MEGI

MESTRADO
Estatística e Gestão de Informação

CARACTERIZAÇÃO E ANÁLISE DA METODOLOGIA DE DESENHO E ESTIMAÇÃO EM INQUÉRITO POR AMOSTRAGEM EM MOÇAMBIQUE

Uma aplicação prática ao Censo Agro-Pecuário 2009/10 (CAP II Moçambique) para pequenas explorações agrícolas.

Maria Alice Chiponde

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre em Estatística e Gestão de Informação

UNIVERSIDADE NOVA ISEGI

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

CARACTERIZAÇÃO E ANÁLISE DA METODOLOGIA DE DESENHO E ESTIMAÇÃO EM INQUÉRITO POR AMOSTRAGEM EM MOÇAMBIQUE

Uma aplicação prática ao censo agro-pecuário 2009/10 (CAP II Moçambique) para pequenas e médias explorações agro-pecuárias.

por

Maria Alice Chiponde

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em Estatística e Gestão de Informação, Especialização em Análise e Gestão da Informação

Orientador: Professor Doutor José António Rui Amaral Santos

2012

AGRADECIMENTOS

Em primeiro lugar, uma palavra de agradecimento e reconhecimento ao Instituto Nacional de Estatística de Moçambique por ter-me proporcionado a bolsa de estudos e a base de dados para o presente trabalho e pelo apoio e compreensão em todos os momentos.

Um gesto de agradecimento especial ao Professor Doutor José António Santos pelo apoio, disponibilidade, sugestões e orientação deste trabalho.

À minha família e amigos pela atenção, apoio, compreensão e paciência durante um longo período de ausência.

Um gesto de agradecimento e reconhecimento especial para os colegas do INE de Moçambique, e em particular para os do INE de Portugal pelo apoio, estímulo e colaboração resignados que tiveram impacto muito relevante para o trabalho que aqui se apresenta.

A todos os demais que, no ISEGI e fora contribuíram de forma direta ou indireta endereço os meus sinceros agradecimentos.

Caracterização e Análise da Metodologia de Desenho e Estimação em Inquérito Por Amostragem em Moçambique*:*

Uma aplicação prática ao censo agro-pecuário 2009/10 (CAP II Moçambique) para pequenas explorações agrícolas.

RESUMO

Os inquéritos por amostragem são cada vez mais utilizados e assumem um papel crucial no desenvolvimento de estatísticas oficiais, captando e disponibilizando informações atualizadas e úteis (que por vezes não existem nos sistemas de informação) e que servem de instrumento para avaliação, previsão e formulação de soluções apropriadas a uma planificação objetiva da ação em diversos campos.

Este trabalho trata de alguns problemas relacionados com a utilização de dados obtidos através de sondagens.

Na maioria das sondagens, emprega-se, geralmente planos de sondagem complexos (que envolvem estratificação, conglomeração, probabilidades distintas de seleção, pesos distintos de observações, ajustamentos para compensar as não respostas, erros de cobertura, entre outros aspetos), uma vez que a amostra aleatória simples revela-se, impraticável dada a indisponibilidade da base de sondagem que inclua todas as unidades da população em estudo. Por outro lado, as amostras complexas têm a vantagem de proporcionar maior informação para um dado custo total fixo em relação à amostra aleatória simples.

Os dados obtidos através de sondagens são utilizados em análises que envolvem cálculos de estimativas para os parâmetros populacionais de interesse, entre outros fins, tais como a construção de modelos. Para isto, a estatística dispõe-se de uma série de ferramentas de análise. Entretanto, não raras vezes, por diversos motivos, a análise estatística é feita sob condições que não refletem a situação complexa da pesquisa, considerando os dados como obtidos por amostras aleatórias simples.

Ao ignorar os aspetos metodológicos, a análise estatística pode fornecer estimativas pontuais incorretas para os parâmetros, bem como para as respetivas variâncias e, comprometer desta forma os resultados do estudo.

Para ilustrar esta questão neste trabalho, é utilizado o Censo Agro-Pecuário 2009/10 em Moçambique que foi realizado por amostragem para as pequenas e médias explorações agro-pecuárias.

Palavras-chave: métodos de amostragem, efeito do plano amostral, estimação de variância em amostras complexas, análise de dados de inquéritos por amostragem.

Caracterização e Análise da Metodologia de Desenho e Estimação em Inquérito Por Amostragem em Moçambique*:*

Uma aplicação prática ao censo agro-pecuário 2009/10 (CAP II Moçambique) para pequenas explorações agrícolas.

ABSTRACT

The sample surveys are increasingly used and play a crucial role in the development of official statistics, collecting and providing useful and updated information (which may not exist in the information systems) and serving as a tool for assessment, prediction and formulation of appropriate solutions to planning action objective concerning various fields.

This thesis discusses issues relevant to the users of data obtained through sample surveys.

In most of survey samples it is used generally complex sampling designs (which involve stratification, clustering, different probabilities of selection, different weights of units, adjustments to compensate the no responses, coverage errors, among other aspects), since the simple random sample proves impractical given the unavailability of frame that includes all units of the population under study. Moreover, the complex samples have the advantage of proving greater information for a given total fixed cost relative to the simple random sample.

The data obtained through survey sample are used in statistical analysis involving calculations of population estimates for the parameters of interest, as well as in building models. For this, the statistical analysis comprises up a series of analysis tools. However, sometimes for different reasons, statistical analysis is carried out under conditions which do not reflect the complex situation of the research, considering the data as obtained by simple random samples.

By ignoring the methodological aspects, the analysis can provide inaccurate estimates of the parameters, as well as their variances, and thus compromise severely the results of the research.

To illustrate the research issue focused on this work it is used the Census of Agriculture and Livestock 2009/10 in Mozambique which was conducted by sampling for small and medium-size farms.

Keywords: sampling methods, design effect, variance estimation in complex surveys, analysis of sample survey data.

ÍNDICE

1. Introdução

Os inquéritos por amostragem assumem papel crucial no desenvolvimento de estatísticas oficiais, captando e disponibilizando informação atualizada e útil (que por vezes não existe nos habituais sistemas de informação) que sirva de subsídio à avaliação, previsão e formulação de soluções adequadas a uma planificação objetiva da ação nos diversos campos de atividade pública, privada, bem como na vida quotidiana.

Uma das grandes preocupações de qualquer instituição produtora de informação estatística tem a ver com a utilização "correta" dos seus dados. Esta utilização por ser interpretada de diversas formas.

Refira-se que o objetivo básico de uma pesquisa por amostragem é de tirar conclusões que transcendam os dados, inicialmente recolhidos, através de um conjunto de técnicas chamado inferência estatística. Para além da estimação das características (uso descritivo), a inferência permite construir intervalos de confiança para essas medidas cujas formulações genéricas têm propriedades relevantes, designadamente, uma elevada probabilidade de incluir os parâmetros que se pretende estimar. Por último, utilizando as estimativas pontuais obtidas na sondagem, realizar uma série de testes de hipóteses (de significância e de ajustamento) a partir dos quais são tomadas as decisões. Este é o chamado uso analítico dos dados que contempla a formulação, seleção, ajustamento e interpretação de modelos onde o foco é basicamente estabelecer a natureza das relações ou associação entre as variáveis com vista a maximizar a obtenção de informação oculta na sua estrutura.

A maioria das amostras empregues em inquéritos por amostragem é caraterizada por uma situação complexa da sua composição envolvendo:

- Estratificação;
- Conglomeração;
- Probabilidades desiguais de seleção;

- Elementos com pesos distintos;
- Ajustes para compensar as não respostas, erros de cobertura, entre outros aspetos.

Acontece porém, por motivos de vária ordem, tais como a complexidade de análise, a falta de conhecimentos dos detalhes amostrais, a falta de recursos a softwares apropriados, os aspetos metodológicos são ignorados, pressupondo que os dados são independentes e identicamente distribuídos, conduzindo assim, a análise a resultados incorretos.

Constitui propósito do presente trabalho, enfatizar que certos cuidados devem ser observados quando se utiliza dados provenientes de inquéritos por amostragem ilustrando as consequências de se ignorar os detalhes metodológicos acima referidos. Para tal, é feita uma abordagem do desenho dos planos amostrais e de seus reflexos na utilização dos dados provenientes de sondagens complexas, apresentando um exemplo de aplicação prática ao II Censo Agro-Pecuário em Moçambique (CAP II Moçambique 2009/10), onde serão calculadas as estimativas de alguns parâmetros populacionais, totais, médias e proporções, bem como as respetivas variâncias.

Os cálculos destas medidas serão feitos sob duas perspetivas:

i) utilizar um procedimento de cálculo que não considera o verdadeiro plano amostral, ou seja, como se de dados independentes e identicamente distribuídos se tratasse;

ii) usar um mecanismo que considere o plano amostral, efetivamente adotado.

A opção deste tema prende-se, por um lado com a preocupação de quem produz dados estatísticos, e por outro lado, com a necessidade de quem usa tais dados, relativamente à utilização e interpretação corretas dos dados.

A nível pessoal, o contacto diário do autor com a atividade estatística durante anos de trabalho sem o conhecimento da dimensão dos problemas reais de uma pesquisa por amostragem constituem um subsídio para abordar questões de natureza metodológica.

Pelas razões acima apontadas e considerando-se que o tema é pouco aprofundado propõe-se desenvolvê-lo esperando que o mesmo traga contributo para a qualidade das pesquisas em Moçambique, nas diferentes esferas da vida socioeconómica.

Para além deste capítulo I da introdução, inclui, contextualização (capítulo II), revisão bibliográfica (capítulo III), estimação de variância em inquéritos por amostragem complexa (capítulo IV), descrição do plano amostral do CAP (capítulo V), estimação (capítulo VI), comparação das metodologias de estimação de variância, testes estatísticos (capítulo VII), Análise dos resultados (capítulo VIII), conclusões e considerações finais (capítulo IX), e anexos.

2. Contextualização

Moçambique situa-se no sudeste Africano, faz fronteiras a Norte, Oeste e a Sul com a Tanzânia, Malawi, Zâmbia, Zimbabwe, África do Sul e Swazilândia e, a Leste, é banhado pelo Oceano Índico. Tem uma superfície total de 799.380 km^2. Dividido em 11 Províncias: Niassa, Cabo Delgado, Nampula, Zambézia, Tete, Manica, Sofala, Inhambane, Gaza, Província de Maputo e Cidade de Maputo (Figura 1).

A projeção da população residente total para 2010, apontava 22 416 881 habitantes, dos quais 10 799 284 homens e 11 617 597 mulheres. A densidade populacional era de cerca de 28,04 hab/km^2. A esperança de vida da população total foi estimada em 51,5 anos, 49,5 anos para os homens e 53,5 anos para as mulheres. A população rural foi estimada em 69,2%, ou seja, 15 508 590 habitantes, sendo 7 414 679 homens e 8 093 911 mulheres.

O Censo populacional de 2007 apontava uma taxa de analfabetismo de 50,3%, sendo 34,5% a dos homens e 64,1% a das mulheres. Em termos de ocupação, o Censo indicou que cerca de 3/4 (ou seja, 73,6%) da população economicamente ativa, durante o período de observação era composta por pessoas empregues em atividades agrícolas. Por sexo, a taxa nas atividades agrícolas dos homens foi estimada em 60,4% e a das mulheres em 86,4% nas atividades agrícolas.

O IFTRAB 2004/05 estimou que 77,4% da População economicamente ativa ocupada era composta por agricultores e pessoas que prestavam trabalhos na agricultura e pescas. Por sexo, a taxa masculina era de 66,1% e a feminina era de 87,0%.

Em 2009, a taxa de prevalência do VIH foi estimada em 11,5% em população adulta (15-49 anos de idade), sendo muito crítica na faixa etária dos 25-29 anos, com 14,2% nos homens e 16,8% nas mulheres. Por áreas de residência, 15,9% nas áreas urbanas e 9,2% nas áreas rurais. Por nível de instrução, a população sem escolaridade apresenta, as cifras mais baixas, 7,2% nos homens e 9,8% nas mulheres, com o nível primário, 9,1% nos homens e 14,4% nas mulheres, e com o nível secundário ou mais, 10,1% nos homens e 15,0% nas mulheres. (fonte: MISAU, Relatório Final do INSIDA 2009).

O PIB nominal Per capita em 2010 foi estimado em 426 dólares americanos. Os principais produtos de exportação eram compostos por: algodão, castanha de caju, açúcar, chá, copra, pescado, alumínio, gás natural e eletricidade.

O clima Moçambicano é influenciado pelas monções do Oceano Índico e pela corrente quente do canal de Moçambique. Em geral, é tropical e húmido com duas estações: a chuvosa, quente e húmida que vai de Novembro a Abril e, a seca e fria a estender-se de Maio a Outubro. As temperaturas variam, sendo as médias aproximadas a 35℃, 20℃ e 11℃ para as máximas, médias e mínimas, respectivamente. A zona sul sofre seca cíclica e ocorrência com frequência, de tempestades tropicais durante a época quente. As condições climáticas variam um pouco de lugar para lugar, a Norte e na zona costeira, o clima é tropical húmido, no sul e na Província de Tete é tropical seco, no interior da Província de Gaza é tropical árido e, nas montanhas é tropical de altitude.

Em termos de altitude, distinguem-se três áreas: a planície costeira com altitude até 200 metros, a zona de planaltos com altitude entre 200 e 600 metros e a mais elevada com altitude de 1.000 metros. A disposição do relevo aliada a um clima tropical condiciona o aparecimento de rios que atravessam o País em paralelo para o Índico. A terra fértil localiza-se ao longo das bacias hidrográficas e no planalto. A região sul, bem como a parte costeira é predominantemente arenosa e com pouca fertilidade.

Moçambique é um país economicamente agrícola, com cerca de 75% da sua população a depender de agricultura como fonte de rendimento (o que é consistente com a taxa de analfabetismo acima indicada, com a proporção da população rural e indicadores do IFTRAB), com predominância da agricultura familiar, onde a mão-de-obra familiar constitui um fator de produção e uma unidade de consumo.

O Censo Agro-Pecuário (CAP) é uma operação estatística que aborda questões sobre a estrutura agro-pecuária, nomeadamente, número de unidades agro-pecuárias, sua distribuição espacial, tipo de propriedade, uso e aproveitamento de terra, posse e uso de meios de produção e tecnologia usada; produção e produtividade das principais culturas agrícolas; efetivos e produção das principais espécies pecuárias. Serve também para produzir bases de sondagem para a realização de pesquisas que abordem variáveis dinâmicas não cobertas pelo censo e contribuir para um sistema integrado de estatísticas agro-pecuárias. O CAP é exaustivo para as grandes

propriedades agro-pecuárias e aquícolas, e para as médias e pequenas explorações, é uma pesquisa por amostragem probabilística. Os dados do CAP serão interconectados com os dados do módulo comum da secção G (informações agro-pecuárias e piscícolas) do III RGPH com a localização geográfica o que permitirá uma base rica de análise investigando também as inter-relações entre as diversas variáveis agro-pecuárias com as características demográficas e socioecónomicas da população.

O CAP serve de instrumento de avaliação de programas e políticas de desenvolvimento agrário que atendam às necessidades e à melhoria das condições de vida da população, para alívio da pobreza podendo apontar-se o Programa Nacional de Desenvolvimento Agrário (PROAGRI), o Programa Alargado de Redução da Pobreza (PARP), e Objetivos do Desenvolvimento do Milénio (ODM). Serve também para fornecer ferramentas às áreas de investigação, entre outros objetivos.

O INE-M, desde a sua criação em 1996, realizou dois censos da população e habitação, dois censos agro-pecuários e um número considerável de inquéritos por amostragem.

Figura 1. Mapa da República de Moçambique

Fonte: Instituto Nacional de Estatística, 2011.

3. Revisão bibliográfica

A abordagem dos principais conceitos da estatística e da teoria de amostragem encontra-se apresentada neste capítulo para consolidar a base teórica com o intuito de fundamentar as discussões.

3.1 Principais conceitos

A existência de uma variedade de campos onde não é possível realizar experiências, tal como acontece no caso de ciências sociais e económicas levou ao desenvolvimento do método estatístico se bem que mais difícil e menos preciso. Este método consiste na recolha, organização, apresentação, descrição, análise e interpretação dos dados para o seu uso na tomada de decisão.

De acordo com vários autores a destacar Särdal et al. (1992), Cochran (1977), Coelho et al., (2010), Levy & Lemeshow (1999), Larson (2004), Crespo (1993) e Santos (2009) são de seguida apresentados alguns conceitos básicos.

Muitas vezes, há interesse de se conhecer o comportamento de certas características de um conjunto de pessoas, famílias, empresas, etc., mas por impossibilidade ou inviabilidade económica ou temporal, limita-se as observações do estudo a apenas uma parte desse conjunto. Existem mecanismos para se definir os elementos que vão ser observados e estimar, com um dado grau de precisão, as características do conjunto a partir das medidas da parte observada.

No entanto, para que os resultados de uma sondagem sejam válidos é necessário que disponham de bons estimadores pontuais dos parâmetros de interesse, assim como bons estimadores de variância desses mesmos estimadores.

A estimação de parâmetros e das variâncias populacionais é um dos aspetos essenciais do processo da inferência estatística, pois as estimativas das variâncias (ou em alternativa, dos desvios padrão ou dos coeficientes de variação) são indicadores muito relevantes da precisão dos resultados de uma sondagem. Cochran (1977), Wolter (1985) e Pessoa & Silva (1998).

Assim, a qualidade das estimativas de uma sondagem depende muito do processo de amostragem adotado, sendo portanto, um indicador de qualidade desse processo.

A finalidade da teoria de amostragem é desenvolver métodos de amostragem eficiente, através do estabelecimento de técnicas de seleção de amostras e de métodos de estimação que forneçam estimativas suficientemente precisas a um custo mais reduzido possível para a pesquisa em causa (Cochran, 1977).

Na maioria das sondagens, faz-se uma combinação de diferentes métodos aleatórios para o desenho de amostras, de modo a tirar partido das vantagens que cada um oferece, isto é, que forneçam resultados, a um custo baixo possível e a um nível de precisão exigido. A combinação de várias técnicas de sondagem resulta em amostras complexas.

As amostras complexas envolvem:

- Estratificação;
- Conglomeração;
- Probabilidades desiguais de seleção;
- Elementos com pesos distintos;
- Ajustes para compensar as não respostas, erros de cobertura, entre outros aspetos.

Em muitas sondagens, especialmente as que abarcam populações humanas, regra geral emprega-se planos de sondagem complexos, uma vez que a amostra aleatória simples revela-se, geralmente, impraticável dada a indisponibilidade da base de sondagem que inclua todas as unidades da população objeto de estudo. Por outro lado, as amostras complexas têm a vantagem de proporcionar maior informação para um dado custo total fixo em relação à amostra aleatória simples, visto que a estratificação consiste em agrupar indivíduos de diferentes categorias de comportamento em relação a uma ou mais variáveis de pesquisa, isto é, assegura a representação de cada subpopulação na amostra, enquanto a conglomeração oferece uma redução de custo por unidade amostral, devido a custos baixos no cadastro e localização das unidades da população.

Portanto, é muito útil incluir na análise dos resultados de uma sondagem os detalhes amostrais para a sua correta utilização.

Uma **sondagem** é entendida como uma técnica estatística de investigação complexa em condições controladas que incide sobre uma fração da população. Uma sondagem envolve operações de amostragem.

População ou universo $\boldsymbol{U}$: conjunto finito ou infinito de entes (portadores de uma característica comum).

População alvo ou população objetivo (de interesse ou ainda de referência): conjunto de entes portadores de informação que se pretende conhecer determinadas características. Ao iniciar-se uma pesquisa de carácter estatístico deverá definir-se explicitamente o conjunto sobre vai incidir o estudo, de tal forma que se possa dizer sem ambiguidade, se um determinado elemento pertence ou não ao conjunto.

População de estudo (ou a inquirir): conjunto de entes donde é retirada a amostra.

A**mostra $\boldsymbol{s}$**: subconjunto finito da população a ser observado. Uma amostra diz-se representativa quando reflete as mesmas características que a população que se pretende investigar, no que se refere ao fenómeno de interesse.

Unidade estatística de inquirição ou de observação: qualquer elemento ou conjunto de elementos da população em estudo portador da informação de interesse.

Unidades de amostragem: unidades estatísticas selecionadas para a amostra.

Base de sondagem: lista, mapa ou outra ferramenta a partir da qual é retirada a amostra. Idealmente, é exaustiva e sem duplicações das unidades objetos de observação. Raras vezes a base de sondagem contém a totalidade dos indivíduos que constituem a população, pois a sua constituição e atualização é um processo extremamente complexo e oneroso devido ao carácter dinâmico da população.

Um dos aspetos cruciais do plano amostral é a definição da base de sondagem, pois a sua qualidade tem implicações nos resultados da pesquisa. Para contornar algumas das limitações das bases de sondagem, são utilizados certos procedimentos que consistem em utilizar a informação auxiliar disponível na base de amostragem ou que possa estar em outras fontes, desde que esteja correlacionada com o fenómeno que se pretenda investigar.

No CAP, a população alvo consiste no conjunto de todas as explorações agrícolas, pecuárias, agro-pecuárias ou aquícolas. A unidade estatística é o agregado familiar ou exploração agrícola, pecuária, agropecuária ou aquícola. Em Moçambique, tal como foi dito na contextualização, as propriedades agrícolas estão, geralmente associadas às famílias como unidades de produção e de consumo. Daí que se utilize as bases dos censos populacionais como base de sondagem do censos agrícolas. Por este motivo e para minimizar aos problemas de base de sondagem, no questionário do Censo

Populacional, foi incorporado um módulo de questões sobre a prática da atividade agropecuária e piscícola (Secção G).

Amostragem: o método usado na seleção de unidades alvo de medição, a amostra.

Método de amostragem: o mecanismo usado na seleção dos elementos da população para a amostra.

Plano amostral: instrumento que comporta a definição da população alvo, base de sondagem, as técnicas de amostragem, o tamanho da amostra e a informação requerida.

Toda a pesquisa estatística está sujeita a erros. A fonte de incerteza de uma pesquisa por amostragem traduz em termos de erros e tal como Särdal, Swensson e Wrettman, 1992, Pág. 16 afirmam, os erros nas estimativas de uma pesquisa são tradicionalmente divididos em duas categorias: os erros amostrais e não amostrais e estão diretamente associada a problemas com:

1. Processo de seleção da amostra;
2. Base de sondagem;
3. Instrumento de medição e;
4. Mecanismo de respostas.

Erros amostrais: o facto de se observar uma fração da população acarreta a perda de informação e, automaticamente incerteza acerca das características da população no processo de generalização dos resultados da amostra. Por outro lado, uma vez que amostras diferentes conduzem, geralmente, a resultados diferentes, as amostras complexas oferecem maior dificuldade de análise. Estes erros decorrem do ponto 1, processo de seleção da amostra.

Erros não amostrais: estão associados aos restantes erros que não dependem do processo de recolha da amostra e podem ocorrer em qualquer etapa da sondagem. É nesta categoria que inserem os itens 2 a 4, ou seja, limitações da Base de sondagem (subcobertura, sobrecobertura, duplicação, informação auxiliar incorreta ou desatualizada), Instrumento de medição (questões sem clareza e objetividade,

questões ambíguas, etc.) e Mecanismo de respostas (não respostas ou respostas incompletas), erros do processamento de dados (edição, digitação, etc.), entre outros.

Os erros amostrais podem ser medidos e a sua grandeza determina a precisão da sondagem enquanto os não amostrais não são mensuráveis.

Inferência é a parte da estatística encarregue de oferecer os métodos para tirar as conclusões dos dados recolhidos de uma amostra, isto é, a partir das estatísticas (características numéricas de uma amostra) generalizar ou estimar os parâmetros populacionais (características numéricas da população). A inferência tem como ferramenta básica a teoria de probabilidades. A estimação pontual, intervalos de confiança e testes de hipóteses são procedimentos da inferência estatística.

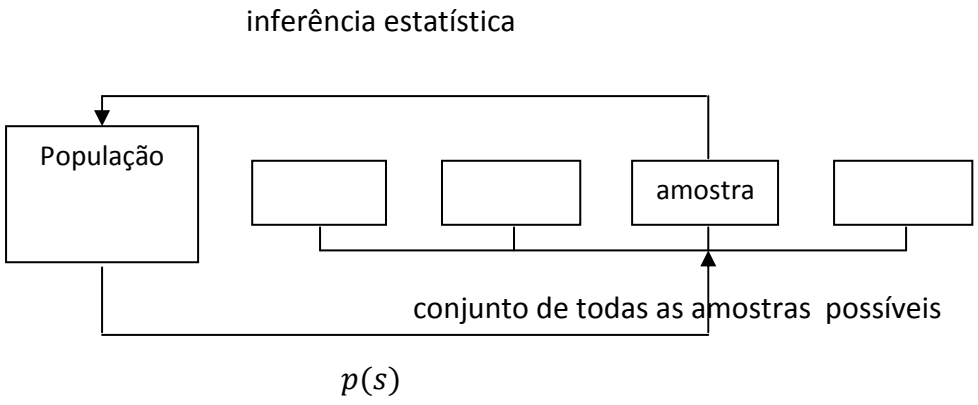

Fig. 2: Inferência estatística

3.2 Definições e notações

Särdal et al. (1992) sugere uma população finita de dimensão N, em que a cada elemento da população está associado um índice, como se os elementos estivessem rotulados pelos número de 1 a N, i.é, $U_i = (i = 1, 2, \dots, k, \dots, N)$. Designando-se por Y a variável de interesse na população, $Y \rightarrow Y_i$ $(i = 1, 2, \dots, N)$, são considerados, por exemplo, os principais parâmetros populacionais, o total (τ), a média (μ), a proporção (P) e a razão, (R) de duas variáveis, Y e X.

O total populacional τ é a soma dos valores do atributo Y das N unidades da população

$$Y = \sum_{i \in U} Y_i = \tau \qquad (3.2.1)$$

A média populacional μ é o valor médio da variável Y dos N elementos da população

$$\bar{Y} = \frac{1}{N} \sum_{i \in U} y_i = \mu \qquad (3.2.2)$$

De acordo com Levy, P., se a característica a ser medida representa a presença ou ausência de algum atributo dicotómico, é frequente desejar-se estimar a proporção de unidades elementares na população que têm o atributo. Se este é representado por $\mathscr{Y}$ e se Y é o número total de unidades elementares na população com o atributo, então P denota a proporção dos elementos da população com o atributo em estudo e é dada por

$$y_i = \begin{cases} 1, se\ o\ i-ésimo\ elemento\ apresenta\ o\ atributo \\ 0, se\ o\ i-ésimo\ elemento\ não\ tem\ o\ atributo \end{cases}$$

$$N_1 = \sum_{i=1}^{N} y_i$$

A média dos valores desta variável define-se como a proporção dos valores com o atributo

$$\mu = \frac{1}{N} \sum_{i \in U} y_i = \frac{1}{N} \sum_{i \in N} 1 + 1 + 0 + 0 + \cdots + 1 + 0 = \frac{N_1}{N} = p \qquad (3.2.3)$$

De modo igual define-se a proporção dos elementos sem o atributo:

$$q = \frac{N_2}{N} = \frac{N - N_1}{N} = 1 - p \qquad (3.2.4)$$

A variância da mesma variável define-se como sendo:

$$\sigma^2 = E(Y^2) - E(Y)^2 = p - p^2 = p(1-p) = pq \qquad (3.2.5)$$

Sejam agora X e Y duas variáveis, $X \neq 0$, define-se o quociente entre essas variáveis:

$$R = \frac{Y}{X} \qquad (3.2.4)$$

3.3 Métodos de sondagem

Existem duas classes de amostras: amostra aleatória ou probabilística e amostra não aleatória ou não probabilística.

3.3.1 Métodos de sondagem não probabilísticos

Na amostragem não probabilística, a seleção das unidades tem um carácter subjetivo, portanto, as probabilidades de seleção dos indivíduos a compor a amostra não são conhecidas, não sendo possível fixar o erro amostral e nem inferir com segurança os resultados para o universo a partir dos dados recolhidos por amostra, pese embora esta técnica seja muito utilizada, sobretudo pelas instituições de pesquisa de opinião. As suas estimativas podem não ser representativas dos parâmetros populacionais desejados. Por esta razão, esta categoria de amostras não será desenvolvida no âmbito deste trabalho.

3.3.2 Métodos de sondagem probabilísticos

Nas amostras aleatórias, cada elemento da população finita tem uma certa probabilidade, diferente de zero, de ser incluído na amostra. Essa probabilidade deve ser conhecida.

O mecanismo usado na seleção da amostra $\boldsymbol{s}$ é caracterizado pela probabilidade $\boldsymbol{p(s)}$ de retirar a amostra do conjunto $\boldsymbol{S}$ de todas as amostras possíveis da população $\boldsymbol{U}$. A seleção da amostra depende das variáveis auxiliares $\boldsymbol{x_i}$ e das variáveis de interesse $\boldsymbol{y_i}$.

O uso de amostras probabilísticas permite exprimir-se o **grau de confiança** (aspeto também fundamental da inferência) que se pode ter nos resultados, isto é, permite prefixar uma fronteira para a precisão, ou seja, mensurar o erro de amostragem (desvio padrão). Tal erro depende das estimativas das probabilidades acima referidas.

Segundo Särndal et al (1992), dado um esquema de seleção da amostra a partir de uma população de dimensão N, supõe-se a existência de uma função $p(.)$ tal que $p(s)$ dá a probabilidade de selecção de uma amostra s ao abrigo do esquema em uso. $p(.)$

é o plano amostral, que desempenha um papel fundamental, pois determina as propriedades estatísticas importantes como a distribuição amostral, o valor esperado e a variância. Portanto,

$$0 \leq p(s) \leq 1, \forall s \in S \qquad (3.3.2.1)$$

e

$$\sum_{s \in S} p(s) = 1 \qquad (3.3.2.2)$$

3.4 Sondagem aleatória simples (SAS)

De acordo com Levy e Lemeshow (1999), Cochran (1977) e Coelho et al., (2010), a sondagem aleatória simples (SAS) é o método de amostragem em que todas as amostras possíveis de dimensão n, fixada a *priori*, a retirar de uma população de dimensão N, têm a mesma probabilidade de selecção. Todos os indivíduos têm a mesma probabilidade de serem selecionados da população.

Conforme o modo de tiragem dos elementos da população, considera-se duas variantes: ***seleção com probabilidades iguais com reposição (PICR)*** e ***seleção com probabilidades iguais sem reposição (PISR)***.

Pode-se realizar a extração dos elementos que vão constituir a amostra enumerando a população de 1 a *N* e sorteando-se, a seguir, através de um dispositivo aleatório[1] *n* elementos da sequência que correspondem aos elementos da amostra.

É fácil notar que a sondagem aleatória simples (SAS) embora seja o método mais simples, quase nunca se usa na prática, dada a exigência de uma base de sondagem que contemple todos os indivíduos da população que permita a identificação individual dos mesmos. Tal lista, geralmente não existe, salvo para casos de populações de dimensão muito reduzida.

[1] Tabela de números aleatórios ou outro procedimento, podendo-se destacar alguns softwares disponível na internet, como é o caso de www.openipi.com, muito utilizado na determinação de amostras em epidemiologia.

3.4.1 Estimadores na sondagem aleatória simples (SAS)

Segundo Cochran (1977) e Crespo (1993), ***estimador*** refere-se à expressão matemática a partir da qual é calculada a estimativa do parâmetro populacional desconhecido com base em observações amostrais. ***Estimativa*** é o valor assumido pelo estimador para uma amostra particular, assim sendo, θ representa um parâmetro de Y, ou seja, θ pode ser representado como função dos valores do parâmetro de interesse Y de todos os elementos da população,

$$\theta = f(y_1, y_2, \dots, y_N)$$

e $\hat{\theta}$ o seu estimador que também pode ser representado como função dos valores das observações da amostra,

$$\hat{\theta} = g(y_1, y_2, \dots, y_n)$$

- **Na sondagem aleatória simples com probabilidades iguais com reposição (SAS-PICR)**

- **Estimador da média μ:**

$$\hat{\mu} = \bar{y} = \frac{1}{n}\sum_{i \in s} y_i \qquad (3.4.1)$$

$$E(\hat{\mu}) = E(\bar{y}) = \frac{1}{n}\sum_{i \in s} n\hat{\mu} = \hat{\mu}$$

- **A variância da média é dada por**

$$V(\hat{\mu}) = V(\bar{y}) = \frac{\sigma^2}{n} \qquad (3.4.2)$$

como σ^2 é desconhecido, usa-se o estimador de variância, $\hat{V}(\hat{\mu})$ que é

$$\hat{V}(\hat{\mu}) = \hat{V}(\bar{y}) = \frac{s^2}{n} \qquad (3.4.3)$$

que é também um estimador centrado de $V(\bar{y})$,onde $s^2 = \frac{1}{n-1}\sum_{i \in s}(y_i - \bar{y})^2$

- **Estimador do total τ:**

Se $\bar{y}$ é o estimador natural de μ então, de $\tau = N\mu$ conclui-se que o estimador do total, $\hat{\tau}$ herde as propriedades do estimador da média, ou seja, $\hat{\tau}$ é estimador centrado do total, τ.

$$\hat{\tau} = N\bar{y} = \frac{N}{n}\sum_{i\in s} y_i \qquad (3.4.4)$$

- **Variância do total:**

$$V(\hat{\tau}) = V(N\bar{y}) = N^2 V(\bar{y}) = N^2\frac{\sigma^2}{n} \qquad (3.4.5)$$

Como a variância da população é, em geral uma incógnita, usa-se o seu estimador

$$\hat{V}(\hat{\tau}) = \frac{N^2}{n}s^2 \qquad (3.4.6)$$

- **Estimador da Proporção P:**

Conforme definido no ponto 3.2 para a população, a proporção p dos indivíduos com um dado atributo na amostra é

$$p = \frac{N_1}{N} \qquad (3.4.7)$$

$$q = \frac{N_2}{N} = \frac{1 - N_1}{N} = 1 - p \qquad (3.4.8)$$

e,

$$\sigma^2 = pq \qquad (3.4.9)$$

- **Variância da proporção:**

$$E(\hat{p}) = p$$

$$V(\hat{p}) = V(y) = \frac{\sigma^2}{n} = \frac{pq}{n} \qquad (3.4.10)$$

$$\hat{V}(\hat{p}) = \hat{V}(\bar{y}) = \frac{s^2}{n} \qquad (3.4.11)$$

Sendo, $s^2 = \frac{1}{n-1}[\sum_{i\in s}(y_i)^2 - n\bar{y}^2]$

Por conveniência, define-se o estimador e as propriedades do estimador da proporção

Por definição, $s^2 = \frac{n\hat{p}\hat{q}}{n-1}$

$$\hat{V}(\hat{p}) = \frac{1}{n}\frac{n\hat{p}\hat{q}}{n-1} = \frac{\hat{p}\hat{q}}{n-1} \qquad (3.4.12)$$

- ***Na sondagem aleatória simples com probabilidades iguais sem reposição (SAS-PISR)***

- **Estimador da média μ:**

$$\hat{\mu} = \bar{y} = \frac{1}{n}\sum_{i\in s} y_i = \frac{1}{n}\sum_{i\in U} y_i\, I_{i\in s} \qquad (3.4.13)$$

$$E(\hat{\mu}) = E(\bar{y}) = \frac{1}{N}N\mu = \mu \qquad (3.4.14)$$

- **A variância da média**

Lembrando que $V(\hat{\theta}) = E\left[\hat{\theta} - E(\hat{\theta})\right]^2$

$$V(\hat{\mu}) = V(\bar{y}) = (1-f)\frac{\sigma'^2}{n} \qquad (3.4.15)$$

onde $f = \frac{n}{N}$ é a taxa de sondagem ou fração amostral.

Como a variância da população é, em geral desconhecida, usa-se o seu estimador, $\hat{V}(\hat{y})$

$$\hat{V}(\bar{y}) = (1-f)\frac{s^2}{n} \qquad (3.4.16)$$

- **Estimador da proporção P e sua variância**

$$\sigma^2 = pq$$

$$E(\hat{p}) = p$$

$$V(\hat{p}) = (1-f)\frac{pq}{n} \qquad (3.4.17)$$

Como p e q são desconhecidos, recorre-se ao estimador da variância

$$\hat{V}(\hat{p}) = (1-f)\frac{\hat{p}\hat{q}}{n-1} = \frac{N-n}{N(n-1)}\hat{p}(1-\hat{p}) \qquad (3.4.18)$$

- **Estimador do total**

$$\hat{\tau} = N\bar{y} \qquad (3.4.19)$$

$$E(\hat{\tau}) = E(N\bar{y}) = NE(\bar{y}) = N\mu = \tau$$

- **A variância do total**

$$V(\hat{\tau}) = V(N\bar{y}) = N^2V(\bar{y}) = N^2(1-f)\frac{\sigma'^2}{n} = \frac{N(N-n)}{n}\sigma'^2 \qquad (3.4.20)$$

- **Estimador da variância do total**

$$\hat{V}(\hat{\tau}) = N^2(1-f)\frac{s^2}{n} = \frac{N(N-n)}{n}s^2 \qquad (3.4.21)$$

3.4.2 Propriedades dos estimadores

Särdal et al. (1992), e Crespo (1993), apontam dois critérios usados para examinar o desempenho de um estimador $\hat{\theta}$:

- enviesamento: $B(\hat{\theta}) = E(\hat{\theta}) - \theta$ e,
- variância: $V(\hat{\theta}) = E[\hat{\theta} - E(\hat{\theta})]^2$ - dispersão em torno da média

onde

$$E(\hat{\theta}) = \sum \hat{\theta}_k P(\hat{\theta}_k)$$

é a média ou valor esperado do estimador $\hat{\theta}$ de θ, $\hat{\theta}_k$ é a estimativa do parâmetro θ de cada amostra selecionada e $p(\hat{\theta}_k)$ é a probabilidade de seleção da amostra s. A partir desta expressão é fácil perceber que as variáveis de aleatorização dependem da distribuição das probabilidades de seleção da amostra.

A precisão como medida de desvio entre os valores do estimador e do verdadeiro parâmetro populacional pode ser quantificado pelo erro quadrático médio, $EQM(\hat{\theta})$,

$EQM(\hat{\theta}) = E[\hat{\theta} - \theta]^2$ - dispersão em torno do verdadeiro parâmetro, verificando-se

$$EQM(\hat{\theta}) = V(\hat{\theta}) + B(\hat{\theta})^2$$

Outra medida de precisão muito utilizada e recomendada é o coeficiente de variação (CV),

$$CV = \frac{\sqrt{V(\hat{\theta})}}{\hat{\theta}}$$

A sua principal propriedade é que permite comparar distribuições diferentes, dado que retira o efeito de escala.

Um estimador diz não enviesado (ou centrado) se o valor esperado ou a média da distribuição amostral das estimativas coincidir com o parâmetro a medir.

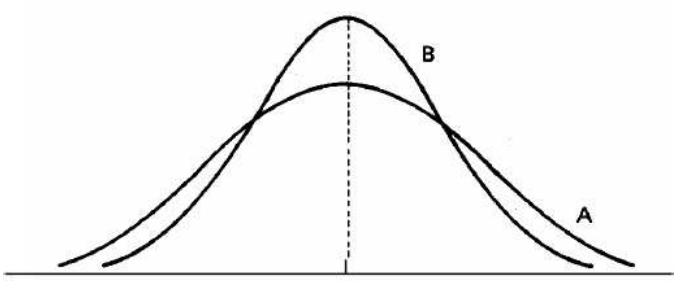

Fig. 3: Estimadores não enviesados

Uma propriedade muito importante é que todo o estimador $\hat{\theta}$ seja não enviesado (ou centrado), isto é, $E(\hat{\theta}) = \theta$ ou $B(\hat{\theta}) = 0$ e que a sua dispersão em torno de θ seja mínima.

As propriedades dos estimadores dependem do plano de sondagem adotado, quer dizer, das probabilidades de inclusão.

3.4.3 Efeito de sondagem

Pessoa e Silva, (1998) e Coelho et al., (2010) indicam o efeito do plano amostral (na língua inglesa, *design effect*, abreviadamente *deff*) para comparar dois planos amostrais, s_1 e s_2 com a mesma dimensão, através da variâncias dos respetivos estimadores $\hat{\theta}_{s_1}$ e $\hat{\theta}_{s_2}$,

$$Efeito\left(\frac{\hat{\theta}_{s_1}}{\hat{\theta}_{s_2}}\right) = \frac{V(\hat{\theta}_{s_1})}{V(\hat{\theta}_{s_2})} \qquad (3.4.3.1)$$

Ou seja,

$$Efeito\left(\frac{\hat{\theta}_{s_1}}{\hat{\theta}_{s_2}}\right) = \frac{V_{PVerd}(\hat{\theta})}{P_{SAS}(\hat{\theta})}$$

Diz-se que s_1 é mais eficiente (preciso) que s_2, se $Efeito\left(\frac{\hat{\theta}_{s_1}}{\hat{\theta}_{s_2}}\right) < 1$ e um *deff*=1 indica que usar um plano de sondagem quanto o outro produz o mesmo impacto.

3.4.4 Efeito do viés dos estimadores sobre os níveis de confiança

Särndal et al., (1992) referem que embora o não enviesamento dos estimadores seja uma característica desejável, a importância do não enviesamento exato não deve ser exagerado e apontam duas razões fundamentais pelos quais em alguns casos se opta pela utilização de estimadores com um enviesamento moderado:

1. Muitos parâmetros têm uma estrutura que torna muito difícil obter estimadores não enviesados;
2. Um estimador com um pequeno enviesamento, muitas vezes pode ter variância e EQM inferiores do que os de um estimador centrado.

Muitos dos estimadores utilizados na prática são parcialmente enviesados, dependendo da ordem de grandeza do seu enviesamento, alguns podem implicar resultados maus, portanto, deve-se evitar utilizar estimadores com enviesamento muito grande.

Lembrando, um estimador ideal é aquele cuja distribuição amostral está fortemente concentrada em torno do valor do parâmetro desconhecido. Uma medida usual de precisão do estimador é o seu EQM, como foi visto,

$$EQM(\hat{\theta}) = E[\hat{\theta} - \theta]^2 = V(\hat{\theta}) + [B(\hat{\theta})]^2 \qquad (3.4.4.1)$$

E que o viés seja o menor possível relativamente ao erro padrão, isto é importante para que os intervalos de confiança sejam válidos. Seja definido o quociente "*bias ratio*"

$$BR(\hat{\theta}) = \frac{B(\hat{\theta})}{\sqrt{V(\hat{\theta})}} \qquad (3.4.4.2)$$

Dependendo da sua grandeza, isto é, se $BR(\hat{\theta})$ for pequeno, apesar de $B(\hat{\theta}) \neq 0$, os intervalos de confiança construídos não apresentarão erro muito grande. Assumindo que a aproximação seguinte é verdadeira,

$$Z = \frac{\hat{\theta} - E(\hat{\theta})}{\sqrt{V(\hat{\theta})}} \sim N(0,1) \qquad (3.4.4.3)$$

a probabilidade do parâmetro desconhecido θ estar contido no intervalo de confiança

$$\hat{\theta} \pm z_{1-\frac{\alpha}{2}} \sqrt{V(\hat{\theta})} \qquad (3.4.4.4)$$

$$Pr\left[\hat{\theta} - z_{1-\frac{\alpha}{2}}\sqrt{V(\hat{\theta})} < \theta < \hat{\theta} + z_{1-\frac{\alpha}{2}}\sqrt{V(\hat{\theta})}\right]$$
$$= Pr\left[-z_{1-\frac{\alpha}{2}} - BR(\hat{\theta}) < Z < z_{1-\frac{\alpha}{2}} - BR(\hat{\theta})\right] \quad (3..4.4.5)$$

Sob a condição da normalidade de Z e se $V(\hat{\theta})$ for conhecido, a probabilidade de cobertura do intervalo de confiança é $1-\alpha$ só quando $BR(\hat{\theta}) = 0$, mas, em geral, $V(\hat{\theta})$ é incógnita, tornando difícil dizer o verdadeiro valor da probabilidade de cobertura. A alternativa é recorrer à sua substituição pelo seu estimador $\hat{V}(\hat{\theta})$. Um valor de $BR(\hat{\theta}) \neq 0$ distorce um pouco a probabilidade de cobertura e o efeito do enviesamento sobre o nível de confiança será pequeno apenas se $BR(\hat{\theta}) \approx 0$.

3.4.5 Consistência e não enviesamento assintótico

Ainda de acordo com Särndal et al. (1992), dado um parâmetro θ estimado por $\hat{\theta}_n$, que é uma função de n variáveis aleatórias independentes e identicamente distribuídas (*v. a. iid)* $\xi_1, \xi_2, \ldots, \xi_n$. O estimador $\hat{\theta}_n$ diz não enviesado assintoticamente se:

$$\lim_{n\to\infty} E(\hat{\theta}_n) = \theta \qquad (3.4.5.1)$$

e $\hat{\theta}_n$ diz-se consistente para θ se $\forall\varepsilon > 0$,

$$\lim_{n\to\infty} Pr(|\hat{\theta}_n - \theta| > \varepsilon) = 0 \qquad (3.4.5.2)$$

$\hat{\theta}_n$ não é o único estimador, mas sim uma sequência $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_n$, assintoticamente não enviesado ou consistente.

Como n é sempre finito, embora muitas vezes grande, quando se sabe que um estimador é assintoticamente não enviesado, então pode ser considerado aproximadamente não enviesado quando n é suficientemente grande. E quando se mantém consistente, então a distribuição amostral de $\hat{\theta}_n$ pode considerar-se fortemente concentrada em torno de θ.

Recorrendo à teoria de amostragem, a definição do não enviesamento e de consistência aqui tratadas não podem ser aplicados diretamente para amostras de populações finitas. Se $\hat{\theta}_n$ é um estimador de baseado numa amostra de dimensão *n* de uma população de *N* elementos, desde que $n \leq N$ e N fixo, então não existe o limite quando $n \to \infty$.

Neste contexto, os resultados assintóticos exigem uma ferramenta matemática mais complexa com sequências de incrementos populacionais, deste modo, ambos *n* e *N*

tendem a infinito. Um tratamento completo destas questões não é apresentado neste trabalho, mas apenas uma ideia basilar do quadro conceptual para o raciocínio assintótico em inquéritos por sondagem.

Se os estimadores $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_g$ dos parâmetros $\theta_1, \theta_2, \dots, \theta_g$ são consistentes, então a função $f = f(\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_g)$ é consistente, ou seja, uma função de estimadores consistentes é uma função consistente.

3.5 Amostragem sistemática

Segundo Cochran (1977) e Klein (2007), os elementos da população devem estar ordenados segundo uma variável relacionada com o atributo em estudo. A amostra é obtida calculando-se primeiro o intervalo de seleção que é dado por

$$k = \frac{N}{n} \qquad (3.5.1)$$

k deve ser inteiro ou arredondado por defeito ao inteiro mais próximo.

A seleção aleatória é só para o primeiro elemento dentre os k primeiros elementos que a partir dele são retirados os elementos seguintes, que distam entre si k unidades até a lista findar.

Na amostragem sistemática, todos os elementos da população têm a mesma probabilidade de serem selecionados. Entretanto, as probabilidades de seleção de diferentes amostras são diferentes, uma vez que os elementos pertencentes a um mesmo intervalo de seleção (definido pelo quociente $k = \frac{N}{n}$, sendo k um número inteiro, ou seja, quando não inteiro, arredonda-se para o inteiro menor mais próximo) têm probabilidade nula de pertencer à mesma amostra. Como consequência, o número total de amostras possíveis de dimensão n a partir de uma população de tamanho N é menor que na SAS. As expressões para o cálculo das medidas de tendência central são as mesmas que as da SAS-PISR mas as dos erros padrões (raízes das variâncias) são diferentes, sendo propostos alguns algoritmos baseados na consideração de que cada intervalo de seleção é constituído num conglomerado de elementos.

Há maiores vantagens da sistemática se existir na base de sondagem, informação sobre a população relacionada com a variável de interesse. A sistemática é mais eficiente que a SAS, sobretudo se a disposição dos elementos é feita de tal modo que,

em relação ao fenómeno em estudo, os elementos homogéneos entre si estejam mais próximo uns dos outros, resultando numa estratificação induzida. Assim, a amostra pode recolher um ou poucos indivíduos de cada grupo de elementos semelhantes e assim formando uma amostra que se traduz numa representação ideal da variabilidade do fenómeno em estudo.

Quando a disposição dos elementos na base de sondagem não está correlacionada com os atributos que se pretende medir, a amostragem resume-se numa SAS.

Se houver periodicidade na ordem dos indivíduos na base de amostragem ela constitui um risco, pois a amostra está sujeita a constituir-se por elementos homogéneos, tornando assim as estimativas menos precisas. Este fenómeno é especialmente vulnerável na amostragem de períodos de tempo e de áreas geográficas. A outra desvantagem da sistemática é que mesmo em circunstâncias em que é mais precisa do que a SAS, as suas propriedades teóricas tornam difíceis mensurar essa precisão.

3.6 Sondagem aleatória estratificada (ST)

Segundo Cochran (1977), Särndal et al (1992) e Coelho et al. (2010), se a população U apresenta categorias de comportamento diferente entre si face a uma ou mais variáveis, e um mesmo comportamento dentro de cada classe, o quadro pode ser estruturado considerando esses subgrupos chamados estratos (U_h) de N_h elementos cada um $(h = 1, 2, \cdots, H)$.

A estratificação consiste em considerar cada estrato U_h como uma população independente donde é extraída uma amostra aleatória s_h de n_h elementos, usando qualquer processo de amostragem probabilístico, objetivando o máximo de precisão das estimativas e avaliação da respetiva precisão.

A amostra global s é constituída pela união de todas as amostras s_h das subpopulações.

Vantagens:

- Melhorar a precisão das estimativas dos parâmetros populacional, desde que os estratos sejam definidos com base na relevância do critério adotado;
- Assegurar na amostra, a representação de cada subpopulação (algumas poderiam estar diluídas numa SAS), assegurando assim a estimação dos parâmetros em cada estrato com uma precisão pré-determinada;

- Excluir amostras que agrupem valores extremos (grupos de *outliers*);
- Reduzir os custos administrativos ou operacionais;
- Dado que cada estrato é tratado como uma população independente, a estratificação permite utilizar diferentes tipos de técnicas de amostragem, aplicando em cada estrato, uma técnica apropriada tendo em conta o custo e benefício;
- Reduzir o efeito dos erros não amostrais.

Cochran (1977, p. 90 e 127) indica os problemas que se colocam no processo de estratificação de uma população:

"What is the best characteristic for the construction of strata? How should the boundaries between the strata be determined? How many strata should there be?"

Ele refere que o problema de estratificação da população foi investigado inicialmente por Dalenius (1957), e depois muitos outros estudos se seguiram.

A seleção e implementação de estratos podem aumentar o custo e complexidade de seleção da amostra, assim como a complexidade das estimativas dos parâmetros populacionais;

A definição de variáveis de estratificação é crucial, mas pode ser difícil. Algumas variáveis de estratificação podem estar relacionadas com o fenómeno em estudo mas outras podem não estar, complicando a pesquisa e reduzindo a utilidade dos estratos;

Em determinadas circunstâncias, especialmente quando o número de estratos é grande, o tamanho da amostra por estrato pode exigir o aumento da amostra global relativamente ao que seria necessário noutros cenários.

A estratificação é mais eficaz mediante três condições:

- A variabilidade dentro dos estratos é mínima;
- A variabilidade entre os estratos é maior;
- Os critérios de estratificação estão fortemente correlacionadas com o fenómeno em estudo.

3.7 Sondagem aleatória com probabilidades desiguais

Das lições de Särndal et al. (1992), Coelho et al. (2010), e Pessoa e Silva, (1998), foi visto que, na SAS, as probabilidades de seleção são iguais para quaisquer unidades da população. Porém, em muitas situações práticas, a população consiste em elementos com dimensão muito diferente, nas quais a SAS revela-se inadequada, uma vez que atribui a mesma probabilidade de seleção para todos os entes, em outras palavras, não considera a importância ou peso das unidades maiores na população. Portanto, a ***sondagem com probabilidades desiguais*** constitui uma ferramenta indispensável no processo de expansão dos resultados da amostra para a população, garantindo estimativas precisas.

A opção pela sondagem com probabilidades desiguais associa-se aos planos adotados, aos objetivos do estudo e se elas estão correlacionadas com o fenómeno em estudo.

Designando por $I(A)$ a função indicatriz que toma o valor 1 se o evento A ocorre e 0, caso contrário, o que quer dizer que o evento é uma função que tem a distribuição de Bernoulli, então:

$$I_{i\in s} = \begin{cases} 1, se\ i \in s, \quad i = 1, 2, \dots, n \\ 0, \qquad\quad caso\ contrário \end{cases}$$

A ***probabilidade de inclusão de ordem 1*** que se denota por $\boldsymbol{\pi_i}$ é a probabilidade de um certo elemento i da população pertencer à amostra, ou seja,

$$\pi_i = Pr(i \in s), \qquad i \in U$$

A ***probabilidade de ordem 2***, indicada por $\boldsymbol{\pi_{ij}}$ é a probabilidade de dois elementos da população i e j $(i \neq j)$, pertencerem, simultaneamente, à amostra, ou seja,

$$\pi_{ij} = Pr(i \in s \wedge j \in s), \quad i, j \in U, \quad i \neq j$$

As propriedades de uma estatística são avaliadas pela distribuição amostral, ou seja, podem ser expressas como funções de $I_{i\in s}$, portanto pelos operadores de esperança $E(.)$ e de variância $V(.)$, pelo que é importante descreve-las. Neste caso,

$$E(I_{i\in s}) = \pi_i \qquad (3.7.1)$$

e

$$V(I_{i\in s}) = E\{[I_{i\in s} - E(I_{i\in s})]^2\} = E(I_{i\in s}^2) - E(I_{i\in s})^2 = \pi_i - \pi_i^2 = \pi_i(1 - \pi_i) \qquad (3.7.2)$$

A ***probabilidade de ordem 2***, π_{ij}, de dois elementos da população i e j $(i \neq j)$, pertencerem, simultaneamente, à amostra é dada por

$$\pi_{ij} = \pi_{ji} = Pr(i \in s \wedge j \in s\,), \qquad \forall i,j \in U \ \wedge \ i \neq j \qquad (3.7.3)$$

Se $i = j$ então $\pi_{ii} = Pr(I_{i\in s}^2 = 1) = Pr(I_{i\in s} = 1) = \pi_i$

De modo análogo, define-se os operadores de esperança e de variância:

$$E\big(I_{i,j\in s}\big) = E\big(I_{i\in s} \wedge I_{j\in s}\big) = \pi_{ij} = \pi_{ji} \qquad (3.7.4)$$

e

$$V\big(I_{i.j\in s}\big) = V\big(I_{i\in s} \wedge I_{j\in s}\big) = \pi_{ij}\big(1 - \pi_{ij}\big) \qquad (3.7.5)$$

A covariância de duas variáveis é dada por:

$$C\big(I_{i\in s}, I_{j\in s}\big) = E\big(I_{i\in s}I_{j\in s}\big) - E(I_{i\in s})E\big(I_{j\in s}\big) = \pi_{ij} - \pi_i\pi_j \qquad (3.7.6)$$

Seja $\Delta_s = [I_{1\in s}, \dots, I_{i\in s}, \dots, I_{N\in s}]'$ um vector aleatório de indicadores dos elementos incluídos na amostra s,

$$\Delta_{ij} = \pi_{ij} - \pi_i\pi_j$$

Então

$$C\big(I_{i\in s}, I_{j\in s}\big) = \pi_{ij} - \pi_i\pi_j = \Delta_{ij} \qquad (3.7.7)$$

Se $i = j$, segue que

$$C(I_{i\in s}, I_{i\in s}) = V(I_{i\in s}) = \pi_i(1 - \pi_i) = \Delta_{ii} \qquad (3.7.8)$$

3.8 Estimação de totais em inquéritos por amostragem

Devido à importância que a estimação de totais representa para as outras estimativas, tais como médias, proporções, razões e taxas, serão aqui apresentadas algumas expressões básicas para os totais populacionais, variâncias e estimadores de variância.

De acordo com Särndal et al., (1992, p.42), seja o problema de estimar o vetor $Y = \tau = \sum_U y_i$ dos totais das R variáveis de interesse de uma

população$Y_1, Y_2, \dots, Y_l, \dots, Y_R$ por meio de uma sondagem. Um estimador amplamente utilizado para o total τ de uma variável de interesse numa população é o estimador de Horvitz-Thompson ($\hat{\tau}_{HT}$), também designado π-ponderado, podendo representar-se também por $\hat{Y}_\pi$ ou $\hat{\tau}_\pi$.

$$\hat{\tau}_{HT} = \sum_{i \epsilon s} \frac{y_i}{\pi_i} \qquad (3.8.1)$$

Onde π_i é a probabilidade de inclusão do indivíduo i na amostra s. Desta relação nota-se que as observações individuais das unidades escolhidas para a amostra são ponderadas pelo inverso das suas probabilidades podendo-se definir este coeficiente como o peso do indivíduo i, w_i

$$w_i = \pi^{-1} \qquad (3.8.2)$$

Segue que,

$$\hat{\tau}_{HT} = \sum_{i \epsilon s} \pi^{-1} y_i \qquad (3.8.3)$$

Assim, o estimador π-ponderado é expresso como uma função linear dos pesos.

As propriedades estatísticas deste estimador são avaliadas em relação à distribuição de probabilidades. Sejam $E(.)$ e $V(.)$ os operadores de esperança e de variância induzidos pelo plano amostral $p(s)$.

Sendo $E(I_i) = \pi_i$ e $\pi_i > 0, \forall i \in U$, então o estimador –ponderado é não enviesado, i. é,

$$E(\hat{\tau}_{HT}) = \sum_{i \in U} I_i \frac{y_i}{\pi_i} = \sum_{U} \pi_i \frac{y_i}{\pi_i} = \sum_{U} y_i = \tau = Y \qquad (3.8.4)$$

ou seja,

$$E(\hat{\tau}_\pi) = \tau$$

A sua variância é definida por:

$$V(\hat{\tau}_\pi) = \sum_{i \in U} \sum_{\substack{j \in U \\ j \neq i}} (\pi_{ij} - \pi_i \pi_j) \frac{y_i y_j}{\pi_i \pi_j} \qquad (3.8.5)$$

Em alternativa, quando o plano de sondagem tem dimensão fixa, a expressão para a variância do estimador pode ser

$$V(\hat{\tau}_\pi) = -\frac{1}{2} \sum_{i \in U} \sum_{j \in U} (\pi_{ij} - \pi_i \pi_j) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)' \qquad (3.8.6)$$

se $i = j$, os valores desses indivíduos não contribuem para a soma. Recomenda-se, frequentemente a utilização de dois estimadores que são:

$$\hat{V}(\hat{\tau}_\pi) = \sum_{i \in s} \sum_{j \in s} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{y_i y_j}{\pi_i \pi_j} \qquad (3.8.7)$$

que é um estimador não enviesado ou centrado, desde que $\pi_{ij} > 0, \forall i, j \in U$, isto é,

$$E\left[\hat{V}(\hat{\tau}_\pi)\right] = V(\hat{\tau}_\pi)$$

O outro é o estimador de Sem-Yates-Grundy

$$\hat{V}_{SYG}(\hat{\tau}_\pi) = -\frac{1}{2} \sum_{i \in s} \sum_{j \in s} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \right) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \qquad (3.8.8)$$

Nota-se que não obstante o facto de as equações da variância acima indicadas, (3.8.5) e (3.8.6) serem equivalentes, em planos de sondagem de dimensão fixa, não se pode dizer o mesmo das expressões dos estimadores das suas variâncias (3.8.7) e (3.8.8) se bem que $\hat{V}_{SYG}(\hat{\tau}_\pi)$ é também não enviesado para amostras de dimensão fixa.

Quando o plano de sondagem é SAS-PISR, as equações dadas para o estimador do total, a sua variância e o estimador da sua variância simplificam-se, dado que as probabilidades de inclusão da primeira e segunda ordem são

$$\pi_i = \frac{n}{N}, \quad \forall i \in U$$

$$\pi_{ij} = \frac{n(n-1)}{N(N-1)}, \quad \forall i \neq j \in U$$

o que leva a

$$\hat{\tau}_\pi = \frac{N}{n} \sum_{i \in s} y_i = N \bar{y}$$

$$V_{SAS-PISR}(\hat{\tau}_\pi) = N^2 \frac{1-f}{n} \frac{N}{N-1} \sigma^2 \qquad (3.8.9)$$

$$\hat{V}_{SAS-PISR}(\hat{\tau}_\pi) = \hat{V}_{SYG}(\hat{\tau}_\pi) = N^2 \frac{1-f}{n} \frac{n}{n-1} s^2 \qquad (3.8.10)$$

Existem muitos estimadores de totais, mas os mais comuns são estimadores ponderados lineares da forma

$$\hat{\tau}_P = \sum_{i \in s} w_i y_i \qquad (3.8.11)$$

onde w_i é o peso associado ao indivíduo i da amostra, $i \in s$. O estimador Horvitz-Thompson é um caso particular destes estimadores em que o peso é dado por

$$w_i = \pi^{-1}, \qquad \forall i \in s$$

O estimador de razão muito utilizado na teoria de amostragem é dado por:

$$\hat{Y}_R = \hat{\tau}_R = \left(\sum_{i \in s} \pi_i^{-1} y_i\right)\left(\sum_{i \in U} x_i\right) \Big/ \left(\sum_{i \in s} \pi_i^{-1} x_i\right) \quad (3.8.12)$$

Onde x é uma variável auxiliar cujo total populacional é $\sum_{i \in U} x_i = X$ que é conhecido. O estimador $\hat{\tau}_R$ também pode ser representados na forma linear,

$$\hat{\tau}_R = \sum_{i \in s} P_i y_i$$

cujo peso é dado por

$$w_i^R = \frac{\pi_i^{-1} \sum_{k \in U} x_k}{\sum_{k \in s} \pi^{-k} x_k} = \frac{\pi_i^{-1} X}{\hat{X}_\pi} \qquad (3.8.13)$$

onde $\hat{X}_\pi = \pi^{-1} x_i$ é o estimador π-ponderado de X

O estimador de razão é enviesado sob a distribuição aleatória em amostras de dimensão reduzida. No caso de amostras grandes, o enviesamento é insignificante e existem expressões assintóticas para tais variâncias da distribuição a partir das quais foram construídos os seus estimadores das variâncias.

A estimação de variâncias para estimadores como os de razão conduz a um problema básico da teoria de amostragem que envolve processos de estimação de variâncias de estimadores complexos.

Existem diferentes métodos de estimação de variância dos parâmetros de interesse a partir dos dados de amostras provenientes de planos de sondagem complexos. A sua opção depende muito das técnicas aplicadas nos ajustamentos para incorporar pesos e planos amostrais no processo de expansão da amostra.

3.9 Estimadores Horvitz-Thompson (π) para diversas variáveis de pesquisa

Segundo Särdal et al. (1992), a maioria dos inquéritos não envolvem uma única variável de estudo, mas sim, um conjunto de variáveis. Seja o problema de analisar o

caso em que o total de cada variável é estimado por seus π estimadores correspondentes. Sejam q variáveis de estudo, designadas por $y_1, y_2, \dots, y_l, \dots, y_q$. Dados os N valores populacionais, $y_{l1}, y_{l2}, \dots, y_{li}, \dots, y_{lN}$ da l–ésima variável $(l = 1, 2, \dots, q)$, estimar os q componentes do vector de totais desconhecidos

$$\tau = \left(\tau_1, \tau_2, \dots, \tau_l, \dots, \tau_q\right)', \qquad \tau_l = \sum_U y_{li} \qquad (3.9.1)$$

Uma vez que a amostra s é extraída da população U segundo um plano amostral $p(s)$ com probabilidade de inclusão π_i e π_{ij}. Para cada $i \in s$, são observados os valores do vector

$$y_i = \left(y_{1i}, y_{2i}, \dots, y_{li}, \dots, y_{qi}\right)'$$

Assumindo que cada total é estimado pelo correspondente estimador π, tem-se:

$$\hat{\tau}_\pi = \left(\hat{\tau}_{1\pi}, \hat{\tau}_{2\pi}, \dots, \hat{\tau}_{l\pi}, \dots, \hat{\tau}_{q\pi}\right)'$$

onde $\hat{\tau}_{j\pi} = \sum_s \check{y}_{li}$, com $\check{y}_{li} = y_{li}/\pi_i$, logo obtém-se um estimador centrado,

$$E(\hat{\tau}_{\pi\pi}) = \tau \qquad (3.9.2)$$

A matriz de variância e covariância associado ao estimador $\hat{\tau}_\pi$ e com o estimador não enviesado da mesma matriz,

$$V(\hat{\tau}_\pi) = E[(\hat{\tau}_\pi - \tau)(\hat{\tau}_\pi - \tau)']$$

É uma matriz simétrica tal que o l–ésimo elemento da diagonal principal é dado pela variância de $\hat{\tau}_{l\pi}$,

$$V(\hat{\tau}_{l\pi}) = \sum\sum_U \Delta_{ij}\check{y}_{li}\check{y}_{lj} \qquad (3.9.3)$$

O elemento ll', fora da diagonal principal é dados pela covariância entre $\hat{\tau}_{l\pi}$ e $\hat{\tau}_{l'\pi}$,

$$Cov\left(\hat{\tau}_{l\pi}, \hat{\tau}_{l'\pi}\right) = \sum\sum_U \Delta_{ij}\check{y}_{li}\check{y}_{l'j} \qquad (3.9.4)$$

A matriz $V(\hat{\tau}_\pi)$ tem como estimador não enviesado a matriz $\hat{V}(\hat{\tau}_\pi)$ tal que o l–ésimo elemento da diagonal principal é

$$\hat{V}(\hat{\tau}_{l\pi}) = \sum\sum_s \check{\Delta}_{ij}\check{y}_{li}\check{y}_{lj} \qquad (3.9.5)$$

O elemento ll', fora da diagonal principal também pode ser dado pelo estimador de covariância entre $\hat{\tau}_{l\pi}$ e $\hat{\tau}_{l'\pi}$,

$$\widehat{Cov}(\hat{\tau}_{l\pi}, \hat{\tau}_{l'\pi}) = \sum\sum_{s} \check{\Delta}_{ij}\check{y}_{li}\check{y}_{l'j} \qquad (3.9.6)$$

Onde $\check{\Delta}_{ij} = \Delta_{ij}/\pi_{ij}$

Os resultados dos elementos da diagonal principal de $V(\hat{\tau}_{\pi})$ e $\widehat{Cov}(\hat{\tau}_{\pi})$ resultam da expressão $\hat{\tau}_{\pi} = \sum_s \left(\frac{y_i}{\pi_i}\right)$. Se $l \neq l'$, o elemento ll' da matriz $V(\hat{\tau}_{\pi})$ é:

$$Cov(\hat{\tau}_{l\pi}, \hat{\tau}_{l'\pi}) = Cov\left(\sum_U I_i\check{y}_{li}, \sum_U I_i\check{y}_{l'i}\right) = \sum\sum_U Cov(I_i, I_j)\check{y}_{li}\check{y}_{l'j}$$
$$= \sum\sum_U \Delta_{ij}\check{y}_{li}\check{y}_{l'j} \qquad (3.9.7)$$

3.10 Amostragem por conglomerado (ou cluster)

Cochran (1977), Coelho et al. (2010) e Klein (2007), referem que é, geralmente, impossível dispor de bases de sondagem completas de indivíduos objetos de pesquisa, tornando quase sempre impraticável a utilização directa de qualquer um dos processos de amostragem anteriormente vistos. Entretanto, na maioria dos casos existem listas de conjuntos de indivíduos formados por algum critério. Tais conjuntos chamam-se conglomerados ou clusters.

Estes grupos constituem as unidades primárias de amostragem (UPA) ou, simplesmente, unidades primárias (UPs), objeto de seleção como se de indivíduos se tratasse num processo de amostragem aleatório e os seus constituintes são chamados unidades secundárias de amostragem (USAs) ou, simplesmente, unidades secundárias (USs).

Na conglomeração, a amostra é composta por todos os elementos pertencentes aos conglomerados selecionados. Assim, a probabilidade de inclusão de um elemento i na amostra coincide com a probabilidade do conglomerado g a que esse elemento pertence e define-se como sendo:

$$\pi_{i,g} = Pr(i \in s) = Pr(g \in S_G) = \pi_g = \frac{m}{M} \qquad (3.10)$$

A base dos conglomerados é que estes devem ser exaustivos e mutuamente exclusivos.

Contrariamente aos estratos, na conglomeração, o ideal é ter-se conglomerados constituídos por elementos com comportamento bem diferente entre si e que os

conglomerados também o sejam entre eles. Assim, um conglomerado seria uma boa representação da população, isto significa que pode-se conseguir boas estimativas com poucos conglomerados, mas na prática é muito difícil conseguir conglomerados de elementos muito diferentes, recorrendo-se muitas vezes a conglomerados naturais que tendem a apresentar elementos com alguma semelhança.

Vantagens: Custos reduzidos

De listagem: não há necessidade de listar todos os elementos da população alvo;

De viagens: o entrevistador não faz visitas a vários aglomerados para inquirir poucos elementos.

Desvantagens:

O erro padrão das estimativas obtidas, ou seja, a variabilidade da amostragem por cluster é, geralmente mais elevado do que nos outros métodos baseados na amostra de dimensão igual, porque as unidades dentro do mesmo cluster tendem a ter comportamento semelhante em relação a muitos atributos. Razão pela qual, a amostragem por conglomerado, regra geral, exige uma amostra maior para conseguir o mesmo nível de precisão.

Se a conveniência administrativa e custos forem os únicos critérios para a escolha da técnica de amostragem, a conglomeração é a técnica ideal, mas se o critério for somente a precisão das estimativas, a conglomeração é uma péssima escolha.

Em geral, o critério assenta na escolha de um método que dá o menor erro padrão (pré definido) a um custo fixo.

3.11 Sondagem multi-etápica

Esta classe de amostragem compreende as chamadas amostragens complexas.

3.11.1 Sondagem bi-etápica

A bi-etápica compreende duas etapas consecutivas: na primeira faz-se a extração de uma amostra aleatória (s_p) de m entre os M conglomerados ou unidades primárias e na segunda, em cada uma das unidades primárias selecionadas faz-se a seleção aleatória de uma amostra (s_i) de n_i indivíduos ou unidades secundárias, USAs que vão compor a amostra global.

Notações e fórmulas

U - População;
U_i - Unidades primárias de amostragem (UPA) ou subpopulações;
M - total de unidades primárias em que a população foi repartida;
m - número de unidades primárias incluídas na amostra;
N_i - dimensão das unidades primárias;
n - dimensão da amostra global (total de unidades secundárias na amostra global);
n_i - dimensão da amostra retirada na i–ésima UPA;
y_{ij} - valor da variável de interesse da j-ésima unidade secundária da i–ésima unidade primária.

A fração amostral da 1ª etapa, i. é, das UPAs é $f_i = \frac{m}{M}$ e das USAs é $f_{2i} = \frac{n_i}{N_i}$.

O processo de seleção de conglomerados de elemento pode ocorrer em mais do que duas etapas, caso multi-etápico.

3.12 Extração com probabilidades iguais sem reposição (PISR) nas duas etapas

O total, a média, a variância e a variância corrigida da variável de interesse são respetivamente,

$$\tau = \sum_{i=1}^{M} \tau_i = \sum_{i=1}^{M} \sum_{j=1}^{N_i} y_{ij} \qquad (3.12.1.1)$$

$$\mu = \frac{1}{N} \sum_{i=1}^{M} \sum_{j=1}^{N_i} y_{ij} \qquad (3.12.1.2)$$

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{M} \sum_{j=1}^{N_i} \left(y_{ij} - \mu\right)^2 \qquad (3.12.1.3)$$

$$\sigma'^2 = \frac{1}{N-1} \sum_{i=1}^{M} \sum_{j=1}^{N_i} \left(y_{ij} - \mu\right)^2 \qquad (3.12.1.4)$$

3.12.1 Extração com probabilidades iguais sem reposição (PISR) nas duas fases

Por simplificação, passar-se-á a usar i para designar a *i*–ésima UPA. O total, a média, a variância e a variância corrigida da variável de interesse na i–ésima UPA são respetivamente,

$$\tau_i = \sum_{j=1}^{N_i} y_{ij} \qquad (3.12.2.1)$$

$$\mu_i = \frac{1}{N_i}\sum_{j=1}^{N_i} y_{ij} \qquad (3.12.2.2)$$

$$\sigma_i^2 = \frac{1}{N_i}\sum_{j=1}^{N_i}\left(y_{ij} - \mu_i\right)^2 \qquad (3.12.2.3)$$

$$\sigma_i'^2 = \frac{1}{N_i - 1}\sum_{j=1}^{N_i}\left(y_{ij} - \mu_i\right)^2 \qquad (3.12.2.4)$$

O **total médio por unidade primária** (que é a média por conglomerado)

$$\mu_\tau = \frac{\tau}{M} = \frac{N\mu}{M} \qquad (3.12.2.5)$$

A média e a variância amostrais da mesma variável nessa UP são dadas por:

$$\bar{y}_i = \frac{1}{n_i}\sum_{j\in s_i} y_{ij}\,, \quad com \quad i \in s_p \qquad (3.12.2.6)$$

$$s_i^2 = \frac{1}{n_i - 1}\sum_{j\in s_i}\left(y_{ij} - \bar{y}_i\right)^2, \quad com \quad i \in s_p \qquad (3.12.2.7)$$

Como se sabe, as probabilidades de seleção das unidades desempenham um papel fundamental nas sondagens.

Define-se a probabilidade do *j*–ésimo elemento do i-ésimo conglomerado pertencer à amostra s como sendo:

$$\pi_j = \pi_i^{(1)}\pi_{j|i}, \qquad com\ j \in U_i, \qquad (3.12.2.8)$$

onde

- $\pi_i^{(1)}$ é a probabilidade de selecionar o i–ésimo conglomerado ou UPA e,

- $\pi_{j|i}$ é a probabilidade de selecionar a j–ésima USA na 2ª etapa dado que foi selecionada a i–ésima UPA na 1ª etapa.

O estimador de *Horvitz-Thompson* é sempre estimador centrado,

$$\hat{\tau} = \sum_{i \epsilon s_p} \sum_{j \epsilon s_i} \frac{y_{ij}}{\pi_j} \qquad (3.12.2.9)$$

onde y_{ij} é o valor do parâmetro Y associado à j–ésima US da i–ésima UPA e π_j, a probabilidade de inclusão na amostra da j–ésima US da i–ésima UPA.

3.12.2 Extração das UP com probabilidades proporcionais ao seu tamanho (PPS) e das US com dimensão constante

$$\pi_i^{(1)} = m\frac{N_i}{N} \qquad (3.12.14)$$

e, $n_i = n_0$,

$$\pi_{j|i} = \frac{n_0}{N_i} \qquad (3.12.15)$$

$$\pi_j = \pi_i^{(1)}\pi_{j|i} = m\frac{N_i}{N}\frac{n_0}{N_i} = \frac{mn_0}{N} = \frac{n}{N} \qquad (3.12.16)$$

e,

$$\hat{\tau} = N\bar{y}$$

A variância do estimador do total é

$$V(\hat{\tau}) = \frac{N}{m}\sum_{i=1}^{M} N_i(\mu_i - \mu)^2 + \frac{N}{n}\sum_{i=1}^{M}(N_i - n_0)\sigma_i'^2 \qquad (3.12.17)$$

O 1º termo corresponde à variabilidade entre as UPAs e o 2º à variabilidade entre as unidades secundárias de um mesmo conglomerado.

O estimador da variância do estimador do total é

$$\hat{V}(\hat{\tau}) = \frac{N^2}{m(m-1)} \sum_{i=1}^{m} (\hat{\mu}_i - \hat{\mu})^2 \qquad (3.12.18)$$

3.13 Estimação de variâncias em inquéritos por amostragem complexas

Pessoa e Silva (1998) referem que como toda a pesquisa por amostragem está sujeita a um grau de incerteza (erro de amostragem), as estimativas da variância dos parâmetros de interesse são indicadores de precisão, ou seja, de qualidade das estimativas. Tais estimativas da variância (ou em alternativa, dos desvios padrão ou coeficiente de variação) são empregues na construção de intervalos de confiança que têm a maior probabilidade de cobertura dos parâmetros de interesse, ou outras formas de inferência.

Dizem ainda que em muitas situações práticas de planos de sondagem complexos, quando se está em presença de estimadores não lineares, as probabilidades de inclusão conjunta podem ser nulas (amostragem sistemática) ou difíceis de estimar (na maioria dos casos de amostras com probabilidades de seleção desiguais), as equações obtidas para estimativas de totais no ponto 8 perdem sentido. Na verdade, os estimadores dos parâmetros de interesse são não lineares cujas expressões para enviesamento e variância são extremamente difíceis de obter ou mesmo impossível (casos de rácios, coeficientes de regressão, entre outros). Para superar estas dificuldades, são utilizadas técnicas apropriadas, sendo de destacar o Método do Conglomerado Primário (*Ultimate Cluster*, na língua inglesa), o Método de Linearização de Taylor e os Métodos de Reamostragem (a destacar Bootstrap e Jackknife).

Apresentar-se-á a seguir alguns métodos de estimação de variâncias em planos de sondagem complexos.

3.13.1 Linearização de Taylor para estimação de variância de estimadores não lineares

Särdal et al. (1992) e Pessoa e Silva (1998) referem que este método consiste em aproximar a função que define o estimador não linear à expressão do desenvolvimento da série de Taylor para encontrar expressões aproximada para a variância do estimador e do respetivo estimador da variância, bem como permite cálculos aproximados para intervalos de confiança para as estimativas. Esta técnica desde há muito tempo que tem sido empregue em muitas áreas da estatística.

Seja o problema de estimar o parâmetro populacional $\theta = (\theta_1, \theta_2, \dots, \theta_l, \dots, \theta_q)$ que pode ser expresso como uma função de q totais populacionais,

$$\theta = f(\tau_1, \tau_2, \dots, \tau_l, \dots, \tau_q) \qquad (3.13.1)$$

onde $\tau_l = \sum_{i \in U} y_{li} \quad (l = 1, 2, \dots, q)$. Admitindo que os valores das variáveis de interesse na população $Y_1, Y_2, \dots, Y_q$ são medidos para o i-ésimo indivíduo na amostra, $(y_{1i}, y_{2i}, \dots, y_{li}, \dots, y_{qi})'$, então:

$$\hat{\tau}_{l\pi} = \sum_s \check{y}_{li} = \sum_s \frac{y_{li}}{\pi_i} \qquad (3.13.2)$$

Assim, pode representar-se $\hat{\theta}$ como $\hat{\theta} = f(\hat{\tau}_{q\pi}) = f(\hat{\tau}_{1\pi}, \hat{\tau}_{2\pi}, \dots, \hat{\tau}_{l\pi}, \dots, \hat{\tau}_{q\pi})$

As propriedades de $\hat{\theta}$ permitem simplificar a sua expressão, se θ for linear, ou seja,

$$\theta = a_0 + \sum_{l=1}^{q} a_l \tau_l \qquad (3.13.3)$$

Que resulta num estimador centrado da forma:

$$\hat{\theta} = a_0 + \sum_{l=1}^{q} a_l \hat{\tau}_{li} \qquad (3.13.4)$$

cuja variância obtém-se de:

$$V(\hat{\theta}) = V\left(\sum_{l=1}^{q} a_l \hat{\tau}_{l\pi}\right) = \sum_{l=1}^{q} \sum_{l'=1}^{q} a_l a_{l'} Cov(\hat{\tau}_{l\pi}, \hat{\tau}_{l'\pi}), l \neq l', \quad l, l' = 1, 2, \dots, q \quad (3.13.5)$$

que é um estimador não enviesado da variância do estimador, estimador dado por:

$$\hat{V}(\hat{\theta}) = \sum_{l=1}^{q} \sum_{l'=1}^{q} a_l a_{l'} \widehat{Cov}(\hat{\tau}_{l\pi}, \hat{\tau}_{l'\pi}) \qquad (3.13.6)$$

Se $l = l'$ então $Cov\,(\hat{\tau}_{l\pi}, \hat{\tau}_{l'\pi}) = V(\hat{\tau}_{l\pi})$ e $\widehat{Cov}(\hat{\tau}_{l\pi}, \hat{\tau}_{l'\pi}) = \hat{V}(\hat{\tau}_{l\pi})$.

Se θ é uma função linear, existem equações alternativa mais simples, em termos de cálculo para as expressões de variância (5.1.5) e do seu estimador (5.1.6). A equação de variância pode ser expressa como:

$$\hat{\theta} = a_0 + \sum_s \breve{u}_i \qquad (3.13.7)$$

onde $\breve{u}_i = u_i/\pi_i$ com $u_i = \sum_{l=1}^{q} a_l y_{li}$, então a variância pode representar-se como:

$$V(\hat{\theta}) = V\left(\sum_s \breve{u}_i\right) = \sum\sum_U \Delta_{ij}\breve{u}_i\breve{u}_j \qquad (3.13.8)$$

cujo estimador é:

$$\hat{V}(\hat{\theta}) = \sum\sum_s \breve{\Delta}_{ij}\breve{u}_i\breve{u}_j \qquad (3.13.9)$$

Se θ for uma função não linear dos q totais, é impossível obter expressões exatas para o enviesamento e variâncias do estimador $\hat{\theta} = f(\hat{\tau}_{1\pi}, \hat{\tau}_{2\pi}, \dots, \hat{\tau}_{l\pi}, \dots, \hat{\tau}_{q\pi})$. Para contornar esta dificuldade, pode aplicar-se a técnica de linearização de Taylor que consiste em transformar o estimador não linear $\hat{\theta}$ através de uma aproximação assintótica a um pseudo-estimador $\hat{\theta}_0$ como função linear de $\hat{\tau}(\hat{\tau}_1, \hat{\tau}_2, \dots, \hat{\tau}_q)$ fácil de manipular, simplificando a estimação de variância, bem como a do seu estimador. $\hat{\theta}_0$ é considerado pseudo-estimador porque, geralmente, depende de certos fatores incógnitos, daí que não é um estimador verdadeiro. Quando se consegue boa aproximação, $\hat{\theta}_0$ pode representar bem $\hat{\theta}$ e podem ser usadas as expressões simples de variância $V(\hat{\theta}_0)$ e de seu estimador $\hat{V}(\hat{\theta}_0)$ como uma aproximação.

Aplicando a expansão de Taylor da primeira ordem para a função $f(\hat{\tau}_\pi)$ em torno de $\tau = (\tau_1, \tau_2, \dots, \tau_q)$, desprezando os restantes termo, obtendo-se

$$\hat{\theta} \cong \hat{\theta}_0 = \theta + \sum_{l=1}^{q} a_l(\hat{\tau}_{l\pi} - \tau_l) \qquad (3.13.10)$$

onde

$$a_l = \left.\frac{\partial f(\hat{\tau}_{1\pi}, \hat{\tau}_{2\pi}, \dots, \hat{\tau}_{q\pi})}{\partial \hat{\tau}_{l\pi}}\right|_{(\hat{\tau}_{1\pi}, \hat{\tau}_{2\pi}, \dots, \hat{\tau}_{q\pi}) = (\tau_1, \tau_2, \dots, \tau_q)} \qquad (3.13.11)$$

Quando existem as derivadas parciais de f dadas por a_0 e não são nulas, para amostras de grande dimensão, nas quais $\hat{\tau}_{1\pi}, \hat{\tau}_{2\pi}, \dots, \hat{\tau}_{q\pi}$ têm maiores probabilidades de assumir valores próximos de $\tau_1, \tau_2, \dots, \tau_q$, o estimador $\hat{\theta}$ assume um

comportamento aproximadamente igual ao da função $\hat{\theta}_0$. Tal aproximação varia em função da dimensão da amostra. Assume-se que o enviesamento e a variância de $\hat{\theta}$ possam ser aproximados pelas equações correspondentes da estatística linear de $\hat{\theta}_0$.

É frequente o uso de variância de uma estatística linear como aproximação de um estimador não linear complexo.

Seja $AV(\hat{\theta})$ a notação da variância aproximada de $\hat{\theta}$ correspondente à variância exacta da estatística linearizada $\hat{\theta}_0$. Seja também $u_i = \sum_{l=1}^{q} a_l y_{li}$. Segue que

$$AV(\hat{\theta}) = V(\hat{\theta}_0) = V\left(\sum_{l=1}^{q} a_l \hat{\tau}_{l\pi}\right) = V\left(\sum_{s} \frac{u_i}{\pi_i}\right) \qquad (3.13.12)$$

Basta lembrar que foi referido que $\hat{\tau}_{l\pi} = \sum_s \breve{y}_{li} = \sum_s \frac{y_{li}}{\pi_i}$

Da definição do $EQM(\hat{\theta})$ segue que:

$EQM(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = E[\hat{\theta} - f(\tau_1, \tau_2, \dots, \tau_q)]^2$, sendo que $E(\hat{\theta}_0) = 0$, então

$$EQM(\hat{\theta}) \cong EQM(\hat{\theta}_0) = V(\hat{\theta}_0) \qquad (3.13.13)$$

Nota-se que com a expressão (3.13.12) pode aproximar-se a variância de um estimador não linear através do estimador π-ponderado u_i sem grandes problemas. No entanto, a dificuldade reside no facto de que uma vez que as quantidades u_i dependem dos valores de $a_1, a_2, \dots, a_q$ que por sua vez dependem dos totais populacionais, que são desconhecidos, então u_i também são desconhecidos. Para contornar esta dificuldade, recorre-se, frequentemente à substituição de cada total desconhecido pelo respetivo estimador π obtendo-se assim também o estimador $\hat{a}_i$ de a_i que permite determinar, $\forall i \in s$, a variável

$$\hat{u}_i = \sum_{l=1}^{q} \hat{a}_l y_{li} \qquad (3.13.14)$$

O estimador de $V(\hat{\theta})$ segue

$$\hat{V}(\hat{\theta}) = \sum\sum_{s} \breve{\Delta}_{ij} \frac{\hat{u}_i \hat{u}_j}{\pi_i \pi_j} \qquad (3.13.15)$$

$\hat{V}(\hat{\theta})$ é uma função do estimador consistente $\hat{u}_i$ e para amostras grandes comporta-se como função dos verdadeiros valores desconhecidos u_i. Assim, $\hat{V}(\hat{\theta})$ assume-se que é consistente para $V(\hat{\theta})$.
Pode dizer-se em bom rigor que $\hat{V}(\hat{\theta})$ corresponde a um estimador de $AV(\hat{\theta})$. No entanto, uma vez que para amostras grandes $\hat{V}(\hat{\theta})$ comporta-se aproximadamente como $V(\hat{\theta})$, pode considerar-se $\hat{V}(\hat{\theta})$ como um estimador consistente de $V(\hat{\theta})$. Isto foi demonstrado em estudos, por simulações em diferentes situações, (Särdal et al., 1992, p. 175).

Com frequência, se o plano amostral é de dimensão finita, dispõe-se de um estimador de variância alternativo dado por $(\hat{V}_{SYG}(\hat{\tau}_\pi), visto\ no\ ponto\ 3.8)$,

$$\hat{V}(\hat{\theta}) = -\frac{1}{2}\sum\sum_{s} \breve{\Delta}_{ij}\left(\frac{\hat{u}_i}{\pi_i} - \frac{\hat{u}_j}{\pi_j}\right)^2 \qquad (3.13.16)$$

3.13.2 Método do conglomerado primário (*Ultimate cluster*)

Segundo Pessoa e Silva (1998), este método é sugestivo para estimação de variâncias de estimadores de totais e médias em situações complexas de amostragem que envolvem dois ou mais estágios. Toma em conta somente as variações dos dados das unidades contidas nas unidades primárias de amostragem, UPAs, ou seja, a nível dos conglomerados primários, supondo que foram retiradas em populações infinitas. Este método tem a vantagem da simplicidade e permite acomodar muitos planos amostrais que envolvem a estratificação e a seleção com probabilidades desiguais (com ou sem reposição). A sua utilização exige estimadores não enviesados dos totais da variável de interesse em cada um das unidades primárias de amostragem selecionados e que pelo menos dois destes estejam selecionados em cada estrato, caso a estratificação seja empregue no primeiro estágio).

A outra grande vantagem do método é que embora tenha sido proposto, originalmente, para estimar as variâncias dos estimadores dos totais, ele pode, sob certas condições, ser usado também para estimar variâncias dos estimadores de medidas de interesse que podem ser aproximadas a funções lineares de estimadores de totais.

O método do conglomerado primário consiste num plano de sondagem em vários estágios em que são selecionadas n_h unidades primárias de amostragem, UPAs no estrato *h (h=1, 2, …, H)*. Indicando por π_{hi} a probabilidade de inclusão da unidade primária de amostragem i do estrato *h*, na amostra e por $\hat{\tau}_{hi}$ o estimador não

enviesado do total da variável de interesse nesse conglomerado primário i. O estimador não enviesado do total populacional $\tau = \sum_h \sum_{i \in U_h} \tau_{hi}$ será definido por

$$\hat{\tau}_{CP} = \sum_{h=1}^{H} \sum_{i=1}^{U_h} \frac{\hat{\tau}_{hi}}{\pi_{hi}} \qquad (3.13.17)$$

Um estimador não enviesado da variância desse estimador é dado por

$$\hat{V}(\hat{\tau}_{CP}) = \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} \left(\frac{\hat{\tau}_{hi}}{\pi_{hi}} - \frac{\hat{\tau}_h}{n_h} \right)^2 \qquad (3.13.18)$$

onde

$$\hat{\tau}_h = \sum_{i=1}^{n_h} \frac{\tau_{hi}}{\pi_{hi}} \qquad (3.13.19)$$

com $h = 1, 2, \dots, H$

Não obstante o facto de que, na maioria de situações práticas, a seleção das unidades primárias ser sem reposição, o estimador de variância do método do conglomerado primário pode resultar em boa aproximação da correspondente variância. Esta propriedade decorre da eficiência das amostras sem reposição em relação às amostras com reposição com mesma dimensão. Devido à sua simplicidade e praticabilidade, este método é muito utilizado em amostragem para a estimação de variâncias de parâmetros de interesse com a devida adaptação.

Seja s uma amostra retirada de uma população em dois estágios, na qual são extraídas n_h UPAs do estrato h $(h = 1, 2, \dots, H)$ com probabilidade proporcional à dimensão, $\hat{\tau}_{hi}$ um estimador não enviesado do total da variável de interesse τ_{hi} na i-ésima UPA (conglomerado primário) do estrato h é definido por

$$\hat{\tau}_{hi} = \frac{N_{hi}}{m_{hi}} \sum_{i=1}^{M_{hi}} \sum_{j=1}^{m_{hi}} \tau_{hij} \qquad (3.13.20)$$

onde N_{hi} é a dimensão da i–ésima UPA do estrato h (conglomerado primário), m_{hi} é a dimensão da amostra s_{hi} (ou seja, número das USA's seleccionadas na i–ésima UPA do estrato h) e τ_{hij} é o valor da variável de interesse y da j-ésima USA selecionada na i-ésima UPA no estrato h.

Obtém-se o estimador não enviesado do total τ que é

$$\hat{\tau} = \sum_{h=1}^{H} \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{1}{P_{hi}} \frac{N_{hi}}{m_{hi}} \sum_{j=1}^{m_{hi}} \tau_{hij} \qquad (3.13.21)$$

O estimador não enviesado da variância do estimador do total é dado por

$$\hat{V}(\hat{\tau}) = \sum_{h=1}^{H} \frac{1}{n_h(n_h-1)} \sum_{i=1}^{n_h} \left(\frac{\hat{\tau}_{hi}}{P_{hi}} - \hat{\tau}_h \right)^2 \qquad (3.13.22)$$

onde $\hat{\tau}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{\hat{\tau}_{hi}}{P_{hi}}$, n_h é a dimensão da amostra s_h da primeira etapa e P_{hi} é a probabilidade da tiragem, num sorteio, da i-ésima UPA do estrato h.

Pode simplificar-se estas expressões representando o estimador em função dos pesos das unidades elementares, W_{hij}

$$W_{hij} = \frac{1}{n_h P_{hi}} \frac{N_{hi}}{m_{hi}} \qquad (3.13.23)$$

Nos planos de sondagem em que a tiragem dos conglomerados primários faz-se sem reposição, π_{hi} indica a probabilidade de inclusão do i-ésimo conglomerados primários do estrato h na amostra s_h e $\pi_{j|hi}$ a probabilidade de inclusão da j-ésima USA dado que a i-ésima UPA foi retirada do estrato h

$$\pi_{hi} = n_h P_{hi}, \qquad \pi_{j|hi} = \frac{m_{hi}}{N_{hi}}, \qquad \pi_{hij} = \pi_{hi}\pi_{j|hi} \qquad \text{então } W_{hij} = \frac{1}{\pi_{hij}} \qquad (3.13.24)$$

Assim, obtém-se o estimador ponderado do total

$$\hat{\tau} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \frac{1}{\pi_{hi}} \frac{1}{\pi_{j|hi}} \tau_{hij} \qquad (3.13.25)$$

e

$$\hat{V}(\hat{\tau}) = \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} (\hat{\tau}_{hi} - \bar{\tau}_h)^2 \qquad (3.13.26)$$

onde $\hat{\tau}_{hi} = \sum_{j=1}^{m_{hi}} w_{hij}\tau_{hij}$ e $\bar{\tau}_{hi} = \frac{1}{n_h} \sum_{i=1}^{n_h} \tau_{hi}, \quad h = 1, 2, \dots, H$

Facilmente, pode observar-se que os estimadores do total para planos de sondagem com ou sem reposição são coincidentes.

Ao representar o estimador do total em função dos pesos associados às unidades elementares, obtém-se uma equação generalizada para planos de sondagem com múltiplos estágios.

Ressalta-se que este método não será aplicado no caso do presente trabalho.

3.13.3 Métodos de Replicação

De acordo com Wolter (1985), os métodos de reamostragem de dados, também conhecidos por método de estimação por grupos aleatórios introduzidos por Mahalanobis (1939, 1946), que mais tarde Deming (1956) e outros autores atribuíram a designação de método de réplicas (*data resampling*, na língua inglesa) e, também referidos por Shao e Tu (1995), constituem os primeiros desenvolvimentos, atualmente muito aplicados na análise estatística para simplificar a estimação de variância em pesquisas amostrais complexas, em alternativa de simplificação com outros métodos, tais como o método de linearização descrito, anteriormente.

A ideia básica dos métodos de reamostragem consiste em construir uma amostra de dimensão n, representativa da população alvo, podendo gerar-se dela, K sub-amostras de dimensão $\frac{n}{K}$, cada uma seleccionada de forma independente e empregando o mesmo esquema amostral, onde K é o número de réplicas. Calcula-se a estimativa do parâmetro de interesse, separadamente, em cada uma das amostras e faz-se a combinação das estimativas de todas as amostras e, estimar a variância entre as diferentes estimativas.

O objetivo do uso de métodos de reamostragem resume-se na estimação ou aproximação da distribuição amostral e suas características. Se θ é o parâmetro de interesse, então:

$$\hat{\theta}_R = \frac{1}{K}\sum_{k=1}^{K}\hat{\theta}_k \qquad (3.13.27)$$

é o seu estimador não enviesado baseado na k-ésima réplica $(k = 1, 2, \dots, K)$, e

$$\hat{V}_R(\hat{\theta}_R) = \frac{1}{K(K-1)}\sum_{k=1}^{K}(\hat{\theta}_k - \hat{\theta}_R)^2 \qquad (3.13.28)$$

é um estimador não enviesado da variância do estimador da réplica, $\hat{\theta}_R$.

Como indicado, as réplicas são construídas de forma independente, isto é, as amostras são selecionadas com reposição e os estimadores obtidos $\hat{\theta}_R$ e $\hat{V}_R(\hat{\theta}_R)$ são não enviesados, independentemente do plano amostral usado. Estas propriedades tornam os métodos de reamostragem muito populares. Além disso, ao contrário do que acontece com a abordagem de linearização, os estimadores aos quais se aplicam as réplicas não precisam de ser expressos como funções de totais. Seu inconveniente é que a sua aplicação prática de forma exata é restrita, pois, geralmente, é pouco eficiente e mais caro selecionar K amostras independentes com o mesmo esquema, em comparação com a seleção directa de uma única amostra de dimensão n. Além disso, se o número de réplicas K for muito pequeno, o estimador de variância pode não ser consistente.

Na aplicação prática, a construção de réplicas a *posteriori* com o intuito de estimar variâncias em situações complexas é de simples aplicação, poderosa e flexível, por adequar-se a muitos esquemas amostrais e situações de interesse. Quando as réplicas são construídas após o inquérito por meio de repartição por sorteio da amostra investigada em K grupos mutuamente exclusivos, da mesma dimensão, elas são chamadas **réplicas dependentes** ou grupos aleatórios (***random groups***).

Em geral, a metodologia de estimação para um parâmetro populacional é a mesma como no caso de grupos aleatórios, as expressões dadas para o estimador da réplica e sua variância são também usadas nesse caso como uma aproximação, mas não possuem as mesmas propriedades que as réplicas independentes.

A repartição da amostra em grupos aleatórios a *posteriori* deve ter em conta o plano amostral e pode não ser aplicável em determinadas circunstâncias. Recomenda-se que a repartição seja feita considerando estratos e alocando unidades primárias inteiras. (Wolter, 1998 Pág. 31)

Quando as réplicas são compostas a *posterior*, é frequente empregar-se um estimador para o parâmetro θ baseado na amostra completa, designado por $\hat{\theta}$, e um estimador de variância

$$\hat{V}_{RK}(\hat{\theta}) = \frac{1}{K}\sum_{k=1}^{K}(\hat{\theta}_k - \hat{\theta})^2 \qquad (3.13.29)$$

mais conservador que o estimador de variância $\hat{V}_R(\hat{\theta}_R)$, visto acima. Uma vez que os estimadores dos grupos aleatórios não são independentes, o estimador $\hat{V}_R(\hat{\theta}_R)$ não é um estimador não enviesado de θ.

De entre os métodos de reamostragem destacam-se os métodos, jackknife introduzido por Quenouille (1949) e bootstrap introduzido por Efron (1979).

3.13.3.1 Estimação de variância pelo método de jackknife

De acordo com Shao e Tu (1995) e Wolter (1985), esta técnica foi introduzida por Quenouille (1949) para estimar o viés de um estimador no contexto de populações infinitas, Wolter diz ainda que o seu uso em populações finitas parece ter sido introduzido por Durbin (1959), na estimação de rácios.

Jackknife, enquanto método de reamostragem, baseia-se em repartir a amostra a *posteriori* em K grupos mutuamente exclusivos e de dimensão igual a $m = \frac{n}{K}$, sendo conveniente que n, m e K sejam inteiros. Em cada subamostra criada, calcula-se o

estimador de θ por omissão dos elementos do K–ésimo grupo, usando a mesma forma funcional que a do estimador $\hat{\theta}$, os ditos ***pseudo-estimadores*** dados por

$$\hat{\theta}_{(k)} = K\hat{\theta} - (K-1)\hat{\theta}_k \qquad (3.13.30)$$

A estimação da variância baseia-se na variabilidade entre as estimativas obtidas a partir das subamostras formadas e a partir da amostra global.

Neste caso, a estimação conhece dois caminhos alternativos, através de estimadores dados por:

$$\hat{V}_{J1}(\hat{\theta}) = \frac{1}{K(K-1)} \sum_{k=1}^{K} \left(\hat{\theta}_{(k)} - \hat{\theta}_J\right)^2 \qquad (3.13.31)$$

ou

$$\hat{V}_{(J2)}(\hat{\theta}) = \frac{1}{K(K-1)} \sum_{k=1}^{K} \left(\hat{\theta}_{(k)} - \hat{\theta}\right)^2 \qquad (3.13.32)$$

onde

$$\hat{\theta}_J = \frac{1}{K} \sum_{k=1}^{K} \hat{\theta}_{(k)} \qquad (3.13.33)$$

é um estimador pontual jackknife para θ, alternativo ao estimador da amostra inteira $\hat{\theta}$.

O estimador $\hat{V}_{J2}(\hat{\theta})$ é mais conservador que o estimador $\hat{V}_{J1}(\hat{\theta})$.

Para conseguir-se uma melhor precisão do estimador resultante, recomenda-se que o número de grupo K seja maior possível, pois a precisão aumenta com o aumento de K. É muito frequente aplicar a técnica igualando o número de grupos à dimensão da amostra, isto é, eliminando, portanto um dado da amostra original de cada vez, e recalcular o pseudo-estimador com base nas restantes observações. Esta regra deve ser usada considerando o número de unidades primárias na amostra quando o plano amostral é em múltiplos estágios, pois as UPAs devem sempre ser eliminadas com todas as unidades a elas pertencentes.

O método jackknife é recomendado para calcular a variância de estimadores não lineares, como é o caso de estimadores baseados numa ponderação de pos-estratificação ou de ajustamento por margens, em que não existe uma fórmula

específica para o cálculo de variância. Este método oferece estimadores com resultado semelhante ao dos estimadores comuns de variância quando usados para situações de estimadores lineares nos dados amostrais. Além disso, as suas propriedades são razoáveis para diversas situações de estimadores não lineares de interesse.

Supõe-se uma população de N unidades donde se extrai uma amostra de dimensão n, aplicando o esquema proposto. Seja π_i, a probabilidade de inclusão da i–ésima unidade na população. Lembrando o estimador HT para o total,

$$\hat{\theta} = \hat{Y} = \sum_{i=1}^{n} \frac{y_i}{\pi_i} \qquad (3.13.34)$$

Neste caso, os pseudo-estimadores tomam a forma

$$\hat{\theta}_k = k\hat{Y} - (k-1)\hat{Y}_k \qquad (3.13.35)$$

E a variância do estimador HT é definida como sendo

$$\hat{V}_j(\hat{\theta}) = \frac{1}{k(k-1)} \sum_{k=1}^{K} (\hat{\theta}_k - \hat{\theta})^2 \qquad (3.13.36)$$

Se $\pi_i = np_i, (i = 1, 2, \dots, n)$ e $k = n$, então,

$$\hat{V}_j(\hat{\theta}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\frac{y_i}{p_i} - \hat{Y}\right)^2 \qquad (3.13.37)$$

Quando $k < n$, o estimador da variância jackknife mantém, algebricamente em forma de valor esperado

$$E[v(\hat{\theta})] = E\left[\frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\frac{y_i}{p_i} - \hat{Y}\right)^2\right] \qquad (3.13.38)$$

onde os valores esperados são relativos ao plano amostral π. O estimador da variância jackknife age como se as amostras fossem selecionadas com probabilidades desiguais segundo um plano com reposição ao invés de um plano sem reposição.
Neste caso, o viés do estimador da variância é definido como:

$$B[v_1(\hat{\theta})] = \frac{[V(\hat{Y}) - V(\hat{Y}_{\pi ps})]n}{n-1} \qquad (3.13.39)$$

Isto significa que o viés do estimador da variância jackknife é $\frac{n}{n-1}$ vezes o ganho (ou perda) de precisão pelo uso de sondagem sem reposição.

Conclui-se que o estimador da variância jackknife é conservador na aplicação prática da sondagem (onde a amostragem π se mostra melhor do que a PPS sem reposição).

3.13.3.2 Estimação de variância pelo método Bootstrap

Esta técnica foi introduzida por Efron (1979) com a finalidade de obter aproximações de estimativas da variância e intervalo de confiança de menor amplitude. A técnica bootstrap é, até agora, pouco explorada para inquéritos por amostragem. Inicialmente, foi concebida para uso com dados independentes. Um problema básico, cuja resposta ainda não foi definitivamente dada, tem a ver com o modo como a técnica deve ser, corretamente modificada para acomodar características especiais de pesquisas amostrais, incluindo a não independência de dados provenientes de amostragem sem reposição e outras complexidades de planos e estimadores.

Para indicar como o método bootstrap funciona, seja dada uma amostra probabilística s, extraída de uma população U por meio de um plano amostral qualquer sem reposição. Seja θ o parâmetro de interesse, $\hat{\theta}$ sua estimativa, e $V(\hat{\theta})$ a estimativa da sua variância:

1. Usando os dados da amostra para construir uma população artificial U^*, assumindo imitar a real, sendo que a população U é desconhecida;
2. Construir um conjunto de amostras independentes "subamostras" ou "amostras bootstrap" da população artificial U^* através do mesmo plano amostral através do qual foi retirada a amostra s da população U. Independência significa que há reposição na população de cada amostra bootstrap antes da retirada da amostra seguinte. Para cada amostra bootstrap, calcula-se uma estimativa $\hat{\theta}_a^*$ $(a = 1, 2, \dots, A)$ da mesma maneira como $\hat{\theta}$ foi calculado;
3. A distribuição observada de $\hat{\theta}_1^*, \hat{\theta}_2^*, \dots, \hat{\theta}_A^*$ é considerada uma "estimativa" da distribuição amostral do estimador $\hat{\theta}$, e $V(\hat{\theta})$ é estimado por

$$\hat{V}_{Boot} = \frac{1}{A-1}\sum_{a=1}^{A}\left(\hat{\theta}_a^* - \hat{\theta}^*\right)^2 \qquad (3.13.40)$$

onde

$$\hat{\theta}^* = \frac{1}{A}\sum_{a=1}^{A}\hat{\theta}_a^* \qquad (3.13.41)$$

O método bootstrap não será objeto de análise no âmbito do presente trabalho.

4. Plano amostral do CAP

Para responder aos objetivos do estudo, ou seja, para ponderar a amostra selecionada, tem interesse fazer uma descrição detalhada do plano amostral empregue.

Na definição do plano amostral do CAP II Moçambique, no que diz respeito a médias e pequenas explorações agro-pecuárias, foram examinadas as informações contidas na base de dados do 3º Recenseamento Geral da População e Habitação (III RGPH), secção G, particularmente, as variáveis definidas como sendo as principais.

Assim, foram excluídas da base de dados do IIIRGPH, as áreas consideradas especiais, os alojamentos coletivos, bem como as áreas de enumeração com um efetivo de agregados familiares com atividade agro-pecuária inferior a 15.

Foi, paralelamente, utilizada a base cartográfica do III RGPH, organizada em diferentes gruas de subdivisão do território até à área de enumeração (AE). Uma AE corresponde à menor subdivisão geográfica do território dimensionada e delimitada pelo INE, ou seja, secção censitária.

Para as grandes explorações, foram usados os cadastros de explorações agrícolas, pecuárias, agro-pecuárias e aquícolas dos Serviços Distritais das Atividades Económicas.

Na indisponibilidade de um cadastro atualizado de agregados familiares residentes em unidades de alojamentos particulares permanentes ocupados, que permitissem uma amostragem directa, optou-se por um plano amostral bietápico. Para garantir a representatividade dos domínios geográficos de interesse, foi considerada a estratificação em três níveis, sendo o primeiro referente à repartição geográfica do País em 11 estratos naturais, as províncias, no segundo nível, a divisão de cada província em distritos, resultando deste modo em 148 estratos naturais, os distritos, e no terceiro nível, a partição do distrito em dois domínios, urbano e rural.

No primeiro estágio, foram considerados como estratos principais os distritos, tomando em conta os dois estratos, urbano e rural. Em cada estrato (distrito), foram sorteadas as áreas de enumeração (AEs) com uma probabilidade proporcional à dimensão (número de agregados familiares na AE), ou seja, **pps** – sistemática, sendo que as áreas de enumeração constituem as unidades primárias de amostragem, UPAs. O número de AEs por distrito é função do tamanho da população. A opção por **pps** visa tornar a amostra eficiente, isto é, atribuir maior peso às unidades com muita informação.

No segundo estágio, em cada AE selecionada, foi feito o sorteio de uma amostra aleatória de dimensão igual (10 agregados familiares, mas no ato da seleção foram extraídos 13, dos quais 3 agregados familiares de reserva para efeitos de substituição, em caso de necessidade para garantir a mesma probabilidade de inclusão), segundo

uma amostragem sistemática com probabilidades iguais. Os agregados familiares compõem as unidades secundárias de amostragem, USAs. As médias explorações foram observadas, exaustivamente, nas AEs selecionadas (conglomeração).

Das 10 unidades observadas em cada AE, foi ainda extraída uma subamostra de duas pequenas explorações para efeitos de medição das áreas de cultivo.

Uma vez que o Agregado familiar constitui a unidade de observação e, como não existisse uma base de sondagem atualizada de agregados familiares, dada a sua dinâmica demográfica, a operação de recolha de dados foi antecedida de uma operação de listagem visando atualizar o cadastro dos agregados familiares de modo a reduzir no máximo a taxa de não respostas, entre outros aspetos. Como foi, anteriormente referido, a seleção das áreas de enumeração nos distritos foi com proporção probabilística baseada no tamanho das AEs. Caso existissem áreas de enumeração com mais de 200 agregados familiares, deveriam ser repartidas, enquanto as que apresentassem menos de 50 agregados familiares deveriam ser agrupadas.

Em suma, o plano amostral obedeceu a dois estágios: i) estratificação das UPAs (AEs) e seleção, com probabilidade proporcional ao tamanho e ii) conglomeração para as médias explorações e seleção de uma amostra sistemática de igual dimensão das USAs (agregados familiares) com probabilidade igual para as pequenas explorações.

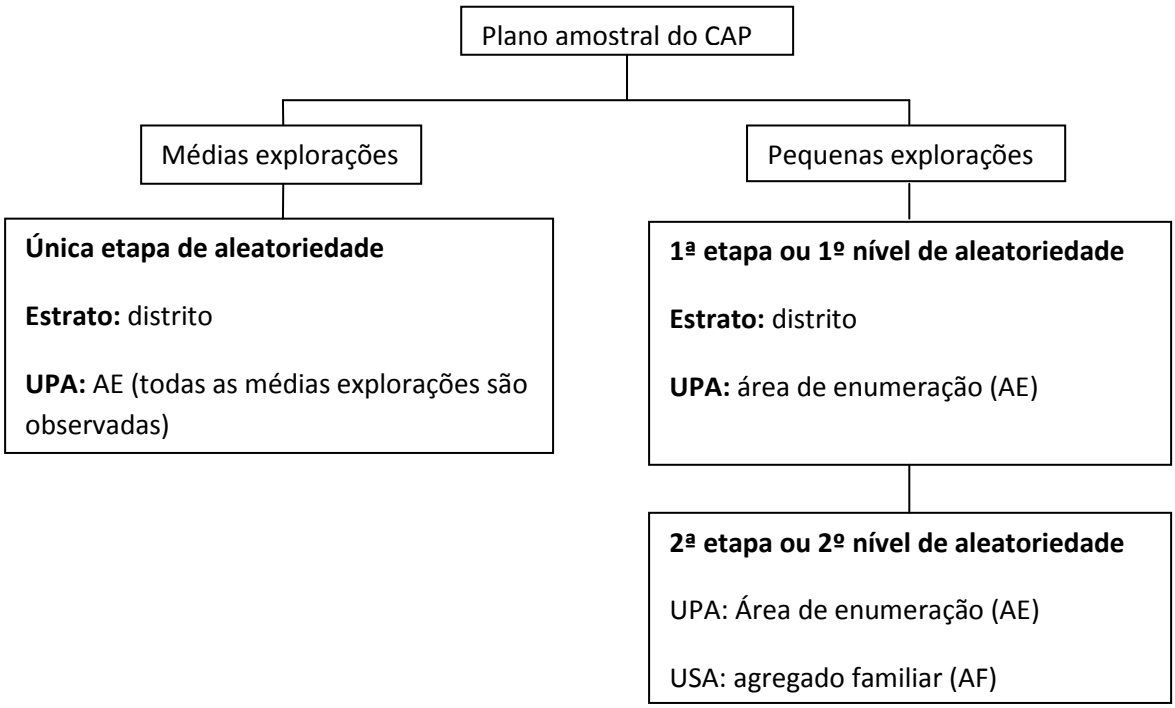

Fig. 4: Plano amostral do CAP

4.1 Definição da amostra

Para assegurar a precisão requerida com a confiança determinada, o cálculo da dimensão da amostra n foi feito considerando a escolha de m, o número de UPAs na 1.ª etapa e na 2.ª etapa, e os n_i, número de USAs em cada UPA selecionada.

Foram também consideradas experiências de censos similares, em que uma estimativa conservadora para o total da amostra consiste em 35.000 a 40.000 famílias agrícolas em cerca de 146 estratos (municípios ou distritos).

Para o cálculo da dimensão da amostra, parte-se da precisão absoluta:

$$d^2 = z^2_{1-\frac{\alpha}{2}}\sigma^2$$

onde σ^2 é a variância, nas duas etapas, do atributo que se pretende conhecer, ou seja, a soma da variância entre as médias das UPAs e a variância entre as unidades secundárias, dentro das UPAs. Uma vez que estas medidas são incógnitas, recorreu-se aos valores da pesquisa piloto.

De um universo de 44.576 áreas de enumeração agrícolas, das quais 10.084 urbanas e 34.492 rurais, foi extraída uma amostra total de dimensão *m* = 3.502 AEs, sendo 668 urbanas e 2834 rurais. A sua alocação pelos diferentes estratos (distritos) que define o número de unidades a selecionar em cada distrito no ato da definição da amostra global, teve em conta a necessidade de se obter estimativas distritais, provinciais, nacionais, urbanas e rurais fiáveis para as principais culturas, espécies animais entre outros itens, bem como considerações de custos. Depois da distribuição das áreas de enumeração pelos distritos foram feitos cálculos para a definição da dimensão da amostra da segunda etapa, que apontaram para amostras de dimensão constante de 10 agregados familiares.

Dado que as UPAs foram selecionadas segundo **pps** e em cada UPA, as USAs foram selecionadas com igual dimensão, isto é, $n_i = n_0$, então $n = mn_0$, o estimador de variância do total é

$$\hat{V}(\hat{\mu}) = \frac{1}{m(m-1)}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

onde

$$\bar{y}_i = \frac{\sum y_i}{n_0}$$

é a média amostral da i-ésima UP.

O estimador da variância corrigida σ'^2_τ é dado por:

$$s_\tau{}^2 = \frac{1}{m-1}\sum_{i \epsilon s_p}\left(\hat{\tau}_i - \frac{\hat{\tau}}{M}\right)^2$$

e chega-se a:

$$m = \frac{t^2_{1-\frac{\alpha}{2}}}{d^2} s^2_\tau$$

5. Metodologia de estimação

As variáveis do inquérito propostas para a análise no presente documento são:

1. População agrícola;
2. Proporção de mulheres;
3. Tamanho médio do agregado familiar;
4. Número de explorações chefiadas por mulheres;
5. Proporão de explorações chefiadas por mulheres;
6. Número de explorações chefiadas por pessoas analfabetas;
7. Número de explorações chefiadas por mulheres analfabetas;
8. Total de mão-de-obra familiar com atividade na exploração;
9. Proporção de explorações que usam mão-de-obra externa a tempo inteiro;
10. Total da área cultivada;
11. Proporção da área cultivada com culturas anuais;
12. Número de explorações que praticam a atividade agrícola;
13. Proporção de explorações com cajueiros;
14. Proporção de explorações com coqueiros;
15. Total de explorações que criam animais;
16. Número de explorações com bovinos;
17. Número médio de bovinos por exploração;
18. Total de explorações que criam suínos;
19. Proporção de explorações que criam suínos;
20. Número médio de suínos por exploração;
21. Proporção de explorações que criam aves;
22. Total de explorações que criam burros;
23. Número médio de burros por exploração;
24. Proporção de explorações com atividade aquícola;
25. Total de explorações com título de uso e aproveitamento de terra.

Importa fazer uma referência a alguns conceitos utilizados no CAP/INE, bem como outros que diferem um pouco dos conceitos habituais.

Foi considerado como agregado familiar, todo o grupo de pessoas ligadas ou não por laços de parentesco, que vivem na mesma casa e compartilham as mesmas refeições e a maior parte das despesas da casa. Agregado familiar agrícola foi utilizado no CAP para se referir a todo o agregado familiar que tivesse pelo menos uma pessoa que se dedicasse à atividade agropecuária por conta própria.

Chefe do agregado familiar foi considerada a pessoa responsável pelo agregado ou aquela que, para efeitos do recenseamento, foi indicada como tal pelos restantes membros. Em cada agregado familiar foi indicado sempre um chefe e devia ser uma

pessoa nele residente, podendo estar presente ou não no momento do censo, desde que a ausência fosse inferior a seis meses.

Assim, no presente trabalho designa-se população agrícola para se referir às pessoas residentes em agregados familiares agrícolas.

Como a questão sobre a posse do título de uso e aproveitamento de terra fosse colocada para cada parcela que o agregado familiar possuísse, neste trabalho considera-se que a exploração tem o título desde que tivesse pelo menos um título.

Para o cálculo das estimativas das variáveis acima propostas, bem como a respetiva precisão foi utilizada a programação macro CALJACK, desenvolvida em software SAS (cedido pelo INE-PT).

De acordo com o documento metodológico do INE-PT, a macro CALJACK consiste em utilizar a informação auxiliar para melhorar as estimativas. Permite expandir os resultados da amostra ponderando as observações, de modo a ajustá-los, por um número de variáveis qualitativas e/ou quantitativas, sobre os efetivos conhecidos sobre toda a população, utilizando para o efeito a informação auxiliar. Permite também calcular as estimativas por ajustamento para as variâncias das variáveis observadas.

No CALJACK pode calcular-se a precisão de um estimador do total, do quociente, da diferença de totais ou da diferença de quocientes, utilizando um estimador da variância do tipo jackknife.

5.1 Aspetos teóricos do ajustamento por margens

5.1.1 Objetivo

A solução teórica do problema caracteriza-se pela escolha de uma função distância, que minimiza as distâncias entre os pesos iniciais e os pesos finais ajustados, sujeita às condições do ajustamento. O ajustamento consiste em substituir os pesos iniciais das unidades amostrais pelos ponderadores finais ajustados, tão próximos quanto possível dos iniciais, utilizando uma determinada função distância que torna os efetivos ponderados após o ajustamento iguais aos efetivos conhecidos.

Seja $U = \{1, 2, \dots, K, \dots, N\}$ uma população de dimensão N, da qual foi selecionada uma amostra de n unidades e π_K a probabilidade de inclusão de qualquer elemento K de U.

Seja θ uma variável para a qual se pretende estimar o total populacional. O estimador de Horvitz-Thompson de θ é

$$\hat{\theta} = \hat{Y}_\pi = \sum_{K \in U} \frac{1}{\pi_k} y_k = \sum_{k \in S} d_k y_k \qquad (5.1)$$

O uso deste estimador não enviesado $\hat{Y}_\pi$ consiste em atribuir, a cada observação, um peso d_k igual ao inverso da sua probabilidade de inclusão.

Sejam $\gamma_1, \gamma_2, \dots, \gamma_j, \dots, \gamma_J$, J variáveis auxiliares conhecidas sobre a amostra s, da qual se conhecem os totais sobre a população:

$$X_J = \sum_{K \in U} x_{jk} \qquad (5.2)$$

Considerando esta informação, vai estimar-se o valor de θ, o total de por meio de um estimador da forma

$$\hat{Y}_w = \sum_{k \in s} w_k y_k$$

em que, os pesos finais ajustados w_k, atribuídos a cada observação são tão próximos quanto possível dos pesos iniciais , no sentido de uma função específica de distância a definir, que verifica as equações de ajustamento

$$\sum_{k \in U} w_k x_{jk} = X_j, \qquad \forall j = 1, 2, \dots, J \qquad (5.3)$$

O problema resume-se à procura de um estimador "semelhante" ao estimador de Horwitz-Thompson que ajuste a amostra aos totais de variáveis auxiliares.

5.1.2 Resolução teórica

Seja G a "função distância" de argumento w_k/d_k que vai medir as distâncias entre os w_k e os d_k, a função deve ser positiva e convexa, $G(1) = G'(1) = 0$.

Uma vez escolhida a função, o problema consiste em determinar os pesos w_k $(k \in s)$ que são soluções de

$$Min \sum_{k \in s} d_k G(w_k/d_k) \qquad (5.4)$$

sob as restrições de ajustamento $\sum_{k \in s} w_k X_k = \mathbf{X}$, ou seja, minimizar uma soma ponderada pelos d_k das distâncias entre os pesos de amostragem d_k e as ponderações

procuradas w_k, sob as restrições de ajustamento, em que, X_k é o vetor linha constituído pelos valores da observação k que tomam as variáveis auxiliares e **X** o vetor linha das margens de ajustamento.

Resolve-se este problema introduzindo um vetor de multiplicadores de Lagrange $\lambda' = (\lambda_1, \lambda_2, \dots, \lambda_J)$, em que o Langrangiano é dado por:

$$L = \sum_{k \in s} d_k G(w_k/d_k) - \lambda'^{(\sum_{k \in s} w_k x_k - X)} \qquad (5.5)$$

Para minimizar a função em causa, deriva-se L e iguala-se a zero, resultando:

$$w_k = d_k F(x_k' \lambda) \qquad (5.6)$$

em que, F é uma função inversa da derivada da função G.

O vetor λ é calculado mediante a resolução do sistema não linear de J equações a J incógnitas determinado pelas equações de ajustamento:

$$\sum_{k \in s} d_k \ F(x_k' \lambda) x_k = X \qquad (5.7)$$

Pode resolver-se numericamente este sistema pelo método iterativo de Newton, calcula-se uma série de vetores $\lambda^{(i)}$ definida por uma relação de recorrência, inicializando o algoritmo com o vetor $\lambda^{(0)} = 0$. A convergência é obtida quando o máximo das diferenças (em valor absoluto) entre as relações dos pesos w_k/d_k, obtidas em duas iterações sucessivas é suficientemente pequeno, inferior a um determinado limiar fixado previamente, ou seja, quando

$$Max \left| \frac{w_k^{(i+1)}}{d_k} - \frac{w_k^{(i)}}{d_k} \right| < \varepsilon \qquad (5.8).$$

5.1.3 As funções distância G disponíveis na macro

Para cada um dos métodos utilizáveis indica-se a função $G(x)$ (em que $x = w_k/d_k$)e a função $F(u)$ (em que $u = x_k' \lambda$). Assim, estão disponíveis 7 métodos a saber, o linear, o ranking ratio, o logit, o linear truncado, o $\frac{1}{1+x}$, o qui-quadrado e o Hellinger, cada um a adequar-se à natureza dos dados e da análise que se pretende.

No caso presente foi usado o método logit

$$G(x) = \begin{cases} \left((x-L)Log\frac{x-L}{1-L} + (U-x)Log\frac{U-x}{U-1}\right)\frac{1}{A}, se\ L < x < U \\ \left((U-L)Log\frac{U-L}{U-1}\right)\frac{1}{A}, se\ x \leq L \\ \left((U-L)Log\frac{U-L}{1-L}\right)\frac{1}{A}, se\ x \geq U \end{cases} \quad (5.9)$$

Em que $A = \frac{U-L}{(1-L)(U-1)}$ e $F(u) = \frac{L(U-1)+U(1-L)exp(Au)}{U-1+(1-L)exp(Au)} \in IR$

A forma logística da função dá o nome a este método que se pode também caracterizar como sendo um método ranking ratio truncado nas duas extremidades, de modo que as relações entre os pesos ajustados e os pesos iniciais (w_k/d_k) sejam limitados inferiormente por L e superiormente por U.

No caso presente, foi utilizada a variável "***pesoin***" para designar o peso inicial das explorações.

O processo de ponderação dos pesos consistiu em:

- Ponderar a amostra selecionada de tal forma que represente a população objeto de estudo, caso as respostas fossem completas (ponderação do desenho amostral);
- Compensar as não-respostas, mediante o ajuste de pesos do desenho inicialmente obtido para as unidades da amostra respondente;
- Calibrar os pesos da amostra respondente para que os totais amostrais extrapolados coincidam com os totais populacionais conhecidos, para minimizar o enviesamento devido a não-respostas.

Tendo em conta as expressões (3.11.14) a (3.11.16) tem-se:

a) Para as médias explorações (nas UPAs onde existam as médias explorações):

$$Pesoin_{ME} = w_{hi} = \frac{1}{\pi_i^{(1)}} = \frac{1}{m\frac{N_i}{N}} = \frac{N}{mN_i} \quad (5.10)$$

Onde, $N = \sum_{i \in h} N_i$ - dimensão do estrato h (distrito), ou seja, o número total de agregados familiares no distrito, N_i - dimensão da i-ésima UPA do estrato h, isto é, o número de agregados familiares na UPA e m - o número de UPAs selecionadas para a amostra no estrato h (w_1 é, exatamente a probabilidade de

seleção da UPA, uma vez que todas as médias explorações na UPA devem ser observados).

Os pesos dos conglomerados assim calculados respeitam a extração de 5 020 UPAs. Dado que uma das UPAs não foi observada devido ao seu desaparecimento, houve necessidade de se corrigir os ponderadores dos 5 019 conglomerados observados de modo a compensar a perda observada, mediante uma correção de não-resposta simples por meio de

$$w_{hi}^{*} = \frac{m_i}{m_i^{(r)}} w_{hi} \qquad (5.11)$$

onde m_i- dimensão da amostra (número de UPAs selecionadas) no estrato h e $m_i^{(r)}$- dimensão das unidades respondentes da mesma amostra.

b) Para as pequenas explorações:

$$Pesoin_{PE} = w_{hij} = \frac{N_{hi}}{n_{hi}} w_{hi}^{*} \qquad (5.12)$$

onde

N_{hi} - total de agregados familiares encontrados considerados para a seleção da amostra da i–ésima UPA do estrato h

n_{hi} - número de agregados familiares selecionados para a amostra da i–ésima UPA do estrato h

O ajuste de não-respostas foi mediante a expressão

$$w_{hij}^{*} = \frac{n_{hi}}{n_{hi}^{(r)}} w_{hij} \qquad (5.13)$$

onde, $n_{hi}^{(r)}$- representa o número de agregados familiares respondentes.

Dado que se trata de extração PPS das UPAs e de tiragens de dimensão constante das USAs ($n_i = n_0, i \in s_P \implies n = mn_0$), e conforme referido no capítulo 3, as probabilidades de inclusão das unidades primárias são dadas por

$$\pi_i^{(1)} = m\frac{N_i}{N}$$

e

$$\pi_{j|i} = \frac{n_0}{N_i}$$

De acordo com a equação (3.11),

$$\pi_j = \pi_i^{(1)} \pi_{j|i} = m \frac{N_i}{N} \frac{n_0}{N_i} = m \frac{n_0}{N} = \frac{n}{N} \qquad (5.14)$$

Como foi anunciado no ponto 3.8, para o cálculo das variâncias das variáveis propostas foram usadas as expressões baseadas em expressões dos totais,

$$\hat{\tau} = \frac{N}{m} \sum_{i=1}^{m} \hat{\mu}_i = \frac{N}{m} \sum_{i \in S_p} \bar{y}_i = N\bar{y}$$

A variância do total é estimada por

$$\hat{V}(\hat{\tau}) = \frac{N^2}{m(m-1)} \sum_{i \in S_P} (\hat{\mu}_i - \hat{\mu})^2 \qquad (5.15)$$

De

$$\hat{\tau} = N\hat{\mu} \implies \hat{\mu} = \frac{\hat{\tau}}{N}$$

chega-se a expressões para o cálculo das estimativas da variância da média e da proporção, dado que a proporção é um caso particular da média

$$\hat{V}(\hat{\tau}) = N^2 V(\hat{\mu}) \implies \hat{V}(\hat{\mu}) = \frac{\hat{V}(\hat{\tau})}{N^2} = \frac{1}{m(m-1)} \sum_{i \in S_P} (\hat{\mu}_i - \hat{\mu})^2 \qquad (5.16)$$

Para o cálculo do ***deff***, foram usadas as expressões para estimar as variâncias de cada variável Y em SAS vistas no ponto 3.4.

Estimador da variância do total (3.4.21)

$$\hat{V}(\hat{\tau}_y) = N^2 \left(\frac{N-n}{N}\right)\left(\frac{s_y^2}{n}\right) = \frac{N(N-n)}{n} s_y^2 \qquad (5.17)$$

Estimador da variância da média (3.4.16)

$$\hat{V}(\hat{\mu}_y) = \hat{V}(\bar{y}) = \left(\frac{N-n}{N}\right)\left(\frac{s_y^2}{n}\right) = \frac{N-n}{Nn} s_y^2 \qquad (5.18)$$

Estimador da variância da proporção (3.4.18)

$$\hat{V}(\hat{P}_y) = \left(\frac{N-n}{N}\right)\left[\frac{P_y(1-P_y)}{n-1}\right] = \frac{N-n}{N(n-1)}P_y(1-P_y) \qquad (5.19)$$

6. Resultados

O Quadro 1 mostra a distribuição dos ponderadores dos dados amostrais. Esses ponderadores foram calculados com base nas probabilidades de inclusão dos indivíduos na amostra, bem como no ajustamento para compensar as não respostas.

Os ponderadores têm utilidade na estimação dos parâmetros populacionais no processo de expansão dos dados da amostra, multiplicando-se cada observação pelo respetivo ponderador.

Como pode ser observado, os pesos individuais na amostra variam substancialmente entre 1,50 e 751,99. A razão entre estes dois valores é de cerca de 501 vezes, enfatizando a grande diferença entre eles.

A grande diferença de pesos deve-se essencialmente às probabilidades de inclusão das unidades na amostra.

Quadro 1: Distribuição dos ponderadores da amostra do CAP

Cod.	Província	Mínimo	Quartil (Q1)	Mediana	Quartil (Q3)	Máximo
01	Niassa	2,08	30,04	48,55	76,02	331,33
02	Cabo delgado	3,38	33,99	51,64	78,99	458,82
03	Nampula	5,20	57,01	82,11	111,70	373,26
04	Zambézia	5,57	70,78	98,44	130,80	462,59
05	Tete	2,85	34,43	70,65	115,47	563,24
06	Manica	2,80	35,42	57,99	94,34	751,99
07	Sofala	3,67	45,33	68,37	106,78	378,94
08	Inhambane	1,50	21,16	35,54	53,89	243,93
09	Gaza	1,75	24,21	53,79	81,18	466,49
10	Maputo Província	2,34	31,13	55,72	98,59	685,29
11	Maputo Cidade	1,91	19,98	34,39	82,80	286,68

Desta forma, fica patente a importância de se incluir na análise de dados a informação sobre os pesos das observações dado que a variabilidade dos ponderadores tem reflexos na estimação pontual, bem como na estimação das variâncias desses estimadores que para além do impacto dos pesos são também influenciados por aspetos do plano amostral, nomeadamente estratificação e conglomeração.

Quadro 2: Resultados

Nº	Parâmetro	Estimativa	CV	*deff*	IC
1	População agrícola	17 468 845	0,8695	0,3952	[17 404 580; 17 533 111]
2	Proporção de mulheres	0,5255	0,5692	1,4562	[0,5197; 0,5314]
3	Tamanho médio do AF	4,8368	0,8695	8,2030	[4,7558; 4,9179]
4	Nº de explorações chefiadas por mulheres	998 499	2,6967	0,5135	[987 311,88815; 1 009 686,2650]
5	Proporção de explorações chefiadas por mulheres	0,2765	2,6967	10,6588	[0,2624; 0,2906]
6	Nº de explorações chefiadas por pessoas analfabetas	1 734 673	1,7994	0,6677	[1 720 446,283; 1 748 900,219]
7	Nº de explorações chefiadas por mulheres analfabetas	714 454	3,1795	0,5575	[704 177,9206; 724 729,7398]
8	Total de mão-de-obra familiar com atividade na exploração	808 767	14,4521	0,2007	[788 012,6933; 829 522,2122]
9	Proporção de explorações que usam mão-de-obra externa a tempo inteiro	0,0724	8,3788	20,1996	[0,0612; 0,0837]
10	Total de área estimada (ha)	5 111 048	2,0941	0,3486	[5 075 089,424; 5 147 007,462]
11	Proporção de área cultivada com culturas anuais	0,9160	0,6169	17,1256	[0,9049; 0,9271]
12	Nº de explorações com atividade agrícola	3 379 687	0,3203	0,2859	[3 374 844,596; 3 384 529,862]
13	Proporção de explorações com cajueiros	0,3549	3,1402	24,4450	[0,3321; 0,3778]
14	Proporção de explorações com coqueiros	0,2264	3,3826	23,7050	[0,2067; 0,2461]
15	Total de explorações que criam animais	2 442 020	1,3107	0,7593	[2 428 081,9704: 2 455 957,1400]
16	Total de explorações com bovinos	190 223	41,4403	0,0034	[189 585,0841; 190 860,8897]
17	Nº médio de bovinos por exploração	5,6975	34,9337	850,8332	[1,1387; 10,2564]
18	Total de explorações que criam suínos	397 680	4,7983	0,4395	[389 736,1334; 405 622,9549]
19	Proporção de explorações com suínos	0,1101	4,7983	10,9702	[0,1001; 0,1201]
20	Nº médio de suínos por exploração	3,0613	3,9108	12,4063	[2,8523; 3,2703]
21	Proporção de propriedades que criam aves	0,6185	1,4432	15,5487	[0,6000; 0,6370]
22	Total de explorações com burros	8 818	100,0288	0,0199	[8 460,3327; 9 176,1100]
23	Nº médio de burros por exploração	1,9482	0,0000	0,0000	
24	Proporção de explorações com atividade aquícola	0,0011	36,7221	8,8534	[0,0001; 0,0020]
25	Total de explorações com título de uso e aproveitamento de terra	89 807	10,4576	0,3405	[86 341,6483; 93 271,5249]

O Quadro 2 mostra os resultados do efeito do plano amostral (*deff*).

A estimação de variância (alternativamente de desvio padrão ou coeficiente de variação) e os testes de hipótese desempenham um papel muito importante em estudos analíticos, visto que na inferência estatística, para além das estimativas pontuais, é necessário transmitir a ideia de precisão associada a essas estimativas através da amplitude dos intervalos de confiança a elas associadas ou através dos valores do desvio padrão (ou em alternativa do coeficiente de variação) que também permitem testar hipóteses relativas aos parâmetros dos modelos. As medidas de precisão permitem avaliar a fiabilidade das estimativas a partir das observações de uma amostra, sendo o coeficiente de variação a medida mais comum, dado que é uma medida relativa, isto é, não leva em consideração a unidade de medida. Embora não exista uma regra consensual a respeito da interpretação desta medida, sabe-se que quanto menor for o valor do CV, melhor é a precisão da estimativa.

Dado que não é possível determinar a verdadeira variância, porque não se tem acesso a toda a população e nem a todas as amostras possíveis, a alternativa é usar a estimativa da variância a partir das observações amostrais.

O efeito do plano amostral como indicador de eficiência dos planos amostrais através da eficiência dos estimadores dos planos de sondagem envolvidos está apresentado no Quadro 2. Como foi dito, valores de *deff* iguais a 1 significam que usar um plano de sondagem complexo ou aleatório simples o impacto é o mesmo, mas quando os valores de *deff* se distanciam de 1 revelam que menosprezar o plano de sondagem efetivamente adotado no ato de estimação de variâncias leva a estimativas enviesadas e comprometem os resultados da pesquisa.

Em suma, foi demonstrado que a inferência estatística é influenciada, por um lado pelos ponderadores individuais das observações, e por outro pelo efeito da estratificação e conglomerado, dado que os ponderadores das observações tem impacto tanto nas estimativas pontuais, quanto na estimação das variâncias dessas estimativas o que revela a necessidade de se incluir na análise informações dos pesos assim como dos detalhes do plano amostral.

Quadro 3: Resumo das estimativas das variáveis propostas

Nome	_NAME_	Niassa	C. Delgado	Nampula	Zambézia	Tete	Manica	Sofala	Inhambane	Gaza	Maputo Prov	Maputo Cid.	Moçambique
Popul. Agrícola	ES_T1	1 008 044	1 488 433	3 421 998	3 771 109	1 583 498	1 323 232	1 484 033	1 265 956	1 163 321	658 762	300 459	17 468 845
Prop. Mulheres	ES_R1	0,5286	0,5224	0,5238	0,5220	0,5146	0,5174	0,5238	0,5432	0,5447	0,5335	0,5307	0,5255
Tamanho med. Agr. Familiar	ES_R2	4,9415	4,4548	4,3432	4,7572	4,8035	5,5554	5,6355	4,8105	5,3301	5,0252	6,1055	4,8368
Expl. Chef. Mulheres	ES_T4	62 747	109 115	182 309	211 658	84 207	58 585	67 721	89 958	75 987	40 606	15 606	998 499
Prop. Expl. Chef. Mulheres	ES_R3	0,3076	0,3266	0,2314	0,2670	0,2554	0,2460	0,2572	0,3418	0,3482	0,3097	0,3171	0,2765
Expl. Chef. Pes. Analfabetas	ES_T5	112 219	198 561	425 667	403 396	165 715	95 570	102 809	105 351	86 641	33 787	4 957	1 734 673
Expl. Chef. Mulh. Analfabetas	ES_T6	48 310	84 098	138 598	164 895	66 268	42 971	47 826	57 151	43 317	17 490	3 529	714 454
Mão-ob. Famil. c/ activ. Expl.	ES_T7	37 581	41 359	134 377	73 271	169 152	124 650	108 141	80 149	29 965	8 709	1 414	808 767
Prop. Expl. c/ Mão-ob ext. tempo int.	ES_R4	0,0940	0,0569	0,0567	0,0686	0,0695	0,1292	0,0860	0,0292	0,1233	0,0663	0,0964	0,0724
Tot. área cultiv. (ha)	ES_T9	361 294	467 262	952 232	995 273	524 470	499 678	440 491	390 582	344 497	108 154	27 115	5 111 048
Prop. Área cultiv. c/ cult. anuais	ES_R5	0,9253	0,8858	0,9245	0,9175	0,9569	0,8746	0,9378	0,9377	0,8817	0,8946	0,7846	0,9160
Expl. c/ activ. Agrícola	ES_T12	192 421	314 155	751 382	771 396	321 062	226 440	245 672	237 511	197 414	94 331	27 903	3 379 687
Prop. Expl. c/ cajueiros	ES_R7	0,0592	0,3084	0,4879	0,3715	0,0050	0,1089	0,2885	0,7706	0,5707	0,3568	0,2075	0,3549
Prop. Expl. c/ coqueiros	ES_R8	0,0068	0,2809	0,1793	0,2970	0,0060	0,0105	0,2688	0,6819	0,2654	0,1528	0,2632	0,2264
Expl. Criam animais	ES_T15	123 538	215 756	476 733	542 538	240 025	184 457	187 748	214 875	162 461	72 997	20 892	2 442 020
Expl. C/ bovinos	ES_T16	1 615	279	9 450	785	46 164	27 676	5 006	43 964	45 557	8 403	1 323	190 223
Nº méd. bovinos/expl.	ES_R9	7,7671	15,0519	4,1646	6,6582	6,0666	5,3717	6,9997	4,0556	6,3697	9,0179	10,9140	5,6975
Expl. criam suínos	ES_T18	5 394	18 944	49 233	84 462	54 753	17 410	19 603	97 334	40 151	7 547	2 847	397 680
Prop. Expl. c/ suínos	ES_R10	0,0264	0,0567	0,0625	0,1065	0,1661	0,0731	0,0744	0,3699	0,1840	0,0576	0,0579	0,1101
Nº med. suínos/expl.	ES_R11	3,3225	3,2012	2,8205	2,7261	3,4583	4,0005	4,8618	2,5185	3,1485	3,7604	5,4429	3,0613
Prop. Expl. c/ aves	ES_R12	0,5689	0,5727	0,5178	0,6544	0,6247	0,7428	0,6952	0,7687	0,6632	0,5168	0,3870	0,6185
Expl. criam burros	ES_T21	5	0	0	183	1 067	509	0	3 820	3 042	170	23	8 818
Nº med. burros/expl.	ES_R13	2,0000			4,0000	2,6384	3,0578		1,7059	1,6959	2,1776	1,0000	1,9482
Prop. Expl. c/ aquacultura	ES_R14	0,0030	0,0000	0,0004	0,0013	0,0016	0,0030	0,0004	0,0010	0,0000	0,0020	0,0003	0,0011
Expl. c/ titulo de uso de terra	ES_T24	2 306	3 191	15 904	16 905	4 354	3 352	6 918	2 838	7 672	14 851	11 515	89 807

Quadro 4: Estimativas dos coeficientes de variação das variáveis propostas

NAME		nias	cabo	namp	zamb	tete	mani	sofa	inha	gaza	mapp	mapc	Pais
Popul. Agrícola	CV_T1	1,122366	1,636184	1,351561	1,198047	2,384667	2,152038	0,770463	1,346150	0,367404	1,382623	1,104899	0,869463
Prop. Mulheres	CV_R1	0,702143	0,768811	0,609819	0,937968	1,469263	1,032956	0,494195	0,692376	0,218809	0,621837	0,700121	0,569195
Tamanho med. Agr. Familiar	CV_R2	1,122366	1,636184	1,351561	1,198047	2,384667	2,152038	0,770463	1,346150	0,367404	1,382623	1,104899	0,869463
Expl. Chef. Mulheres	CV_T4	3,758630	4,127035	2,461500	4,395237	5,001850	4,402313	2,866065	4,500898	1,132958	3,846461	3,999076	2,696682
Prop. Expl. Chef. Mulheres	CV_R3	3,758630	4,127035	2,461500	4,395237	5,001850	4,402313	2,866065	4,500898	1,132958	3,846461	3,999076	2,696682
Expl. Chef. Pes. Analfabetas	CV_T5	1,736512	3,305712	2,405704	3,806984	12,426480	5,983215	1,661869	2,642405	0,797985	3,399431	2,935043	1,799389
Expl. Chef. Mulh. Analfabetas	CV_T6	4,349567	5,721837	3,317626	6,138799	13,661676	8,296032	3,353133	5,449894	1,422734	4,881233	4,636592	3,179491
Mão-ob. Famil. c/ activ. Expl.	CV_T7	15,158116	14,592177	8,035767	7,907554	57,080033	25,390620	8,820195	14,136290	3,462690	9,451039	8,284981	14,452098
Prop. Expl. c/ Mão-ob ext. tempo int.	CV_R4	11,626918	7,396873	11,249867	8,177155	15,655761	13,062819	7,839701	10,309073	3,020359	8,449069	8,818085	8,378841
Tot. área cultiv. Estimada (ha)	CV_T9	2,477949	2,624983	2,697421	3,582960	10,004308	6,230116	1,716828	2,681283	0,817507	2,106307	2,442686	2,094105
Prop. Área cultiv. c/ cult. anuais	CV_R5	1,138792	0,832784	0,734186	1,042911	4,626990	1,364754	0,455738	0,591155	0,237677	0,603117	0,420402	0,616946
Expl. c/ activ. Agrícola	CV_T12	0,606472	0,855007	0,900025	0,929750	4,276854	3,492182	0,357852	0,774804	0,213929	0,984510	0,549331	0,320293
Prop. Expl. c/ cajueiros	CV_R7	3,971093	3,177299	1,414425	13,186628	9,667141	7,997786	2,024735	11,071073	1,179870	4,780576	37,640819	3,140153
Prop. Expl. c/ coqueiros	CV_R8	5,082605	6,850899	2,234555	24,592621	9,926518	9,206546	4,274359	46,404708	1,631814	5,855890	35,360025	3,382577
Expl. Criam animais	CV_T15	2,070310	1,459150	1,126376	1,714251	5,245564	3,857716	1,391104	2,185535	0,537589	1,356589	1,528560	1,310694
Expl. C/ bovinos	CV_T16	52,776502	5,670604	4,894149	10,112025	18,828199	11,867332	15,154642	27,303142	2,887406	16,508521	5,698181	41,440265
Nº méd. bovinos/expl.	CV_R9	61,026704	4,273322	2,598693	4,608018	18,504388	8,249797	10,461695	12,125356	1,960839	8,490422	4,102392	34,933656
Expl. criam suínos	CV_T18	11,837633	8,223784	3,607744	12,076547	23,293501	13,234088	6,475799	23,208295	2,217527	9,954790	6,242508	4,798317
Prop. Expl. c/ suínos	CV_R10	11,837633	8,223784	3,607744	12,076547	23,293501	13,234088	6,475799	23,208295	2,217527	9,954790	6,242508	4,798317
Nº med. suínos/expl.	CV_R11	7,958620	5,120119	2,386103	8,933729	25,616489	10,111541	6,455538	15,933402	1,777152	6,610033	5,163548	3,910830
Prop. Expl. c/ aves	CV_R12	2,380411	1,896100	1,278966	1,774992	5,587627	4,053009	1,644017	2,396962	0,609304	1,418287	1,956577	1,443246
Expl. criam burros	CV_T21		20,643135	18,671915	39,514608	68,428262	53,947291		100,103509	11,981870		33,898272	100,028790
Nº med. burros/expl.	CV_R13		11,214565	8,577529	26,633807	0,000000	13,695558		0,000000	6,625187		18,511055	0,000000
Prop. Expl. c/ aquacultura	CV_R14			77,451340	39,172390	100,242678	58,964376	55,078751	40,683739	17,310717	72,575334	45,139532	36,722088
Expl. c/ titulo de uso de terra	CV_T24	19,572677	14,522928	15,736546	20,738656	8,904243	11,174037	10,879873	24,366019	4,311962	19,086329	19,185051	10,457629

Quadro 5: *DEFF*

		Niassa	C. Delgado	Nampula	Zambézia	Tete	Manica	Sofala	Inhambane	Gaza	Maputo Prov	Maputo Cid.	Moçambique
Popul. Agrícola	T1	2,662787	2,419781	0,700269	0,443332	0,291327	1,035879	3,041380	1,177030	17,441223	4,596690	14,422969	0,395180
Prop. Mulheres	R1	0,170799	0,282166	0,287603	0,535778	0,983618	0,341122	0,085736	0,283710	0,017272	0,091684	0,057164	1,456244
Tamanho med. Agr. Familiar	R2	0,992605	5,671047	6,277138	4,910387	13,072944	3,419765	0,339743	1,958797	0,063693	1,139123	0,321412	8,203030
Expl. Chef. Mulheres	T4	5,977259	1,619768	0,286914	0,309116	0,130400	0,870118	6,636558	2,750879	40,843020	4,112185	23,313468	0,513457
Prop. Expl. Chef. Mulheres	R3	2,237841	3,768851	2,603068	3,388995	5,346126	2,811902	0,734670	4,476611	0,140043	0,984273	0,528576	10,658775
Expl. Chef. Pes. Analfabetas	T5	3,683109	1,233450	0,272133	0,478255	0,056537	0,834037	9,468719	2,808388	54,383482	7,160273	97,504042	0,667684
Expl. Chef. Mulh. Analfabetas	T6	5,570278	1,170409	0,259775	0,388481	0,059638	0,722387	6,452069	3,251101	48,112600	5,719559	55,141715	0,557547
Mão-ob. Famil. c/ activ. Expl.	T7	2,277276	1,125073	0,499619	1,812887	0,006772	0,062940	1,922111	0,903667	34,238348	12,148628	638,537175	0,200748
Prop. Expl. c/ Mão-ob ext. tempo int.	R4	1,622484	6,226689	1,320796	9,919335	14,243450	1,872435	0,801630	17,411312	0,143399	1,833592	0,473016	20,199600
Tot. área cultiv. Estimada (ha)	T10	2,873607	1,531225	0,727708	1,437586	0,039571	0,347758	2,380799	1,836401	31,202411	1,374508	24,039051	0,348643
Prop. Área cultiv. c/ cult. anuais	R5	4,658855	2,139159	4,438182	6,240876	129,097873	3,812191	0,970549	2,690357	0,147217	0,729751	0,105036	17,125646
Expl. c/ activ. Agrícola	T12	5,440133	1,673281	1,035407	1,392271	1,565221	13,384205	7,033887	2,034698	39,422687	4,562877	5,953341	0,285868
Prop. Expl. c/ cajueiros	R7	8,517557	6,187657	3,108686	5,012582	329,219574	23,535721	1,517132	1,277771	0,231891	1,783143	0,023294	24,444961
Prop. Expl. c/ coqueiros	R8	94,802218	6,570115	10,317660	0,179724	458,854710	53,637361	0,953598	0,244500	0,226734	4,115733	0,025865	23,705001
Expl. Criam animais	T15	6,348029	0,876449	0,258506	0,418296	0,256798	2,334721	10,997040	3,797439	66,141026	3,561884	24,433876	0,759281
Expl. C/ bovinos	T16	0,121338	95,473749	3,770432	30,738940	0,011391	0,357149	2,209829	0,090080	9,065457	0,545407	62,365066	0,003384
Nº méd. bovinos/expl.	R9	653,560169	0,253603	0,588819	1,717156	61,771295	7,668629	1,698605	51,232486	0,147166	1,804098	0,122508	850,833250
Expl. criam suínos	T18	18,169853	9,351233	2,071420	0,408360	0,103868	0,653523	5,620008	0,489965	35,770006	7,065350	71,837544	0,439513
Prop. Expl. c/ suínos	R10	5,535433	17,180312	19,874173	4,645188	5,299693	2,403047	0,758417	0,849193	0,128653	2,176745	2,163251	10,970225
Nº med. suínos/expl.	R11	0,787361	4,178705	1,562188	52,582979	206,638259	11,580308	1,594068	116,374475	0,238472	0,649718	0,360279	12,406325
Prop. Expl. c/ aves	R12	2,397221	2,593994	2,530986	4,362683	8,057613	6,440389	1,091565	5,494250	0,205980	0,837420	0,690717	15,548729
Expl. criam burros	T21	0,000000			2,059239	0,000830	0,115675		0,000108	4,330132	0,000000	31,873843	0,019895
Nº med. burros/expl.	R13	0,000000			234,926812	0,000000	22,930598		0,000000	6,537250	0,000000	281,097747	0,000000
Prop. Expl. c/ aquacultura	R14	0,000000		10,052398	6,149131	0,218035	1,270269	0,428786	7,464452		0,084427	2,015241	8,853378
Expl. c/ titulo de uso de terra	T24	2,747613	4,832582	0,108588	0,178970	3,041489	9,126803	5,441007	1,953850	24,324970	2,361851	1,659415	0,340463

7. Conclusões

Como foi anteriormente referido, o coeficiente de variação (*CV*) é a medida mais usada para tirar conclusões sobre a fiabilidade das estimativas. De acordo com os valores dos *CV*s do Quadro 2, pode concluir-se que os resultados do Censo Agro-Pecuário II são bons, os CV na sua maioria apresentam-se bastante baixo (muitos a situarem-se abaixo de 10%). No entanto, um aspeto a ressaltar na interpretação deste indicador está relacionado com fenómenos pouco frequentes em pesquisas estatísticas e que têm grande impacto no valor do *CV*. Regra geral, em pesquisas estatísticas, fenómenos raros resultam em maiores variâncias, e consequentemente valores altos de *CV* o que não permite fazer melhor juízo da sua precisão é o caso do variável 22 (total de explorações com burros) que apresenta um valor de *CV* muito alto. Tais fenómenos requerem um aumento da dimensão da mostra o que acarreta custos.

Quanto às estimativas, os aspetos demográficos aqui analisados apresentam resultados muito próximos aos apurados pelo censo populacional 2007. A população agrícola apresenta-se com um valor relativamente alto, à projeção da população rural dado que a população agrícola inclui também alguns residentes das áreas urbanas com pelo menos um elemento na família a praticar agricultura por conta própria. Em 2007 cerca de 31,0% das famílias rurais moçambicanas eram chefiadas por mulheres, os resultados do CAP aqui apresentados revelam 31,8%, o censo demográfico indicava que em 2007, em média, as famílias moçambicanas era compostas por 4,4 pessoas e 4,3 para as áreas rurais, e os resultados aqui apurados do CAP apontam 4,8 pessoas, em média nas famílias rurais.

Os resultados do CAP publicados pelo INE apresentam-se relativamente superiores aos dados apresentados neste trabalho. Uma das razões prende-se com a informação auxiliar utilizada neste trabalho, por um lado, porque o número das médias explorações não estava disponível e, por outro lado, verificou-se que o número das pequenas explorações em alguns conglomerados superou o número que era esperado devido à reclassificação das explorações.

Quadro 6: Comparação dos resultados das pequenas e médias explorações do CAP

Nº	Explorações (nº) por espécie pecuária	CAP 2009/10 (divulgados)	Estimativas do Trabalho	Diferença	%
01	Bovinos	204 912	190 223	14 689	**7,72**
02	Suínos	434 010	397 680	36 330	**9,14**
03	Burros	9 436	8 818	618	**7,01**

Relativamente ao efeito do plano amostral (*deff*), como se pode contatar a partir do Quadro 2, há uma variação muito grande dos valores do *deff*. Em certos casos, os

valores do *deff* são elevados, o que justifica a necessidade de se considerar o verdadeiro plano amostral quando se utiliza um plano de sondagem complexo para a obtenção de dados. Esses valores acontecem porque as estimativas de variância tomando as observações como independentes e identicamente distribuídas, ou seja, como se fossem obtidas por um plano de sondagem aleatório simples subestimam as variâncias corretas.

Uma vez que um dos propósitos do CAP é servir de base de sondagem para as pesquisas intercensitárias, os valores do *deff* podem ser usados para planificar aqueles inquéritos, dado que permitem comparar e prever o impacto do uso de plano amostrais alternativos sobre a precisão de estimadores de totais de variáveis relevantes. Podem, também, ser usados para determinar dimensões amostrais. Portanto, os valores de *deff* podem ser utilizados como informação auxiliar na planificação de inquéritos por amostragem antes da seleção das amostras.

8. Recomendações

Em futuras ocasiões, propõe-se um estudo comparativo sobre o desempenho dos estimadores de variância dos métodos de linearização de Taylor, Ultimate cluster e Bootstrap, focados no presente trabalho e apresentar um plano de sondagem alternativo;

Uma vez que os problemas da base de sondagem têm implicações nas estimativas dos inquéritos, recomenda-se um estudo para tratar dos problemas de base de sondagem dos Censos Agro-Pecuários.

Uma vez que os dados revelaram a existência de uma proporção muito pequena de pessoas que praticam a atividade agro-pecuária nas áreas urbanas o que é consistente com os resultados do IFTRAB 2004/2005, recomenda-se que no futuro os inquéritos pilotos abarquem as áreas urbanas, propriamente dita e que sejam adotados planos de sondagem adequados;

Maior rigor na seleção de pessoal a trabalhar na recolha de dados e que seja dada uma melhor formação, bem como maior supervisão do trabalho de campo de modo a reduzir os erros de medição;

Dado que a operação de listagem consome recursos, e por outro lado como os Censos agro-pecuários são também utilizados como base de sondagem para os inquéritos infra-anuais, recomenda-se que se melhore a identificação das explorações agro-pecuárias na base de sondagem com a utilização de endereços físicos e/ou coordenadas geográficas;

Uma vez que as médias e grandes explorações são em número reduzido, os Serviços Distritais das Atividades Económicas deveriam possuir listas atualizadas daquelas com endereços bem identificados.

9. Referências bibliográficas

BARNETT, V. - Sample Survey: Principle & Methods. Earl Babbie, Secound Edition. Califórnia: Belmont, 1998.

COCHRAN, W. G. - Sampling Tecniques. John Wiley & Sons, Third Edition, 1977.

COELHO, P. S.; PNHEIRO, J. A.; XUFRE, P. - Métodos de Sondagem. Lisboa: Instituto Superior de Estatística e Gestão de Informação. Universidade Nova de Lisboa, 2010. Documento policiado e distribuído no quadro da disciplina de Métodos de Sondagem no Curso de Mestrado em Estatística e Gestão de Informação ministrado pelo ISEGI. UNL.

CORDEIRO, R. - Efeito do desenho em amostragem de conglomerados para estimar a distribuição de ocupações em trabalhadores. Rev. Saúde Pública. 35:1 (2010) 10-5. Disponível em http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0034-89102001000100002

Crespo, Maria Teresa. - Técnicas de Amostragem. Cursos de Curta Duração. Centro Europeu de Estatística Para os Países em Vias de Desenvolvimento, (1993)

DEAN, A.G; SULLIVAN, K.M; SOE, M.M. - OpenEpi: Open Source Epidemiologic Statistics for Public Health, version 2.3.1. [Em linha] USA [Consult. 14 Marc. 2011] disponível em

http://www.openepi.com/OE2.3/Menu/OpenEpiMenu.htm

EFRON, B. – The Jackknife, The Bootstrap and Other Resampling Plans. – Society for Industrial and Applied Mathematics. Philadelphia 1982.

http://www.ine.gov.mz – indicadores básicos, Moçambique

http://www.measuredhs.com/pubs/pdf/SR179/SR179p.pdf)

http://www.portalsaofrancisco.com.br/alfa/mocambique/macambique.php

http://www.portalsaofrancisco.com.br/alfa/mocambique/macambique.2-php

GUJARATI, D., Tradução de Monteiro, M.J.C. – Ecometria Básica. 4ª edição. – Rio de Janeiro: Elsevier, 2006.

KLEIN, C. H. – Estudos Seccionais – Inquéritos Epidemiológicos: Escola Nacional de Saúde Pública – Fiocruz – RJ - Brasil, 2007. Documento cedido aos estudantes no âmbito do curso de Mestrado em Saúde Pública.

LARSON, R.; FARBER, B. - Estatística Aplicada. 2ª ed. São Paulo: Pearson Prentice Hall, 2004.

LEVY, P. S.; LEMESHOW, S. - Sampling of Populations – Methods and Applications. Third Edition. John Wiley & Sons, Inc. New York, 1999.

LOHR, S. L. – Sampling: Design and Analysis. Pacific Grove: Duxbury Press. USA, 1999.

MURTEIRA, B.; RIBEIRO, C. S.; SILVA, J. A.; PIMENTA, C. - Introdução à Estatística. Lisboa, 2010.

Pessoa, D. G. C. e Silva, P. L. N. – Análise de Dados Amostrais Complexos. IBGE, 1998.

REIS, E., MELO, P., ANDRADE, R., CALAPAZ, T. – Estatística Aplicada. 2.ª Edição. Lisboa: Edições Sílabo,

SÃO, J. e TU, D. – The jackknife and Bootstrap. New York: Springer-Verlag, 1996.

SANTOS, J. A. – Métodos de Sondagem. Lisboa: Instituto Superior de Estatística e Gestão de Informação. Universidade Nova de Lisboa, 2009. Documento policiado e distribuído no quadro da disciplina de Estatística no Curso de Mestrado em Estatística e Gestão de Informação ministrado pelo ISEGI. UNL.

SÄRDAL, C. E., SWENSSON, B. e WRETMAN, J. – Model Assisted Survey Sampling - Springer Seriesn in Statistics. – Spring-Vertag. New York 1991

VICENTE, P.; REIS, E.; FERRÃO, F. - Sondagem: A amostragem como factor decisivo da qualidade. Lisboa: Edições Sílabo, 1996.

Wikipédia. - Sondagem. [Em linha] A enciclopédia livre, 2010 [Consult. 18 Dez. 2010]. Disponível em http://pt.wikipedia.org/wiki/Sondagem.

WOLTER, K. M. – Introduction to Variance Estimation. – Springer-Verlag 1985.

10. Anexos

10.1 Outputs

Quadro A1.1: Estimativas de totais e quocientes

NAME	nias	cabo	namp	zamb	tete	mani	sofa	inha	gaza	mapp	mapc	Pais
ES_T1	1 008 044,4409	1 488 433,2498	3 421 998,4250	3 771 109,1431	1 583 497,8236	1 323 231,7552	1 484 032,9199	1 265 956,4262	1 163 320,7780	658 761,6892	300 458,7771	17 468 845,4280
ES_T2	532 892,3649	777 554,8177	1 792 504,9707	1 968 676,3891	814 820,9029	684 686,4219	777 345,2982	687 608,9811	633 636,9177	351 453,0094	159 456,6113	9 180 636,6848
ES_T3	203 997,0000	334 121,0000	787 904,0000	792 708,0000	329 654,0000	238 188,0000	263 338,0000	263 163,0000	218 253,0000	131 092,0000	49 211,0000	3 611 629,0000
ES_T4	62 746,9123	109 114,9152	182 308,5222	211 658,2923	84 207,3873	58 585,1756	67 721,3106	89 957,8960	75 986,5397	40 605,6462	15 606,4761	998 499,0734
ES_T5	112 219,4390	198 560,8037	425 666,7531	403 395,7300	165 715,4324	95 570,1885	102 809,1357	105 351,0845	86 641,2371	33 786,9219	4 956,5246	1 734 673,2506
ES_T6	48 310,1111	84 098,3053	138 598,3913	164 894,6703	66 268,1946	42 971,4749	47 825,5667	57 151,3390	43 317,1051	17 490,0402	3 528,6317	714 453,8302
ES_T7	37 580,6969	41 358,8049	134 377,3337	73 270,7674	169 151,5542	124 649,9573	108 141,0217	80 149,3342	29 964,6801	8 708,8966	1 414,4057	808 767,4528
ES_T8	19 181,7334	18 995,3366	44 660,8182	54 344,7426	22 916,7886	30 772,5264	22 649,6060	7 687,7799	26 907,7185	8 697,3209	4 746,3068	261 560,6779
ES_T9	361 293,5383	467 261,8892	952 231,9531	995 272,8757	524 469,9440	499 678,4857	440 491,4887	390 581,8851	344 497,2505	108 154,1211	27 115,0118	5 111 048,4433
ES_T10	334 319,5086	413 882,9598	880 311,0810	913 203,1431	501 859,2404	437 021,6851	413 075,5974	366 236,1002	303 747,6329	96 756,7694	21 274,3669	4 681 688,0847
ES_T12	192 420,5950	314 155,1913	751 381,6920	771 396,0653	321 061,9465	226 440,4219	245 672,2290	237 511,4340	197 414,0001	94 331,0714	27 902,5821	3 379 687,2288
ES_T13	12 081,5327	103 055,6860	384 429,6863	294 499,0186	1 636,6531	25 939,1063	75 983,7812	202 780,6105	124 548,7607	46 774,6992	10 212,5400	1 281 942,0745
ES_T14	1 396,5652	93 838,7398	141 261,7096	235 455,4393	1 993,5257	2 490,5170	70 777,6461	179 450,8282	57 919,1186	20 029,7725	12 951,6737	817 565,5357
ES_T15	123 538,3348	215 755,6773	476 733,3629	542 537,7559	240 024,5675	184 456,6517	187 748,1569	214 874,7276	162 461,3870	72 997,2502	20 891,6851	2 442 019,5571
ES_T16	1 614,7772	279,2723	9 449,6592	785,3729	46 163,5116	27 676,2929	5 006,3976	43 964,1004	45 557,2322	8 402,9037	1 323,4669	190 222,9869
ES_T17	12 542,0642	4 203,5801	39 354,2338	5 229,1535	280 057,5356	148 667,5160	35 043,4730	178 299,3552	290 186,0167	75 776,2968	14 444,2762	1 083 803,5009
ES_T18	5 394,3574	18 944,4413	49 233,3885	84 462,1228	54 753,3653	17 409,7365	19 602,7678	97 334,2462	40 150,6659	7 547,0676	2 847,3848	397 679,5441
ES_T19	17 922,7542	60 645,5195	138 861,0818	230 254,7799	189 355,4017	69 648,0280	95 304,9678	245 134,0576	126 414,3256	28 380,1514	15 497,9090	1 217 418,9766
ES_T20	116 046,5789	191 357,5756	407 996,5207	518 772,4629	205 919,7848	176 930,7919	183 085,2280	202 280,8430	144 738,9977	67 744,7466	19 042,8540	2 233 916,3841
ES_T21	4,9646	0,0000	0,0000	182,5437	1 066,8872	508,9658	0,0000	3 820,3928	3 041,7075	169,9696	22,7902	8 818,2214
ES_T22	9,9292	0,0000	0,0000	730,1747	2 814,9240	1 556,3357	0,0000	6 517,2705	5 158,4996	370,1250	22,7902	17 180,0489

ES_T23	605,6046	0,0000	325,0209	1 037,0587	538,3549	723,4543	101,2141	270,5268	0,0000	260,3634	14,5607	3 876,1584
ES_T24	2 306,4806	3 191,0836	15 904,0142	16 904,6505	4 354,3240	3 352,3426	6 917,6205	2 837,8719	7 671,7523	14 851,3259	11 515,1205	89 806,5866
ES_T25	26 768,0000	40 474,0000	129 340,0000	88 064,0000	26 620,0000	39 698,0000	73 122,0000	42 557,0000	36 269,0000	57 171,0000	49 211,0000	609 294,0000
ES_T26	177 229,0000	293 647,0000	658 564,0000	704 644,0000	303 034,0000	198 490,0000	190 216,0000	220 606,0000	181 984,0000	73 921,0000	0,0000	3 002 335,0000
ES_R1	0,5286	0,5224	0,5238	0,5220	0,5146	0,5174	0,5238	0,5432	0,5447	0,5335	0,5307	0,5255
ES_R2	4,9415	4,4548	4,3432	4,7572	4,8035	5,5554	5,6355	4,8105	5,3301	5,0252	6,1055	4,8368
ES_R3	0,3076	0,3266	0,2314	0,2670	0,2554	0,2460	0,2572	0,3418	0,3482	0,3097	0,3171	0,2765
ES_R4	0,0940	0,0569	0,0567	0,0686	0,0695	0,1292	0,0860	0,0292	0,1233	0,0663	0,0964	0,0724
ES_R5	0,9253	0,8858	0,9245	0,9175	0,9569	0,8746	0,9378	0,9377	0,8817	0,8946	0,7846	0,9160
ES_R7	0,0592	0,3084	0,4879	0,3715	0,0050	0,1089	0,2885	0,7706	0,5707	0,3568	0,2075	0,3549
ES_R8	0,0068	0,2809	0,1793	0,2970	0,0060	0,0105	0,2688	0,6819	0,2654	0,1528	0,2632	0,2264
ES_R9	7,7671	15,0519	4,1646	6,6582	6,0666	5,3717	6,9997	4,0556	6,3697	9,0179	10,9140	5,6975
ES_R10	0,0264	0,0567	0,0625	0,1065	0,1661	0,0731	0,0744	0,3699	0,1840	0,0576	0,0579	0,1101
ES_R11	3,3225	3,2012	2,8205	2,7261	3,4583	4,0005	4,8618	2,5185	3,1485	3,7604	5,4429	3,0613
ES_R12	0,5689	0,5727	0,5178	0,6544	0,6247	0,7428	0,6952	0,7687	0,6632	0,5168	0,3870	0,6185
ES_R13	2,0000			4,0000	2,6384	3,0578		1,7059	1,6959	2,1776	1,0000	1,9482
ES_R14	0,0030	0,0000	0,0004	0,0013	0,0016	0,0030	0,0004	0,0010	0,0000	0,0020	0,0003	0,0011

QuadroA1.2: Estimativas das variâncias

NAME	nias	cabo	namp	zamb	tete	mani	sofa	inha	gaza	mapp	mapc	Pais
VART1	279079343,6	362295977,9	292757975,4	251315818,7	51336390,18	200981650,4	695126056,5	184139081,7	4119235721	421011925,1	306111808,7	1075079692
VART2	105777644,6	98096098,58	103976459,7	100184713,9	17273205,42	59389798,6	251638425,4	64841661,82	1473724931	148443067,3	117733306,8	406370548,7
VART3	3,3456E-15	1,42542E-13	1,67408E-15	5,72953E-16	6,72921E-18	4,85091E-15	3,45705E-14	7,12226E-16	2,5426E-12	4,65744E-16	2,31856E-12	3,52922E-14
VART4	16820053,66	9834437,895	4903184,699	6630404,367	609355,8737	3195470,181	27301474,25	7975962,52	127974305,6	6785368,429	11340172,64	32578421,13
VART5	11888924,17	8203122,871	6423362,137	13237533,83	379359,3233	4086641,309	50041802,27	8792953,905	191613052,7	12214508,15	23656740,46	52688104,24
VART6	13380323,92	6143131,78	3595072,336	6958687,483	232391,7236	2105340,825	21598218,7	6931897,356	103322701,5	5449785,057	9440784,954	27487067,38
VART7	39303064,76	19118747,8	41481504,31	97155761,01	651803,1323	4889594,945	140478044,8	28222830,81	784286383,4	104457595,7	196397063,2	112130373
VART8	4877791,551	3961412,745	747991,4241	6331851,935	552153,6663	1290757,983	12258922,78	3910344,041	62411115,47	3662174,49	4083721,467	20733993,39
VART9	134061900,4	81775773,18	110999692,5	320527395,6	7358575,165	45402352,1	267263082,3	93843794,87	1745832801	86083206,33	164125719	434391309,1
VART10	92524441,17	62294063,19	98955528,13	243484440,1	6354304,895	36993178,03	231507617,9	81551150,14	1401980722	69758327,54	141965872,4	336591798,4
VART12	3630032,088	2849019,376	4569597,973	4432412,076	1424090,953	10851834,43	7229838,256	2222732,173	52274630,34	5849953,844	3110607,542	6104511,631
VART13	16748042,44	15660121,36	8226456,521	11699774,79	974683,5619	13994635,83	60585734,33	1789054,148	228773133,9	13194780,89	379517,1267	85520332,92
VART14	22747680,3	15744868,85	16079514,5	375135,9197	1652896,435	3400521,334	36457838,36	419997,3621	177986220,8	17178212,87	496899,8005	63432655,12
VART15	19952412,58	5619534,064	5857842,609	9998565,943	1200967,814	7930005,522	43981531,23	7289859,492	172345237,6	6487055,9	13460972,74	50566489,71
VART16	21723,85177	6673801,083	4629677,506	7832350,835	62093,15833	994408,7546	2050801,515	194379,514	30167660,5	683072,8517	6919426,833	105924,5994
VART17	3294663,512	422788868,3	88518269,36	237743984,4	8408347,331	83883641,78	46119097,22	9801107,314	1278676816	35252819,57	337741831,8	5124185,132
VART18	5029135,442	10902571,36	12331127,62	4420485,413	439907,9876	997572,6107	10164970,87	1567349,622	77768615,44	3808018,311	11682607,82	16424868,39
VART19	63157800,15	170481446	111585886,6	93974175,72	16003616,97	17652043,7	159463376,5	18264427,33	1205307476	103122651,7	242449470,4	209152580,9
VART20	20748909,42	7531712,684	6693098,972	9862783,455	1132188,744	7538872,462	44990980,03	7737256,05	185268792,1	6742718,106	16232669,89	56057602,25
VART21	0	394263,1574	508854,6876	40447,57527	243,2019165	8407,79197	0	24,69815902	1116377,794	0	130795,298	33341,38409
VART22	0	1294001,281	1716747,594	501485,5309	243,2019165	42167,10991	0	98,79263609	5328466,265	0	1240260,609	533462,1454
VART23	0	0	43901,39264	80312,35744	213,0434848	23568,90185	32047,25945	60704,38102	450228,6544	5395,855166	59054,28446	145031,1788
VART24	390100,7668	1241359,248	199436,3889	483345,8529	1051310,472	2753916,03	2994064,297	315841,2064	14995684,76	1743244,432	697858,5073	3125207,555
VART25	4,28757E-17	1,43366E-13	1,46358E-17	7,82497E-18	6,72921E-18	4,79083E-15	2,49714E-16	4,46868E-18	1,48624E-13	2,19042E-17	6,01873E-17	5,80576E-17
VART26	2,50938E-15	4,0661E-16	1,39208E-15	4,0484E-16	0	1,55862E-17	2,8739E-14	6,51281E-16	2,37807E-12	2,12161E-16	2,31527E-12	2,84643E-14
VARD1	93636749,12	126446688,4	78642688,12	107466557	19175144,96	66512050,61	252044428	54721717,24	1391762634	110137567,8	100759314,9	382219727,6

VARD2	279079343,4	362295977,9	292757975,3	251315818,8	51336390,18	200981650,4	695126056,5	184139081,7	4119235719	421011925	306111807,8	1075079692
VARD3	16820053,66	9834437,896	4903184,699	6630404,367	609355,8737	3195470,181	27301474,25	7975962,52	127974305,6	6785368,429	11340172,64	32578421,13
VARD4	4877791,551	3961412,745	747991,4241	6331851,935	552153,6663	1290757,982	12258922,78	3910344,041	62411115,47	3662174,49	4083721,467	20733993,39
VARD5	28333438,11	8538781,395	8032936,296	29010545,42	1013988,187	2361746,855	18652371,7	4636481,93	153256840,5	7318862,191	5317422,998	40040265,41
VARD6	8,7753E+11	1,12393E+12	7,66004E+11	4,29123E+11	34947295853	5,035E+11	2,29982E+12	4,344E+11	1,09658E+13	5,94527E+11	9,65869E+11	2,93614E+12
VARD7	16748042,44	15660121,36	8226456,521	11699774,79	974683,5619	13994635,83	60585734,33	1789054,148	228773133,9	13194780,89	379517,1269	85520332,92
VARD8	22747680,3	15744868,85	16079514,5	375135,9197	1652896,435	3400521,334	36457838,36	419997,3622	177986220,8	17178212,87	496899,8006	63432655,12
VARD9	2951295,09	344608609,5	57831603,67	167847620,8	7642275,829	70857271,47	32470777,89	7471994,972	991804830,9	27301230,61	268716011,1	4106139,979
VARD10	5029135,442	10902571,36	12331127,62	4420485,413	439907,9876	997572,6109	10164970,87	1567349,622	77768615,44	3808018,311	11682607,82	16424868,39
VARD11	39457281,07	105976470,3	62119200,23	66654697,28	13625933,11	12351626,82	112673230,1	11403936,89	804184608,4	75169250,72	171034637,4	133718344,5
VARD12	20748909,42	7531712,684	6693098,972	9862783,454	1132188,744	7538872,462	44990980,03	7737256,05	185268792,1	6742718,106	16232669,89	56057602,25
VARD13	0	442182,4196	515287,9673	291708,9087	0	13564,01042	0	24,69815902	2199827,472	0	636987,0113	300072,4568
VARD14	3,3456E-15	1,42542E-13	43901,39264	80312,35744	213,0434848	23568,90184	32047,25945	60704,38101	450228,6543	5395,855167	59054,28446	145031,1789
VARR1	1,34541E-05	1,75356E-05	1,0971E-05	2,35552E-05	6,08015E-05	3,03698E-05	6,70129E-06	1,33969E-05	1,32235E-06	1,06095E-05	1,29788E-05	8,82944E-06
VARR2	0,002499886	0,007605768	0,004227261	0,004429759	0,021198296	0,011695102	0,001119739	0,004424849	0,000315799	0,006071099	0,002816848	0,001710859
VARR3	0,000150668	0,000206457	7,07992E-05	0,000116869	0,000251621	0,000185944	4,39784E-05	0,000191662	9,81107E-06	9,78467E-05	0,000104353	5,18446E-05
VARR4	4,36934E-05	8,31629E-05	1,08006E-05	0,000111607	0,000228	7,51091E-05	1,97472E-05	9,39653E-05	4,78471E-06	5,28095E-05	3,75785E-05	3,29956E-05
VARR5	0,000101747	5,39162E-05	4,73926E-05	8,31992E-05	0,001317925	0,000149068	1,77508E-05	2,99231E-05	4,7398E-06	3,1988E-05	1,61827E-05	3,20439E-05
VARR7	0,000150023	0,000328757	0,000118785	0,000206223	0,000402475	0,000814346	9,75941E-05	4,29908E-05	1,75388E-05	0,000190272	3,49233E-06	0,000136095
VARR8	0,000203765	0,000330536	0,000232179	6,61224E-06	0,000682529	0,000197876	5,87279E-05	1,00925E-05	1,36452E-05	0,000247714	4,57248E-06	0,000100945
VARR9	84,37676291	0,074091713	0,011107418	0,061269541	4,078634731	0,553470567	0,18982557	0,886955447	0,012481275	0,353201227	0,06194001	5,410022414
VARR10	4,50491E-05	0,00022888	0,000178055	7,79166E-05	0,000181651	5,80487E-05	1,63742E-05	3,76633E-05	5,96208E-06	5,49126E-05	0,000107504	2,61382E-05
VARR11	0,064909664	0,025987656	0,003611223	0,127731745	1,943985776	0,144579814	0,033151756	0,28025095	0,002959804	0,103276975	0,03188822	0,011366616
VARR12	0,000185861	0,000158115	9,66446E-05	0,000173844	0,000467514	0,000438686	7,24734E-05	0,000185926	1,42035E-05	9,72317E-05	0,000149373	8,92089E-05
VARR13		0,036172352	0,021411141	0,663278496	0	0,0889434		0	0,016660309		0,238538599	0
VARR14	0	0	6,33912E-07	1,41561E-06	8,79719E-08	1,37147E-06	5,16231E-08	1,45872E-06	3,45165E-08	7,78096E-08	5,43419E-07	2,308E-07

Quadro A1.3: Estimativas dos coeficientes de variação

NAME	nias	cabo	namp	zamb	tete	mani	sofa	inha	gaza	mapp	mapc	Pais
CV_T1	1,122365952	1,636184489	1,35156087	1,198047392	2,384667145	2,152037672	0,770463339	1,346149541	0,367404203	1,382622639	1,104898976	0,869463244
CV_T2	1,322713937	1,563095076	1,482948171	1,461870882	2,606417234	2,192747678	0,884969108	1,511080393	0,41815319	1,567349984	1,331642078	1,023968932
CV_T3	1,73114E-11	1,72986E-10	1,55476E-11	1,00494E-11	5,27132E-12	5,31295E-11	2,35982E-11	1,30823E-11	4,41505E-11	8,19521E-12	4,61903E-10	2,36988E-11
CV_T4	3,758629941	4,127034537	2,461499983	4,395236735	5,00184983	4,402313421	2,866065286	4,500898227	1,132957777	3,846461257	3,99907584	2,696681952
CV_T5	1,736511871	3,305711613	2,405704016	3,806984413	12,42647959	5,983215326	1,661868823	2,642404528	0,797985249	3,399431417	2,935043202	1,799388703
CV_T6	4,349567442	5,721837128	3,317625854	6,138798889	13,66167632	8,296031633	3,353132984	5,449893624	1,422733991	4,88123326	4,636592344	3,179490889
CV_T7	15,15811604	14,59217658	8,035766891	7,907553609	57,08003334	25,39062012	8,820195416	14,13629034	3,462690471	9,451038761	8,284980718	14,45209785
CV_T8	11,6269184	7,396873154	11,24986653	8,177154869	15,65576067	13,06281923	7,839700635	10,30907346	3,020358665	8,44906916	8,81808536	8,378840852
CV_T9	2,47794878	2,624983449	2,697421351	3,582959727	10,0043083	6,230115578	1,716827795	2,681282579	0,817506712	2,106307121	2,442686408	2,094104719
CV_T10	2,324077944	2,598426444	2,71618208	3,570529447	11,84888216	6,286073603	1,728408823	2,701178756	0,799776327	2,021941054	2,374160378	2,009020495
CV_T12	0,60647233	0,855007163	0,900024796	0,929749834	4,276854429	3,492181541	0,357852192	0,774804286	0,213928662	0,984510049	0,54933058	0,32029344
CV_T13	3,971093087	3,177299477	1,414425174	13,18662762	9,667140921	7,997786338	2,024735376	11,07107285	1,179869865	4,780575592	37,64081853	3,140153221
CV_T14	5,082605371	6,850899225	2,234554744	24,59262089	9,926517863	9,206546175	4,274359042	46,40470869	1,631813852	5,855889528	35,36002538	3,382577203
CV_T15	2,070310458	1,459150195	1,126376316	1,714251495	5,245564268	3,857715928	1,391104088	2,185535351	0,537589164	1,356588521	1,528560126	1,310694377
CV_T16	52,77650233	5,67060371	4,894148913	10,1120253	18,82819884	11,86733176	15,1546416	27,30314189	2,887405765	16,50852127	5,698180778	41,44026496
CV_T17	43,18035521	7,08574128	5,276752015	10,37143102	20,07518406	12,08663093	17,25634985	24,96137785	3,299361071	16,94297965	6,562135056	43,28934539
CV_T18	11,8376328	8,223784412	3,607744111	12,07654717	23,29350065	13,23408793	6,475798726	23,20829458	2,217527076	9,954790355	6,242507623	4,798317228
CV_T19	13,10432865	10,32861925	4,309242954	13,91859628	25,81285053	14,80412653	9,093894958	23,8450524	2,851732979	10,65519733	8,223049817	6,280915629
CV_T20	2,380411158	1,896100319	1,278965615	1,774991704	5,587626624	4,053009069	1,64401686	2,396961965	0,60930427	1,418287243	1,956576815	1,443245938
CV_T21		20,64313475	18,67191523	39,51460786	68,42826156	53,94729113		100,1035092	11,98186981		33,89827247	100,0287899
CV_T22		22,05179506	20,1042315	45,50152247	68,42826156	55,48025618		100,1035092	13,43620788		39,563054	100,0287899
CV_T23			77,4513385	39,17239026	100,2426919	58,96437459	55,07875546	40,68373918	17,31071678	72,57533207	45,13953085	36,72208706
CV_T24	19,5726766	14,52292761	15,73654572	20,73865586	8,904243079	11,17403732	10,8798725	24,36601939	4,31196236	19,08632932	19,18505069	10,45762914
CV_T25	1,61782E-11	1,04397E-09	8,98953E-12	7,04649E-12	5,27132E-12	1,21068E-10	1,22177E-11	7,89721E-12	6,32728E-11	6,40053E-12	2,91437E-11	8,65229E-12
CV_T26	1,70591E-11	1,10804E-11	1,69128E-11	1,01368E-11		5,34076E-12	2,57417E-11	1,43996E-11	5,13633E-11	7,65747E-12	5,02123E-10	2,39431E-11
CV_R1	0,702143375	0,768811146	0,609818519	0,937967952	1,46926302	1,03295632	0,494195438	0,692376057	0,218808672	0,621837406	0,700121142	0,569195395

CV_R2	1,122365952	1,636184489	1,35156087	1,198047392	2,384667145	2,152037672	0,770463339	1,346149541	0,367404203	1,382622639	1,104898976	0,869463244
CV_R3	3,758629942	4,127034537	2,461499983	4,395236735	5,00184983	4,40231342	2,866065287	4,500898227	1,132957777	3,846461258	3,99907584	2,696681951
CV_R4	11,62691839	7,396873156	11,24986652	8,17715487	15,65576067	13,06281923	7,83970064	10,30907346	3,020358665	8,449069157	8,818085362	8,378840858
CV_R5	1,138792207	0,832783911	0,734186477	1,042910683	4,626989645	1,364754443	0,455738166	0,591155436	0,237677277	0,603117082	0,420401943	0,616946336
CV_R7	3,971093088	3,177299477	1,414425173	13,18662762	9,667140922	7,997786338	2,024735376	11,07107285	1,179869865	4,780575592	37,64081892	3,140153221
CV_R8	5,082605372	6,850899226	2,234554744	24,59262088	9,926517863	9,206546175	4,274359042	46,40470839	1,631813852	5,855889529	35,36002508	3,382577203
CV_R9	61,02670447	4,273322481	2,598692991	4,60801778	18,50438774	8,249797219	10,46169464	12,12535618	1,960838975	8,490422466	4,102392489	34,93365577
CV_R10	11,83763279	8,22378441	3,60774411	12,07654716	23,29350065	13,23408793	6,475798724	23,20829457	2,217527076	9,954790356	6,242507622	4,798317229
CV_R11	7,958619843	5,120118551	2,386102758	8,933729484	25,61648851	10,11154141	6,455538096	15,93340192	1,777152215	6,610032763	5,163548492	3,910829935
CV_R12	2,380411158	1,896100319	1,278965615	1,774991704	5,587626624	4,053009069	1,64401686	2,396961965	0,60930427	1,418287243	1,956576815	1,443245938
CV_R13		11,21456474	8,577528679	26,63380728	0	13,69555826		0	6,625186895		18,51105474	0
CV_R14			77,45134009	39,17239031	100,2426776	58,96437584	55,07875084	40,68373911	17,31071678	72,57533444	45,13953183	36,72208808

Quadro A1.4: Estimativas das variâncias na SAS

prov	nias	cabo	namp	zamb	tete	mani	sofa	inha	gaza	mapp	mapc	pais
va1	7,969408493	5,385673532	4,404172011	5,12075627	6,554196246	9,594963654	10,52108285	11,87291515	16,06640715	11,46801723	9,622512415	8,577420311
va2	2,752100727	2,207225352	1,730003194	2,024070095	2,49439179	3,601732374	3,909216354	4,341554918	5,696144127	4,426478821	3,727055907	3,206933416
va3	0	0	0	0	0	0	0	0	0	0	0	0
va4	0,213973547	0,218398308	0,180030555	0,193758687	0,173807796	0,181615235	0,189369828	0,220044435	0,213149339	0,20660456	0,220534154	0,200049157
va5	0,245449881	0,239227089	0,248657698	0,250029126	0,249571269	0,242313142	0,243281605	0,237616437	0,239682945	0,213592281	0,110000677	0,248801186
va6	0,18265206	0,18880131	0,145790863	0,161808327	0,144935626	0,1441279	0,154094268	0,161815601	0,146088483	0,119304255	0,077623086	0,155438082
va7	1,312337606	0,611267757	0,8746526	0,48410664	3,579913396	3,841865931	3,364322028	2,370233661	1,558263612	1,076593925	0,139447866	1,761095239
va8	0,084433473	0,046512508	0,055571897	0,064268144	0,123830807	0,126292808	0,092958242	0,037313185	0,138270041	0,085588534	0,130149896	0,082029642
va9	3,547421355	1,921048398	1,606884481	2,014070411	6,916619032	6,456504284	5,167538916	3,87825997	3,806217024	7,841703433	3,095440024	3,92835679
va10	2,859701026	0,985801653	1,198074808	1,307350422	6,273449172	3,610726968	4,24265848	3,376170968	2,876509924	5,592412884	1,646711824	2,985226454
va11	7020,025932	7044,480687	8414,884297	11175,02549	5989,150976	6298,853144	12314,47836	16709,63961	17156,32092	14932,67545	9615,878341	11258,98287
va12	0,05073828	0,061246209	0,04649295	0,028758054	0,033840534	0,040096498	0,047315206	0,082906011	0,090203586	0,160528729	0,236890425	0,067328083
va13	0,062038111	0,218218651	0,249832924	0,233630552	0,005728336	0,110864973	0,215666009	0,167820666	0,249695057	0,230227352	0,194249055	0,231009123
va14	0,005110062	0,221203711	0,142580342	0,209838262	0,010655546	0,009645421	0,182464446	0,205254989	0,164944149	0,099831412	0,206303009	0,183434851
va15	0,238996367	0,230635543	0,238719467	0,215922202	0,173946064	0,167970988	0,184103533	0,145689118	0,177258641	0,228037857	0,24977395	0,209976632
va16	0,013613679	0,002514444	0,012935399	0,002301688	0,202742073	0,137692936	0,042720132	0,163764806	0,226376307	0,156814088	0,050302759	0,098689214
va17	5,70331505	3,030058403	1,617984623	0,471657862	76,34506102	33,53699012	16,21393692	18,85903202	96,04538162	84,81894922	30,11629969	29,3999263
va18	0,021046347	0,041938482	0,062712619	0,097784461	0,157527803	0,075488281	0,083260077	0,242772708	0,147898716	0,067484572	0,073731215	0,117825956
va19	5,592917104	1,092149281	1,016882314	1,514742658	7,46339649	2,884347311	6,100253787	5,261810041	7,253085344	24,98922218	7,848913834	5,154967881
va20	0,245750612	0,24223629	0,249585787	0,22560517	0,225360312	0,188600497	0,194715966	0,177637657	0,220667547	0,241498518	0,242888332	0,230559061
va21	0,000320924	0	0	0,000177431	0,010900589	0,003594494	0	0,017351121	0,017538365	0,004703574	0,001860462	0,005283867
va22	0,001283697	0	0	0,002838893	0,148569992	0,039565455	0	0,080739466	0,148039807	0,03202806	0,001860462	0,041587301
va23	0,003518834	0	0,000770416	0,001417683	0,001999993	0,003952513	0,000633914	0,000969555	0	0,001886787	0,000931099	0,001350785
va24	0,01079583	0,009239977	0,019348334	0,024396184	0,012856417	0,01492203	0,025330829	0,012268075	0,041936583	0,092415525	0,190666972	0,028941408

10.2 Anexo II

10.2.1 Classificação das explorações agro-pecuárias

As variáveis consideradas para a classificação de explorações foram: área de terra sob cultivo de culturas anuais e permanentes e efetivos animais das principais espécies, uso da rega, práticas de horticulturas e fruticultura.

Quadro A2: Classificação das explorações agro-pecuárias

Fatores	Limite 1	Limite 2
Área cultivada não irrigada (ha)	10	50
Área cultivada irrigada, pomares em produção, plantações, hortícolas, floricultura (ha)	5	10
Número de cabeças de gado bovino	10	100
Número de caprinos/ovinos/suínos	50	500
Número de aves (1)	2 000	10 000

Notas:

1. **Pequenas explorações:** se todos os fatores forem menores que o limite 1;
   Médias explorações: se pelo menos um dos fatores for igual ou maior que o limite 1 e menor que o limite 2;
   Grandes explorações: Se um dos fatores for igual ou superior ao limite 2.

2. Se a exploração é apenas aquícola, é considerada **grande**, a exploração comercial com mais de 5 hectares e uma produção de 100 toneladas por ano. Considera-se **pequena exploração aquícola**, aquela que tem menos de 5 hectares.
   Não existe critério a prior para a distinção das médias explorações aquícolas.

10.2.2 Questionário do III Recenseamento Geral da População e Habitação - 2207

Secção G – Actividade Agropecuária e piscícola (Secção G: RGPH)

G1: Algum membro deste agregado familiar pratica actividade agrícola por conta própria?

1 |__| Sim 2 |__| Não

G2: Este agregado familiar tem tanques de aquacultura?

1 |__| Sim ⟶ |__|__|

2 |__| Não

Se sim, quantos? ____

G3: Algum membro deste agregado familiar pratica a pesca artesanal?

1 |__| Sim 2 |__| Não

G4: Este agregado possui cajueiros?

1 |__| Sim ⟶ |__|__|__|

2 |__| Não

G5: Este agregado possui coqueiros?

1 |__| Sim ⟶ |__|__|__|

2 |__| Não

G6: Quantos destes animais o agregado cria?

Sim ____ Não ____

G6.1: Bois/Vacas |__|__|__|__|

G6.2: Cabritos |__|__|__|__|

G6.3: Ovelhas/carneiros |__|__|__|__|

G6.3: Porcos |__|__|__|__|

G6.5: Galinhas |__|__|__|__|

G6.6: Patos |__|__|__|__|

10.3 Abreviaturas

AE – área de enumeração

CAP – censo agro-pecuário

CV – coeficiente de variação

deff – efeito do plano amostral

EQM – erro quadrático médio

INE – Instituto Nacional de Estatística

PIB – produto interno bruto

PPT – probabilidade proporcional ao tamanho

RGPH - recenseamento geral da população e habitação

SAS – sondagem aleatória simples

SAS-PICR - sondagem aleatória simples, probabilidades iguais com reposição

SAS-PISR - sondagem aleatória simples, probabilidades iguais sem reposição

UP (UPA) – unidade primária (unidade primária de amostragem)

US (USA) – unidade secundária (unidade secundária de amostragem)