Universidade Federal de Juiz de Fora

Instituto de Ciências Exatas

Curso de Estatística

Vítor da Fonseca Vieira

Impactos do Desenho Amostral Complexo na Análise de Dados da PNAD

Juiz de Fora

2015

Vítor da Fonseca Vieira

Impactos do Desenho Amostral Complexo na Análise de Dados da PNAD

Monografia apresentada ao Curso de Estatística da Universidade Federal de Juiz de Fora, como requisito parcial para obtenção do título de Bacharel em Estatística.

Orientador: Marcel de Toledo Vieira

Juiz de Fora

2015

Ficha catalográfica elaborada através do programa de geração
automática da Biblioteca Universitária da UFJF,
com os dados fornecidos pelo(a) autor(a)

Vieira, Vítor da Fonseca.
Impactos do Desenho Amostral Complexo na Análise de Dados da PNAD / Vítor da Fonseca Vieira. -- 2015.
44 p. : il.

Orientador: Marcel de Toledo Vieira
Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Juiz de Fora, Instituto de Ciências Exatas, 2015.

1. Amostragem complexa. 2. EPA. 3. Inferência. 4. Renda domiciliar. I. Vieira, Marcel de Toledo, orient. II. Título.

Vítor da Fonseca Vieira

Impactos do Desenho Amostral Complexo na Análise de Dados da PNAD

Monografia apresentada ao Curso de Estatística da Universidade Federal de Juiz de Fora, como requisito parcial para obtenção do título de Bacharel em Estatística.

Aprovada em: 24/06/2015

BANCA EXAMINADORA

_____

Prof. Dr. Marcel de Toledo Vieira – Orientador

Universidade Federal de Juiz de Fora

_____

Prof. Dr. Ronaldo Rocha Bastos

Universidade Federal de Juiz de Fora

_____

Prof. Dr. Ricardo Freguglia

Universidade Federal de Juiz de Fora

AGRADECIMENTOS

Agradeço à minha mãe Meire e meu pai Fernando, por serem a base de tudo que eu sou, por terem me transmitido valores essenciais durante a vida e por não terem medido esforços para que eu tivesse a melhor educação possível.

Ao meu irmão Vinícius, pela enorme amizade e companheirismo, e também por ser uma referência de pessoa para mim.

À minha cunhada Carol, por ser um exemplo de garra e determinação e pelos conselhos e conversas.

À minha tia Vera, por estar sempre presente e na torcida pelas minhas vitórias.

À minha tia Cormarie, pelo estímulo que deu aos meus estudos.

À Lúcia, por todo carinho e dedicação desde a minha infância.

À minha namorada Júlia, por alegrar meus dias e por me motivar a ser cada vez melhor.

Ao meu orientador Marcel, por ter me conduzido na realização deste trabalho com atenção e paciência.

Aos amigos Jack, Motoboy, Camila, Carol e Isabela e demais colegas que compartilharam comigo as dificuldades deste curso, pelos momentos de descontração, de ajuda e incentivo. Cada um deles teve sua importância nessa conquista.

Aos meus amigos de longa data, pela parceria no dia a dia. Nossa união é para sempre.

Aos meus professores, por terem alimentado meu interesse pela Estatística e por socializarem seus conhecimentos.

À Faculdade de Economia por ter permitido o uso do software *Stata* em suas instalações e ao Instituto Brasileiro de Geografia e Estatística por ter disponibilizado os dados da Pesquisa Nacional por Amostra de Domicílios.

“The greatest enemy of knowledge is not ignorance, it is the illusion of knowledge”

Stephen Hawking

RESUMO

Pesquisas de todos os tipos são realizadas no Brasil e no mundo atualmente. A quantidade de dados disponíveis é enorme. Os métodos de Amostragem são de grande relevância neste contexto e por isso foram objeto de estudo neste trabalho. Neste sentido, torna-se essencial tomar as devidas precauções durante todo o processo de amostragem, desde antes da coleta dos dados até a fase de análise dos mesmos. Nesta monografia adotamos a Pesquisa Nacional por Amostra de Domicílios (PNAD) realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE) como motivação para a aplicação dos métodos estudados. O principal objetivo deste estudo é avaliar os possíveis efeitos da não consideração das características do desenho amostral na análise de dados da PNAD. Apresentamos também um resumo dos principais tipos de planos amostrais e dos métodos de inferência para planos amostrais complexos para introduzir o assunto. O software estatístico *Stata* auxiliou na estimação da média e do seu erro padrão da variável *"renda domiciliar mensal per capita"* levando o planejamento amostral em consideração ou não. No caso da PNAD, algumas das consequências observadas ao ignorarmos o desenho foram: (i) a subestimação dos erros padrão e consequentemente intervalos de confiança enviesados; e (ii) estimativas de médias diferentes das encontradas quando se considera o planejamento amostral.

Palavras-chave: Amostragem complexa; EPA; Inferência; Renda domiciliar.

ABSTRACT

All kinds of researches are performed in Brazil and in the world currently. The amount of data available is huge. Sampling methods are very relevant in this context and, therefore, they were the subject of this study. In this sense, it is essential to take precautions throughout the sampling process, from before collecting data until the analysis phase. In this work we adopt the National Household Sample Survey (PNAD) conducted by the Brazilian Institute of Geography and Statistics (IBGE) as motivation for the application of the studied methods. The main goal of this work is to evaluate the potential effects of the lack of consideration of sample design features in the data analysis of PNAD. We also present a summary of the main types of sampling designs and inference methods for complex sampling in order to introduce the subject. The statistical software *Stata* has helped in the estimation of the mean and its standard error of the variable *"monthly household income per capita"* taking sample design into account or not. In the case of PNAD, some of the consequences observed when the design was ignored were: (i) underestimation of standard errors and, therefore, biased confidence intervals; and (ii) different estimated means compared to those found when sampling design was considered.

Keywords: Complex sampling; Design effect; Inference; Household income.

SUMÁRIO

1 - INTRODUÇÃO

Vivemos em um mundo que vem se modificando bastante a partir das últimas décadas. No que se refere à tecnologia e informação, o salto foi gigantesco. Junto com essas mudanças vieram novas necessidades. Tal evolução possibilitou ao homem entender cada vez mais como se comporta o ambiente ao seu redor e o que vemos atualmente é uma quantidade massiva de dados que precisam ser interpretados e transformados em informação útil, o que só acontece quando estes são coletados e analisados corretamente.

Neste contexto se encontram as pesquisas de grande porte, que são realizadas no Brasil e no mundo com diversas finalidades, gerando dados valiosos que podem ser trabalhados sob diferentes óticas. Essas pesquisas podem ser conduzidas através de um censo, que envolve a coleta de informações sobre todas as unidades da população, ou por amostragem, que é um conjunto de métodos que permitem a observação de informações de algumas unidades selecionadas aleatoriamente com o objetivo de inferir parâmetros para a população.

Talvez possamos pensar intuitivamente que a melhor maneira de se estudar uma população seja através de um censo. Porém, inúmeras situações nos conduzem à adoção de métodos de amostragem. Quando se quer que um resultado seja produzido de forma mais ágil, ou a seleção de uma unidade implica na perda da mesma, ou mesmo quando não se tem recursos financeiros suficientes para se realizar um censo, levantamentos por amostragem se mostram muito eficientes.

Para conduzir um estudo estatístico por meio de amostragem é necessário conhecer o conceito de planos amostrais probabilísticos. Eles são planos nos quais todas as unidades da população tem uma probabilidade não nula de pertencer à amostra, e essa probabilidade é conhecida ou calculável. Além disso, podemos enumerar todas as possíveis amostras dado o procedimento empregado para a seleção. Tais planos tem a vantagem de que as informações obtidas para a amostra podem ser generalizadas para a população, pois a seleção aleatória garante a representatividade da mesma. Outro ponto positivo é que os erros amostrais podem ser mensurados (Vieira, 2013).

São exemplos de planos amostrais probabilísticos: (i) amostragem aleatória simples (com ou sem reposição); (ii) amostragem estratificada; (iii) amostragem por conglomerados; (iv) amostragem com probabilidades desiguais (por exemplo, com

probabilidades proporcionais a uma medida de tamanho – PPT); e (v) amostragem em múltiplos estágios.

É de suma importância que escolhamos o melhor plano amostral para o caso estudado, pois isso terá uma grande influência nos resultados obtidos. Além de levar em consideração aspectos como o tempo e recursos disponíveis, devemos conhecer bem a população para decidir qual plano amostral é o mais adequado, de forma a trazer estimativas mais eficientes para os parâmetros de interesse. De maneira geral, buscamos um pequeno erro padrão e um baixo efeito do plano amostral, além de estimadores não viciados, conceitos estes que serão abordados ao longo desta monografia.

Podemos calcular o possível impacto do uso de um determinado plano amostral sobre a precisão dos estimadores das variáveis de interesse e também calcular tamanhos amostrais que garantem determinado nível de eficiência. Desta forma podemos comparar e planejar estudos futuros. Se a amostragem complexa já foi realizada, a recomendação é que sempre se considere as características do plano amostral para inferir sobre os parâmetros, ao invés de utilizar a hipótese ingênua de que a amostragem realizada foi aleatória simples.

Quando as características do plano amostral são desconsideradas, temos consequências prejudiciais. Ignorar os pesos na estimação da média provoca vícios substanciais que não podem ser descartados, o mesmo acontece ao ignorar os pesos na estimação da variância do estimador. Conglomeração e seleção com probabilidades desiguais, ou seja, pesos amostrais desiguais resultam em aumentos no efeito do plano amostral, enquanto que a estratificação tem como efeito uma redução nesta medida (Silva e Pessoa, 1998).

O Instituto Brasileiro de Geografia e Estatística (IBGE) é o órgão responsável pelas estatísticas oficiais do nosso País. Várias pesquisas são realizadas a fim de conhecer melhor o cenário em que vivemos e assim ser instrumento de auxílio na tomada de decisões dos governantes e gestores, formulando, validando e avaliando políticas públicas voltadas para o desenvolvimento socioeconômico e para a melhoria das condições de vida da população de uma forma geral.

Um dos levantamentos de dados realizados pelo IBGE é a Pesquisa Nacional por Amostra de Domicílios (PNAD) que investiga anualmente, de forma permanente, características gerais da população, de educação, trabalho e rendimento, além de investigar com periodicidade variável outras características de acordo com a

necessidade do momento. A PNAD é uma das pesquisas de grande porte realizadas através da adoção de técnicas de amostragem no País (IBGE, 2015).

Neste estudo, escolhemos a variável renda como sendo a de principal interesse, aqui definida como "*renda domiciliar mensal per capita*" e utilizamos variáveis auxiliares para efeito de comparação entre grupos. A variável de interesse tem caráter contínuo e as auxiliares - região, sexo, cor, escolaridade e faixa etária - são todas categóricas e formam estratos naturais ou domínios de estimação de interesse.

A PNAD é realizada por meio da seleção de uma amostra probabilística e por isso servirá como base de estudo deste trabalho que tem como maior objetivo descrever, analisar e comparar métodos de inferência estatística que consideram as características do desenho amostral (dentre elas, os pesos amostrais) e aqueles que desconsideram tais características. Nesta monografia, um maior enfoque será dado às questões metodológicas em relação às questões de ordem substantiva.

No Capítulo 2 é apresentada uma revisão sobre planos amostrais. O Capítulo 3 aborda Inferência para dados amostrais complexos. Já o Capítulo 4 trata da aplicação dos procedimentos aos dados da PNAD 2013. No Capítulo 5 temos as considerações finais.

2 - PLANOS AMOSTRAIS PROBABILÍSTICOS

Planos amostrais probabilísticos garantem que todas as unidades da população possuam uma probabilidade não nula de serem selecionados para a amostra. Além disso, permitem a definição de um conjunto com todas as amostras possíveis e suas respectivas probabilidades de seleção, de acordo com o processo probabilístico determinado.

2.1 - AMOSTRAGEM ALEATÓRIA SIMPLES

A amostragem aleatória simples é o método mais elementar e ao mesmo tempo mais importante que pode ser adotado para a seleção de uma amostra, pois serve como um plano próprio e também é usado em procedimentos de múltiplos estágios, fornecendo a base para esquemas amostrais complexos, ou seja, ele mesmo pode ser usado como plano amostral sem a necessidade de outros planos, ou então pode ser combinado com outro tipo de planejamento amostral.

Podemos selecionar a amostra de duas maneiras. Se, ao sortearmos uma unidade da população, excluirmos tal unidade do próximo sorteio, chamamos de amostragem aleatória simples sem reposição (AASs). Já se uma unidade sorteada puder ser repetida na amostra, chamamos de amostragem aleatória simples com reposição (AASc).

O procedimento, para o caso “com reposição”, consiste em sortear com igual probabilidade 1/*N* uma unidade da população. Repetimos o processo até que sejam obtidos os *n* elementos que irão compor a amostra, sendo que o tamanho da amostra é previamente definido. Dessa forma, estará garantido que todas as possíveis amostras de tamanho *n* tenham a mesma probabilidade de serem escolhidas (Cochran, 1965, p.38).

A amostragem aleatória simples sem reposição (AASs) é mais intuitiva e eficiente, o que resulta em um menor efeito do plano amostral. Desta forma, podemos afirmar que a AASs é sempre “melhor”, exceto quando o tamanho da amostra é igual a 1 e não existe diferença. Porém a AASc, por resultar em independência entre as observações, tem vantagens estatísticas e matemáticas pois facilita a determinação das propriedades dos estimadores e das quantidades populacionais de interesse. Portanto, a AASc é bastante adotada como pressuposto

básico para os métodos estatísticos apresentados na maioria dos livros de Estatística. Quando a população é muito grande, a diferença entre AASs e AASc se torna desprezível (Vieira, 2013).

Consideramos $\bar{y}$ como estimador da média populacional,

$$\bar{y} = \frac{1}{n}\sum_{i\epsilon s} Y_i \text{ , e}$$

$$. Var[\bar{y}\,] = \frac{\sigma^2}{n}.$$

Adotamos *T(s)* como estimador do total populacional,

$$T(s) = N\bar{y}, \text{ e}$$

$$Var[T] = N^2\frac{\sigma^2}{n}.$$

Julgamos s^2 como estimador da variância populacional,

$$s^2 = \frac{1}{n-1}\sum_{i\epsilon s}(Y_i - \bar{y}\,)^2.$$

Todos os estimadores acima são válidos para AASc e são não viesados (Bolfarine e Bussab, 2005, p.76).

2.2 - AMOSTRAGEM ESTRATIFICADA

A amostragem estratificada consiste basicamente na divisão da população em subpopulações bem definidas (estratos), formando grupos naturais ou substantivos de interesse, o que garante um maior espalhamento da amostra em comparação com a amostragem aleatória simples. Para que estes estratos sejam escolhidos de maneira adequada é necessário que se tenha conhecimento das variáveis que auxiliam no processo de estratificação.

De cada estrato, retiramos unidades, usualmente de forma independente, permitindo estimação tanto para a população como um todo quanto para os subgrupos, o que é muito útil para o pesquisador, pois dá liberdade de pré-estabelecer estratos que fornecerão resultados mais precisos de acordo com seus objetivos iniciais.

Além disso, podemos ou não utilizar o mesmo plano amostral para sortear uma amostra dentro de cada estrato, apesar de não ser comum a utilização de planos diferentes. Já que cada estrato pode ser tratado como uma população distinta, temos que para cada amostra utilizamos os estimadores convenientes para os parâmetros do estrato. Também é possível montar um estimador para a população considerando os estimadores de cada estrato e determinando suas propriedades.

A amostragem estratificada quase sempre é estatisticamente mais eficiente que a amostragem aleatória simples, sendo que quanto mais homogêneos são os subgrupos, maior a eficiência do plano amostral. O fato dos subgrupos serem mais homogêneos internamente do que a população como um todo, proporciona uma redução do erro amostral no geral. Este aumento da precisão das estimativas permite a diminuição da amostra para um nível de precisão fixo.

A eficiência do plano amostral pode ser influenciada por vários fatores, sendo que é considerado mais eficiente o plano estratificado no qual a variância dentro dos estratos é menor. A escolha das variáveis de estratificação, o número de estratos, a determinação dos limites dos estratos, a alocação da amostra nos estratos e o método de seleção em cada estrato são alguns pontos importantes na busca do melhor desempenho (Vieira, 2013).

Consideramos $\bar{y}_{es}$ como estimador da média populacional,

$$\bar{y}_{es} = \sum_{h=1}^{H} W_h \, \bar{y}_h,$$

e T_{es} como estimador do total populacional,

$$T_{es} = \sum_{h=1}^{H} N_h \bar{y}_h,$$

onde $W_h = \frac{N_h}{N}$ e *H* é o número de estratos.

A alocação da amostra nos estratos pode ser feita de formas distintas. A amostragem estratificada proporcional considera o tamanho dos estratos e distribuem as *n* unidades da amostra proporcionalmente a este tamanho, sendo, (Bolfarine e Bussab, 2005, p.102)

$$n_h = n \frac{N_h}{N}.$$

A amostragem estratificada uniforme utiliza o mesmo tamanho de amostra para cada estrato, ou seja, (Bolfarine e Bussab, 2005, p.103)

$$n_h = \frac{n}{H}.$$

A alocação ótima de Neyman mostra que o número ideal de unidades a serem observadas no estrato *h* é diretamente proporcional a $N_h \sigma_h$, sendo, (Bolfarine e Bussab, 2005, p.106)

$$n_h = n \frac{N_h \sigma_h}{\sum_{h=1}^{H} N_h \sigma_h}.$$

Este último método é muito utilizado, pois faz uso do conhecimento de que quanto maior a variância do estrato, maior deve ser também o tamanho da amostra a ele designado.

2.3 – AMOSTRAGEM POR CONGLOMERADOS

A amostragem por conglomerados é utilizada na maioria das vezes quando é inviável fazer uma lista digna de confiança que identifique cada elemento da população ou quando é muito trabalhoso e custoso o deslocamento para se observar cada elemento, devido às distâncias geográficas entre as mesmas, por exemplo. (Cochran, 1965, p.318).

A importância de falarmos primeiramente da motivação ao fazermos amostragem por conglomerados é que ela é menos eficiente que a amostragem aleatória simples, logo seria lógico pensar em AAS antes de tudo. Apesar disso, a amostragem por conglomerados gera estimativas com precisão aceitável se for bem conduzida, o que inclui a busca por maior heterogeneidade dentro dos conglomerados e maior homogeneidade entre os conglomerados, sendo assim é muito útil, especialmente quando a população for extensa (Cochran, 1965, p.318).

O procedimento consiste em dividir a população em conglomerados, contendo mais de um elemento populacional, que podem ser regiões, estados, setores censitários, hospitais, escolas entre outros e selecionar uma amostra de conglomerados de acordo com um plano amostral qualquer. Feito isso, todos os elementos são selecionados caracterizando uma amostragem por conglomerados em um estágio. Alternativamente, um segundo (ou mais estágios) de seleção poderia ser conduzido até que no último estágio todos os elementos fossem selecionados, caracterizando uma amostragem por conglomerado em múltiplos estágios. Os sorteios de cada estágio podem seguir ou não os mesmos planos amostrais (Vieira, 2013).

Ao contrário da amostragem estratificada, na amostragem por conglomerados, quanto maior a heterogeneidade dentro do conglomerado mais eficiente é o procedimento amostral, e este é um dos motivos deste método ser pior em termos de eficiência, pois as unidades dentro de um mesmo conglomerado tendem a ter alta correlação quanto às variáveis pesquisadas. Uma das soluções para o caso em que os conglomerados são muito homogêneos é fazer a seleção da amostra em mais estágios (Vieira, 2013).

Seja y_{ij} o valor da variável de pesquisa para unidade j do conglomerado *i*. O total de unidades no conglomerado é definido como M_i. O total de conglomerados é N. O total de unidades é $M_0 = \sum_{i=1}^{N} M_i$.

O total no conglomerado *i* é dado por:

$$y_i = \sum_{j=1}^{M_i} y_{ij}.$$

A média no conglomerado *i* é dada por:

$$\bar{Y}_i = \sum_{j=1}^{M_i} \frac{y_{ij}}{M_i}.$$

O total populacional é dado por:

$$Y = \sum_{i=1}^{N} y_i.$$

A média por conglomerado é dada por:

$$\bar{Y}_C = \sum_{i=1}^{N} \frac{y_i}{N}.$$

A média por unidade é dada por:

$$\bar{Y} = \sum_{i=1}^{N} \frac{y_i}{M_0}.$$

As expressões que apresentamos acima são válidas para amostragem por conglomerados em um estágio.

Quando os conglomerados são muito homogêneos, o uso da amostragem por conglomerados em um estágio se torna menos recomendável, pois como as unidades são muito parecidas elas tendem a fornecer o mesmo tipo de informação, aumentando a variação amostral. Para contornar essa situação, adotamos a amostragem por conglomerados em múltiplos estágios, sendo mais comum a utilização de dois ou três estágios (Bolfarine e Bussab, 2005, p.197).

No caso da amostragem por conglomerados em dois estágios, selecionamos no primeiro estágio conglomerados, que são unidades primárias de amostragem (UPAs), seguindo algum plano amostral. Feito isso, sorteamos elementos, que são unidades secundárias de amostragem (USAs), utilizando ou não o mesmo plano amostral.

A amostragem por conglomerados em três estágios segue a mesma linha de raciocínio da amostragem por conglomerados em dois estágios, porém após a seleção dos elementos, que são unidades secundárias de amostragem, é realizado mais um sorteio para a escolha de unidades elementares de cada uma das USAs selecionadas, sendo que o plano amostral adotado em cada uma das etapas é definido pelo pesquisador (Vieira, 2013). Expressões para estimadores que levam em consideração amostragem por conglomerados em 2 estágios podem ser encontradas em Bolfarine e Bussab (2005).

2.4 – AMOSTRAGEM COM PROBABILIDADES PROPORCIONAIS A UMA MEDIDA DE TAMANHO

Vimos até aqui apenas esquemas probabilísticos que consideram que todas as amostras tem a mesma probabilidade de seleção. Entretanto, as unidades de amostragem podem apresentar grande variação de tamanho e ignorar tal fato pode fazer com que o plano amostral seja menos eficiente.

Neste contexto aparece a amostragem com probabilidades desiguais, para os casos em que a variação de tamanho das unidades de amostragem é grande. Abordaremos aqui a amostragem proporcional a uma medida de tamanho (PPT), que é um dos métodos que utiliza probabilidades desiguais de seleção de uma amostra, e é muito utilizado em pesquisas domiciliares e educacionais.

É necessário que tenhamos uma variável auxiliar associada a uma medida de tamanho de cada elemento que nos ajude na construção da amostra. Se esta variável for correlacionada à variável de interesse, a amostragem PPT é mais eficiente que a amostragem aleatória simples (Vieira, 2013).

Os planos amostrais estudados anteriormente são casos particulares em que se considera igual probabilidade de seleção para todas as unidades. Para que o estimador de total continue sendo não viesado na amostragem PPT, os pesos das

unidades devem ser o inverso das respectivas probabilidades de inclusão na amostra.

Existem algumas formas diferentes de se fazer a amostragem com probabilidades proporcionais ao tamanho como, por exemplo, o método dos totais cumulativos (com reposição) e a amostragem PPT de Poisson. Para maiores informações, consultar Kish (1995) e Bolfarine e Bussab (2005). O método dos totais cumulativos (com reposição) consiste basicamente em criar uma lista com os totais acumulados da variável auxiliar *X* associada ao tamanho e dessa forma estabelecer intervalos de seleção com base no tamanho de cada unidade. Por exemplo, se a unidade 1 tem 100 elementos e a unidade 2 tem 50 elementos, o primeiro intervalo de seleção é de 1 a 100 e o segundo de 101 a 150 e assim sucessivamente até que todas as unidades tenham seu respectivo intervalo. Feito isso, basta sortear um número aleatório *r* com distribuição uniforme entre 1 e a soma da variável *X* e selecionar a unidade no qual o número *r* faz parte do intervalo de seleção. Devemos repetir este procedimento até que se tenha a quantidade de elementos pretendidos na amostra.

Na amostragem PPT de Poisson, para cada unidade é designada uma probabilidade de inclusão π_i de acordo com a variável auxiliar *X* que está associada ao tamanho da unidade. O cálculo de π_i é dado por:

$$\pi_i = nx_i/X.$$

A partir daí, sorteamos de forma independente para cada unidade um número aleatório entre 0 e 1 segundo uma distribuição uniforme e se este número for menor ou igual a π_i, a unidade *i* é incluída na amostra. Sendo assim, o tamanho da amostra não é fixo, mas o seu valor esperado pode ser estimado.

Os pesos amostrais normalmente refletem o número de unidades populacionais que cada unidade amostral representa. Inicialmente ele pode ser igual ou proporcional ao inverso da probabilidade de seleção da unidade, porém ao final da análise ele pode incorporar outras informações. É bastante útil quando queremos tratar o problema de não resposta, e o que fazemos é ajustar os pesos para as respostas dos respondentes de tal forma que eles representem os não respondentes. Também podemos usar informações para fazer ajustes de modo que

a distribuição amostral ponderada para certas variáveis se assemelhe com distribuições populacionais para as mesmas variáveis, através de métodos de calibração.

3- INFERÊNCIA PARA DADOS AMOSTRAIS

A Inferência Estatística é baseada nos procedimentos de estimação e testes de hipóteses. A estimação para um parâmetro pode ser feita pontualmente ou por intervalos de acordo com alguns métodos, como por exemplo, o Método dos Momentos, o Método da Máxima Verossimilhança e o Método dos Mínimos Quadrados. Os testes de hipóteses são usados em modelos estatísticos (Vieira, 2013).

3.1 - INFERÊNCIA ESTATÍSTICA CLÁSSICA

> "Seja *Y* uma variável aleatória de interesse, e sejam $y_1, \dots, y_n$, *n* observações desta variável. Em Inferência Estatística, o modelo usual considera $y_1, \dots, y_n$, como valores (realizações) de variáveis aleatórias $Y_1, \dots, Y_n$. Aqui $Y_1, \dots, Y_n$ são variáveis aleatórias independentes e identicamente distribuídas (IID), com a mesma distribuição de *Y*, digamos com função de densidade ou de frequência $f(y;\theta)$, onde $\theta \in \Theta$ é o parâmetro indexador da distribuição *f*, e Θ é o espaço paramétrico." (Silva e Pessoa, 1998, p.17)

> "O Método da Máxima Verossimilhança consiste em estimar θ com estatística cujo valor maximize a verossimilhança da amostra em relação a θ. Isto é, o estimador de máxima verossimilhança para θ é a função dos dados amostrais $Y_1, \dots, Y_n$ que torna máxima a função de verossimilhança $l(y;\theta)$ em θ." (Vieira, 2013).

Seja a equação de verossimilhança da amostra

$$l(\theta; y) = \prod_{i=1}^{n} f(y_i; \theta), \text{ e}$$

a Log-verossimilhança

$$L(\theta; y) = \sum_{i=1}^{n} log\ [f(y_i, \theta)].$$

As equações de verossimilhança são dadas por:

$$\sum_{i=1}^{n} \frac{\partial log[f(y_i,\theta)]}{\partial \theta} = \sum_{i=1}^{n} u_i(\theta) = 0.$$

.

A solução $\hat{\theta}$ é o estimador de máxima verossimilhança de θ .

Podemos estimar variâncias por máxima verossimilhança para grandes amostras de acordo com as seguintes expressões,

$$V(\hat{\theta}) \approx [J(\theta)]^{-1},$$

$$J(\theta) = -\sum_{i=1}^{n} \frac{\partial^2 log[f(y_i,\theta)]}{\partial \theta^2} = -\sum_{i=1}^{n} \frac{\partial u_i(\theta)}{\partial \theta},$$

$$\hat{V}(\hat{\theta}) \approx [J(\hat{\theta})]^{-1}, \text{ e}$$

$$J(\hat{\theta}) = J(\theta)|_{\theta=\hat{\theta}}.$$

A Inferência Estatística Clássica é analítica e baseada no modelo *f(y,θ)* em que os dados obtidos são utilizados para a descrição de populações infinitas e foi proposta para ser utilizada em situações onde o pesquisador tem certo controle e faz sentido replicar o experimento. Os planos amostrais neste caso são ignorados e os dados recebem pesos iguais.

3.2 - DADOS AMOSTRAIS COMPLEXOS

Seja uma população finita *U={1,...,N}* e uma amostra $s=\{i_1, \dots, i_n\}$ retirada dessa população de acordo com algum plano amostral caracterizado por *p(s)*, sendo que a probabilidade de cada possível amostra é calculável. Os valores $y_1, \dots, y_n$ de uma variável de interesse *Y* são considerados fixos, porém desconhecidos. De

forma alternativa, podemos *reindexar* a população de maneira que a amostra seja formada pelos índices *s={1,...,n}.*

Desde que se tenha a amostra $y_1, \dots, y_n$, o objetivo é inferir sobre funções $g(y_1, \dots, y_n)$ dos parâmetros populacionais, tais como totais, médias, proporções, etc.

Esta abordagem é descritiva e utilizada no caso de populações finitas. Foi proposta inicialmente para ser utilizada em estudos socioeconômicos. Além disso, é essencialmente não paramétrica por não assumir nenhuma distribuição paramétrica para as observações da amostra. As características dos planos amostrais neste caso são consideradas na análise e os dados recebem pesos diferentes. Uma das desvantagens é que não se pode generalizar, sendo que inferências só são realizadas para a particular população de onde a amostra foi selecionada (Silva e Pessoa, 1998, p.21).

3.3 - MODELAGEM DE SUPERPOPULAÇÃO

A modelagem de superpopulação surge no contexto da amostragem com o objetivo de combinar a tradição modelista com a tradição amostrista, considerando um modelo parametrizado e aproveitando a estrutura do planejamento amostral em um estudo analítico.

Não se pode ignorar o planejamento amostral e modelar os dados como independentes e identicamente distribuídos, pois isso em geral causa diferenças nas estimativas pontuais, estimativas de variância e nas distribuições de estatísticas de teste. A solução é o Método da Máxima-Pseudoverossimilhança (MPV) que incorpora o planejamento amostral e pesos, e tem como objetivo inferir sobre θ (Vieira, 2013).

> "Sejam os valores $y_1, \dots, y_N$, da variável de interesse *Y* na população finita, considerados observações ou realizações das variáveis aleatórias $Y_1, \dots, Y_N$, supostamente independente e identicamente distribuídas (IID) com distribuição *f(y;* θ), onde $\theta \in \Theta$. Utilizando um plano amostral definido por *p(s)*, obtemos os valores na amostra $y_1, \dots, y_n$. A partir de $y_1, \dots, y_n$ (não considerados IID, em geral) queremos fazer inferência sobre o

parâmetro θ, considerando características do planejamento amostral.” (Silva e Pessoa, 1998).

Dada a equação de verossimilhança do censo:

$$\sum_{i=1}^{N} u_i(\theta) = T(\theta) = 0, T(\theta) \text{ é um total populacional.}$$

A estimativa de totais com amostra é:

$$\hat{T}(\theta) = \sum_{i \in s} w_i . u_i(\theta), \text{ e}$$

a equação de verossimilhança da amostra é:

$$\hat{T}(\theta) = \sum_{i \in s} w_i . u_i(\theta) = 0$$

A solução $\hat{\theta}_{\text{MPV}}$ é o estimador de MPV usando amostra ponderada (Vieira, 2013).

Este procedimento é aplicável a muitos modelos paramétricos e planos amostrais, sendo que o estimador pode variar de acordo com os pesos dados às observações.

3.4 - LINEARIZAÇÃO DE *TAYLOR*

Em algumas situações, temos o interesse de estimar parâmetros não lineares, como por exemplo, razões, correlações, coeficientes de regressão, quantis de distribuições, etc. A linearização de *Taylor* foi um dos métodos desenvolvidos para tornar possível a estimação de variâncias para tais parâmetros (Wolter, 2007, p.226).

O método consiste em considerar uma função de *K* totais populacionais que represente o parâmetro populacional, isto é, $\theta=f(Y_1, \ldots, Y_K)$ e através de uma expansão em séries de *Taylor* (considerando apenas o termo de primeira ordem) obter um estimador $\hat{\theta}_L$ linearizado que seja uma boa aproximação para $\hat{\theta}$. Para

grandes amostras, $\hat{\theta}$ e $\hat{\theta}_L$ tem comportamento semelhante e, portanto, podemos tomar o estimador linearizado como uma boa aproximação para o estimador não linear (Silva e Pessoa, 2007, p.39).

A desvantagem deste procedimento é que nem sempre é fácil escrever uma estatística de interesse como função linear de totais ou médias populacionais, pois o processo envolve muitas derivações e cálculos específicos. Entretanto, com o auxílio de pacotes matemáticos computacionais isto pode ser feito sem maiores problemas. (Vieira, 2013).

3.5 - MÉTODO *JACKKNIFE*

O método *jackknife* é um método alternativo para estimar variâncias de estimadores. A ideia básica é repartir a posteriori uma amostra de *n* elementos em *G* grupos de *n/G* elementos mutuamente exclusivos e calcular os pseudo-valores $\hat{\theta}_{(g)}$, dados por:

$$\hat{\theta}_{(g)} = G\hat{\theta} - (G-1)\hat{\theta}_g,$$

onde $\hat{\theta}_g$ é um estimador não viesado para θ obtido da amostra após a exclusão de todas as unidades do grupo *g* e usando os mesmos procedimentos que seriam usados para calcular $\hat{\theta}$ considerando a amostra completa.

Feito isso, estimamos a variância usando um dos estimadores abaixo, sendo o segundo mais conservador. Temos que: (Vieira, 2013)

$$\hat{V}_{J1}(\hat{\theta}) = \frac{1}{G(G-1)} \sum_{g=1}^{G} (\hat{\theta}_{(g)} - \hat{\theta}_{JK})^2, \text{ ou}$$

$$\hat{V}_{J2}(\hat{\theta}) = \frac{1}{G(G-1)} \sum_{g=1}^{G} (\hat{\theta}_{(g)} - \hat{\theta})^2, \text{ onde}$$

$$\hat{\theta}_{JK} = \frac{1}{G} \sum_{g=1}^{G} \hat{\theta}_g.$$

É importante saber que a descrição acima não é válida para planos amostrais estratificados, pois estes requerem cálculos mais complexos. Outra observação é

que no caso de planos amostrais em múltiplos estágios, se uma unidade primária de amostragem (UPA) é excluída na aplicação do *jackknife*, então todas as unidades subordinadas a ela também deverão ser eliminadas da amostra.

Esta técnica para redução de vício de estimadores tem a vantagem de ser flexível e genérica, além dos estimadores não precisarem ser expressos em função de totais populacionais. Entretanto é preciso ficar atento, pois em geral ela é menos eficiente, principalmente se *G* for pequeno, pois o estimador pode ser instável (Wolter, 2007, p.151).

Os estimadores de variância do método *jackknife* fornecem os mesmos resultados que os estimadores lineares usuais de variância, além disso, suas propriedades são razoáveis para alguns casos de estimadores não lineares. Porém, para o caso de quantis de distribuições e estatísticas de ordem esta técnica não deve ser utilizada (Silva e Pessoa, 1998, p.45).

3.6 - ERRO PADRÃO

O erro padrão é uma medida usada na amostragem probabilística para indicar a imprecisão associada a uma estimativa (Cruz, 1978, p.740). O erro padrão da média, por exemplo, é uma estimativa do desvio padrão da distribuição das médias de todas as amostras possíveis com o mesmo tamanho provenientes da mesma população (Lunet *et al*, 2006).

Dada a distribuição amostral de um estimador, podemos calcular sua variância. Caso a distribuição não seja conhecida, mas uma aproximação da distribuição possa ser estimada, a variância adotada será a variância dessa aproximação. Chamamos de erro padrão a raiz quadrada dessa variância. Se, por exemplo, $\hat{\theta}$ for estimador do parâmetro θ, o erro padrão de $\hat{\theta}$ é: (Morettin e Bussab, 2013, p.321).

$$EP(\hat{\theta}) = \sqrt{Var(\hat{\theta})}.$$

Como a variância de $\hat{\theta}$ é desconhecida, calculamos o erro padrão estimado para $\hat{\theta}$, sendo:

$$\widehat{EP}(\hat{\theta}) = \sqrt{\widehat{Var}(\hat{\theta})}.$$

No caso específico da média, a estimativa do erro padrão é:

$$\widehat{EP}(\bar{X}) = \frac{S}{\sqrt{n}},$$

sendo S o desvio padrão amostral de X.

O erro padrão diminui com o aumento do tamanho da amostra, ou seja, quanto maior o tamanho da amostra mais precisa será a estimativa do erro padrão.

3.7 - EFEITO DO PLANO AMOSTRAL

Em 1965, Kish propôs uma medida do efeito do plano amostral sobre a variância de um estimador, um método para comparar ganhos ou perdas de precisão de planos amostrais alternativos antes da seleção da amostra. Essa medida ficou conhecida como efeito do plano amostral (EPA) ou design *effect (deff)*. A expressão do EPA de Kish é representada por:

$$EPA_{Kish}(\hat{\theta}) = \frac{V_{verd}(\hat{\theta})}{V_{AAS}(\hat{\theta})}.$$

Essa equação equivale a razão entre a variância verdadeira de um estimador, isto é, considerando o plano amostral complexo e a variância do estimador considerando amostragem aleatória simples. Se o EPA calculado for maior que 1, o plano amostral do numerador é menos eficiente, caso contrário, ele é mais eficiente. Esta medida fornece informações para o apoio ao planejamento de novas pesquisas amostrais, porém, ela perde sua importância quando a amostra já foi selecionada.
Daí surgiu a necessidade da criação de uma medida com a capacidade de avaliar a tendência de um estimador consistente, calculado sob a hipótese de ser independente e identicamente distribuído (IID), subestimar ou superestimar a variância verdadeira do estimador pontual. Skinner, Holt e Smith (1989) propuseram

o efeito do plano amostral ampliado ou *misspecification effect (meff)*, capaz de medir os efeitos da especificação incorreta, tanto do plano amostral, quanto do modelo ajustado. O EPA ampliado é definido por:

$$EPA(\hat{\theta}, V_0) = \frac{V_{verd}(\hat{\theta})}{E_{verd}(\widehat{V_0})}, \text{ onde}$$

$\hat{V}_0 = \hat{V}_{IID}(\hat{\theta})$ é um estimador consistente da variância do estimador, considerando a hipótese de que as observações são IID, $V_{verd}(\hat{\theta})$ é a variância verdadeira de um estimador, considerando o plano amostral e $E_{verd}(V_0)$ é a esperança de um estimador verdadeiro, considerando o plano amostral (Silva e Pessoa, 1998, p.47).

Em aplicações para dados reais, como é o caso desta monografia, adotamos o estimador do EPA que é definido como:

$$\widehat{EPA}(\hat{\theta}, V_0) = \frac{\hat{V}_{verd}(\hat{\theta})}{\widehat{V_0}}.$$

3.8 - ESTIMADORES NÃO VICIADOS

Seja uma amostra $X_1, X_2, \dots, X_n$ de uma variável aleatória que descreve uma característica de interesse de uma população e θ um parâmetro que queremos estimar.

Um estimador $\hat{\theta}$ do parâmetro θ é qualquer função das observações da amostra, ou seja,$\hat{\theta} = g(X_1, \dots, X_n)$. Nosso objetivo é encontrar uma função de $\hat{\theta}$ que se aproxime de θ. Um dos critérios adotados para medir essa proximidade é o vício do estimador.

O estimador $\hat{\theta}$ é considerado não viesado ou não viciado para θ se $E(\hat{\theta}) = \theta$.

Caso contrário, o estimador $\hat{\theta}$ é considerado viciado, e o viés de $\hat{\theta}$ pode ser calculado, sendo: (Morettin e Bussab, 2013, p.302)

$$V(\hat{\theta}) = E(\hat{\theta}) - \theta.$$

4 - APLICAÇÃO AOS DADOS DA PNAD

4.1 - PNAD

A Pesquisa Nacional por Amostra de Domicílios (PNAD), realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE), é uma pesquisa de larga escala implantada progressivamente no Brasil a partir de 1967, e ocorre anualmente com o objetivo de coletar dados econômicos, sociais e demográficos da população brasileira. A necessidade da realização deste estudo se tornou evidente nessa época devido à demanda por informações detalhadas da situação do País não estar sendo atendida pelo Censo Demográfico, que ocorre apenas de 10 em 10 anos, pois estas eram insuficientes ou já estavam defasadas.

A PNAD foi a alternativa encontrada pelo IBGE, visto que, por ser realizada através de métodos de amostragem permitiu o estudo e planejamento do desenvolvimento socioeconômico do País com menor uso de recursos financeiros, economia de tempo e com maior controle das fases operacionais e ainda consegue estimar parâmetros com pequena margem de erro.

A estrutura da PNAD abrange três tipos de pesquisa: básica, suplementar e especial. As pesquisas básicas são realizadas continuamente e se caracterizam por buscar conhecimentos sobre os assuntos mais relevantes para mensurar o nível socioeconômico da população, tais como habitação e mão-de-obra, além de características demográficas e educacionais. As pesquisas suplementares estudam de maneira aprofundada os temas da pesquisa básica e também tópicos relacionados à mesma. Já as pesquisas especiais são independentes da pesquisa básica e tratam de assuntos mais complexos.

A PNAD é realizada através de questionários preenchidos de acordo com uma entrevista pessoal, atualmente assistida por computador, e abrange a população residente em domicílios particulares permanentes[1] e em unidades de habitação em domicílios coletivos, com o foco nos indivíduos e nas famílias. Uma característica muito útil da PNAD é que os indicadores produzidos podem ser facilmente comparados com os obtidos em outras pesquisas domiciliares, incluindo o

[1] O domicílio particular localizado em unidade que se destina a servir de moradia (casa, apartamento e cômodo).

Censo Demográfico, por seguir conceitos e definições operacionais muito semelhantes à delas, o que é uma enorme vantagem (IBGE, 2015).

Vários pesquisadores já utilizaram as bases de dados fornecidas pela PNAD considerando o planejamento amostral em estudos de renda como Moura (2008) e posteriormente Barbosa (2013), entre outros.

4.2 - PLANO AMOSTRAL DA PNAD

A PNAD é uma pesquisa por amostragem complexa, pois envolve estratificação, conglomeração e amostragem com probabilidades proporcionais a uma medida de tamanho, em algumas de suas etapas de seleção. O plano amostral adotado pela PNAD é estratificado e conglomerado com um, dois ou três estágios de seleção dependendo do estrato.

O processo de construção do desenho amostral da PNAD consiste inicialmente em dividir o Brasil em 36 estratos naturais, sendo que 27 deles correspondem às unidades da federação e os outros 9 são municípios das regiões metropolitanas com sede na capital (PA, CE, PE, BA, MG, RJ, SP, PR, RS).

Para os 27 estratos que representam as unidades da federação, é realizada uma amostragem por conglomerados em 3 estágios, sendo as unidades primárias de amostragem (UPAs) os municípios, as unidades secundárias de amostragem (USAs) os setores censitários, estes dois primeiros selecionados com probabilidade proporcional ao seu tamanho, e as unidades terciárias de amostragem (UTAs) os domicílios, escolhidos por amostragem sistemática, e todos os moradores de um domicílio da amostra também a compõem.

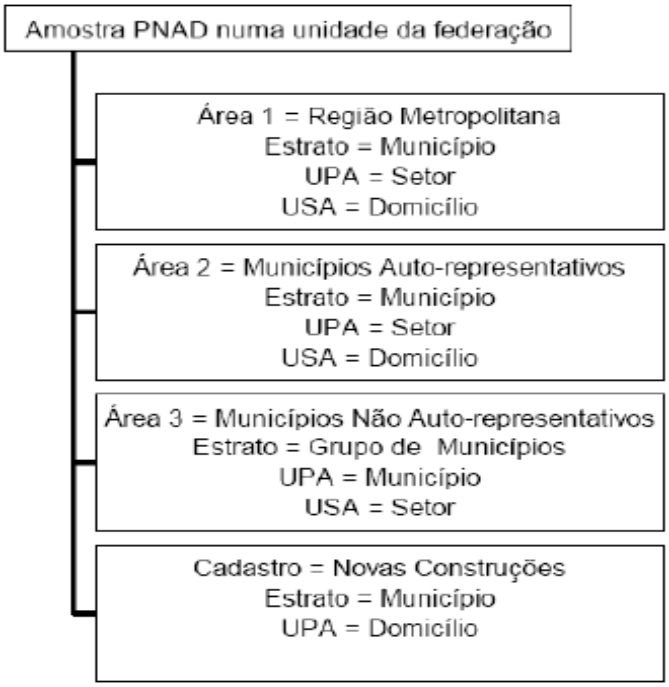
Alguns municípios são considerados auto representativos por possuírem 80% ou mais da população total da unidade da federação em que se localiza e então, estes incorporarão a amostra com certeza. Neste caso, os setores censitários que farão parte da amostra são selecionados com probabilidade proporcional à quantidade de domicílios nele contido. Feito isso, os domicílios são selecionados.

Quando os municípios não são auto representativos, estes são conglomerados por tamanho e proximidade geográfica de forma que os estratos tenham um tamanho de população semelhante. Depois de sorteados os grupos de municípios, definimos municípios que irão compor a amostra e então o mesmo procedimento é adotado para a escolha dos setores censitários e dos domicílios.

Para os 9 estratos representados por municípios das regiões metropolitanas com sede na capital, é realizada uma estratificação por município, sendo conglomerados em 2 estágios, de tal forma que as UPAs sejam os setores censitários e as USAs os domicílios, obedecendo às mesmas regras de seleção já descritas acima.

A última etapa do processo é para a escolha de unidades domiciliares selecionadas a partir do Cadastro de Projetos de Novas Construções, que inclui conjuntos habitacionais com mais de 30 domicílios que foram construídos depois do último censo. Neste caso, estes novos conjuntos habitacionais são estratificados por municípios, e para estes estratos o plano amostral é conglomerado em um estágio, sendo que as UPAs são os próprios domicílios. A seleção é feita por amostragem sistemática. A Figura 1 abaixo representa o plano amostral da PNAD (Silva *et al*, 2002).

Figura 1: Construção do plano amostral da PNAD.



Fonte: Silva *et al*. (2002).

O IBGE apresenta vários estudos realizados através da utilização de dados provenientes da PNAD, todos eles de maneira correta, considerando o planejamento amostral. Entretanto, apesar de fornecer as informações necessárias para que um

pesquisador não vinculado ao instituto também faça uso dos mesmos, não mostra de maneira clara a importância de considerarmos o planejamento amostral e as consequências de o ignorarmos em sua página na internet.

4.3 - VARIÁVEIS

Nesta monografia, como já dito anteriormente, foi dado um maior enfoque às questões metodológicas em relação às questões de ordem substantiva. Sendo assim, a escolha das variáveis foi feita de forma a possibilitar a ilustração da importância dos conceitos apresentados neste estudo, e assim facilitar a compreensão dos leitores.

A variável escolhida como sendo a de principal interesse foi renda, que tem caráter contínuo e é aqui definida como “renda domiciliar mensal *per capita*”. Este é um importante indicador social e por isso desperta a curiosidade pelo tema. Outras variáveis auxiliares foram também selecionadas com o objetivo de permitir comparações entre alguns grupos. Estas variáveis são todas categóricas e são as seguintes: região, sexo, cor, escolaridade e faixa etária. Tais variáveis formam estratos naturais ou domínios de estimação de interesse e por isto foram escolhidas. Nas análises conduzidas, foram consideradas as informações prestadas apenas pela pessoa de referência de cada domicílio. Portanto, todos os resultados produzidos dizem respeito a elas. Estes dados são referentes à PNAD 2013, ano mais recente cujos dados estão disponíveis.

O Quadro 1 a seguir apresenta uma breve descrição das variáveis consideradas.

Quadro 1: Descrição das variáveis auxiliares

Variáveis	Categorias	Descrição das categorias
Região	Norte	Reside na Região Norte
	Nordeste	Reside na Região Nordeste
	Sul	Reside na Região Sul
	Sudeste	Reside na Região Sudeste
	Centro-Oeste	Reside na Região Centro-Oeste
Sexo	Masculino	É do sexo masculino
	Feminino	É do sexo feminino
Cor	Branco	É da cor branca
	Preto / Pardo	É da cor preta ou parda
Escolaridade	Sem instrução	Não tem instrução
	Fundamental incompleto	Possui como mais alto grau de instrução Ensino Fundamental incompleto
	Fundamental completo / Médio incompleto	Possui como mais alto grau de instrução Ensino Fundamental completo ou Ensino Médio incompleto
	Médio completo	Possui como mais alto grau de instrução Ensino Médio completo
	Superior incompleto ou acima	Possui como mais alto grau de instrução Ensino Superior incompleto ou acima
Faixa Etária	18 a 27 anos	Tem entre 18 e 27 anos
	28 a 40 anos	Tem entre 28 e 40 anos
	41 a 60 anos	Tem entre 41 e 60 anos
	61 ou mais anos	Tem 61 ou mais

Fonte: Elaboração própria, com base nos dados da PNAD 2013.

4.4 - RESULTADOS

Os resultados obtidos foram na sua totalidade gerados através do *software Stata* versão 12 (licença de propriedade da Faculdade de Economia da UFJF).

Nas tabelas apresentadas abaixo, temos as estimativas, em reais, da média da renda domiciliar mensal *per capita* e os respectivos erros padrão e intervalos de confiança de 95%, para todas as categorias descritas no quadro acima, quando

consideramos e quando não consideramos o Plano Amostral da PNAD 2013, além do Efeito do Plano Amostral (EPA).

Inicialmente traçamos o perfil dos domicílios brasileiros. Podemos observar que a estimativa da média da renda domiciliar mensal per capita é de R$1.119,38 sem considerar o plano amostral e de R$1.129,19 considerando o plano amostral, sugerindo uma subestimação da média quando o desenho amostral é desconsiderado. O Erro Padrão é de R$5,44 sem considerar o plano amostral e R$10,99 considerando o plano amostral, ou seja, muito menor quando não se considera o plano amostral. Por consequência disso, o intervalo de confiança de 95% é mais estreito quando não consideramos o plano amostral, tendo como limite inferior R$1.108,72 e limite superior R$1.130,04. Já quando consideramos o plano amostral, o intervalo de confiança de 95% tem como limite inferior R$1.107,64 e limite superior de R$1.150,73. O Efeito do Plano Amostral verificado foi de 4,08, o que pode ser considerado muito alto e estar indicando a presença de efeitos de conglomeração mais fortes do que os efeitos da estratificação no processo de estimação.

Tabela 1: Resultados para o Brasil

	Sem considerar o plano amostral				Considerando o plano amostral				
País	Média	Erro Padrão	IC (95%) LI	IC (95%) LS	Média	Erro Padrão	IC (95%) LI	IC (95%) LS	$\widehat{EPA}$
Brasil	1119,38	5,44	1108,72	1130,04	1129,19	10,99	1107,64	1150,73	4,08

LI e LS representam o limite inferior e superior do intervalo de confiança respectivamente

Fonte: Elaboração própria, com base nos dados da PNAD 2013.

Na tabela a seguir, os domicílios são separados pelas regiões do Brasil. Neste caso, vemos uma grande disparidade entre a estimativa da média da renda domiciliar mensal per capita na região Norte e Nordeste em comparação com a região Sudeste, Sul e Centro-Oeste. A estimativa da média da renda domiciliar mensal *per capita* ora é maior sem considerar o plano amostral, ora é maior considerando o plano amostral. O Erro Padrão é maior em todos os casos que se considera o plano amostral e por consequência disso os intervalos de confiança de 95% também são mais alargados. Um resultado muito importante de se observar é a diferença entre a estimativa da média de Sudeste e Sul que é significativa quando

ignoramos o desenho amostral e deixa de ser quando o consideramos. Os Efeitos do Plano Amostral calculados variaram entre 2,14 e 4,05, indicando novamente a presença de efeitos de conglomeração mais fortes do que os efeitos da estratificação no processo de estimação.

Tabela 2: Resultados por região

	Sem considerar o plano amostral				Considerando o plano amostral				
Região	Média	Erro Padrão	IC (95%) LI	IC (95%) LS	Média	Erro Padrão	IC (95%) LI	IC (95%) LS	$\widehat{EPA}$
Norte	808,48	9,01	790,82	826,14	784,16	14,85	755,00	813,32	2,72
Nordeste	781,64	7,99	765,98	797,30	727,83	13,46	701,42	754,24	2,84
Sudeste	1320,13	10,93	1298,71	1341,55	1327,55	21,99	1284,41	1370,69	4,05
Sul	1401,74	14,85	1372,63	1430,86	1325,14	21,71	1282,53	1367,75	2,14
Centro-Oeste	1415,80	21,26	1374,13	1457,46	1350,41	35,78	1280,15	1420,67	2,83

LI e LS representam o limite inferior e superior do intervalo de confiança respectivamente

Fonte: Elaboração própria, com base nos dados da PNAD 2013.

A próxima tabela mostra as diferenças entre domicílios cuja pessoa de referência é do sexo masculino, e domicílios cuja pessoa de referência é do sexo feminino. É possível notar que a estimativa da média da renda domiciliar mensal *per capita* é maior no caso de ser do sexo masculino. O Erro Padrão foi menor quando não se considerou o plano amostral. Os Efeitos do Plano Amostral foram 2,42 e 3,09.

Tabela 3: Resultados por sexo

	Sem considerar o plano amostral				Considerando o plano amostral				
Sexo	Média	Erro Padrão	IC (95%) LI	IC (95%) LS	Média	Erro Padrão	IC (95%) LI	IC (95%) LS	$\widehat{EPA}$
Masculino	1159,19	7,34	1144,80	1173,57	1169,48	12,90	1144,19	1194,77	3,09
Feminino	1056,42	7,89	1040,96	1071,88	1062,85	12,27	1038,80	1086,90	2,42

LI e LS representam o limite inferior e superior do intervalo de confiança respectivamente

Fonte: Elaboração própria, com base nos dados da PNAD 2013.

Temos agora na tabela seguinte a divisão por cor da pele. A estimativa da média da renda domiciliar mensal per capita de domicílios cuja pessoa de referência

é branca é muito maior que de domicílios cuja pessoa de referência é preta ou parda. Outra vez o Erro Padrão encontrado foi menor quando não se considerou o plano amostral. Os Efeitos do Plano Amostral foram 1,67 e 3,32.

Tabela 4: Resultados por cor

Cor	Sem considerar o plano amostral				Considerando o plano amostral				$\widehat{EPA}$
	Média	Erro Padrão	IC (95%) LI	IC (95%) LS	Média	Erro Padrão	IC (95%) LI	IC (95%) LS	
Branco	1512,69	10,84	1491,44	1533,94	1495,09	19,75	1456,38	1533,80	3,32
Preto / Pardo	817,59	4,59	808,59	826,60	804,81	5,94	793,16	816,46	1,67

LI e LS representam o limite inferior e superior do intervalo de confiança respectivamente

Fonte: Elaboração própria, com base nos dados da PNAD 2013.

A tabela que se segue é dividida por grau de escolaridade e mostra que quanto maior a escolaridade da pessoa de referência do domicílio, maior é a estimativa da média da renda domiciliar mensal per capita, como era de se esperar. É possível observar que domicílios cuja pessoa de referência possui Ensino Superior incompleto ou acima têm uma estimativa da média da renda domiciliar mensal per capita muito maior que os demais domicílios. Mais uma vez os Erros Padrão foram menores quando não se considerou o plano amostral. Os Efeitos do Plano Amostral variaram entre 1,42 e 2,85.

Tabela 5: Resultados por escolaridade

Escolaridade	Sem considerar o plano amostral				Considerando o plano amostral				$\widehat{EPA}$
	Média	Erro Padrão	IC (95%) LI	IC (95%) LS	Média	Erro Padrão	IC (95%) LI	IC (95%) LS	
Sem instrução	594,39	4,77	585,03	603,75	602,41	5,71	591,22	613,60	1,43
Fundamental incompleto	703,19	4,51	694,35	712,04	725,86	5,40	715,28	736,44	1,43
Fundamental completo / Médio incompleto	816,69	7,52	801,94	831,43	843,08	9,62	824,21	861,95	1,64
Médio completo	1099,70	8,76	1082,52	1116,88	1118,50	10,44	1098,02	1138,97	1,42
Superior incompleto ou acima	2958,16	28,17	2902,95	3013,38	2950,49	47,55	2857,27	3043,71	2,85

LI e LS representam o limite inferior e superior do intervalo de confiança respectivamente

Fonte: Elaboração própria, com base nos dados da PNAD 2013.

A última tabela desta sequência apresenta a estimativa da média da renda domiciliar *per capita* divida por faixas de idade da pessoa de referência do domicílio. Foram desconsiderados domicílios cuja pessoa de referência fosse menor de idade. Os resultados mostram que quanto mais velha é a pessoa de referência do domicílio maior é a estimativa da média da renda domiciliar mensal per capita. Além disso, os Erros Padrão continuam sendo menores quando não se considera o plano amostral. Os Efeitos do Plano Amostral variaram entre 1,41 e 2,42.

Tabela 6: Resultados por faixa etária

Faixa Etária	Sem considerar o plano amostral				Considerando o plano amostral				$\widehat{EPA}$
	Média	Erro Padrão	IC (95%) LI	IC (95%) LS	Média	Erro Padrão	IC (95%) LI	IC (95%) LS	
18 a 27 anos	767,54	10,02	747,90	787,19	776,60	11,92	753,22	799,97	1,41
28 a 40 anos	999,63	9,71	980,60	1018,65	1010,86	15,10	981,26	1040,46	2,42
41 a 60 anos	1158,02	8,71	1140,94	1175,09	1162,53	12,88	1137,27	1187,78	2,19
61 ou mais anos	1323,84	12,85	1298,64	1349,03	1325,23	19,51	1286,99	1363,47	2,30

LI e LS representam o limite inferior e superior do intervalo de confiança respectivamente

Fonte: Elaboração própria, com base nos dados da PNAD 2013.

5 - CONSIDERAÇÕES FINAIS

O objetivo desta monografia é mostrar a importância da consideração do planejamento amostral na análise de dados amostrais coletados por pesquisas de grande porte como a PNAD do IBGE. O uso da variável *"renda domiciliar mensal per capita"* serviu como suporte para a exemplificação de uma situação real. Sendo assim, não procuramos analisar a fundo as características de renda dos domicílios brasileiros.

A partir dos resultados apresentados na seção anterior, podemos fazer alguns comentários. Em todas as tabelas apresentadas, o Erro Padrão foi menor quando não se considerou o plano amostral em comparação com o mesmo caso, porém considerando o plano amostral. Tal fato confirma o que foi dito por Vieira (2009), por exemplo. Isto aconteceu porque o planejamento amostral da PNAD envolve conglomeração em algumas de suas etapas e um dos efeitos conhecidos da amostragem por conglomerados é que a variância dentro dos conglomerados é pequena, pois são grupos naturalmente homogêneos.

Ao ignorarmos o planejamento amostral podemos subestimar ou superestimar os Erros Padrão associados, sendo que no caso da PNAD eles são subestimados. As consequências dessa subestimação podem ser graves. Uma delas é que o intervalo de confiança passa a ter a sua cobertura comprometida. Nos exemplos mostrados, os intervalos de confiança de 95% considerando o plano amostral são bem mais largos do que os intervalos de confiança de 95% quando não se considera o plano amostral.

Uma das graves consequências de se ignorar o planejamento amostral foi vista ao compararmos as estimativas da média da variável *"renda domiciliar mensal per capita"* por regiões. Suponhamos que um governante fosse se basear nestes dados para tomar medidas políticas. Neste caso, sua decisão seria comprometida, pois a diferença que parecia ser significativa entre Sul e Sudeste quando não consideramos planejamento amostral, na verdade não era significativa, o que foi possível observar após a consideração do plano amostral.

A estimativa da média da variável *"renda domiciliar mensal per capita"* foi diferente nos casos onde consideramos o planejamento amostral, sendo ora superior, ora inferior, sugerindo a ocorrência de viés ao se desconsiderar o desenho amostral. Por se tratar de uma estimativa, poderíamos argumentar que não sabemos

qual valor mais se aproxima da realidade. Porém, a literatura da Amostragem confirma a qualidade dos resultados produzidos levando-se em consideração o plano amostral, ver por exemplo, Isaki e Fuller (1982).

Outro ponto que podemos observar é que os Efeitos do Plano Amostral calculados foram sempre superiores a 1. Não considerar o planejamento amostral no estudo significa julgar que todo o processo de amostragem foi feito por Amostragem Aleatória Simples. Portanto estes EPAs maiores do que 1 também sugerem que a Amostragem Aleatória Simples traria melhores resultados do que o planejamento amostral utilizado pela PNAD, que envolve conglomeração em algumas de suas etapas. O motivo de a PNAD utilizar tal planejamento amostral passa por questões de custo, tempo, logística, cadastro, entre outros. Neste caso, cabe a quem vai estudar os dados da PNAD aceitar o planejamento amostral escolhido pelo IBGE e fazer as considerações necessárias nas análises.

Outra observação que fizemos foi de que o Efeito do Plano Amostral foi menor para as estimativas em situações em que o número de categorias da variável auxiliar era maior, sendo este um possível efeito da estratificação. No caso das tabelas apresentadas na seção anterior podemos observar que as variáveis com menor número de subgrupos e também para o País como um todo, apresentam EPAs maiores do que as variáveis com maior número de subgrupos, como por exemplo, a variável *"Escolaridade"* que foi dividida em 5 categorias. Estes resultados confirmam, por exemplo, as amostragens descritas no texto de Skinner e Vieira (2007).

Para um trabalho futuro, podemos repetir a mesma aplicação aqui apresentada, para a PNAD Contínua, que tem periodicidade menor e comparar os resultados. Também é possível aprofundar as análises políticas e econômicas.

Esta monografia pretende encorajar as pessoas que estudam dados obtidos por técnicas de amostragem, principalmente os dados da PNAD que foram usados aqui como suporte, a considerar o planejamento amostral adotado. A amostragem é muito útil quando utilizamos os procedimentos corretos desde antes da coleta dos dados até a parte de análise dos mesmos. Caso contrário, podemos chegar a conclusões equivocadas e disseminar tal conhecimento enviesado ou ainda tomar decisões erradas que podem prejudicar empresas privadas, órgãos públicos e a população em geral.

6 - APÊNDICES

6.1 - *DO FILE STATA*

Para realizar as análises utilizando o *software Stata* é preciso antes de tudo que se tenha a base de dados organizada em um formato que possa ser acessada pelo programa (Fraga, 2010). Depois deste passo, devemos colar o código abaixo na caixa de comandos, com as devidas modificações que se façam necessárias (Santos, 2010). Observe que linhas de comandos apresentadas abaixo que se iniciam com **"*"** referem-se à comentários.

```
*Carregar arquivo com a base de dados da PNAD
use "C:\nome_do_arquivo.dta"

*Manter variáveis que vamos utilizar
keep UF V0401 V4750 V0302 V0404 V4745 V8005 V4618 V4617 V4611

*Renomear variáveis
rename V0401 condicao_familia
rename V4750 renda
rename V0302 sexo
rename V0404 cor
rename V4745 escolaridade
rename V8005 idade
rename V4618 psu
rename V4617 strat
rename V4611 peso

*Manter apenas a pessoa de referência da família
keep if condicao_familia==1

*Apagar valores sem declaração
drop if renda>9999999

*Gerar dummies regiões
gen norte=.
replace norte=0 if (UF < 10 | UF > 19)
replace norte=1 if (UF > 9 & UF < 20)

gen nordeste=.
replace nordeste=0 if (UF < 20 | UF > 29)
replace nordeste=1 if (UF > 19 & UF < 30)

gen sudeste=.
replace sudeste=0 if (UF < 30 | UF > 39)
replace sudeste=1 if (UF > 29 & UF < 40)

gen sul=.
replace sul=0 if (UF < 40 | UF > 49)
replace sul=1 if (UF > 39 & UF < 50)

gen centro_oeste=.
```

```
replace centro_oeste=0 if (UF < 50 | UF > 59)
replace centro_oeste=1 if (UF > 49 & UF < 60)

*Gerar dummies sexo
gen masculino=.
replace masculino=0 if sexo~=2
replace masculino=1 if sexo==2

gen feminino=.
replace feminino=0 if sexo~=4
replace feminino=1 if sexo==4

*Gerar dummies raça
gen branco=.
replace branco=0 if cor~=2
replace branco=1 if cor==2

gen preto_pardo=.
replace preto_pardo=0 if (cor~=4 &cor~=8)
replace preto_pardo=1 if (cor==4 | cor==8)

*Gerar dummies escolaridade
gen sem_instrucao=.
replace sem_instrucao=0 if escolaridade~=1
replace sem_instrucao=1 if escolaridade==1

gen fundamental_incompleto=.
replace fundamental_incompleto=0 if escolaridade~=2
replace fundamental_incompleto=1 if escolaridade==2

gen fundamental_medio_incompleto=.
replace fundamental_medio_incompleto=0 if (escolaridade~=3 &
escolaridade~=4)
replace fundamental_medio_incompleto=1 if (escolaridade==3 |
escolaridade==4)

gen medio_completo=.
replace medio_completo=0 if escolaridade~=5
replace medio_completo=1 if escolaridade==5

gen superior_incompleto_mais=.
replace superior_incompleto_mais=0 if (escolaridade~=6 & escolaridade~=7)
replace superior_incompleto_mais=1 if (escolaridade==6 | escolaridade==7)

*Gerar dummies faixa etária
gen faixa_etaria_18a27=.
replace faixa_etaria _18a27=0 if (idade < 18 | idade > 27)
replace faixa_etaria _18a27=1 if (idade > 17 & idade < 28)

gen faixa_etaria _28a40=.
replace faixa_etaria _28a40=0 if (idade < 28 | idade > 40)
replace faixa_etaria _28a40=1 if (idade > 27 & idade < 41)

gen faixa_etaria _41a60=.
replace faixa_etaria _41a60=0 if (idade < 41 | idade > 60)
replace faixa_etaria _41a60=1 if (idade > 40 & idade < 61)

gen idade_61mais=.
replace faixa_etaria _61mais=0 if idade < 61
replace faixa_etaria _61mais=1 if idade > 60
```

```
*Declarar plano amostral
svyset psu [pweight=peso],strata(strat) vce(linearized)
singleunit(centered) ||_n

*Comparar médias
mean renda
svy:mean renda
mean renda if norte==1
svy:mean renda if norte==1
mean renda if nordeste==1
svy:mean renda if nordeste==1
mean renda if sudeste==1
svy:mean renda if sudeste==1
mean renda if sul==1
svy:mean renda if sul==1
mean renda if centro_oeste==1
svy:mean renda if centro_oeste==1

mean renda if masculino==1
svy:mean renda if masculino==1
mean renda if feminino==1
svy:mean renda if feminino==1

mean renda if branco==1
svy:mean renda if branco==1
mean renda if preto_pardo==1
svy:mean renda if preto_pardo==1

mean renda if sem_instrucao==1
svy:mean renda if sem_instrucao==1
mean renda if fundamental_incompleto==1
svy:mean renda if fundamental_incompleto==1
mean renda if fundamental_medio_incompleto==1
svy:mean renda if fundamental_medio_incompleto==1
mean renda if medio_completo==1
svy:mean renda if medio_completo==1
mean renda if superior_incompleto_mais==1
svy:mean renda if superior_incompleto_mais==1

mean renda if faixa_etaria_18a27==1
svy:mean renda if faixa_etaria _18a27==1
mean renda if faixa_etaria _28a40==1
svy:mean renda if faixa_etaria _28a40==1
mean renda if faixa_etaria _41a60==1
svy:mean renda if faixa_etaria _41a60==1
mean renda if faixa_etaria _61mais==1
svy:mean renda if faixa_etaria _61mais==1
```

7 – REFERÊNCIAS

BARBOSA, Ana Luiza Neves de Holanda e BARBOSA FILHO, Fernando de Holanda. Diferencial de salários entre os setores público e privado no Brasil: Um modelo de escolha endógena. Pesquisa e Planejamento Econômico, 2013.

BOLFARINE, Heleno e BUSSAB, Wilton O. Elementos de Amostragem. São Paulo: Blucher, 2005.

COCHRAN, William G. Técnicas de Amostragem. Rio de Janeiro: Fundo de Cultura, 1965.

CRUZ, José. Amostragem Estatística – Noções Básicas. Aracaju ed. Universidade Federal de Sergipe, 1978

IBGE. Pesquisa Nacional por Amostra de Domicílios. Disponível em: www.metadados.ibge.gov.br/consulta/dthPesquisa.aspx?codPesquisa=PD. Acesso em 9 de Junho de 2015.

ISAKI, C. T. FULLER, W. A. Survey Design Under the Regression Superpopulation Model. Journal of the American Statistical Association: Vol. 77, n. 377, 89-96, 1982.

KISH, Leslie. Survey Sampling. New York: John Wiley, 1995.

LUNET, Nuno. SEVERO, Milton. BARROS, Henrique. Desvio Padrão ou Erro Padrão. Notas Metodológicas. Serviço de Higiene e Epidemiologia da Faculdade de Medicina da Universidade do Porto: Arquivos de Medicina. Vol. 20, Nº 1/2. Portugal, 2006.

MORETTIN, Pedro A. BUSSAB, Wilton de O. Estatística Básica. São Paulo: Saraiva, 2013.

MOURA, Rodrigo L. Testando as Hipóteses do Modelo de Mincer para o Brasil. Revista Brasileira de Economia, 2008.

SANTOS, Gilnei Costa. Tratamento e extração dos microdados da PNAD. Aula prática. Viçosa, 2010.

SILVA, Pedro Luis do Nascimento. PESSOA, Djalma Galvão Carneiro. LILA, Maurício Franca. Análise estatística de dados da PNAD: incorporando a estrutura do plano amostral. Ciência & Saúde Coletiva: Vol.7, nº4, Rio de Janeiro, 2002.

SILVA, Pedro Luis do Nascimento; PESSOA, Djalma Galvão Carneiro. Análise de Dados Amostrais Complexos. IBGE, 1998.

SKINNER C. VIEIRA M. D. T. Variance estimation in the analysis of clustered longitudinal survey data. Survey Methodology, 2007.

VIEIRA, M. D. T. Notas de aula de Amostragem. Juiz de Fora, Universidade Federal de Juiz de Fora, 2013.

VIEIRA, M. D. T. Analysis of Longitudinal Survey Data. 1. Saarbrücken: VDM Verlag Dr. Müller, 2009.

WOLTER, Kirk M. Introduction to Variance Estimation. New York: Springer, 2007.

FRAGA, Roney. Tutorial extrair dados da PNAD com Stata. Disponível em: www.youtube.com/watch?v=G5RiNKkh7Hs . Acesso em 9 de Junho de 2015.