

DECLARAÇÃO DE HONRA

Declaro que este trabalho é resultado da minha própria investigação, que não foi submetido para outro grau que não seja o indicado – **Licenciatura em Estatística**, da Universidade Eduardo Mondlane.

Maputo, aos ____ de _____ de 2015

(Nelsa Braz Impene Combo)

DEDICATÓRIA

Dedico este trabalho ao meu pai

Braz Impene Combo

À minha mãe

Mariato Macário

E a toda família Combo.

AGRADECIMENTOS

Agradeço primeiramente ao nosso Senhor todo poderoso Deus pela vida, saúde e acompanhamento durante o meu percurso académico, aos meus pais pelo amor, carinho, educação, apoio moral e financeiro em todos momentos da minha vida.

Aos meus irmãos Isilda, Dulce, Braz e Edvilce pelo amor, carinho, compreensão e força de enfrentar todos momentos bons e maus desta longa caminhada.

Agradecer também ao meu marido Benessone e a minha tia Muachema por nunca me terem deixado cair em fracasso e proporcionarem uma vida feliz e de amparo para mim, de amarem e cuidarem de mim, e de me darem sabedoria e ensinamento das coisas da vida.

Ao meu supervisor Carlos Creva Singano por ter-me dado toda atenção e orientação no trabalho, pela paciência, dedicação e competência que teve no percurso do trabalho e privilegiá-lo por ser meu supervisor.

O meu especial agradecimento vai para meus colegas e amigos Rito, Gélcio, Narciso e Scothy por terem-me ajudado academicamente e moralmente em todos momentos vividos como estudante, e por ajudarem-me a enfrentar as barreiras da vida.

Por fim agradecer as minha colegas do quarto Faiza, Iveth, Angelina e Cultódia e a todos que de forma directa ou indirectamente contribuíram e fizeram parte da minha vida académica e social.

Meu muito obrigado!

RESUMO

O método de amostragem estratificada tem a vantagem de ser mais eficiente do que os métodos de amostragem simples ou sistemática, pois é mais económico em termos de tempo e dinheiro e fornece resultados com menor erro associado, sendo assim, faz-se estudo dos métodos para validação e controle de qualidade do Inquérito Contínuo aos Agregados Familiares (INCAF) 2012-13 da província de Nampula com domínio de base o plano de sondagem estratificada.

Usam-se os métodos de re-amostragem de Jackknife e BRR e o método linear de conglomerados últimos para estimação do erro-padrão, do CV, do Deff e dos intervalos de confiança das estimativas para validação do inquérito. Os dados para a realização do estudo são secundários, correspondentes ao INCAF_2012/3 (trimestre I) da província de Nampula, obtidos no INE de Moçambique. Do estudo foi concluído que os melhores métodos de desenho de amostragem que aproveitam as particularidades da sondagem estratificada são os métodos de re-amostragem ou replicados, os métodos de re-amostragem são mais precisos em relação aos métodos lineares e que o inquérito é credível pois os resultados do INCAF/2012/13 – trimestre I são válidos e representativos da população de Nampula.

Palavras-chaves: Amostragem, Plano de Sondagem Aleatório Estratificado, Método Jackknife, Método BRR, e Conglomerados Últimos.

Lista de Abreviatura

AAE	Amostragem Aleatória Estratificada
AAS	Amostragem Aleatória Simples
AASSR	Amostragem Aleatória Simples Sem Reposição
AC	Amostragem por Conglomerado
AE	Área de Enumeração
AF	Agregados Familiares
AM	Amostragem Multi-etápica
AS	Amostragem Sistemática
BRR	Replicação Repetida Balanceda
CV	Coeficiente de Variação
Deff	Efeito de Desenho
INCAF	Inquérito Contínuos aos Agregados Familiares
INE	Instituto Nacional de Estatística
UPA	Unidade Primária de Amostragem

Lista de Tabela

Tabela 1	Parâmetros para cálculo de conglomerados.....	28
Tabela 2	Estatísticas para cálculo de conglomerados.....	29
Tabela 3	Diferença entre os tipos de erros.....	37
Tabela 4	Uma amostra aleatória estratificada pequena, usada para ilustrar o BRR.....	48
Tabela 4.1	Estáticas descritivas da variável sexo.....	57
Tabela 4.2	Estáticas descritivas da variável idade.....	58
Tabela4.3	Comparação dos métodos.....	60
Tabela4.4	Qualidade dos métodos para variável idade usando como estimativa a média.....	61
Tabela 4.5	Qualidade dos métodos para variável sexo usando como estimativa a proporção..	61
Tabela4.6	Qualidade dos métodos para variável sexo usando como estimativa o total.....	62
Tabela4.7	Qualidade dos métodos para variável idade usando como estimativa o total.....	62

Lista de Figura

Figura 4.1	Coeficiente de variação para idade média dos métodos BRR e Jackknife61
------------	--

ÍNDICE

I.	INTRODUÇÃO.....	1
1.1.	Definição do Problema de Estudo.....	2
1.2.	Justificação do Estudo.....	3
1.3.	Objectivos.....	3
1.3.1.	Geral.....	3
1.3.2.	Específicos.....	3
1.4.	Estrutura do Trabalho.....	4
II.	REVISÃO DA LITERATURA.....	5
2.1.	Conceitos de Teoria da Amostragem.....	5
2.2.	Plano Amostral.....	8
2.3.	Amostras Aleatórias ou Probabilísticas.....	9
2.4.	Amostras Não Aleatórias ou Não Probabilística.....	10
2.5.	As Fases de um Processo de Amostragem ou do Plano Amostral.....	11
2.5.1.	A Identificação da População Alvo e População Inquirida.....	12
2.5.2.	O Método de Selecção da Amostra ou Método de Amostragem.....	12
2.5.3.	Dimensionamento da Amostra.....	13
2.6.	Ponderações de Amostragem.....	31
2.7.	Critérios para Formação das Upas.....	32
2.8.	Erros de Amostragem.....	32
2.8.1.	Diferença entre os Tipos de Erros.....	34
2.8.2.	Formas de Evitar Erros:.....	34
2.9.	Probabilidade de inclusão: Estimador do Tipo Horvitz-Thompson.....	35
2.10.	Principais Métodos de Estimação de Variâncias em Populações Finitas.....	37
2.11.	Métodos de Re-amostragem ou Replicados.....	38
2.11.1.	Método de Jackknife.....	39
2.12.2.	Replicação Repetida Balanceada (BRR).....	44
III.	MATERIAL E MÉTODOS.....	51
3.1.	Material.....	51
3.2.	Métodos.....	51
IV.	RESULTADOS E DISCUSSÃO.....	54
V.	CONCLUSÕES E RECOMENDAÇÕES.....	61

5.1. Conclusões.....	61
5.2. Recomendações.....	61
REFERÊNCIAS BIBLIOGRÁFICAS.....	62

I. INTRODUÇÃO

Para fazer face a crescente necessidade de informação, tanto por parte das empresas e instituições, como por parte de particulares, surgiu a necessidade de desenvolver métodos estatísticos que permitem recolher informações a partir da observação de uma parte da população. Esses métodos são conhecidos como métodos de amostragem e quando associados a uma probabilidade requerem a existência de uma base de sondagem (Costa, 2000).

Uma sondagem é um estudo científico da parte de uma população com o objectivo de estudar atitudes, hábitos, preferências, acontecimentos, circunstâncias e assuntos de interesse comum de modo que os resultados sejam generalizados para a população. Sendo assim, base de sondagem é uma listagem dos elementos da qual se vai seleccionar a amostra (Campos, nd).

O conceito população significa conjunto de unidades individuais, com uma ou mais características comuns, que se pretendem estudar (Marenda, 2010).

A determinação ou definição da população alvo é uma das fases mais importantes na realização de uma sondagem. Pois é sobre essa população que o estudo vai incidir ou incorrer (Costa, 2000).

Por definição, população alvo é a totalidade dos elementos sobre os quais incide uma análise e dos quais se pretende obter informação. Para definir correctamente a população alvo, primeiro deve-se conhecer qual é o objectivo do estudo, e depois a incidência do estudo (Campos, nd).

Um inquérito pode ser considerado como uma interrogação particular ou colectiva acerca de uma situação englobando indivíduos, com o objectivo de obter informação e quando possível generalizar (Campos, nd).

Quando se realiza um estudo por amostragem, é preciso avaliar a qualidade dos resultados, quer dizer, avaliar a precisão das estimativas resultantes. A avaliação de qualidade da implementação de uma pesquisa faz-se habitualmente, através da avaliação da qualidade do plano de sondagem ou do plano de amostragem (Lemm, 2013).

Um plano de sondagem é o processo a adoptar na recolha de elementos a incluir na amostra. E um método de sondagem é uma técnica ou método adoptado na recolha dos elementos a incluir na amostra (Sousa, 2011), incluindo o processo de avaliação da qualidade dos resultados.

1.1. Definição do Problema de Estudo

Nos últimos tempos, é cada vez maior a aplicação de inquéritos por amostragem. Vários inquéritos são realizados anualmente por Institutos Oficiais de Estatística (IOE), instituições académicas, agentes privados e governamentais a fim de fornecer informação estatística fiável (Lemm, 2013).

De acordo com Lemm (2013), os primeiros resultados de um inquérito por amostragem probabilística são as estimativas pontuais. Contudo, com igual importância, é a estimação das variâncias associadas a estas estimativas. A importância da estimação de variâncias e os erros padrão correspondentes, justifica-se pelo facto da variância ser uma medida de qualidade das estimativas pontuais.

O Instituto Nacional de Estatística (INE) de Moçambique realizou no período 2007-2014 quatro inquéritos de grande escala (IOF 2008/09-Inquérito aos Orçamentos Familiares; MICS 2008-Inquérito sobre Indicadores Múltiplos; MCP-Inquérito sobre Parceria Sexual Múltipla e INCAF 2012/13-Inquérito Contínuo aos Agregados Familiares). Contudo, com a excepção dos primeiros três em que foi feita estimação dos erros de amostragem das estimativas principais para os domínios planeados (Nacional, Nacional Urbano e Rural e provincial), no último inquérito (INCAF 2012/13) não foi feita a estimação dos erros de amostragem das estimativas principais ao nível de cada uma das principais variáveis, devido a várias razões (Lemm, 2013)¹.

Assim, pretende-se fazer uma aplicação prática dos métodos de estimação de variâncias programados nos vários softwares para validar um inquérito, utilizando dados provenientes do INE, especificamente do INCAF 2012/13, província de Nampula como um domínio de análise.

Para além dos métodos analíticos (ou directos) de estimação de variâncias, existem duas categorias de métodos de estimação que competem entre si: o método de linearização de Taylor que se baseia na série de expansão de 1ª ordem de Taylor e os métodos replicados que re-utilizam a informação amostral (Rust, 1985).

¹ Lemm, Anchile & D'Agostino, Antonella (2013). Evaluation Report INCAF 2012-2013 survey Maputo, 15th-27th September 2013 Mission Report.

1.2. Justificação do Estudo

As áreas de aplicação das sondagens estatísticas são muito diversas, tendo especial destaque os estudos das populações humanas, sob a forma de estudos pré-eleitorais, de opinião pública ou para planificação de políticas para o bem-estar.

A motivação para a escolha do tema reside no facto de, conhecer os métodos práticos de estimação de variâncias utilizados para fazer o controle de qualidade e validação dos inquéritos realizados, pois, os métodos analíticos (ou directos) de estimação de variâncias não estão programados directamente nos vários softwares de estimação da estatística.

Como metodologia de estudo, a sondagem possibilita o conhecimento momentâneo de um universo de elementos, numa perspectiva descritiva e quantificada. A escolha e análise de dados é feita com base numa amostra de elementos que deverá permitir a extrapolação das interpretações à totalidade do universo.

1.3. Objectivos

1.3.1. Geral:

Estudar os métodos de estimação de erros de amostragem das variáveis para validação e controle de qualidade do INCAF 2012-13 na província de Nampula com domínio de base o plano de sondagem estratificada.

1.3.2. Específicos:

- Identificar os métodos usados na validação de um inquérito;
- Identificar os melhores métodos que tirem proveito da sondagem estratificada;
- Estimar a precisão das estimativas principais do INCAF 2012-13 na província de Nampula;
- Fazer uma análise comparativa da qualidade das estimativas usando vários métodos e identificar aquele que melhor tira proveito do plano de sondagem estratificada.

1.4. Estrutura do Trabalho

A introdução feita acima representa o capítulo um do trabalho. Após esta introdução, o trabalho está estruturado do seguinte modo:

- O capítulo dois, trata da revisão da literatura onde se definem os conceitos da teoria de amostragem, o plano amostral, o tipo de amostras que se podem usar num plano amostral, as fases de um plano amostral, o dimensionamento da amostra usando os diferentes métodos da amostragem probabilística e os critérios para elaboração da base de sondagem, os métodos de re-amostragem e vantagens do seu uso.
- O capítulo três, trata sobre material e métodos utilizados para a obtenção e análise dos dados e os diferentes métodos aplicados no processamento dos dados.
- O capítulo quatro, mostra as discussões e os resultados obtidos no processamento dos dados.
- No capítulo cinco e último do trabalho, mostra as conclusões chegadas e as recomendações para futuros estudos.

II. REVISÃO DA LITERATURA

2.1. Conceitos de Teoria da Amostragem

A teoria da amostragem estuda as relações existentes entre uma população e as amostras extraídas dessa população (Cavalcante, 2009).

O objectivo principal da teoria da amostragem é obter uma amostra que seja uma representação da população e que conduza à estimação das características da população com grande precisão (Marenda, 2010).

De acordo com Costa (2000), o termo amostragem refere-se ao processo de seleccionar uma amostra de uma população, bem como, fazer inferência de estimativas para a população. O objectivo da teoria de amostragem é de seleccionar uma parte (amostra) de um conjunto de elementos (população) e com base na informação recolhida dessa amostra, inferir sobre determinada característica de interesse da população.

Segundo Tavares (2008), amostragem é todo o processo de recolha de uma parte, geralmente pequena, dos elementos que constituem um dado conjunto. Da análise dessa parte obtém-se informações para todo o conjunto.

De acordo com as definições dadas entende-se que, amostragem é um conjunto de técnicas aplicadas para seleccionar uma parte da população, com a finalidade de fazer generalizações sem precisar examinar todos os elementos de um dado grupo.

Enquanto que a técnica da amostragem selecciona parte de uma população e observa-a, com vista a estimar uma ou mais características para a totalidade da população, um censo envolve um exame a todos os elementos da população (Costa, 2000).

De acordo com Marenda (2010), o censo apresenta dificuldades que tornam a amostragem um processo mais atraente. Entre as dificuldades que o censo apresenta pode enumerar-se as seguintes:

- Quando a população for infinita, o censo seria impossível;
- A amostra pode ser actualizada mais facilmente que o censo;
- O custo do censo pode torná-lo proibitivo;
- A precisão de um censo varia de acordo com o tamanho da população examinada;
- Factores tempo e custo podem apontar pela preferência de uma amostra que um censo.

Porém, ainda com base em Marenda (2010), há ocasiões que o levantamento do censo é vantajoso:

- Quando a população for pequena e o custo entre o censo e a amostra forem praticamente iguais;
- Se o tamanho da amostra necessária tiver que ser muito grande em relação à população que se deseja examinar;
- Nas ocasiões em que se exige precisão completa;
- Nas ocasiões em que já existe informação completa.

Segundo Bolfarine (2005), população ou universo é o conjunto de todas as unidades elementares de interesse. É indicado por

$$U = \{1, 2, 3, \dots, N\} \quad (1)$$

Onde N é um número fixo e conhecido da população.

Elemento populacional é a nomenclatura usada para denotar qualquer elemento $i \in U$. É também conhecido por unidade elementar. Unidade elementar ou simplesmente elemento de uma população é o objecto ou entidade portadora das informações que pretende-se recolher. (Bolfarine, 2005).

Unidade de amostragem ou unidade estatística é o elemento da população considerada e sobre o qual vai ser estudada a característica de interesse.

Segundo Costa (2000), uma característica é aquilo que caracteriza, é a propriedade específica de um ser ou de uma classe de seres; é o que se pretende estudar (exemplo: peso, idade, etc).

Com base em Marendá (2010), as características podem ser de natureza quantitativa e neste caso consideram-se escalas numéricas nas quais as variáveis se podem classificar em:

- Contínuas (referem-se a medições, pesagens, etc);
- Discretas (referem-se a contagens).

Ou de natureza qualitativa e neste caso classificam-se em:

- Nominais (ex: sexo, espécie de uma dada planta ou animal, etc);
- Ordinais (ex: itens de valores de uma dada classificação).

Segundo Bolfarine (2005), uma amostra é uma sequência qualquer de n unidades de U , isto é,

$$s = (k_1, k_2, \dots, k_n) \quad (2)$$

Seja $f_i(s)$, a variável que indica o número de vezes (frequência) que a i -ésima unidade populacional aparece na amostra S . Seja $\delta_i(s)$ a variável binária que indica a presença ou não da i -ésima unidade na amostra S , isto é,

$$\delta_i(s) = \begin{cases} 1, & \text{se } i \in s \\ 0, & \text{se } i \notin s \end{cases}$$

De acordo com Costa (2000), dimensão ou tamanho da amostra é o número de elementos que constituem a amostra.

Bolfarine (2005) define tamanho $n(s)$ da amostra S como a soma das frequências das unidades populacionais na amostra, isto é,

$$n(s) = \sum_{i=1}^N f_i(s) \quad (3)$$

Chama-se tamanho efectivo $v(s)$ da amostra (S) o número de unidades populacionais distintas presentes na amostra S , isto é,

$$v(s) = \sum_{i=1}^N \delta_i(s)$$

De acordo com Bolfarine (2005), chama-se dados da amostra S a matriz ou vector das observações pertencentes a amostra, isto é,

$$d_s = (Y_{k_1}, Y_{k_2}, \dots, Y_{k_n}) = (Y_{k_i}, k_i \in s)$$

Quando S percorre todos os pontos possíveis de um plano amostral s_A , tem-se associado um vector aleatório que é representado por

$$d = y = (y_1, y_2, \dots, y_i, \dots, y_n)$$

Um estimador é uma função que estima o valor de um parâmetro, baseando-se nas observações de uma amostra. Estimativa é o valor do estimador, calculado com base na amostra que se recolheu. E amostra enviesada é uma amostra que não é representativa da população (Vieira, 2008).

Um estimador ponderado é uma função que estima o valor de um parâmetro associado ao peso de cada observação de uma amostra.

Um plano de amostragem é o processo a adoptar na recolha de elementos a incluir na amostra. E um método de amostragem é uma técnica ou método adoptado na recolha dos elementos a incluir na amostra (Sousa, 2011).

2.2. Plano Amostral

Depois de identificados a população sobre a qual se pretende recolher os dados e o instrumento a utilizar para a recolha, define-se um método de amostragem adequado ao tipo de dados e ao instrumento de análise, isto é, é necessário que se estabeleça, à priori, um plano de amostragem de acordo com a população alvo, a definição da população a inquirir e, de acordo com um processo adequado de administração do instrumento de análise.

O plano de amostragem começa por determinar o nível de extensão geográfica em que o processo de amostragem deverá ser conduzido: mundial, nacional, regional, urbano, rural, grupo de indivíduos, etc (Vieira, 2008).

Segundo Vieira (2008), os critérios para escolha de um plano amostral envolve três etapas:

1. A identificação da população alvo e população inquirida;
2. O método de selecção da amostra ou método de amostragem;
3. A dimensão da amostra.

2.3. Amostras Aleatórias ou Probabilísticas

Segundo Sousa 2011:

- Nos métodos de amostragem aleatória, a selecção de elementos ou grupo de elementos da população é feita de modo que, cada elemento da população, tenha uma probabilidade de inclusão na amostra, calculável e diferente de zero, ou seja, cada elemento da população tem uma probabilidade conhecida de ser seleccionado;
- Só usando este tipo de amostra, é possível conhecer o grau de confiança (isto é, o grau de certeza que se tem a respeito da precisão da estimativa) dos resultados;

- Os critérios de selecção dos elementos estão, rigorosamente, definidos não permitindo que a subjectividade ou arbítrio do julgamento humano intervenha na recolha dos elementos;
- Possibilita calcular, matematicamente, a dimensão da amostragem em função da precisão e grau de confiança desejados para os resultados;

O mais importante é obter estimativas próximas dos parâmetros a estimar e isto só se consegue, se se tiver uma amostra, a mais representativa possível do universo.

De acordo com Sousa (2011), os principais casos de amostragem probabilística são:

- *Amostragem Aleatória Simples (AAS)*: É indicada quando uma amostra de tamanho n é seleccionada de uma população de tamanho N , na qual todas as amostras possíveis de tamanho n , têm a mesma probabilidade de ser seleccionada.
- *Amostragem Sistemática (AS)*: É um tipo de amostragem em que o plano de amostragem é obtido por um critério pelo qual intervalos regulares de mesmo tamanho entre unidades da amostra são tomados até se compor uma amostra de tamanho n . Para implementar o processo de amostragem sistemática, deve-se conhecer inicialmente o tamanho de sua população e determinar o tamanho da amostra desejada, e pela razão entre o tamanho da população N e da amostra n determina-se o número de intervalos.
- *Amostragem Aleatória Estratificada (AAE)*: É obtida por separar os elementos da população em grupos não sobrepostos, chamados de estratos, e então selecciona-se uma amostra aleatória de cada estrato formado.
- *Amostragem por Conglomerado ou Clusters (AC)*: É uma amostra probabilística em que cada unidade amostral é uma colecção, ou grupo de elementos. Ex.: Um quarteirão de uma cidade, que consiste de um conjunto (ou conglomerado) de domicílios;
- *Amostragem Multi-etápica (AM)*: É uma amostra probabilística em que apenas as unidades finais dos conglomerados são estudadas.

2.4. Amostras Não Aleatórias ou Não Probabilística

De acordo co Sousa 2011:

- Nos métodos de amostragem não aleatória, a selecção de elementos da população permite a escolha dos indivíduos a incluir na amostra, segundo determinado critério, mais ou menos subjectivo. Nesta forma de amostragem, não se conhece a probabilidade de determinado elemento ser seleccionado.
- São este tipo de amostras que possibilitam a conclusão mais rápida do estudo e com menos custos;
- Mas quando um critério subjectivo é aplicado na selecção da amostra, há o inconveniente de não se saber com que graus de confiança são as conclusões obtidas e generalizáveis à população.

Segundo Tavares (2008), os principais casos de amostragem não-probabilística são:

- *Inacessibilidade a toda população*: quando o investigador não tem acesso a toda população de estudo, somente uma parte dela está disponível.
Exemplo: população usuárias de drogas da Coop (não existe cadastro).
- *Material contínuo*: devido a característica da continuidade é impossível realizar sorteio.
Exemplo: Retirar amostras de água em diferentes pontos de um rio para avaliar a poluição.
- *Amostragem por quotas*: Inclui unidades amostrais na amostra segundo diversas características da população e nas mesmas proporções que figuram na população.
Exemplo: idade, sexo, nível sócio-económico, etc.
- *Amostragem por julgamento (ou conveniência)*: Inclui na amostra as unidades estatísticas que poderão proporcionar uma representatividade da população, de acordo com a lógica, senso comum ou um julgamento equilibrado.

Esse é o processo de amostragem que ocorre quando se faz um estudo de opinião pública em uma cidade, os investigadores vão amostrando as pessoas que vão aparecendo e passando por eles. Esse processo de amostragem ocasiona muitas distorções na amostra e por isso ela deve ser usada apenas em casos especiais, como quando a população estudada é muito homogênea.

- *Amostragem Intencional*: é realizada quando o investigador intencionalmente escolhe um grupo de elementos para compor a amostra e se dirige intencionalmente a esses elementos interessados em sua opinião. De acordo com determinado critério, é escolhido intencionalmente um grupo de elementos que irão compor a amostra.

- *Amostragem por voluntário*: Quando o indivíduo se apresenta para fazer parte da amostra. É um método muito aplicado em investigações médicas.
- *A esmo (ou sem norma)*: O investigador procura ser aleatório sem, no entanto, realizar propriamente o sorteio.
Exemplo: Misturar 10000 parafusos e retirar 100.

2.5. As Fases de um Processo de Amostragem ou do Plano Amostral

Como visto de acordo com Vieira (2008), a construção de um plano amostral envolve três etapas que serão detalhadas a seguir:

2.5.1. A Identificação da População Alvo e População Inquirida

A População alvo é a totalidade dos elementos sobre os quais se deseja obter determinado conjunto de informações. Mas, em muitas situações, não é operacional estudar uma amostra, retirada da população alvo, sendo, por isso necessário definir qual a população a inquirir, não coincidente com a população alvo, da qual se irá retirar a amostra (Vieira, 2008).

Por exemplo: um estudo telefónico sobre a opinião da população mocambicana acerca de uma nova marca de leite, colocada à venda no mercado. A população alvo é constituída por todos os mocambicanos, mas a população inquirida só pode ser constituída por todos os mocambicanos que têm telefone.

2.5.2. O Método de Selecção da Amostra ou Método de Amostragem

De acordo com Vieira (2008), existem dois grupos de métodos para seleccionar e recolher amostras: os métodos aleatórios ou probabilísticos e os métodos não aleatórios ou não probabilísticos.

Uma amostragem probabilística é a mais recomendável sempre que possível, pois dessa forma pode-se garantir a representatividade da amostra e as possíveis discrepâncias entre a população e a amostra será apenas obra do acaso.

Os métodos de amostragem não-probabilística são efectuados de forma tendenciosa, onde o desejo do investigador ou de quem está recolhendo os dados interfere directamente na escolha dos elementos que compoem a amostra. O facto das escolhas serem direccionadas não permite que os resultados obtidos a partir das amostras representem a população a qual elas pertencem.

Os métodos de sondagem probabilísticos são caracterizados por atribuírem a toda e qualquer elemento da população uma probabilidade, calculável e diferente de zero², de ser escolhida para integrar a amostra (Särndal et al. 1992). Cochran (1977) e apresenta quatro propriedades matemáticas necessárias para a obtenção de uma amostra probabilística de uma determinada população:

- ser possível definir o conjunto de todas as amostras $S = (s_1, s_2, s_3, \dots, s_m)$, que podem ser seleccionadas para uma população específica.
- cada amostra s_i tem uma probabilidade $p(s_i)$ conhecida de ser seleccionada.
- cada amostra s pode ser seleccionada através de um processo aleatório, onde a cada s é atribuída uma probabilidade π de ser seleccionada.
- o processo de estimação dos parâmetros estatísticos e dos erros de amostragem associados se baseia na amostra seleccionada.

Crítérios de Comparação de Estimadores

Reis (1998), define o erro quadrático médio (EQM) do estimador como sendo o valor esperado do erro amostral ao quadrado, isto é

$$EQM(T, \theta) = E(e^2) = E((T - \theta)^2)$$

Quando os estimadores não são centrados utiliza-se o erro quadrado médio (EQM) para comparar o melhor estimador. Pode parecer intuitivo que, ao utilizar estimadores não-viciados, tenhamos um erro quadrático médio pequeno, porém nem sempre isso ocorre, ou seja, controlar o vício do estimador não garante um controle do erro quadrático médio. As vezes, um estimador com um pequeno aumento no vício pode gerar um grande decrescimento na variância e, conseqüentemente, um erro quadrático médio menor (Reis, 1998).

² De acordo com Vicente et al. (2001:48), “calculável e diferente de zero” significa que todos os elementos da população têm alguma possibilidade de ser seleccionada para a amostra e que essa probabilidade de inclusão pode ser determinada, mesmo que não seja conhecida à priori.

Como o erro quadrático médio é uma função do parâmetro, não pode-se, em geral, dizer que existe um melhor estimador a partir da comparação de seus erros quadráticos médios. Porém, sua informação pode, eventualmente, fornecer um guia a seguir na escolha entre estimadores (Reis, 1998).

2.5.3. Dimensionamento da Amostra

De acordo com Cochran (1977), o dimensionamento da amostra varia de acordo com o plano amostral probabilístico desejado.

2.5.3.1. Dimensionamento da Amostra na Amostragem Aleatória Simples (AAS)

Segundo Lutz (2007), do ponto de vista da quantidade de informação contida na amostra, amostragem sem reposição é mais adequado que amostragem com reposição, visto que, a amostragem sem reposição é mais eficiente que a amostragem com reposição e reduz a variabilidade, uma vez que não é possível retirar elementos extremos mais do que uma vez.

Segundo Cochran (1977), para o caso da Amostragem Aleatória Simples (AAS) tem-se o seguinte para a população considerada infinita:

Sendo \bar{y} a média das observações de uma AAS, deve-se ter

$$\Pr(|\bar{y} - \bar{Y}| \geq d) = \alpha \quad (4)$$

Onde

d é a margem de erro escolhida

α é uma pequena probabilidade e

Admitindo-se que \bar{y} seja normalmente distribuído, seu erro padrão é

$$\sigma_{\bar{y}} = \sqrt{\frac{N-n}{N}} * \frac{S}{\sqrt{n}} \quad (5)$$

Portanto a fórmula que relaciona n com o desejado grau de precisão é

$$d = t * \sqrt{\frac{N-n}{N}} * \frac{S}{\sqrt{n}} \quad (6)$$

Onde t é a abscissa da curva de frequência normal, que define uma área α , na extremidade de seus ramos. Tirando o valor de n nessa expressão, tem-se

$$n = \frac{\left(\frac{t^* S}{d}\right)^2}{1 + \frac{1}{N} \left(\frac{t^* S}{d}\right)^2} \quad (7)$$

No caso em que N é grande, uma primeira aproximação é

$$n_0 = \left(\frac{t^* S}{d}\right)^2 = \frac{S^2}{V} \quad (8)$$

Onde

$$V = \frac{d^2}{t^2} = \text{variância desejada da amostra}$$

Na prática calcula-se primeiramente o n_0 . Esse valor n_0 é aceitável, salvo quando $\frac{n_0}{N}$ tem valor desprezível, caso contrário calcula-se n pela fórmula

$$n = \frac{n_0}{1 + \frac{n_0}{N}} \quad (9)$$

Cálculo de n na AAS pelas Proporções

Ainda de acordo com Cochran (1977), admite-se uma certa margem de erro d , na proporção estimada P , e há um pequeno risco α , que está-se disposto a aceitar de que o erro real seja maior que d . Assim sendo deseja-se que

$$\Pr(|P - P| \geq d) = \alpha$$

Presume-se que a amostra é aleatória simples e admite-se que P seja normalmente distribuído. Então

$$\sigma_p = \sqrt{\frac{N-n}{N-1}} * \sqrt{\frac{P*Q}{n}} \quad (10)$$

Portanto a fórmula que relaciona n com o grau de precisão desejado é

$$d = t * \sqrt{\frac{N-n}{N-1}} * \sqrt{\frac{P*Q}{n}} \quad (11)$$

Tirando o valor de n nessa expressão, tem-se

$$n = \frac{\frac{t^2 * P * Q}{d^2}}{1 + \frac{1}{N} * \left(\frac{t^2 * P * Q}{d^2} - 1 \right)} \quad (12)$$

Para o uso na prática, substitui-se P , na fórmula acima por uma estimativa antecipada, P . No caso em que N é grande, uma primeira aproximação é

$$n_0 = \frac{t^2 * p * q}{d^2} = \frac{p * q}{V} \quad (13)$$

Na prática, calcula-se primeiramente n_0 . Se $\frac{n_0}{N}$ for desprezível, então, n_0 é uma expressão satisfatória para o n apresentado. Em caso contrário, é evidente, pela comparação das expressões, que n é obtido pela fórmula

$$n = \frac{n_0}{1 + (n_0 - 1) / N} = \frac{n_0}{1 + (n_0 / N)} \quad (14)$$

2.5.3.2. Dimensionamento usando o Plano de Amostragem Aleatória Estratificada (AE)

Segundo Cochran (1977), na amostragem estratificada, a população de N unidades é dividida em subpopulações de N_1, N_2, \dots, N_L unidades respectivamente. Essas população não se sobrepõem e juntas abrangem a totalidade da população, de tal modo que:

$$N_1 + N_2 + \dots + N_L = N \quad (15)$$

As subpopulações são denominadas estratos. Para obter todo o proveito da estratificação, os valores de N_h devem ser conhecidos. Depois de determinar os estratos, selecciona-se uma amostra em cada um deles, sendo as selecções feitas separadamente, nos diferentes estratos. As grandezas das amostras dentro dos estratos são denominadas n_1, n_2, \dots, n_L respectivamente.

De acordo com Cochran (1977), a estratificação é uma técnica principalmente usada nos seguintes casos:

- Quando se desejam dados de determinada precisão sobre certas subdivisões da população, é aconselhável tratar cada uma das subdivisões como uma população;
- Os problemas da amostragem podem-se apresentar sensivelmente diferentes em partes diversas da população. Nas populações humanas, as pessoas que vivem em instituições como hotéis, hospitais, prisões, lares, etc, são frequentemente colocadas em um estrato separado das pessoas que vivem em domicílios normais, porque a cada uma das duas situações correspondem a uma maneira adequada de conduzir a amostragem.
- A estratificação pode proporcionar um aumento de precisão nas estimativas das características da totalidade da população.

A estratificação divide uma população heterogênea em subpopulação que, isoladamente, são homogêneas (estratos). Se os estratos forem homogêneos, no sentido de que o valor das medidas variem pouco de uma unidade para outra, pode-se obter uma estimativa precisa do valor médio de um estrato qualquer mediante uma pequena amostra desse estrato. Depois, essas estimativas podem ser combinadas para constituírem uma estimativa precisa do conjunto da população (Cochran, 1977).

Notação de Estratificação

O índice inferior h indica o estrato, e o i , a unidade dentro do estrato. Todos os símbolos referem-se ao estrato h :

N_h	número total das unidades
n_h	número de unidades da amostra
y_{i_h}	valor obtido para a unidade de ordem i
$W_h = \frac{N_h}{N}$	peso do estrato
$f_h = \frac{n_h}{N_h}$	fracção amostral do estrato
$\bar{Y}_h = \frac{\sum_{i=1}^{N_h} y_{h_i}}{N_h}$	valor médio verdadeiro
$\bar{y}_h = \frac{\sum_{i=1}^{n_h} y_{h_i}}{n_h}$	valor médio da amostra
$S_h^2 = \frac{\sum_{i=1}^{N_h} (y_{h_i} - \bar{Y}_h)^2}{N_h - 1}$	variância verdadeira

Propriedade das Estimativas

Com base em Cochran (1977), para o valor médio por unidade da população, a estimativa usada na amostragem estratificada é representada por \bar{y}_{st} (st do inglês stratified), onde

$$\bar{y}_{st} = \frac{\sum_{h=1}^L N_h * \bar{y}_h}{N} \quad (16)$$

Na qual $N = N_1 + N_2 + \dots + N_L$.

A estimativa \bar{y}_{st} , de modo geral, não é o mesmo que valor médio amostral. O valor médio amostral, \bar{y} , é dado pela fórmula

$$\bar{y} = \frac{\sum_{h=1}^L n_h * \bar{y}_h}{n} \quad (17)$$

A diferença é que em \bar{y}_{st} , as estimativas individuais recebem seu peso correcto, $\frac{N_h}{N}$. \bar{y} coincide com \bar{y}_{st} , sob a condição de que em todos os estratos se tenha

$$\frac{n_h}{n} = \frac{N_h}{N} \text{ ou } \frac{n_h}{N_h} = \frac{n}{N} \text{ ou } f_h = f \quad (18)$$

Isso significa que a fracção amostral é a mesma em todos os estratos. Essa estratificação é denominada *estratificação com repartição proporcional* dos n_h e fornece uma amostra *autoponderada*. Quando tem que fazer-se numerosas estimativas, uma amostra autoponderada poupa tempo.

Ainda segundo Cochran (1977), as principais propriedades da estimativas \bar{y}_{st} estão indicadas nos teoremas seguintes:

Teorema 1: se, em todos os estratos, a estimativa amostral \bar{y}_h for sem tendência, então \bar{y}_{st} é uma estimativa sem tendência do valor médio da população \bar{Y} .

Demonstração:

$$E(\bar{y}_{st}) = E \frac{\sum_{h=1}^L N_h * \bar{y}_h}{N} = \frac{\sum_{h=1}^L N_h * \bar{Y}_h}{N} \quad (19)$$

Uma vez que as estimativas são sem tendências nos estratos individuais, o valor médio da população \bar{Y} pode ser dado pela fórmula

$$\bar{Y} = \frac{\sum_{h=1}^L \sum_{i=1}^{N_h} y_{hi}}{N} = \frac{\sum_{h=1}^L N_h * \bar{Y}_h}{N} \quad (20)$$

Teorema 2: para a amostragem estratificada, a variância de \bar{y}_{st} , sendo este uma estimativa do valor médio da população \bar{Y} , é dada pela fórmula

$$V(\bar{y}_{st}) = \frac{\sum_{h=1}^L N_h^2 * V(\bar{y}_h)}{N^2} = \sum_{h=1}^L W_h^2 * V(\bar{y}_h) \quad (21)$$

onde

$$V(\bar{y}_h) = E(\bar{y}_h - \bar{Y}_h)^2 \quad (22)$$

Há duas restrições ao teorema:

- a) \bar{y}_h deve ser uma estimativa sem tendência de \bar{Y}_h ;
- b) As amostras devem ser seleccionadas independentemente nos diferentes estratos.

Teorema 3: para a amostra aleatória estratificada, a variância da estimativa \bar{y}_{st} é

$$V(\bar{y}_{st}) = \frac{1}{N^2} * \sum_{h=1}^L N_h * (N_h - n_h) * \frac{S_h^2}{n_h} = \sum_{h=1}^L W_h^2 * \frac{S_h^2}{n_h} * (1 - f_h) \quad (23)$$

Demonstração: uma vez que \bar{y}_h é uma estimativa sem tendência de \bar{Y}_h , pode-se aplicar o teorema 2.

$$V(\bar{y}_h) = \frac{S_h^2}{n_h} * \frac{N_h - n_h}{N_h} \quad (24)$$

Substituindo, na expressão do teorema 2, $V(\bar{y}_{st})$ pelo valor acima, obtém-se:

$$\begin{aligned} V(\bar{y}_{st}) &= \frac{1}{N^2} * \sum_{h=1}^L N_h^2 * V(\bar{y}_h) = \frac{1}{N^2} * \sum_{h=1}^L N_h * (N_h - n_h) * \frac{S_h^2}{n_h} \\ &= \sum_{h=1}^L W_h^2 * \frac{S_h^2}{n_h} * (1 - f_h) \end{aligned} \quad (25)$$

Alguns casos particulares da fórmula acima:

1. Se as fracções da amostragem n_h / N_h forem desprezíveis em todos os estratos

$$V(\bar{y}_{st}) = \frac{1}{N^2} * \sum_{h=1}^L \frac{N_h^2 * S_h^2}{n_h} = \sum_{h=1}^L \frac{W_h^2 * S_h^2}{n_h} \quad (26)$$

Esta fórmula é apropriada quando se pode desprezar as correções das população finitas.

2. No caso em que a repartição é proporcional, pode-se substituir n_h por seu valor $\frac{n * N_h}{N}$, desse

modo reduzindo a variância para

$$V(\bar{y}_{st}) = \sum \frac{N_h}{N} * \frac{S_h}{n} \left(\frac{N-n}{N} \right) = \frac{1-f}{n} * \sum W_h S_h^2 \quad (27)$$

3. Se a amostragem for proporcional e as variâncias de todos os estratos tiverem o mesmo valor

S_w^2 , obtém-se a fórmula simplificada

$$V(\bar{y}_{st}) = \frac{S_w^2}{n} * \left(\frac{N-n}{N} \right) \quad (28)$$

Teorema 4: se $\hat{Y}_{st} = N * \bar{y}_{st}$ é a estimativa do valor total da população Y , então tem-se

$$V(\hat{Y}_{st}) = \sum N_h * (N_h - n_h) * \frac{S_h^2}{n_h} \quad (29)$$

Repartição Óptima

Segundo Cochran (1977), na amostragem estratificada, os valores das grandezas amostrais, n_h , nos estratos são escolhidos pelo seleccionador da amostra. A escolha pode ter em vista tornar mínimo o valor de $V(\bar{y}_{st})$ dentro de determinado limite de custo para a selecção da amostra, ou tornar mínimo o custo para um valor específico de $V(\bar{y}_{st})$.

A função-custo tem a seguinte forma:

$$custo = C = c_0 + \sum c_h * n_h \quad (30)$$

Dentro de um estrato qualquer, o custo é proporcional á grandeza da amostra, mas o custo por unidade, c_h , pode variar de um estrato para o outro. O termo c_0 , representa as despesas gerais. A função do custo acima é adequado a principal parcela do custo é a que corresponde a realização da medida de cada unidade.

Teorema 5: com a função-custo, a variância do valor médio \bar{y}_{st} , é mínima quando n_h é proporcional

a $\frac{N_h * S_h}{\sqrt{c_h}}$.

Demonstração:

$$V(\bar{y}_{st}) = \sum_{h=1}^L \frac{W_h^2 * S_h^2}{n_h} * (1 - f_h) = \sum_{h=1}^L \frac{W_h^2 * S_h^2}{n_h} - \sum_{h=1}^L \frac{W_h^2 * S_h^2}{N_h} \quad (31)$$

Respeitando a restrição de que

$$c_1 n_1 + c_2 n_2 + \dots + c_L n_L = C - c_0$$

Usando o processo de cálculo dos coeficientes de Lagrange, escolhe-se o n_h e o coeficiente λ , de modo a tornar mínimo o valor de: $V(\bar{y}_{st}) + \lambda * (\sum c_h * n_h - C + c_0)$, que fazendo algumas transformações resulta na equação

$$\frac{n_h}{n} = \frac{W_h * S_h / \sqrt{c_h}}{\sum (W_h * S_h / \sqrt{c_h})} = \frac{N_h * S_h / \sqrt{c_h}}{\sum (N_h * S_h / \sqrt{c_h})} \quad (32)$$

Este teorema conduz a seguinte regra de conduta: em um determinado estrato, seleccione uma amostra maior se:

- O estrato é maior;
- O estrato tem maior variação interna;
- A amostragem é mais barata no estrato.

Para completar a repartição é necessário uma outra providência. A equação acima dá n_h em função de n , mas ainda não sabe-se o valor que n tem. A solução depende de saber se a amostra é seleccionada para atender a um determinado custo total, C , ou a uma determinada variância, V , para \bar{y}_{st} . Se o custo é fixado, substitui-se, na expressão da função-custo, n_h por seu valor óptimo e resolve-se a equação para n . Isso dá

$$n = \frac{(C - c_0) * \sum (N_h * S_h / \sqrt{c_h})}{\sum (N_h * S_h * \sqrt{c_h})} \quad (33)$$

Se V é fixado, substitui-se o valor óptimo de n_h na fórmula que dá o valor de $V(\bar{y}_{st})$. Então

$$n = \frac{\left(\sum W_h * S_h * \sqrt{c_h}\right) * \sum W_h * S_h / \sqrt{c_h}}{V + (1/N) * \sum W_h * S_h^2} \quad (34)$$

Onde $W_h = \frac{N_h}{N}$.

Um caso especial importante surge quando $c_h = c$, isto é, quando o custo por unidade é o mesmo para todos os estratos. O custo total se torna $C = c_0 + c * n$, e a repartição óptima para um custo fixado, se reduz á repartição óptima para uma grandeza amostral fixada. O resultado nesse caso especial é o seguinte:

- Na amostragem aleatória estratificada, $V(\bar{y}_{st})$ é mínima, para uma grandeza total da amostra, n , se

$$n_h = n * \frac{W_h * S_h}{\sum W_h * S_h} = n * \frac{N_h * S_h}{\sum N_h * S_h} \quad (35)$$

Essa repartição é as vezes chamada *repartição de Neyman*.

A fórmula da variância mínima, quando fixado o valor de n , é dada pela substituição do valor de n_h dado na expressão acima, na fórmula geral de $V(\bar{y}_{st})$. O resultado é

$$V_{\min}(\bar{y}_{st}) = \frac{\left(\sum W_h * S_h\right)^2}{n} - \frac{\sum W_h * S_h^2}{N} \quad (36)$$

O segundo termo do membro da direita da expressão representa a função cumulativa de probabilidade (cpf).

Estimativa da Grandeza da Amostra com Dados Contínuos

De acordo com Cochran (1977), as fórmulas para a determinação de n para uma repartição qualquer com alguns casos particulares úteis, presume-se que a estimativa tem uma variância fixada, V .

Estimativa do Valor Médio da População, \bar{Y}

Seja s_h a estimativa de S_h , e $n_h = w_h * n$, onde w_h é escolhido.

A variância prevista $V(\bar{y}_{st})$ é dada por

$$V = \frac{1}{n} * \sum \frac{W_h^2 * s_h^2}{w_h} - \frac{1}{N} * \sum W_h s_h^2 \quad (37)$$

Sendo $W_h = N_h / N$. Isso dá como fórmula geral de n

$$n = \frac{\sum \frac{W_h^2 * s_h^2}{w_h}}{V + \frac{1}{N} * \sum W_h * s_h^2} \quad (38)$$

Se as cpf forem desprezadas, tem-se como uma primeira expressão

$$n_0 = \frac{1}{V} * \sum \frac{W_h^2 * s_h^2}{w_h} \quad (39)$$

Se n_0 / N não for desprezível, pde-se calcular n pela fórmula

$$n = \frac{n_0}{1 + \frac{1}{NV} * \sum W_h * s_h^2} \quad (40)$$

Ainda de acordo com Cochran (1977), nos casos particulares, as fórmulas tomam várias formas, que podem ser mais cômodas para os cálculos. Algumas delas são:

Repartição óptima presumível (sendo n fixado): w_h equivalente a $W_h * s_h$.

$$n = \frac{(\sum W_h * s_h)^2}{V + \frac{1}{N} \sum W_h * s_h^2} \quad (41)$$

Repartição proporcional: $w_h = W_h = N_h / N$.

$$n_0 = \frac{\sum W_h * s_h^2}{V} \quad (42)$$

$$n = \frac{n_0}{1 + \frac{n_0}{N}} \quad (43)$$

Estimativa do Valor Total da População

Segundo Cochran (1977), se V for a desejada $V(\bar{Y}_{st})$, as principais fórmulas são as seguintes:

Geral:

$$n = \frac{\sum \frac{N_h^2 * s_h^2}{w_h}}{V + \sum N_h * s_h^2} \quad (44)$$

Repartição ótima presumível (sendo n fixado):

$$n = \frac{(\sum N_h * s_h)^2}{V + \sum N_h * s_h^2} \quad (45)$$

Repartição proporcional:

$$n_0 = \frac{N}{V} * \sum N_h * s_h^2 \quad (46)$$

$$n = \frac{n_0}{1 + \frac{n_0}{N}} \quad (47)$$

Quando a Estratificação Produz Grande Aumento de Precisão?

Com base em Cochran (1977), a variável ideal para a estratificação é o próprio valor \mathcal{Y} , a quantidade a ser medida no levantamento. Se podesse-se estratificar de acordo com os valores de \mathcal{Y} , não haveria superposição entre os estratos, e a variância dentro dos estratos seria muito menor que a variância do conjunto, principalmente se houvesse muitos estratos.

Na prática, não se pode estratificar de acordo com os valores de \mathcal{Y} . Entretanto, algumas aplicações importantes se aproximam bastante dessa situação e portanto permitem grande aumento de precisão, mediante a situação das três condições seguintes:

- A população seja constituída de instituições cujos tamanhos variem amplamente;
- As principais variáveis a serem medidas se relacionem, estritamente, com o tamanho das instituições;
- Disponha-se de uma boa medida dos tamanhos para o estabelecimento dos estratos.

2.5.3.3. Amostragem Sistemática (AS)

Vários métodos têm sido propostos para determinar a melhor aproximação do erro de amostragem de uma amostra sistemática. Uma amostra sistemática constituída de unidades equidistantes entre si e bem sorteada pode ser considerada como uma amostra aleatória simples sem reposição, ou estratificada, e o erro de amostragem calculado com uma amostra aleatória. Entretanto, o erro calculado desse modo estima o erro máximo provável, o qual pode superestimar o erro real. Um procedimento simples e útil para obter a maior aproximação do erro verdadeiro é o método das diferenças sucessivas (Bolfarine, 2005).

Segundo Bolfarine (2005), a vantagem da amostragem sistemática é a facilidade de sua execução. Também, é menos sujeita a erros do entrevistador que os outros esquemas de amostragem. Quanto à sua precisão, existem situações em que ela é mais precisa que a AAS. Mas, na maioria dos casos, a sua eficiência é próxima da AAS, principalmente quando o sistema de referência está numa ordem aleatória. Em outros casos, quando existem tendências do tipo linear ou existem periodicidade na população, sua precisão pode ser bem diferente do planeamento AAS. A AS pode ser muito prejudicada por ciclos presentes na população.

2.5.3.4. Amostragem Por Conglomerados ou Por Clusters (AC)

Com base em Bolfarine (2005), os planos amostrais simples e por estratificação sorteiam unidades elementares directamente da população ou de estratos dessa mesma população. Quando os sistemas de referência não são adequados e o custo de actualizá-los é muito elevado ou quando a movimentação para identificar as unidades elementares no campo é cara e consome muito tempo, a tarefa amostral pode ser facilitada se forem seleccionados grupos de unidades elementares, os chamados *conglomerados*. O que caracteriza bem o planeamento amostral de conglomerados é que a unidade

amostral contém mais de um elemento populacional. Nos conglomerados deve existir heterogeneidade dentro dos conglomerados para garantir a eficiência.

Uma das inconveniências para o uso da amostragem por conglomerados prende-se ao facto de que as unidades, dentro de um mesmo conglomerado, tendem a ter valores parecidos em relação às variáveis que se investigam, e isso torna estes planos menos eficientes. Comparando-se amostragem de elementos com a de conglomerados de mesmo tamanho, a de conglomerados tende a: ter custo por elemento menor, ter menor variância e maiores problemas para análises estatísticas (Bolfarine, 2005).

De acordo com Vieira (2008) os passos para obtenção de uma amostra por conglomerado são:

- Especificar os conglomerados ou clusters (população é dividida em conglomerados);
- Seleccionar uma amostra de conglomerados aleatoriamente usando a AAS;
- São observados todos os elementos que pertencem aos conglomerados seleccionados;
- Ou é realizado um segundo estágio (ou são realizados mais estágios) até que no último estágio todos os elementos são observados;
- Deve haver uma lista identificando grupos de elementos (conglomerados) da população.

Notação (para população):

A : número de conglomerados em que a população foi dividida

N_α : tamanho do conglomerado α ($\alpha = 1, 2, \dots, A$)

$\sum_{\alpha=1}^A N_\alpha = N$: tamanho da população

$\bar{N} = \frac{\sum_{\alpha=1}^A N_\alpha}{A} = \frac{N}{A}$: tamanho médio dos conglomerados

Notação (para amostra):

a : número de conglomerados sorteados

$n_\alpha = N_\alpha$: tamanho do conglomerado α ($\alpha = 1, 2, \dots, a$)

$\sum_{\alpha=1}^a n_\alpha = n$: tamanho da amostra

$\bar{n} = \frac{\sum_{\alpha=1}^a n_\alpha}{a} = \frac{n}{a}$: tamanho médio dos conglomerados sorteado

Tabela 1: Parâmetros para cálculo de conglomerados

no conglomerado α	Globais
$\mu_\alpha = \frac{\sum_{i=1}^{N_\alpha} Y_{\alpha i}}{N_\alpha}$	$\mu = \frac{\sum_{\alpha=1}^A \sum_{i=1}^{N_\alpha} Y_{\alpha i}}{N}$
$\tau_\alpha = \sum_{i=1}^{N_\alpha} Y_{\alpha i}$	$\tau = \sum_{\alpha=1}^A \sum_{i=1}^{N_\alpha} Y_{\alpha i}$
$\sigma_\alpha^2 = \frac{\sum_{i=1}^{N_\alpha} (Y_{\alpha i} - \mu_\alpha)^2}{N_\alpha}$	$\sigma^2 = \frac{\sum_{\alpha=1}^A \sum_{i=1}^{N_\alpha} (Y_{\alpha i} - \mu)^2}{N}$
$S_\alpha^2 = \frac{\sum_{i=1}^{N_\alpha} (Y_{\alpha i} - \mu_\alpha)^2}{N_\alpha - 1}$	$S^2 = \frac{\sum_{\alpha=1}^A \sum_{i=1}^{N_\alpha} (Y_{\alpha i} - \mu)^2}{N - 1}$
$(\alpha = 1, 2, \dots, A)$	

Fonte: Cavalcante (2009)

Tabela 2: Estatísticas para cálculo de conglomerados

para o conglomerado sorteado	para toda amostra
$\bar{y}_\alpha = \frac{\sum_{i=1}^{n_\alpha} y_{\alpha i}}{n_\alpha}$	$\bar{\bar{y}} = \frac{\sum_{\alpha=1}^a \bar{y}_\alpha}{a} : \text{média das médias dos conglomerados}$
$T_\alpha = \sum_{i=1}^{n_\alpha} y_{\alpha i}$	$\bar{T} = \frac{\sum_{\alpha=1}^a T_\alpha}{a} : \text{total médio dos conglomerados}$
$(\alpha = 1, 2, \dots, a)$	Sorteados.
Para $\alpha = 1, 2, \dots, a$	Correspondem a $\bar{\mu}$ e $\bar{\tau}$, respectivamente.
$\bar{y}_\alpha = \mu_\alpha$	Se conglomerados de mesmo tamanho, $\bar{\bar{y}}$ é
$T_\alpha = \tau_\alpha$	um estimador não enviesado de μ .

Fonte: Cavalcante (2009)

A amostragem em conglomerados é vista como uma variação da amostragem em dois estágios, onde o segundo estágio é sistematicamente organizado dentro do primeiro estágio de amostragem.

Estimativas

Notações:

N = número total potencial de conglomerados na população;

M = número de subunidades do conglomerado;

n = número de conglomerados amostrados;

X_{ij} = variável de interesse.

Média da população por subunidade

$$\bar{X} = \frac{\sum_{i=1}^n \sum_{j=1}^M X_{ij}}{nM}$$

Média das subunidades por conglomerado

$$\bar{X}_i = \frac{\sum_{j=1}^M X_{ij}}{M}$$

Variância da população por subunidade

$$s_x^2 = \frac{1}{nM - 1} \sum_{i=1}^n \sum_{j=1}^M (x_{ij} - \bar{X})^2$$

2.5.3.5. Amostragem Multi-Etápica

Segundo Battisti (2008), quando a amostra por conglomerado compreende mais de duas etapas para a selecção da amostra final, tem-se um tipo de dimensão chamada *amostra de etapas múltiplas ou multi-etápica*. Na amostragem Multi-etápica, apenas as unidades finais são estudadas.

Com base em Battisti (2008), os passos para obtenção de uma amostra na amostragem multi-etapas são:

- Definir os conglomerados, tendo em conta duas condições: a proximidade geográfica dos elementos dentro do conglomerado e a dimensão dos conglomerados;
- Seleccionar através de amostragem aleatória simples, alguns dos conglomerados definidos;
- Seleccionar uma amostra, dentro de cada conglomerado, sempre utilizando um processo aleatório e, em fases sucessivas, até alcançar uma amostra de unidades elementares.

De acordo com Battisti (2008), em alguns planos amostrais, algumas unidades da população são mais importantes, por terem uma contribuição maior no valor do parâmetro, neste caso estabelece-se probabilidades desiguais de selecção às diferentes unidades da população.

Nos casos em que a probabilidades de selecção é proporcional à uma medida de tamanho da população, o procedimento amostral é definido como amostragem com probabilidade proporcional ao tamanho (PPT). A vantagem em seleccionar a unidade de amostragem com PPT é obter uma amostra mais representativa da população e assim aumentar a precisão dos estimadores quando comparados a AAS.

2.6. Ponderações de Amostragem

a) Ponderação de Base

Segundo Corlett (2000), para que as estimativas da amostra reflitam o universo, é preciso aplicar factores de ponderação ou de extrapolação. Para o efeito dos programas de tabulação, cada registo de elemento da amostra deve ter incorporado um ponderador, para ser multiplicado aos valores amostrais durante o processo de tabulação.

O ponderador de base para o i -ésimo elemento da amostra é igual ao inverso da sua probabilidade de selecção para cada etapa de amostragem.

Segundo Reis (1998), na análise dos dados, obtidos em planos amostrais complexos, é necessário o uso de técnicas desenvolvidas especificamente para esse fim, ou seja, a ponderação pelo efeito de estratificação e de conglomeração (ou seja, efeito de desenho). Na amostragem por conglomerados, a fim de se compensar as probabilidades (f) desiguais de selecção das unidades em cada um dos estágios (f_1, f_2, \dots, f_i) devem-se atribuir ponderações diferenciadas aos elementos da amostra, correspondentes ao inverso do produto das probabilidades de inclusão nos diversos estágios de selecção, estimando, então, novos pesos (w) para cada elemento da amostra:

$$w = 1/f$$

b) Ajustamento aos Factores de Ponderação

Ao final, cada inquérito terá de ajustar as ponderações para se tomar em conta a não-resposta em cada AE. A ponderação final (w'_{hij}) para os elementos amostrais na j -ésima AE amostral dentro da i -ésima UPA amostral no estrato h pode-se expressar da seguinte maneira:

$$w'_{hij} = w * \frac{k_{e,hij}}{k_{r,hij}}$$

Onde:

$k_{e,hij}$ é o número total de efectivos na amostra dentro da j -ésima AE da amostra;

$k_{r,hij}$ é número de efectivos na AE que fornecem resposta, quer dizer, para os quais obtive-se entrevistas.

O ajustamento de não -resposta supõe que os elementos que respondem e os que não respondem, em média, não diferem de forma significativa em relação as características sócio-económicas de interesse e que a taxa de não -resposta dentro da AE é baixa.

2.7. Critérios para Formação das Upas

Os critérios segundo a sua ordem de importancia, são os seguintes:

- Primeiro e sobretudo, limites identificáveis e duráveis;
- Segundo, tamanho alvo em termos do número de elementos; e
- Terceiro, tamanho limitado em termo de superfície.

Segundo Corlett (2000), as UPAs são urbano ou rural na sua totalidade. Não se pode juntar AEs urbanas com rurais.

2.8. Erros de Amostragem

O uso de amostras requer alguns cuidados, a fim de evitar erros. Em todos estudos pode ocorrer erros amostrais e não amostrais. Os erros não-amostrais aparecem em qualquer etapa do levantamento amostral, e se não forem identificados e avaliadas as possíveis distorções introduzidas, podem comprometer seriamente um plano amostral tecnicamente perfeito (Lutz, 2007).

Segundo Lutz (2007), erros amostrais são erros que ocorrem quando existe uma diferença entre o valor obtido na amostra e o parâmetro de interesse na população. E erros não - amostrais (sistemáticos) ocorrem quando os dados amostrais são recolhidos, registrados ou analisados incorrectamente. Em inquéritos por amostragem, pressupõe que estes erros são nulos dado a melhor formação ou treinamento aos inquiridores, inquiridores seleccionados entre os melhores, melhor supervisão, melhores métodos de recolha de dados. Contudo, o erro de amostragem sempre existem e deve ser avaliado.

De acordo com Cochran (1977), os erros não amostrais classificam-se em: *Efeito do erro de não resposta ou falta de resposta.*

Segundo Cochran (1977), a expressão não resposta usa-se para designar a impossibilidade de se medirem algumas das unidades da amostra seleccionada. No estudo de não resposta, é conveniente que se pense na população, considerando-a dividida em dois estratos, o primeiro constituído por todas as unidades que se conseguiriam medir, se acontecesse elas seriam incluídas na amostra, e o segundo composto pelas unidades que não se conseguiriam medir.

Um levantamento em que, quando necessário, se fazem pelo menos três visitas a cada casa, e em que um supervisor com excepcionais poderes de persuadir visita todas as pessoas que se recusam a fornecer dados, terá um estrato sem resposta muito menor do que um outro, no qual só se faça uma única tentativa em cada casa.

Ainda de acordo com Cochran (1977), os tipos de não resposta são:

Falta de cobertura – impossibilidade da localização ou da visita a algumas unidades da amostra. Decorre do uso de relações incompletas e algumas vezes as condições meteorológicas e as deficiências dos meios de transportes tornam impossível atingir-se certas unidades, durante o período de levantamento.

Não encontrados – este grupo contém as pessoas que residem no endereço mas se encontram temporariamente fora de casa. As famílias em que ambos os cônjuges trabalham fora, ou que não têm crianças, são mais difíceis de se atingir que as famílias que têm crianças de pouca idade ou pessoas idosas confinadas em casa.

Não sabem responder – o respondente pode não possuir a informação desejada em certas perguntas, ou talvez não queira fornecê-la. A redacção habilidosa do questionário e a sua verificação são garantias, neste caso.

A turma dura - é constituído pelas pessoas que se recusam, inflexivelmente, a ser entrevistadas, pelos fisicamente incapacitados e pelos que se mantêm fora de casa, durante todo o tempo disponível para o trabalho de campo. Representa uma fonte de tendência, que persiste seja qual for o esforço que se faça para completar as declarações.

2.8.1. Diferença entre os Tipos de Erros

Tabela 3: Diferença entre os tipos de erros

Erros amostrais	Erros não amostrais (ou Sistemáticos)
1. Ocorre apenas devido ao processo amostral e não a problemas de mensuração e obtenção das informações.	1. Ocorre devido a factores independentes do plano amostral e que ocorreriam mesmo se a população toda fosse investigada.
2. Podem ser controlados e medidos.	2. Não podem ser controlados nem medidos
3. Tendem a desaparecer com o crescimento do tamanho da amostra.	3. Podem alterar radicalmente os resultados e, conseqüentemente a interpretação de uma análise de estudo.

Fonte: Vieira (2008)

De acordo com Sousa (2011), as principais fontes de erros não-amostrais são:

- Definição errada do problema de estudo;
- Definição errada da população de estudo;
- Definição parcial da população de estudo;
- Não-resposta;
- Instrumentos de recolha de dados;
- Escalas;
- Entrevistadores;
- Entrevistados;
- Inferências causais impróprias;
- Processamento;
- Análises;
- Interpretação.

2.8.2. Formas de Evitar Erros:

De acordo com Alves (2006), as formas de evitar erros de amostragem são:

- Utilizar recursos humanos adequados às necessidades do estudo;
- Treinamento adequado dos investigadores;
- Elaborar manual de instruções (estudos complexas);
- Distribuir instruções ao longo do instrumento;
- Exercer contínuo controle da qualidade;
- Verificar por amostragem a veracidade e qualidade das entrevistas realizadas.

Nos inquéritos por amostragem, estes erros são minimizados através do tratamento correcto dos aspectos acima, sendo, em regra nulos.

2.9. Probabilidade de inclusão: Estimador do Tipo Horvitz-Thompson

Segundo Reis (1998), assume-se que, para toda unidade i na população, uma vizinhança A_i é definida como sendo um conjunto de unidades incluindo i . Essas vizinhanças não dependem dos valores y populacionais e, no caso do conglomerado adaptativo, a vizinhança de cada unidade consiste dos vizinhos geograficamente mais próximos. A relação de vizinhança é simétrica: se a unidade j está na vizinhança da unidade i , a unidade i está na vizinhança da unidade j . Seleccionada a amostra inicial, se uma unidade seleccionada satisfaz a condição $C=\{y : y \geq c\}$, todas as unidades dentro da sua vizinhança são adicionadas à amostra e observadas. O valor de C é positivo, arbitrário e fixo e, foi definido igual a 1. Para as unidades vizinhas que satisfizerem a condição, as unidades em suas vizinhanças também serão incluídas na amostra e assim por diante.

Define-se um conglomerado como o conjunto de todas as unidades que são observadas devido à selecção inicial da unidade i . Tal conjunto pode consistir da união de várias vizinhanças. Dentro de cada conglomerado, está um subconjunto de unidades, chamada rede, com a propriedade de que a selecção de uma unidade dentro da rede levará à inclusão na amostra de todas as unidades dessa rede.

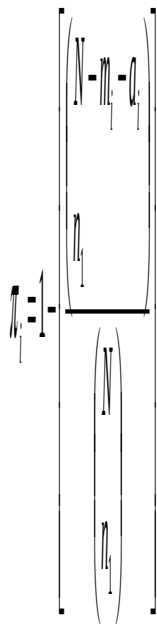
Uma unidade que não satisfaz a condição mas está na vizinhança de uma unidade que a satisfaz, é chamada de unidade de fronteira. Enquanto a selecção de uma unidade na rede resulta na inclusão de todas as unidades dessa rede e de todas as unidades de fronteira, a selecção de uma unidade de fronteira não resulta na inclusão de outras unidades. É conveniente considerar uma unidade que não satisfaz a condição como uma rede de tamanho um, tal que, dados os valores y , a população pode ser particionada unicamente em redes (Reis, 1998).

No estudo, trabalhar-se com amostragem sem reposição. Assim, quando a amostra inicial de n unidades for seleccionada, essas n unidades serão distintas. A unidade i será incluída na amostra, se uma unidade da rede à qual ela pertence for seleccionada na amostra inicial, ou se uma unidade de uma rede da qual i é uma unidade de fronteira for seleccionada. Seja m_i o número de unidades na rede da qual i faz parte e a_i , o número total de unidades nas redes das quais i é uma unidade de fronteira (Reis, 1998).

Note-se que, se a unidade i satisfaz a condição, tem-se $a_i = 0$, enquanto que, se a unidade i não satisfaz a condição, $m_i = 1$. A probabilidade de selecção da unidade i em uma das n retiradas é

$$p_i = \frac{(m_i + a_i)}{n} \quad \text{para } i=1,2,\dots,n \quad (48)$$

A probabilidade de que a unidade i seja incluída na amostra é:



$$(49)$$

O tamanho da amostra final, V , é uma variável aleatória, que pode ser pensada como a soma de N variáveis aleatórias Bernoulli, cada uma com média p_i , $i = 1, 2, \dots, N$. Assim, no caso dos desenhos de conglomerado adaptativo, o valor esperado de V é dado por

$$E(V) = \sum_{i=1}^N \pi_i$$

Ainda segundo Reis (1998), nos desenhos amostrais de conglomerado adaptativo, as probabilidades de selecção não são conhecidas para cada unidade na amostra. Um estimador não viciado, sugerido por Thompson (1990), é uma modificação do estimador de Hansen-Hurwitz, usando as unidades que não satisfazem a condição somente quando elas forem seleccionadas na amostra inicial. O estimador modificado é baseado nas probabilidades de que uma unidade da rede seja interceptada pela amostra

inicial. Seja y_i a rede que inclui a unidade i e m_i o número de unidades na rede. Seja w_i a média das observações na rede que inclui a i -ésima unidade da amostra inicial, tal que

$$w_i = \frac{1}{m_i} \sum_{j \in \psi} y_{ij} \quad (50)$$

O estimador modificado para amostragem sem reposição é

$$\hat{\mu}_1 = \frac{1}{n} \sum_{i=1}^n w_i \quad (51)$$

A variância é dada por

$$\text{Var}(\hat{\mu}_1) = \frac{N-n}{Nn(N-1)} \sum_{i=1}^N (w_i - \hat{\mu}_1)^2 \quad (52)$$

Um estimador não viciado para essa variância é

$$\hat{\text{Var}}(\hat{\mu}_1) = \frac{N-n}{Nn(N-1)} \sum_{i=1}^N (w_i - \hat{\mu}_1)^2 \quad (2.11.7)$$

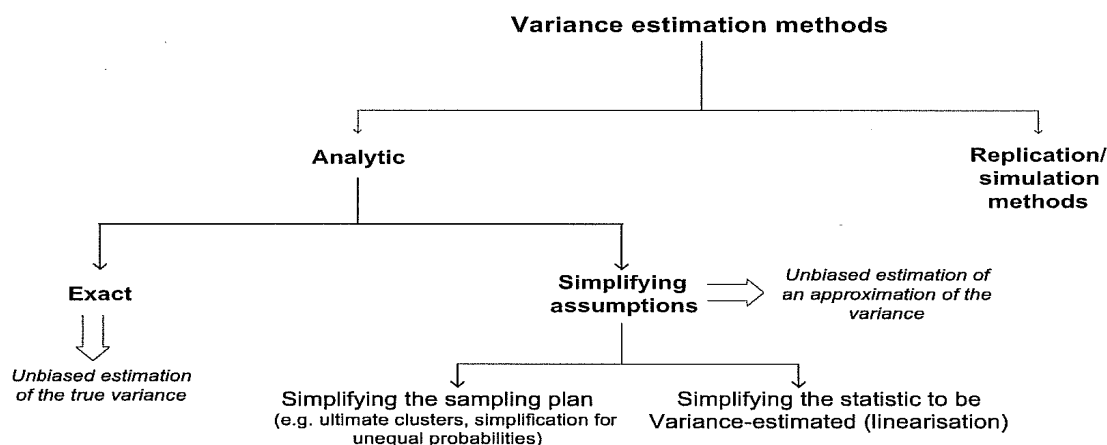
2.10. Principais Métodos de Estimação de Variâncias em Populações Finitas

De acordo com Rust (1985), o estimador de uma variância é afectado pela estrutura da variável em estudo na população, complexidade do plano amostral e a forma do estimador (linear: médias, totais e proporções e não lineares: rácios, quantis, etc). Segundo Rust (1985), geralmente, estimadores lineares são do tipo $\hat{\theta} = \sum_h a_h \hat{\theta}_h$ (isto é, combinação linear de estimadores), em que θ_h é um parâmetro do estrato h ($\theta_h \in \Theta$) e $\theta = \sum_h a_h \theta_h$. A constante $a_h \in R$ depende apenas da unidade amostral seleccionada e não dos valores assumidos pela variável de interesse.

Quando o plano de sondagem é complexo (ou desenvolvido em mais de uma etapa, utiliza probabilidades desiguais de selecção e várias etapas de ponderação), não existem fórmulas directas para o cálculo da variância dos estimadores. Mesmo para um plano de sondagem simples, a utilização de ponderação múltipla dos dados faz com que as fórmulas de cálculo das variâncias dos estimadores

(por exemplo, para o total) sejam complicadas. Nestes casos em que não há métodos directos (exactos) para calcular os erros padrão centrados dos estimadores pontuais, a alternativa é fazer uma aproximação dos mesmos (Rust 1985).

Na figura abaixo, ilustra-se os vários métodos alternativos para a estimação de variâncias numa população U de dimensão N .



Fonte: European Communities (2002)

2.11. Métodos de Re-amostragem ou Replicados

De acordo com Rennó (2011), a re-amostragem é o nome que se dá a um conjunto de técnicas ou métodos que se baseiam em calcular estimativas a partir de repetidas amostragens dentro da mesma amostra (única).

Segundo Lohr (1999), métodos de re-amostragem trata a amostra como se se tratasse de uma população em si; toma-se diferentes amostras a partir desta nova população e usa-se as sub-amostras para estimar a variância.

Estes métodos baseiam-se na ideia de que a amostra obtida é representativa da população alvo, podendo extrair-se novas e repetidas amostras a partir da amostra original, com o objectivo de estimar variâncias ou intervalos de confiança.

De acordo com Sarndal (1992), os métodos replicados podem ser utilizados para uma vária gama de planos de sondagem, incluindo os planos de sondagem multi-etápicas, estratificados e amostras seleccionadas com probabilidades desiguais. Os mesmos autores afirmam que os métodos replicados combinam várias técnicas de estimação, desde os ajustamentos devido as não respostas e a pós-estratificação. A sua grande desvantagem é a intensidade computacional para grandes amostras.

Sarndal (1992), refere ainda que, a forma mais simples para calcular $\hat{\theta}_a$ (estimações em cada amostra replicada) consiste em: a) criar ponderadores replicados e ligá-los a cada observação, procedimento igual se faz normalmente para ajustar os ponderadores da amostra global s , associando-os depois as observações. Cada estimativa $\hat{\theta}_a$ é calculada usando os ponderadores das réplicas, tal como os ponderadores da amostra global s são utilizados para calcular $\hat{\theta}$;

b) se ajustamentos dos ponderadores finais são realizados devido as não respostas, estes mesmos ajustamentos devem ser aplicados separadamente em cada replicado para a estimação $\hat{\theta}_a$.

Alguns métodos replicados de estimação de variâncias são o Jackknife, proposto por Quenouille (1949), o Bootstrap, introduzido por Efron (1979) e o método de Replicação Repetida Balanceda (BRR). Abordasse apenas os métodos Jackknife e Replicação Repetida Balanceda (BRR).

2.11.1. Método de Jackknife

Segundo Sarndal (1992), o método de jackknife originou fora do campo de investigação por amostragem. A primeira ideia, desenvolvido por Quenouille (1949, 1956), era usar jackknifing para reduzir a tendência de um estimador, numa população infinita. Tukey (1958), posteriormente, sugeriu que o método também pode ser usado para produzir estimativas da variância. Para população finita, o método de jackknife foi considerado pela primeira vez por Durbin (1959). Aqui, da-se uma revisão do método de jackknife, como é comumente utilizado para estimar a da variância na amostragem de um estudo.

Seja s uma amostra de n elementos (o total da amostra), obtidos a partir de amostragem aleatória. Seja θ o parâmetro da população a ser estimado por $\hat{\theta}$, que é um estimador baseado em dados do total amostral s . O objectivo é o de estimar $V(\hat{\theta})$.

O método de jackknife começa com o particionamento da amostra s em A grupos aleatórios dependentes de igual tamanho $m (= n / A)$.

Assume-se que, para qualquer s , cada grupo é uma amostragem aleatória simples sem reposição (AASS) da amostra s , mesmo que s em si é uma amostra não AASS. Em seguida, para cada grupo ($a = 1, 2, \dots, A$), calcula-se $\hat{\theta}_a$, um estimador de θ da mesma forma como função $\hat{\theta}$, mas com base apenas em dados que permanecem após omitindo o grupo. Para $a = 1, 2, \dots, A$, define-se

$$\hat{\theta}_a = A * \hat{\theta} - (A - 1) * \hat{\theta}_{(a)} \quad (53)$$

O estimador jackknife de θ (uma alternativa para a estimativa para os $\hat{\theta}$) é

$$\hat{\theta}_{ij} = \frac{1}{A} \sum_{a=1}^A \hat{\theta}_a \quad (54)$$

E o estimador jackknife da variância é definido como

$$\hat{V}_{ij1} = \frac{1}{A(A-1)} \sum_{a=1}^A (\hat{\theta}_a - \hat{\theta}_{ij})^2 \quad (55)$$

Na prática, \hat{V}_{ij1} é usado como estimador de $\hat{V}(\hat{\theta})$ bem como de $V(\hat{\theta}_{ij})$. Por vezes a alternativa usada para, \hat{V}_{ij1} é

$$\hat{V}_{ij2} = \frac{1}{A(A-1)} \sum_{a=1}^A (\hat{\theta}_a - \hat{\theta})^2 \quad (56)$$

Seja θ a população total, $\theta = t = \sum_U y_i$. suponha que a amostra s de tamanho fixo n , é obtida pela probabilidade proporcional ao tamanho (PPT) de AASS com probabilidades de inclusão $\pi_1, \pi_2, \dots, \pi_N$.

Seja $\hat{\theta}$ o estimador π com base na amostra total s , que é

$$\hat{\theta} = \hat{t}_\pi = \sum_s y_i / \pi_i$$

Seja s divididos em grupos aleatórios de tamanhos iguais A tal como descrito acima. Define-se

$$\hat{\theta}_{(a)} = \hat{t}_{\pi(a)} = \left[\frac{A}{(A-1)} \right] \sum_{s-s_a} y_i / \pi_i \quad (57)$$

Segue-se a partir de (48), (49) e (52) que

$$\hat{\theta}_a = \hat{t}_a = A\hat{t} - (A-1)\hat{t}_{\pi(a)} = A \sum_{s_a} y_i / \pi_i$$

E

$$\hat{\theta}_{ij} = \hat{t}_{ij} = \frac{1}{A} \sum_{a=1}^A \hat{t}_a = \hat{t}_\pi$$

Assim, os dois estimadores da variância (50) e (51) coincidem

$$\hat{V}_{ij1} = \hat{V}_{ij2} = \frac{1}{A(A-1)} \sum_{a=1}^A (\hat{t}_a - \hat{t}_\pi)^2 = \hat{V}_{ij} \quad (58)$$

Nota-se que (58) é idêntica aos grupos aleatórios da variância estimada para o mesmo desenho amostral e agrupamento aleatório mostrando que $\hat{V}_{ij1} = \hat{V}_{ij2} = \hat{V}_{DRG}$. Em outras palavras, quando o estimador π é envolvido, o método de jackknife é uma forma alternativa de calcular \hat{V}_{DRG} . Resulta que para π estimador

$$E(\hat{V}_{ij}) = E(\hat{V}_0) \quad (59)$$

Onde \hat{V}_0 é o estimador de variância simplificada. A seguinte expressão familiar é obtida para o viés de \hat{V}_{ij} .

$$E(\hat{V}_{ij}) - V(\hat{t}) = \frac{n}{n-1} [V_0 - V(\hat{t})] \quad (60)$$

No caso especial em que $A = n$ e $m = 1$ (comum em prática), é possível mostrar que o estimador jackknife da variância é igual a \hat{V}_0 não é só expectativa, mas de forma idêntica,

$$\hat{V}_{ij} = \hat{V}_0$$

Portanto, nesse caso, o jackknife é uma forma alternativa de computação \hat{V}_0 .

No caso de duas fase de amostragem (SI), os resultados do exemplo anterior são simplificadas como se segue

$$\hat{t}_\pi = N\bar{y}_s$$

$$\hat{t}_{\pi(a)} = \frac{N}{n-m} \sum_{s-s_a} y_i = N\bar{y}_{s-s_a}$$

$$\hat{t}_a = \frac{N}{m} \sum_{s_a} y_i = N\bar{y}_{s_a}$$

$$\hat{t}_{ij} = \hat{t}_\pi = N\bar{y}_s$$

$$\hat{V}_{ij} = \frac{N^2}{A(A-1)} \sum_{a=1}^A (\bar{y}_{s_a} - \bar{y}_s)^2$$

$$E(\hat{V}_{ij}) = E(N^2 S_{ys}^2 / n)$$

$$E(\hat{V}_{ij}) - V(N\bar{y}_s) = NS_{sU}^2$$

Segue-se que no caso da amostragem de SI, que pode remover os viés do estimador de jackknife simplesmente multiplicando por $1 - f = 1 - n / N$. Em particular, quando $A=n$ e $m=1$ se obtém

$$\hat{V}_{ij} = N^2 S_{ys}^2 / n$$

E

$$(1 - f) \hat{V}_{ij} = N^2 (1 - f) S_{ys}^2 / n$$

Que é o estimador de variância imparcial comum.

2.11.1.1. Método de Jackknife para Amostragem Estratificada

Ainda de acordo com Sarndal (1992), quando o método de jackknife é aplicado a uma amostra estratificada, se utiliza outros estimadores da variância diferentes do visto na fórmula (50) ou (51).

De acordo com Wolter (1985) citado por Sarndal (1992), deve-se ter um cuidado especial para não aplicar os estimadores clássicos de jackknife a problemas de amostragem estratificada.

Com base em Sarndal (1992), assumi-se que a amostra de estrato $h (h = 1, 2, \dots, H)$ é dividida ao acaso

em grupos A_h , para um total de $A = \sum_{h=1}^H A_h$ grupos. Como antes, seja $\hat{\theta}$ o estimador amostral de θ .

Seja $\hat{\theta}_{(ha)}$ o estimador de θ com base no que resta da amostra de estrato h após omitindo o grupo. Um estimador que tem sido sugerido para o $V(\hat{\theta})$ é

$$\hat{V}_{ij3} = \sum_{h=1}^H [(A_h - 1) / A_h] \sum_{a=1}^{A_h} [\hat{\theta}_{ha} - \hat{\theta}]^2 \quad (61)$$

O que seria imparcial se a seleção da amostra fosse com reposição e se $\hat{\theta}$ é o estimador do total da população $\theta = t = \sum_U y_i$ com π probabilidade.

Em particular, suponha que é utilizado a amostragem SI em cada estrato. Seja

$$\hat{\theta} = \sum_{h=1}^H N_h \bar{y}_{s_h} \quad (62)$$

$$\hat{\theta}_{(ha)} = N_1 \bar{y}_{s_1} + \dots + N_{h-1} \bar{y}_{s_{h-1}} + N_{h+1} \bar{y}_{s_{h+1}} + \dots + N_H \bar{y}_{s_H} \quad (63)$$

Em seguida obtem-se

$$\hat{V}_{ij3} = \sum_{h=1}^H \frac{N_h^2}{A_h(A_h-1)} \sum_{a=1}^{A_h} (\bar{y}_{s_{ha}} - \bar{y}_{s_h})^2 \quad (64)$$

Sob o presasuposto de que $A_h = n_h$, tem-se simplesmente

$$\hat{V}_{ij3} = \sum_{h=1}^H N_h^2 S_{ys_h}^2 / n_h \quad (65)$$

que é o estimador da variância tradicional, com a correção de população finita omitido.

2.11.1.2. Método de Jackknife para Amostragem Multietápica

De acordo com o pensamento de Sarndal (1992), na amostragem de múltiplos estágios, o método de jackknife é normalmente aplicado ao nível UPAs. Suponha que no primeiro estágio, s_1 amostra contendo n_1 UPAs é tirada do conjunto U_1 composto de N_1 UPAs. Seja s_1 dividida aleatoriamente em A grupos de UPAs, com m UPAs em cada grupo. Seja $\hat{\theta}$ o estimador de amostragem total de θ , e seja $\hat{\theta}_{(a)}$ o estimador a denotar de θ , que se baseia no restante grupo de dados após remoção da UPA. Usando as equações (53) à (56), obtem-se

$$\hat{\theta}_a = A\hat{\theta} - (A-1)\hat{\theta}_{(a)} \quad a = 1, 2, \dots, A \quad (66)$$

$$\hat{\theta}_{ij} = \frac{1}{A} \sum_{a=1}^A \hat{\theta}_a \quad (67)$$

E os estimadores para variância são

$$\hat{V}_{ij1} = \frac{1}{A(A-1)} \sum_{a=1}^A (\hat{\theta}_a - \hat{\theta}_{ij})^2 \quad (68)$$

$$\hat{V}_{ij2} = \frac{1}{A(A-1)} \sum_{a=1}^A (\hat{\theta}_a - \hat{\theta})^2 \quad (69)$$

$$\hat{\theta} = \hat{t} = \sum_{s_{ia}} \hat{t}_i / \pi_{li}$$

$$\hat{\theta}_{(a)} = \hat{t}_{(a)} = [A/(A-1)] \sum_{s_i - s_{ia}} \hat{t}_i / \pi_{li}$$

$$\hat{\theta}_a = \hat{t}_a = A \sum_{s_{ia}} \hat{t}_i / \pi_{li}$$

$$\hat{\theta}_{ij} = \hat{t}_{ij} = \hat{t}$$

$$\hat{V}_{ij1} = \hat{V}_{ij2} = \frac{1}{A(A-1)} \sum_{a=1}^A (\hat{t}_a - \hat{t})^2 = \hat{V}_{ij} \quad (70)$$

2.12.2. Replicação Repetida Balanceada (BRR)

De acordo com Lohr (1999), algumas investigações são estratificadas ao ponto que apenas duas UPAs são selecionadas a partir de cada estrato. O método de Replicação Repetida Balanceada (BRR) exige que o total de amostras a ser desenhada use um desenho de amostragem estratificada com duas unidades primárias de amostragem (UPAs) por estrato. Isto dá ao mais alto grau de estratificação possível, permitindo o cálculo da variância estimada em cada estrato.

2.12.2.1. BRR numa Amostra Aleatória Estratificada

Com base em Lohr (1999), ilustra-se o cálculo da variância para \bar{y}_{str} a partir de uma amostra aleatória estratificada usando BRR.

Suponha que uma amostra aleatória estratificada de duas unidades de observação é escolhido a partir de cada um dos sete estratos. Arbitrariamente rotula-se uma das unidades da amostra no estrato h como y_{h_1} , e o outro como y_{h_2} . Os valores da amostra são apresentados na tabela abaixo

Tabela 4: Uma amostra aleatória estratificada pequena, usada para ilustrar o BRR.

Estrato	$\frac{N_h}{N}$	y_{h_1}	y_{h_2}	\bar{y}_h	$y_{h_1} - y_{h_2}$
1	0.3	2000	1792	1896	208
2	0.1	4525	4735	4630	-210
3	0.05	9550	14060	11805	-4510
4	0.1	800	1250	1025	-450
5	0.2	9300	7264	8282	2036
6	0.05	13286	12840	13063	446
7	0.2	2106	2070	2088	36

Fonte :Lohr (1999)

A média da população é estimada em

$$\bar{y}_{str} = \sum_{h=1}^H \frac{N_h}{N} \bar{y}_h = 4451.7 \quad (71)$$

Ignorando as correções população finita tem-se o estimador de variância como

$$V_{str}(\bar{y}_{str}) = \sum_{h=1}^H \left(\frac{N_h}{N} \right)^2 \frac{s_h^2}{n_h} \quad (72)$$

Quando $n_h = 2$, como é o caso, $s_h = (y_{h_1} - y_{h_2})^2 / 2$, então

$$V_{str}(\bar{y}_{str}) = \sum_{h=1}^H \left(\frac{N_h}{N} \right)^2 \frac{(y_{h_1} - y_{h_2})^2}{4} \quad (73)$$

Logo, $V_{str}(\bar{y}_{str}) = 55892.75$. Isto pode superestimar a variância se a amostragem é sem substituição.

De acordo com Lohr (1999), para usar o método de grupo aleatório, escolhe-se aleatoriamente uma das observações em cada estrato para o grupo 1 e atribuir a outra para o grupo 2. Os grupos neste situação são meias-amostras. Por exemplo, o grupo 1 pode consistir de $\{y_{11}, y_{22}, y_{32}, y_{42}, y_{51}, y_{62}, y_{71}\}$ e o grupo 2 das outras sete observações. Então

$$\hat{\theta}_1 = (0.3)(2000) + (0.1)(4735) + \dots + (0.2)(2106) = 4824.7$$

$$\hat{\theta}_2 = (0.3)(1792) + (0.1)(4525) + \dots + (0.2)(2070) = 4078.7$$

BRR usa a variabilidade entre R replicar meias- amostras que são seleccionados de forma equilibrada para estimar a variância de $\hat{\theta}$.

Para definir o equilíbrio, vamos introduzir a seguinte notação. Metade da amostra r podem ser definido por um vector $\alpha_r = (\alpha_{r1}, \dots, \alpha_{rH})$: Logo

$$y_h(\alpha_r) = \begin{cases} y_{h_1} & \text{se } \alpha_{rh} = 1 \\ y_{h_2} & \text{se } \alpha_{rh} = -1 \end{cases} \quad (74)$$

Equivalente a

$$y_h(\alpha_r) = \frac{\alpha_{rh} + 1}{2} y_{h_1} - \frac{\alpha_{rh} - 1}{2} y_{h_2} \quad (75)$$

Grupo 1 contém observações $\{y_{11}, y_{22}, y_{32}, y_{42}, y_{51}, y_{62}, y_{71}\}$ como acima, então $\alpha_1 = (1, -1, -1, -1, 1, -1, 1)$.

Da mesma forma, $\alpha_2 = (-1, 1, 1, -1, 1, -1, -1)$. O conjunto de R replicar

meias- amostras é equilibrado se

$$\sum_{r=1}^R \alpha_{rh} \alpha_{rl} = 0 \quad \text{para todo } l \neq h$$

Por repetição r , calcula-se $\hat{\theta}(\alpha_r)$ da mesma forma que $\hat{\theta}$ mas usando apenas as observações no meio da amostra seleccionada por α_r . Para estimar a média de uma amostra aleatória estratificada, $\hat{\theta}(\alpha_r) = \sum_{h=1}^H (N_h / N) y_h(\alpha_r)$. Defini-se o estimador de variância de BRR como

$$\hat{V}_{BRR}(\hat{\theta}) = \frac{1}{R} \sum_{r=1}^R [\hat{\theta}(\alpha_r) - \hat{\theta}]^2 \quad (76)$$

onde R é o número total de repetições.

Se o conjunto de meias-amostras é equilibrado, para amostragem aleatória estratificada então $\hat{V}_{BRR}(\bar{y}_{str}) = \hat{V}_{str}(\bar{y}_{str})$.

Matriz de Hadamard

Segundo Lohr (1999), uma matriz de Hadamard \mathbf{A} de dimensão R é uma matriz quadrada que tem todos os elementos iguais a 1 ou -1. A matriz de Hadamard deve satisfazer a exigência de que $\mathbf{A}'\mathbf{A} = R\mathbf{I}$, Onde \mathbf{I} é uma matriz de identidade. A dimensão de uma matriz de Hadamard deve ser igual a 1, 2, ou um múltiplo de quatro.

Por exemplo, a sequência de matriz é uma matriz de Hadamard de dimensão $k = 8$:

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{bmatrix}$$

Nota que pode-se calcular a estimativa de variância BRR, criando uma nova variável de pesos para cada repetição meia-amostra. O peso de amostragem para a observação i no estrato h é $w_{hi} = N_h / n_h$, e

$$\bar{y}_{str} = \frac{\sum_{h=1}^H \sum_{i=1}^2 w_{hi} y_{hi}}{\sum_{h=1}^H \sum_{i=1}^2 w_{hi}} \quad (77)$$

Defini-se por

$$w_{hi}(\alpha_r) = \begin{cases} 2w_{hi} & \text{se a observação do estrato } h \text{ é o meio da amostra seleccionada por } \alpha_r \\ 0 & \text{outros casos} \end{cases}$$

Então

$$\bar{y}_{str}(\alpha_r) = \frac{\sum_{h=1}^H \sum_{i=1}^2 w_{hi}(\alpha_r) y_{hi}}{\sum_{h=1}^H \sum_{i=1}^2 w_{hi}(\alpha_r)} \quad (78)$$

Da mesma forma, para qualquer estatística $\hat{\theta}$ calculado usando os w_{hi} pesos, $\hat{\theta}(\alpha_r)$ é calculado exactamente da mesma maneira, mas usando os novos pesos $w_{hi}(\alpha_r)$.

2.12.2.2. BRR em Investigações de Múltiplos Estágios Estratificados

Segundo Lohr (1999), o valor da BRR em uma investigação complexa vem da sua capacidade de estimar a variância de uma quantidade θ da população em geral, onde θ pode ser uma razão de duas variáveis, um coeficiente de correlação, um quantil, ou outra quantidade de interesse.

Suponha que na população de estratos H são seleccionados duas UPAs a partir do estrato h com probabilidades desiguais e com substituição (em métodos de replicação, usa-se muitas vezes

amostragem com a substituição , porque o desenho da subamostragem não afecta o estimador de variância). O mesmo método pode ser usado quando a amostragem é feito sem

substituição em cada estrato , mas a variância estimada é esperado ser maior do que a variância sem substituição .

O vector α_r define o meio da amostra r da seguinte forma: Se $\alpha_{rh} = 1$, então todas as unidades de observação na UPA 1 do estrato h estão no meio da amostra r . Se $\alpha_{rh} = -1$, então todas as unidades de observação em UPA 2 do estrato h estão no meio de amostra r . Os vectores α_r são seleccionados de forma equilibrada exactamente como na amostragem aleatória estratificada. Agora , por meio de amostra r , cria-se uma nova coluna de pesos $w(\alpha_r)$:

$$w_i(\alpha_r) = \begin{cases} 2w_i & \text{se a unidade de observação } i \text{ é meio - amostra } r \\ 0 & \text{outros casos} \end{cases} \quad (79)$$

A estimativa do total da população de \mathcal{Y} para o total da amostra é $\sum_{i \in S} w_i y_i$; a estimativa do total da população de \mathcal{Y} por meio de amostra r é $\sum_{i \in S} w_i(\alpha_r) y_i$. Se $\theta = t_y / t_x$, então $\hat{\theta} = \sum_{i \in S} w_i y_i / \sum_{i \in S} w_i x_i$, e $\hat{\theta}(\alpha_r) = \sum_{i \in S} w_i(\alpha_r) y_i / \sum_{i \in S} w_i(\alpha_r) x_i$.

A função de distribuição empírica é calculado usando os pesos :

$$\hat{F}(y) = \frac{\text{soma de } w_i \text{ para todas as observações com } y_i \leq y}{\text{soma de } w_i(\alpha_r) \text{ para todas as observações}} \quad (80)$$

A distribuição empírica utilizando metade da amostra r é

$$\hat{F}_r(y) = \frac{\text{soma de } w_i(\alpha_r) \text{ para todas as observações com } y_i \leq y}{\text{soma de } w_i(\alpha_r) \text{ para todas as observações}} \quad (81)$$

Se θ é a mediana da população, então $\hat{\theta}$ pode ser definido como o menor valor de y para $\hat{F}(y) \geq 1/2$, e $\hat{\theta}(\alpha_r)$ é o menor valor de y para o qual $\hat{F}_r(y) \geq 1/2$.

Para qualquer quantidade θ , define-se

$$(9.7) \quad \hat{V}_{BRR}(\hat{\theta}) = \frac{1}{R} \sum_{r=1}^R [\hat{\theta}(\alpha_r) - \hat{\theta}]^2 \quad (82)$$

De acordo com Lohr (1999), o método de BRR também pode ser utilizado para estimar as covariâncias de estatísticas: Se θ e η são duas quantidades de interesse, então

$$\hat{Cov}_{BRR}(\hat{\theta}, \hat{\eta}) = \frac{1}{R} \sum_{r=1}^R [\hat{\theta}(\alpha_r) - \hat{\theta}] * [\hat{\eta}(\alpha_r) - \hat{\eta}] \quad (83)$$

Quando um método de replicação, como BRR é usado, para análise de dados pode-se calcular variações de arquivos de dados, sem a necessidade de saber a estratificação e agrupamento das informações.

III. MATERIAL E MÉTODOS

3.1. MATERIAL

Para o trabalho usaram-se dados secundários recolhidos no Inquérito Contínuo aos Agregados Familiares (INCAF trimestre I) da província de Nampula, fornecidos pelo Instituto Nacional de Estatística de Moçambique.

Para o processamento dos dados uso-se o pacote estatístico *Microsoft Excel* versão 2010, software estatístico *WESVAR* versão 5.1 e software estatístico *SPSS* versão 20.0, e para a digitação do texto usa-se o pacote *Microsoft Word* versão 2010.

Quanto à necessidade de aumento do tamanho da amostra, ao se utilizar conglomerados, torna-se indispensável a correção por um valor conhecido como *deff* (*design effect* = efeito de desenho), calculado pela razão entre a estimativa da variância determinada pelo plano amostral e a estimativa da variância obtida por uma amostra aleatória simples de mesmo tamanho. Além de ser utilizado para verificar a perda de precisão da estimativa, o efeito de desenho é utilizado também para o planeamento de estudos futuros, no cálculo do tamanho de amostra.

3.2. MÉTODOS

Para compreender o uso e aplicação do plano de sondagem estratificado e desenho da base de sondagem usaram-se métodos de amostragem probabilística tais como amostragem aleatória simples, amostragem estratificada, amostragem por conglomerados últimos e amostragem multi-etápica referidos no capítulo anterior.

Esses métodos de amostragem probabilística são executados através de técnicas impíricas para análises estatísticas conhecidas como técnicas ou métodos de re-amostragem, incorporadas nos softwares utilizados no trabalho. Os métodos de re-amostragem mais conhecidos são: o método de Bootstrap, o método de Jackknife, e o método de Replicação Repetida Balançada ou equilibrada (BRR).

3.2.1. Métodos para Validação dos Inquéritos

Os métodos de re-amostragem usados para obtenção dos resultados no trabalho são: o método de Jackknife, e o método de Replicação Repetida Balançada ou equilibrada (BRR).

O método Jackknife é baseado no princípio de “deixe um de fora” que consiste em separar uma observação da amostra original, estimar os coeficientes com base no restante da amostra ($n - 1$) e classifica a observação separada utilizando a nova equação. O procedimento é repetido para todas as amostras (n vezes) depois que as previsões de pertinência de cada grupo forem feitas, uma por vez, uma matriz de classificação é construída e a percentagem de classificação correcta é acumulada para todas as observações da amostra. O método Jackknife é muito sensível para amostras pequenas, orientações sugerem que ele seja usado quando o tamanho do grupo menor é pelo menos três vezes o número de variáveis independentes.

O método BRR é usado para estimar variâncias de quantis. O analista de dados precisa apenas das colunas de replicas de pesos, e não precisa a informação original do desenho amostral para calcular os desvios. Ela requer um número relativamente pequeno de cálculos (e relativamente poucas colunas de pesos réplicados) quando comparado com o jackknife e bootstrap. Exige que o total de amostras a ser desenhada use um desenho de amostragem estratificada com duas unidades primárias de amostragem (UPAs) por estrato.

Vantagens : BRR dá um estimador de variância que é assintoticamente equivalente a obtida a partir de métodos de linearização de funções suaves de totais populacionais.

Desvantagens: Tal como definido acima, BRR só pode ser usado em situações em que há duas UPAs por estrato. Na prática, porém, muitas vezes é estendida a outros projetos de amostragem usando esquemas de balanceamento mais complicados. BRR, como o jackknife e bootstrap, calcula a variância com substituição, e pode superestimar a variância sem substituição.

Na ponderação pelo efeito de desenho, cada elemento da amostra teve associado um peso (w) que é o inverso da sua probabilidade de inclusão.

$$w = \frac{1}{f}$$

Onde: w = peso e

f = probabilidade de inclusão

De acordo com Costa (2002), para comparar a precisão de diferentes variáveis foi proposta a utilização do coeficiente de variação, dado pela expressão (expresso em %):

$$CV = \frac{\sqrt{QME}}{\bar{x}} * 100$$

O coeficiente de variação (CV) é definido como a estimativa do erro experimental em percentagem da estimativa da média, é uma das medidas estatísticas mais utilizadas pelos investigadores na avaliação da precisão dos experimentos. Na prática, o QME é substituído por variância da estimativa para grandes amostras.

3.2.2. Regra de Decisão

Segundo Rust (1985), em cada teste que sera feito, se o coeficiente de variação for inferior a 10% considera-se o mesmo como baixo, ou seja, o experimento tem alta precisão, de 10% a 20% os CVs são considerados médios, implicando em boa precisão, de 20% a 30% são julgados altos, significando baixa precisão e acima de 30% são tidos como muito altos, indicando baixíssima precisão. Em resumo:

Baixo : $0\% \leq CV < 10\%$

Médio : $10\% \leq CV < 20\%$

Alto : $20\% \leq CV < 30\%$

Muito alto: $CV \geq 30\%$

Para efeitos de análise o nível de confiança estabelecido é de 5% de significância.

Foram seleccionadas algumas variáveis (sexo, idade e estado civil) que compoem a base de dados, para fazer a avaliação da qualidade e comparabilidade dos métodos. Na prática, utilizam-se estimativas cujos CV são inferiores a 15% e associados ao intervalo de confiança.

IV. RESULTADOS E DISCUSSÃO

4.1 Análise descritiva

As estimativas de re-amostragem ajustam a precisão dos resultados de cada variável que compõe a base de sondagem fazendo o controle de qualidade e validação do inquérito de modo a tirar conclusões fiáveis e fazer generalização.

Interpretação de estimativas de algumas variáveis obtidas usando amostragem por conglomerados últimos:

A tabela 4.1 mostra que:

Em média, 77.5% dos AF da província de Nampula são chefiados por um indivíduo do sexo masculino. Analisando o coeficiente de variação, verifica-se que existe uma pequena variabilidade na variável sexo na chefia dos agregados, ou seja uma alta precisão da variável em análise, pois, o coeficiente de variação está entre 0 á 10% (CV=2.4%), o verdadeiro valor da estimativa encontra-se no intervalo de 73.87 a 81.20%, a 95% de confiança. Num plano de AASS para se ter a mesma precisão (2.4%) é necessário aumentar a actual amostra (874 casos) em 1.697 vezes.

763413.49 AF da província de Nampula são chefiados por um indivíduo do sexo masculino. Verifica-se que existe uma pequena variabilidade da variável sexo na chefia dos agregados (CV=5.2%), num intervalo de confiança de 684898.18 a 841928.80 que na prática contém o verdadeiro valor da estimativa a 95% de confiança. Num plano de AASS para se ter a mesma precisão (5.2%) é necessário aumentar a actual amostra (874) em 8.039 vezes.

Tabela 4.1: Estatísticas descritivas da variável sexo.

Estatísticas Univariadas									
Sexo		Estimativa	Erro padrão	Intervalo de confiança 95%		Coeficiente de Variação	Efeito de desenho	Raiz quadrada do efeito de desenho	Tamanho da população
				Inferior	Superior				Amostra não ponderada
Média	Homem	.7754	.01839	.7387	.8120	.024	1.697	1.303	984572.638
Total	Homem	763413.49	39413.29459	684898.18	841928.80	.052	8.039	2.835	984572.638

Com base na tabela 4.2, em média os AF da província de Nampula são chefiados por um indivíduo de 41.16 anos de idade. Analisando o coeficiente de variação, verifica-se que existe uma pequena variabilidade na variável idade na chefia dos AF, ou seja uma alta precisão da variável em análise, pois, o coeficiente de variação está entre 0 á 10% (CV=1.7%), o verdadeiro valor da estimativa encontra-se no intervalo de 39.75 a 42.57 que na prática contém a estimativa a 95% de confiança. Num plano de AASS para se ter a mesma precisão (1.7%) é necessário aumentar a actual amostra (874 casos) em 1.935 vezes.

E que no total 40524010 AF da província de Nampula são chefiados por um indivíduo de 41.16 anos de idade. Verifica-se que existe uma alta precisão da variável em análise, pois, o coeficiente de variação está entre 0 á 10% (CV=5.5%), concretizando-se num intervalo de confiança de 36095441 a 44952579 que na prática contém a estimativa a 95% de confiança. Num plano de AASS para se ter a mesma precisão (5.5%) é necessário aumentar a actual amostra (874) em 19.689 vezes.

Tabela 4.2: Estatísticas descritivas da variável idade

Estatísticas Univariadas									
Idade		Estimativa	Erro padrão	Intervalo de confiança 95%		Coeficiente de Variação	Efeito de desenho	Raiz quadrada do efeito de desenho	Tamanho da população
				Inferior	Superior				Amostra não ponderada
Média	Idade	41.16	.708	39.75	42.57	.017	1.935	1.391	984572.638
Total	Idade	40524010	2223063.108	36095441	44952579	.055	19.689	4.437	984572.638
									874

De acordo com a tabela 1 dos anexos 68.86% dos chefes dos AF da província de Nampula vivem em estado civil de união marital. Verifica-se que existe uma pequena variabilidade na variável estado civil dos AF, ou seja uma alta precisão da variavel em análise, pois, o coeficiente de variação está entre 0 á 10% (CV=2.9%),o verdadeiro valor da estimativa encontra-se no intervalo de confiança de 64.83 a 72.88% a 95% de confiança. Num plano de AASS para se ter a mesma precisão (2.9%) é necessário aumentar a actual amostra (874 casos) em 1.661 vezes.

E que 677935.92 AF da província de Nampula vivem em estado civil de união marital. Verificando-se uma alta precisão da variável estado civil com CV=5.9%, concretizando-se num intervalo de confiança de 597977.93 a 757893.91 que na prática contém a estimativa a 95% de confiança. Num plano de AASS para se ter a mesma precisão (5.9%) é necessário aumentar a actual amostra (874) em 6.772 vezes.

A intepetação das estimativas das restantes variáveis é idêntica a interpretação feita acima variando nos valores da estimativa do Coeficiente de Variação (CV), do erro-padrão, do efeito de desenho, do tamahno da amostra não ponderada e dos intervalos de confiança que se encontra o verdadeiro valor da estimativa da média ou proporção. As tabelas que mostram essas estimativas são 1, 2,...9 dos anexos.

Usando o ajustamento por meio de métodos de re-amostragem verifica-se que há melhor precisão das estimativas em relação ao método de conglomerados últimos interpretado acima pois apresentam menores CVs.

Os resultados obtidos mostram que para o caso do método de BRR tem- se:

77.538% dos AF da província de Nampula são chefiados por um indivíduo do sexo masculino. Analisando o coeficiente de variação, verifica-se que existe uma pequena variabilidade na variável sexo na chefia dos agregados, ou seja uma alta precisão da variável em análise, pois, o coeficiente de variação está entre 0 á 10% ($CV=0.08794\%$), o verdadeiro valor da estimativa encontra-se no intervalo de 77.244 a 77.831% a 95% de confiança. Num plano de AASS para se ter a mesma precisão (0.08794%) é necessário aumentar a actual amostra (874 casos) em 0.00233 vezes. (Tabela 10 dos anexos)

763413.49 AF da província de Nampula são chefiados por um indivíduo do sexo masculino. Verifica-se que existe uma pequena variabilidade da variável sexo na chefia dos agregados ($CV=0.24366\%$), num intervalo de confiança de 755410.063 a 771416.9241 que na prática contém o verdadeiro valor da estimativa a 95% de confiança. (Tabela 10 dos anexos)

A tabela 11 dos anexos mostra que os AF da província de Nampula são chefiados por um indivíduo de 41.159 anos de idade. Analisando o coeficiente de variação, verifica-se que existe uma pequena variabilidade na variável idade na chefia dos AF, ou seja uma alta precisão da variável em análise, pois, o coeficiente de variação está entre 0 á 10% ($CV=0.99114\%$), o verdadeiro valor da estimativa encontra-se no intervalo de 39.404 a 42.914 que na prática contém a estimativa a 95% de confiança. Num plano de AASS para se ter a mesma precisão (0.99114%) é necessário aumentar a actual amostra (874 casos) em 1.321 vezes.

40524010 AF da província de Nampula são chefiados por um indivíduo de 41.159 anos de idade. Verifica-se que existe uma alta precisão da variável em análise, pois, o coeficiente de variação está entre 0 á 10% ($CV=1.32051\%$), concretizando-se num intervalo de confiança de 38221559.8 a 42826460.46 que na prática contém a estimativa a 95% de confiança. (Tabela 11 dos anexos)

De acordo com a tabela 12 dos anexos 68.856% dos chefes dos AF da província de Nampula vivem em estado civil de união marital. Verifica-se que existe uma pequena variabilidade na variável estado civil dos AF, ou seja uma alta precisão da variável em análise, pois, o coeficiente de variação está entre 0 á 10% ($CV=1.70622\%$), o verdadeiro valor da estimativa encontra-se no intervalo de confiança de 63.801 a 73.911% a 95% de confiança. Num plano de AASS para se ter a mesma precisão (1.70622%) é necessário aumentar a actual amostra (874 casos) em 0.563 vezes.

E que no total 677935.92 AF da província de Nampula vivem em estado civil de união marital. Verificando-se uma alta precisão da variável estado civil com $CV=1.71308\%$, concretizando-se num intervalo de confiança de 627966.61 a 727905.24 que na prática contém a estimativa a 95% de confiança. (Tabela 12 dos anexos)

A interpretação das estimativas das restantes variáveis do método BRR é idêntica a interpretação feita acima variando nos valores da estimativa do Coeficiente de Variação (CV), do erro-padrão, do efeito de desenho, do tamanho da amostra não ponderada e dos intervalos de confiança que se encontra o verdadeiro valor da estimativa da média ou proporção. As tabelas que mostram essas estimativas são 13, 14,...20 dos anexos.

Para o caso do método de re-amostragem jackknife a interpretação das estimativas de cada variável é idêntica a do método de BRR variando apenas nos valores da estimativa do Coeficiente de Variação(CV), do erro-padrão, do efeito de desenho e dos intervalos de confiança que se encontra o verdadeiro valor da estimativa da média ou proporção com igual valor estimado para a média ou proporção para cada variável. As tabelas que mostram essas estimativas são 21, 22,...31 dos anexos.

4.2. Comparabilidade dos Métodos Aplicados

Tabela4.3: Comparação dos métodos

variável quantitativa (idade)				variável qualitativa (sexo)		
Estimativa	Método			Método		
	Métodos lineares	Métodos replicados		Métodos lineares	Métodos replicados	
	Conglomerados ultimos	BRR	JKn	Conglomerados ultimos	BRR	JKn
\bar{x}	—	—	—	77.53755	77.53755	77.53755
	41.15898	41.15898	41.15898	—	—	—
	40524010.11	40524010.11	40524010.11	763413.49	763413.49	763413.49

A tabela acima mostra que as estimativas da média e do total de variáveis quantitativas para métodos lineares ou replicados como da proporção e total de variáveis qualitativas para métodos lineares ou replicados não varia de um método para o outro, pois a fórmula para a estimação da média ou proporção e total é idêntica em cada método garantindo que tanto usando métodos lineares ou métodos replicados em dados multietápico-estratificados as estimativas sejam aplicáveis mas as medidas de qualidade em cada método podem variar devido ao tamanho de réplicas e do frame sugerindo assim o método mais preciso para validar os inquéritos.

4.3. Medidas de Qualidade dos Métodos

Tabela 4.4: Qualidade dos métodos para variável idade usando como estimativa a média

variável quantitativa (idade)						
Método	Estimativa (media)	Erro padrão	Intervalo de confiança 95%		Coeficiente de Variação %	Efeito de desenho
			Inferior	Superior		
Conglomerados Ultimos	41.15898	0.70778	39.74901	42.56896	1.71963	1.93469
BRR	41.15898	0.40794	39.40375	42.91422	0.99114	1.32051
JKn	41.15898	0.40793	39.40379	42.91418	0.99112	0.64284

O método de Jackknife apresenta melhor precisão para a estimativa da média, da variável idade, visto que apresenta menor valor para o erro-padrão, para o coeficiente de variação e para o efeito de desenho dando a este método melhor qualidade das estimativas para validação do inquerito.

Figura 4.1: coeficiente de variação para idade média dos métodos BRR e Jackknife

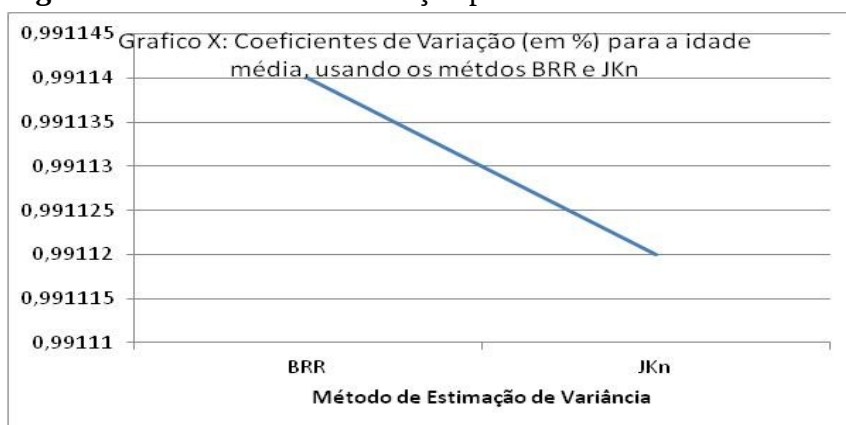


Tabela 4.5: Qualidade dos métodos para variável sexo usando como estimativa a proporção

variável qualitativa (sexo)						
Método	Estimativa (proporção)	Erro padrão	Intervalo de confiança 95%		Coeficiente de Variação %	Efeito de desenho
			Inferior	Superior		
Conglomerados Ultimos	77.53755	.018390	73.87406	81.20104	2.37176	1.69667
BRR	77.53755	0.06819	77.24417	77.83093	0.08794	0.00233
JKn	77.53755	0.06819	77.24417	77.83093	0.08794	0.00233

Tanto o método de jackknife como o método de BRR apresentam melhor precisão para a variável sexo visto que apresentam igual e menor valor para o coeficiente de variação e para o efeito de desenho podendo um deles ser considerado melhor método para precisão e validação do inquerito.

Em geral o método de jackknife apresenta melhores resultados da precisão de estimativas da média para variáveis quantitativas e da proporção para variáveis qualitativas visto que apresenta menores valores para o erro-padrão, para o coeficiente de variação e para o efeito de desenho dando a este método melhor qualidade das estimativas para validação do inquerito.

Tabela4.6: Qualidade dos métodos para variável sexo usando como estimativa o total

variável qualitativa (sexo)					
Método	Estimativa (total)	Erro padrão	Intervalo de confiança 95%		Coeficiente de Variação %
			Inferior	Superior	
Conglomerados Ultimos	763413.49	39413.29459	684898.18	841928.80	5.16277
BRR	763413.49	1860.115394	755410.06	771416.92	0.24366
JKn	763413.49	1860.115394	755410.06	771416.92	0.24366

Tabela4.7: Qualidade dos métodos para variável idade usando como estimativa o total

variável quantitativa (idade)					
Método	Estimativa (total)	Erro padrão	Intervalo de confiança 95%		Coeficiente de Variação %
			Inferior	Superior	
Conglomerados Ultimos	40524010	2223063.108	36095441	44952579	5.48579
BRR	40524010	535123.4490	38221560	42826460	1.32051
JKn	40524010	535123.4490	38221560	42826460	1.32051

Tanto para variáveis qualitativas assim como para as quantitativas a estimativa do total usando o método de Jackknife ou BRR a precisão não varia pois apresenta menor e iguais resultados para a estimativa erro-padrão, para o coeficiente de variação e para o efeito de desenho.

Os métodos replicados são mais precisos em relação aos métodos lineares e a melhoria da precisão depende do frame de actualização das unidades da base de sondagem.

Os vários pressupostos teóricos de que ambos métodos de estimação da variância resultam em estimativas consistentes e próximas umas das outras, para médias, proporções e totais, que vieram a confirmar-se na prática, para amostras grandes como o INCAF 2012-2013. De facto, as variações existentes entre as várias estimativas (desvios padrão, CV e Precisão Relativa), são muito pequenas conforme se pode ver nos quadros apresentados. Por exemplo, para o rácio (média, proporção...), obtiveram-se 2.4%, 0.08794% e 0.08794% de Precisões Relativas (PR), respectivamente, para o método de conglomerados últimos, BRR e Jackknife, para variável sexo, valores próximos uns dos outros.

Embora todos os métodos replicados forneçam estimativas de variâncias não significativamente distintas das obtidas pelo método de linearização de Taylor, estes métodos sobrestimam as verdadeiras variâncias em relação ao método de Linearização de Taylor. Esta situação justifica-se pelo facto do processo de formação de réplicas para a estimação de variâncias se realizar ao nível das UPAs.

Em geral, as estimativas de variância obtidas por Bootstrap e Jackknife são menores que as obtidas pelo BRR, provavelmente devido ao número de estratos na amostra não ser suficientemente grande. Os

métodos Bootstrap e Jackknife fornecem intervalos de confiança $\hat{\theta} \pm 1.96 \sqrt{\hat{\text{Var}}(\hat{\theta})}$ com amplitudes

pequenos e similares.

No final dos anexos apresenta-se gráficos de estatísticas descritivas das variáveis usadas no estudo.

V. CONCLUSÕES E RECOMENDAÇÕES

5.1. CONCLUSÕES

Depois da investigação e análises feitas conclui-se que :

- Os métodos replicados BRR e Jackknife são usados para validar os inquéritos.
- Os melhores métodos de desenho de amostragem que tirem proveito das particularidades da sondagem estratificada são os métodos de re-amostragem ou replicados (BRR e Jackknife), pois apresentam menores CVs. Por exemplo, os CVs para a média da variável idade quando se utiliza o método dos conglomerados últimos ($CV=1.71963$) são maiores que os obtidos por métodos de re-amostragem ($CV=0.99114$ ou $CV=0.99112$).
- Os métodos replicados são mais precisos para estimar a média, total e a proporção (1.71963 , 0.99114 e 0.99112) devido a descartar o tipo de distribuição amostral assumida por uma estatística e calcula uma distribuição empírica combinando várias técnicas de estimação, desde os ajustamentos devido as não respostas e a pós-estratificação.
- Uma comparação directa entre os métodos replicados e lineares mostram que as estimativas para a média, o total e a proporção (41.15898 ; 40524010.11 e 77.53755% respectivamente) não variam de um método para outro, garantindo assim, que as estimativas sejam aplicáveis para o estudo desejado no INCAF 2012-13 da província de Nampula.
- O método de Jackknife apresenta melhor precisão dando a este método melhor qualidade das estimativas para validação do inquérito.

5.2. RECOMENDAÇÕES

- Recomenda-se que em cada inquérito realizado se seleccione as variáveis de maior relevância para o estudo e fazer-se análise da precisão usando os métodos de re-amostragem para validar os inquéritos;
- Recomenda-se que se use o método de re-amostragem de jackknife para estimativas como média e proporção no ajustamento e validação do inquérito realizado;
- Recomenda-se que se façam estudos usando planos de sondagem estratificado quando se está perante dados em que a base está diferenciada em urbano e rural.

REFERÊNCIAS BIBLIOGRÁFICAS

1. ALVES, N. A.C. (2006), Investigação por Inquérito, Universidade dos Açores, Departamento de Matemática, Ponta Delgada;
2. BATTISTI, I.D.E. (2008), Análise de Dados Epidemiológicos Incorporando Planos Amostrais Complexos, Faculdade de Medicina da Universidade Federal do Rio Grande do Sul, Porto Alegre;
3. BOLFARINE, H. e BUSSAB, W. O. (2005), Elementos de Amostragem, Editora Blucher, São Paulo;
4. CAMPOS, M. J. F. P. (nd), O Inquérito Estatístico: uma introdução à elaboração de questionário, amostragem, organização e apresentação dos resultados, ALEA - Acção Local Estatística Aplicada, Portugal;
5. CAVALCANTE, F. e ZEPPELINI, P. D. (2009), O que é Amostragem?, Cavalcante e Associados, Brasil;
6. COCHRAN, W.G. (1977), Sampling Techniques, 3rd Edition, a Willey Publication in Applied Statistics, John Wiley e Sons, New York;
7. CORLETT, M. e MABOTE, P (2000), Desenho da Amostra Mãe Derivada do Recenseamento Populacional de 1997 de Moçambique, Instituto Nacional de Estatística, Moçambique;
8. COSTA, A. C. M. (2000), Técnicas de Estimação no Âmbito da Pós-estratificação, Instituto Superior de Estatística e Gestão de Informação da Universidade Nova de Lisboa, Lisboa;
9. EUROPEAN COMMUNITIES (2002). Monographs of Official Statistics: Variance Estimation Methods in the European Union. Luxembourg: Office for Official Publication of the European Communities. 2002 Edition. Luxembourg;
10. LEMM, A. e D'AGOSTINO, A. (2013), Evaluation Report INCAF 2012-2013 Survey Maputo, 15th-27th September 2013 Mission Report, Italia;
11. LOHR, S. L. (1999), Sampling: Design and Analysis, Second Edition, Arizona State University, United States of America;

12. LUTZ, M. (2007), Amostragem, Instituto Federal de Farroupilha Campus Alegre, São Paulo;
13. MARENDIA, F. R. B. (2010), Amostragem, Universidade Tecnológica Federal do Paraná, Ponta Grossa;
14. REIS, I. A. E ASSUNÇÃO, R. M. (1998), SCIENTIA FORESTALIS: Comparando três métodos de amostragem: métodos de distâncias, contagem de quadrats e conglomerado adaptativo, IPEF – ESALQ Universidade de São Paulo.
15. RENNO C. D. (2011), Jackknife, Bootstrap e outros métodos de reamostragem, São José dos Campos, 8 de dezembro de 2011;
16. RUST, KEITH (1985). Variance Estimation for Complex Estimators in Sample Survey, Journal of Official Statistics, Vol 1, Nº 4; 381-397.
17. SARNDAL, C. E, SWENSSON, B. e WRETMAN, J. (1992), Model Assisted Survey Sampling, Springer Series in Statistics, New York;
18. SOUSA, P. C. O. S. (2011), Introdução a Amostragem Estatística, Universidade Federal do Vale do São Francisco, Juazeiro- BA;
19. TAVARES, R (2008), Noções de Amostragem, Brasil;
20. THOMPSON, S. K. (1993), Sampling, a Wiley Interscience Publication, John Wiley e Sons, New York;
21. TUKEY, J. W. (1958), The Annals of Mathematical Statistics, Bias and confidence in not-quite large samples (abstract).
22. VIEIRA, M. T. F. de A. da S. (2008), Amostragem, Universidade de Aveiro: Departamento de Matemática;

ANEXOS

Tabelas de resultados obtidos a partir do método de conglomerados últimos:

Tabela 1: variável estado civil

Estatísticas Univariadas										
Estado civil		Estimativa	Erro padrão	Intervalo de confiança 95%		Coeficiente de Variação	Efeito de desenho	Raiz quadrada do efeito de desenho	Tamanho da população	Amostra não ponderada
				Inferior	Superior					
Média	Uniao marital	.6886	.02019	.6483	.7288	.029	1.661	1.289	984572.638	874
Total	Uniao marital	677935.92	40137.49406	597977.93	757893.91	.059	6.772	2.602	984572.638	874

Tabela 2: variável total de pessoas no AF

Estudo do plano de sondagem estratificada: aplicação ao inquérito contínuo aos agregados familiares 2012-2013 (trimestre I) da província de Nampula.

Estatísticas Univariadas									
Agregado Familiar	Estimativa	Erro padrão	Intervalo de confiança 95%		Coeficiente de Variação	Efeito de desenho	Raiz quadrada do efeito de desenho	Tamanho da população	Amostra não ponderada
Média Total de pessoas no AF	4.63	.106	4.42	4.84	.023	1.987	1.410	984572.638	874
Total Total de pessoas no AF	4555795	261106.784	4035644	5075947	.057	12.535	3.541	984572.638	874

Tabela 3: variável total de pessoas de 5 e mais anos no AF

Estatísticas Univariadas									
Agregado Familiar	Estimativa	Erro padrão	Intervalo de confiança 95%		Coeficiente de Variação	Efeito de desenho	Raiz quadrada do efeito de desenho	Tamanho da população	Amostra não ponderada
Média Total de pessoas de 5 anos e mais	3.66	.082	3.49	3.82	.022	1.779	1.334	984572.638	874
Total Total de pessoas de 5 anos e mais	3598707	208023.563	3184303	4013111	.058	11.837	3.441	984572.638	874

Tabela 4: variável total de pessoas de 5 à 17 anos no AF

Estatísticas Univariadas									
Agregado Familiar	Estimativa	Erro padrão	Intervalo de confiança 95%		Coeficiente de Variação	Efeito de desenho	Raiz quadrada do efeito de desenho	Tamanho da população	Amostra não ponderada
Média Total de pessoas de 5-17 anos	1.65	.065	1.52	1.78	.040	1.716	1.310	984572.638	874
Total Total de pessoas de 5-17 anos	1622042	114286.167	1394372	1849712	.070	5.412	2.326	984572.638	874

Tabela 5: variável ocupação principal

Estatísticas Univariadas									
Ocupacao principal	Estimativa	Erro padrão	Intervalo de confiança 95%		Coeficiente de Variação	Efeito de desenho	Raiz quadrada do efeito de desenho	Tamanho da população	Amostra não ponderada
Média Campones	.7620	.02548	.7112	.8128	.033	2.494	1.579	844328.706	697
Total Campones	643381.09	44257.87535	555195.34	731566.83	.069	10.554	3.249	844328.706	697

Tabela 6: variável ramo de actividade

Estatísticas Univariadas									
Ramo de actividade	Estimativa	Erro padrão	Intervalo de confiança 95%		Coeficiente de Variação	Efeito de desenho	Raiz quadrada do efeito de desenho	Tamanho da população	Amostra não ponderada
Média Agricultura, silv. Pesca	.7960	.02285	.7505	.8415	.029	2.262	1.504	852519.190	704
Total Agricultura, silv. Pesca	678611.45	45518.38033	587914.10	769308.81	.067	12.352	3.515	852519.190	704

Tabela 7: variável a quem pertence a habitação

Estatísticas Univariadas									
A quem pertence a habitacao	Estimativa	Erro padrão	Intervalo de confiança 95%		Coeficiente de Variação	Efeito de desenho	Raiz quadrada do efeito de desenho	Tamanho da população	Amostra não ponderada
Média propria	.9441	.01041	.9233	.9648	.011	1.791	1.338	984221.170	873
Total propria	929171.50	48126.72341	833298.15	1025044.85	.052	39.521	6.287	984221.170	873

Tabela 8: variável fonte de iluminação

Estatísticas Univariadas										
Fonte de iluminação		Estimativa	Erro padrão	Intervalo de confiança 95%		Coeficiente de Variação	Efeito de desenho	Raiz quadrada do efeito de desenho	Tamanho da população	Amostra não ponderada
				Inferior	Superior					
Média	electricidade	.1901	.02396	.1424	.2378	.126	2.796	1.672	767809.048	750
Total	electricidade	145949.41	18553.16582	108989.61	182909.21	.127	2.843	1.686	767809.048	750

Tabela 9: variável fonte de água

Estatísticas Univariadas										
Fonte de água		Estimativa	Erro padrão	Intervalo de confiança 95%		Coeficiente de Variação	Efeito de desenho	Raiz quadrada do efeito de desenho	Tamanho da população	Amostra não ponderada
				Inferior	Superior					
Média	canalizada na casa do vizinho	.0863	.01503	.0564	.1163	.174	2.500	1.581	984221.170	873
Total	canalizada na casa do vizinho	84968.04	14769.23791	55546.21	114389.87	.174	2.492	1.578	984221.170	873

Tabelas de resultados obtidos a partir do método de re-amostragem BRR:

Tabela 10: variável sexo

Estatísticas Univariadas									
Sexo		Estimativa	Erro padrão	Intervalo de confiança 95%		Coeficiente de Variação %	Efeito de desenho	Tamanho da população	Amostra não ponderada
				Inferior	Superior				
Média (%)	Homem	77.53755	0.06819	77.24417	77.83093	0.08794	0.00233	984572.638	874
Total	Homem	763413.49	1860.115394	755410.063	771416.9241	0.24366	—	984572.638	874

Tabela 11: variável idade

Estatísticas Univariadas									
Idade		Estimativa	Erro padrão	Intervalo de confiança 95%		Coeficiente de Variação %	Efeito de desenho	Tamanho da população	Amostra não ponderada
				Inferior	Superior				
Média	Idade	41.15898	0.40794	39.40375	42.91422	0.99114	1.32051	984572.638	874
Total	Idade	40524010	535123.449	38221559.8	42826460.46	1.32051	—	984572.638	874

Tabela 12: variável estado civil

Estatísticas Univariadas									
Estado civil		Estimativa	Erro padrão	Intervalo de confiança 95%		Coeficiente de Variação %	Efeito de desenho	Tamanho da população	Amostra não ponderada
				Inferior	Superior				
Média (%)	Uniao marital	68.85586	1.17483	63.80095	73.91076	1.70622	0.56253	984572.638	874
Total	Uniao marital	677935.92	11613.606	627966.61	727905.24	1.71308	—	984572.638	874

Tabela 13: variável total de pessoas no AF

Estatísticas Univariadas									
Agregado Familiar		Estimativa	Erro padrão	Intervalo de confiança 95%		Coeficiente de Variação%	Efeito de desenho	Tamanho da população	Amostra não ponderada
				Inferior	Superior				
Média	Total de pessoas no AF	4.62718	0.13868	4.03048	5.22388	2.99712	3.42885	984572.638	874
Total	Total de pessoas no AF	4555795.5	141569.182	3946672.43	5164918.47	3.10745	—	984572.638	874

Tabela 14: variável total de pessoas de 5 e mais anos no AF

Estatísticas Univariadas									
Agregado Familiar		Estimativa	Erro padrão	Intervalo de confiança 95%		Coeficiente de Variação %	Efeito de desenho	Tamanho da população	Amostra não ponderada
				Inferior	Superior				
Média	Total de pessoas de 5 anos e mais	3.65510	0.08845	3.27453	4.03566	2.41990	2.07503	984572.638	874
Total	Total de pessoas de 5 anos e mais	3598706.8	92271.958	3201692.61	3995720.98	2.564031	—	984572.638	874

Tabela 15: variável total de pessoas de 5 à 17 anos no AF

Estatísticas Univariadas									
Agregado Familiar		Estimativa	Erro padrão	Intervalo de confiança 95%		Coeficiente de Variação %	Efeito de desenho	Tamanho da população	Amostra não ponderada
				Inferior	Superior				
Média	Total de pessoas de 5 a 17 anos	1.64746	0.06304	1.37622	1.91869	3.82642	1.59653	984572.638	874
Total	Total de pessoas de 5 a 17 anos	1622041.8	64722.891	1343561.69	1900521.94	3.99021	–	984572.638	874

Tabela 16: variável ocupação principal

Estatísticas Univariadas									
Ocupacao principal		Estimativa	Erro padrão	Intervalo de confiança 95%		Coeficiente de Variação %	Efeito de desenho	Tamanho da população	Amostra não ponderada
				Inferior	Superior				
Média (%)	Campones	76.20031	2.54410	65.25392	87.14670	3.33870	2.48756	844328.706	697
Total	Campones	643381.09	26675.72144	528604.721	758157.45	4.14618		844328.706	697

Tabela 17: variável ramo de actividade

Estatísticas Univariadas									
Ramo de actividade		Estimativa	Erro padrão	Intervalo de confiança 95%		Coeficiente de Variação %	Efeito de desenho	Tamanho da população	Amostra não ponderada
				Inferior	Superior				
Média (%)	Agricultura, silv. Pesca	79.60072	1.39102	73.61566	85.58579	1.74749	0.83889	852519.190	704
Total	Agricultura, silv. Pesca	678611.45	19535.77	594555.82	762667.09	2.87879	2.277	852519.190	704

Tabela 18: variável a quem pertence a habitação

Estatísticas Univariadas									
A quem pertence a habitacao		Estimativa	Erro padrão	Intervalo de confiança 95%		Coeficiente de Variação %	Efeito de desenho	Tamanho da população	Amostra não ponderada
				Inferior	Superior				
Média (%)	propria	94.40678	0.53525	92.10380	96.70975	0.56696	0.47365	984221.170	873
Total	propria	929171.50	48126.72341	833298.15	1025044.85	0.30941		984221.170	873

Tabela 19: variável fonte de iluminação

Estatísticas Univariadas									
Fonte de iluminação		Estimativa	Erro padrão	Intervalo de confiança 95%		Coeficiente de Variação %	Efeito de desenho	Tamanho da população	Amostra não ponderada
				Inferior	Superior				
Média (%)	electricidade	19.00856	0.44001	17.11532	20.90179	2.31483	0.09432	767809.048	750
Total	electricidade	145949.41	2810.976612	133854.755	158044.0669	1.92599		767809.048	750

Tabela 20: variável fonte de água

Estatísticas Univariadas									
Fonte de água		Estimativa	Erro padrão	Intervalo de confiança 95%		Coeficiente de Variação %	Efeito de desenho	Tamanho da população	Amostra não ponderada
				Inferior	Superior				
Média (%)	canalizada na casa do vizinho	8.63302	0.22663	7.65791	9.60814	2.62517	0.05685	984221.170	873
Total	canalizada na casa do vizinho	84968.04	14769.23791	55546.21	114389.87	2.38822	—	984221.170	873

Tabelas de resultados obtidos a partir do método de re-amostragem Jackknife:

Tabela 21: variável sexo

Estatísticas Univariadas									
Sexo		Estimativa	Erro padrão	Intervalo de confiança 95%		Coeficiente de Variação %	Efeito de desenho	Tamanho da população	Amostra não ponderada
				Inferior	Superior				
Média (%)	Homem	77.53755	0.06819	77.24417	77.83093	0.08794	0.00233	984572.638	874
Total	Homem	763413.49	1860.115394	755410.063	771416.9241	0.24366		984572.638	874

Tabela 22: variável idade

Estatísticas Univariadas									
Idade		Estimativa	Erro padrão	Intervalo de confiança 95%		Coeficiente de Variação %	Efeito de desenho	Tamanho da população	Amostra não ponderada
				Inferior	Superior				
Média	Idade	41.15898	0.40793	39.40379	42.91418	0.99112	0.64284	984572.638	874
Total	Idade	40524010	535123.449	38221559.8	42826460.46	1.32051		984572.638	874

Tabela 23: variável estado civil

Estatísticas Univariadas									
Estado civil		Estimativa	Erro padrão	Intervalo de confiança 95%		Coeficiente de Variação %	Efeito de desenho	Tamanho da população	Amostra não ponderada
				Inferior	Superior				
Média (%)	Uniao marital	68.85586	1.17484	63.80091	73.91080	1.70624	0.56254	984572.638	874
Total	Uniao marital	677935.922	11613.60602	627966.609	727905.2352	1.71308		984572.638	874

Tabela 24: variável total de pessoas no AF

Estatísticas Univariadas									
Agregado Familiar		Estimativa	Erro padrão	Intervalo de confiança 95%		Coeficiente de Variação%	Efeito de desenho	Tamanho da população	Amostra não ponderada
				Inferior	Superior				
Média	Total de pessoas no AF	4.62718	0.13868	4.03048	5.22388	3.42885	3.42885	984572.638	874
Total	Total de pessoas no AF	4555795.5	141569.182	3946672.43	5164918.47	3.10745	—	984572.638	874

Tabela 25: variável total de pessoas de 5 e mais anos no AF

Estatísticas Univariadas									
Agregado Familiar		Estimativa	Erro padrão	Intervalo de confiança 95%		Coeficiente de Variação %	Efeito de desenho	Tamanho da população	Amostra não ponderada
				Inferior	Superior				
Média	Total de pessoas de 5 anos e mais	3.65510	0.08845	3.27453	4.03566	2.41989	2.07502	984572.638	874
Total	Total de pessoas de 5 anos e mais	3598706.8	92271.958	3201692.61	3995720.98	2.564031	—	984572.638	874

Tabela 26: variável total de pessoas de 5 à 17 anos no AF

Estatísticas Univariadas									
Agregado Familiar		Estimativa	Erro padrão	Intervalo de confiança 95%		Coeficiente de Variação %	Efeito de desenho	Tamanho da população	Amostra não ponderada
				Inferior	Superior				
Média	Total de pessoas de 5 a 17 anos	1.64746	0.06304	1.37623	1.91869	3.82640	1.59652	984572.638	874
Total	Total de pessoas de 5 a 17 anos	1622041.8	64722.891	1343561.69	1900521.94	3.99021	—	984572.638	874

Tabela 27: variável ocupação principal

Estatísticas Univariadas									
Ocupacao principal		Estimativa	Erro padrão	Intervalo de confiança 95%		Coeficiente de Variação %	Efeito de desenho	Tamanho da população	Amostra não ponderada
				Inferior	Superior				
Média (%)	Campones	76.20031	2.54410	65.25394	87.14668	3.33870	2.48755	844328.706	697
Total	Campones	643381.09	26675.72144	528604.721	758157.45	4.14618		844328.706	697

Tabela 28: variável ramo de actividade

Estatísticas Univariadas									
Ramo de actividade		Estimativa	Erro padrão	Intervalo de confiança 95%		Coeficiente de Variação %	Efeito de desenho	Tamanho da população	Amostra não ponderada
				Inferior	Superior				
Média (%)	Agricultura, silv. Pesca	79.60072	1.39097	73.61587	85.58558	1.74743	0.83883	852519.190	704
Total	Agricultura, silv. Pesca	678611.45	19535.76994	594555.821	762667.0866	2.87879	—	852519.190	704

Tabela 29: variável a quem pertence a habitação

Estatísticas Univariadas									
A quem pertence a habitacao		Estimativa	Erro padrão	Intervalo de confiança 95%		Coeficiente de Variação %	Efeito de desenho	Tamanho da população	Amostra não ponderada
				Inferior	Superior				
Média (%)	própria	94.40678	0.53524	92.10385	96.70971	0.56695	0.47363	984221.170	873
Total	própria	929171.5	2874.93832	916801.641	941541.3627	0.30941		984221.170	873

Tabela 30: variável fonte de iluminação

Estatísticas Univariadas									
Fonte de iluminacao		Estimativa	Erro padrão	Intervalo de confiança 95%		Coeficiente de Variação %	Efeito de desenho	Tamanho da população	Amostra não ponderada
				Inferior	Superior				
Média (%)	electricidade	19,00856	0.44000	17.11539	20.90172	2.31475	0.09431	767809.048	750
Total	electricidade	145949.41	2810.976612	133854.755	158044.0669	1.92599		767809.048	750

Tabela 31: variável fonte de água

Estatísticas Univariadas									
Fonte de água		Estimativa	Erro padrão	Intervalo de confiança 95%		Coeficiente de Variação %	Efeito de desenho	Tamanho da população	Amostra não ponderada
				Inferior	Superior				
Média (%)	canalizada na casa do vizinho	8.63302	0.22663	7.65791	9.60813	2.62515	0.05685	984221.170	873
Total	canalizada na casa do vizinho	84968.039	2029.223576	76236.9943	93699.08286	2.38822	—	984221.170	873