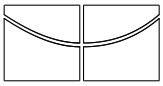

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Dissertação de Mestrado

Intervalo de Confiança para o parâmetro estimado pelo estimador de Horvitz-Thompson

por

Thuany de Aguiar Santos

Orientador: Prof. Dr. Alan Ricardo da Silva

Maio de 2016

Thuany de Aguiar Santos

Intervalo de Confiança para o parâmetro estimado pelo estimador de Horvitz-Thompson

Dissertação apresentada ao Departamento de Estatística do Instituto de Ciências Exatas da Universidade de Brasília como requisito parcial à obtenção do título de Mestre em Estatística.

Universidade de Brasília

Brasília, Maio de 2016

A Deus, que é a minha força e esperança em todos os momentos.

Agradecimentos

- Agradeço a Deus que me guiou em todos os momentos.
- Ao Professor **Alan Ricardo da Silva** pela dedicação e extremo comprometimento ao longo do trabalho, e acima de tudo pela generosidade em ensinar.
- A minha mãe **Alda** pelo apoio incondicional e pela motivação.
- Ao meu pai **Valdir** e minha amiga **Josane**, por todo o suporte e apoio.
- Ao meu marido **Gabriel**, pelo carinho, companherismo e amor.
- A minha irmã **Airina** pelo cuidado e amizade.
- Aos meus amigos de mestrado **Carolina Andrade**, **Raquel Araújo** e **Yuri Sampaio**, por estarem presente nas horas mais difíceis.
- Aos meus amigos e chefes do trabalho que tornaram este trabalho possível.

Sumário

Lista de Figuras

Lista de Tabelas

Resumo

Em um problema de amostragem sem reposição com probabilidades desiguais de seleção, Horvitz e Thompson (1952) propuseram um estimador não viesado capaz de estimar o total, a média de uma variável de interesse ou o tamanho populacional. Além dos estimadores pontuais, pode-se estimar intervalos de possíveis estimativas do parâmetro estudado por meio de estimadores intervalares. Portanto, este trabalho visa apresentar uma revisão da metodologia utilizada para calcular os Intervalos de Confiança (IC) baseados no estimador de Horvitz-Thompson. Verificou-se que o intervalo de confiança clássico, baseado na distribuição normal, é o mais utilizado na literatura. Este IC necessita da variância populacional para ser calculado, por isso, utilizou-se também o IC baseado na distribuição *t-student*, devido a variância ser estimada e comparou-se o desempenho de diferentes estimadores da variância apresentados na literatura com relação ao tamanho amostral. Verificou-se que nem sempre os IC baseados na distribuição normal atingiram a cobertura nominal e que os estimadores da variância, que independem da probabilidade de seleção conjunta, tiveram desempenho parecido ao estimador proposto por Yates e Grundy (1953) e Sen (1953), em populações pequenas.

Palavras Chave: *Amostragem, Estimador de Horvitz-Thompson, Intervalo de Confiança, Variância, Probabilidades desiguais de seleção, Captura-Recaptura.*

Abstract

In a problem of sampling without replacement with unequal probability of selection, Horvitz e Thompson (1952) have given an estimator without bias capable of estimates the total, the mean of a variable or the population size. Besides the point estimators, it is possible to estimate the ranges of possibles estimates of the parameter analyzed, by the intervals estimators. Therefore, this research has the main purpose to show an overview about the methodologies used to compute the confidence interval based on the Horviz-Thompson estimator. It was found that the classic confidence interval, based on the normal distribution, it is the most used in the literature. This interval needs that the population variance must be calculated, that is why, it was also used the confidence interval based on the t-student distribution, because the variance is estimated and then it was compared the performance of the different estimators of variance showed in the literature with relation to the sample size. It was concluded that not always the confidence interval based on the normal distribution reached the nominal cover and that the estimators of variance that are independent of the joint probability of selection had similar performance to the estimator given by Yates e Grundy (1953) and Sen (1953) in small populations.

key words: *sampling, Horvitz-Thompson estimator, Confidence bands, unequal probability sampling, Capture-Recapture.*

Introdução

Nem sempre é possível pesquisar todos os elementos da população quando se deseja obter informações populacionais. Uma solução passível de utilização é selecionar uma parte dos elementos da população (amostras) e basear-se no resultado dessas amostras para fazer inferências com base nas informações levantadas.

Quando se coletam amostras de uma população, o elemento pesquisado pode retornar ou não para a população. Se o elemento retorna, há a chance de ser selecionado novamente e a probabilidade de selecionar qualquer elemento posteriormente não é alterada, o que facilita os cálculos dos estimadores populacionais.

Entretanto, para o caso da amostragem sem reposição, as probabilidades de selecionar os elementos são alteradas a cada retirada, o que dificulta um pouco os cálculos, porém a variância do estimador do parâmetro é sempre menor do que no caso com reposição (Cochran, 1977).

Outra variação na metodologia de amostragem é que a probabilidade de seleção dos elementos da população pode ser igual ou desigual. Sabe-se que o uso de probabilidades desiguais na seleção da amostra pode trazer reduções consideráveis na variância do estimador do parâmetro comparado à amostragem com probabilidades iguais de seleção (Rao, 1963).

A partir das informações levantadas na amostra, os parâmetros populacionais de interesse podem ser estimados por um único valor (estimativas pontuais) ou podem ser estimados por um intervalo de valores (estimativas intervalares).

Existem diversos estimadores pontuais na literatura, estimadores do tipo razão, regressão, estimador de Hájek (1964), entre outros.

Um exemplo de estimador que produz estimativas pontuais para a média, para o total ou para o tamanho populacional, é conhecido como estimador de Horvitz-

Thompson (HT).

Suponha que o universo (U) seja formado por N elementos $(1, 2, ..., N)$. Horvitz e Thompson (1952) propuseram o estimador para o total populacional ($\tau = \sum_{i \in U} y_i$) de uma variável de interesse Y, para amostras (s) com n unidades amostrais, selecionadas sem reposição com probabilidades desiguais de seleção, dado por:

$$\hat{Y}_{HT} = \sum_{i \in s} \frac{y_i}{\pi_i}. \tag{1}$$

sendo que y_i representa a medida da variável de interesse do i-ésimo elemento e π_i é a probabilidade de inclusão do i-ésimo elemento na amostra.

Considere por exemplo uma população formada por três domicílios em que estão sendo observadas as variáveis renda bruta familiar mensal (em salários mínimos) e o número de trabalhadores em cada domicílio (Tabela 1).

Tabela 1: População de três domicílios

Variável	Valor
Domicílio	A B C
Renda	12 30 18
N^o Trabalhadores	1 3 2
Proporção do N^o de Trabalhadores	$\frac{1}{6}\frac{3}{6}\frac{2}{6}$

Com o intuito de estimar a renda total dos domicílios, selecionaram-se aleatoriamente duas unidades sem reposição de acordo com a variável suplementar proporção do número de trabalhadores no domicílio. Portanto, as amostras possíveis de serem geradas são $S = \{AB, AC, BA, BC, CA, CB\}$. Desse modo, a $P(AB) = P(A \text{ no } 1^o \text{ sorteio})P(B \text{ no } 2^o \text{ sorteio } | A \text{ no } 1^o \text{ sorteio}) = \frac{1}{6}\frac{3}{5} = \frac{1}{10}$, que é diferente de $P(BA) = \frac{3}{6}\frac{1}{3} = \frac{1}{6}$. Sendo assim é possível construir a tabela da distribuição amostral de $\hat{Y}_{HT}$ (Tabela 2) e a tabela das probabilidades de seleção dos domicílios (Tabela 3).

Tabela 2: Distribuição amostral de $\hat{Y}_{HT}$

s	AB	AC	BA	BC	CA	CB
$P(s)$	$\frac{6}{60}$	$\frac{4}{60}$	$\frac{10}{60}$	$\frac{20}{60}$	$\frac{5}{60}$	$\frac{15}{60}$
δ_A	1	1	1	0	1	0
δ_B	1	0	1	1	0	1
δ_C	0	1	0	1	1	1
$\bar{y}$	21	15	21	24	15	24
$\hat{Y}_{HT}$	$64,1$	$53,4$	$64,1$	$59,8$	$53,4$	$59,8$

Tabela 3: Probabilidades de seleção conjunta

Domicílios	A	B	C	π
A	0	0,27	0,15	0,42
B	0,27	0	0,58	0,85
C	0,15	0,58	0	0,73
π	0,42	0,85	0,73	2

É possível verificar que a esperança do estimador de HT é igual ao total populacional, neste caso, igual a $\tau = \sum_{i \in U} y_i = 60$. No exemplo apresentado, percebe-se que a probabilidade de selecionar o domicílio i, $i = A, B, C$, na primeira retirada é a proporção de trabalhadores em cada domicílio.

Horvitz e Thompson (1952) apresentaram o estimador para o total populacional, sua variância e o estimador desta variância, mas não fizeram menção ao Intervalo de Confiança (IC) para o total populacional.

Autores como Overton (1985), Yip et al. (1999), Stehman e Overton (1987) e Yip et al. (2001) basearam-se na distribuição assintótica do estimador de Horvitz-Thompson para construir intervalos de confiança da forma

$$\hat{Y}_{HT} \pm z_{\alpha/2}\sqrt{\hat{V}}, \tag{2}$$

em que $z_{\alpha/2}$ é o quantil da distribuição normal a um nível de confiança de $1 - \alpha$, $\hat{Y}_{HT}$ é a forma genérica para o estimador de Horvitz-Thompson para um parâmetro

populacional e $\hat{V}$ é a forma genérica para a variância estimada do estimador de Horvitz-Thompson.

Desse modo, para a construção deste intervalo de confiança necessita-se do estimador da variância de HT. Os dois estimadores da variância de HT mais conhecidos são: o estimador proposto por Horvitz e Thompson (1952) e o estimador proposto por Yates e Grundy (1953) e Sen (1953). Porém, o uso destes estimadores para amostragens complexas, tal como amostragem com probabilidades desiguais de seleção, é complicado, pois esses estimadores abarcam em sua fórmula a probabilidade de seleção conjunta dos elementos i e j (π_{ij}). Para o cálculo de π_{ij} é necessário o conhecimento de todas as probabilidades dos pares de unidades serem incluídas na amostra, o que gera enorme esforço computacional em algumas situações. Um caminho para contornar este problema, é a utilização de aproximações para a probabilidade de seleção conjunta e posteriormente estimação da variância utilizando os estimadores supracitados. Outra alternativa consiste em utilizar estimadores da variância que não necessitem da probabilidade conjunta de seleção para serem calculados Deville (1999); Brewer e Donadio (2003); Hájek (1964).

Portanto, este trabalho tem por objetivo primeiramente fazer uma revisão dos intervalos de confiança baseados no estimador de Horvitz-Thompson. Posteriormente, serão investigadas as coberturas dos intervalos de confiança baseados na distribuição assintótica do estimador de Horvitz-Thompson (distribuição normal) e na distribuição *t-student*, à medida que o tamanho amostral aumenta. Serão verificados também o comportamento dos diferentes estimadores da variância, que dependem ou não da probabilidade de seleção conjunta, nas simulações do trabalho para pequenas amostras.

Esta dissertação está organizada da seguinte forma: O Capítulo 1 apresenta o estimador de Horvitz-Thompson, o cálculo exato e aproximado das probabilidade de seleção, diferentes estimadores da variância. O Capítulo 2 faz uma revisão dos intervalos de confiança utilizados para o parâmetro estimado pelo estimador de Horvitz-Thompson. O Capítulo 3 apresenta a metodologia do trabalho e o Capítulo 4 descreve os resultados das simulações realizadas. Por fim, o Capítulo 5 apresenta as conclusões, limitações do trabalho e recomendações para trabalhos futuros.

Capítulo 1

Estimador de Horvitz-Thompson

Em um problema de amostragem aleatória sem reposição com probabilidades desiguais de seleção, Horvitz e Thompson (1952) propuseram uma teoria geral para estimar o total populacional e a média de uma variável de interesse. Sendo assim, o presente Capítulo tem por desiderato apresentar os conceitos do estimador de HT.

1.1 O Estimador de Horvitz-Thompson

Horvitz e Thompson (1952) demonstraram uma teoria geral para estimar o total populacional a partir de planos amostrais sem reposição, com probabilidade desigual de seleção. Sendo assim, seja o universo (U) formado por N elementos $(1, 2, ..., N)$ de onde se retira uma amostra (s), de tamanho n, sem reposição. O estimador linear não viesado do total populacional ($\tau = \sum_{i \in U} y_i$), de uma característica Y proposto por Horvitz e Thompson (1952) é dado por

$$\hat{Y}_{HT} = \sum_{i \in s} \frac{y_i}{\pi_i} = \sum_{i \in U} I_i \frac{y_i}{\pi_i}, \tag{1.1}$$

em que $I_i = \begin{cases} 1 & , \text{ se o elemento } i \text{ está incluído na amostra} \\ 0 & , \text{ se o elemento } i \text{ não está incluído na amostra} \end{cases}$
é a variável aleatória que especifica se o elemento i está ou não incluído na amostra, e y_i representa a medida da variável de interesse do i-ésimo elemento.

O estimador de HT poderá ser aplicado, por exemplo, para obter a renda total de uma população ou o peso total de animais em determinado local. Outro parâmetro

de singular interesse nos estudos em geral é a média. Quando se divide a Equação (1.1) por N, tem-se a equação do estimador de HT para a média

$$\hat{\mu}_{HT} = \frac{1}{N} \sum_{i \in s} \frac{y_i}{\pi_i} \tag{1.2}$$

em que $\hat{\mu}_{HT}$ permite estimar a renda média de uma população, o peso médio de animais ou a média de filhos das mulheres brasileiras.

Nas Equações (1.1) e (1.2) a probabilidade de inclusão na amostra do i-ésimo elemento é representada por π_i, em que $0 < \pi_i \leq 1$ (Horvitz e Thompson, 1952).

Seja s uma amostra de tamanho n extraída da população U com probabilidade $p(s)$, então a probabilidade de inclusão da unidade i na amostra é dada por (Fuller, 2009)

$$\pi_i = \sum_{s:i \in s} p(s), \tag{1.3}$$

e a probabilidade de inclusão conjunta das unidades i e j na amostra é dada por

$$\pi_{ij} = \sum_{s:(i,j) \in s} p(s). \tag{1.4}$$

Erdós e Rényi (1959) e Hájek (1964) provaram que o estimador de Horvitz-Thompson é assintoticamente Gaussiano para amostragem aleatória simples sem reposição, de acordo com Cardot e Josserand (2011). Esta prova assintótica é utilizada por diversos autores (Cardot e Josserand, 2011; Cardot et al., 2013; Berger e Torres, 2012) para construir intervalos de confiança baseados no estimador de Horvitz-Thompson.

1.2 Variância do Estimador de Horvitz-Thompson

Horvitz e Thompson (1952) apresentaram o estimador do total populacional juntamente com a variância deste estimador dada por:

$$V(\hat{Y}_{HT}) = \sum_{i=1}^{N} \frac{1 - \pi_i}{\pi_i} y_i^2 + \sum_{i=1}^{N} \sum_{j \neq i}^{N} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} y_i y_j. \tag{1.5}$$

Para amostras de tamanho fixo, a variância também pode ser expressa por Yates

e Grundy (1953); Sen (1953)

$$V(\hat{Y}_{HT}) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j\neq i}^{N}(\pi_i\pi_j - \pi_{ij})\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2. \tag{1.6}$$

As expressões são algebricamente idênticas, mas quando π_i ou π_{ij} são desiguais, substituindo as quantidades amostrais nas expressões da variância (1.5) e (1.6), elas conduzem a diferentes estimadores da variância (Lohr, 2010). Tem-se que π_{ij} é a probabilidade de seleção conjunta dos elementos i e j, em que $\sum_{i=1}^{N}\pi_i = n$ e $\sum_{i\neq j}^{N}\pi_{ij} = (n-1)\pi_i$ e a probabilidade de seleção pode ser obtida a partir de uma variável auxiliar.

(Horvitz e Thompson, 1952). O estimador da variância da forma (1.5), sugerido por Horvitz e Thompson (1952) é dado por

$$\hat{V}(\hat{Y}_{HT}) = \sum_{i\in s}\frac{1-\pi_i}{\pi_i^2}y_i^2 + \sum_{i\in s}\sum_{j\in s;j\neq i}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}. \tag{1.7}$$

Trabalhando com a segunda expressão da variância (1.6), sugerida por Yates e Grundy (1953) e Sen (1953), o estimador da variância é obtido a partir da equação

$$\hat{V}(\hat{Y}_{HT}) = \frac{1}{2}\sum_{i\in s}\sum_{j\in s;j\neq i}\frac{(\pi_i\pi_j - \pi_{ij})}{\pi_{ij}}\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2, \tag{1.8}$$

em que $\pi_{ij} > 0$ para todas as unidades da amostra.

Apesar de ambos os estimadores, (1.7) e (1.8), serem não viesados, eles são difíceis de serem utilizados na prática, pois para que seu cálculo seja realizado, necessita-se que a probabilidade de seleção conjunta seja conhecida. Este é um desafio para muitos estudos, especialmente para grandes tamanhos amostrais. Além do problema relatado, em algumas situações, esses estimadores ainda apresentam a desvantagem de produzir estimativas negativas para a variância (Lohr, 2010).

Uma alternativa, sugerida por Durbin (1953), para estimar a variância do estimador de HT e contornar as dificuldades dos estimadores (1.7) e (1.8), é supor que os elementos da amostra foram selecionados com reposição, e utilizar o estimador da variância com reposição dado por:

$$\hat{V}_D = \frac{n}{n-1} \sum_{i \in s} \left(\frac{y_i}{\pi_i} - \frac{\hat{Y}_{HT}}{n} \right)^2 . \tag{1.9}$$

Esse estimador (1.9) é sempre não negativo, e a probabilidade conjunta não necessita ser conhecida para que seu cálculo seja efetuado. No caso em que a amostragem sem reposição é mais eficiente do que a amostragem com reposição, espera-se que o estimador da variância no caso com reposição (1.9) superestime a variância. Porém, espera-se que o viés deste estimador seja pequeno se a fração $\frac{n}{N}$ for pequena (Lohr, 2010).

Rao (1963) utilizando uma aproximação assintótica, derivou uma expressão para o estimador da variância de Horvitz-Thompson independente de π_{ij} . Para amostras sem reposição e quando $\pi_i = np_i$ em que $p_i = \frac{x_i}{\sum_{i \in U} x_i}$ e sob a restrição de que $x_i \leq \frac{\sum_{i \in U} x_i}{n}$, sendo x_i uma medida relacionada ao indivíduo i, a fórmula de aproximação da probabilidade de seleção conjunta da ordem de N^{-3} ($O(N^{-3})$) é

$$\pi_{ij} = n(n-1)p_i p_j + n(n-1)(p_i^2 p_j + p_i p_j^2) - n(n-1)p_i p_j \sum_{t \in U} p_t^2. \tag{1.10}$$

Substituindo a aproximação de π_{ij} na equação de Yates e Grundy (1953) e Sen (1953) (1.8) retendo todos os termos para $O(N^1)$, tem-se o estimador de Rao para a variância de Horvitz-Thompson [1]

$$\hat{V}_R = \frac{1}{n-1} \sum_{i \in s; i<j} \left[1 - (\pi_i + \pi_j) + \frac{1}{n} \sum_{t=1}^{N} \pi_t^2 \right] \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 . \tag{1.11}$$

Existem outras alternativas para o estimador da variância que não necessitam da probabilidade conjunta para serem calculados como apresentado em Brewer e Donadio (2003). Uma dessas alternativas, para amostras de tamanho fixo, foi proposta por Hájek (1964)

$$\hat{V}_{HAJ} = \frac{n}{n-1} \sum_{i \in s} (1 - \pi_i) \left(\frac{y_i}{\pi_i} - A \right)^2 , \tag{1.12}$$

e outra por Deville (1999), que apresentou uma modificação do estimador (1.12)

[1] O somatório de π_t não aparece na Equação (23) do artigo de Rao (1963) que trata de uma amostra de n elementos, mas aparece na Equação (17) que trata de uma amostra de 2 elementos. Por isso se acredita que houve um esquecimento deste somatório por parte do autor.

proposto por Hájek (1964)

$$\hat{V}_{DEV} = \frac{1}{1-\sum_{i\in s} a_i^2} \sum_{i\in s}(1-\pi_i)\left(\frac{y_i}{\pi_i} - A\right)^2, \tag{1.13}$$

em que $A = \sum_{i\in s} a_i \frac{y_i}{\pi_i}$ e $a_i = (1-\pi_i)/\sum_{j\in s}(1-\pi_j)$ para as fórmulas (1.12) e (1.13).

Outra opção para amostras de tamanho fixo, do estimador da variância de Horvitz-Thompson, foi proposta por Brewer e Donadio (2003). Suponha que existam constantes c_i em que $\pi_{ij} \approx \pi_i\pi_j \frac{(c_i+c_j)}{2}$ em que existem três diferentes escolhas para o cálculo de c_i

$$c_i = \frac{(n-1)}{(n-\pi_i)}, \tag{1.14}$$

$$c_i = \frac{(n-1)}{n - \frac{1}{n}\sum_{j\in U}\pi_j^2}, \tag{1.15}$$

$$c_i = \frac{(n-1)}{\left(n - 2\pi_i + \frac{1}{n}\sum_{j\in U}\pi_j^2\right)}, \tag{1.16}$$

em que a terceira forma de calcular c_i é baseada na expressão assintótica para π_{ij} obtida por Hartley e Rao (1962).

Assim, o estimador da variância é dado por:

$$\hat{V}_{BD} = \sum_{i\in s}\left(\frac{1}{c_i} - \pi_i\right)\left(\frac{y_i}{\pi_i} - \frac{\hat{Y}_{HT}}{n}\right)^2. \tag{1.17}$$

As diferentes formas de estimar a variância do estimador de Horvitz-Thompson, (1.7), (1.8), (1.9), (1.12), (1.13) e (1.17), também serão alvo de estudo neste trabalho.

Capítulo 2

Intervalo de Confiança

A área de amostragem procura obter informações populacionais baseadas em resultados de amostras. Para isso, os parâmetros populacionais desconhecidos, como média, total de uma variável de interesse e tamanho populacional, são estimados por estimadores pontuais ou intervalares, calculados a partir de informações da amostra.

Os estimadores pontuais têm por resultado um único valor como estimativa. Já os estimadores intervalares têm por resultado do processo de estimação, intervalos de possíveis valores do parâmetro estudado. A estimação intervalar possui o diferencial de possibilitar o dimensionamento da magnitude do erro que se comete na estimação do parâmetro de interesse.

Um exemplo de estimador pontual presente na literatura, e que é foco deste trabalho, é o estimador de Horvitz-Thompson (apresentado na Seção 1.1). Com o intuito de estudar os estimadores intervalares baseados no estimador de HT, apresentados na literatura, fez-se uma revisão dos intervalos de confiança para os parâmetros, total, média e tamanho populacional, baseados no estimador de HT.

2.1 Intervalo de Confiança

O estimador de Horvitz-Thompson, apresentado no Capítulo 1, é um exemplo de estimador pontual, ou seja, fornece um único valor numérico como estimativa para o total populacional de uma variável de interesse. Em algumas situações pode ser apresentada uma estimativa intervalar para o parâmetro de interesse.

Definição (Mood et al. (1974)): Uma estimação intervalar do verdadeiro valor do parâmetro θ é qualquer par de funções, $L(x_1, ..., x_n)$ e $U(x_1, ...x_n)$, de uma amostra que satisfaça $L(\mathbf{x}) \leq U(\mathbf{x})$ para todo $\mathbf{x} \in X$, em que $\mathbf{x} = (x_1, ..., x_n)$. Se $\mathbf{X} = \mathbf{x}$ é observado, a inferência $L(\mathbf{x}) \leq \theta \leq U(\mathbf{x})$ é obtida. O intervalo aleatório $[L(\mathbf{X}), U(\mathbf{X})]$ é chamado de estimador intervalar.

Destaca-se que estimadores intervalares junto com uma medida de confiança são por vezes conhecidos como intervalos de confiança do tipo

$$P(L(x_1, ...x_n) \leq \theta \leq U(x_1, ..., x_n)) = 1 - \alpha, \tag{2.1}$$

em que $1 - \alpha$ é chamado de coeficiente de confiança.

Uma interpretação para intervalo de confiança de acordo com Magalhães e Lima (2013) é: se obtivermos várias amostras de mesmo tamanho e, para cada uma delas, calcularmos os correspondentes intervalos de confiança com coeficiente de confiança $1 - \alpha$, esperamos que a proporção de intervalos que contenham o valor de θ seja igual a $1 - \alpha$.

2.2 Intervalos de confiança baseados no estimador de Horvitz-Thompson

Stehman e Overton (1987) compararam o desempenho de dois estimadores da variância de Horvitz-Thompson (estimador proposto por Horvitz e Thompson (1952) (1.7) e o estimador proposto por Yates e Grundy (1953) e Sen (1953) (1.8)), para situações de amostragem aleatória sistemática, com probabilidades desiguais e tamanho amostral fixo (n). Como visto na Seção 1.2, ambos os estimadores necessitam da probabilidade de seleção conjunta, ocasionando um enorme esforço computacional para serem calculados. Por este motivo o artigo de Stehman e Overton (1987) testou duas fórmulas de aproximação de π_{ij} no cálculo dos estimadores (1.7) e (1.8):

- $\hat{V}^0_{HT}$ = estimador da variância, proposto por Horvitz-Thompson (1.7) , calculado utilizando π^0_{ij};
- $\hat{V}^0_{SYG}$ = estimador da variância, proposto por Yates-Grundy e Sen (1.8) , cal-

culado utilizando π_{ij}^0;

- $\hat{V}_{HT}^{hr}$ = estimador da variância, proposto por Horvitz-Thompson (1.7) , calculado utilizando π_{ij}^{hr};

- $\hat{V}_{SYG}^{hr}$ = estimador da variância, proposto por Yates-Grundy e Sen (1.8) , calculado utilizando π_{ij}^{hr}.

em que a fórmula para aproximação da probabilidade de seleção conjunta, proposta por Overton (1985) é

$$\pi_{ij}^0 = \frac{(n-1)\pi_i\pi_j}{n-\frac{1}{2}(\pi_i+\pi_j)} = \frac{2(n-1)\pi_i\pi_j}{2n-\pi_i\pi_j}, \tag{2.2}$$

e a outra fórmula aproximada da probabilidade de seleção conjunta é a fórmula truncada de Hartley e Rao (1962)

$$\pi_{ij}^{hr} = \frac{(n-1)\pi_i\pi_j}{\left[n-\pi_i-\pi_j+\sum_{k=1}^{N}\frac{\pi_k^2}{n}\right]}. \tag{2.3}$$

As amostras foram selecionadas de tal modo que a probabilidade de seleção da unidade i é proporcional a x_i, sendo $\pi_i = n\frac{x_i}{\sum_{i\in U} x_i}$.

Com o objetivo de verificar as propriedades dos estimadores da variância ($\hat{V}_{HT}^0$, $\hat{V}_{SYG}^0$, $\hat{V}_{HT}^{hr}$ e $\hat{V}_{SYG}^{hr}$), os autores utilizaram dois conjuntos de simulações. Para o primeiro conjunto de simulação, denominado de Grupo 1 (Tabela 2.1), eles examinaram dois conjuntos de dados (*stream survey*) e mais dois conjuntos de dados apresentados na literatura.

Tabela 2.1: Populações do Grupo 1

População	N	$cv(x)$	$cv(y)$	$\rho(x,y)$	$cv(y/x)$
Sales [1]	327	1,20	1,19	0,99	0,14
Paddy [2]	108	0,69	0,78	0,79	0,39
*Stream*1 [3]	100	0,92	0,72	0,86	0,71
*Stream*2 [3]	100	0,66	0,52	0,81	0,41

[1] Cumberland e Royall (1981), x = vendas brutas da empresa em 1974, y = vendas em 1975.

[2] Murthy (1967), x = área geográfica, y = área sob as nuvens no inverno.

[3] x = área da bacia direta, y = comprimento do alcance.

Para o segundo conjunto de simulações, denominado de Grupo 2 (Tabela 2.2), os autores geraram, a partir da estrutra dos dados *Stream*1, dados simulados com $N = 72$, em que B são chamadas populações limites e I populações interiores. As populações limites tem um alto coeficiente de variação $cv(\frac{y}{x})$ e as populações interiores tem um baixo $cv(\frac{y}{x})$.

Tabela 2.2: Coeficiente de variação ($cv(y/x)$) das populações do Grupo 2

População	$\rho(x,y) = 0,53$	$\rho(x,y) = 0,82$	$\rho(x,y) = 0,99$
B_1	$0,88$	$0,80$	$0,49$
B_2	$1,11$	$0,59$	$0,12$
B_3	$0,61$	$0,56$	$0,44$
I_1	$0,07$	$0,05$	$0,01$
I_2	$0,11$	$0,08$	$0,05$
I_3	$0,13$	$0,08$	$0,02$
I_4	$0,12$	$0,09$	$0,05$

Fonte: Stehman e Overton (1987)

Na Tabela 2.2, $\rho(x,y)$ é o coeficiente de correlação linear de Pearson entre as variáveis x e y.

Um dos critérios considerados, para comparação dos estimadores da variância foi

a cobertura dos intervalos de confiança baseados em cada estimador da variância, em que os intervalos são baseados na distribuição normal

$$\hat{Y} \pm z_{\alpha/2}\sqrt{\hat{V}}, \tag{2.4}$$

sendo $z_{\alpha/2}$ o quantil da distribuição normal padrão a um nível de 95% de confiança e $\hat{V}$ é uma forma genérica para qualquer um dos estimadores da variância ($\hat{V}^0_{HT}$, $\hat{V}^0_{SYG}$, $\hat{V}^{hr}_{HT}$ e $\hat{V}^{hr}_{SYG}$). Os resultados da cobertura dos IC alcançados para o Grupo 1 são apresentados na Tabela 2.3.

Tabela 2.3: Cobertura dos intervalo de confiança (95%), ($n = 16$) - Grupo 1

População	$\hat{V}^0_{HT}$	$\hat{V}^0_{SYG}$	$\hat{V}^{hr}_{HT}$	$\hat{V}^{hr}_{SYG}$
Sales	63	94	95	94
Paddy	92	93	94	93
Stream1	87	88	89	88
Stream2	87	87	89	87

Fonte: Stehman e Overton (1987)

Os resultados da cobertura dos IC alcançados para o Grupo 2 são apresentados nas Tabelas 2.4 e 2.5 .

Tabela 2.4: Cobertura dos intervalo de confiança (95%), ($n = 16$) para o Grupo 2 - Utilizando π^{hr}_{ij}

Pop.	$\rho = 0,53$		$\rho = 0,82$		$\rho = 0,99$	
	$\hat{V}^{hr}_{HT}$	$\hat{V}^{hr}_{SYG}$	$\hat{V}^{hr}_{HT}$	$\hat{V}^{hr}_{SYG}$	$\hat{V}^{hr}_{HT}$	$\hat{V}^{hr}_{SYG}$
B_1	87	85	87	85	90	89
B_2	90	90	92	93	59	93
B_3	93	93	93	93	93	93
I_1	76	93	62	94	49	93
I_2	84	94	75	94	63	93
I_3	86	93	69	93	52	93
I_4	88	93	82	93	70	93

Fonte: Stehman e Overton (1987)

Tabela 2.5: Cobertura dos intervalo de confiança (95%), ($n = 16$) para o Grupo 2 - Utilizando π_{ij}^0

Pop.	$\rho = 0,53$		$\rho = 0,82$		$\rho = 0,99$	
	$\hat{V}_{HT}^0$	$\hat{V}_{SYG}^0$	$\hat{V}_{HT}^0$	$\hat{V}_{SYG}^0$	$\hat{V}_{HT}^0$	$\hat{V}_{SYG}^0$
B_1	88	84	89	84	92	89
B_2	91	89	93	92	92	93
B_3	93	93	92	93	93	93
I_1	95	93	95	94	93	93
I_2	96	93	97	94	98	93
I_3	95	93	95	93	92	93
I_4	93	93	91	93	88	93

Fonte: Stehman e Overton (1987)

Percebe-se que os IC baseados na distribuição normal, não atingem sempre a cobertura nominal esperada de 95%. Em alguns situações os intervalos de confiança não atingem sequer 80% de cobertura.

Särndal et al. (1992) comparando o desempenho de diferentes estimadores da média populacional, realizaram simulações com a base de dados MU281, em que a variável de interesse y era a receita municipal recebida em 1985 e as variáveis auxiliares x_1 o número de assentos do partido conservador no conselho municipal e x_2 o número de assentos do partido social-democrata no conselho municipal. Os estimadores da média comparados foram:

- O estimador de Horvitz-Thompson (1.2);
- O estimador do tipo razão, $\hat{Y}_{ra1} = \sum_{i \in U} x_{1i} \frac{\sum_{i \in s} y_i}{\sum_{i \in s} x_{1i}}$;
- O estimador do tipo razão, $\hat{Y}_{ra2} = \sum_{i \in U} x_{2i} \frac{\sum_{i \in s} y_i}{\sum_{i \in s} x_{2i}}$;
- O estimador do tipo regressão $\hat{Y}_{rg1} = N[\bar{y}_s + \hat{B}_1(\bar{x}_{1U} - \bar{x}_{1s})]$; e
- O estimador do tipo regressão $\hat{Y}_{rg1} = N[\bar{y}_s + \hat{B}_2(\bar{x}_{2U} - \bar{x}_{2s})]$.

Os autores calcularam o IC para a média baseado na distribuição normal,

$$\hat{Y} \pm z_{\alpha/2}\sqrt{\hat{V}}, \tag{2.5}$$

sendo $z_{\alpha/2}$ o quantil da distribuição normal padrão a um nível de 95% de confiança e $\hat{Y}$ uma forma genérica para qualquer um dos estimadores da média e $\hat{V}$ o estimador da variância. Os IC baseados na normal ficaram abaixo da cobertura nominal esperada, sendo assim, os autores verificaram que obtiveram um ganho na cobertura dos intervalos quando utilizaram o IC baseado na distribuição *t-student*,

$$\hat{Y} \pm t_{(n-1),1-\alpha/2}\sqrt{\hat{V}}, \tag{2.6}$$

em que $t_{(n-1),1-\alpha/2}$ é o quantil da distribuição *t-student* com $n-1$ graus de liberdade a um nível de 95% de confiança.

Seguindo o trabalho de Särndal et al. (1992), Matei e Tillé (2005) utilizaram a cobertura dos intervalos de confiança baseado na distribuição *t-student* como uma das medidas para avaliar o desempenho de vinte estimadores da variância, inclusive dos mencionados no Capítulo 1, em amostragem com máxima entropia com probabilidades desiguais e amostras fixas. Os autores fizeram simulações com três grandes populações e obtiveram o resultado empírico de que o conhecimento da probabilidade de seleção conjunta não é necessário a fim de obter estimativas precisas da variância. Os IC baseados na distribuição *t-student* atingiram a cobertura nominal em quase todas as simulações de Matei e Tillé (2005).

Cardot e Josserand (2011), motivados pelo problema de estimar a curva média ($\mu_N(t)$) de consumo de energia elétrica de um grande número de consumidores em um intervalo fixo de tempo, propuseram o estimador de Horvitz-Thompson para a curva média baseado em observações discretas.

Tem-se que a curva média de consumo da população é dada por:

$$\mu_N(t) = \frac{1}{N}\sum_{i\in U} Y_i(t), t \in [0,T], \tag{2.7}$$

em que U é uma população finita de tamanho N, sendo possível associar uma função única de consumo $Y_i(t)$ a cada unidade i da população, para $t \in [0,T]$, com $T < \infty$. Para estimar a curva média de consumo de energia elétrica da população, Cardot e

Josserand (2011) definiram o estimador de Horvitz-Thompson para a curva média de consumo baseado em observações discretas

$$\hat{\mu}_N(t) = \frac{1}{N}\sum_{i\in s}\frac{\tilde{Y}_i(t)}{\pi_i}, t \in [0,T]. \tag{2.8}$$

Como a função de consumo $Y_i(t)$ não é medida a cada instante t em $[0,T]$ e sim em alguns pontos discretos deste intervalo, para cada unidade i da amostra s utilizou-se uma interpolação para estimar a curva de consumo em cada instante t

$$\tilde{Y}_i(t) = Y_i(t_k) + \frac{Y_i(t_{k+1}) - Y_i(t_k)}{t_{k+1} - t_k}(t - t_k), t \in [t_k, t_{k+1}]. \tag{2.9}$$

A função de covariância de $\hat{\mu}_N(t)$, denotada por $\gamma(s,t) = cov(\hat{\mu}_N(s), \hat{\mu}_N(t))$ é estimada por

$$\hat{\gamma}(s,t) = \frac{1}{N^2}\sum_{i\in s}\sum_{j\in s}\frac{\tilde{Y}_i(s)}{\pi_i}\frac{\tilde{Y}_j(t)}{\pi_j}\frac{\triangle_{ij}}{\pi_{ij}}, \tag{2.10}$$

em que $\triangle_{ij} = \pi_{ij} - \pi_i\pi_j$ se $i \neq j$ e $\triangle_{ii} = \pi_i(1-\pi_i)$ para todo $(s,t) \in [0,T]\times[0,T]$.

Com o desiderato de derivar intervalos de confiança para a curva média de consumo, os autores consideraram a estrutura de superpopulações, que pode ser encontrada com mais detalhes em Isaki e Fuller (1982) e Fuller (2009), em que se tem uma sequência crescente de populações U com tamanho N tendendo a infinito e uma sequência de amostras s de tamanho fixo n retiradas de U. Sob o ponto de vista assintótico foram construídos intervalos de confiança para a curva média de consumo, da forma

$$P\left[|\hat{\mu}_N(t) - \mu_N(t)| < \left(2ln\left(\frac{2}{\alpha}\right)\hat{\gamma}(t,t)\right)^{1/2}, t \in [0,T]\right] \simeq 1-\alpha \tag{2.11}$$

com nível de confiança de $1-\alpha$.

Cardot et al. (2013), ainda trabalhando no problema de estimar a curva média de consumo baseado em observações discretas (2.7), ao invés de utilizar uma interpolação para estimar a trajetória para cada instante t, como feito anteriormente, propuseram suavizar as trajetórias com polinômios locais e, em seguida, estimar a curva média de consumo com um estimador do tipo Horvitz-Thompson (2.8). Sendo assim, para

todas as unidades $i \in s$, eles obtiveram

$$X_{ki} = Y_i(t_k) + \epsilon_{ki} \tag{2.12}$$

em que as curvas foram medidas em pontos discretos $0 = t_1 < t_2 < ... < t_d = T$ sendo $1 < k < d$, em que d é a quantidade de pontos de medição da curva de consumo e ϵ_{ki} são erros aleatórios. Cardot et al. (2013) utilizaram uma suavização linear na curva de consumo baseada em dados discretos,

$$\tilde{Y}_i(t) = \sum_{i=1}^{d} W_k(t) X_{ki} \tag{2.13}$$

em que $W_k(t)$ é a função peso, expressa por

$$W_k(t) = \frac{(1/dh)[q_2(t) - (t_k - t)q_1(t)]M\left(\frac{(t_k - t)}{h}\right)}{q_2(t)q_0(t) - q_1^2(t)}, \tag{2.14}$$

e M é uma função kernel não negativa, $h > 0$ é a largura da banda ou *bandwidth*, e

$$q_l(x) = \frac{1}{dh}\sum_{j=1}^{d}(t_k - t)^l M\left(\frac{t_k - t}{h}\right), l = 0, 1, 2. \tag{2.15}$$

Sob a mesma estrutura de superpopulação utilizada em Cardot e Josserand (2011), os autores propuseram o intervalo de confiança assintótico para a curva média de consumo ($\mu_N(t)$) dado por:

$$\left\{\left[\hat{\mu}_N(t) \pm c\frac{\sqrt{\hat{\gamma}_N(t,t)}}{\sqrt{N}}\right], t \in [0, T]\right\} \tag{2.16}$$

em que c é uma constante que aproximadamente satisfaz

$$P(|G(t)| \leq c\sqrt{\gamma(t,t)}, \forall t \in [0, T]) = 1 - \alpha \tag{2.17}$$

Tem-se que $G(t)$ é um processo gaussiano com média zero e função covariância $\gamma(t, t)$. A constante c pode ser obtida por um método de simulação, em que para todo $c > 0$ a medida que $N \to \infty$, $P(|\hat{G}_N(t)| \leq c\sqrt{\hat{\gamma}_N(t,t)}, \forall t \in [0, T]|\hat{\gamma}_N)$ converge em

probabilidade para $P(|G(t)| \leq c\sqrt{\gamma(t,t)}, \forall t \in [0,T])$. Sendo $\hat{G}_N(t)$ uma sequência de processos gaussianos para cada N, com média zero e covariância $\hat{\gamma}_N$ definida em (2.10). A constante c é obtida como sendo o quantil de ordem $(1-\alpha)$ da distribuição

$$P\left(|\hat{G}_N(t)| \leq c\sqrt{\gamma_N(\hat{t},t)}, \forall t \in [0,T]|\hat{\gamma_N}\right) = 1-\alpha \tag{2.18}$$

Berger e Torres (2012) propuseram uma nova abordagem de verossimilhança empírica que pode ser usada para construir intervalos de confiança sob amostragem, sem reposição, com probabilidades desiguais de seleção. Suponha uma população finita U formada por N unidades, com N não necessariamente conhecido. Suponha que o parâmetro de interesse θ_0 é a solução da equação:

$$G(\theta) = 0, \text{ com } G(\theta) = \sum_{i \in U} g_i(\theta), \tag{2.19}$$

em que $g_i(\theta)$ é uma função de θ e das características da unidade i. Deseja-se estimar θ a partir de uma amostra s de tamanho fixo n selecionada sem reposição, em um único estágio com probabilidade desigual de seleção. O estimador de máxima verossimilhança empírica $\hat{\theta}$ de θ_0 é a solução da equação

$$\hat{G}(\theta) = 0, \text{ com } \hat{G}(\theta) = \sum_{i \in s} \hat{m}_i g_i(\theta), \tag{2.20}$$

em que $\hat{m}_i$ são os valores que maximizam o logaritmo da função de máxima verossimilhança

$$l(m) = log\left(\Pi_{i=1}^{n}\left(\frac{\pi_i m_i}{\sum_{j=1}^{n}\pi_j m_j}\right)\right) \tag{2.21}$$

sendo $m_i = NP_i$ a unidade de massa da unidade i da população e P_i a massa de probabilidade da unidade i. Quando $\hat{m}_i = \pi_i^{-1}$ e $g_i(\theta) = y_i - n^{-1}\theta\pi_i$, o estimador de máxima verossimilhança empírica $\hat{\theta}$ é o estimador de Horvitz-Thompson (1.1) do total, sendo y_i o conjunto de valores da variável de interesse relacionados ao indivíduo i.

Para derivar intervalos de confiança baseados na abordagem de verossimilhança

empírica, os autores assumiram a seguinte suposição

$$\hat{G}(\theta)V(\hat{G}(\theta))^{-1/2} \rightarrow N(0,1). \tag{2.22}$$

Sendo assim, a um nível $1-\alpha$ de confiança, os IC para o parâmetro θ_0 são da forma

$$[min\{\theta|\hat{r}(\theta) \leq \chi^2_1(\alpha)\}; max\{\theta|\hat{r}(\theta) \leq \chi^2_1(\alpha)\}] \tag{2.23}$$

em que $\chi^2_1(\alpha)$ é o quantil superior da distribuição qui-quadrado com 1 grau de liberdade, $\hat{r}(\theta) = 2\{l(\hat{m}) - l(\hat{m^*}, \theta)\}$ é a razão de máxima verossimilhança, uma função convexa, não simétrica, com o mínimo quando θ é o estimador de máxima verossimilhança empírica, e $l(\hat{m^*}, \theta)$ é o valor máximo do logaritmo da função de máxima verossimilhança.

A função proposta por Berger e Torres (2012) pode ser utilizada para construir intervalos de confiança para parâmetros populacionais como média, coeficiente de regressão, quantis e indicadores de pobreza. Berger e Torres (2014) aplicaram a abordagem da máxima verossimilhança empírica para construir intervalos de confiança para uma medida de pobreza. Para essa situação específica o estimador de máxima verossimilhança empírica reduziu-se ao estimador de Hájek (1964).

Berger e De La Riva Torres (2016) extenderam o trabalho de Berger e Torres (2012) para amostras complexas, gerando ainda um código no *software* R para a construção dos intervalos (Berger, 2015).

Capítulo 3

Material e Métodos

Este capítulo tem por intuito apresentar os métodos e os bancos de dados que serão utilizados neste trabalho para verificar o desempenho dos estimadores da variância de Horvitz-Thompson e comparar os intervalos de confiança para o total populacional baseados no estimador de Horvitz-Thompson.

3.1 Material

Neste estudo serão utilizados dados de populações apresentados ao longo da literatura para as simulações do desempenho dos estimadores das variâncias de Horvitz-Thompson e para calcular a cobertura dos IC estudados neste trabalho. A Tabela 3.1 apresenta a descrição das populações que serão utilizadas, em que $CV(x)$ e $CV(y)$ são, respectivamente, os coeficientes de variação das variáveis x e y, e ρ o coeficiente de correlação de Pearson entre x e y.

Tabela 3.1: Descrição das populações

Pop.	Fonte	y	x	N	$CV(y)$	$CV(x)$	ρ
1	Cochran (1963, pg 325)	Quantidade de pessoas por bloco	Quantidade de quartos por bloco	10	0,15	0,14	0,65
2	Horvitz e Thompson (1952, pg 682)	Quantidade de famílias	Estimativa da quantidade de famílias	20	0,44	0,40	0,87
3	Sukhatme (1954, pg 279) Círculos 1 -20	Área plantada de trigo	Quantidade de aldeias	20	0,63	0,50	0,59

A justificativa para a utilização dessas populações se deve ao fato de que Brewer e Donadio (2003) utilizaram-as para verificar o desempenho dos estimadores da variância quando $n = 2$, entretanto, os autores não contruíram IC para o total populacional utilizando estes estimadores da variância. Aém de verificar o desempenho da variância para $n > 2$ também serão construídos intervalos de confiança para o total populacional.

3.2 Métodos

Com o desiderato de comparar algumas metodologias de obtenção dos intervalos de confiança estudados, serão realizadas simulações baseadas em 1.000 replicações do procedimento amostral, em que serão calculadas as coberturas dos três diferentes intervalos de confiança para o total populacional, sendo que as unidades foram selecionadas com probabilidades proporcionais a variável auxiliar x_i, essa metodologia de seleção foi mencionada por Yates e Grundy (1953).

- IC baseado na distribuição normal, $\hat{Y}_{HT} \pm z_{\alpha/2}\sqrt{\hat{V}}$, utilizados por trabalhos estudados no Capítulo 2;
- IC baseado na distribuição t de *Student* , $\hat{Y}_{HT} \pm t_{\alpha/2}\sqrt{\hat{V}}$, devido ao fato de se utilizar estimativas das variâncias; e

- Intervalo de confiança *Bootstrap*.

Primeiramente, decidiu-se conduzir alguns estudos empíricos com o intuito de avaliar o desempenho dos estimadores da variância apresentados no Capítulo 1. Serão comparadas as probabilidades de seleção calculadas de forma exata e aproximada no cálculo do estimador da variância (1.8). No caso exato, seja s uma amostra de tamanho n da população U com probabilidade $p(s)$, então a probabilidade de inclusão da unidade i na amostra é dada por $\pi_i = \sum_{s:i\in s} p(s)$ e a probabilidade de inclusão conjunta das unidades i e j na amostra é dada por $\pi_{ij} = \sum_{s:(i,j)\in s} p(s)$. Para o propósito do estudo, a probabilidade de seleção conjunta (π_{ij}) será obtida pela aproximação (1.10) e a probabilidade de seleção (π_i) é proporcional a x, onde, $\pi_i = n\frac{x_i}{\sum_{i\in U} x_i}$, sob a restrição de que $x_i \leq \frac{\sum_{i\in U} x_i}{n}$.

A Tabela 3.2 apresenta a nomeclatura dos estimadores da variância utilizados nas simulações deste trabalho.

Tabela 3.2: Nomeclatura dos estimadores da variância

Nomeclatura do estimador	Descrição
varBD1	Estimador da variância (1.17) com a constante c_i calculada pela equção (1.14)
varBD2	Estimador da variância (1.17) com a constante c_i calculada pela equção (1.15)
varBD3	Estimador da variância (1.17) com a constante c_i calculada pela equção (1.16)
varD	Estimador da variância (1.9)
varDEV	Estimador da variância (1.13)
varHAJ	Estimador da variância (1.12)
varR	Estimador da variância (1.11)
varht	Estimador da variância (1.7)
varsyg	Estimador da variância (1.8)

Os resultados expressos no Capítulo 4 foram obtidos por meio do *software* SAS 9.2 utilizando o algoritmo desenvolvido por Nascimento e Silva (2013).

Capítulo 4

Resultados

Este capítulo apresenta os resultados das simulações dos estimadores da variância de Horvitz-Thompson e dos intervalos de confiança para o total populacional baseados no estimador HT.

Não foi calculada a cobertura do IC baseado no estimador $\hat{V}_{HT}$ (1.7) devido a este apresentar estimativas negativas.

4.1 Verificação da qualidade da aproximação da probabilidade de inclusão

O cálculo da probabilidade de seleção conjunta dos elementos i e j (π_{ij}) necessita do conhecimento de todas as probabilidades dos pares de unidades estarem incluídos na amostra. Para amostras com probabilidades desiguais de seleção, que é o caso das simulações deste Capítulo, o cálculo destas probabilidades ocasiona um enorme esforço computacional.

Primeiramente, procurou-se verificar o comportamento da fórmula de aproximação do π_{ij}. Desta forma, comparou-se a distribuição das estimativas do estimador da variância (1.8) quando calculado com as probabilidades de seleção exatas $\pi_{ij} = \sum_{s:(i,j)\in s} p(s)$, e quando calculado a apartir da aproximação (1.10), em todas as três populações estudadas.

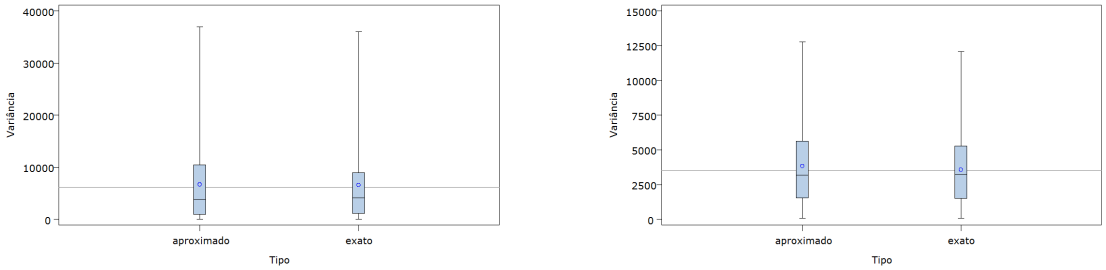

Figura 4.1: Boxplot das estimativas da variância (1.8), utilizando o cálculo exato (1.4) e a fórmula (1.10) - População 1 (esquerda - $n = 2$, direita - $n = 3$)

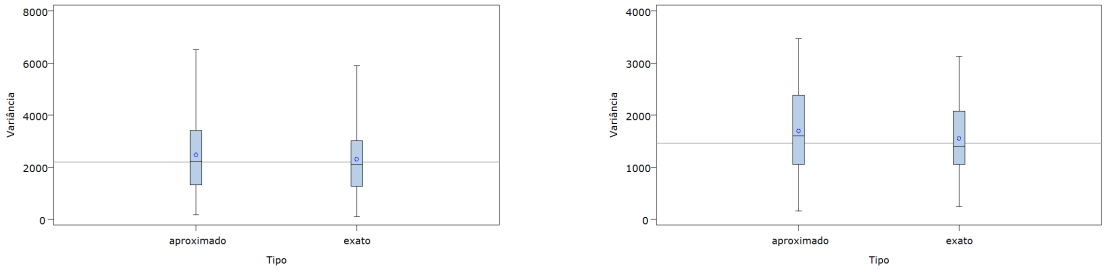

Figura 4.2: Boxplot das estimativas da variância (1.8), utilizando o cálculo exato (1.4) e a fórmula (1.10) - População 1 (esquerda - $n = 4$, direita - $n = 5$)

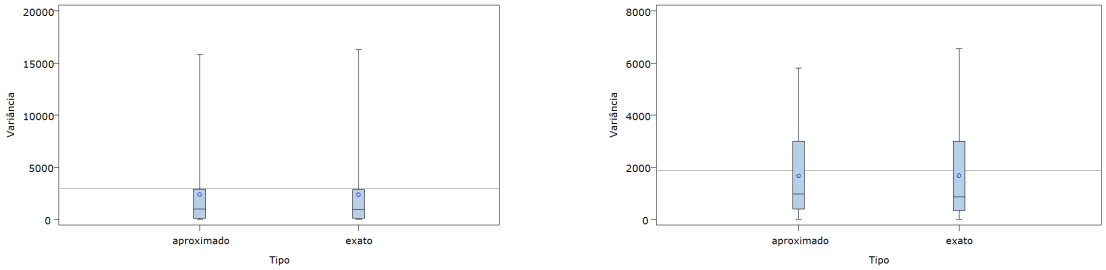

Figura 4.3: Boxplot das estimativas da variância (1.8), utilizando o cálculo exato (1.4) e a fórmula (1.10) - População 2 (esquerda - $n = 2$, direita - $n = 3$)

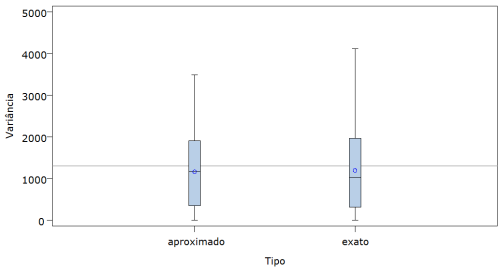

Figura 4.4: Boxplot das estimativas da variância (1.8), utilizando o cálculo exato (1.4) e a fórmula (1.10) - População 2 ($n = 4$)

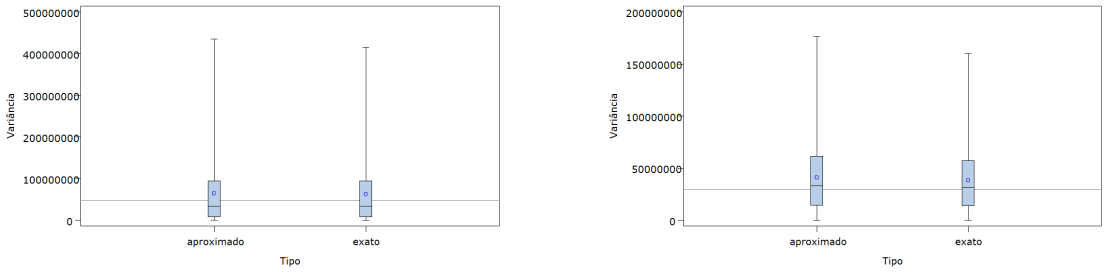

Figura 4.5: Boxplot das estimativas da variância (1.8), utilizando o cálculo exato (1.4) e a fórmula (1.10) - População 3 (esquerda - $n = 2$, direita - $n = 3$)

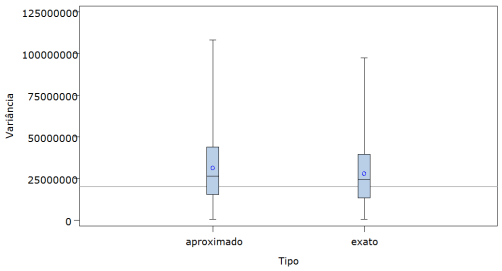

Figura 4.6: Boxplot das estimativas da variância (1.8), utilizando o cálculo exato (1.4) e a fórmula (1.10) - População 3 ($n = 4$)

Nas Figuras 4.1 a 4.6, é possível identificar que quando $n = 2$, em todas as populações, a distribuição das estimativas do estimador da variância são praticamente iguais. Existe uma maior diferença quando $n > 2$, mesmo assim, a aproximação aparenta ser adequada. Quando se compara a distribuição das estimativas, percebe-se que a simetria e/ ou assimetria são iguais em todas as Figuras, quando $n > 2$ a alteração que se percebe é na amplitude das estimativas, que geralmente é menor quando a variância (1.8) é calculada utilizando a probabilidade de seleção exata.

Dessa forma, como a aproximação (1.8) requer menos esforço computacional e a partir dos resultados mostrados acima, as próximas seções utilizarão tal aproximação para o cálculo das probabilidades de seleção.

4.2 População 1

Primeiramente, procura-se comparar a distribuição das estimativas dos diferentes estimadores da variância de Horvitz-Thompson. Destaca-se que nas Figuras 4.7 e 4.8 o verdadeiro valor da variância do estimador de Horvitz-Thompson é representado por uma linha horizontal nos gráficos.

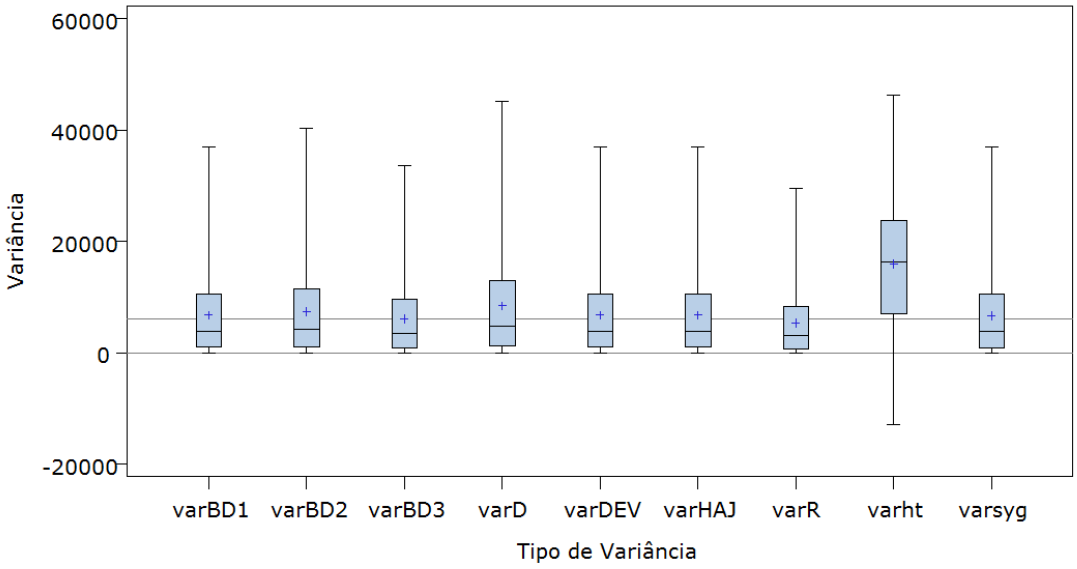

Figura 4.7: Boxplot das estimativas da variância de Horvitz-Thompson - $n = 2$

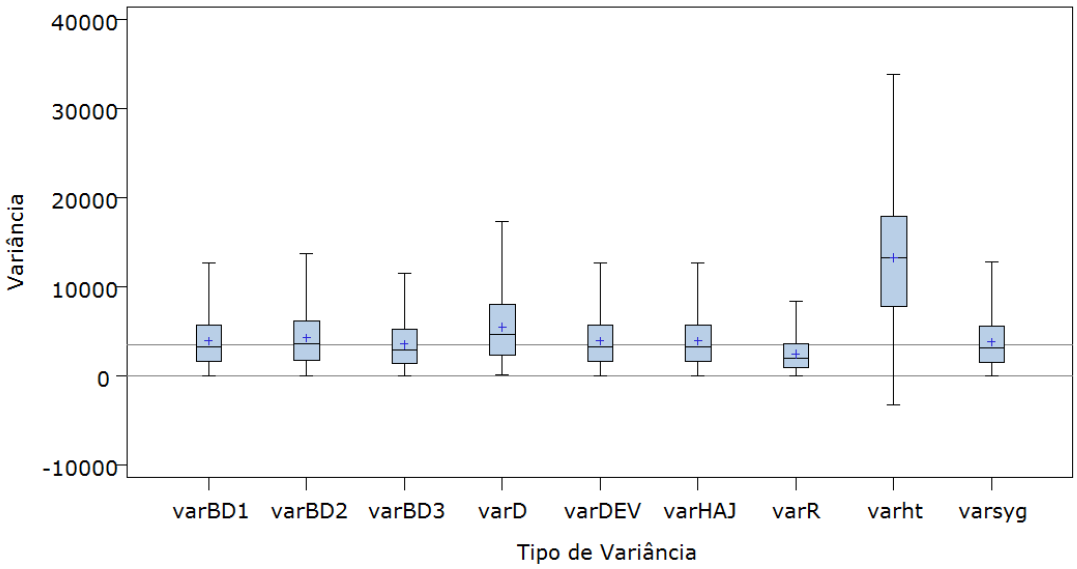

Figura 4.8: Boxplot das estimativas da variância de Horvitz-Thompson - $n = 3$

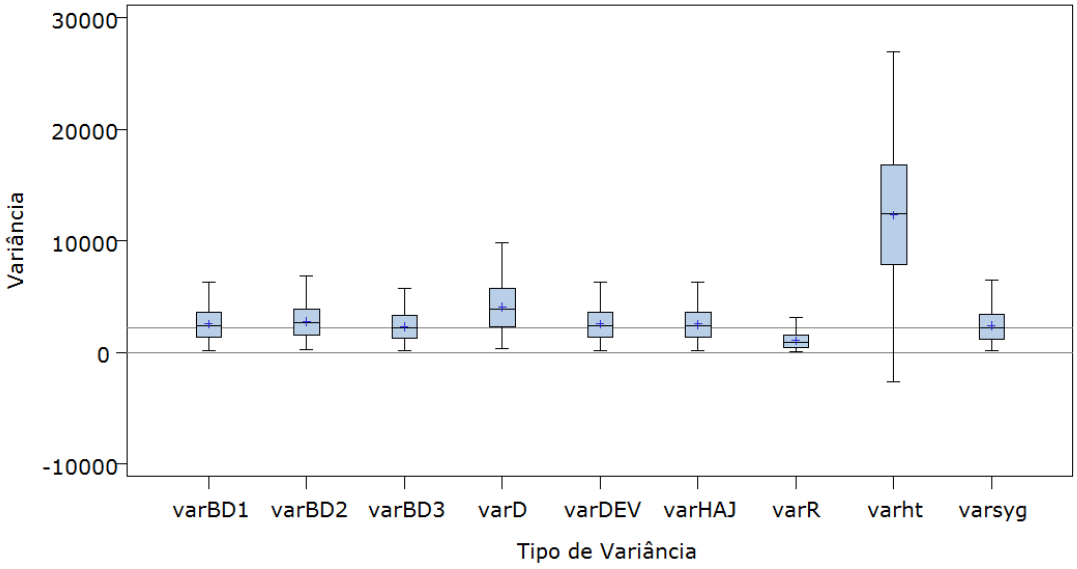

Figura 4.9: Boxplot das estimativas da variância de Horvitz-Thompson - $n = 4$

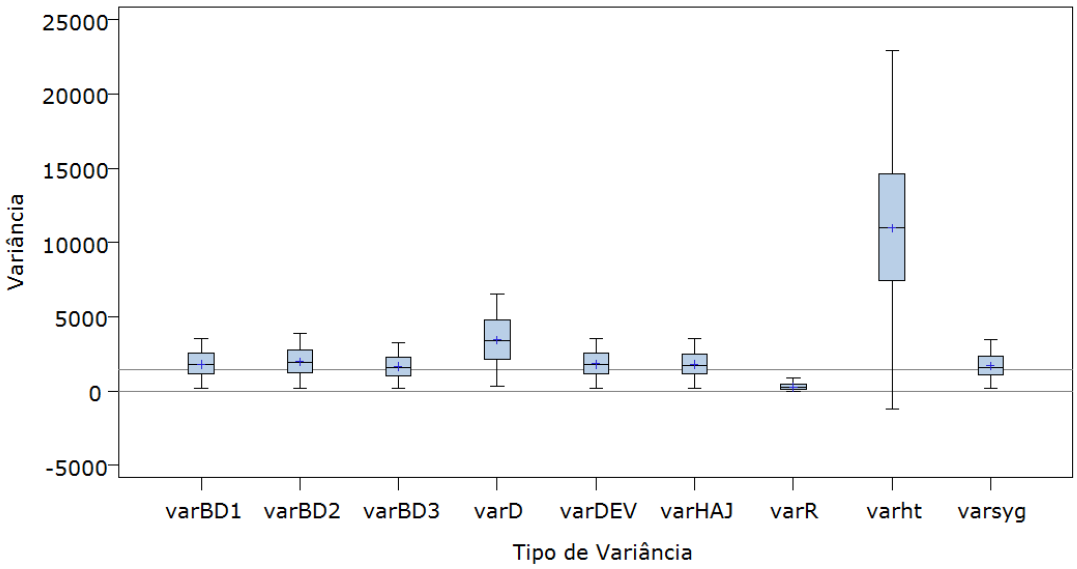

Figura 4.10: Boxplot das estimativas da variância de Horvitz-Thompson - $n = 5$

As Figuras 4.7, 4.8, 4.9 e 4.10 apresentam a distribuição das estimativas da variância produzidas por cada um dos nove estimadores analisados. Primeiramente, verifica-se que o estimador com menor amplitude interquartílica é $\hat{V}_R$ (1.11), porém percebe-se que este estimador subestima o verdadeiro valor da variância do estimador de HT (representado pela linha horizontal). Observa-se também que o estimador $\hat{V}_{HT}$ (1.7) apresenta a maior amplitude interquartílica e produz estimativas negativas da variância. Sendo assim, este estimador não foi utilizado para construir os intervalos de confiança apresentados posteriormente.

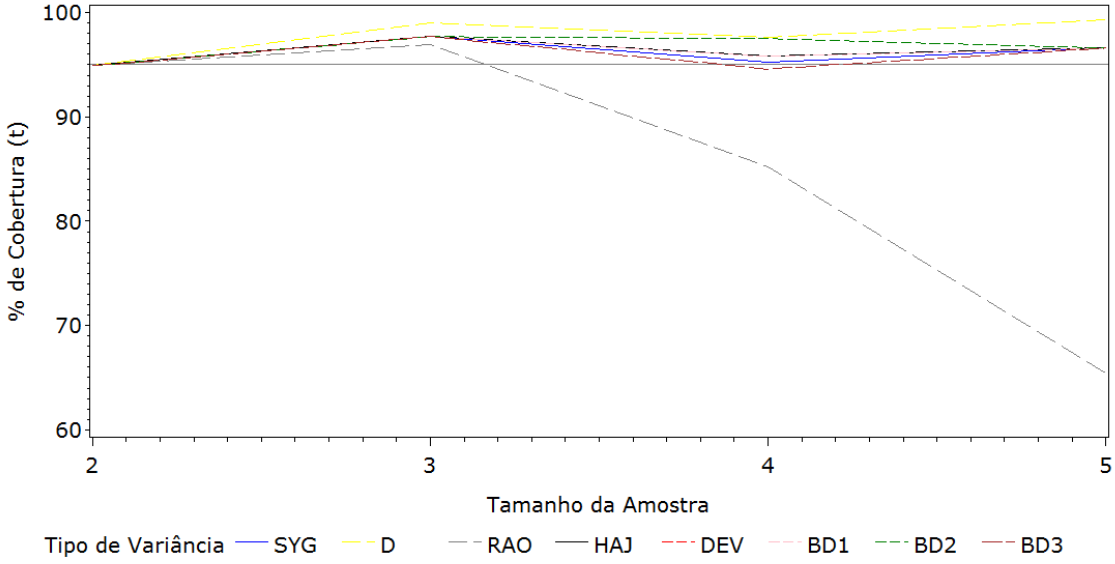

Figura 4.11: Cobertura dos IC de 95% para o total populacional, baseados na distribuição *t-student*

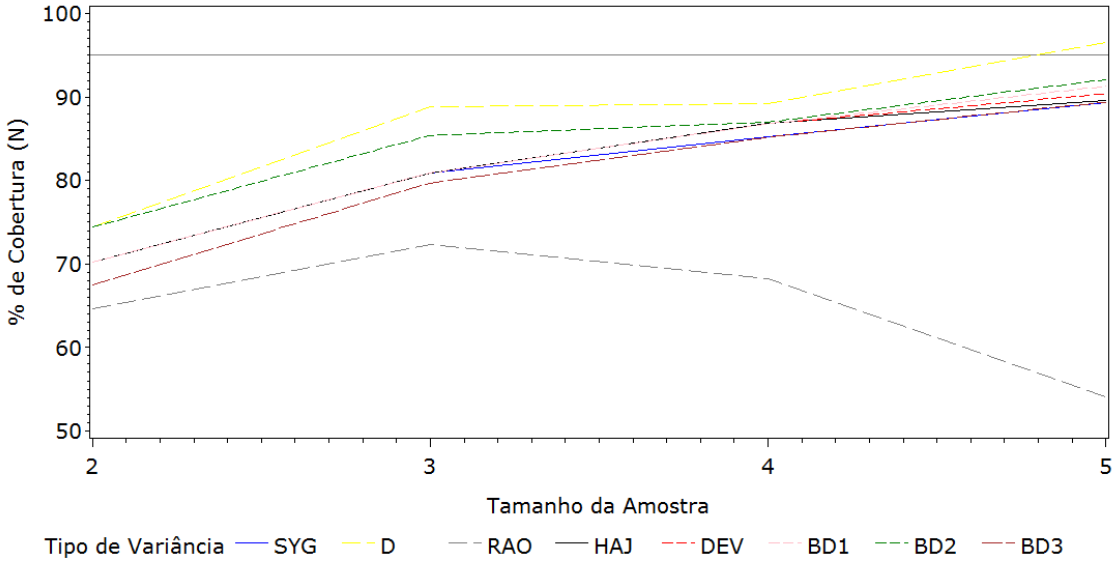

Figura 4.12: Cobertura dos IC de 95% para o total populacional, baseados na distribuição normal

As Figuras 4.11 e 4.12 apresentam, respectivamente, as coberturas dos IC baseados na distribuição *t-student* e na distribuição normal, à medida que o tamanho amostral aumenta. Comparando as duas Figuras, é possível verificar que à medida

que o tamanho amostral aumenta as coberturas dos IC também aumentam, exceto as coberturas dos IC baseados no estimador $\hat{V}_R$, o que pode ser explicado pois este estimador subestima a variância, como mostram as Figuras 4.7, 4.8, 4.9 e 4.10. Percebe-se ainda, que a cobertura dos IC baseados na distribuição normal é bem menor do que a cobertura dos IC baseados na distribuição *t-student*, para todos os IC simulados.

Nos intervalos baseados na distribuição *t-student*, as coberturas dos IC são bem parecidas, mantendo-se acima da cobertura nominal em quase todos os casos, exceto no caso do IC baseado no estimador $\hat{V}_R$.

As coberturas dos intevalos baseados na distribuição normal aumentam junto com o tamanho amostral, para este caso, o IC que utiliza o estimador $\hat{V}_D$ tem cobertura superior aos demais. Novamente, o IC baseado no estimador $\hat{V}_R$ diminue à medida que o tamanho da amostra aumenta.

Deste modo, para a População 1, quando se compara as coberturas dos IC dos estimadores que independem da probabilidade de seleção conjunta e as coberturas dos IC do estimador $\hat{V}_{SYG}$, que depende da probabilidade conjunta de seleção, estas coberturas tem desempenho parecido para todos os tamanhos amostrais simulados, exceto o estimador de $\hat{V}_R$, que tem cobertura bem abaixo da cobertura nominal de 95% analisada.

Quanto a cobertura dos intervalos de confiança *Bootstrap* verificou-se que para todos os tamanhos amostrais as coberturas foram superiores a $99,5\%$. Portanto, resolveu-se avaliar a amplitude destes intervalos frente às amplitudes dos IC das Figuras 4.11 e 4.12 .

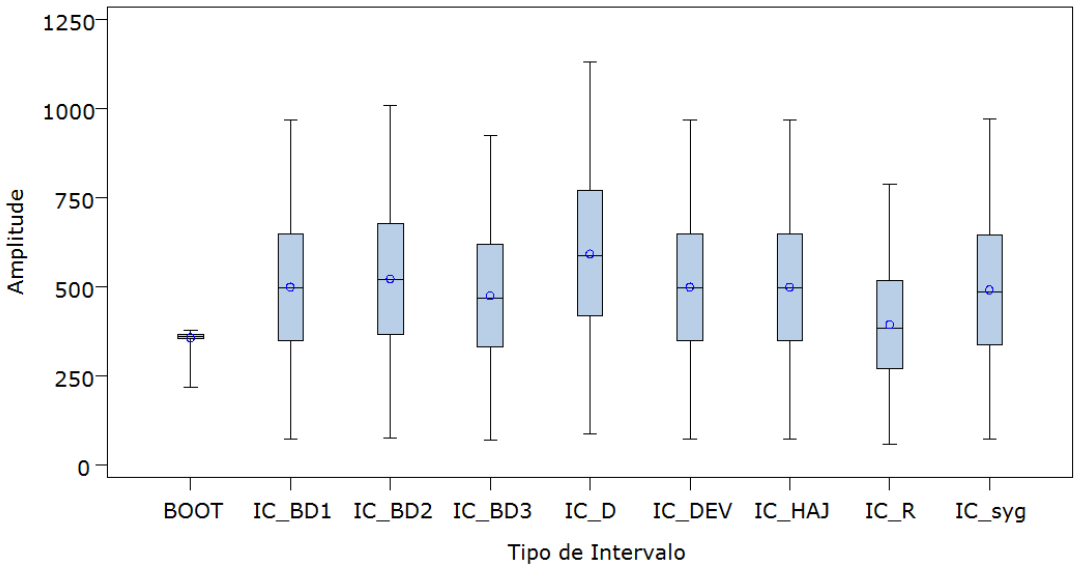

Figura 4.13: Distribuição das amplitudes dos IC baseados na distribuição *t-student* e dos IC *Bootstrap* - $n = 3$

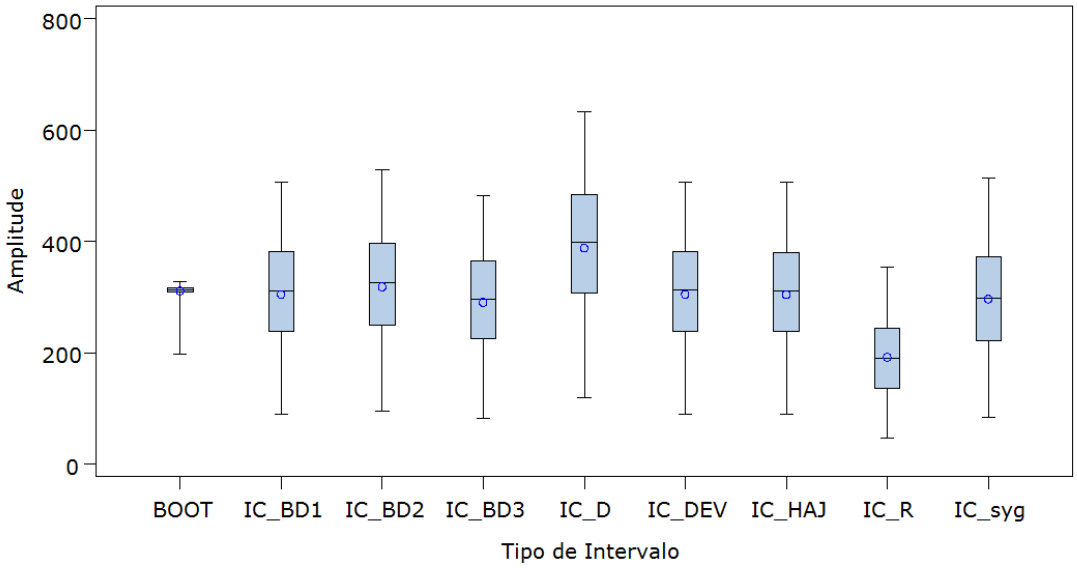

Figura 4.14: Distribuição das amplitudes dos IC baseados na distribuição *t-student* e dos IC *Bootstrap* - $n = 4$

A amplitude dos IC *Bootstrap* são menores ou iguais às amplitudes dos IC baseados na distribuição *t-student*. Portanto, os intervalos *Bootstrap* têm cobertura um pouco maior e amplitude menor do que os IC *t-student*. Verifica-se também que os

IC que utilizam o estimador $\hat{V}_D$ apresentaram amplitudes maiores do que todos os outros intervalos comparados nas Figuras 4.13 e 4.14.

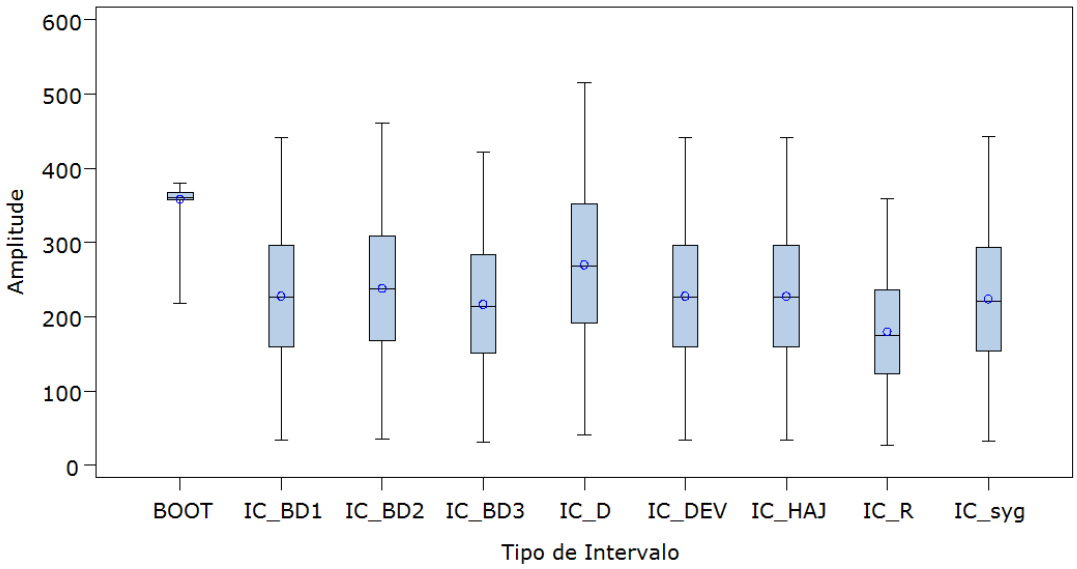

Figura 4.15: Distribuição das amplitudes dos IC baseados na distribuição normal e dos IC *Bootstrap*- $n = 3$

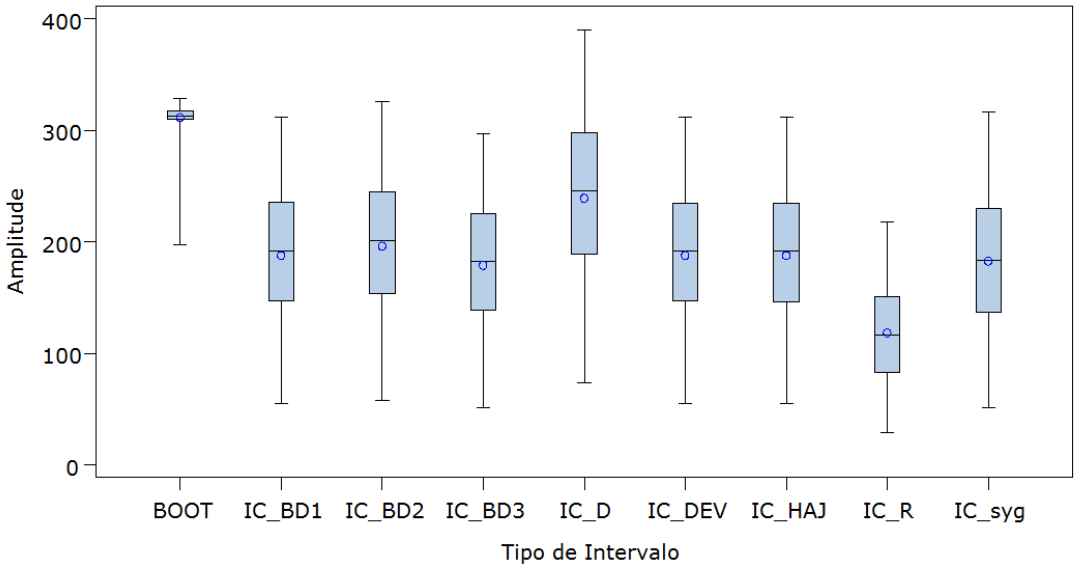

Figura 4.16: Distribuição das amplitudes dos IC baseados na distribuição normal e dos IC *Bootstrap* - $n = 4$

A partir das Figuras 4.15 e 4.16 verifica-se que os IC *Bootstrap* tem desvio inter-

quartilico da amplitude bem acima do que os IC baseados na distribuição normal. Porém, as coberturas dos IC *Bootstrap* são superiores às coberturas dos demais IC baseados na distribuição normal.

Sendo assim, os IC *Bootstrap* têm cobertura maior do que os IC baseados na distribuição normal e na *t-student* e têm desempenho melhor ou igual, quanto a amplitude, quando comparados com os IC baseados na distribuição *t-student*, porém as amplitudes dos IC *Bootstrap* são maiores do que as amplitudes dos IC baseados na distribuição normal.

4.3 População 2

Na população 2, cujos parâmetros foram apresentados na Tabela 3.1, em que deseja-se estimar o total de famílias (y) e tem-se a variável auxiliar (x) estimativa da quantidade de famílias, as variáveis x e y apresentam uma forte correlação positiva.

As Figuras 4.17, 4.18 e 4.19 apresentam a distribuição das estimativas da variância de Horvitz-Thompson produzidas por cada um dos nove estimadores analisados.

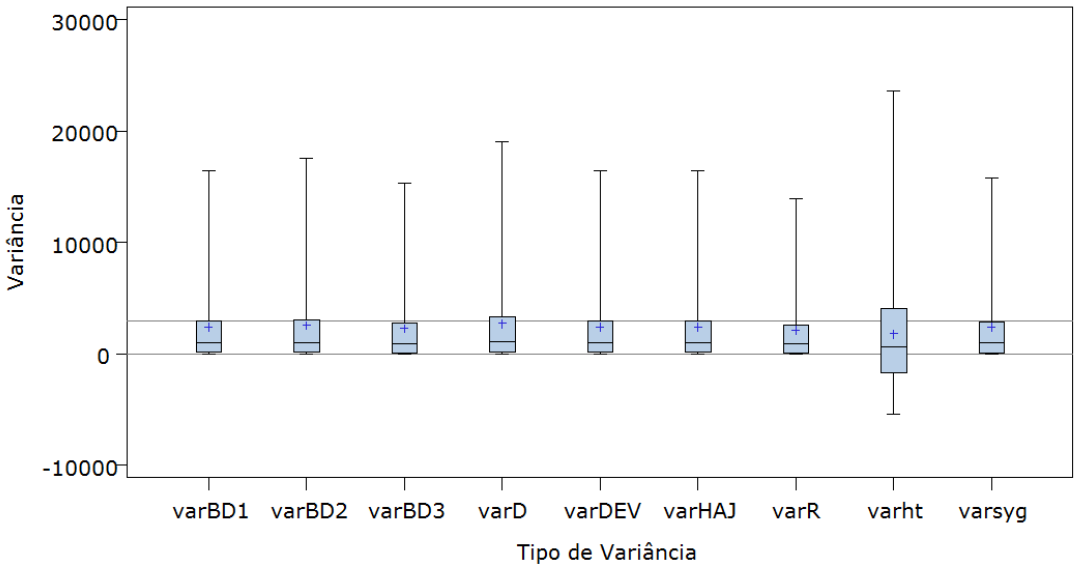

Figura 4.17: Boxplot das estimativas da variância de Horvitz-Thompson ($n = 2$)

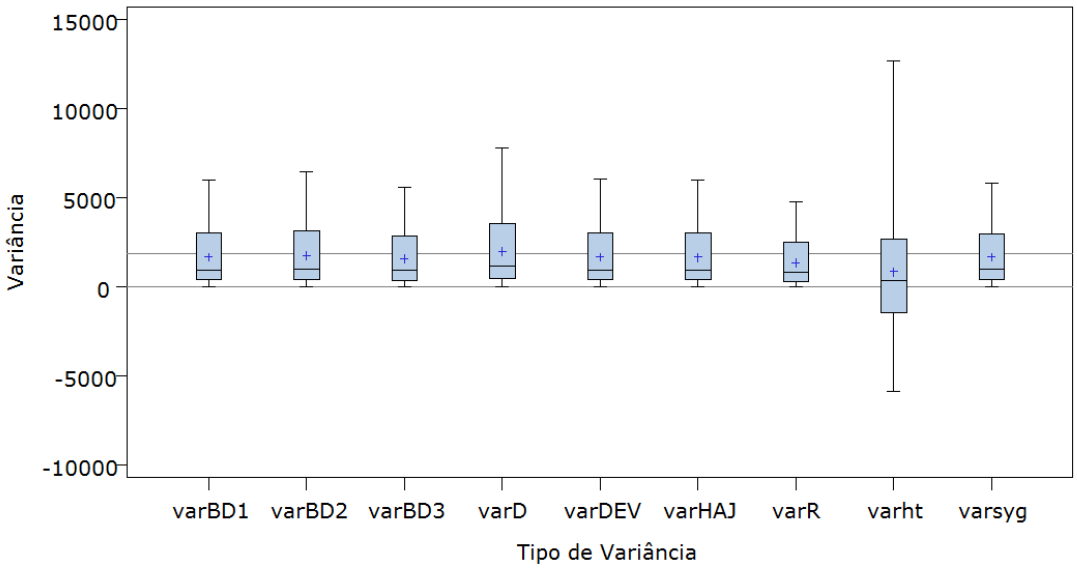

Figura 4.18: Boxplot das estimativas da variância de Horvitz-Thompson ($n = 3$)

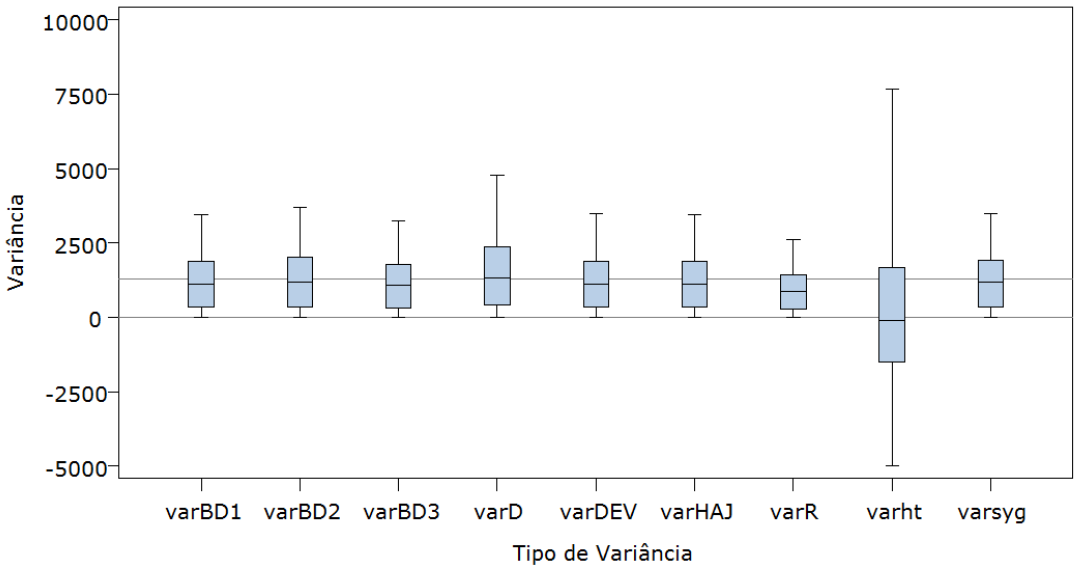

Figura 4.19: Boxplot das estimativas da variância de Horvitz-Thompson ($n = 4$)

Percebe-se que as amplitudes das estimativas produzidas pelos estimadores da variância tendem a diminuir à medida que o tamanho amostral aumenta, para todos os estimadores estudados. O estimador que apresenta menor amplitude interquartílica é $\hat{V}_R$ (1.11), embora todos os estimadores apresentem semelhanças em suas

distribuições, exceto o estimador $\hat{V}_{HT}$ (1.7) que assim como apresentado na literatura, produz estimativas negativas da variância. Por este motivo, não se utilizou este estimador para construir os intervalos de confiança apresentados posteriormente.

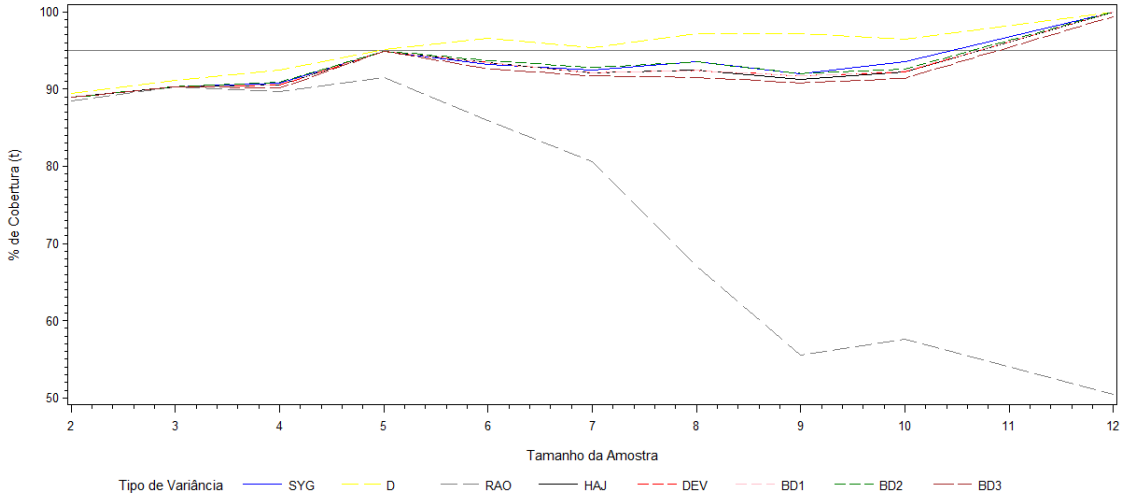

Figura 4.20: Cobertura dos IC de 95% para o total populacional, baseados na distribuição *t-student*

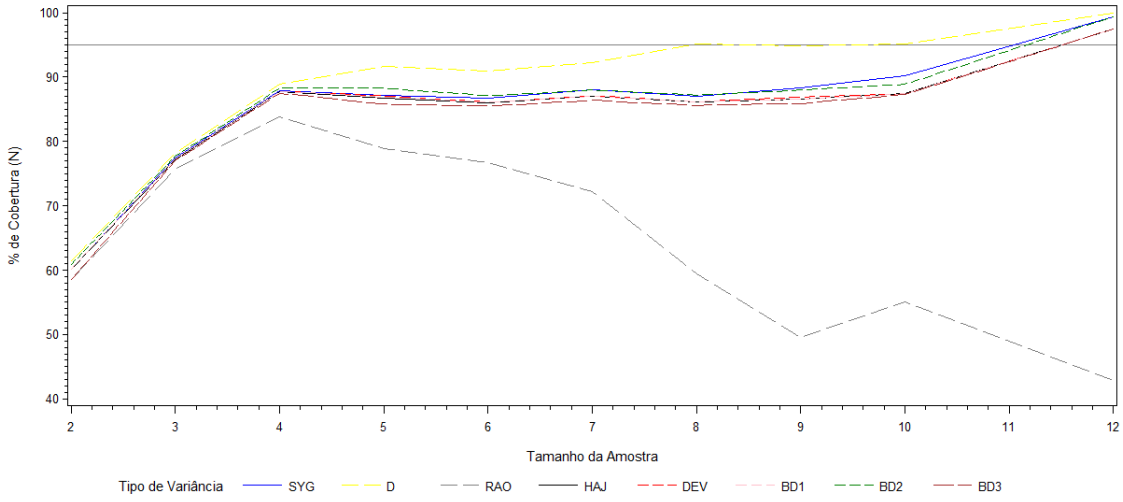

Figura 4.21: Cobertura dos IC de 95% para o total populacional, baseados na distribuição normal

As Figuras 4.20 e 4.21 apresentam, respectivamente, as coberturas dos IC baseados na distribuição *t-student* e na distribuição normal, à medida que o tamanho amostral

aumenta. Decidiu-se utilizar amostras de até 60% do tamanho populacional, como feito na população 1. Comparando as duas Figuras, é possível verificar que à medida que o tamanho amostral aumenta, as coberturas dos IC também aumentam, exceto as coberturas dos IC baseados no estimador $\hat{V}_R$. Percebe-se ainda, que as coberturas dos IC baseados na distribuição normal é menor do que as coberturas dos IC baseados na distribuição *t-student*, para todos os IC simulados.

Quanto aos intervalos baseados na distribuição *t-student*, percebe-se que até $n = 5$ todos os IC aumentam sua cobertura junto com o tamanho amostral. Entretanto, a cobertura do IC que utiliza a estimativa da variância calculada pelo estimador $\hat{V}_R$ diminui à medida que o tamanho amostral aumenta. Quando $n > 5$ a cobertura do IC que utiliza a estimativa da variância do estimador $\hat{V}_D$, se mantém acima da cobertura nominal, destacando-se dos demais intervalos de confiança. Quanto aos IC baseados nos outros seis estimadores da variância,$\hat{V}_{HAJ}$, $\hat{V}_{DEV}$, $\hat{V}_{BD1}$, $\hat{V}_{BD2}$, $\hat{V}_{BD3}$ e $\hat{V}_{SYG}$, as coberturas desses IC se mantêm parecidas, entre 90% e 95%.

Em relação aos intervalos de confiança baseados na distribuição normal, até $n = 4$ todos os IC aumentam sua cobertura e estão próximos entre si, exceto pelo estimador $\hat{V}_R$ que começa a se distanciar dos demais quando $n = 3$. Para $n > 4$ a cobertura do IC que utiliza as estimativas do estimador $\hat{V}_R$ diminui à medida que o tamanho amostral aumenta, sugerindo ineficiência deste estimador. Entretanto, a medida que o tamanho amostral aumenta, a cobertura do IC baseado no estimador $\hat{V}_D$ aumenta e se distancia dos demais IC, atingindo a cobertura nominal quando $n = 8$. As coberturas dos demais IC baseados nos outros seis estimadores da variância, $\hat{V}_{HAJ}$, $\hat{V}_{DEV}$, $\hat{V}_{BD1}$, $\hat{V}_{BD2}$, $\hat{V}_{BD3}$ e $\hat{V}_{SYG}$, se mantêm próximas, em torno de 85% a 90% e só atingem a cobertura esperada, 95%, quando $n > 10$.

Deste modo, quando se comparam as coberturas dos IC dos estimadores que independem da probabilidade de seleção conjunta e as coberturas dos IC do estimador, $\hat{V}_{SYG}$, que depende da probabilidade conjunta de seleção, estas coberturas têm desempenhos parecidos para todos os tamanhos amostrais simulados. Destaca-se apenas, a cobertura dos IC do estimador de $\hat{V}_R$, que têm desempenho muito abaixo do esperado, e a cobertura dos IC do estimador $\hat{V}_D$, que atinge cobertura superior aos demais IC.

Quanto a cobertura dos intervalos de confiança *Bootstrap* verificou-se que para todos os tamanhos amostrais as coberturas foram superiores a $99,5\%$. Portanto,

resolveu-se avaliar as amplitudes destes intervalos frente às amplitudes dos IC das Figuras 4.20 e 4.21.

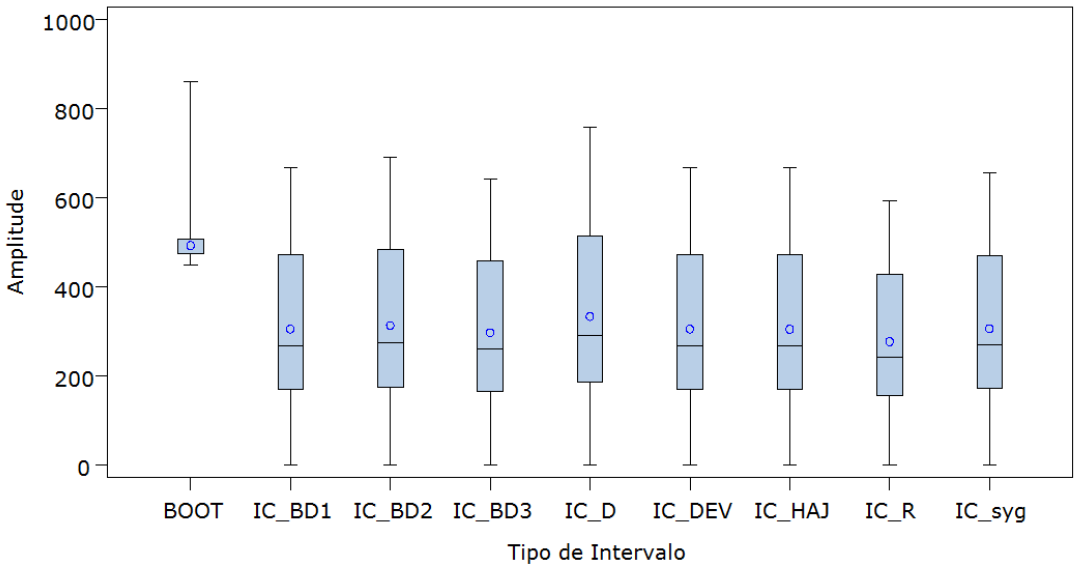

Figura 4.22: Distribuição das amplitudes dos IC baseados na distribuição *t-student* e dos IC *Bootstrap* - $n = 3$

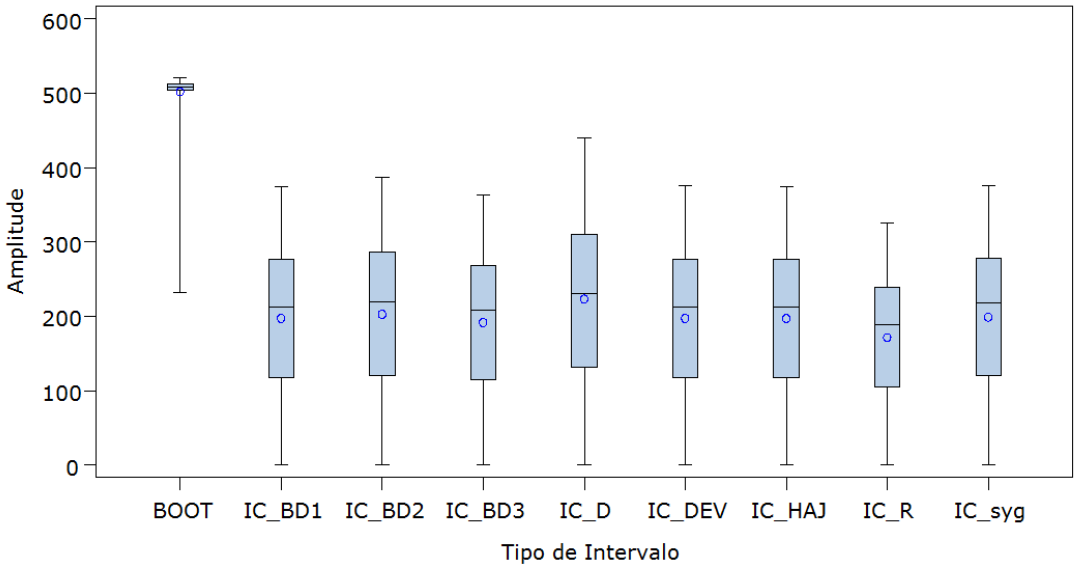

Figura 4.23: Distribuição das amplitudes dos IC baseados na distribuição *t-student* e dos IC *Bootstrap* - $n = 4$

As amplitudes dos IC *Bootstrap* são maiores do que as amplitudes dos IC baseados

na distribuição *t-student*. Portanto, os intervalos *Bootstrap* têm coberturas maiores e amplitudes maiores do que os IC *t-student*.

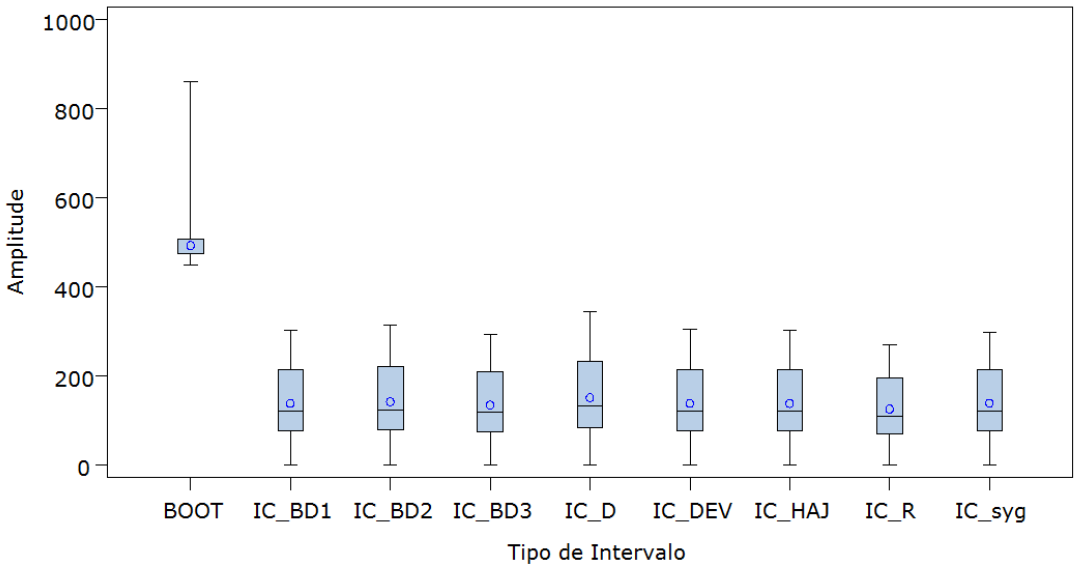

Figura 4.24: Distribuição das amplitudes dos IC baseados na distribuição normal e dos IC *Bootstrap*- $n = 3$

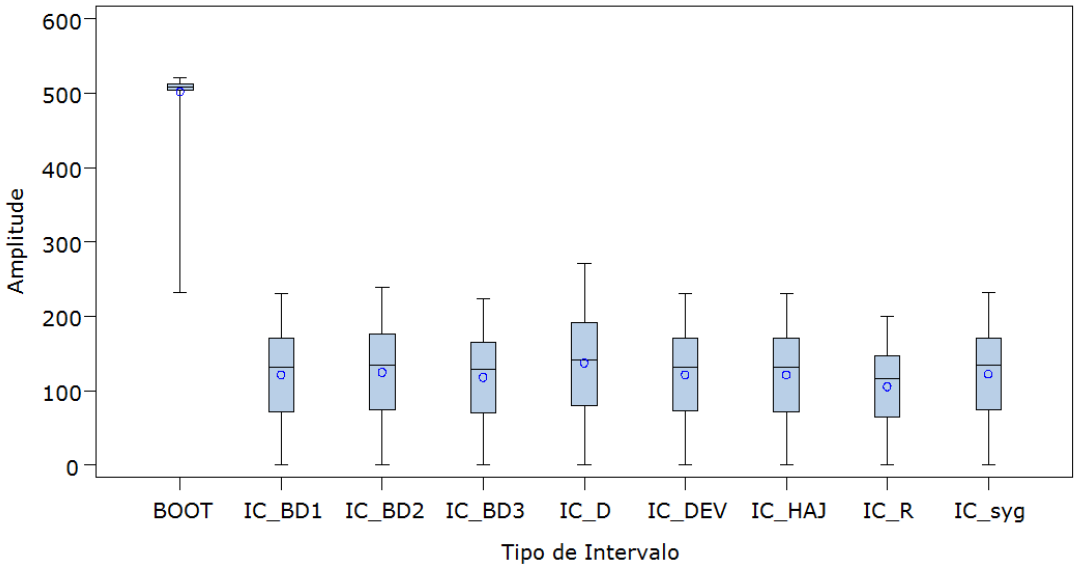

Figura 4.25: Distribuição das amplitudes dos IC baseados na distribuição normal e dos IC *Bootstrap* - $n = 4$

A partir das Figuras 4.24 e 4.25 verifica-se que os IC de *Bootstrap* têm amplitudes

maiores do que os IC baseados na distribuição normal. Porém, as coberturas dos IC *Bootstrap* são superiores as coberturas dos demais IC baseados na distribuição normal.

Sendo assim, os IC *Bootstrap* apresentam coberturas maiores do que os IC baseados nas distribuições normal e *t-student*, entretanto as amplitudes de seus intervalos são maiores do que as amplitudes dos demais IC.

4.4 População 3

Na população estudada nesta seção, , cujos parâmetros foram apresentados na Tabela 3.1, deseja-se estimar o total da área de trigo plantada e a variável auxiliar nesta população é a quantidade de aldeias. As variáveis apresentam uma correlação positiva moderada.

As Figuras 4.26, 4.27 e 4.28 apresentam a distribuição das estimativas da variância de Horvitz-Thompson produzidas por cada um dos nove estimadores analisados.

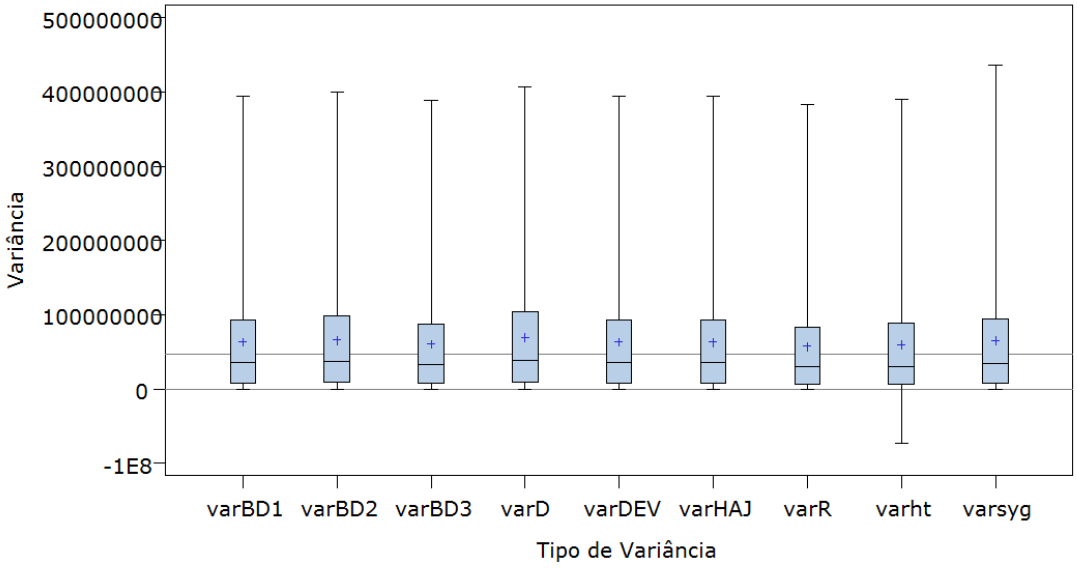

Figura 4.26: Boxplot das estimativas da variância de Horvitz-Thompson ($n = 2$)

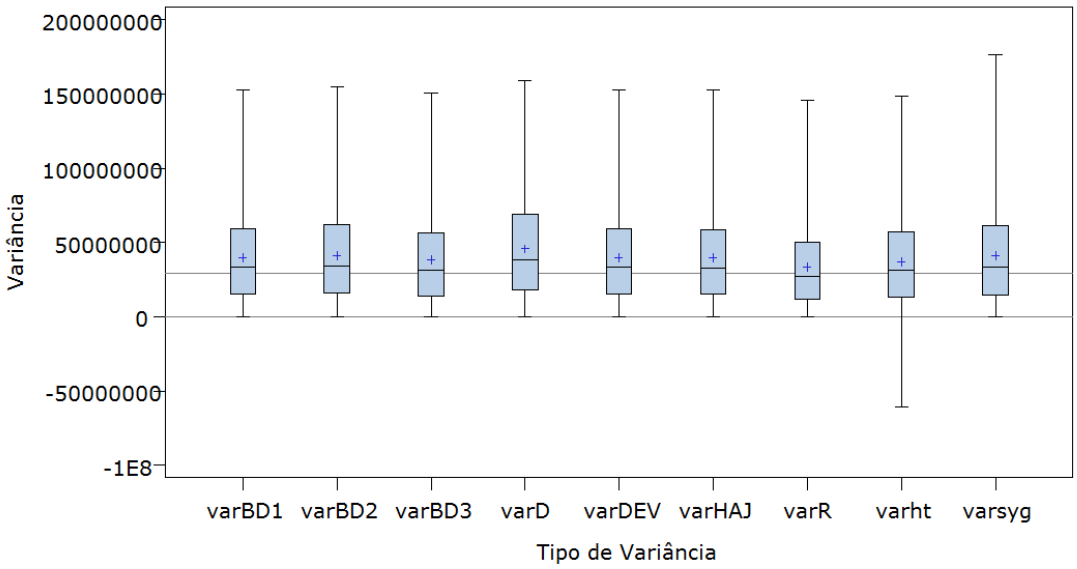

Figura 4.27: Boxplot das estimativas da variância de Horvitz-Thompson ($n = 3$)

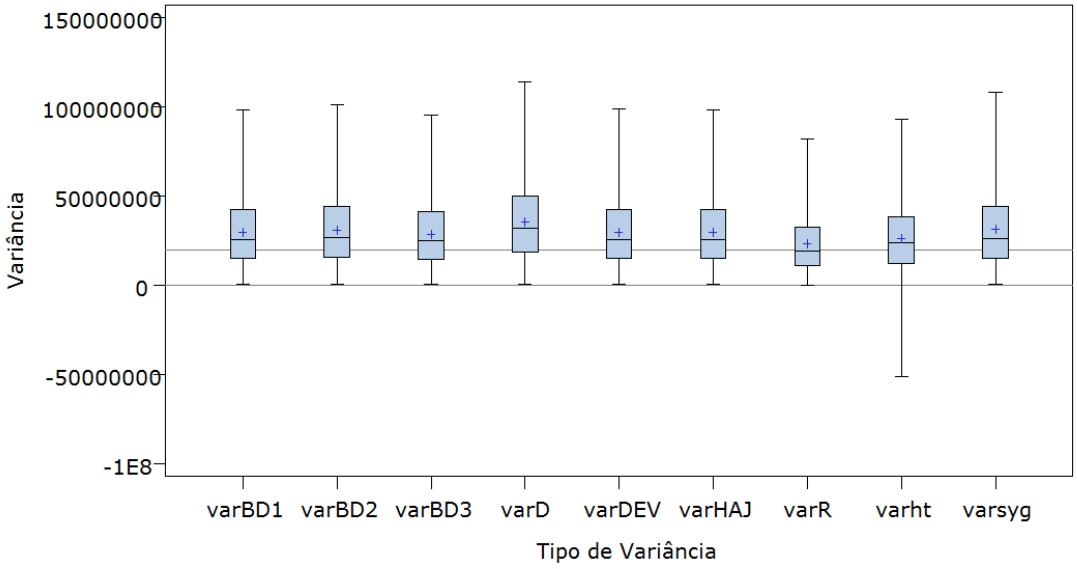

Figura 4.28: Boxplot das estimativas da variância de Horvitz-Thompson ($n = 4$)

Quanto às amplitudes destas estimativas, percebe-se que elas tendem a diminuir à medida que o tamanho amostral aumenta, para todos os estimadores estudados. Embora todos os estimadores apresentem semelhanças em suas distribuições de estimativas, o estimador $\hat{V}_{HT}$ (1.7), assim como apresentado na literatura, produz estimativas

negativas da variância. Novamente, não se utilizou este estimador na construção dos intervalos de confiança apresentados posteriormente.

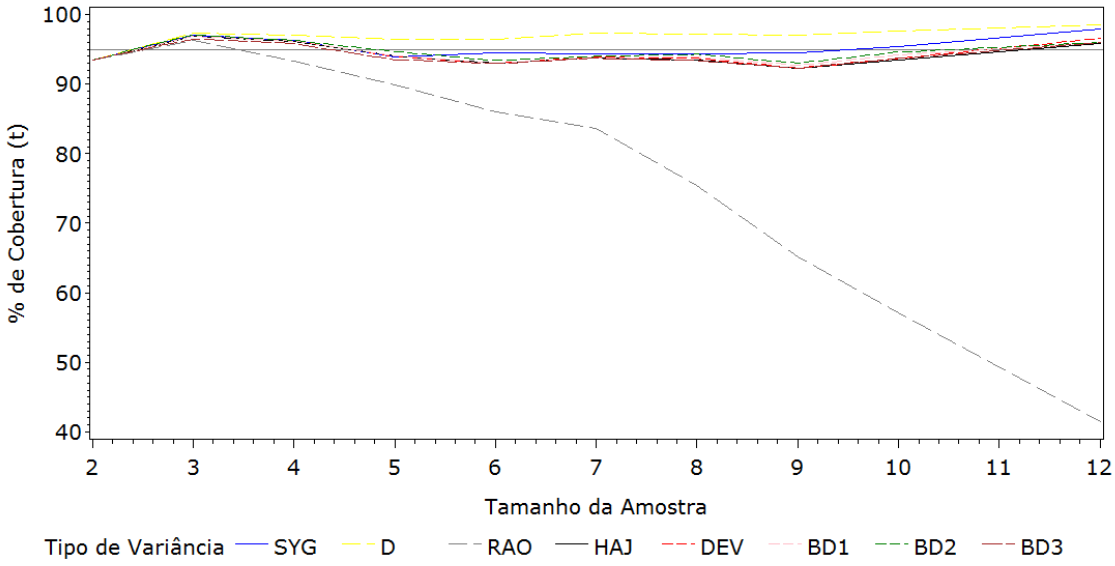

Figura 4.29: Cobertura dos IC de 95% para o total populacional, baseados na distribuição *t-student*

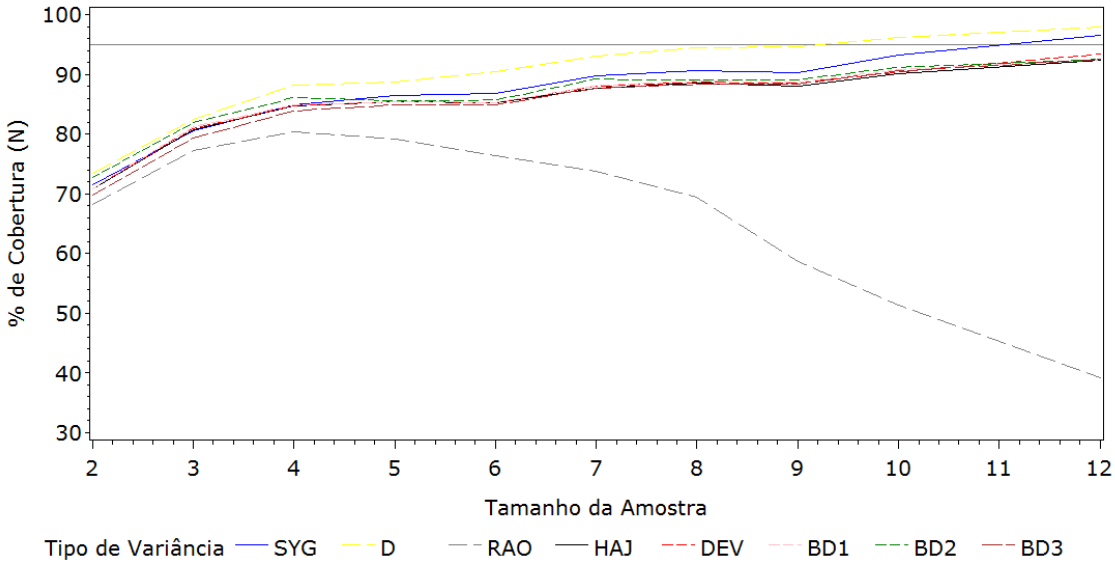

Figura 4.30: Cobertura dos IC de 95% para o total populacional, baseados na distribuição normal

As Figuras 4.29 e 4.30 apresentam, respectivamente, as coberturas dos IC baseados

na distribuição *t-student* e na distribuição normal, à medida que o tamanho amostral aumenta. Da mesma forma que na população 22, decidiu-se utilizar amostras de até 60% do tamanho populacional. Comparando as duas Figuras, verifica-se que as coberturas dos IC baseados na distribuição *t-student* são superiores às coberturas dos IC baseados na distribuição normal, para todos os IC simulados.

Todos os intervalos baseados na distribuição *t-student* até $n = 3$ têm cobertura parecida, contudo, quando $n > 3$, a cobertura do IC baseado no estimador $\hat{V}_R$ diminue à medida que o tamanho amostral aumenta. O IC que mantém cobertura mais elevada à medida que n aumenta, em relação aos demais IC, é o IC baseado no estimador $\hat{V}_D$. As coberturas dos demais IC têm desempenhos parecidos, se mantendo entre 90% e 95%, porém o IC $\hat{V}_{SYG}$ se diferencia dos outros seis intervalos e se aproxima da cobertura do IC de $\hat{V}_D$, quando $n > 7$.

Em relação aos intervalos de confiança baseados na distribuição normal, a cobertura do IC que utiliza o estimador $\hat{V}_D$ se destaca dos demais e atinge cobertura nominal quando $n > 8$. O segundo IC com melhor cobertura é o IC baseado no estimador $\hat{V}_{SYG}$ que se destaca dos demais quando $n > 8$. No caso da distribuição normal, exceto pelo IC baseado em $\hat{V}_R$, todas as coberturas dos IC aumentam à medida que o tamanho amostral aumenta.

Deste modo, o IC baseado no estimador que necessita da probabilidade de seleção conjunta tem desempenho superior aos demais IC à medida que o tamanho populacional aumenta, exceto quando comparado ao IC que utiliza o estimador $\hat{V}_D$.

Quanto a cobertura dos intervalos de confiança *Bootstrap* verificou-se que para todos os tamanhos amostrais a cobertura foi superior a $99,5\%$. Portanto, resolveu-se avaliar a amplitude destes intervalos frente às amplitudes dos IC das Figuras 4.29 e 4.30.

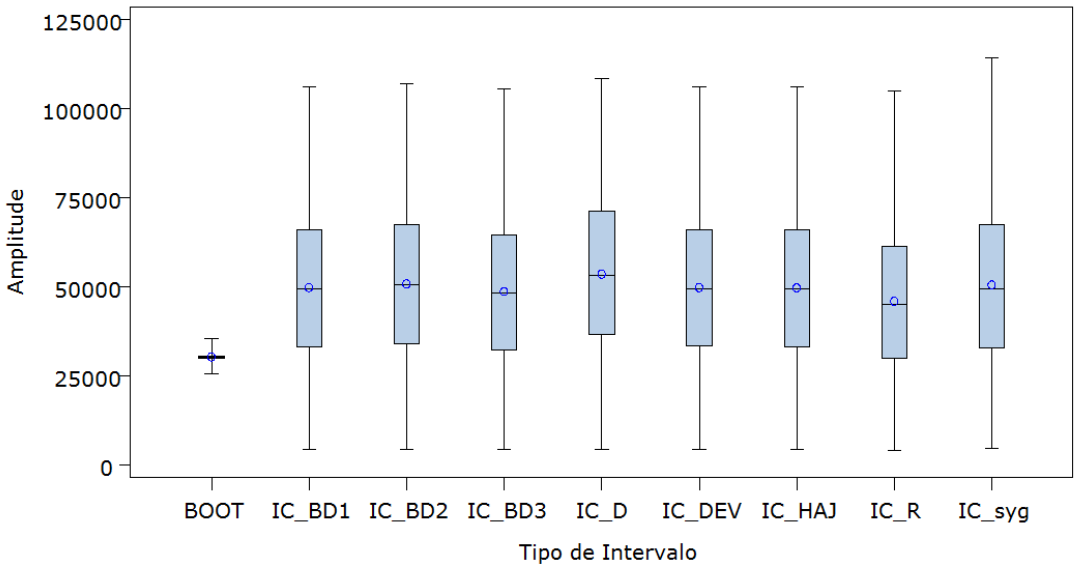

Figura 4.31: Distribuição das amplitudes dos IC baseados na distribuição *t-student* e dos IC *Bootstrap* - $n = 3$

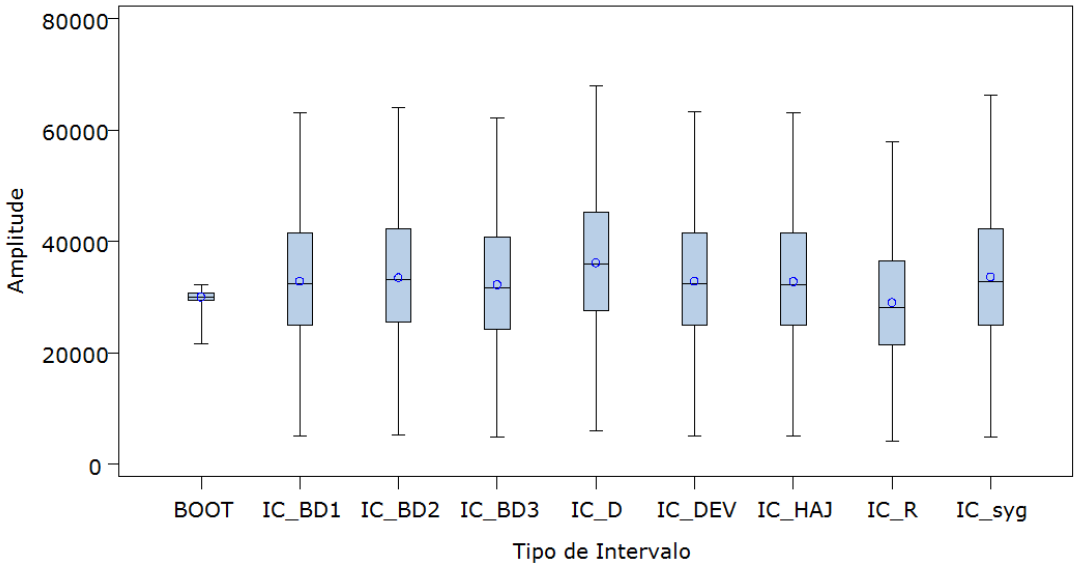

Figura 4.32: Distribuição das amplitudes dos IC baseados na distribuição *t-student* e dos IC *Bootstrap* - $n = 4$

As amplitudes dos IC *Bootstrap* são menores ou iguais as amplitudes dos IC baseados na distribuição *t-student*. Portanto, os IC *Bootstrap* têm coberturas maiores e amplitudes menores do que os IC *t-student*.

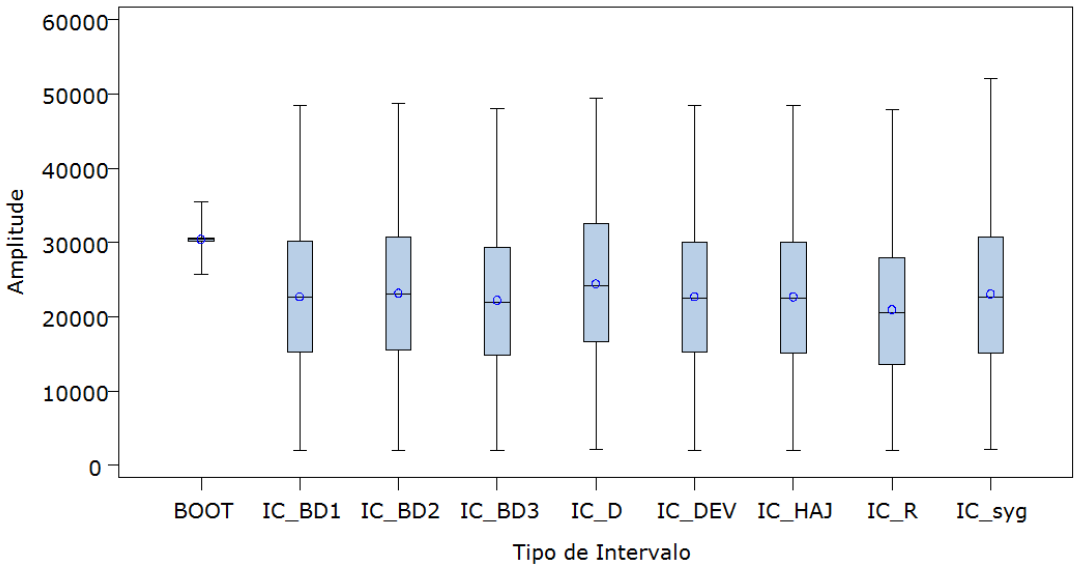

Figura 4.33: Distribuição das amplitudes dos IC baseados na distribuição normal e dos IC *Bootstrap*- $n = 3$

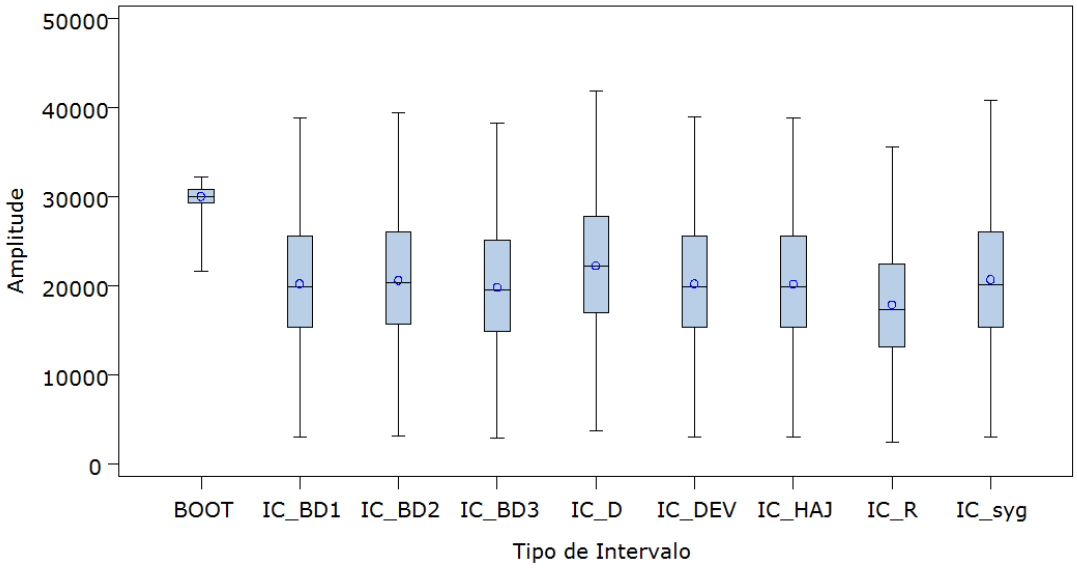

Figura 4.34: Distribuição das amplitudes dos IC baseados na distribuição normal e dos IC *Bootstrap* - $n = 4$

A partir das Figuras 4.33 e 4.34 verifica-se que os IC de *Bootstrap* têm amplitudes maiores do que os IC baseados na distribuição normal. Porém, as coberturas dos IC *Bootstrap* são superiores as coberturas dos demais IC baseados na distribuição

normal.

Sendo assim, os IC *Bootstrap* apresentam coberturas superiores aos IC baseados na distribuição normal e na *t-student* e têm desempenhos melhores ou iguais, quanto a amplitude, quando comparado com os IC baseados na distribuição *t-student*. Porém, as amplitudes dos IC *Bootstrap* são maiores do que os IC baseados na distribuição normal.

Capítulo 5

Conclusão

O estimador de Horvitz-Thompson (HT) é muito utilizado na literatura, pois é um estimador não viesado, que foi proposto para estimar o total populacional ou a média em amostras sem reposição, sendo as unidades selecionadas com probabilidades desiguais de seleção. O estimador de Horvitz-Thompson é utilizado em diferentes áreas do conhecimento e com diferentes abordagens.

Neste trabalho, procurou-se fazer uma revisão de literatura para verificar as metodologias utilizadas para calcular intervalos de confiança (IC) para o parâmetro estimado pelo estimador de Horvitz-Thompson. Verificou-se que quando o estimador de Horvitz-Thompson é utilizado para estimar o total ou a média populacional, geralmente se utiliza o intervalo clássico baseado na distribuição normal, entretanto são apresentados na literatura outros métodos, tais como *Bootstrap* e Intervalos de máxima verossimilhança empírica. Quanto ao intervalo clássico se pode utilizar diferentes estimadores da variância para construí-lo, o que impactará no resultado do IC.

Sendo assim, primeiramente se buscou comparar alguns estimadores da variância de Horvitz-Thompson apresentados na literatura e posteriormente se realizou um estudo empírico a cerca do comportamento dos intervalos de confiança para o total populacional estimado pelo estimador de Horvitz-Thompson. Utilizou-se a distribuição normal e a distribuição *t-student* para calcular os intervalos de confiança para o total populacional.

Matei e Tillé (2005) quando verificaram o desempenho dos estimadores da variância

utilizaram também a cobertura dos intervalos de confiança produzidos por estes estimadores para avaliar os resultados, porém os resultados desses autores foram obtidos para grandes populações.

Os resultados empíricos deste trabalho foram obtidos a partir de populações pequenas. Ratificando os resultados de Matei e Tillé (2005), em pequenas populações, verificou-se que não houve muitas diferenças entre o desempenho dos estimadores que independem da probabilidade de seleção conjunta (π_{ij}) e do estimador (1.8), que necessita de π_{ij}. Sugerindo portanto, que também em populações pequenas a utilização dos estimadores que independem de π_{ij} é adequada e tem desempenho parecido ao do estimador (1.8), possibilitando contornar a dificuldade do cálculo da probabilidade de seleção conjunta.

Brewer e Donadio (2003) utilizaram populações pequenas para verificar o desempenho dos estimadores da variância quando $n = 2$, entretanto, os autores não contruíram IC para o total populacional utilizando estes estimadores da variância. Os resultados apontaram que os estimadores que independem de π_{ij} tendem a ser mais eficientes do que o estimador (1.8), entretanto o ganho deles é pequeno. Esta conclusão não foi ratificada neste trabalho, haja vista que não se verificou uma diferença significativa entre estes estimadores.

Apesar da aproximação de Rao (1963) ter sido utilizada e ter apresentado bons resultados, o estimador proposto por ele não teve bom desempenho, pois, a cobertura do IC baseado neste estimador diminuiu à medida que o tamanho amostral aumentou.

Outro resultado que ratificou o que foi expresso na literatura, foram as estimativas negativas produzidas pelo estimador da Variância (1.7). Entretanto, o estimador (1.8) não apresentou resultados negativos nas simulações, não sendo possível ratificar o que foi mencionado por Lohr (2010).

Quanto ao desempenho dos intervalos de confiança, verificou-se que os IC baseados na distribuição *t-student* apresentaram coberturas maiores do que os IC baseados na distribuição normal, ratificando o resultado de Särndal et al. (1992), e os IC *Bootstrap* tiveram maiores coberturas do que todos os demais IC. Entretanto, em diversas situações os IC *Bootstrap* apresentaram amplitudes maiores do que os IC *t-student* e normal. Vale ressaltar que dentre os IC baseados nas distribuições normal e *t-student*, o IC com melhor cobertura para todos os tamanhos amostrais e populações

foi o IC baseado na distribuição *t-student* que utilizou o estimador com reposição proposto por Durbin (1953)(1.9).

5.1 Limitações do Trabalho

Uma das limitações enfrentadas neste trabalho é que, devido ao grande esforço computacional à medida que o tamanho amostral aumenta, não foi possível utilizar nas simulações a probabilidade de seleção exata no cálculo do estimador de Horvitz-Thompson. A aproximação (1.10) utilizada não funciona quando a condição $x_i \leq \frac{\sum_{i \in U} x_i}{n}$ não é satisfeita, principalmente com grandes populações.

Tentou-se aplicar a metodologia do Capitulo 3 à base apresentada em Kish (1965), no entanto as coberturas dos intervalos de confiança diminuíam à medida que o tamanho amostral aumentava. Isso parece um pouco estranho, uma vez que à medida que a amostra aumenta, as estimativas pontuais convergem para o parâmetro e as variâncias convergem para zero. O problema identificado foi que as probabilidades de seleção não convergiam para 1 no caso em que a amostra se aproximava da população, ou no caso extremo, quando a amostra era igual à população, as probabilidades de seleção não eram exatamente iguais a 1 (de fato, eram bem maiores do que 1, visto que a restrição $x_i \leq \frac{\sum_{i \in U} x_i}{n}$ não era satisfeita). Por outro lado, quando se utiliza a probabilidade de seleção calculada de forma exata ($\pi_i = \sum_{s: i \in s} p(s)$), à medida que a amostra aumenta, as probabilidades de seleção convergem para 1, e no caso extremo, são iguais a 1 como pode ser visto na Tabela 5.1, que apresenta a mesma população da Tabela 1, retirada uma amostra de tamanho 3 (ou seja, igual a população).

Tabela 5.1: Probabilidades de seleção conjunta, $n = 3$

Domicílios	A	B	C	π
A	0	1	1	1
B	1	0	1	1
C	1	1	0	1
π	1	1	1	3

Outro óbice enfrentado no trabalho foi que a *Library* do Software R para cons-

trução de intervalos de máxima verossimilhança empírica, apesar de publicada no artigo Berger (2015) não está disponível para uso. Isso restringiu o estudo comparativo no que se refere ao cálculo dos intervalos de confiança, apesar do estudo de Berger e Torres (2012) apontar que sua proposta de intervalo possui maior cobertura do que o intervalo clássico. No entanto, a aplicação feita utilizando os intervalos com a distribuição *t-student* e a distribuição normal, além de comparar as variâncias amostrais com a variância populacional para $n > 2$, proporcionou um ganho no entendimento na utilização de intervalos de confiança clássico para o estimador de Hovitz-Thompson.

Uma possivel causa para os intervalos *Bootstrap* terem cobertura próximas a 100% deve se ao fato de ter sido selecionada a primeira amostra com probabilidades desiguais e sem reposição e as amostras *Bootstrap* com probabilidades iguais e com reposição. Isto pode ser mais investigado posteriormente.

5.2 Recomendações para Trabalhos Futuros

Recomenda-se para trabalhos futuros:

- Utilizar o intervalo de Berger e Torres (2012) para comparar com o intervalo clássico utilizando dados reais e utilizando a probabilidade de seleção exata, a fim de verificar sua performance em pequenas amostras.
- Desenvolver algoritmo computacional mais eficiente para o cálculo das probabilidade de seleção exatas, a fim de se utilizar em amostras grandes.

Referências Bibliográficas

Bailey, N. T. (1952). Improvements in the interpretation of recapture data. *The Journal of Animal Ecology*, pages p. 120–127.

Berger, Y. G. (2015). An r library to construct empirical likelihood confidence intervals for complex estimators. *Proceeding of the conference on New Techniques and Technologies for Statistics. NTTS*.

Berger, Y. G. e De La Riva Torres, O. (2016). Empirical likelihood confidence intervals for complex sampling designs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78:p. 319–341.

Berger, Y. G. e Torres, O. D. L. R. (2012). Estimation of confidence intervals using a new empirical likelihood approach. Disponível em: URL `http://www.q2012.gr/articlefiles/sessions/25.1_Berger_Estimation%20of%20confidence%20intervals_Q2012.pdf`.

Berger, Y. G. e Torres, O. D. L. R. (2014). *Empirical Likelihood Confidence Intervals: An Application to the EU-SILC Household Surveys. In Contributions to Sampling Statistics*. Springer International Publishing.

Bolfarine, H. e Bussab, W. O. (2005). *Elementos de Amostragem*. ABE - Projeto Fisher.

Brewer, K. R. W. e Donadio, M. E. (2003). The high entropy variance of the horvitz-thompson estimator. *Survey Methodology*, 29:p. 189–196.

Cardot, H., Degras, D., e Josserand, E. (2013). Confidence bands for horvitz-thompson estimators using sampled noisy functional data. *Bernoulli*, 19:p. 2067–2097.

Cardot, H. e Josserand, E. (2011). Horvitz-thompson estimators for functional data: asymptotic confidence bands and optimal allocation for stratified sampling. *Biometrika*, 98:p. 107–118.

Casella, G. e Berger, R. L. (2002). *Statistical Inference*, (2th ed.). Duxbury-Thomson Learning.

Chao, A. (2001). An overview of closed capture-recapture models. *Journal of Agricultural, Biological, and Environmental Statistics*, 6.2:p.158–175.

Chapman, D. (1951). Some properties of the hypergeometric distribution with applications to zoological sample censuses. *University of California Press*, 7.1:p. 131–60.

Cochran, W. G. (1977). *Sampling Techniques*, (3rd ed.). John Wiley & Sons.

Craig, C. C. (1953). On the utilization of marked specimens in estimating populations of flying insects. *Biometrika*, pages p. 170–176.

Cumberland, W. G. e Royall, R. W. (1981). Prediction models and unequal probability sampling. *Journal of the Royal Statistical Society, Ser. B*, 43:p. 353–367.

Deville, J. C. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology*, 25:p. 193–203.

Durbin, J. (1953). Some results in sampling theory when the units are sampled with unequal probabilities. *Journal of the Royal Statistical Society, Series B*, 15:p.262–269.

Erdós, P. e Rényi, A. (1959). On the central limit theorem for sample from a finite population. *Annals of Mathematical Statistics*, 4:p. 49–61.

Fuller, W. A. (2009). *Sampling Statistics*. New York: Wiley.

Gaskell, T. J. e George, B. J. (1972). A bayesian modification of the lincoln index. *Journal of Applied Ecology*, pages p. 377–384.

Hájek, J. (1964). Assymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics*, 35:p. 1491–1523.

Hansen, M. H. e Hurwitz, W. N. (1943). On the theory of sampling from finite population. *The Annals of Mathematical Statistics*, 14(04):p. 333–362.

Hartley, H. O. e Rao, J. N. K. (1962). Sampling with unequal probability and without replacement. *Annals of Mathematical Statistics*, 33:p. 350–374.

Horvitz e Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:p. 663–685.

Huggins, R. M. (1989). On the statistical analysis of capture experiments. *Biometrika*, pages p. 133–140.

Hwang, W. e Huang, S. Y. H. (2003). Estimation in capture-recapture models when covariates are subject to measurement erros. *Biometrics*, 59:p. 1113–1122.

Hwang, W. e Huggins, R. (2005). An examination of the effect of heterogeneity on the estimation of population size using capture-recapture data. *Biometrika*, 92:p. 229–233.

Isaki, C. T. e Fuller, W. A. (1982). Survey design under regression superpopulation model. *Journal of the American Statistical Association*, 77:p.89–96.

Kish, L. (1965). *Survey Sampling*. John Wiley & Sons.

Lohr, S. L. (2010). *Sampling: Design and Analysis*, (2th ed.). Cengage Learning.

Magalhães, M. N. e Lima, A. C. P. (2013). *Noções de Probabilidade e Estatística*. EDUSP.

Mascioli, F. (2008). Capture- recapture methods to estimate prevalence indicators for the evaluation of drug policies. *Bulletin on Narcotics - UNODC*, LX.

Matei, A. e Tillé, Y. (2005). Evaluation of variance approximations and estimators in maximum entropy sampling with unequal probability and fixed sample size. *Journal of Official Statistics*, 21.4:p.543.

Midha, C. (1988). The horvitz-thompson estimator and estimates of its variance. *Annual Meeting of the American Statistical Association*, (82).

Mood, A. M., Boes, D. C., e Graybill, F. A. (1974). *Introduction to the Theory of Statistics*, (3th ed.). McGraw-Hill.

Murthy, M. N. (1967). Sampling theory and methods. *Statistical Publishing Society*.

Nascimento, I. F. e Silva, A. R. (2013). A sas macro for generating a set of all possible samples with unequal probabilites without replacement. *SAS Global Forum*.

Overton, W. S. (1985). A sampling plan for streams in the national stream survey. Oregon State University.

Overton, W. S. e Davis., D. E. (1969). Estimating the numbers of animals in wildlife populations. *The Wildlife Society*, 623:p. 403–455.

Pollock, K. H. (2002). The use of auxiliary variables in capture-recapture modelling: an overview. *Journal of Applied Statistics*, 29:p.85–102.

Rao, J. N. K. (1963). On three procedures of unequal probability sampling without replacement. *Journal of the American Statistical Association*, 58.301:p. 202–215.

Särndal, C. E., Swensson, B., e Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer Verlag.

Sen, P. K. (1953). On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Verlag*, 5:p. 119 – 127.

Sirken, M. G. (2001). The hansen-hurwitz estimator revisited: Pps sampling without replacement. *Annual Meeting of the American Statistical Association*.

Southwood, T. R. E. e Henderson, P. A. (2000). *Ecological Methods*. Blackwell Science.

Stehman, S. V. e Overton, W. S. (1987). Estimating the variance of the horvitz-thompson estimator in variable probability, systematic samples. Disponível em: URL `http://www.amstat.org/sections/srms/Proceedings/papers/1987_132.pdf`.

Van Der Heijden, P. G., Bustami, R., Cruyff, M. J., Engbersen, G., e Van Houwelingen, H. C. (2003). Point and interval estimation of the population size using the truncated poisson regression model. *Statistical Modelling*, 3.4:p. 305–322.

Welch, H. E. (1960). Two applications of a method of determining the error of population estimates of mosquito larvae by the mark and recapture technique. *Ecology*, pages p. 228–229.

Yates, F. e Grundy, P. M. (1953). Selection without replacement from within strata with probability proportional to size. *Royal Statistical Society*, pages p. 253 – 261.

Yip, P. S., Huggins, R. M., e Lin, D. Y. (1995). Inference for capture-recapture experiments in continuous time with variable capture rates. *Biometrika*, 83:p.447–483.

Yip, P. S., Wan, E. C., e Chan, K. S. (2001). A unified approach for estimating population size in capture-recapture studies with arbitrary removals. *Journal of agricultural, biological, and environmental statistics*, 6.2:p.183–194.

Yip, P. S. F., Zhou, Y., Lin, D. Y., e Fang, X.-Z. (1999). Estimation of population size based on additive hazards models for continuous-time recapture experiments. *Biometrics*, 55.3:p. 904–908.