UNIVERSIDADE DE ÉVORA

ESCOLA DE CIÊNCIAS E TECNOLOGIA

DEPARTAMENTO DE MATEMÁTICA

Avaliação de Métodos de Estimação da Variância em Amostras Complexas

ELÁDIO ANTÓNIO MUIANGA

Orientadora: Professora Doutora Anabela Cristina Cavaco Ferreira Afonso

Mestrado em Modelação Estatística e Análise de Dados

Área de especialização: **Modelação Estatística e Análise de Dados**

Dissertação

Évora, 2016

Esta dissertação inclui as críticas e as sugestões feitas pelo júri

UNIVERSIDADE DE ÉVORA

ESCOLA DE CIÊNCIAS E TECNOLOGIA

DEPARTAMENTO DE MATEMÁTICA

Avaliação de Métodos de Estimação da Variância em Amostras Complexas

ELÁDIO ANTÓNIO MUIANGA

Orientadora: Professora Doutora Anabela Cristina Cavaco Ferreira Afonso

Mestrado em Modelação Estatística e Análise de Dados

Área de especialização: **Modelação Estatística e Análise de Dados**

Dissertação

Évora, 2016

Esta dissertação inclui as críticas e as sugestões feitas pelo júri

"Se o ser humano não pensar que quer sempre mais, fatalmente teremos sempre menos. O homem só fracassa quando desiste de tentar. Todos os dias me levanto para vencer".

Aristóteles Onassis

RESUMO

A necessidade de conhecer uma população impulsiona um processo de recolha e análise de informação. Usualmente é muito difícil ou impossível estudar a totalidade da população, daí a importância do estudo com recurso a amostras. Conceber um estudo por amostragem é um processo complexo, desde antes da recolha dos dados até a fase de análise dos mesmos. Na maior parte dos estudos utilizam-se combinações de vários métodos probabilísticos de amostragem para seleção de uma amostra, que se pretende representativa da população, denominado delineamento de amostragem complexo.

O conhecimento dos erros de amostragem é necessário à correta interpretação dos resultados de inquéritos e à avaliação dos seus planos de amostragem. Em amostras complexas, têm sido usadas aproximações ajustadas à natureza complexa do plano da amostra para a estimação da variância, sendo as mais utilizadas: o método de linearização *Taylor* e as técnicas de reamostragem e replicação.

O principal objetivo deste trabalho é avaliar o desempenho dos estimadores usuais da variância em amostras complexas. Inspirado num conjunto de dados reais foram geradas três populações com características distintas, das quais foram sorteadas amostras com diferentes delineamentos de amostragem, na expectativa de obter alguma indicação sobre em que situações se deve optar por cada um dos estimadores da variância.

Com base nos resultados obtidos, podemos concluir que o desempenho dos estimadores da variância da média amostral de *Taylor*, *Jacknife* e *Bootstrap* varia com o tipo de delineamento e população. De um modo geral, o estimador de *Bootstrap* é o menos preciso e em delineamentos estratificados os estimadores de *Taylor* e *Jackknife* fornecem os mesmos resultados.

Palavras-chave: Amostras complexas, inferência e estimadores da variância.

Evaluation of variance estimation methods in complex samples

ABSTRACT

The need to know a population drives a process of collecting and analyzing information. Usually is to hard or even impossible to study the whole population, hence the importance of sampling. Framing a study by sampling is a complex process, from before the data collection until the data analysis. Many studies have used combinations of various probabilistic sampling methods for selecting a representative sample of the population, calling it complex sampling design.

Knowledge of sampling errors is essential for correct interpretation of the survey results and evaluation of the sampling plans. In complex samples to estimate the variance has been approaches adjusted to the complex nature of the sample plane. The most common are: the linearization method of Taylor and techniques of resampling and replication.

The main objective of this study is to evaluate the performance of usual estimators of the variance in complex samples. Inspired on real data we will generate three populations with distinct characteristics. From this populations will be drawn samples using different sampling designs. In the end we intend to get some lights about in which situations we should opt for each one of the variance estimators.

Our results show that the performance of the variance estimators of sample mean Taylor, Jacknife and Bootstrap varies with the design and population. In general, the Bootstrap estimator is less precise and in stratified design Taylor and Jackknife estimators provide the same results.

Keywords: Complex samples, inference, variance estimation.

AGRADECIMENTOS

A Deus por estar comigo nesta caminhada, por me ter concedido o dom da vida e a coragem de lutar. Por sempre guiar os meus caminhos e abençoar-me a cada dia, dando-me sabedoria para superar todos os obstáculos da vida, fazendo-me acreditar que posso vencer sempre, sendo assim forte na fé.

A Nádia e ao Christian pelo carinho, paciência e compressão, durante a minha ausência. Hoje, eles podem ver mais um dos meus sonhos, tornando-se realidade.

Uma palavra especial de agradecimento e reconhecimento vai ao Instituto Nacional de Estatística de Moçambique (INE-M) por ter-me proporcionado a bolsa de estudo, uma oportunidade única de concretizar um projeto de vida.

A todos os professores que compõem o corpo docente do Departamento de Matemática da Universidade de Évora – DMAT, que partilharam o seu tempo e seus conhecimentos para fazer de mim a pessoa que hoje sou. Em especial à Professora Doutora Anabela Cristina Cavaco Ferreira Afonso orientadora deste trabalho, por sua competência e dedicação como professora, transmitindo segurança e conhecimentos.

Aos camaradas do Mestrado e amigos em Évora, os quais compartilhamos momentos especiais diariamente durante a formação.

A todos o meu sincero muito obrigado.

Índice

Lista de figuras

Lista de tabelas

Lista de abreviaturas e símbolos

Ao longo deste trabalho utilizar-se-ão as seguintes abreviaturas

AAE	Amostra aleatória estratificada.
AAS	Amostra aleatória simples.
AASc	Amostra aleatória simples com reposição.
AASs	Amostra aleatória simples sem reposição.
AC	Amostragem por conglomerados (grupos).
CAE	Classificador de atividade económica.
CAE 45	Empresas de comércio, manutenção e reparação de veículos automóveis e motociclos (divisão 45 da CAE).
CAE 46	Empresas de comércio por grosso, exceto de veículos automóveis e motociclos (divisão 46 da CAE).
CAE 47	Empresas de comércio a retalho, exceto de veículos automóveis e motociclos (divisão 47 da CAE).
CV	Coeficiente de variação.
$\widehat{CV}$	Estimador ou estimativa do coeficiente de variação.
Env	Enviesamento.
EPS	Escala de pessoal ao serviço.
EQM	Erro quadrático médio.
$\widehat{\text{EQM}}$	Estimador ou estimativa do erro quadrático médio.
EQME	Erro quadrático médio escalado.
$\widehat{\text{EQME}}$	Estimador ou estimativa do erro quadrático médio escalado.
FCP	Fator de correção da população finita.
FUE	Ficheiro de unidade estatística.
IAE	Índice de atividade económica.
i.i.d	Independentes e identicamente distribuídas.
IC	Intervalo de Confiança.
INE-M	Instituto Nacional de Estatística de Moçambique.
Metical (MT)	Unidade Monetária de Moçambique.
NPS	Número de pessoal ao serviço.
UPA	Unidade primária de amostragem.

USA	Unidade secundária de amostragem.
UTA	Unidade terciária de amostragem.
VVN	Volume de negócios mensal da empresa.

Serão ainda utilizados as seguintes notações

Cov	Covariáveis.
$E(\hat{\mu})$	Esperança matemática do estimador da média populacional.
f	Fração de amostragem.
K	Dimensão das amostras por estrato.
M	Número dos grupos na população.
m	Número dos grupos na amostra.
n	Dimensão da amostra.
n_i	Número dos elementos do estrato ou grupo *i* na amostra.
N	Dimensão da população.
NR	Número de réplicas.
N_i	Número dos elementos dos grupos *i* na população.
s	Amostra.
S^2	Variância amostral.
Var	Variância.
$\widehat{Var}$	Estimador ou estimativa da variância.
w_i	Peso de amostragem.
W_i	Peso do estrato *i*.
$\bar{Y}_i$	Média da variável de interesse no estrato ou grupos *i*.
Y_T	Total variável de interesse.
$\bar{y}_i$	Média amostral da variável de interesse no estrato ou grupos *i*.
π	Probabilidade de inclusão.
μ	Média da variável de interesse na população.
$\hat{\mu}$	Estimador ou estimativa da média populacional.
σ	Desvio padrão da variável de interesse na população.
σ^2	Variância da variável de interesse na população.

Capítulo 1

Introdução

Investigações de todos os tipos são realizadas atualmente no mundo. A enorme quantidade de dados disponíveis condiciona as investigações, além de que, esta enorme quantidade de dados precisam ser interpretados e transformados em informação. Neste contexto são várias as investigações que são realizadas em Moçambique e no Mundo dando origem a dados que podem ser analisados sob diferentes óticas. Estas podem ser conduzidas através de um censo, que envolve a colheita de informações sobre todas as unidades da população, ou por amostragem, que é um conjunto de métodos que permitem a observação de informações de algumas unidades selecionadas aleatoriamente com o objetivo de inferir parâmetros para a população. Estes métodos selecionam as unidades que serão observadas e estimam, com um grau de precisão, as características dos parâmetros à partir das medidas da parte observada (amostra).

Para o desenvolvimento de um estudo por amostragem probabilística é necessário ter um conhecimento sobre o conceito plano amostral probabilístico. Este resume-se num instrumento que comporta a definição da população alvo, a base de amostragem, as técnicas de amostragem, o tamanho da amostra e a informação requerida. "As técnicas estatísticas tradicionais formuladas para modelar e analisar dados são realizações de variáveis aleatórias independentes e identicamente distribuídas (i.i.d) (Scott e Holt, 1982; Skinner, 1986). No entanto, dados desse tipo são raros na prática (Chambers e Skinner, 2003). Essas técnicas não levam em consideração os planos amostrais complexos que frequentemente são empregados para a obtenção dos dados" (Skinner *et al*., 1989).

Uma amostra complexa consiste numa combinação de vários métodos probabilísticos de amostragem para a seleção de uma amostra representativa da população (Szwarcwald *et al*., 2008). Estas amostras têm pelo menos uma das seguintes características: estratos, conglomerados, probabilidades de seleção desiguais, ajustamentos para compensar as

não respostas e outras pós-estratificações (Lavrakas, 2008, pág. 113-115). Com estes procedimentos, a fórmula para estimar a variância do estimador tende a ser complicada, principalmente quando a amostra foi retirada em múltiplas etapas de grupos sem reposição (Lohr, 2010, pág. 281). Os pesos amostrais e o efeito do delineamento de amostragem são geralmente usados para solucionar o problema. Logo, é necessário incorporar estas características do plano amostral na análise descritiva ou analítica dos dados (Heeringa *et al.*, 2010).

No entanto, para que os resultados de uma amostra sejam válidos é necessário que disponham de bons estimadores pontuais dos parâmetros de interesse, assim como bons estimadores de variância desses mesmos estimadores. O cálculo das estimativas pontuais, como da média populacional e do total podem ser facilmente estimados a partir dos pesos da amostra. Estimar variâncias é mais complexo, uma vez que de um modo geral não existe uma expressão analítica para um estimador centrado e eficiente da variância do estimador.

Os estimadores da variância são de grande relevância neste contexto e por isso foram objecto de estudo deste trabalho. Neste sentido, torna-se essencial tomar as devidas precauções durante todo o processo de análise das amostras, desde os métodos de seleção até a fase de análise dos dados. Nesta dissertação, devido a confidencialidade no acesso aos dados reais, foi adoptada uma metodologia de geração de números pseudoaleatórios para a criação do universo de estudo, "Empresas do sector do comércio", com base na informação de uma amostra representativa da estrutura do sector fornecida pelo Instituto Nacional de Estatística de Moçambique (INE-M) como motivação para a aplicação dos estimadores avaliados.

O INE-M é o órgão responsável pelas estatísticas oficiais de Moçambique. Várias são as investigações realizadas pelo INE-M a fim de conhecer melhor o cenário em que vivemos e assim servir de instrumento para auxílio na tomada de decisões dos governantes e gestores, na formulação, validação e avaliação de políticas públicas voltadas para o desenvolvimento socioeconómico e para a melhoria das condições de vida da população de uma forma geral. Uma das investigações realizadas pelo INE é o Índice de Atividade Económica (IAE) que visa analisar mensalmente, o comportamento das atividades económicas no País, nomeadamente, Comércio, Transporte, Indústria,

Hotelaria e Turismo, além de investigar com periodicidade variável características de acordo com as necessidades. O IAE é uma das pesquisas de grande porte realizada através do uso de delineamentos de amostragem complexos no País.

Nesta dissertação, escolhemos a variável "Volume de negócios mensal da empresa" (VVN) como sendo a de principal interesse, e utilizámos variáveis auxiliares para efeito de estratificação e comparação entre grupos. A variável de interesse é contínua e as auxiliares, Região, Classificador de atividade económica (CAE) e Escala de pessoal ao serviço (EPS), são todas categóricas e formam estratos naturais no domínio de estimação de interesse.

O *software* estatístico *RStudio* auxiliou na estimação da média, variância e do desvio padrão do VVN tomando em consideração quatro delineamentos de amostragem adotados para três populações em estudo.

1.1 Objectivos

O IAE é realizado por meio da seleção de uma amostra complexa representativa da população e por isso serviu como base de estudo neste trabalho que teve como objetivo principal de avaliar o desempenho, e respetivas propriedades, dos estimadores usuais da variância da média amostral em amostras complexas. Para esse efeito, com base em três conjuntos de dados, simulados a partir de dados reais, serão sorteadas 10 000 réplicas de amostras, de cada um desses conjuntos, de acordo com quatro delineamentos de amostragem.

Conhecer as propriedades dos estimadores da variância de medidas amostrais e as suas características constituem um passo primordial para intervenção em inquéritos amostrais, para garantir a precisão das estimativas consoante a estrutura do plano amostral. Iremos analisar o possível impacto do uso de um determinado plano amostral sobre a precisão dos estimadores das variáveis de interesse. Para esse efeito, vamos estudar o enviesamento e a precisão das estimativas considerando três populações com diferentes características. Pretende-se, assim, contribuir para a qualidade das investigações de amostragem em Moçambique, nas diferentes esferas da vida socioeconómica.

1.2 Estrutura do trabalho

Este trabalho encontra-se organizado em 5 capítulos, para além de uma apresentação da notação e abreviaturas utilizadas e dos apêndices.

Para além deste primeiro capítulo da introdução, contextualização e definição dos objetivos do estudo, no segundo capítulo é apresentada uma revisão bibliográfica sobre amostras aleatórias e estimadores usuais da variância em planos amostrais complexos como forma de introduzir a temática.

O terceiro capítulo aborda a metodologia usada para a geração das populações em estudo, os métodos utilizados para seleção das amostras segundo os vários delineamentos propostos, e os mecanismos para avaliar e comparar os estimadores da variância da média amostral.

No quarto capítulo descrevem-se as populações em estudo, apresentam-se e discutem-se os resultados referentes às distribuições dos estimadores obtidos por simulação.

Um resumo das conclusões acompanhado pelas recomendações em relação ao estudo desenvolvido é apresentado no quinto e último capítulo.

Capítulo 2

Conceitos fundamentais

A necessidade de conhecer uma população no que respeita a uma ou várias características impulsiona um processo de recolha e análise de informação. A dificuldade e/ou impossibilidade de estudar a totalidade da população ditou a importância do estudo por recurso a amostras. A utilização de procedimentos de amostragem probabilística permite aos investigadores utilizar as respostas dadas a entrevistas feitas a uma pequena fração de uma população (amostra) para fazer inferências sobre a população. O processo de amostragem probabilística envolve duas etapas:

- A primeira etapa baseia-se nos princípios estatísticos da teoria de probabilidade para a seleção das unidades a incluir na amostra (identificação e/ou construção da base de amostragem e escolha aleatória dos sujeitos);
- A segunda etapa está ligada à efetivação da amostra de uma forma válida e confiável (taxas de resposta e ajustamentos posteriores).

A base de amostragem operacionaliza a definição da população que se pretende estudar identificando os sujeitos da população e fornecendo as informações que possibilitem que sejam encontrados, caso sejam selecionados para a amostra.

2.1 Amostragem probabilística

Na amostragem probabilística, cada indivíduo da população tem uma probabilidade conhecida e não nula de ser selecionado. Além disso, permitem a definição de um conjunto com todas as amostras possíveis e suas respetivas probabilidades de seleção, de acordo com o processo probabilístico determinado. Existem vários processos de amostragem probabilística, sendo os mais utilizados a amostragem aleatória simples, a amostragem estratificada, e a amostragem por conglomerados (Cochran, 1977).

2.1.1 Amostragem aleatória simples

A amostra aleatória simples (AAS) é um subconjunto de indivíduos selecionados totalmente ao acaso a partir da população com *N* elementos por um processo que garante que:

1. Todos os indivíduos da população têm a mesma probabilidade de ser selecionados para a amostra; e
2. Cada subconjunto possível de indivíduos tem a mesma probabilidade de ser selecionado que qualquer outro subconjunto de indivíduos.

Este é o método mais elementar e em simultâneo o mais importante que pode ser adotado para a seleção de uma amostra. Além de ser um processo independente é também usado em procedimentos de múltiplas etapas, fornecendo a base para delineamentos complexos.

A seleção dos elementos da amostra pode ser feita de duas maneiras:

1. Amostragem aleatória simples sem reposição (AASs): ao sortearmos uma unidade da população, excluímos esta unidade do próximo sorteio. Dizemos então que as tiragens não são independentes.
2. Amostragem aleatória simples com reposição (AASc): no caso em que a unidade sorteada pode ser repetida na amostra. Nesta situação, as tiragens são independentes.

É uma forma de amostragem que dá a cada elemento da população a mesma possibilidade de ser escolhido. Portanto, a possibilidade de uma amostra *s* ser selecionada é (probabilidade de seleção):

$$P(s) = \frac{1}{C_n^N} \quad \text{(AASs)}, \tag{2.1}$$

$$P(s) = \frac{1}{N^n} \quad \text{(AASc)}. \tag{2.2}$$

A probabilidade do indivíduo *i* ser incluído na amostra *s* é (probabilidade de inclusão de 1ª ordem):

$$\pi_i = \frac{C_{n-1}^{N-1}}{C_n^N} = \frac{n}{N} \quad \text{(AASs)}, \tag{2.3}$$

$$\pi_i = 1 - \left(1 - \frac{1}{N}\right)^n \quad \text{(AASc)}. \tag{2.4}$$

O inverso da probabilidade de inclusão de 1ª ordem designa-se por peso de amostragem:

$$w_i = \frac{1}{\pi_i}, \tag{2.5}$$

e pode ser interpretado como o número de unidades da população que são representadas pela unidade *i*. Na AAS, com ou sem reposição, qualquer unidade da amostra representa o mesmo número de unidades na população.

De notar que na AASs, $\sum_{i \in s} w_i = \sum_{i \in s} \frac{N}{n} = N$ (Lohr, 2010, pág. 39).

Um estimador é uma função dos elementos da amostra, que se utiliza para estimar parâmetros. Ao valor do estimador calculado para uma amostra que se recolheu, dá-se o nome de estimava. Com base numa AAS de dimensão *n*, o estimador não enviesado para a média populacional da variável *y* em estudo é:

$$\hat{\mu} = \sum_{i=1}^{n} \frac{y_i}{n} \tag{2.6}$$

onde y_i representa o valor da variável de interesse no *i*-ésimo elemento da amostra. Este estimador pode ser reescrito à custa dos pesos de amostragem:

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{n} \frac{N}{n} y_i = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} \tag{2.7}$$

A variância do estimador da média é:

$$Var(\hat{\mu}) = \frac{\sigma^2}{n}(1 - f), \tag{2.8}$$

onde $f = \frac{n}{N}$ representa a fração de amostragem, $(1 - f)$ o fator de correção de população finita e $\sigma^2 = \frac{1}{N-1} \sum_{k=1}^{N} (Y_k - \bar{Y})^2$ é a variância da população (corrigida).

O estimador não enviesado para a variância de $\hat{\mu}$ é:

$$\widehat{Var}(\hat{\mu}) = \frac{S^2}{n}(1-f), \tag{2.9}$$

onde $S^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(y_j - \bar{y}\right)^2$ é a variância da amostra (corrigida).

No caso da AASc o termo f é aproximadamente zero, e por essa razão $(1-f)$ deixa de fazer parte nas expressões (2.8) e (2.9).

A AASs é mais intuitiva e eficiente do que AASc, exceto quando o tamanho da amostra é igual a 1 e não existe diferença. Porém a AASc, por resultar em independência entre as observações, tem vantagens estatísticas e matemáticas que facilitam a determinação das propriedades dos estimadores e das quantidades populacionais de interesse. Portanto, a AASc é muito adotada como pressuposto básico para os métodos estatísticos apresentados na maioria dos Manuais de Estatística. Quando a população é muito grande, a diferença entre AASs e AASc torna-se desprezável (Vieira, 2013).

A AAS é um processo de amostragem muito simples. No entanto, numa situação que exija entrevistas pessoais, corre-se o risco de obter uma amostra muito dispersa geograficamente, morosa sem ganhos significativos em termos de precisão dos resultados comparativamente a processos de amostragem mais elaborados (Vicente *et al.*, 2001, pág. 52).

2.1.2 Amostragem aleatória estratificada

A amostragem aleatória estratificada (AAE) consiste basicamente na divisão da população em *K* grupos bem definidos (estratos) mútua e exaustivamente exclusivos, sendo retirada uma amostra aleatória de n_i elementos de cada estrato. A amostra total de *n* elementos é o somatório das subamostras retiradas de cada estrato (Levy e Lemeshow, 1991).

O objetivo da estratificação de uma população é de reduzir a variabilidade dos estimadores e assim obter estimativas mais precisas. O que se pretende é a criação de estratos/grupos que originem grupos muito homogéneos internamente mas muito

diferentes dos outros estratos, ou seja, que a variância total seja essencialmente explicada pela variância entre os estratos (Vicente *et al.*, 2001, pág. 58).

Consideremos $\hat{\mu}_E$ como estimador não enviesado para a média populacional.

$$\hat{\mu}_E = \sum_{i=1}^{K} W_i \bar{y}_i \tag{2.10}$$

onde $W_i = \frac{N_i}{N}$ representa o peso do estrato *i*, N_i o número total dos elementos do estrato i, e $\bar{y}_i$ a média amostral do estrato i, $i = 1, \dots, K$.

A alocação da amostra nos estratos pode ser feita de formas distintas. A alocação igual considera amostras de igual dimensão para todos os estratos, sendo:

$$n_i = \frac{n}{K}. \tag{2.11}$$

Na alocação proporcional, as dimensões das amostras são proporcionais aos tamanhos dos estratos, ou seja:

$$n_i = \left(\frac{N_i}{N}\right) n\,. \tag{2.12}$$

A alocação ótima de *Neyman* mostra que o número ideal de unidades a serem observadas no estrato i é diretamente proporcional a $N_i\sigma_i$, isto é:

$$n_i = \frac{N_i\sigma_i}{\sum_{k=1}^{K} N_k\sigma_k}, \tag{2.13}$$

sendo σ_i o desvio padrão da variável de interesse no estrato *i.*

A variância do estimador é:

$$Var(\hat{\mu}_E) = \sum_{i=1}^{K} W_i^2 \frac{\sigma_i^2}{n_i}(1 - f_i), \tag{2.14}$$

com $f_i = \frac{n_i}{N_i}$ taxa de amostragem por estrato.

Um estimador não enviesado para a variância do estimador é:

$$\widehat{Var}(\hat{\mu}_E) = \sum_{i=1}^{K} W_i^2 \frac{S_i^2}{n_i}(1 - f_i). \tag{2.15}$$

sendo S_i^2 a variância amostral por estrato *i*.

A AAE quase sempre é estatisticamente mais eficiente do que a AAS, sendo que quanto mais homogêneos forem os subgrupos maior é a eficiência do plano amostral. O facto de os subgrupos serem mais homogêneos internamente do que a população como um todo, proporciona uma redução do erro amostral no geral. Este aumento da precisão das estimativas permite reduzir a amostra para um nível de precisão fixo.

Na maioria dos inquéritos por amostragem, os pesos são usados para calcular estimativas pontuais. Na AAE o peso amostral da unidade j do estrato i é $w_{j|i} = \left(\frac{N_i}{n_i}\right)$, pois a probabilidade de inclusão é $\pi_{j|i} = \frac{n_i}{N_i}$, e o estimador da média populacional (2.10) pode ser reescrito da seguinte forma:

$$\hat{\mu}_E = \frac{\sum_{i=1}^{K} \sum_{j=1}^{n_i} w_{j|i}\, y_{ij}}{\sum_{i=1}^{K} \sum_{j=1}^{n_i} w_{j|i}} \tag{2.16}$$

onde n_i é a dimensão da amostra por estrato.

Tendo presente que a variância total de uma população pode ser decomposta em variância entre estratos somada à variância dentro dos estratos, na estratificação pretende-se que a variância total seja fundamentalmente explicada pela variância entre os estratos. A estratificação é eficaz quando na população existem valores extremos para a característica em estudo, sendo possível agregá-los num estrato separado, o que permite estimar os parâmetros dentro de cada estrato (Vicente *et al.*, 2001, pág. 58)

2.1.3 Amostragem por conglomerados (grupos)

A amostragem por conglomerados (AC) é utilizada na maioria das vezes quando não temos acesso a uma base de amostragem digna de confiança que identifique cada elemento da população ou quando é muito trabalhoso ou dispendioso o deslocamento para se observar cada elemento, devido às distâncias geográficas entre as mesmas, por exemplo (Cochran, 1977).

A AC é menos eficiente que a AAS, logo seria lógico pensar na utilização da AAS antes de tudo. No entanto, a AC gera estimativas com precisão aceitável se for bem conduzida, o que inclui a busca por maior heterogeneidade dentro dos conglomerados e maior homogeneidade entre os conglomerados, sendo assim muito útil, especialmente quando a população for extensa (Cochran, 1965, pág. 318).

Na AC os indivíduos estão agrupados naturalmente de acordo com algum critério. Designam-se por unidades primárias de amostragem (UPA) os conglomerados, isto é, os agrupamentos de indivíduos, e por unidades secundárias (USA) os indivíduos que compõem os conglomerados. A amostragem por conglomerados consiste em selecionar uma amostra aleatória simples inicial de *n* conglomerados (UPA) em vez de elementos individuais.

O tamanho e forma dos grupos podem afetar a eficiência. Para produzir uma boa estimativa, o tamanho do grupo pode ser usado como informação auxiliar para a seleção de grupos com probabilidades desiguais.

- **Grupos de igual dimensão**

No caso dos estimadores para grupos de igual dimensão temos que, o estimador para a média populacional $\hat{\mu}_C$ é,

$$\hat{\mu}_C = \frac{1}{m}\sum_{i=1}^{m} \bar{y}_i \tag{2.17}$$

onde *m* designa o número de grupos amostrados e $\bar{y}_i$ a média amostral do grupo *i*.

A variância do estimador da média é:

$$Var(\hat{\mu}_C) = \frac{1-f}{m}\frac{1}{M-1}\sum_{i=1}^{M}(\bar{Y}_i - \bar{Y})^2 \tag{2.18}$$

onde $f = \frac{m}{M}$ é a taxa de amostragem dos grupos, *M* representa o número de grupos na população, $\bar{Y}_i$ a média do grupo *i*, $i=1,\ldots,M$ e $\bar{Y}$ é a média global.

Um estimador não enviesado para a variância do estimador é dado por:

$$\widehat{Var}(\hat{\mu}_C) = \frac{1-f}{m}\frac{1}{m-1}\sum_{i=1}^{m}(\bar{y}_i - \bar{y})^2 \tag{2.19}$$

sendo $\bar{y}$ a média global estimada.

- **Grupos de dimensão diferente**

Quando os grupos têm dimensão diferente esta informação auxiliar pode ser usada para selecionar grupos com probabilidades diferentes ou para usar estimadores rácio. Passamos então a ter duas alternativas para estimar a média, total e variância dos estimadores.

❖ Opção A (estimador rácio)

O estimador para a média populacional é:

$$\hat{\mu}_C = \frac{1}{\sum_{i=1}^{m} n_i}\sum_{i=1}^{m} y_{T_i} \tag{2.20}$$

onde y_{T_i} é o total da característica de interesse dos indivíduos do grupo *i*.

A variância do estimador da média é:

$$Var(\hat{\mu}_C) = \frac{(M-m)M}{mN^2}\frac{1}{M-1}\sum_{i=1}^{M}\left(Y_{T_i} - N_i\bar{Y}\right)^2 \tag{2.21}$$

sendo N_i o número total dos elementos do grupo *i*.

Um estimador não enviesado para a variância do estimador é dado por:

$$\widehat{Var}(\hat{\mu}_C) = \frac{(M-m)M}{mN^2}\frac{1}{m-1}\sum_{i=1}^{m}\left(y_{T_i} - N_i\hat{\mu}_C\right)^2 \tag{2.22}$$

❖ Opção B (usa a teoria da AAS)

O estimador não enviesado para a média populacional é:

$$\hat{\mu}_C = \frac{M}{Nm}\sum_{i=1}^{m} y_{T_i} \tag{2.23}$$

A variância do estimador da média é:

$$Var(\hat{\mu}_C) = \frac{(M-m)M}{mN^2}\frac{1}{M-1}\sum_{i=1}^{M}\left(Y_{T_i} - \bar{Y}_T\right)^2 \quad (2.24)$$

onde Y_{T_i} é o total da característica de interesse do grupo *i* e $\bar{Y}_T$ é a média dos totais dos grupos.

Um estimador não enviesado para a variância do estimador é dado por:

$$\widehat{Var}(\hat{\mu}_C) = \frac{(M-m)M}{mN^2}\frac{1}{m-1}\sum_{i=1}^{m}\left(y_{T_i} - \bar{y}_T\right)^2 \quad (2.25)$$

sendo $\bar{y}_T$ a média dos totais estimados para os grupos.

Estando em causa amostras de unidades elementares da mesma dimensão, o desvio padrão das estimativas obtidas a partir de um delineamento de amostragem por grupos é maior comparativamente ao obtido com outros esquemas de amostragem (Levy e Lemeshow, 1991).

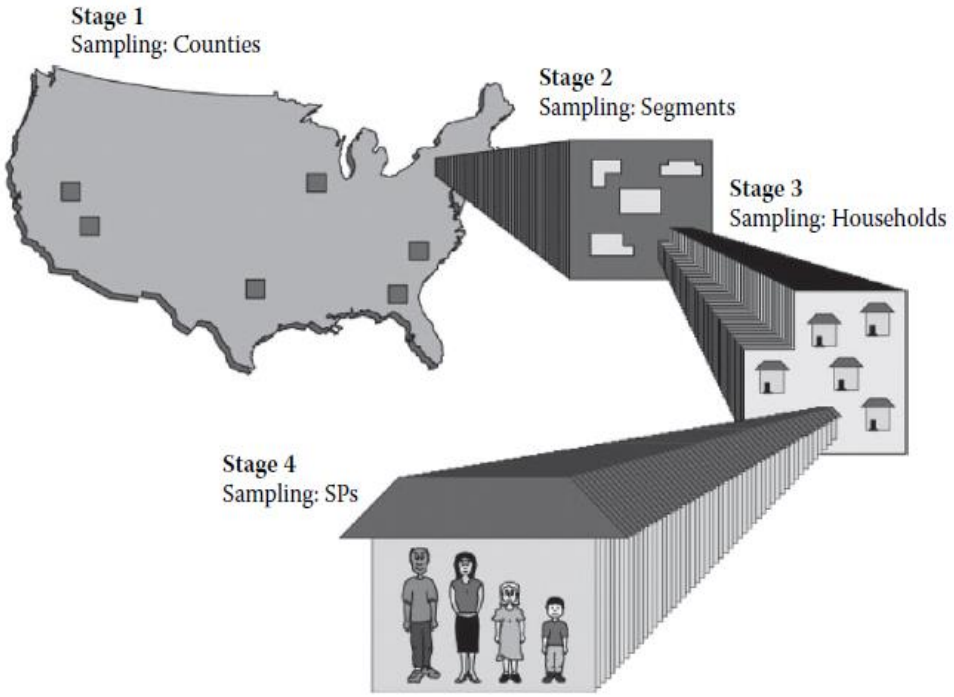
2.1.4 Amostragem multietápica

A amostragem multietápica é uma extensão da amostragem por grupos, com maior flexibilidade. Este esquema de amostragem consiste em:

- Numa 1.ª etapa selecionamos segundo um esquema aleatório simples (ou outro) os grupos da população a inquirir, a que chamamos unidades primárias de amostragem (*UPA*s);
- Numa 2.ª etapa sorteamos elementos, ou subconjuntos de elementos de cada uma das *UPA*s selecionadas na 1ª etapa, que são unidades secundárias de amostragem (*USA*s), utilizando ou não o mesmo plano amostral.
- O processo pode repetir-se ao longo de várias etapas.

Na Ilustração 2.1 exemplifica-se o delineamento acima descrito muito aplicado na amostragem aos domicílios. Na fase primária de amostragem são selecionados municípios ou grupos dos municípios adjacentes, na segunda fase de amostragem são selecionados segmentos de área, na terceira fase escolhem-se unidades habitacionais

dentro dos segmentos de área selecionados na etapa anterior, e conclui-se com a seleção aleatória dos elementos elegíveis nas unidades habitacionais selecionadas. Este tipo de esquemas, com pequenas alterações no número de etapas e na escolha de unidades de amostragem, são usados em toda América, África, Ásia e Europa (Heeringa and O'Muircheartaigh, 2010).



Fonte: Heeringa and O'Muircheartaigh, 2010

Ilustração 2.1 Esquema de amostragem probabilística em múltiplas etapas

No caso da amostragem por conglomerados bietápica, quando a amostra é por conglomerados com probabilidades desiguais temos que, π_i representa a probabilidade da *i*-ésima *UPA* estar na amostra, e $\pi_{j|i}$ a probabilidade do elemento de ordem *j* da *i*-ésima *UPA* estar na amostra, dado que esta *UPA* pertence a amostra. Então os pesos de amostragem são dados por:

$$w_{ij} = \frac{1}{(\pi_i \pi_{j|i})}. \tag{2.26}$$

Na amostragem por conglomerados em três estágios, o princípio estende-se:

$$w = w_p \times w_{s|p} \times w_{t|s,p} \tag{2.27}$$

onde w_p representa os pesos da *UPA*, $w_{s|p}$ representa os pesos da *USA* e $w_{t|s,p}$ são os pesos associados a *UTA* (unidade terciaria de amostragem).

O estimador genérico para a média da população é:

$$\widehat{\bar{\mu}} = \frac{\hat{t}_y}{\sum_{i \in s} w_i}. \tag{2.28}$$

onde $\hat{t}_y = \sum_{i \in s} w_i \, y_i$ é o estimador geral do total da população e $\sum_{i \in s} w_i$ estima o número de unidades de observação na população.

Mas os pesos de amostragem não dão nenhuma informação sobre como encontrar o erro padrão da estimativa, uma vez que este depende das probabilidades conjuntas de qualquer par de unidades de observação ser selecionado para estar na amostra. Portanto, para a estimação do erro padrão é requerido mais conhecimento sobre o plano de amostragem do que somente os pesos de amostragem.

2.2 Amostras complexas

Uma amostra complexa consiste numa combinação de vários métodos probabilísticos de amostragem para a seleção de uma amostra representativa da população (Szwarcwald *et al.*, 2008). Estas amostras têm pelo menos uma das seguintes características: estratos, conglomerados, probabilidades de seleção desiguais e ajustamentos para compensar as não respostas e outras pós-estratificações (Lavrakas, 2008, pág. 113-115). Com estes delineamentos, a fórmula para estimar a variância do estimador tende a ser complicada, principalmente quando a amostra foi retirada em múltiplas etapas de grupos sem reposição (Lohr, 2010, pág. 281). Os pesos amostrais e o efeito do delineamento de amostragem são geralmente usados para solucionar o problema. Logo, é necessário incorporar estas características do plano amostral na análise descritiva ou analítica dos dados (Heeringa *et al.*, 2010).

A análise de dados proveniente de amostras complexas apresenta dois desafios principais:

- Obter estimativas pontuais corretas;
- Estimar corretamente a variância e o desvio padrão;

As três principais características – estratos, conglomerados e pesos – possíveis de estarem presentes numa amostragem complexa têm diferentes efeitos na estimativa pontual e na estimativa da variância (Kreuter e Valliant, 2007).

2.3 Estimadores da variância em amostras complexas

Em amostragem, assim como na Estatística Clássica, a estimação de variâncias é uma componente essencial da abordagem inferencial adotada: sem estimativas da variância, nenhuma indicação da precisão (e portanto, da qualidade) das estimativas de interesse está disponível. Nesse contexto, uma tentação que assola muitos utilizadores incautos é esquecer que os resultados são baseados em dados apenas de uma amostra da população, e portanto sujeitos a incerteza, que não pode ser quantificada sem medidas de precisão amostral (Pessoa e Silva, 1998, pág. 37).

O cálculo da média populacional e do total podem ser facilmente estimados a partir dos pesos da amostra. Estimar variâncias é um processo mais complexo, pois em amostras complexas com várias etapas de estratificação e conglomerados, a variância do estimador da média e do total deve ser calculada para cada nível e, em seguida, deve ser combinado segundo o delineamento do inquérito (Lohr, 2010, pág. 365).

A teoria e aplicação de investigações por amostragem têm crescido dramaticamente nos últimos 60 anos. Estas investigações incidem sobre quase todos campos de estudo científico, incluindo agricultura, demografia, educação, energia, transporte, saúde, economia, política, sociologia e assim por diante. Um requisito básico em todas formas de análise, senão a principal exigência nas investigações práticas, é que uma medida da precisão deve ser fornecida para cada estimativa derivada dos dados de investigação. A medida mais usada da precisão é a variância do estimador (Wolter, 2007, pág. 1).

Como um preliminar para qualquer discussão, é importante reconhecer que a variância de um levantamento estatístico é uma função tanto da forma da estatística como da natureza do delineamento amostral (Wolter, 2007, pág. 1).

Estimar a variância em amostras complexas é um tema atual e bastante complexo em termos gerais, uma vez que de um modo geral não existe uma expressão analítica para um estimador centrado e eficiente da variância do estimador. Neste contexto de dimensões de investigações complexas, surge a questão, como determinar um estimador adequado da variância de um dado estimador. A escolha é normalmente difícil,

envolvendo a estimação da variância, do enviesamento e do erro quadrático médio devido à complexidade dos esquemas de amostragem.

Neste âmbito existem vários métodos para estimar a variância dos totais estimados, mas basicamente existem duas abordagens:

- Analítica usando o método de linearização Taylor;
- Métodos de reamostragem ou replicação (*Jackknife,* réplicas equilibradas repetidas (*BRR*) e *Bootstrap*).

A escolha de um método para estimar a variância envolve um equilíbrio de fatores tais como a precisão, custo e flexibilidade. Nenhum dos métodos para estimar a variância do estimador é o melhor no geral (Wolter, 2007, pág. 366). Por isso, num bom julgamento no qual está envolvida a escolha de um método para estimar a variância, não será surpresa se o estatístico recomendar métodos diferentes para diferentes aplicações da investigação.

Neste trabalho iremos tomar em consideração os mais usuais para estimação da variância, o método de *Linearização em série de Taylor* e os métodos de replicação (reamostragem) *Jackknife* e *Bootstrap,* mostrando em que circunstâncias cada um deles melhor se adequa ao delineamento.

2.3.1 Método de Linearização de Taylor

A maior parte das fórmulas de variância que conhecemos são para estimadores de médias e totais. Mas estas fórmulas podem ser utilizadas para encontrar variações para qualquer combinação linear de médias e totais, ou para parâmetros que não sendo combinações linear de totais ou médias podem ser aproximados por uma dessas combinações lineares (Lohr, 2010, pág. 366). Esta é a abordagem de linearização de Taylor.

A seguir as etapas para construção de um estimador de linearização da variância de uma função não-linear de médias ou totais:

- Expressar a quantidade de interesse em função da média ou total das variáveis na amostra. Em geral, $\theta = h(t_1, \dots, t_k)$ ou $\theta = h(\bar{y}_1, \dots, \bar{y}_k)$.
- Encontrar a derivada parcial de *h* no que diz respeito a cada argumento. Com as derivadas parciais avaliadas em quantidades populacionais, formar as constantes de linearização a_j.
- Aplicar a linearização de Taylor para a estimativa:

$$h(\hat{t}_1, \hat{t}_2, \dots, \hat{t}_k) \approx h(t_1, \dots, t_k) + \sum_{j=1}^{k} a_j(\hat{t}_j - t_j) \tag{2.29}$$

onde:

$$a_j = \frac{\partial h(c_1, c_2, \dots, c_k)}{\partial c_j}_{|t_1, t_2, \dots, t_k} \tag{2.30}$$

- Definir a nova variável *q* através de:

$$q_i = \sum_{j=1}^{k} a_j y_{ij} \tag{2.31}$$

- Estimar o total com:

$$\hat{t}_q = \sum_{i \in s} w_i q_i = \sum_{j=1}^{k} a_j \hat{t}_j \tag{2.32}$$

A variância estimada de $\hat{t}_q$, pode ser obtida fazendo a substituição do estimador $\hat{t}_j$ pelas estimativas aproximadas da população.

$$Var\left(\sum_{j=1}^{k} a_j \hat{t}_j\right) = Var(\hat{t}_q) = \sum_{j=1}^{k} a_j^2 \, Var(\hat{t}_j) + 2 \sum_{j=1}^{k-1} \sum_{l=j+1}^{k} a_j a_l Cov(\hat{t}_j, \hat{t}_l). \tag{2.33}$$

cujo estimador é:

$$\widehat{Var}_T(\hat{\theta}) = \sum_{j=1}^{k-1} \sum_{l=j+1}^{k} a_j a_l \widehat{Cov}(\hat{t}_j, \hat{t}_l). \tag{2.34}$$

Neste método, sabemos que, se as derivadas parciais são conhecidas, a linearização quase sempre dá uma estimativa de variância para uma estatística e podem ser aplicadas em modelos gerais de amostragem. Os métodos de linearização têm um longo histórico de utilização e a teoria está bem desenvolvida.

Uma desvantagem deste método reside no facto dos cálculos poderem ser confusos, e o método ser difícil de aplicar para funções complexas que envolvem pesos. Temos que encontrar expressões analíticas para as derivadas parciais de *h* ou calcular as derivadas parciais numericamente, o que pode exigir muita programação porque para o cálculo de cada estatística é necessário um método diferente. Além disso, nem todas as estatísticas podem ser expressas como uma função suave dos totais populacionais, sendo a mediana e outros quartis exemplo disso (Koop, 1972). Finalmente a precisão aproximada da linearização depende do tamanho da amostra o que origina a que a variância do estimador por vezes seja subestimada se a amostra não for grande o suficiente.

A aproximação de primeira ordem é amplamente utilizada em investigações por amostragem de populações finitas. A experiência tem mostrado que, quando o tamanho da amostra é suficientemente grande e onde os conceitos do delineamento de investigação eficiente são aplicados com êxito, a primeira ordem de expansão em série de *Taylor* proporciona frequentemente aproximações confiáveis (Wolter, 2007, pág. 233).

2.3.2 Métodos de Reamostragem e Réplicas

O método de reamostragem requer que se selecionem duas ou mais subamostras a partir de uma determinada população, ou eventualmente, de uma amostra, e calcular em separado as estimativas do parâmetro populacional de interesse para cada amostra. As estimativas da variância são feitas a partir da combinação de todas amostras. Os métodos de reamostragem diferem na maneira de gerar réplicas de amostras em planos de investigação complexos (Münnich, 2005, pág. 69).

Nos métodos de reamostragem calculam-se as estimativas da variância de uma amostra em que são amostradas *UPAs* com reposição. Se *UPAs* são amostrados sem reposição, estes métodos ainda podem ser usados, mas espera-se que sobrestimem a variância o que resulta em intervalos de confiança conservativos (Lohr, 2010, pág. 373).

2.3.2.1. Réplicas Repetidas de Jackknife

Este método de reamostragem é uma extensão do método dos grupos aleatórios, que permite replicar grupos que se sobrepõem. Foi criado por *Quenouille* (1956) como um método para reduzir o enviesamento dos estimadores, num contexto da Estatística Clássica. Tukey (1958) propôs a usá-lo para estimar variâncias e calcular intervalos de confiança.

Para uma amostra aleatória simples (AAS), seja $\hat{\theta}_{(j)}$ um estimador com a mesma forma de $\hat{\theta}$ sem a observação *j*. Definimos o estimador *Jackknife* "*delete* - 1" (assim chamado porque temos de eliminar uma observação em cada repetição) como:

$$\widehat{Var}_{JK}(\hat{\theta}) = \frac{n-1}{n}\sum_{j=1}^{n}\left(\hat{\theta}_{(j)} - \hat{\theta}\right)^2. \tag{2.35}$$

Quando $\hat{\theta} = \bar{y}$ temos que:

$$\hat{\theta}_{(j)} = \bar{y}_{(j)} = \frac{1}{n-1}\sum_{i\neq j} y_i = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i - y_j\right) = \bar{y} - \frac{1}{n-1}\left(y_j - \bar{y}\right). \tag{2.36}$$

e

$$\sum_{j=1}^{n}\left(\bar{y}_{(j)} - \bar{y}\right)^2 = \frac{1}{(n-1)^2}\sum_{j=1}^{n}\left(y_j - \bar{y}\right)^2 = \frac{1}{n-1}s_y^2. \tag{2.37}$$

Assim, $\widehat{Var}_{JK}(\hat{\mu}) = \frac{s_y^2}{n}$, o que corresponde ao estimador da variância de $\hat{\mu}$ da AASc, e justifica a razão de ser do multiplicador $\frac{n-1}{n}$ na expressão (2.35).

O estimador *Jackknife* é um método para todos os fins. O mesmo procedimento é usado para estimar a variância para cada estatística para o qual o *Jackknife* pode ser usado. É aplicado para amostras de múltiplas etapas, estratificadas onde mais de duas *UPAs* são amostradas em cada estrato.

O método *Jackknife* fornece um estimador consistente da variância quando $\hat{\theta}$ é uma função suave dos totais populacionais (Krewski e Rao, 1981). Além disso, pode ser

usado para quantificar os efeitos da imputação nas estimativas de variância (Rao e Shao, 1992). No entanto, este estimador tem um fraco desempenho na estimação de variâncias de algumas estatísticas que não são funções de suaves dos totais populacionais (por ex., exemplo, não fornece um estimador consistente da variância dos quartis em uma AAS). Para alguns delineamentos de amostragem, como o estratificado, o *Jackknife* pode exigir muita computação. Além disso, suas propriedades são razoáveis para vários outros casos de estimadores não lineares de interesse (veja, por exemplo, Cochran, 1977, pág. 321 e Wolter, 2007, pág. 306).

2.3.2.2. Réplicas de Bootstrap

Assim como acontece com o estimador *Jackknife*, os resultados teóricos para o *Bootstrap* foram inicialmente desenvolvidos para áreas de investigação diferentes da amostragem estatística. Para Shao e Tu (1995) os resultados teóricos para o *Bootstrap* resumem-se num método de amostragem complexo.

Suponhamos que *s* é uma AASc de tamanho *n*. Esperamos que, no delineamento da amostra, possamos reproduzir as propriedades de toda a população. Vamos então, tratar a amostra *s*, como se fosse uma população, e retirar réplicas de amostras de *s*. Se a amostra é realmente semelhante à população, então as amostras geradas a partir da função de massa de probabilidade empírica devem se comportar como amostras retiradas da população (Lohr, 2010, pág. 384).

Se a AAS original é sem reposição, Gross (1980) propõe a criação $\frac{N}{n}$ cópias da amostra de modo a formar uma "pseudo-população" e, em seguida, retirar *R* AASs da pseudo-população. Se $\frac{n}{N}$ é pequeno, com ou sem reposição as distribuições de *Bootstrap* devem ser semelhantes.

De seguida apresentam-se os passos de Rao e Wu, (1988) para realizar uma reamostragem *Bootstrap* no caso de um delineamento com múltiplas etapas, descrito em Lohr (2010, pág. 285). Seja n_i o número de *UPAs* amostradas no estrato *i* e *R* o número de réplicas *Bootstrap* a serem criadas (tipicamente, R = 500 ou 1000).

1. Para *r*-ésima réplica de *Bootstrap r* (*r* = 1,…,*R*), selecionar uma AASc de n_i - 1 *UPAs* da amostra com n_i *UPAs* no estrato *i*. Realizar este processo de

forma independente para cada estrato. Seja $m_{ij}(r)$ o número de vezes que a *j*-ésima *UPA* do estrato *i* é selecionada na réplica *r*.

2. Para cada uma das réplicas *r*, criar o vetor de réplicas dos pesos dados por:

$$w_l(r) = w_l * \frac{n_i}{n_i - 1} m_{ij}(r) \tag{2.38}$$

para a observação *l* da *j*-ésima UPA no estrato *i*. Portanto teremos *R* vectores de pesos replicados;

3. Usar os vectores de pesos replicados para, estimar $Var(\hat{\theta})$. Seja $\hat{\theta}_r^*$ o estimador de θ, calculado da mesma forma que $\hat{\theta}$ mas usando os pesos $w_l(r)$ em vez dos pesos originais w_l. Então,

$$\widehat{Var}_B(\hat{\theta}) = \frac{1}{R-1} \sum_{l=1}^{R} (\hat{\theta}_r^* - \hat{\theta})^2 \tag{2.39}$$

O *Bootstrap* pode ser usado tanto para funções suaves da média populacional como para algumas funções não suavizadas tais como quartis em delineamentos gerais de amostragem. Permite encontrar intervalos de confiança diretamente a partir $\hat{\theta}_1^*, \hat{\theta}_2^*, ..., \hat{\theta}_R^*$ bastando considerar para os limites do intervalo os percentis 5 e 95, $\hat{\theta}_1^*, \hat{\theta}_2^*, ..., \hat{\theta}_R^*$ ou utilizar o método de *Bootstrap-t* (Efron, 1982).

Para algumas configurações, o *Bootstrap* pode exigir mais cálculos do que o método de *Jackknife*, uma vez que tipicamente *R* representa um número muito grande de réplicas. No entanto, em outras investigações de larga escala se, por exemplo, for retirada uma AAE, o *Bootstrap* pode exigir menos cálculos do que o *Jackknife*. De notar que, a estimativa de variância *Bootstrap* difere quando é retirado um conjunto diferente de amostras *Bootstrap*.

2.4 Propriedades e critérios de avaliação dos estimadores

Quando se pretende fazer inferência da amostra para a população e se dispõem de várias técnicas de estimação, existem critérios que ajudam na escolha do estimador mais

adequado, no sentido de que este devera fornecer estimativas mais próximas do valor do parâmetro desconhecido da população. Geralmente, para se analisar a qualidade dos estimadores recorre-se a duas propriedades fundamentais: o enviesamento e a precisão.

- Não-enviesamento

Um estimador $\hat{\theta}$ diz-se não-enviesado ou centrado para o parâmetro θ se: $E(\hat{\theta}) = \theta$. O enviesamento amostral é dado pela distância entre os valores médios, ou valor esperado, da distribuição de amostragem, e o verdadeiro valor do parâmetro, isto é,

$$Env(\hat{\theta}) = E(\hat{\theta}) - \theta. \tag{2.40}$$

Naturalmente, o desejável é que o estimador seja centrado ou não-enviesado, ou seja, que $Env(\hat{\theta}) = 0$.

- Precisão

Para avaliar a dispersão da distribuição amostral do estimador usamos geralmente a variância ou o desvio padrão que se definem, respectivamente por:

$$Var(\hat{\theta}) = E\left[\left(\hat{\theta} - E(\hat{\theta})\right)^2\right], \tag{2.41}$$

$$\sigma_{\hat{\theta}} = \sqrt{Var(\hat{\theta})}. \tag{2.42}$$

Geralmente o desvio padrão do estimador designa-se por erro padrão. Um estimador é tanto mais eficiente quanto menor for a sua variância.

O quociente entre o desvio padrão do estimador e o seu valor esperado designa-se coeficiente de variação do estimador:

$$CV(\hat{\theta}) = \frac{\sigma_{\hat{\theta}}}{E(\hat{\theta})} \times 100. \tag{2.43}$$

A precisão de um estimador é, habitualmente expressa pelo erro quadrático médio (*EQM*) o qual é definido por:

$$EQM(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right] = Var(\hat{\theta}) + \left[Env(\hat{\theta})\right]^2. \tag{2.44}$$

Um estimador é tanto melhor quanto menor for o seu *EQM*. Portanto, o estimador com menor *EQM* é o preferido (Wolter, 2007, pág. 3). O *EQM* indica o quão próximo está o estimador do verdadeiro valor.

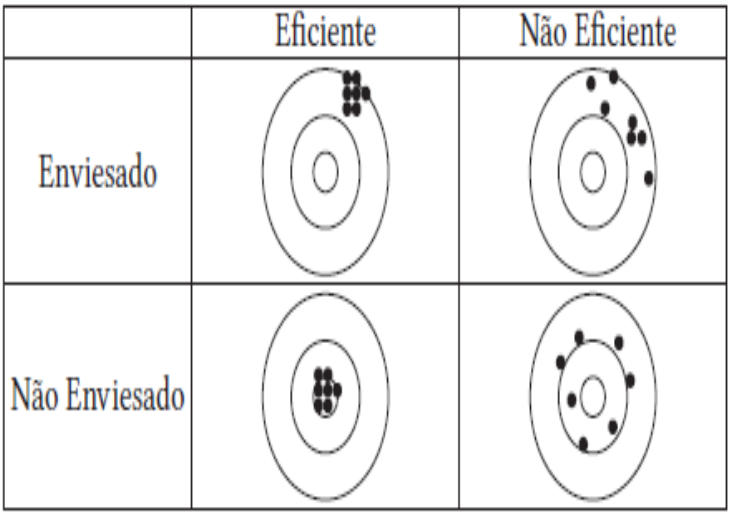
Quando se pretende comprar o desempenho de um estimador em populações distintas com características diferentes, devemos tomar em atenção as escalas de medida de acordo com as características da variável analisada. Para tornar os resultados comparáveis, precisamos escalar a $\sqrt{EQM(\hat{\theta})}$ dividindo pelo $E(\hat{\theta})$, que habitualmente é designada pela raiz do erro quadrático médio escalado ($REQME$) a qual é definida por:

$$REQME(\hat{\theta}) = \frac{\sqrt{EQM(\hat{\theta})}}{E(\hat{\theta})} \times 100. \tag{2.45}$$

As medidas de desempenho escalado combinam o enviesamento e a precisão para definir o desempenho do estimador sempre que os resultados de populações com diferentes características são comparados (Walther e Moore, 2005).

A Ilustração 2.2 ilustra duas das propriedades desejáveis nos estimadores: o não enviesamento e a eficiência. No entanto, em algumas situações, a utilização de um estimador com um enviesamento moderado é preferível, pelos seguintes motivos:

- Muitos parâmetros têm uma estrutura formal que dificulta a determinação de um estimador centrado;
- Um estimador com um enviesamento moderado pode muitas vezes ter variância e erro quadrático médio inferior ao estimador centrado.



Fonte: Afonso e Nunes, 2010.

Ilustração 2.2 Esquema do enviesamento e da precisão, sendo o verdadeiro valor o centro da circunferência menor.

Capítulo 3

Materiais e métodos

A abordagem acerca da metodologia utilizada para a geração da população de estudo, a informação de cada um dos planos de amostragem adotados e a metodologia usada para avaliação dos estimadores da variância encontram-se apresentados neste capítulo para consolidar os fundamentos teóricos com o intuito de fundamentar as discussões.

3.1 Caracterização geral da população real

A população alvo corresponde às empresas do sector formal com localização em Moçambique. Existem várias maneiras de definir o sector formal, sendo a mais objetiva aquela que define como sector formal as empresas registadas nas fontes administrativas que vão servir de base para atualização do FUE, que é o quadro estatístico utilizado como base de amostragem para a conceção e realização de inquéritos pelo INE de Moçambique.

A unidade estatística de observação e inquirição foram empresas que se encontram classificadas com atividade principal na secção G, do Classificador de Atividade Económica (CAE Rev-2), e que corresponde ao comércio por grosso e a retalho, reparação de veículos automóveis, motociclos e de bens de uso pessoal e doméstico:

CAE 45 - Empresas de Comércio, manutenção e reparação de veículos automóveis e motociclos;

CAE 46 - Empresas de Comércio por grosso (inclui agentes), exceto de veículos automóveis e motociclos;

CAE 47 - Empresas de Comércio a retalho, exceto de veículos automóveis e motociclos.

O território Moçambicano está dividido em 11 províncias distribuídas por três regiões: Norte, Centro e Sul. Tendo por base os resultados do Censo em 2012, verifica-se que o

sector do comércio era composto por 14961 empresas, distribuídas por todas as províncias do País, mas com uma maior concentração na região Sul e predominando as empresas da CAE 47 (Tabela 3.1).

Tabela 3.1 Distribuição do número das empresas classificadas na secção G por região, província, CAE.

Região	Província	CAE			Número total de empresas
		45	46	47	
Norte	Niassa	23	22	292	337
	Cabo Delegado	16	20	720	756
	Nampula	59	54	775	888
Centro	Zambézia	25	25	415	465
	Tete	33	18	766	817
	Manica	35	19	1175	1229
	Sofala	124	127	3289	3540
Sul	Inhambane	31	22	967	1020
	Gaza	67	19	1127	1213
	Maputo Província	52	28	869	949
	Maputo Cidade	388	378	2981	3747
Total		853	732	13376	14961

Fonte: INE, CENSO 2012

Atendendo à confidencialidade no acesso aos dados reais por parte do INE relativamente à variável de interesse e o número de pessoal ao serviço, optou-se por gerar um universo fictício de empresas, inspirado na informação real existente, que será tomado como a população de referência para o desenvolvimento do estudo.

3.2 Simulações

Com o objetivo de estudar o comportamento dos estimadores iremos considerar para o estudo três populações e quatro delineamentos de amostragem distintos.

3.2.1 Geração da população

Como ponto de partida para a geração de três populações, com 10 000 empresas fictícias em cada uma, foram consideradas quatro variáveis fundamentais:

- ***Região*** - a região do país onde a empresa está localizada,
- ***CAE*** - o Classificador de Atividade Económica CAE Rev-2,
- ***NPS*** - o número do pessoal ao serviço, e

- ***VVN*** - o volume de negócios da empresa.

As empresas fictícias foram distribuídas pelas três regiões do país, Norte, Centro e Sul, de acordo com a distribuição de probabilidade:

Região	Norte	Centro	Sul
f(Região)	p_1R	p_2R	$1-(p_1R+p_2R)$

onde $0 < p_1R < 1$ e $0 < p_2R < 1$ são as probabilidades de uma empresa estar localizada nas regiões Norte e Centro, respetivamente, e o termo $1-(p_1R+p_2R)$ a probabilidade de uma empresa estar localizada na região Sul.

Cada uma das empresas fictícias foi afeta aleatoriamente a um código CAE, de acordo com a distribuição de probabilidade:

CAE	45	46	47
f(CAE)	p_1C	p_2C	$1-(p_1C+p_2C)$

onde $0 < p_1C < 1$ e $0 < p_2C < 1$ são as probabilidades de uma empresa pertencer as CAEs 45 e 46, respetivamente, e o termo $1-(p_1C+p_2\,C)$ a probabilidade de uma empresa pertencer a CAE 47.

Em Moçambique a distribuição do número de pessoal ao serviço (*NPS*) é assimétrica existindo uma maior concentração de trabalhadores nas regiões onde há um maior número de empresas. Para tentar replicar este comportamento na população fictícia de empresas recorreu-se à distribuição binomial negativa, cuja função massa de probabilidade é:

$$P(X=x)=\binom{x-1}{k-1}p^k(1-p)^{x-k}, para\ x=k,k+1,\ldots \qquad (3.1)$$

onde *p* representa a probabilidade de sucesso, *k* o número de sucessos pretendidos e *x* o número total de provas a realizar.

Além disso, há um maior número empresas com menos de 50 trabalhadores (caso de novos investimentos nas pequenas empresas), um número muito reduzido de empresas com no mínimo 50 mas menos de 250 trabalhadores (situação que carece de uma análise profunda), e um número acentuado de empresas com pelo menos 250 trabalhadores

(consequência dos grandes investimentos que estão a ser realizados no País). Portanto, para replicar este comportamento foram considerados três pares de valores distintos para a probabilidade de sucesso e números de sucessos para cada um destes grupos.

Posteriormente, a variável *NPS* foi categorizada em escala de pessoal ao serviço (*EPS*) de acordo com a dimensão da empresa (Tabela 3.2).

Tabela 3.2 Categorias da variável EPS

Categoria de EPS	Condição	Classificação das empresas
1	*NPS < 50*	Pequenas
2	50 ≤ NPS < 250	Média
3	NPS ≥ 250	Grandes

O volume de negócios da empresa (*VVN*) foi gerado assumindo que este dependia linearmente do número de trabalhadores na empresa:

$$VVN = \alpha + \beta * NPS + \varepsilon \tag{3.2}$$

com os valores de $\alpha\ e\ \beta$ a variarem segundo a população e o $\varepsilon \sim N(0; \sigma \times NPS)$ e o σ é uma constante maior do que 0.

3.2.2 Extração das amostras

Nas três populações em estudo foram retiradas amostras aleatórias de acordo com quatro delineamentos de amostragem probabilísticos, assumindo sempre como parâmetro de interesse a média da variável VVN das empresas.

Nos quatro delineamentos de amostragem a seguir apresentados para cada uma das populações, procurou-se obter sempre amostras que correspondem a 10% da população em estudo, com vista a posteriormente fazer uma comparação dos resultados obtidos com os diferentes estimadores da variância dos estimadores nos esquemas adotados.

3.2.2.1 Delineamento de amostragem I

Para a seleção da amostra considerou-se um esquema de amostragem estratificado por Região e CAE, em que as unidades de amostragem (empresas) foram selecionadas sem reposição aleatoriamente dentro de cada estrato.

A distribuição da amostra de 1000 empresas pelos estratos foi feita por afetação ótima de *Neyman* de acordo com a expressão (2.13), uma vez que:

- Os estratos têm dimensões diferentes (número de empresas por estrato);
- A variância da variável de interesse varia entre os estratos

3.2.2.2 Delineamento de amostragem II

A seleção da amostra seguiu um esquema de amostragem estratificado por Região, em que as unidades primárias de amostragem (empresas) foram selecionadas sem reposição aleatoriamente dentro de cada estrato.

A distribuição da amostra foi feita de modo proporcional em cada estrato, adequada uma vez que:

- Os estratos têm dimensões diferentes (número de empresas por estrato);
- A variância da variável de interesse (VVN) é similar entre estratos.

3.2.2.3 Delineamento de amostragem III

A seleção da amostra seguiu um esquema de amostragem por grupos em duas etapas. Assumiu-se como unidade primária de amostragem cada uma das combinações entre a Região e a CAE (RegCAE), e as empresas a unidade secundária de amostragem. Foram selecionados sem reposição aleatoriamente seis dos nove grupos de RegCAE existentes na população de empresas. O número de unidades secundárias (empresas) a selecionar sem reposição aleatoriamente dentro das unidades primárias selecionadas foi determinado com recurso à alocação ótima de *Neyman*.

Em suma, este delineamento consistiu em:

1ª etapa: escolher aleatoriamente $m = 6$ das $M = 9$ RegCAE.

2ª etapa: escolher aleatoriamente n_{ij} empresas de entre N_{ij} empresas existentes nas RegCAE selecionadas na etapa anterior.

Uma vez que a seleção das empresas foi efetuada em duas etapas, o cálculo das respetivas probabilidades de inclusão teve em conta as seguintes componentes:

- A probabilidade de inclusão das UPA;

- A probabilidade (condicionada) de inclusão das empresas de cada UPA;

3.2.2.4 Delineamento de amostragem IV

A seleção da amostra seguiu um esquema de amostragem estratificado e multietápico. A dimensão das amostras a retirar em cada estrato é proporcional ao número de empresas nesse estrato (variável Região). Dentro de cada estrato, foram selecionadas duas UPA (variável CAE) com probabilidade proporcional ao tamanho, totalizando seis das nove combinações entre estrato e UPA (variável RegCAE) existentes na população de empresas. O número de USA (empresas) a selecionar, aleatoriamente, dentro das UPA de cada estrato obtido com recurso à alocação proporcional aos grupos da UPA.

Em suma, este delineamento consistiu em:

1ª etapa: estratificar a população por Região.

2ª etapa: escolher por alocação proporcional $m = 2$ das $M = 3$ CAE por região, totalizando $m = 6$ das $M = 9$ RegCAE.

3ª etapa: escolher aleatoriamente n_{ij} empresas de entre N_{ij} empresas existentes nas RegCAE selecionadas na etapa anterior.

Uma vez que a seleção das empresas foi efetuada em três etapas, o cálculo das respetivas probabilidades de inclusão teve em conta as seguintes componentes:

- A probabilidade de inclusão das UPA;
- A probabilidade (condicionada) de inclusão das empresas nas UPA.

3.3 Metodologia de avaliação dos estimadores da variância

No presente trabalho, com o objetivo de mostrar o desempenho das técnicas de estimação da variância foram utilizados: o método de *Linearização de Taylor* e as técnicas de replicação *Jackknife* e *Bootstrap*, em amostras que reflitam a estrutura das empresas no sector do comércio no território Moçambicano.

Para avaliar a precisão dos estimadores de variância em função de cada um dos esquemas (delineamento de amostragem), foram realizadas *NR* = 10 000 réplicas de cada um dos delineamentos.

Para cada uma dessas réplicas, $nr = 1, \dots, NR$ foram calculadas:

- A estimativa da media do VVN, $\hat{\mu}_{nr}$;
- As estimativas da variância da média do VVN pelos métodos:
  a) *Linearização de Taylor* $\widehat{Var}_T(\hat{\mu}_{nr})$ assumindo a expressão (2.34);
  b) *Jackknife* $\widehat{Var}_{JK}(\hat{\mu}_{nr})$ de acordo com a expressão (2.35);
  c) *Bootstrap* $\widehat{Var}_B(\hat{\mu}_{nr})$ usando a expressão (2.39).

A partir da distribuição de amostragem obtida com as *NR* réplicas, estimou-se:

- A média

$$\hat{E}(\hat{\mu}) = \frac{1}{NR}\sum_{nr=1}^{NR}\hat{\mu}_{nr},$$

- A variância da estimativa da média

$$\widehat{Var}(\hat{\mu}) = \frac{1}{NR-1}\sum_{nr=1}^{NR}[\hat{\mu}_{nr} - \hat{E}(\hat{\mu})]^2.$$

- O enviesamento do estimador foi calculado por meio da expressão (2.40);
- O erro quadrático médio, dado pela expressão (2.44).
- A raiz do erro quadrado médio escalado, dado pela expressão (2.45).

Para a geração da população de estudo e a estimação da média e respetiva variância foi utilizado o *software* RStudio, versão R-3.1.1.

Capítulo 4

Apresentação e discussão de resultados

A informação do perfil demográfico das populações em estudo (empresas do sector do comércio), a avaliação dos estimadores e a análise comparativa dos mesmos foram feitas no presente capítulo, como forma de responder aos objetivos do estudo.

4.1 Populações em estudo

Para a realização deste trabalho geraram-se três populações de empresas com características diferenciadas:

1. Para a primeira população assumimos uma imagem da população real de empresas do sector do comércio no território Moçambicano, no que se refere a distribuição das variáveis em estudo.
2. Na segunda população procurámos manter a distribuição do número de trabalhadores assim como a distribuição do volume de negócios por empresa da primeira população, mas garantiu-se um equilíbrio no número de empresas por Região e CAE.
3. Para a terceira população procurámos garantir uma variação na distribuição do NPS entre as empresas, mas mantendo a distribuição das restantes variáveis definidas para a primeira população.

4.1.1 População I

As empresas fictícias foram distribuídas por três regiões: Norte, Centro e Sul, de acordo com a distribuição apresentada na Ilustração 4.1. À semelhança da realidade Moçambicana, localizaram-se mais empresas na região Sul do que Centro e Norte.

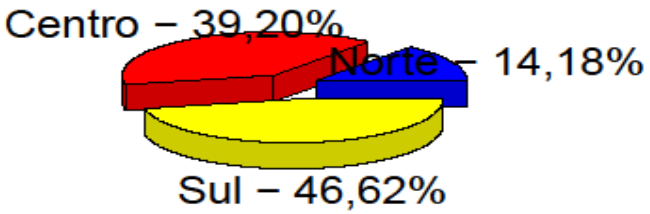


Ilustração 4.1 Distribuição das empresas por Região (população I).

Cada uma das empresas fictícias foi afeta aleatoriamente a um código CAE, de acordo com a distribuição apresentada Ilustração 4.2. A sua distribuição em Moçambique também é assimétrica, verificando-se um maior número de empresas na "CAE 47" nas três regiões do País.

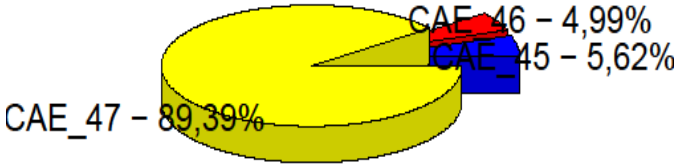


Ilustração 4.2 Distribuição das empresas por CAE (população I).

Como foi descrito no capítulo III, para a distribuição da variável NPS no país assumiu-se que temos uma maior concentração de empresas pequenas, seguido das médias e por último as empresas grandes. Assim, para 70% das empresas considerou-se que *NPS* ~ BN(0,4; 4), para 5% das empresas considerou-se *NPS* ~ BN(0,35; 40) e para as restantes *NPS* ~ BN(0,3; 90). Na Ilustração 4.3 temos a distribuição do NPS pelo número de empresas.

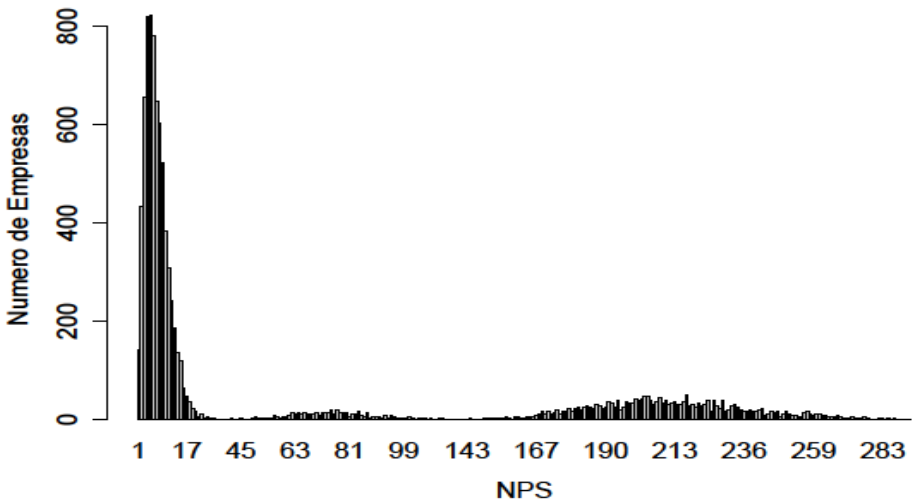


Ilustração 4.3 Distribuição do NPS por Empresa (População I).

Posteriormente foi criada a variável EPS que corresponde à variável NPS categorizada conforme foi descrito no capítulo III. Na Ilustração 4.4, podemos verificar que se obteve o padrão pretendido, isto é, maior número de empresas com menos de 50 trabalhadores (pequena), e menor número de empresas grandes.

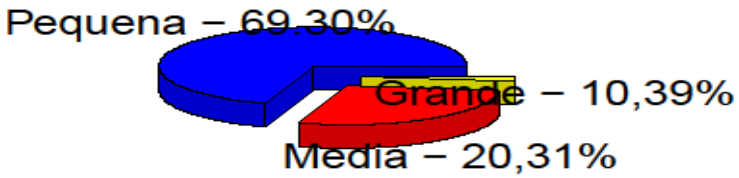


Ilustração 4.4 Distribuição das empresas por EPS (população I).

Tabela 4.1 Estatísticas descritivas da variável NPS (População I).

Medidas Resumo	NPS
N.º de observações	10000
Mínimo	1
Maximo	312
1º quartil	5,0
3º quartil	101,0
Média	61,1
Mediana	9,0
Variância	7848,8
Desvio-padrão	88,6
Assimetria	1,2
Achatamento	-0,4

Na população I verificou-se que o NPS varia de 1 a 312 trabalhadores, com uma média tendenciosa de aproximadamente 61 trabalhadores por empresa um desvio em relação à média de quase 89 trabalhadores o que indica que a população possui uma elevada dispersão na distribuição do NPS por empresa (Tabela 4.1). Apresenta uma assimetria à direita, que indica a predominância de pequenas empresas, isto é, com poucos trabalhadores.

Para gerar os valores do volume de negócios das empresas (VVN), considerou-se o modelo (3.2) apresentado no capítulo III, no qual assumimos que a distribuição da variável VVN está relacionada com o NPS nas empresas, isto é, empresas com poucos trabalhadores tem valores de VVN mais baixos do que as empresas com muitos trabalhadores. Considerou-se $\alpha = 6800, \beta = 3200$ e $\varepsilon \sim N(0; 800 \times NPS)$. O critério utilizado para a escolha destes valores foi que a maior parte dos valores de VVN estivessem compreendidos entre 10 000 e 1 000 000 de meticais (MT) (Ilustração 4.5).

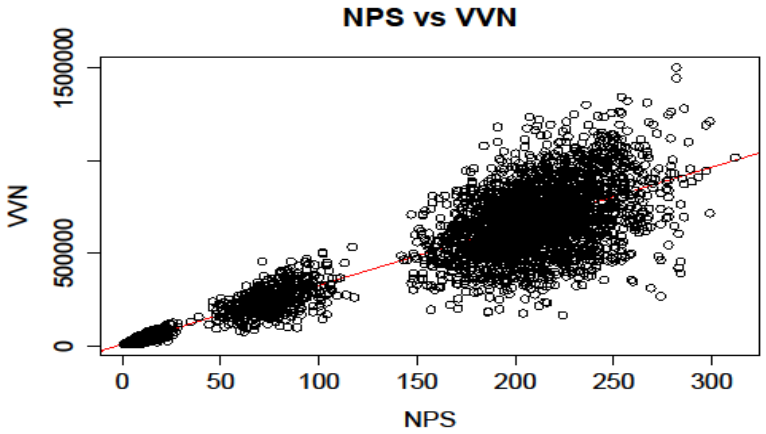


Ilustração 4.5 Relação entre NPS e VVN (População I).

Na Tabela 4.2 apresentam-se algumas medidas resumo da variável *VVN* para a população das 10 000 empresas fictícias geradas, onde podemos verificar que a distribuição das empresas por região e EPS é desequilibrada (isto é, não uniforme), com uma maior concentração das empresas na EPS pequena nas três regiões. O valor da média assim como o desvio padrão da variável VVN variam muito por EPS e região na população.

Tabela 4.2 Número de empresas em cada Região e EPS, média, desvio-padrão da variável VVN (população I).

Região	***EPS***	***Dimensão*** (N_i)	***Média*** ($\bar{Y}_i$)	***Desvio-padrão*** (σ_i)
Norte	Pequena	1002	29914,7	15370,5
	Média	390	596229,5	228705,0
	Grande	26	846084,2	241062,8
	Subtotal	1418	200636,3	294749,1
Centro	Pequena	2761	29487,2	14687,5
	Média	1076	591780,4	236200,1
	Grande	83	821224,2	223539,8
	Subtotal	3920	200594,8	295466,3
Sul	Pequena	3263	29284,9	15991,5
	Média	1305	589535,4	223955,8
	Grande	94	875162,6	215391,9
	Subtotal	4662	203167,3	295307,1
Total		**10000**	**201800,0**	**295263,7**

A distribuição das empresas por região e CAE também é desequilibrada (Tabela 4.3), com uma maior concentração das empresas na CAE 47 nas três regiões. O valor da média assim como o desvio padrão da variável VVN pouco variam por CAE e região na população em estudo.

Tabela 4.3 Número de empresas em cada estrato, desvio-padrão da variável VVN e dimensão da amostra (população I).

Região	***CAE***	***Dimensão (N_i)***	***Média ($\bar{Y}_i$)***	***Desvio-padrão (σ_i)***
Norte	45	77	193908,0	311956,8
	46	69	182830,0	266453,1
	47	1272	202009,5	295325,6
	Subtotal	1418	200636,3	294749,1
Centro	45	236	182132,7	277123,7
	46	188	224092,6	307947,5
	47	3496	200577,5	295974,3
	Subtotal	3920	200594,8	295466,3
Sul	45	249	209346,7	310282,8
	46	242	215275,8	312165,1
	47	4171	202095,9	293436,5
	Subtotal	4662	203167,3	295307,1
Total		**10000**	**201800,0**	**295263,7**

Na Tabela 4.4 apresenta-se o tamanho das amostras a retirar por região e/ou CAE com cada um dos quatro delineamentos definidos no capítulo III.

No delineamento I podemos verificar, pela dimensão das amostras por estrato, que em cada uma das combinações entre a Região e a CAE temos amostras com dimensões menores em todas as regiões nas CAEs 45 e 46 e dimensões maiores na CAE 47, devido ao tamanho dos estratos na população e à variabilidade da variável de interesse nos respectivos estratos, este facto é confirmado pela dimensão da população nos respetivos estratos. Para este delineamento, a margem de erro relativa é de 8,56% para a média do VVN para um nível de confiança de 5%.

No delineamento II a dimensão da amostra por região é proporcional a dimensão da população por região. Com este delineamento, para uma amostra de 1000 empresas, a

margem de erro relativa é de 8,71% para a média do VVN para um nível de significância de 5%.

No delineamento III a dimensão da amostra em cada grupo é aleatória pois depende dos grupos selecionados em cada simulação, garantindo-se apenas que a dimensão da amostra final é de 1000 empresas. Para este delineamento a margem de erro relativa é de 8,24% para a média do VVN para um nível de significância de 5%.

No delineamento IV, a dimensão da amostra final definida por estrato na primeira etapa é proporcional a dimensão da população por região, a dimensão da amostra em cada grupo é proporcional aos grupos selecionados em cada simulação. Para este delineamento a margem de erro relativa é de 9,36% para a média do VVN para um nível de significância de 5%.

Tabela 4.4 Número de empresas a amostrar por região e/ou CAE, com cada delineamento (população I).

Região	***CAE***	***Dimensão da amostra (n_i)***				
		Delineamento I	***Delineamento II***	***Delineamento III***	***Delineamento IV***	
Norte	45	8	142	*	142	*
	46	6		*		*
	47	127		*		*
Centro	45	22	392	*	392	*
	46	20		*		*
	47	350		*		*
Sul	45	26	466	*	466	*
	46	26		*		*
	47	414		*		*
Total		**1000**	**1000**	**1000**	**1000**	

* A dimensão da amostra varia entre simulações.

4.1.2 População II

As empresas fictícias foram distribuídas pelas três regiões, Norte, Centro e Sul, de acordo com a distribuição apresentada na Ilustração 4.6, garantindo um equilíbrio na distribuição das empresas por região.

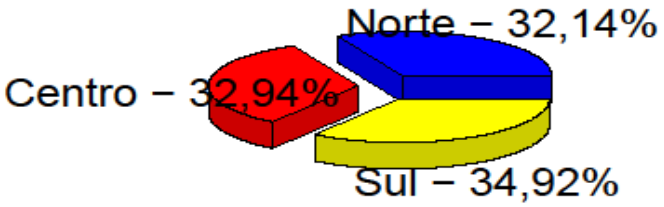


Ilustração 4.6 Distribuição das empresas por Região (população II).

A cada uma das empresas fictícias foi afeta aleatoriamente a um código CAE, de acordo com a distribuição apresentada na Ilustração 4.7. A distribuição equilibrada das empresas visa avaliar os estimadores em populações com características distintas.

Ilustração 4.7 Distribuição das empresas por CAE (população II).

Para a população II a distribuição das variáveis NPS e VVN manteve-se o valor dos parâmetros de distribuição usados na população I.

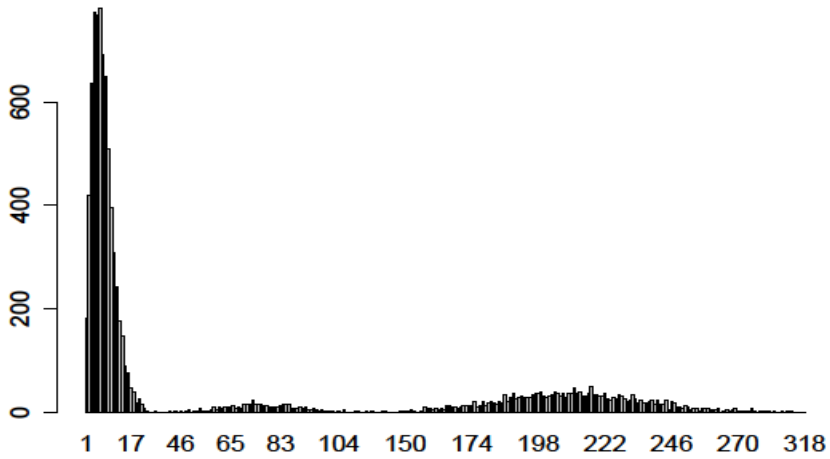


Ilustração 4.8 Distribuição do NPS por Empresa (População II).

Na Ilustração 4.9, podemos verificar que se obteve o padrão pretendido, mantendo o maior número de empresas com menos de 50 trabalhadores (pequena), e menor número de empresas grandes.

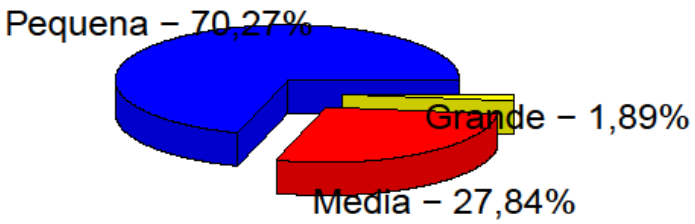


Ilustração 4.9 Distribuição das empresas por EPS (população II).

Tabela 4.5 Estatísticas descritivas da variável NPS (População II).

Medidas Resumo	NPS
N.º de observações	10000
Mínimo	1
Maximo	315
1º quartil	5,0
3º quartil	154,0
Média	62,4
Mediana	9,0
Variância	7936,1
Desvio-padrão	89,1
Assimetria	1,1
Achatamento	-0,5

Na população II temos que o NPS varia de 1 a 315 trabalhadores, com uma média tendenciosa de aproximadamente 62 trabalhadores por empresa um desvio em relação à média de quase 89 trabalhadores o que indica que a população possui uma elevada dispersão na distribuição do NPS por empresa (Tabela 4.5). Apresenta uma assimetria à direita, que indica a predominância de pequenas empresas, isto é, com poucos trabalhadores.

Para gerar os valores do VVN, foi usado o modelo (3.2) com os parâmetros de distribuição usados na população I (Ilustração 4.10).

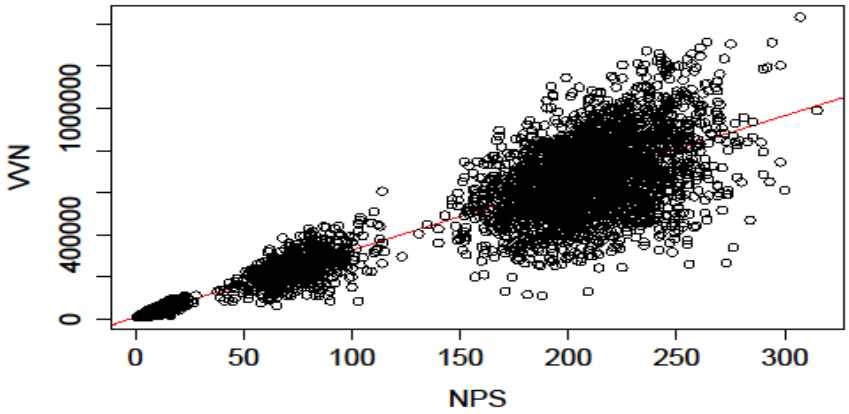


Ilustração 4.10 Relação entre NPS e VVN (População II).

Na Tabela 4.6 apresentam-se algumas medidas resumo da variável *VVN* para a população das 10 000 empresas fictícias geradas, onde podemos verificar que a distribuição das empresas por região e EPS é desequilibrada, com uma maior concentração das empresas na EPS pequena nas três regiões. O valor da média assim como o desvio padrão da variável VVN variam muito por EPS e região na população.

Tabela 4.6 Número de empresas em cada Região e EPS, média, desvio-padrão da variável VVN (população II).

Região	***EPS***	***Dimensão (N_i)***	***Média ($\bar{Y}_i$)***	***Desvio-padrão (σ_i)***
Norte	Pequena	2231	29411,8	15124,4
	Média	920	594784,6	229694,4
	Grande	63	838003,3	224282,8
	Subtotal	1418	207098,3	298302,2
Centro	Pequena	2305	29525,3	15133,8
	Média	920	593291,0	234449,1
	Grande	69	804941,4	246830,8
	Subtotal	3920	203225,4	296600,8
Sul	Pequena	2399	29140,7	14855,9
	Média	1024	582451,5	232135,3
	Grande	69	841302,0	194911,2
	Subtotal	4662	207442,4	296178,4
Total		**10000**	**205942,7**	**296978,0**

A distribuição das empresas por região e CAE é equilibrada (Tabela 4.7). O valor da média assim como o desvio padrão da variável VVN pouco variam por CAE e região na população em estudo.

Tabela 4.7 Número de empresas em cada estrato, desvio-padrão da variável VVN e dimensão da amostra (população II).

Região	***CAE***	***Dimensão (N_i)***	***Média ($\bar{Y}_i$)***	***Desvio-padrão (σ_i)***
Norte	45	1014	192815,7	281667,1
	46	1086	209021,1	306173,8
	47	1114	218224,3	304915,3
	Subtotal	3214	207098,3	298302,2
Centro	45	1047	205983,4	300807,9
	46	1119	208365,5	296299,4
	47	1128	195566,3	293054,0
	Subtotal	3294	203225,4	296600,8
Sul	45	1117	208115,1	292387,1
	46	1162	192414,8	283067,4
	47	1213	221218,6	311109,5
	Subtotal	3492	207442,4	296178,4
Total		**10000**	**205942,7**	**296978,0**

Na Tabela 4.8 apresenta-se o tamanho das amostras a retirar por região e/ou CAE com cada um dos quatro delineamentos definidos no capítulo III.

No delineamento I podemos verificar pela dimensão das amostras por estrato, que em cada uma das combinações entre a Região e a CAE temos amostras com dimensões muito próximas, devido as características homogéneas existentes entre as empresas nestes estratos populacionais. A dimensão das amostras por estrato são muito similares contrariamente ao que acontece na população I. Para este delineamento a margem de erro relativa é de 8,59% para a média do VVN para um nível de significância de 5%.

No delineamento II a amostra por região possui dimensões proporcionais à da população por região, contrariamente ao que acontece na população I. De acordo com este delineamento, para uma amostra de aproximadamente 1000 empresas, a margem de erro relativa é de 8,53% para a média do VVN para um nível de significância de 5%.

Para o delineamento III, como vimos na população I, a dimensão da amostra em cada grupo é aleatória pois depende dos grupos selecionados em cada simulação, com a amostra final de aproximadamente 1000 empresas. A margem de erro relativa é de 8,52% para a média do VVN para um nível de significância de 5%.

No delineamento IV, a dimensão da amostra final definida por estrato na primeira etapa é proporcional a dimensão da população por região, que diferem muito comparativamente a população I. A amostra a retirar na zona Norte é superior a indicada na população I, acontecendo o contrário nas outras duas regiões. Para este delineamento a margem de erro relativa é de 8,67% para a média do VVN para um nível de significância de 5%.

Tabela 4.8 Número de empresas a amostrar por região e/ou CAE, com cada delineamento (população II).

Região	***CAE***	***Dimensão da amostra (n_i)***				
		Delineamento I	***Delineamento II***	***Delineamento III***	***Delineamento IV***	
Norte	45	96	321	*	321	*
	46	112		*		*
	47	115		*		*
Centro	45	106	329	*	329	*
	46	112		*		*
	47	111		*		*
Sul	45	110	349	*	349	*
	46	111		*		*
	47	127		*		*
Total		**1000**	**1000**	**1000**	**1000**	

* A dimensão da amostra varia entre simulações

4.1.3 População III

Na população III, as empresas fictícias foram distribuídas pelas três regiões, Norte, Centro e Sul, de acordo com a distribuição apresentada na Ilustração 4.1. Para cada uma das empresas fictícias foi afeta aleatoriamente a um código CAE, de acordo com a distribuição apresentada na Ilustração 4.2. A distribuição das variáveis acima descritas segue o esquema usado na população I.

Nesta população III procurou-se diminuir a variabilidade da variável NPS, relativamente à população I, reduzindo as probabilidades de sucesso entre as subamostras da população de 10 000 empresas fictícias. Ao contrário da população I, assumiu-se uma maior concentração de médias empresas, seguido das pequenas e por último as grandes. Assim, para 40% das empresas considerou-se que $NPS \sim BN(0{,}4; 164)$, para 30% das empresas considerou-se que $NPS \sim BN(0{,}35; 30)$ e para as restantes $NPS \sim BN(0{,}41; 4)$. Na Ilustração 4.11 temos a distribuição do NPS pelo número de empresas.

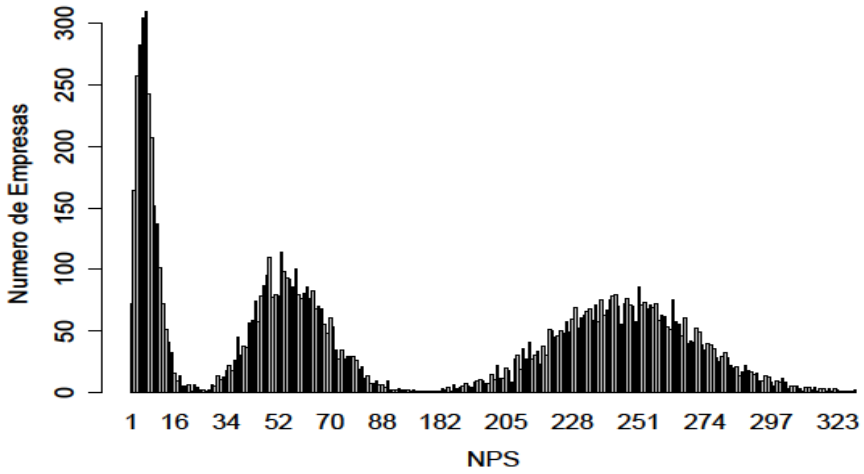


Ilustração 4.11 Distribuição do NPS por Empresa (População III).

Posteriormente foi criada a variável EPS que corresponde à variável NPS categorizada conforme foi descrito no capítulo III. Na Ilustração 4.12, podemos verificar que se obteve o padrão pretendido, isto é, um menor desequilíbrio na distribuição de empresas por EPS, relativamente a população I, com número de empresas médias próximo do número de empresas pequenas, mantendo-se um menor número de empresas grandes. A distribuição menos desequilibrada do NPS visa avaliar os estimadores em populações com características distintas.

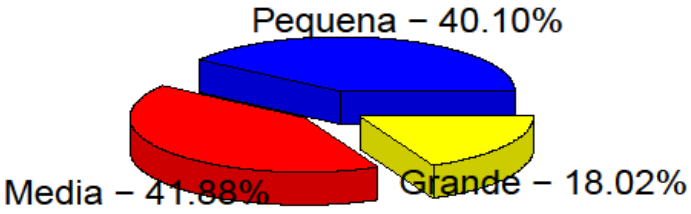


Ilustração 4.12 Distribuição das empresas por EPS (população III)

Tabela 4.9 Estatísticas descritivas da variável NPS (População III).

Medidas Resumo	NPS
N.º de observações	10000
Mínimo	1
Maximo	336
1º quartil	29,0
3º quartil	243,0
Média	130,0
Mediana	69,0
Variância	11876,0
Desvio-padrão	108,9
Assimetria	0,2
Achatamento	-1,8

Para a população III verificou-se que o NPS varia de 1 a 336 trabalhadores, com uma média de 130 trabalhadores por empresa, um desvio em relação a média de aproximadamente 109 trabalhadores o que indica que a população possui uma elevada dispersão na distribuição do NPS por empresa (Tabela 4.9). Apresenta uma simetria entre as bossas de distribuição dos trabalhadores, que indica que, há um equilíbrio na distribuição dos trabalhadores por EPS.

Para gerar os valores do VVN, foi usado o modelo (3.2) com os parâmetros de distribuição usados na população I (Ilustração 4.13).

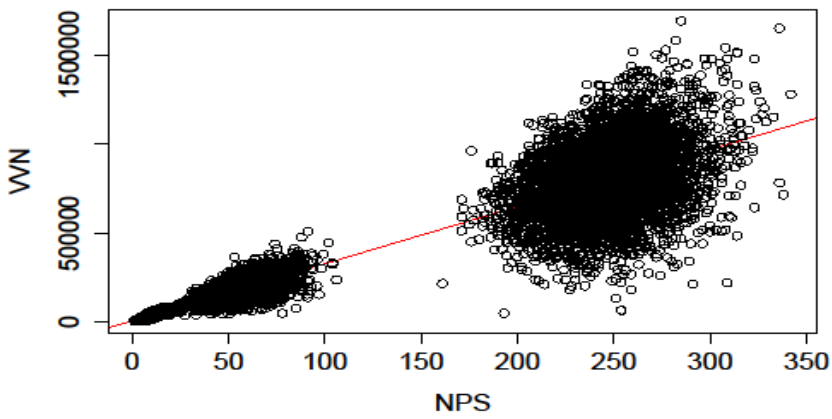


Ilustração 4.13 Relação entre NPS e VVN (População III).

Na Tabela 4.10 apresentam-se algumas medidas resumo da variável VVN para a população das 10 000 empresas fictícias geradas, onde podemos verificar que a distribuição das empresas por região e EPS é equilibrada nas EPS pequena e média em todas as regiões. O valor da média assim como o desvio padrão da variável VVN variam muito por EPS e região na população.

Tabela 4.10 Número de empresas em cada Região e EPS, média, desvio-padrão da variável VVN (população III).

Região	***EPS***	***Dimensão (N_i)***	***Média ($\bar{Y}_i$)***	***Desvio-padrão (σ_i)***
Norte	Pequena	583	53201,9	52306,7
	Média	590	489340,3	306513,9
	Grande	245	877953,0	218697,6
	Subtotal	1418	377169,3	374424,0
Centro	Pequena	1525	53350,7	52361,3
	Média	1665	479268,5	299125,6
	Grande	730	877042.7	217060,2
	Subtotal	3920	387648,8	373481,9
Sul	Pequena	1902	52003,6	51222,5
	Média	1933	480839,1	301109,9
	Grande	827	871168,0	219304,6
	Subtotal	4662	375124,2	371581,8
Total		**10000**	**380323,8**	**372740,9**

Conforme se pode observar na Tabela 4.11, o número de empresas por região e CAE manteve-se relativamente à população I. O valor da média assim como o desvio padrão da variável VVN variam por CAE e região na população em estudo.

Tabela 4.11 Número de empresas em cada estrato, desvio-padrão da variável VVN e dimensão da amostra (população III).

Região	***CAE***	***Dimensão (N_i)***	***Média ($\bar{Y}_i$)***	***Desvio-padrão (σ_i)***
Norte	45	77	381465,2	376040,3
	46	69	353838,6	372995,8
	47	1272	378174,8	374655,6
	Subtotal	1418	377169,3	374424,0
Centro	45	236	377563,8	379452,0
	46	188	341306,3	354508,7
	47	3496	390821,6	373995,8
	Subtotal	3920	387648,8	373481,9
Sul	45	249	398026,5	398669,6
	46	242	364405,1	370646,5
	47	4171	374378,9	370000,0
	Subtotal	4662	375124,2	371581,8
Total		**10000**	**380323,8**	**372740,9**

Na Tabela 4.12 apresenta-se o tamanho das amostras a retirar por região e/ou CAE com cada um dos quatro delineamentos definidos no capítulo III.

Nos delineamentos I e II a dimensão da amostra a retirar em cada estrado é igual ao obtido para a população I. No delineamento I, a margem de erro relativa é de 5,86% para a média do VVN para um nível de confiança de 5% mostra que existem poucas diferenças na dimensão das amostras em relação a população I. No delineamento II a margem de erro relativa é de 5,79% para a média do VVN para um nível de significância de 5%, temos amostras de dimensão igual às da população I.

No delineamento III, a dimensão da amostra em cada grupo é aleatória pois depende dos grupos selecionados em cada simulação, garantindo-se apenas que a dimensão da amostra final é de 1000 empresas. Para este delineamento a margem de erro relativa de 5,96% para a média do VVN para um nível de significância de 5%.

No delineamento IV, podemos verificar que a amostra por região tem as mesmas dimensões da população I, pois foram usados os mesmos parâmetros para gerar a distribuição das empresas por região e na amostragem também foi usada a alocação

proporcional ao tamanho. Para este delineamento a margem de erro relativa é de 6,47% para a média do VVN para um nível de significância de 5%.

Tabela 4.12 Número de empresas a amostrar por região e/ou CAE, em cada delineamento (população III).

Região	***CAE***	***Dimensão da amostra (n_i)***				
		Delineamento I	***Delineamento II***	***Delineamento III***	***Delineamento IV***	
Norte	45	8	142	*	142	*
	46	7		*		*
	47	128		*		*
Centro	45	24	392	*	392	*
	46	18		*		*
	47	351		*		*
Sul	45	27	466	*	466	*
	46	24		*		*
	47	414		*		*
Total		**1000**	**1000**	**1000**	**1000**	

* A dimensão da amostra varia entre simulações.

4.2 Distribuições amostrais da média

Para cada um dos quatro delineamentos definidos, foram retiradas 10 000 amostras que permitiram obter estimativas da média do VVN e construir as distribuições de amostragem empíricas do estimador da média e dos estimadores da variância da média estimada do VVN. A variância das médias estimativas corresponderá a uma aproximação da variância real do estimador da variância do VVN, para cada um dos delineamentos (Bean, 1975 and Kish, 1974).

Pela análise da Ilustração 4.14, podemos observar que nos quatro delineamentos o estimador da média para a variável de interesse é não-enviesado. Apesar das diferenças existentes na dispersão das estimativas entre os delineamentos, temos que, para cada população temos um delineamento com a menor dispersão, o que é favorável para a precisão dos resultados das estimativas. Temos a presença de *outliers* superiores e inferiores para as três populações analisadas. Os diagramas de caixa e bigodes são simétricos indicando que as distribuições amostrais da média para as estimativas são simétricas. Apesar das diferenças existentes nos parâmetros usados para a geração das três populações em estudo existe uma simetria semelhante na distribuição das estimativas da média do VVN nas três populações (Tabela A.1, A.2 e A.3).

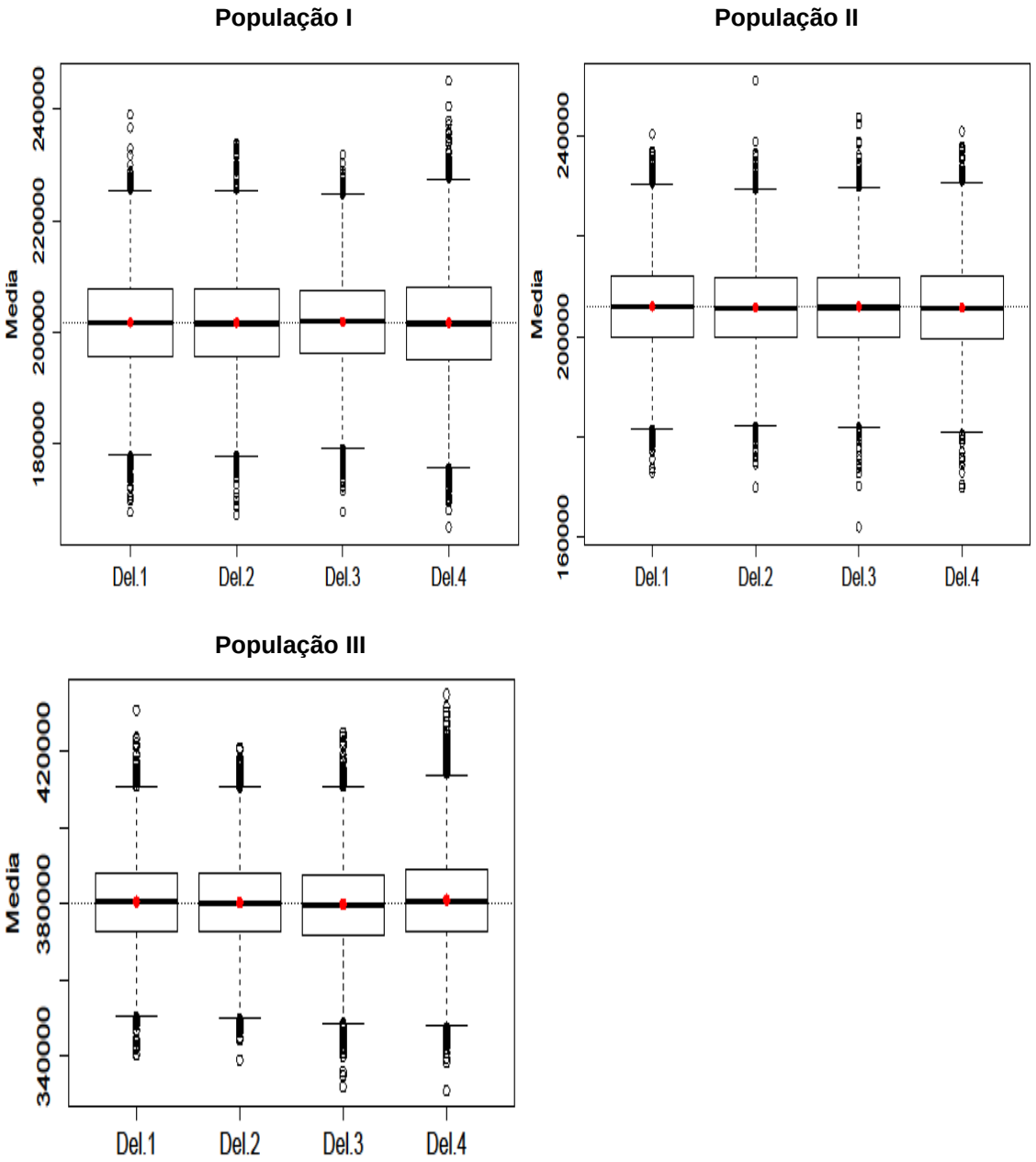


Ilustração 4.14 Diagrama de caixas e bigodes das distribuições amostrais da média para as três populações sob os delineamentos I a IV. Os pontos a vermelho representam a média das estimativas do VVN, por delineamento. A linha horizontal pontilhada representa a média do VVN, por população.

4.3 Distribuição dos estimadores da variância da média estimada

As distribuições de amostragem obtidas para os estimadores *Taylor*, *Jackknife*, *Bootstrap* para a variância de $\hat{\mu}$ para as três populações são apresentados nas figuras que se seguem (Ilustração 4.15 a Ilustração 4.18). Além disso, também é apresentada a distribuição de amostragem do estimador usualmente indicado na literatura para os delineamentos de amostragem definidos, e que foram apresentados no capítulo III.

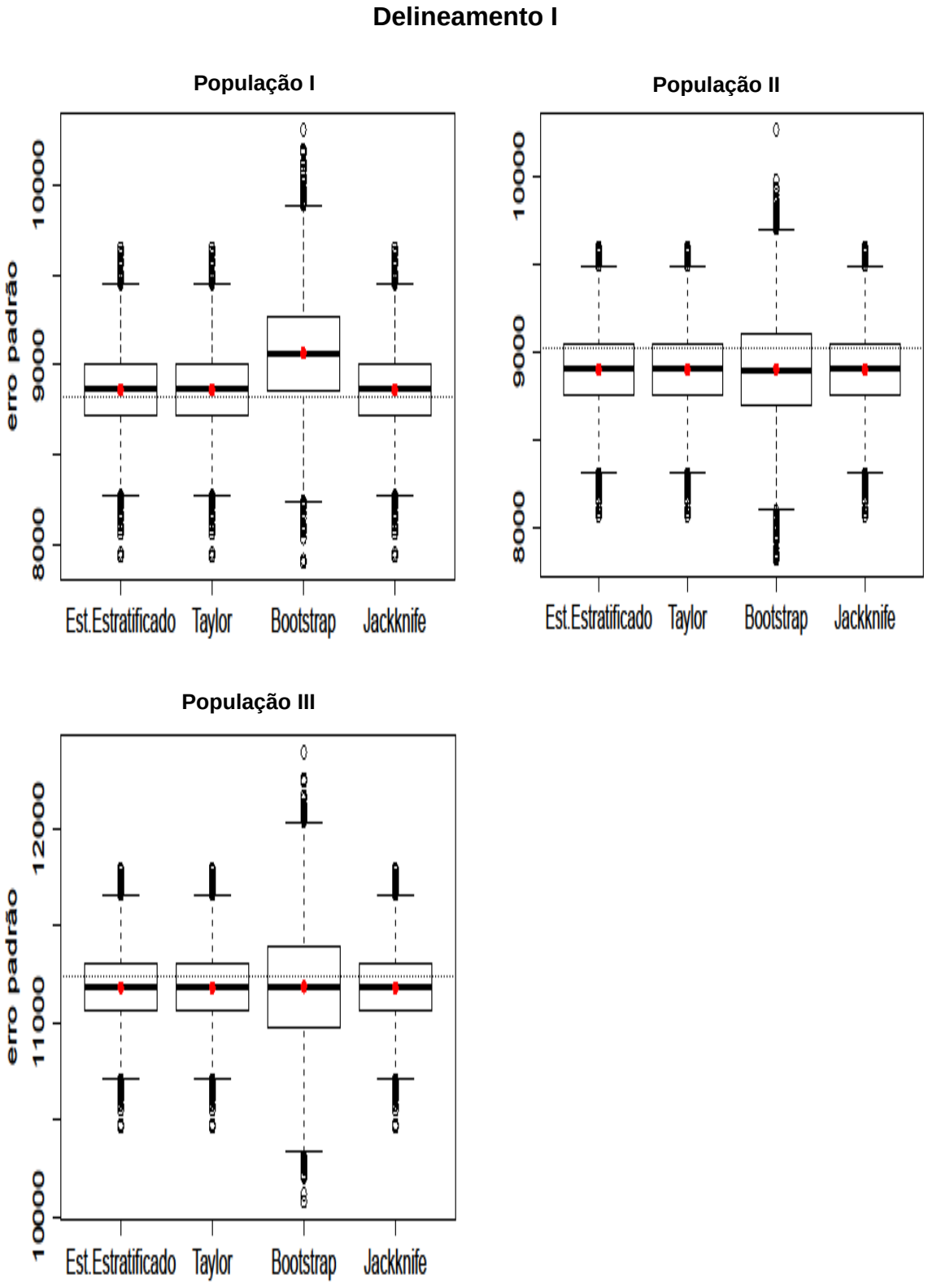


Ilustração 4.15 Diagrama de caixas e bigodes das distribuições amostrais dos estimadores da amostragem estratificada (*Etapa1*), de *Taylor*, *Bootstrap e Jackknife* para a variância do estimador média, sob delineamento estratificado por Região e CAE. Os pontos a vermelho representam a média das estimativas do desvio padrão da média estimada do VVN por estimador. A linha horizontal pontilhada representa o desvio padrão da estimativa média do VVN, por população.

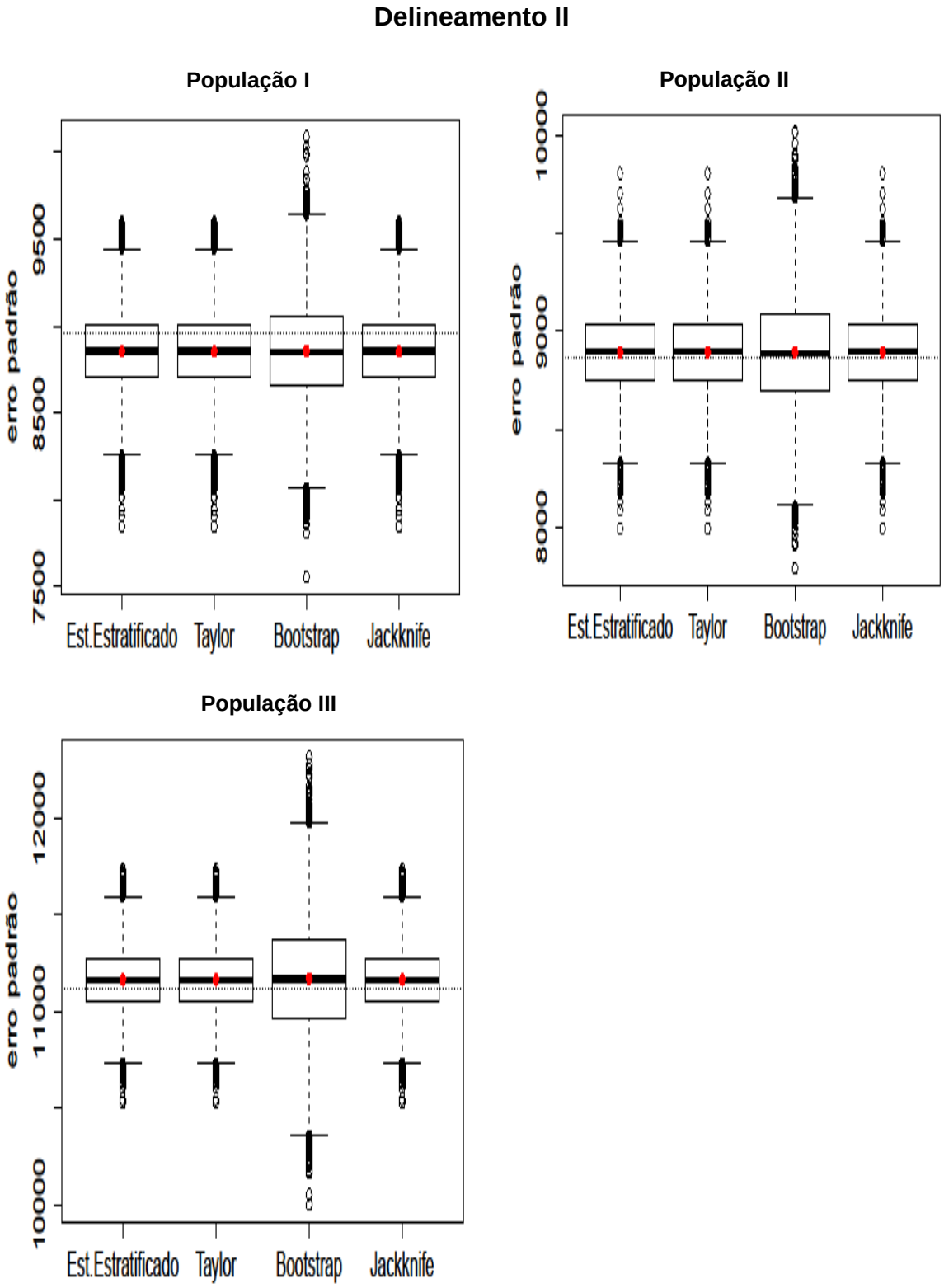


Ilustração 4.16 Diagrama de caixas e bigodes das distribuições amostrais dos estimadores da amostragem estratificada (*Etapa1*), de *Taylor, Bootstrap e Jackknife* para a variância do estimador média, sob delineamento estratificado proporcional ao tamanho da Região. Os pontos a vermelho representam a média das estimativas do desvio padrão da média estimada do VVN por estimador. A linha horizontal pontilhada representa o desvio padrão da estimativa média do VVN, por população.

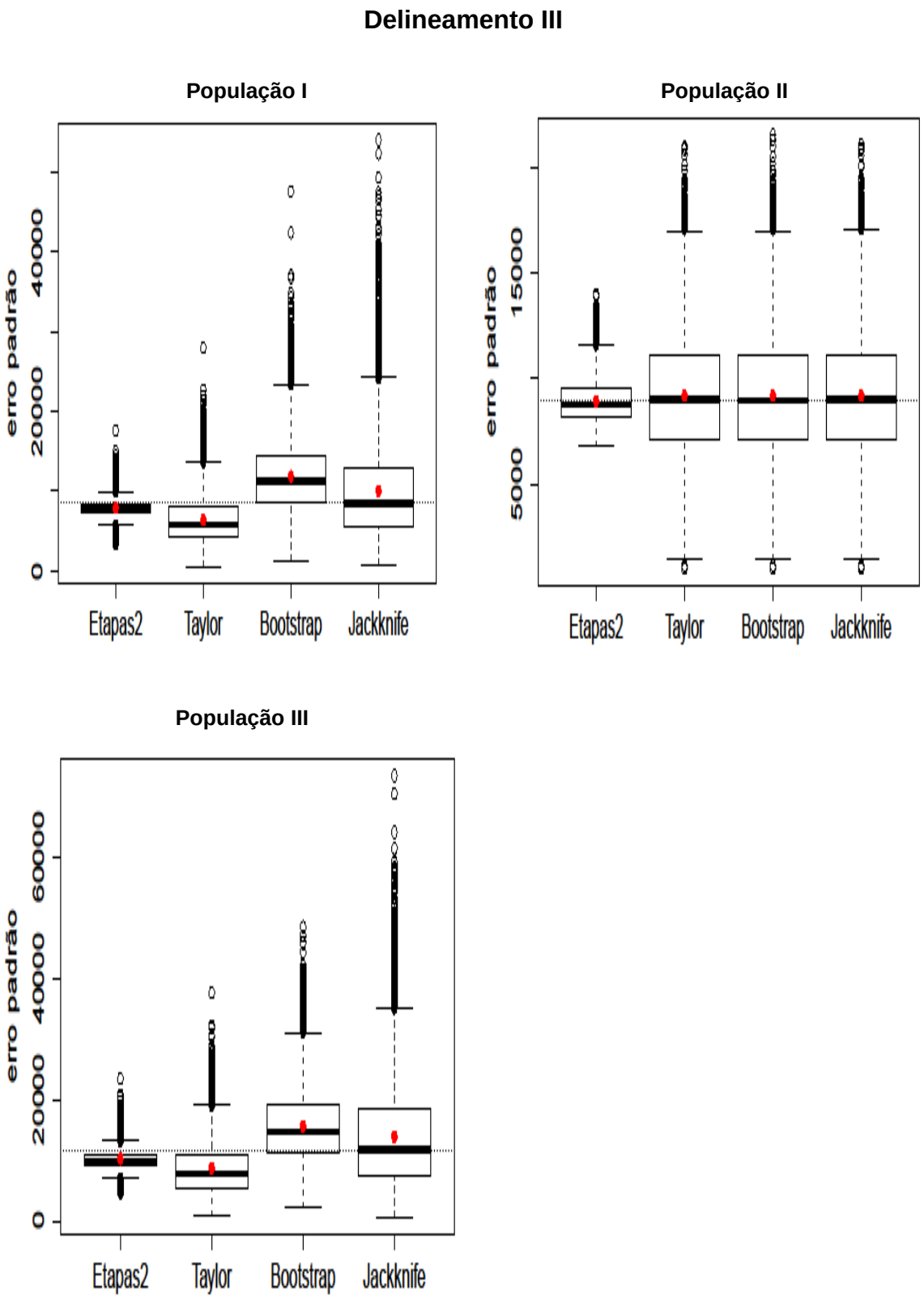


Ilustração 4.17 Diagrama de caixas e bigodes das distribuições amostrais dos estimadores, indicado na literatura *(Etapas2)*, de *Taylor*, *Bootstrap e Jackknife* para a variância do estimador média, sob delineamento em grupos em duas etapas (RegCAE, Id). Os pontos a vermelho representam a média das estimativas do desvio padrão da média estimada do VVN por estimador. A linha horizontal pontilhada representa o desvio padrão da estimativa média do VVN, por população.

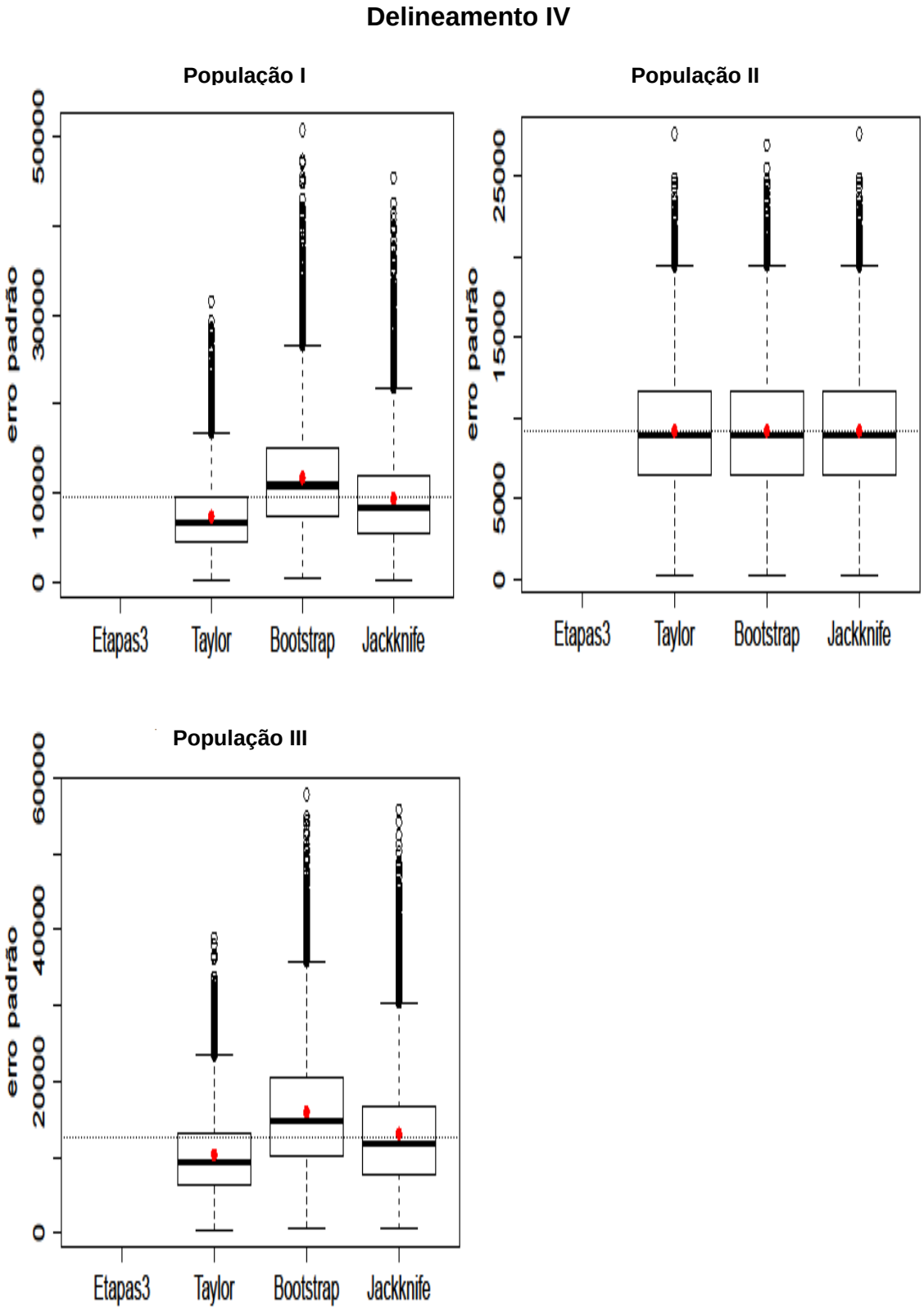


Ilustração 4.18 Diagrama de caixas e bigodes das distribuições amostrais dos estimadores, indicado na literatura *(Etapas3)*, de *Taylor*, *Bootstrap e Jackknife* para a variância do estimador média, sob delineamento estratificado (Região) em grupos duas em etapas (RegCAE, Id). Os pontos a vermelho representam a média das estimativas do desvio padrão da média estimada do VVN por estimador. A linha horizontal pontilhada representa o desvio padrão da estimativa média do VVN, por população.

A distribuição dos estimadores da variância varia consoante o delineamento e a população. De um modo geral, os estimadores indicados na literatura para os delineamentos apresentados e os estimadores *Taylor* e *Jackknife* possuem distribuições idênticas, com exceção nos delineamentos III e IV, onde as estimativas variam entre os quatro estimadores. Há situações em que os estimadores são não enviesados ou apresentam um enviesamento pequeno e tolerável, mas há outras em que o enviesamento não é tolerável. O estimador *Bootstrap* é o que apresenta maior dispersão nas estimativas. Nos delineamentos I e II as distribuições amostrais são aproximadamente simétricas, enquanto nos delineamentos III e IV para algumas populações a distribuição é assimétrica com presença de *outliers* superiores e inferiores. Em alguns delineamentos há uma maior dispersão entre os *outliers* inferiores do que nos superiores, enquanto em outros acontece o contrário. Nos delineamentos III e IV temos uma maior concentração de *outliers* superiores do que inferiores. A assimetria positiva e a presença de *outliers* nos quatro estimadores é comum nos delineamentos III e IV. Um cenário aproximadamente simétrico verifica-se nos delineamentos I e II.

4.4 Análise comparativa dos estimadores da variância

Em cada uma das populações consideradas, e para cada cenário descrito acima, temos em anexo, nas Tabela A.1, A.2 e A.3, os resultados obtidos, com as simulações realizadas, para as medidas de precisão e eficiência dos estimadores.

Pela análise dos resultados do delineamento I, nas três populações temos que o estimador *Bootstrap* é o menos eficiente apresentando os maiores valores de $REQME$, e os restantes três estimadores tem um desempenho similar e com valores de $REQME$ semelhantes. A distribuição de todos estimadores é simétrica com presença de *outliers* superiores e inferiores nas três populações. Nas populações I e III os estimadores *Taylor*, *Jackknife* e o indicado na literatura (Estratificado) têm estimativas do enviesamento quase nulas. No entanto, na população I o estimador *Bootstrap* é enviesado mas é não enviesado na população III. Na população II todos os estimadores têm o mesmo enviesamento.

No delineamento II as estimativas do enviesamento entre os estimadores para as três populações são semelhantes às do delineamento I, com o estimador *Bootstrap* com a

maior dispersão e maior valor do $REQME$. A distribuição dos estimadores é simetrica com presença de *outliers* superiores e inferiores nas três populações.

A observar que o estimador *Jackknife* em planos estratificados simples (delineamentos I e II) fornece os mesmos resultados que os estimadores lineares usuais de variância, além disso, as suas propriedades são razoáveis para alguns casos de estimadores não lineares (Silva e Pessoa, 1998, p.45).

Quanto ao delineamento III, temos que a distribuição dos estimadores varia nas três populações em estudo. O estimador indicado na literatura (Etapas2) é não enviesado e é o mais preciso. Nas populações I e III, o estimador *Taylor* tem um ligeiro enviesamento negativo e os estimadores *Boostrap* e *Jackknife* apresentam algum enviesamento positivo, que é superior no estimador *Boostrap,* mas o estimador *Boostrap* é mais eficiente do que o estimador *Jackknife.* As distribuições dos quatro estimadores são assimétricas positivas e apenas existem *outliers* superiores. Os resultados globais $REQME$ indicam o estimador indicado na literatura (Etapas2) como o mais eficiente e preciso, seguido do estimador *Boostrap* com os melhores resultados das medidas a situarem-se pelo triplo dos resultados do estimador indicado na literatura. Na população II a distribuição dos estimadores *Taylor, Boostrap* e *Jackknife,* sendo aproximadamente simétrica e não enviesada.

No delineamento IV, não foi possível obter os resultados das estimativas para o estimador indicado na literatura (Etapa3) devido à dificuldade do cálculo das probabilidades de inclusão de segunda ordem nas diferentes etapas do processo. Este cálculo além de ser complicado por vezes é até impossível para determinados procedimentos de amostragem. À semelhança do que se verificou no delineamento III, no delineamento IV a distribuição dos estimadores varia com a população em estudo. Nas populações I e III os estimadores *Taylor* e *Boostrap* são enviesados, sendo o estimador *Boostrap* o mais enviesado e menos preciso. A distribuição dos três estimadores é assimétrica positiva. Na população II, o comportamento dos três estimadores é semelhante sendo todos centrados e com distribuição quase simétrica. De salientar em todas as populações a prevalência de *outliers* superiores.

Em suma, o estimador indicado na literatura é o mais eficiente nas três populações em estudo, seguido do estimador *Taylor* e por último o estimador *Bootstrap* caracterizado por ter o maior enviesamento, maior dispersão e o maior erro quadrático médio das estimativas.

Capítulo 5

Conclusões e recomendações

Esta dissertação pretende encorajar aos estudantes e investigadores que analisam dados obtidos por técnicas de amostragem complexas, como os que foram aqui usados como suporte. A amostragem é útil quando utilizamos os procedimentos corretos para a condução do trabalho prático (desde antes da recolha dos dados até à parte da análise). Caso contrário, podemos chegar a conclusões equivocadas e disseminar esse conhecimento enviesado ou ainda tomar decisões erradas que podem prejudicar o funcionamento de empresas privadas, órgãos públicos e a população geral.

O presente estudo visou avaliar o desempenho dos estimadores da variância da média amostral quando se utilizam delineamentos de amostragem complexos. Com estes procedimentos, pretendemos contribuir para o conhecimento e divulgação das alternativas existentes para estimação da variância, as quais são fundamentais para a realização de inferências estatísticas feitas a partir de amostras complexas.

Nesse âmbito foram geradas três populações com características diferentes e foram considerados quatro delineamentos de amostragem: I) estratificado por 1 variável, II) estratificado por duas variáveis, III) por grupos a duas etapas, e IV) multietápico (estratificado e por grupos a duas etapas).

5.1 Conclusões

Os resultados empíricos do presente estudo mostraram que não existe um padrão consistente no comportamento dos estimadores da variância por população.

Nos delineamentos de amostragem estratificados (delineamento I e II), os resultados dos estimadores *Taylor*, *Jackknife* e o indicado na literatura foram iguais. Apresentaram melhores resultados quanto à eficiência e à precisão das estimativas do que o estimador *Bootstrap.* Verificou-se assim, à semelhança do referido por Silva e Pessoa (1998, pág.

45), que em delineamentos estratificados o estimador *Jackknife* forneceu os mesmos resultados que o estimador *Taylor*.

Resultados diferentes podem ser visualizados nos planos multietápicos (delineamentos III e IV), que são os planos que melhor se aproximam dos planos implementados pelas instituições responsáveis pelas estatísticas oficiais.

No delineamento por grupos a duas etapas (delineamento III), o estimador indicado na literatura (Etapas2) para a variância da média amostral é sempre não enviesado e foi o mais eficiente dentre os avaliados, seguido do estimador *Taylor* com valores de precisão ligeiramente inferiores aos do *Jackknife* e *Bootstrap.*

Kovar *et al.,* (1998) compararam os estimadores *Taylor, Balanced Replicated Replication (BRR), Jackknife* e *Bootstrap* com um estudo de simulação, baseado em populações hipotéticas, construídas para se assemelharem à população do estudo *National Assessment of Educational Progress*. Usaram um delineamento por grupos em duas etapas, consistindo a primeira etapa na estratificação de acordo com uma certa variável. As unidades da segunda etapa são assumidas como sendo observações individuais. Os autores concluíram, em relação à precisão das estimativas de variância do estimador não linear da razão, que os estimadores *Taylor* e *Jackknife* tiveram os melhores desempenhos, equivalentes entre si. No nosso estudo, a igualdade no desempenho destes dois estimadores (*Taylor* e *Jackknife*) só foi verificada nos delineamentos estratificados (delineamento I e II) nas três populações.

Resultados similares ao do delineamento anterior foram obtidos no delineamento multietápico (delineamento IV) com o enviesamento, ou não, dos estimadores *Taylor, Bootstrap* e *Jackknife* a depender das características da população em estudo. O estimador *Taylor* foi o mais eficiente, seguido pelo estimador *Jackknife* sendo o estimador *Bootstrap* o que apresentou menor precisão. Por meio do enviesamento, podemos observar que estes foram irrelevantes comparativamente à grandeza do erro-padrão das estimativas da variância. Este facto leva à conclusão de que, nas condições em que foi feito o estudo, os problemas de precisão dos estimadores estão mais associados à não eficiência das estimativas do que ao enviesamento.

Estes resultados coincidem com os observados por Kish & Frankel (1974) e Bean (1975), que concluíram que não existe um padrão consistente de enviesamento nos estimadores *Taylor, Balanced Replicated Replication (BRR) e Jackknife*. No nosso estudo, dependendo do delineamento e população em estudo, verificou-se que os estimadores da variância *Taylor, Bootstrap* e *Jackknife* tanto podiam ser enviesados, como ter um enviesamento pequeno e tolerável, como ter um enviesamento não tolerável.

Rao e Wu (1985, 1988) demonstraram que os resultados do estimador *Jackknife* estão mais próximos do estimador linearização *Taylor* do que o estimador *Bootstrap*. No geral, os métodos *Jackknife* e *Taylor* tendem a apresentar um desempenho similar. Eles são mais estáveis para funções suaves mas inconsistentes para funções não-suaves.

Podemos assim concluir que pela natureza dos planos implementados no estudo, os estimadores indicados na literatura para os delineamentos estratificados e por grupos em duas etapas foram os que apresentaram o menor enviesamento nos resultados, e garantem a maior precisão e confiabilidade nas estimativas. Segue-se o estimador *Taylor* com resultados das aproximações mais fiáveis.

5.2 Recomendações

Considerando que alguns estimadores avaliados exibiram resultados equivalentes em relação à precisão e ao enviesamento nos delineamentos estratificados, as questões relacionadas à operacionalização passam a ter peso preponderante na decisão sobre qual o estimador a utilizar.

Nos delineamentos por grupos em duas etapas, recomenda-se o uso dos estimadores usuais indicados na literatura, seguido do estimador *Taylor*.

Em futuros estudos propomos que se analisem outros planos multietápicos procurando sempre evidenciar o melhor desempenho dentre os estimadores da variância. Sugere-se o desenvolvimento das fórmulas para cada um destes cálculos das estimativas em múltiplas etapas e a comparação com os estimadores usuais da variância.

Referências Bibliográficas

- Afonso, A., Nunes, C. (2010). Estatística e probabilidades. Aplicações e soluções em SPSS. Escolar Editora, Lisboa.
- Bean JA. (1975). Distribution and properties of variance estimators for complex multistage probability samples. An empirical distribution. *Vital Health Statistical 2*. (65), i-iv, 1–46.
- Cochran, W. G. (1977). *Sampling Techniques*, 3rd Edition. New York: John Wiley & Sons.
- Chambers, R.L. and Skinner, C.J., eds, (2003). *Analysis of Survey Data*. Chichester: John Wiley.
- Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. CBMS-NSF Regional Conference Series in Applied Mathematics, Monograph 38, SIAM, Philadelphia.
- Gross, S. T. (1980). Median Estimation in Sample Surveys. In Proceedings of the Section on Survey Research Methods, American Statistical Association, 181–184.
- Heeringa, S. G., West, B. T. and Berglund, P. A. (2010). *Applied Survey Data analysis*. Boca Raton, FL: Champman & Hall / CRC.
- Kish L., Frankel M. R. (1974), Inference from complex samples. *Journal of the Royal Statistical Society. Series B (Methodological),* 36(1):1–37.
- Kovar J. G., Rao J. N. K., Wu C. F. J. (1998). Bootstrap and other methods to measure errors in survey estimates. *Canadian Journal of Statistics,* 16(Supl), 25–45.
- Koop, J. C. (1972). On the deviation of expected values and variance of ratios without the use of infinite series expansions. *Metrika* 19, 156–170
- Kreuter F., Valliante R. (2007). A Survey on Survey Statistics: What is done and can be done in Stata. *The Stata Journal*, 7(1), 1–21.

- Krewski, D., and Rao, J. N. K. (1981). Inference from Stratified Samples: Properties of the linearization, jackknife and balanced repeated replication methods. *The Annals of Statistics*, 9, 1010–1019.
- Lavrakas, P. J. (2008). *Encyclopedia of survey research methods*. SAGE Publications.
- Levy, P.S., Lemeshow, S. (1991). *Sampling of Populations Methods and Applications*. Second Edition. New York: Wiley.
- Lorh, S. L. (2010). *Sampling: Design and Analysis*. Second Edition. Boston: Michelle Julet.
- Münnich, R. (2005): *Datenqualität in komplexen Stichprobenerhebungen*, Unpublished thesis, university Tübingen.
- Pessoa, D. G. C. e Silva P. L. N. (1998). *Análise de Dados Amostrais Complexos*. São Paulo: Associação Brasileira de Estatística.
- Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika*, 43, 353–360.
- Rao, J. N. K., and Wu, C. F. J. (1985). Inference from Stratified Samples: Second-order analysis of three methods for nonlinear statistics. *Journal of the American Statistical Association*, 80, 620–630.
- Rao, J. N. K., and Shao, J. (1992). Jackknife Variance Estimation with Survey Data under hot deck imputation. *Biometrika*, 79, 811–822.
- Rao, J. N. K., and Wu, C. F. J. (1988*).* Resampling Inference with Complex Survey Data. *Journal of the American Statistical Association*, 83, 231–241.
- Shao, J., and Tu, D. (1995). *The jackknife and bootstrap*. NewYork: Springer.
- Skinner, C.J., Holt, D. and Smith, T.M.F. (1989). *Analysis of Complex Surveys*. Chichester, England: John Wiley & Sons.
- Skinner, C.J., Holt, D. and Smith, T.M.F. (1989). General Introduction. In C.J. Skinner, D. Holt, and T.M.F. Smith (Eds.), *Analysis of Complex Surveys* (pp. 1-20). Chichester, England: John Wiley & Sons.

- Szwarcwald, C. L. e Damacena, G. N. (2008). Amostras Complexas em Inquéritos Populacionais: Planeamento e implicações na análise estatística dos dados. *Rev Bras Epidemiol,* 11(l1), 38–45.
- Tukey, J. W. (1958). Bias and Confidence in Not-quite large Samples [Abstract]. *Annals of Mathematical Statistics*, 29, 614.
- Viera, M. D. T. (2013). Notas de aula de Amostragem. *Universidade Federal de Juiz de Fora*.
- Vicente, P., Reis, E., e Ferrão, F. (2001). *Sondagens. A amostragem com factor decisivo de qualidade*. Lisboa: Edições Sílabo.
- Wolter, K. M. (2007). *Introduction to Variance Estimation*. Second Edition. Springer.
- Walther, B. A. e Moore, J. L. (2005). The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. *Ecography,* 28, 815–829.

Anexos

Apêndice – A

Resultados das simulações do Capítulo III

População I

Tabela A.1 Esperança, variância, coeficiente de variação, enviesamento, raiz do erro quadrático médio de $\widehat{Var}(\hat{\mu})$ e raiz do erro quadrático médio da $\widehat{Var}(\hat{\mu})$ escalado. O resultado da $\widehat{Var}\left(\widehat{Var}(\hat{\mu})\right)$ foi dividido por 1 000 000 000 e da $\sqrt{\widehat{EQM}\left(\widehat{Var}(\hat{\mu})\right)}$ por 1 000 000 (População I).

Estimador	Estimador de Variância	Plano Amostral			
		I	II	III	IV
Esperança $\hat{E}(\hat{\mu})$		201767,3	201743,4	201948,0	201729,6
Enviesamento $\widehat{Env}(\hat{\mu})\%$		-0,0162	-0,0280	0,0733	-0,0349
Variância $\widehat{Var}(\hat{\mu})$		77785753	80370546	71968081	92935042
Esperança $\hat{E}\left(\widehat{Var}(\hat{\mu})\right)$	*Est.1,2,3* *Taylor* *Bootstrap* *Jackknife*	78535726 78535726 82231164 78535726	78433206 78433206 78540415 78433206	64080462 49588617 162622217 141550293	- 69025976 174264372 116997721
Variância $\widehat{Var}\left(\widehat{Var}(\hat{\mu})\right)$	*Est.1,2,3* *Taylor* *Bootstrap* *Jackknife*	15158,38 15158,38 30503,92 15158,38	15402,72 15402,72 27493,36 15402,72	340162 2478297 18033059 40669573	- 5923723 36678363 22297808
Enviesamento $\widehat{Env}\left(\widehat{Var}(\hat{\mu})\right)\%$	*Est.1,2,3* *Taylor* *Bootstrap* *Jackknife*	0,96 0,96 5,71 0,96	-2,41 -2,41 -2,28 -2,41	-10,96 -31,10 125,96 96,68	- -25,73 87,51 25,89
Coeficiente de Variação $\widehat{CV}\left(\widehat{Var}(\hat{\mu})\right)\%$	*Est.1,2,3* *Taylor* *Bootstrap* *Jackknife*	4,96 4,96 6,72 4,96	5,00 5,00 6,68 5,00	28,78 100,39 82,58 142,47	- 111,50 109,90 127,63
Erro quadrático médio $\sqrt{\widehat{EQM}\left(\widehat{Var}(\hat{\mu})\right)}$	*Est.1,2,3* *Taylor* *Bootstrap* *Jackknife*	3,9650 3,9650 7,0898 3,9650	4,3768 4,3768 5,5536 4,3768	20,0593 54,5815 162,0223 213,3337	- 80,5938 208,0693 151,2509
Erro quadrático médio escalado $\frac{\sqrt{\widehat{EQM}(\widehat{Var}(\hat{\mu}))}}{\hat{E}(\widehat{Var}(\hat{\mu}))}\%$	*Est.1,2,3* *Taylor* *Bootstrap* *Jackknife*	5,0486 5,0486 8,6218 5,0486	5,5802 5,5802 7,0710 5,5802	31,3033 110,0686 99,6311 150,7123	- 116,7587 119,3986 129,2768

População II

Tabela A.2 Esperança, variância, coeficiente de variação, enviesamento, raiz do erro quadrático médio de $\widehat{Var}(\hat{\mu})$ e raiz do erro quadrático médio da $\widehat{Var}(\hat{\mu})$ escalado. O resultado da $\widehat{Var}\left(\widehat{Var}(\hat{\mu})\right)$ foi dividido por 1 000 000 000 e da $\sqrt{\widehat{EQM}\left(\widehat{Var}(\hat{\mu})\right)}$ por 1 000 000 (População II).

Estimador	Estimador de Variância	Plano Amostral			
		I	II	III	IV
Esperança $\hat{E}(\hat{\mu})$		205975,7	205759,1	205944,7	205705,5
Enviesamento $\widehat{Env}(\hat{\mu})\%$		0,01603	-0,0892	0,0009	-0,1152
Variância $\widehat{Var}(\hat{\mu})$		81395228	78604798	80114827	83094434
Esperança $\hat{E}\left(\widehat{Var}(\hat{\mu})\right)$	*Est.1,2,3* *Taylor* *Bootstrap* *Jackknife*	79237303 79237303 79315454 79237303	79137559 79137559 79207276 79137559	80882139 92960504 92982893 93219186	- 98736576 98767468 98746558
Variância $\widehat{Var}\left(\widehat{Var}(\hat{\mu})\right)$	*Est.1,2,3* *Taylor* *Bootstrap* *Jackknife*	14927,29 14927,29 27718,90 14927,29	14235,39 14235,39 26594,42 14235,39	372533 3297198 3321897 3331121	- 6169070 6199742 6170350
Enviesamento $\widehat{Env}\left(\widehat{Var}(\hat{\mu})\right)\%$	*Est.1,2,3* *Taylor* *Bootstrap* *Jackknife*	-2,65 -2,65 -2,56 -2,65	0,68 0,68 0,77 0.68	0,96 16,03 16,06 16,36	- 18,82 18,86 18,84
Coeficiente de Variação $\widehat{CV}\left(\widehat{Var}(\hat{\mu})\right)\%$	*Est.1,2,3* *Taylor* *Bootstrap* *Jackknife*	4,88 4,88 6,64 4,88	4,77 4,77 6,51 4,77	23,86 61,77 61,99 61,91	- 79,54 79,72 79,55
Erro quadrático médio $\sqrt{\widehat{EQM}\left(\widehat{Var}(\hat{\mu})\right)}$	*Est.1,2,3* *Taylor* *Bootstrap* *Jackknife*	4,4254 4,4254 5,6608 4.4254	3,8104 3,8104 5,1920 3,8104	19,3164 58,8405 59,0549 59,1848	- 80,0858 80,2832 80,0958
Erro quadrático médio escalado $\frac{\sqrt{\widehat{EQM}(\widehat{Var}(\hat{\mu}))}}{\hat{E}(\widehat{Var}(\hat{\mu}))}\%$	*Est.1,2,3* *Taylor* *Bootstrap* *Jackknife*	5,5850 5,5850 7,1370 5,5850	4,8150 4,8150 6,5550 4,8150	23,8821 63,2963 63,5116 63,4900	- 81,1106 81,2850 81,1125

População III

Tabela A.3 Esperança, variância, coeficiente de variação, enviesamento, raiz do erro quadrático médio de $\widehat{Var}(\hat{\mu})$ e raiz do erro quadrático médio da $\widehat{Var}(\hat{\mu})$ escalado. O resultado da $\widehat{Var}\left(\widehat{Var}(\hat{\mu})\right)$ foi dividido por 1 000 000 000 e da $\sqrt{\widehat{EQM}\left(\widehat{Var}(\hat{\mu})\right)}$ por 1 000 000 (População III).

Estimador	Estimador de Variância	Plano Amostral			
		I	II	III	IV
Esperança $\hat{E}(\hat{\mu})$		380654,2	380346,7	379971,5	381110,8
Enviesamento $\widehat{Env}(\hat{\mu})\%$		0,0869	0,0060	-0,0926	0,2069
Variância $\widehat{Var}(\hat{\mu})$		129392304	126328741	133779640	157486066
Esperança $\hat{E}\left(\widehat{Var}(\hat{\mu})\right)$	*Est.1,2,3*	124957067	125059890	106966297	-
	Taylor	124957067	125059890	94042398	133901442
	Bootstrap	130600440	125211683	280023995	315820155
	Jackknife	124957067	125059890	270591623	225394318
Variância $\widehat{Var}\left(\widehat{Var}(\hat{\mu})\right)$	*Est.1,2,3*	15650,3	15736,8	1326409	-
	Taylor	15650,3	15736,8	10128625	20099030
	Bootstrap	51009,0	47955,2	51780019	103376925
	Jackknife	15650,3	15736,8	138955806	73628106
Enviesamento $\widehat{Env}\left(\widehat{Var}(\hat{\mu})\right)\%$	*Est.1,2,3*	-3,43	-1,00	-20,04	-
	Taylor	-3,43	-1,00	-29,70	-14,98
	Bootstrap	0,93	-0,88	109,32	100,54
	Jackknife	-3,43	-1,00	102,27	30,13
Coeficiente de Variação $\widehat{CV}\left(\widehat{Var}(\hat{\mu})\right)\%$	*Est.1,2,3*	3,17	3,17	34,05	-
	Taylor	3,17	3,17	107,02	105,88
	Bootstrap	5,47	5,53	81,26	101,81
	Jackknife	3,17	3,17	137,76	120,39
Erro quadrático médio $\sqrt{\widehat{EQM}\left(\widehat{Var}(\hat{\mu})\right)}$	*Est.1,2,3*	5,9432	4,1649	45,2257	-
	Taylor	5,9432	4,1649	108,2020	143,7194
	Bootstrap	7,2435	7,0145	270,4948	358,3945
	Jackknife	5,9432	4,1649	397,0810	279,7135
Erro quadrático médio escalado $\frac{\sqrt{\widehat{EQM}(\widehat{Var}(\hat{\mu}))}}{\hat{E}(\widehat{Var}(\hat{\mu}))}\%$	*Est.1,2,3*	4,7562	3,3304	42,2803	-
	Taylor	4,7562	3,3304	115,0566	107,3322
	Bootstrap	5,5463	5,6021	96,5970	113,4806
	Jackknife	4,7562	3,3304	146,7455	124,0996

Apêndice – B

Código R

B.1 Simulação geração da população I

```
N <- 10000
set.seed(8)
Id <- c(1:N)

Regiao <- sample(1:3, N, rep=TRUE, p=c(.137,.397,.466))
Regiao <- factor(Regiao, labels= c("Norte", "Centro", "Sul"))
table(Regiao)
prop.table(table(Regiao))*100

library("plotrix")
pie3D(table(Regiao),col=c("blue","red","yellow"),cex.main=1.5,labels=c("Norte - 14.18%",
"Centro - 39.20%", "Sul - 46.62%"), cex=0.7,explode=0.2)
title("Distribuição das empresas por Região",cex=0.4)

CAE<- sample(45:47, N, rep=TRUE, p=c(.057,.049,.894))
CAE <- factor(CAE, labels = c("CAE_45","CAE_46", "CAE_47"))
prop.table(table(CAE))*100
pie3D(table(CAE),col=c("blue","red","yellow"),cex.main=1.5,labels=c("CAE_45","CAE_46","C
AE_47"), cex=0.7,explode=0.2)
title("Distribuicao das empresas por CAE",cex=0.4)

pBN <- sample(c(0.3, 0.35, 0.4), size=N, replace=T, prob=c(.25, 0.05, .70))
kBN <- ifelse(pBN==0.3, 90, ifelse(pBN==0.35, 40, 4))
NPS <- rnbinom(N, kBN, pBN)+1
length(NPS)
tab<-table(NPS)
barplot(tab)
title("Numero de pessoal ao servico por Empresa",cex=0.4)
summary(NPS)

  EPS <-NA
  EPS[NPS < 50]<-1
  EPS[NPS >= 50 & NPS < 250]<-2
  EPS[NPS >= 250]<-3

  table(EPS)
  prop.table(table(EPS))
  EPS <- factor(EPS, labels = c("Pequena", "Media", "Grande"))
  prop.table(table(EPS))*100
  pie3D(table(EPS),col=c("blue","red","yellow"),cex.main=1.5,labels=c("Pequena          -
69.30%","Media - 20,31%","Grande - 10,39%"), cex=0.7,explode=0.2)
  title("Distribuicao das empresas por EPS",cex=0.4)

m <- 3200
b <- 6800
S <- 800
VVN <- m*NPS + b + rnorm(N, 0, sd=S*NPS)
summary(VVN)

  mod<-lm(VVN~NPS)
  mod
  plot(NPS, VVN, abline(reg = mod, col = 2))
  title("NPS vs VVN",cex=0.4)

empresa <- data.frame(Id, Regiao, CAE, NPS, EPS, VVN)

  empresa$RegCAE<-paste(empresa$Regiao,empresa$CAE,sep=".")
  empresa$CAE.EPS<-paste(empresa$CAE,empresa$EPS,sep=".")
  empresa$RegEPS<-paste(empresa$Regiao,empresa$EPS,sep=".")

write.table(empresa, file="Universo_2015.csv")
universo1 <- read.table("Universo_2015.csv")
head(universo1)
str(universo1)

  FCP1 <- data.frame(table(universo1$EPS))
  EPS <- FCP1$Var1
  Ni.EPS <- FCP1$Freq
  FCP1 <- data.frame(EPS, Ni.EPS)
  universo1 <- merge(universo1, FCP1, by.universo1=EPS, by.FCP1=EPS, All=T)
  FCP2 <- data.frame(table(universo1$Regiao))
  Regiao <- FCP2$Var1
  Ni.Regiao <- FCP2$Freq
  FCP2 <- data.frame(Regiao, Ni.Regiao)
```

```
  universo1 <- merge(universo1, FCP2, by.universo1=Regiao, by.FCP2=Regiao, All=T)

  FCP3 <- data.frame(table(universo1$CAE))
  CAE <- FCP3$Var1
  Ni.CAE <- FCP3$Freq
  FCP3 <- data.frame(CAE, Ni.CAE)
  universo1 <- merge(universo1, FCP3, by.universo1=CAE, by.FCP3=CAE, All=T)

  FCP4 <- data.frame(table(universo1$RegCAE))
  RegCAE <- FCP4$Var1
  Ni.RegCAE <- FCP4$Freq
  FCP4 <- data.frame(RegCAE, Ni.RegCAE)
  universo <- merge(universo1, FCP4, by.universo1=RegCAE, by.FCP4=RegCAE, All=T)

  FCP5 <- data.frame(table(universo$CAE.EPS))
  FCP5 <- data.frame(CAE.EPS=FCP5$Var1, Ni.CAE.EPS=FCP5$Freq)
  universo <- merge(universo, FCP5, by.universo=CAE.EPS, by.FCP5=CAE.EPS, All=T)

  FCP6 <- data.frame(table(universo$RegEPS))
  FCP6 <- data.frame(RegEPS=FCP6$Var1, Ni.RegEPS=FCP6$Freq)
  universo <- merge(universo, FCP6, by.universo=RegEPS, by.FCP6=RegEPS, All=T)

write.table(universo, file="universo.csv")
universo <- read.table("universo.csv")

table(universo$CAE, universo$Regiao)
table(universo$EPS, universo$CAE)

Media.R <- by(universo$VVN, universo$Regiao, mean)
Media.T <- by(universo$VVN, universo$RegCAE, mean)
Media.T2 <- by(universo$VVN, universo$RegEPS, mean)

basicStats(universo$VVN)
basicStats(universo$NPS)
```

B.2 Delineamento I

```
N.I <- nrow(universo)
k <- length(unique(universo$RegCAE))
N.I <- by(universo$VVN, universo$RegCAE,length)
sigmai.I <- by(universo$VVN, universo$RegCAE, sd)

d.I <- 17370
np.I     <-     stratasize(e=d.I,as.numeric(N.I),     Sh=as.numeric(sigmai.I),level=0.95,
type="opt")
np.I$n
nip.I <- stratasamp(np.I$n,as.numeric(N.I), Sh=as.numeric(sigmai.I), type="opt")
nip.I

dados1<-universo[order(as.numeric(universo$RegCAE)),]
dados1$probin<-inclusionprobastrata(as.numeric(dados1$RegCAE),nip.I[2,])
dados1$wij <- (1/dados1$probin)

fi.I<-nip.I[2,]/as.numeric(N.I)
Wi.I<-as.numeric(N.I)/sum(as.numeric(N.I))

set.seed(92)
Est.media <- NA, Var.Taylor <- NA, SE.Taylor <- NA, Var.Bootstrap <- NA, SE.Bootstrap <-
NA
Var.JKn <- NA, SE.JKn <- NA, mean.Est <- NA, Var.Est<-NA, SE.Est<-NA

MM <- 10000
for (i in 1:MM){
  s.I <- strata(dados1,"RegCAE",size=nip.I[2,], method="srswor")
  amostra.1<-getdata(dados1,s.I)

    mean.i<-as.numeric(by(amostra.1$VVN,amostra.1$RegCAE,mean))
    mean.Est[i]<-sum(Wi.I*mean.i)
    si2<-as.numeric(by(amostra.1$VVN,amostra.1$RegCAE,var))
    ni<-as.numeric(table(amostra.1$RegCAE))
    Var.Est[i]<-sum(Wi.I^2*si2/ni*(1-fi.I))
    SE.Est[i]<-sqrt(Var.Est[i])

nI.plan <- svydesign(ids = ~1, strata = ~RegCAE, weights = ~wij, fpc =~Ni.RegCAE, data =
amostra.1)
```

```
summary(nI.plan)
    mediaT <- svymean(~ VVN, nI.plan)
    Est.media[i] <- coef(mediaT)
    Var.Taylor[i] <- SE(mediaT)^2
    SE.Taylor[i] <- SE(mediaT)

    rnI.planBoot <- as.svrepdesign(nI.plan, type = "bootstrap", replicates = 1000)
    mediaB <- svymean(~VVN, rnI.planBoot)
    Var.Bootstrap[i] <- SE(mediaB)^2
    SE.Bootstrap[i] <- SE(mediaB)

    rnI.planJkn <- as.svrepdesign(nI.plan, type = "JKn")
    mediaJ <- svymean(~VVN, rnI.planJkn)
    Var.JKn[i] <- SE(mediaJ)^2
    SE.JKn[i] <- SE(mediaJ)

print(c(i,Est.media[i]))
}

Esq.1.10M <- data.frame(Est.media, Var.Taylor, SE.Taylor, Var.Bootstrap, SE.Bootstrap,
Var.JKn, SE.JKn, Var.Est, SE.Est)
write.table(Esq.1.10M, file="Esq.1.10M.csv", dec=",")

  M.Media <- mean(Est.media)
  Var.Media <- var(Est.media)
  SE.Media <- sd(Est.media)
SE.1   <-   cbind(Est.Estratificado   =   SE.Est,   Taylor=   SE.Taylor,   Bootstrap=
SE.Bootstrap,Jackknife= SE.JKn)
  medias<-cbind(mean(SE.Est), mean(SE.Taylor), mean(SE.Bootstrap), mean(SE.JKn))
  boxplot.matrix(SE.1, main ="Estimativas da Variancia", ylab = "erro padrão")
  abline(h= SE.Media, v=0, lty= "dotted")
  points(medias[1,], pch =16, col ="red")

    M.Est.Estratificado <- mean(Var.Est)
    Var.M.Est.Estratificado <- var(Var.Est)
    Var.M.Est.Estratificado.p <- var(Var.Est)/1000000000
    CV.Est.Estratificado <- sqrt(Var.M.Est.Estratificado)/M.Est.Estratificado*100

    M.Taylor <- mean(Var.Taylor)
    Var.M.Taylor <- var(Var.Taylor)
    Var.M.Taylor.p <- var(Var.Taylor)/1000000000
    CV.Taylor <- sqrt(Var.M.Taylor)/M.Taylor*100

    M.Bootstrap <- mean(Var.Bootstrap)
    Var.M.Bootstrap <- var(Var.Bootstrap)
    Var.M.Bootstrap.p <- var(Var.Bootstrap)/1000000000
    CV.Bootstrap <- sqrt(Var.M.Bootstrap)/M.Bootstrap*100

    M.JKn <- mean(Var.JKn)
    Var.M.JKn <- var(Var.JKn)
    Var.M.JKn.p <- var(Var.JKn)/1000000000
    CV.JKn <- sqrt(Var.M.JKn)/M.JKn*100

Est.mediasI   <-   c(M.Media,   Var.Media,   M.Taylor,   Var.M.Taylor,   M.Bootstrap,
Var.M.Bootstrap, M.JKn, Var.M.JKn,  M.Est.Estratificado, Var.M.Est.Estratificado)
Est.mediasI

  Env.M <- M.Media - mean(universo$VVN)
  Env.Mp <- (Env.M/mean(universo$VVN))*100

  Env.Est.I <- M.Est.Estratificado - Var.Media
  Env.Est.Ip <- (Env.Est.I/Var.Media )*100

  Env.T.I <- M.Taylor - Var.Media
  Env.T.Ip <- (Env.T.I/Var.Media )*100

  Env.B.I <- M.Bootstrap - Var.Media
  Env.B.Ip <- ( Env.B.I/Var.Media)*100

  Env.J.I <- M.JKn - Var.Media
  Env.J.Ip <- (Env.J.I /Var.Media)*100

  EQM.Est.I <- Var.M.Est.Estratificado + (Env.Est.I)^2
  EQM.T.I <- Var.M.Taylor + (Env.T.I)^2
  EQM.B.I <- Var.M.Bootstrap + (Env.B.I)^2
  EQM.J.I <- Var.M.JKn + (Env.J.I)^2
```

```
  REQM.Est.I <- sqrt(EQM.Est.I)/1000000
  REQM.T.I <- sqrt(EQM.T.I)/1000000
  REQM.B.I <- sqrt(EQM.B.I)/1000000
  REQM.J.I <- sqrt(EQM.J.I)/1000000

  REQME.Est.I <- sqrt(EQM.Est.I)/M.Est.Estratificado*100
  REQME.T.I <- sqrt(EQM.T.I)/M.Taylor*100
  REQME.B.I <- sqrt(EQM.B.I)/M.Bootstrap*100
  REQME.J.I <- sqrt(EQM.J.I)/M.JKn*100

  alfa <- 0.05
  zalfa <- qnorm(1-alfa/2)
  MerroI <- zalfa * sqrt(77785753)
```

B.3 Delineamento II

```
N.II <- nrow(universo)
N.IIi <- by(universo$VVN, universo$Regiao,length)
sigmai.II <- by(universo$VVN, universo$Regiao, sd)

d.II <- 17370
np.II <- stratasize(e=d.II,as.numeric(N.IIi), Sh=as.numeric(sigmai.II),level=0.95)
np.II$n
nip.II <- stratasamp(np.II$n,as.numeric(N.IIi), Sh=as.numeric(sigmai.II))
nip.II

dados2<-universo[order(as.numeric(universo$Regiao)),]
dados2$probin<-inclusionprobastrata(as.numeric(dados2$Regiao),nip.II[2,])
dados2$wij <- (1/dados2$probin)

fi.II<- nip.II[2,]/as.numeric(N.IIi)
Wi.II<-as.numeric(N.IIi)/sum(as.numeric(N.II))

set.seed(12)
Est.mediaII <- NA, Var.TaylorII <- NA, SE.TaylorII <- NA, Var.BootstrapII <- NA,
SE.BootstrapII <- NA, Var.JKnII <- NA, SE.JKnII <- NA, mean.EstII <-NA, Var.EstII<-NA,
SE.EstII<-NA

for (i in 1:MM){
  s.II <- strata(dados2,"Regiao",size=nip.II[2,], method="srswor")
  amostra.2<-getdata(dados2,s.II)

    mean.i2<-as.numeric(by(amostra.2$VVN,amostra.2$Regiao,mean))
    mean.EstII[i]<-sum(Wi.II*mean.i2)
    si22<-as.numeric(by(amostra.2$VVN,amostra.2$Regiao,var))
    nIIi<-as.numeric(table(amostra.2$Regiao))
    Var.EstII[i]<-sum(Wi.II^2*si22/nIIi*(1-fi.II))
    SE.EstII[i]<-sqrt(Var.EstII[i])

    nII.plan <- svydesign(id=~1, strata = ~Regiao, weights = ~wij, fpc = ~Ni.Regiao,
data = amostra.2)
    summary(nII.plan)

    mediaT.II <- svymean(~ VVN, nII.plan)
    Est.mediaII[i] <- coef(mediaT.II)
    Var.TaylorII[i] <- SE(mediaT.II)^2
    SE.TaylorII[i] <- SE(mediaT.II)

    rnII.planBoot <- as.svrepdesign(nII.plan, type = "bootstrap", replicates = 1000)
    mediaB.II <- svymean(~VVN, rnII.planBoot)
    Var.BootstrapII[i] <- SE(mediaB.II)^2
    SE.BootstrapII[i] <- SE(mediaB.II)

    rnII.planJkn <- as.svrepdesign(nII.plan, type = "JKn")
    mediaJ.II <- svymean(~VVN, rnII.planJkn)
    Var.JKnII[i] <- SE(mediaJ.II)^2
    SE.JKnII[i] <- SE(mediaJ.II)

print(c(i,Est.mediaII[i]))
}

Esq.2.10M <- data.frame(Est.mediaII, Var.TaylorII, SE.TaylorII, Var.BootstrapII,
SE.BootstrapII, Var.JKnII, SE.JKnII, Var.EstII, SE.EstII)
write.table(Esq.2.10M, file="Esq.2.10M.csv")
  M.MediaII <- mean(Est.mediaII)
```

```
  Var.MediaII <- var(Est.mediaII)
  SE.MediaII <- sd(Est.mediaII)
SE.2<cbind(Est.Estratificado=SE.EstII,Taylor=SE.TaylorII,Bootstrap=SE.BootstrapII,Jackkn
ife=SE.JKnII)
  medias.2<-cbind(mean(SE.EstII),        mean(SE.TaylorII),        mean(SE.BootstrapII),
mean(SE.JKnII))
  boxplot.matrix(SE.2, main ="Estimativas da Variancia", ylab = "erro padrão")
  abline(h=SE.MediaII, v=0, lty= "dotted")
  points(medias.2[1,], pch =16, col ="red")

  M.Est.EstratificadoII <- mean(Var.EstII)
  Var.M.Est.EstratificadoII <- var(Var.EstII)
  Var.M.Est.EstratificadoII.p <- var(Var.EstII)/1000000000
  CV.Est.EstratificadoII <- sqrt(Var.M.Est.EstratificadoII)/M.Est.EstratificadoII*100

  M.TaylorII <- mean(Var.TaylorII)
  Var.M.TaylorII <- var(Var.TaylorII)
  Var.M.TaylorII.p <- var(Var.TaylorII)/1000000000
  CV.TaylorII <- sqrt(Var.M.TaylorII)/M.TaylorII*100

  M.BootstrapII <- mean(Var.BootstrapII)
  Var.M.BootstrapII <- var(Var.BootstrapII)
  Var.M.BootstrapII.p <- var(Var.BootstrapII)/1000000000
  CV.BootstrapII <- sqrt(Var.M.BootstrapII)/M.BootstrapII*100

  M.JKnII <- mean(Var.JKnII)
  Var.M.JKnII <- var(Var.JKnII)
  Var.M.JKnII.p <- var(Var.JKnII)/1000000000
  CV.JKnII <- sqrt(Var.M.JKnII)/M.JKnII*100

Est.mediasII  <-  c(M.MediaII,  Var.MediaII,  M.TaylorII,  Var.M.TaylorII,  M.BootstrapII,
Var.M.BootstrapII,         M.JKnII,         Var.M.JKnII,         M.Est.EstratificadoII,
Var.M.Est.EstratificadoII)
Est.mediasII

  Env.MII <- M.MediaII - mean(universo$VVN)
  Env.MIIp <- (Env.MII/mean(universo$VVN))*100

  Env.Est.II <- M.Est.EstratificadoII - Var.MediaII
  Env.Est.IIp <- (Env.Est.II/Var.MediaII )*100

  Env.T.II <- M.TaylorII - Var.MediaII
  Env.T.IIp <- (Env.T.II/Var.MediaII)*100

  Env.B.II <- M.BootstrapII - Var.MediaII
  Env.B.IIp <- (Env.B.II/Var.MediaII)*100

  Env.J.II <- M.JKnII - Var.MediaII
  Env.J.IIp <- (Env.J.II/Var.MediaII)*100

  EQM.Est.II <- Var.M.Est.EstratificadoII + (Env.Est.II)^2
  EQM.T.II <- Var.M.TaylorII + (Env.T.II)^2
  EQM.B.II <- Var.M.BootstrapII + (Env.B.II)^2
  EQM.J.II <- Var.M.JKnII + (Env.J.II)^2

  REQM.Est.II <- sqrt(EQM.Est.II)/1000000
  REQM.T.II <- sqrt(EQM.T.II)/1000000
  REQM.B.II <- sqrt(EQM.B.II)/1000000
  REQM.J.II <- sqrt(EQM.J.II)/1000000


 REQME.Est.II <- sqrt(EQM.Est.II)/M.Est.EstratificadoII*100
 REQME.T.II <- sqrt(EQM.T.II)/M.TaylorII*100
 REQME.B.II <- sqrt(EQM.B.II)/M.BootstrapII*100
 REQME.J.II <- sqrt(EQM.J.II)/M.JKnII*100

 MerroII <- zalfa * sqrt(80370546)
```

B.4 Delineamento III

```
N.3 <- nrow(universo)
N.III <- by(universo$VVN, universo$RegCAE,length)
universo<-universo[order(as.numeric(universo$RegCAE)),]

set.seed(7)
```

```
Est.mediaIII <- NA, Var.TaylorIII <- NA, SE.TaylorIII <- NA, Var.BootstrapIII <- NA,
SE.BootstrapIII <- NA, Var.JKnIII <- NA, SE.JKnIII <- NA, media.Est <- NA, Var.Est.III
<- NA, SE.Est.III <- NA

  for (i in 1:MM){
    nUniv <- cluster(universo, clustername=c("RegCAE"), size=6, method="srswor")
    amostranUniv <- getdata(universo,nUniv)
    N.3i <- nrow(amostranUniv)
    amostranUniv$RegCAE <- factor(amostranUniv$RegCAE)
    table(amostranUniv$RegCAE)

    NUniv.I <- by(amostranUniv$VVN, amostranUniv$RegCAE,length)
    sigmai.IU <- by(amostranUniv$VVN, amostranUniv$RegCAE, sd)
    np.U <- 1000
    nip.U <- stratasamp(np.U, as.numeric(NUniv.I), Sh=as.numeric(sigmai.IU), type="opt")
    nip.U

    amostranUniv$wi.III <- (1/amostranUniv$Prob)
    table(amostranUniv$wi.III)

    dados<-amostranUniv
    dados$peso.ij <- inclusionprobastrata(as.numeric(dados$RegCAE),nip.U[2,])
    dados$wij.III <- 1/dados$peso.ij
    table(dados$wij.III)

    dados$wij.g <- dados$wi.III * dados$wij.III
    table(dados$wij.g)

    s.Iu <- strata(dados,"RegCAE", size=nip.U[2,], method="srswor")
    amostra.3<-getdata(dados, s.Iu)
    table(amostra.3$wij.g)
    sum(amostra.3$wij.g)

    nIII.plan <- svydesign(ids  = ~ RegCAE + Id, weights = ~wij.g, data = amostra.3)
    summary(nIII.plan)

    mediaT.III <- svymean(~ VVN, nIII.plan)
    Est.mediaIII[i] <- coef(mediaT.III)
    Var.TaylorIII[i] <- SE(mediaT.III)^2
    SE.TaylorIII[i] <- SE(mediaT.III)

    rnIII.planBoot <- as.svrepdesign(nIII.plan, type = "bootstrap", replicates = 1000)
    mediaB.III <- svymean(~VVN, rnIII.planBoot)
    Var.BootstrapIII[i] <- SE(mediaB.III)^2
    SE.BootstrapIII[i] <- SE(mediaB.III)

    rnIII.planJkn <- as.svrepdesign(nIII.plan, type = "JK1")
    mediaJ.III <- svymean(~VVN, rnIII.planJkn)
    Var.JKnIII[i] <- SE(mediaJ.III)^2
    SE.JKnIII[i] <- SE(mediaJ.III)

    M <- length(levels(universo$RegCAE))
    m <- 6
    Ni.III <- as.numeric(table(amostranUniv$RegCAE))
    ni.III <- nip.U[2,]

    Si2 <- sum(as.numeric(by(amostranUniv$VVN, amostranUniv$RegCAE, var)))
    Med <- as.numeric(by(amostra.3$VVN, amostra.3$RegCAE, mean))
    si2 <- as.numeric(by(amostra.3$VVN, amostra.3$RegCAE, var))
    table(amostra.3$RegCAE)
    f1 <- m/M
    f2 <- ni.III/Ni.III

    ti<-Ni.III*as.numeric(Med)
    media <- sum(ti)/sum(Ni.III)
    media.Est[i] <- media

    sr2<- sum((ti - Ni.III * media)^2)/ (m-1)
    parte1 <- 1/(mean(Ni.III)^2) * (1-f1) * sr2/ m
    parte2 <- 1/(m * M *(mean(Ni.III)^2)) * sum((Ni.III^2)*(1-f2)*si2/ni.III)
    Var.Est.III[i] <- parte1 + parte2
    SE.Est.III[i] <- sqrt(Var.Est.III[i])

print(c(i,Est.mediaIII[i]))
  }
```

```
Esq.3.10M  <-  data.frame(Est.mediaIII,  Var.TaylorIII,  SE.TaylorIII,  Var.BootstrapIII,
SE.BootstrapIII, Var.JKnIII, SE.JKnIII, Var.Est.III, SE.Est.III)

  write.table(Esq.3.10M , file="Esq.3.10M.csv")

  M.MediaIII <- mean(Est.mediaIII)
  Var.MediaIII <- var(Est.mediaIII)
  SE.MediaIII <- sd(Est.mediaIII)

  SE.3  <-  cbind(Etapas2=SE.Est.III,  Taylor=SE.TaylorIII,  Bootstrap=SE.BootstrapIII,
Jackknife=SE.JKnIII)
  medias.3<-cbind(mean(SE.Est.III),      mean(SE.TaylorIII),      mean(SE.BootstrapIII),
mean(SE.JKnIII))
  boxplot.matrix(SE.3, main ="Estimativas da Variancia", ylab = "erro padrão")
  abline(h=SE.MediaIII, v=0, lty= "dotted")
  points(medias.3[1,], pch =16, col ="red")

  M.Etapas2III <- mean(Var.Est.III)
  Var.M.Etapas2  <- var(Var.Est.III)
  Var.M.Etapas2.p <- var(Var.Est.III)/1000000000
  CV.Etapas2 <- sqrt(Var.M.Etapas2)/M.Etapas2III *100
  M.TaylorIII <- mean(Var.TaylorIII)
  Var.M.TaylorIII <- var(Var.TaylorIII)
  Var.M.TaylorIII.p <- var(Var.TaylorIII)/1000000000
  CV.TaylorIII <- sqrt(Var.M.TaylorIII)/M.TaylorIII*100

  M.BootstrapIII <- mean(Var.BootstrapIII)
  Var.M.BootstrapIII <- var(Var.BootstrapIII)
  Var.M.BootstrapIII.p <- var(Var.BootstrapIII)/1000000000
  CV.BootstrapIII <- sqrt(Var.M.BootstrapIII)/M.BootstrapIII*100

  M.JKnIII <- mean(Var.JKnIII)
  Var.M.JKnIII <- var(Var.JKnIII)
  Var.M.JKnIII.p <- var(Var.JKnIII)/1000000000
  CV.JKnIII <- sqrt(Var.M.JKnIII)/M.JKnIII*100

  Est.mediasIII   <-   c(M.MediaIII,   Var.MediaIII,   M.TaylorIII,   Var.M.TaylorIII,
M.BootstrapIII, Var.M.BootstrapIII, M.JKnIII, Var.M.JKnIII)
  Est.mediasIII

  Env.M.III <- M.MediaIII- mean(universo$VVN)
  Env.M.IIIp <- (Env.M.III /mean(universo$VVN))*100

  Env.Etapas2.III <- M.Etapas2III - Var.MediaIII
  Env.Etapas2.IIIp <- (Env.Etapas2.III/Var.MediaIII)*100

  Env.T.III <- M.TaylorIII - Var.MediaIII
  Env.T.IIIp <- (Env.T.III/ Var.MediaIII)*100

  Env.B.III <- M.BootstrapIII - Var.MediaIII
  Env.B.IIIp <- (Env.B.III/ Var.MediaIII)*100

  Env.J.III <- M.JKnIII - Var.MediaIII
  Env.J.IIIp <- (Env.J.III/ Var.MediaIII)*100

  EQM.Etapas2.III <- Var.M.Etapas2 + (Env.Etapas2.III)^2
  EQM.T.III <- Var.M.TaylorIII + (Env.T.III)^2
  EQM.B.III <- Var.M.BootstrapIII + (Env.B.III)^2
  EQM.J.III <- Var.M.JKnIII + (Env.J.III)^2

  REQM.Etapas2.III <- sqrt(EQM.Etapas2.III)/1000000
  REQM.T.III <- sqrt(EQM.T.III)/1000000
  REQM.B.III <- sqrt(EQM.B.III)/1000000
  REQM.J.III <- sqrt(EQM.J.III)/1000000

 REQME.Est.III <- sqrt(EQM.Etapas2.III)/M.Etapas2III*100
 REQME.T.III <- sqrt(EQM.T.III)/M.TaylorIII*100
 REQME.B.III <- sqrt(EQM.B.III)/M.BootstrapIII*100
 REQME.J.III <- sqrt(EQM.J.III)/M.JKnIII*100

 MerroIII <- zalfa * sqrt(71968081)
```

B.5 Delineamento IV

```
 N.IV <- nrow(universo)
```

```
  N.IVi <- by(universo$VVN, universo$Regiao,length)
  sigmai.IV <- by(universo$VVN, universo$Regiao,sd)

  d.IV <- 17370
  np.IV <- stratasize(e=d.IV,as.numeric(N.IVi), Sh=as.numeric(sigmai.IV),level=0.95)
  np.IV$n
  nip.IV <- stratasamp(np.IV$n,as.numeric(N.IVi), Sh=as.numeric(sigmai.IV))
  nip.IV

  univR<-universo[order(as.numeric(universo$Regiao), as.numeric(universo$CAE)),]

  set.seed(7)
  Est.mediaIV <- NA, Var.TaylorIV <- NA, SE.TaylorIV <- NA, Var.BootstrapIV <- NA,
SE.BootstrapIV <- NA, Var.JKnIV <- NA, SE.JKnIV <- NA, media.Est <- NA, Var.EstIV <- NA,
SE.EstIV <- NA

  for (i in 1:MM){
    NCentro.IV <- subset(univR, Regiao%in%c("Centro"))
    NC.IV <- as.numeric(nrow(NCentro.IV))

    pik1 <- inclusionprobabilities(as.numeric(table(NCentro.IV$CAE)), 2)

    nC45 <- cluster(NCentro.IV, clustername=c("CAE"), size=2, method="systematic",
pik=pik1)
    amostra45.IV <- getdata(NCentro.IV, nC45)
    amostra45.IV$CAE<-factor(amostra45.IV$CAE)
    table(amostra45.IV$CAE)

    Nc45 <- as.numeric(nrow(amostra45.IV))
    Nc45.I <- by(amostra45.IV$VVN, amostra45.IV$CAE,length)
    sigmai.IC45 <- by(amostra45.IV$VVN, amostra45.IV$CAE, sd)
    np.c45 <- 392
    nip.c45 <- stratasamp(np.c45, as.numeric(Nc45.I), Sh=as.numeric(sigmai.IC45))
    nip.c45

    amostra45.IV$wi.C <- (1/as.numeric(amostra45.IV$Prob))
    amostra45.IV$peso.ij                                                    <-
inclusionprobastrata(as.numeric(amostra45.IV$CAE),nip.c45[2,])
    amostra45.IV$wij.C <- 1/amostra45.IV$peso.ij
    amostra45.IV$wij.g <- amostra45.IV$wi.C * amostra45.IV$wij.C

    s.Ice <- strata(amostra45.IV,"CAE", size=nip.c45[2,], method="srswor")
    amostra45.F <- getdata(amostra45.IV, s.Ice)

    NNorte.IV <- subset(univR, Regiao%in%c("Norte"))
    NN.IV <- as.numeric(nrow(NNorte.IV))

    pik2 <- inclusionprobabilities(as.numeric(table(NNorte.IV$CAE)), 2)

    nC46 <- cluster(NNorte.IV, clustername=c("CAE"), size=2, pik=pik2,
method="systematic")
    amostra46.IV <- getdata(NNorte.IV,nC46)
    amostra46.IV$CAE<-factor(amostra46.IV$CAE)
    table(amostra46.IV$CAE)

    Nc46 <- as.numeric(nrow(amostra46.IV))
    Nc46.I <- by(amostra46.IV$VVN, amostra46.IV$CAE,length)
    sigmai.IC46 <- by(amostra46.IV$VVN, amostra46.IV$CAE, sd)
    np.c46 <- 142
    nip.c46 <- stratasamp(np.c46, as.numeric(Nc46.I), Sh=as.numeric(sigmai.IC46))
    nip.c46

    amostra46.IV$wi.C <- (1/as.numeric(amostra46.IV$Prob)) # peso dos grupos amostrados
    amostra46.IV$peso.ij                                                    <-
inclusionprobastrata(as.numeric(amostra46.IV$CAE),nip.c46[2,])
    amostra46.IV$wij.C <- 1/amostra46.IV$peso.ij # peso das amostras
    amostra46.IV$wij.g <- amostra46.IV$wi.C * amostra46.IV$wij.C # peso final

    s.In <- strata(amostra46.IV,"CAE", size=nip.c46[2,], method="srswor")
    amostra46.F <- getdata(amostra46.IV, s.In)

    NSul.IV <- subset(univR, Regiao%in%c("Sul"))
    NS.IV <- as.numeric(nrow(NSul.IV))

    pik3 <- inclusionprobabilities(as.numeric(table(NSul.IV$CAE)), 2)
    nC47 <- cluster(NSul.IV, clustername=c("CAE"), size=2, pik=pik3, method="srswor")
    amostra47.IV <- getdata(NSul.IV,nC47)
```

```
    amostra47.IV$CAE<-factor(amostra47.IV$CAE)
    table(amostra47.IV$CAE)

    Nc47 <- as.numeric(nrow(amostra47.IV))
    Nc47.I <- by(amostra47.IV$VVN, amostra47.IV$CAE,length)
    sigmai.IC47 <- by(amostra47.IV$VVN, amostra47.IV$CAE, sd)
    np.c47 <- 466
    nip.c47 <- stratasamp(np.c47, as.numeric(Nc47.I), Sh=as.numeric(sigmai.IC47))
    nip.c47
    amostra47.IV$wi.C <- (1/as.numeric(amostra47.IV$Prob)) # peso dos grupos amostrados
    amostra47.IV$peso.ij                                                    <-
inclusionprobastrata(as.numeric(amostra47.IV$CAE),nip.c47[2,])
    amostra47.IV$wij.C <- 1/amostra47.IV$peso.ij # peso das amostras
    amostra47.IV$wij.g <- amostra47.IV$wi.C * amostra47.IV$wij.C # peso final

    s.Is <- strata(amostra47.IV,"CAE", size=nip.c47[2,], method="srswor")
    amostra47.F <- getdata(amostra47.IV, s.Is)

    amostrauniv <- rbind(amostra45.IV, amostra46.IV, amostra47.IV)
    amostrauniv$RegCAE<-factor(amostrauniv$RegCAE)
    table(amostrauniv$RegCAE)

    amostra.4 <- rbind(amostra45.F, amostra46.F, amostra47.F)
    amostra.4$RegCAE<-factor(amostra.4$RegCAE)
    table(amostra.4$wij.g)
    sum(amostra.4$wij.g)

    nIV.plan <- svydesign(ids = ~ CAE + Id, strata = ~Regiao, weights = ~wij.g, data =
amostra.4, nest = T)
    summary(nIV.plan)

    mediaT.IV <- svymean(~ VVN, nIV.plan)
    Est.mediaIV[i] <- coef(mediaT.IV)
    Var.TaylorIV[i] <- SE(mediaT.IV)^2
    SE.TaylorIV[i] <- SE(mediaT.IV)

    rnIV.planBoot <- as.svrepdesign(nIV.plan, type = "bootstrap", replicates = 1000)
    mediaB.IV <- svymean(~VVN, rnIV.planBoot)
    Var.BootstrapIV[i] <- SE(mediaB.IV)^2
    SE.BootstrapIV[i] <- SE(mediaB.IV)

    rnIV.planJkn <- as.svrepdesign(nIV.plan, type = "JKn")
    mediaJ.IV <- svymean(~VVN, rnIV.planJkn)
    Var.JKnIV[i] <- SE(mediaJ.IV)^2
    SE.JKnIV[i] <- SE(mediaJ.IV)

    print(c(i,Est.mediaIV[i]))
  }

  Esq.4.10MF  <-  data.frame(Est.mediaIV,  Var.TaylorIV,  SE.TaylorIV,  Var.BootstrapIV,
SE.BootstrapIV, Var.JKnIV, SE.JKnIV, Var.EstIV, SE.EstIV)

  write.table(Esq.4.10MF, file="Esq.4.10MF.csv", dec=",")
  Esq.4 <- read.table("Esq.4.10MF.csv", dec=",")

  M.MediaIV <- mean(Esq.4$Est.mediaIV)
  Var.MediaIV <- var(Esq.4$Est.mediaIV)
  SE.MediaIV <- sd(Esq.4$Est.mediaIV)

  SE.4             <-             cbind(Etapas3=0,             Taylor=Esq.4$SE.TaylorIV,
Bootstrap=Esq.4$SE.BootstrapIV,Jackknife=Esq.4$SE.JKnIV)
  medias.4<-cbind(mean(0),      mean(Esq.4$SE.TaylorIV),      mean(Esq.4$SE.BootstrapIV),
mean(Esq.4$SE.JKnIV))
  boxplot.matrix(SE.4, main ="Estimativas da Variancia", ylab = "erro padrão")
  abline(h=SE.MediaIV, v=0, lty= "dotted")
  points(medias.4[1,], pch =16, col ="red")

  M.Etapas2IV <- mean(Esq.4$Var.EstIV)
  Var.M.Etapas2  <- var(Esq.4$Var.EstIV)
  Var.M.Etapas2.p <- var(Esq.4$Var.EstIV)/1000000000
  CV.Etapas2 <- sqrt(Var.M.Etapas2)/M.Etapas2IV *100

  M.TaylorIV <- mean(Esq.4$Var.TaylorIV)
  Var.M.TaylorIV <- var(Esq.4$Var.TaylorIV)
  Var.M.TaylorIV.p <- var(Esq.4$Var.TaylorIV)/1000000000
  CV.TaylorIV <- sqrt(Var.M.TaylorIV)/M.TaylorIV*100
```

```
  M.BootstrapIV <- mean(Esq.4$Var.BootstrapIV)
  Var.M.BootstrapIV <- var(Esq.4$Var.BootstrapIV)
  Var.M.BootstrapIV.p <- var(Esq.4$Var.BootstrapIV)/1000000000
  CV.BootstrapIV <- sqrt(Var.M.BootstrapIV)/M.BootstrapIV*100

  M.JKnIV <- mean(Esq.4$Var.JKnIV)
  Var.M.JKnIV <- var(Esq.4$Var.JKnIV)
  Var.M.JKnIV.p <- var(Esq.4$Var.JKnIV)/1000000000
  CV.JKnIV <- sqrt(Var.M.JKnIV)/M.JKnIV*100

  Est.mediasIV <- c(M.MediaIV, Var.MediaIV, M.TaylorIV, Var.M.TaylorIV, M.BootstrapIV,
Var.M.BootstrapIV, M.JKnIV, Var.M.JKnIV, M.Etapas2IV, Var.M.Etapas2)
  Est.mediasIV

  Env.M.IV <- M.MediaIV - mean(universo$VVN)
  Env.M.IVp <- (Env.M.IV /mean(universo$VVN))*100

  Env.Etapas2.IV <- M.Etapas2IV - Var.MediaIV
  Env.Etapas2.IVp <- (Env.Etapas2.IV/Var.MediaIV)*100

  Env.T.IV <- M.TaylorIV - Var.MediaIV
  Env.T.IVp <- (Env.T.IV/Var.MediaIV)*100

  Env.B.IV <- M.BootstrapIV - Var.MediaIV
  Env.B.IVp <- (Env.B.IV /Var.MediaIV)*100

  Env.J.IV <- M.JKnIV - Var.MediaIV
  Env.J.IVp <- (Env.J.IV/Var.MediaIV)*100

  EQM.Etapas2.IV <- Var.M.Etapas2 + (Env.Etapas2.IV)^2
  EQM.T.IV <- Var.M.TaylorIV + (Env.T.IV)^2
  EQM.B.IV <- Var.M.BootstrapIV + (Env.B.IV)^2
  EQM.J.IV <- Var.M.JKnIV + (Env.J.IV)^2

  REQM.Etapas2.IV <- sqrt(EQM.Etapas2.IV)/1000000
  REQM.T.IV <- sqrt(EQM.T.IV)/1000000
  REQM.B.IV <- sqrt(EQM.B.IV)/1000000
  REQM.J.IV <- sqrt(EQM.J.IV)/1000000

  REQME.Est.IV <- sqrt(EQM.Etapas2.IV)/M.Etapas2IV*100
  REQME.T.IV <- sqrt(EQM.T.IV)/M.TaylorIV*100
  REQME.B.IV <- sqrt(EQM.B.IV)/M.BootstrapIV*100
  REQME.J.IV <- sqrt(EQM.J.IV)/M.JKnIV*100

  MerroIV <- zalfa * sqrt(92935042)

  Mediapo <- mean(universo$VVN)
  Media         <-          cbind(Del.1=Esq.1$Est.media,         Del.2=Esq.2$Est.mediaII,
Del.3=Esq.3$Est.mediaIII, Del.4=Esq.4$Est.mediaIV)
  medias             <-cbind(mean(Esq.1$Est.media),             mean(Esq.2$Est.mediaII),
mean(Esq.3$Est.mediaIII),     mean(Esq.4$Est.mediaIV))
  boxplot.matrix(Media, main ="Estimativa media", ylab = "Media")
  abline(h=Mediapo, v=0, lty= "dotted")
  points(medias[1,], pch =16, col ="red")
```