P}} \qquad (4)$$

Com esses pesos, o estimador de razão (3) para o total populacional da variável de interesse *y* pode ser escrito como um estimador linear, da forma

$$\hat{Y}^R = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} w_{hij} \, y_{hij},$$

e fica portanto igualmente simples de calcular usando qualquer pacote estatístico padrão, desde que os pesos corretos w_{hij} sejam considera-

Ciência & Saúde Coletiva, 7(4):659-670, 2002

dos, motivo da simplicidade da estimação pontual de totais, médias, taxas e razões partindo da amostra da PNAD.

Todas as pessoas residentes num domicílio recebem o peso w_{hij} calculado para o domicílio onde residem. Os pesos assim ajustados, quando usados para estimar o total da população em cada estrato natural, produzem uma estimativa que é igual ao valor da população residente projetada para o estrato natural pelo IBGE, conferindo assim a propriedade de *calibração* no total populacional à amostra da PNAD.

A variância do estimador $\hat{Y}^R$ pode ser estimada usando o método de linearização de Taylor (Pessoa e Nascimento Silva, 1998; Korn e Graubard, 1999) através da expressão:

$$\hat{V}(\hat{Y}^R) = \sum_{h=1}^{H} \frac{1}{n_h} \left[s_{hy}^2 + \hat{R}^2 s_{hp}^2 - 2\hat{R}\, s_{hpy} \right] \quad (5)$$

onde $s_{hp}^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} \left(\frac{\hat{P}_{hi}}{P_{hi}} - \hat{P}_h \right)^2$,

$$s_{hpy}^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} \left(\frac{\hat{P}_{hi}}{P_{hi}} - \hat{P}_h \right) \left(\frac{\hat{Y}_{hi}}{P_{hi}} - \hat{Y}_h \right),$$

$$\hat{P}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{\hat{P}_{hi}}{p_{hi}}, \text{ e}$$

$$\hat{P}_{hi} = \sum_{j=1}^{n_{hi}} n_h\, p_{hi}\, d_{hij}\, x_{hij}.$$

Usando (5), estimativas dos desvios padrão (DPs) e coeficientes de variação (CVs) associados às estimativas de totais da PNAD podem ser facilmente calculadas usando, respectivamente, $dp\,(\hat{Y}^R) = \sqrt{\hat{V}(\hat{Y}^R)}$ e $cv\,(\hat{Y}^R) = \sqrt{\hat{V}(\hat{Y}^R)} \,/\, \hat{Y}^R$.

Para obter estimativas de total e das respectivas variâncias para áreas definidas como agregações de estratos naturais (como por exemplo, os totais de unidades da federação ou os totais nacionais), basta somar as estimativas dos totais e das respectivas variâncias obtidas usando (3) e (5) para todos os estratos naturais componentes da área de interesse.

Vale aqui notar que os procedimentos usuais dos pacotes estatísticos padrão não permitem estimar diretamente as variâncias e os desvios padrão das estimativas de totais considerando as fórmulas aqui apresentadas. Entretanto, já há vários pacotes estatísticos especializados para estimação em pesquisas amostrais complexas, entre os quais se destaca o SUDAAN (ver a revisão no último capítulo de Pessoa e Nascimento Silva, 1998). Mais recentemente, começaram a ficar disponíveis procedimentos implementando essa metodologia de estimação de totais e suas variâncias incorporando o plano amostral em alguns dos pacotes estatísticos padrão, entre os quais o SAS, o STATA, e as funções em R desenvolvidas por Pessoa (2002).

Ajuste de modelos considerando o plano amostral

Esta seção descreve resumidamente o método de Máxima Pseudoverossimilhança (MPV), devido a Binder (1983), comumente empregado para ajuste de modelos paramétricos quando se considera o plano amostral (estratificação, conglomeração, etc.) e os pesos no processo de inferência com dados de amostras complexas. O material aqui apresentado é resumido da discussão apresentada em Pessoa e Nascimento Silva (1998).

Seja $y_j=(y_{j1}, ..., y_{jR})'$ o vetor Rx1 das variáveis de pesquisa observadas para a unidade elementar j, gerado por um vetor aleatório Y_j, para $j \in U$, onde $U=\{1, ..., N\}$ é o conjunto de rótulos das unidades elementares da população de interesse. Suponha também que $Y_1,...,Y_N$ são independentes e identicamente distribuídos com densidade $f(y;\theta)$, onde $\theta = (\theta_1, \theta_2,..., \theta_K)$ é o vetor Kx1 de parâmetros desconhecidos de interesse. Se todas as unidades elementares da população finita U fossem pesquisadas, a função de log-verossimilhança populacional seria dada por:

$$L_U(\theta) = \sum_{j \in U} \log[f(y_j; \theta)] \quad (6)$$

Sob certas condições de regularidade, igualando-se as derivadas parciais de $L_U(\theta)$ com relação a cada componente de θ a 0, temos um sistema de equações $\sum_{j \in U} u_j(\theta) = \mathbf{0}$,

onde $u_j(\theta) = \partial \log[f(y_j; \theta)]/\partial\theta$ é o vetor Kx1 dos escores da unidade elementar j, para $j \in U$. A solução θ_U deste sistema seria o estimador de Máxima Verossimilhança de θ no caso de um censo. Podemos considerar θ_U uma quantidade desconhecida da população finita, sobre a qual se deseja fazer inferências baseadas em informações da amostra. Para populações onde N for grande, θ_U será muito próximo de θ, e fazer inferência para θ_U será o mesmo que fazer inferência para θ.

Seja $T(\theta) = \sum_{j \in U} u_j(\theta)$

a soma dos escores, que é um vetor de totais populacionais. Para estimar este vetor de totais,

pode-se usar um estimador linear ponderado da forma $\hat{T}(\theta) = \sum_{j \in s} w_j u_j(\theta)$,

onde os w_j são pesos amostrais adequados para a estimação de totais populacionais a partir da amostra *s*, tais como os implicados pelos estimadores (1) ou (3) por exemplo. O vetor de parâmetros θ do modelo definido por $f(y;\theta)$ para a população finita pode ser estimado usando o estimador de Máxima Pseudoverossimilhança $\hat{\theta}_{MPV}$ que é um valor de θ que serve de solução das equações dadas por

$$\hat{T}(\theta) = \sum_{j \in s} w_j u_j(\theta) = 0 \qquad (7)$$

A variância assintótica do estimador $\hat{\theta}_{MPV}$, sob a distribuição conjunta gerada pelo modelo e o plano amostral, pode ser estimada por:

$$\hat{V}(\hat{\theta}_{MPV}) = [\hat{J}(\hat{\theta}_{MPV})]^{-1}\, \hat{V}\left[\sum_{j \in s} w_j u_j(\hat{\theta}_{MPV})\right] [\hat{J}(\hat{\theta}_{MPV})]^{-1} \qquad (8)$$

onde $\hat{J}(\hat{\theta}_{MPV}) = \dfrac{\partial \hat{T}(\theta)}{\partial \theta}\Big|_{\theta = \hat{\theta}_{MPV}} =$

$$= \sum_{j \in s} w_j \frac{\partial u_j(\theta)}{\partial \theta}\Big|_{\theta = \hat{\theta}_{MPV}} \text{ e}$$

$$\hat{V}\left[\sum_{j \in s} w_j u_j(\hat{\theta}_{MPV})\right]$$

é um estimador consistente para a matriz de variância (do desenho) do estimador do total populacional dos escores, obtido por exemplo usando (5) no caso da PNAD.

Muitos modelos paramétricos podem ser ajustados empregando o método da Máxima Pseudoverossimilhança para estimar os parâmetros, com dados obtidos através de diferentes planos amostrais. Os estimadores de MPV não serão únicos, entretanto, já que existem diversas maneiras de se definir os pesos w_j correspondentes a diferentes estimadores de totais. Os pesos mais usados são os do estimador simples para totais-estimador (1). No caso da PNAD, são usados os pesos (4) correspondentes ao estimador de razão (3). Dependendo do modelo que se quer ajustar, basta calcular os escores $u_j(\theta)$ adequados e usar os estimadores de total (3) e da correspondente variância (5) para calcular as estimativas pontuais $\hat{\theta}_{MPV}$ dos parâmetros θ do modelo e as estimativas da matriz de variâncias $\hat{V}(\hat{\theta}_{MPV})$, mediante as expressões (7) e (8) devidamente adaptadas. Tais estimativas de $\hat{\theta}_{MPV}$ e $\hat{V}(\hat{\theta}_{MPV})$ podem então ser usadas para calcular intervalos de confiança ou estatísticas de teste baseadas na distribuição assintótica normal para fazer inferência sobre os componentes de θ (Binder, 1983).

Para amostras autoponderadas (como é o caso da PNAD dentro de um estrato natural qualquer), os pesos w_j serão constantes e o estimador pontual $\hat{\theta}_{MPV}$ será idêntico ao estimador usual de Máxima Verossimilhança (MV) em uma amostra de observações independentes e identicamente distribuídas com distribuição $f(y;\theta)$. Porém o mesmo não ocorre quando se trata da variância do estimador de θ, pois esta é afetada por outros aspectos do plano amostral, tais como a estratificação e conglomeração. Mesmo para amostras em que o estimador pontual coincide com o estimador usual de Máxima Verossimilhança, a estimativa da variância obtida pelo procedimento de MPV é preferível à estimativa usual da variância baseada no método de MV, pois esta última desconsidera os efeitos do plano amostral usado para obter os dados. Além disso, para áreas definidas por agregações de estratos naturais com frações amostrais distintas, nem mesmo as estimativas pontuais θ de obtidas por MPV coincidirão com as estimativas obtidas por Máxima Verossimilhança.

O procedimento de MPV proporciona estimativas consistentes e razoavelmente simples de calcular tanto para os parâmetros como para as variâncias dos estimadores pontuais dos parâmetros. Este procedimento é a base para o desenvolvimento de vários pacotes computacionais especializados, tais como SUDAAN, ou de procedimentos capazes de incorporar adequadamente os efeitos de planos amostrais complexos já disponíveis em pacotes padrão tais como SAS e STATA, entre outros.

Por outro lado, o procedimento de MPV requer conhecimento de informações detalhadas sobre a estrutura do plano amostral para cada uma das unidades da amostra, tais como pertinência a estratos e conglomerados ou unidades primárias de amostragem, e seus respectivos pesos. Além disso, as propriedades dos estimadores de MPV não são conhecidas para pequenas amostras. Este problema não será obstáculo em análises que usam os dados da amostra inteira da PNAD, ou, no caso de domínios de estudo separados, quando estes tiverem amostras suficientemente grandes. Porém, tal dificuldade deve ser considerada quando as amostras nos domínios de interesse forem pequenas em termos do número de unidades primárias amostradas no domínio. Outra dificuldade do

procedimento é que não podem ser utilizados métodos usuais de diagnóstico e outros procedimentos da inferência clássica, tais como gráficos de resíduos e testes estatísticos de Razões de Verossimilhança. Entretanto, há recursos alternativos para diagnóstico que consideram os efeitos dos diferentes aspectos do desenho amostral complexo empregado (Eltinge, 1999 ou Korn e Graubard, 1999).

Estimativas de efeitos do plano amostral para variáveis selecionadas na PNAD/1998

Como forma de ilustrar o efeito de ignorar o plano amostral e os pesos na análise de dados da PNAD, foram calculadas estimativas para algumas medidas descritivas, juntamente com os respectivos desvios padrão, usando os dados da PNAD/1998 e aplicando os métodos descritos nas seções 3 e 4. Tais estimativas foram calculadas utilizando o pacote SUDAAN (Shah *et al.*, 1995), de forma que foram incorporados os efeitos do plano amostral (estratificação, conglomeração, sorteio PPT das UPAs) e do ajuste dos pesos para calibração nos totais populacionais de pessoas por estrato natural ao calcular as estimativas de variâncias e desvios padrão das estimativas pontuais de médias e proporções.

Qualquer sistema empregado para estimar os desvios padrão das estimativas amostrais com dados da PNAD (SUDAAN não foge à regra) requer informação sobre três aspectos do plano amostral para poder calcular corretamente as estimativas. Primeiro, é preciso indicar qual o tipo de plano amostral e/ou estimador de variância deve ser usado. A opção adequada de plano amostral e estimador de variância a ser utilizada quando se emprega o SUDAAN é DESIGN=WR, que corresponde à aproximação do plano amostral PPT sistemático adotado para seleção da amostra por um plano PPT com reposição no momento de estimar variâncias das estimativas, e à aplicação das fórmulas relevantes para estimação de variâncias apresentadas nas seções 3 e 4 deste artigo. Segundo, é necessário identificar a estrutura do plano amostral, isto é, a que estrato e unidade primária de amostragem pertence cada unidade amostral elementar (domicílio ou pessoa). Para este fim, devem ser usadas as variáveis ESTRATO e UPA construídas com o algoritmo apresentado no anexo 1. Por último, falta indicar qual é o peso da unidade amostral a ser usado no cálculo das estimativas. Os arquivos de microdados da PNAD fornecem essa informação já pronta. Para 1998, trata-se da variável V4729 do arquivo de pessoas, ou V4611 do arquivo de domicílios. Esses pesos já são os pesos ajustados (ou calibrados) definidos em (4).

Usando essas informações e considerando os dados de pessoas e domicílios da PNAD/1998 foram produzidas as estimativas das tabelas 1 e 2, respectivamente. Nessas tabelas, a última coluna apresenta estimativas do EPA (Efeito do Plano Amostral – ver Pessoa e Nascimento Silva, 1998), definido como a razão da variância obtida considerando o plano amostral através da metodologia descrita na seção 3, e a variância obtida ignorando o plano amostral (isto é, a variância estimada como se a amostra fosse AASC). Valores de EPA afastados de 1 indicam que ignorar o plano amostral na estimação da variância leva a estimativas viciadas e incorretas. Valores grandes (> 1) de EPA indicam que o estimador "ingênuo" da variância obtido ignorando o plano amostral complexo leva a subestimar a variância verdadeira do estimador.

As estimativas apresentadas nas tabelas 1 e 2 se referem ao total do país menos a zona rural da região Norte (área de abrangência da PNAD). Um exame dos valores dos EPAs apresentados nessas tabelas revela com clareza que ignorar o plano amostral é contra-indicado no caso da PNAD/1998. Para as variáveis de pessoas consideradas, os EPAs variam de 1,9 a 13,7, com um valor médio de 5,5. Isto indica que estimativas ingênuas de variância teriam valor esperado muito menor que os valores das variâncias sob o plano amostral efetivamente utilizado. Este efeito é maior para variáveis com grande homogeneidade intraconglomerados, como é o caso das variáveis nas linhas 1 e 2 da tabela 1. Nota-se também que o efeito do plano amostral pode variar bastante de uma variável para outra.

Já para as variáveis de domicílio (tabela 2), os EPAs variam entre 2,3 e 8,4, com média de 4,7. Embora menos dispersos, os valores dos EPAs para domicílios também indicam que é inadequada a opção de ignorar o plano amostral ao tentar estimar a precisão de estimativas derivadas da PNAD/1998. Verifica-se também a mesma diferenciação do EPA entre distintas variáveis, tendo maiores valores ocorrido para as variáveis cuja homogeneidade intraconglomerados é maior (linhas 8, 10, 14 e 15 da tabela 2).

Todas as estimativas apresentadas nas tabelas 1 e 2, como derivam do uso da amostra in-

Tabela 1
Estimativas, desvios padrão, coeficientes de variação e efeitos do plano amostral para variáveis de pessoas – PNAD – 1998.

Linha	Descrição da variável	Estimativa	Desvio padrão	CV(%)	EPA
1	Proporção de pessoas brancas	53,8%	0,3%	0,6	13,7
2	Proporção de pessoas negras ou pardas	45,4%	0,3%	0,7	13,7
3	Proporção de pessoas analfabetas	24,4%	0,2%	0,7	5,8
4	Proporção de pessoas que freqüentam escola	30,9%	0,1%	0,4	2,3
5	Proporção de pessoas exercendo trabalho infantil	2,8%	0,2%	5,2	2,6
6	Proporção de pessoas que trabalham	54,8%	0,2%	0,3	3,4
7	Proporção de pessoas empregadas	2,7%	0,1%	2,9	8,4
8	Proporção de pessoas conta própria	2,7%	0,1%	2,5	6,2
9	Proporção de pessoas empregadoras	0,3%	0,0%	5,3	3,0
10	Proporção de pessoas com auxílio-moradia	7,8%	0,2%	2,4	4,5
11	Proporção de pessoas com auxílio-alimentação	37,2%	0,3%	0,8	3,3
12	Proporção de pessoas com auxílio-transporte	34,2%	0,3%	0,9	3,7
13	Proporção de pessoas com auxílio-creche/educação	2,6%	0,1%	2,8	1,9
14	Proporção de pessoas com auxílio-saúde	16,5%	0,3%	1,6	4,8
15	Renda média do trabalho principal	512,8	5,8	1,1	5,4
16	Proporção de pessoas com previdência	44,2%	0,3%	0,7	5,6

Tabela 2
Estimativas, desvios padrão, coeficientes de variação e efeitos do plano amostral para variáveis de domicílios – PNAD – 1998.

Linha	Descrição da variável	Estimativa	Desvio padrão	CV(%)	EPA
1	Proporção com paredes de material adequado	96,0%	0,2%	0,2	6,1
2	Proporção com cobertura de material adequado	97,1%	0,1%	0,1	5,8
3	Número médio de cômodos por domicílio	5,65	0,0166	0,3	4,9
4	Número médio de cômodos servindo de dormitório	1,97	0,0043	0,2	2,3
5	Proporção de domicílios próprios	74,3%	0,2%	0,3	2,8
6	Proporção de domicílios alugados	13,5%	0,2%	1,3	2,4
7	Média do aluguel	223,2	3,0	1,4	2,5
8	Proporção com terreno próprio	92,3%	0,3%	0,3	7,4
9	Proporção com água canalizada pelo menos um cômodo	84,8%	0,3%	0,3	4,3
10	Proporção com água de rede geral	89,0%	0,3%	0,4	8,4
11	Proporção com água canalizada de rede geral	23,9%	0,9%	3,6	5,7
12	Proporção com água de poço ou nascente	52,0%	1,3%	2,5	6,6
13	Proporção com ao menos um banheiro	91,0%	0,2%	0,2	5,7
14	Proporção com esgotamento adequado	70,2%	0,4%	0,6	7,7
15	Proporção com energia elétrica	94,2%	0,2%	0,2	7,2
16	Proporção com telefone	31,7%	0,3%	1,0	4,6
17	Proporção com filtro d'água	56,2%	0,3%	0,5	3,0
18	Proporção com rádio	90,4%	0,2%	0,2	2,6
19	Proporção com TV em cores	78,0%	0,3%	0,3	3,8
20	Proporção com TV em preto e branco	43,6%	0,6%	1,4	2,8
21	Proporção com geladeira	81,7%	0,3%	0,3	3,7
22	Proporção com freezer	19,5%	0,2%	1,2	3,0
23	Proporção com máquina de lavar roupa	32,0%	0,3%	1,0	3,9

Ciência & Saúde Coletiva, 7(4):659-670, 2002

teira da PNAD/1998 em nível nacional (90.913 domicílios com entrevistas realizadas e 344.975 pessoas entrevistadas), apresentam elevado grau de precisão (seus coeficientes de variação estimados variam entre 0,1% e 5,3%, com valor médio de 1,2%). Quando a amostra da PNAD for utilizada para estimar para domínios de estudo mais detalhados (estados, regiões metropolitanas, e outros), há que prestar maior atenção aos valores dos desvios padrão e/ou coeficientes de variação das estimativas, pois estas podem ser imprecisas. Nascimento Silva e Pessoa (2002) observaram, por exemplo, que estimativas diretas e indiretas das taxas de mortalidade infantil obtidas dos dados de fecundidade da PNAD podem ser bastante imprecisas para alguns estados da federação.

Como os efeitos do plano amostral sobre as estimativas de variância não são uniformes para diferentes variáveis, ao contrário, são bastante diversos, a prática recomendada é sempre buscar calcular estimativas das medidas de precisão das estimativas de interesse considerando todos os aspectos relevantes do plano amostral. Hoje em dia, isso não representa mais um problema sério, de vez que estão disponíveis recursos computacionais adequados para esse fim.

Comentários finais

Uma das principais dificuldades que os usuários da PNAD têm para considerar adequadamente os efeitos do plano amostral complexo utilizado na hora de fazer suas análises é a pouca exposição aos métodos e técnicas necessários para fazer uso correto dos dados. Este artigo busca enfrentar essa dificuldade, apresentando uma exposição compreensiva, embora resumida, dos métodos e técnicas disponíveis para estimação e análise de dados de pesquisas amostrais complexas, como é o caso da PNAD.

Outra dificuldade enfrentada pelos usuários é a decodificação das informações sobre a metodologia da PNAD de maneira a aplicarem corretamente os métodos aqui expostos, com auxílio dos pacotes computacionais especializados disponíveis. Esta dificuldade também foi atacada com a exposição detalhada dos métodos de amostragem e estimação usados na PNAD, e de como as informações sobre a estrutura do plano amostral podem ser trabalhadas para uso num pacote estatístico especializado (SUDAAN). Usuários de outros pacotes podem aproveitar imediatamente grande parte da informação para uso com seus pacotes preferidos, desde que baseados em metodologia similar para estimação de variâncias.

Por último, outra dificuldade dos usuários é aceitar que a idéia de usar os pacotes estatísticos padrão nas análises pode levar a resultados incorretos na inferência. Foi demonstrada de maneira incontestável com os valores das estimativas de EPA apresentados para uma amostra intencional de variáveis da PNAD que tais efeitos não podem ser ignorados, sob pena de inferências grosseiramente viciadas. Como tais efeitos são importantes para um número grande de variáveis de tipos diferentes (tanto características de pessoas como de domicílios foram consideradas), e variam bastante de uma variável para outra, a lição a ser extraída é que as análises devem sempre considerar os aspectos relevantes do plano amostral da PNAD.

Referências bibliográficas

Bianchini ZM & Albieri S 1999. Uma revisão dos principais aspectos dos planos amostrais das pesquisas domiciliares realizadas pelo IBGE. *Revista Brasileira de Estatística* 60(213):7-23.

Binder DA 1983. On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review* 51:279-292.

Cochran WG 1977. *Sampling techniques*. (3ª ed.) John Wiley and Sons, Nova York.

Eltinge J 1999. Assessment of information capacity and sensitivity in the analysis of complex surveys. *Bulletin of the International Statistical Institute*, Proceedings of the 52nd session, Tomo LVIII. Helsinque.

IBGE 1981. *Metodologia da Pesquisa Nacional por Amostra de Domicílios na Década de 70*. Rio de Janeiro. Série Relatórios Metodológicos, volume 1.

Korn EL & Graubard BI 1999. *Analysis of health surveys*. John Wiley and Sons, Nova York.

Lehtonen R & Pahkinen EJ 1995. Practical methods for design and analysis of complex surveys. John Wiley & Sons, Chichester.

Leite PGPG 2001. *Análise da situação ocupacional de crianças e adolescentes nas regiões Sudeste e Nordeste do Brasil utilizando informações da PNAD/1999*. Dissertação de mestrado da Escola Nacional de Ciências Estatísticas, Rio de Janeiro.

Leote RMD 1996. *Um perfil socioeconômico das pessoas ocupadas no setor informal na área urbana do Rio de Janeiro*. Relatórios Técnicos nº 02/96. Escola Nacional de Ciências Estatísticas, Rio de Janeiro.

Nascimento Silva PL 1996. *Utilizing auxiliary information for estimation and analysis in sample surveys*. Tese de doutorado, Universidade de Southampton.

Nascimento Silva PL & Pessoa DGC 2002. *Estimando a precisão das estimativas indiretas das taxas de mortalidade obtidas a partir da PNAD*. Trabalho aceito para o XIII Encontro da ABEP.

Pessoa DGC 2002. ADAC: Biblioteca de Funções em R para a Análise de Dados Amostrais Complexos. 15º Simpósio Nacional de Probabilidade e Estatística. Associação Brasileira de Estatística, São Paulo.

Pessoa DGC, Nascimento Silva PL & Duarte RPN 1997. Análise estatística de dados de pesquisas por amostragem: problemas no uso de pacotes padrões. *Revista Brasileira de Estatística* 58(210):53-75.

Pessoa DGC & Nascimento Silva PL 1998. *Análise de dados amostrais complexos*. Associação Brasileira de Estatística, São Paulo.

Reis EJ, Tafner P & Reiss LO 2001. *Distribuição de riqueza imobiliária e de renda no Brasil: 1992-1999*. IPEA-DIMAC, Rio de Janeiro.

Särndal CE, Swensson B & Wretman JH 1992. *Model assisted survey sampling*. Springer-Verlag, Nova York.

Shah BV *et al.* 1995. *Statistical methods and mathematical algorithms used in SUDAAN*. Research Triangle Institute.

Skinner CJ, Holt D & Smith TMF (eds.). 1989. *Analysis of complex surveys*. John Wiley & Sons, Chichester.

Artigo apresentado em 18/9/2002
Aprovado em 31/10/2002
Versão final apresentada em 11/11/2002