Universidade Federal da Paraíba
Centro de Ciências Exatas e da Natureza
Departamento de Estatística

AVALIAÇÃO DE ERROS DE MENSURAÇÃO EM MÉTODOS DE ESTIMAÇÃO SOB A ABORDAGEM DE CADASTRO DUPLO.

Clarissa de Oliveira Cavalcanti

Junho/2018

Clarissa de Oliveira Cavalcanti

AVALIAÇÃO DE ERROS DE MENSURAÇÃO EM MÉTODOS DE ESTIMAÇÃO SOB A ABORDAGEM DE CADASTRO DUPLO

Trabalho de Conclusão de Curso II apresentado ao Curso de Bacharelado em Estatística na Universidade Federal da Paraíba, como requisito para obtenção do Grau de Bacharel. Áreas de Concentração: Estatística aplicada e Amostragem.

Orientador: Profº Dr. Hemílio Fernandes Campos Coelho

João Pessoa
Junho de 2018

Ata da Sessão Pública de Defesa de Monografia de Clarissa de Oliveira Cavalcanti, realizada em 08 de junho de 2018 no Departamento de Estatística da Universidade Federal da Paraíba - UFPB.

Aos oito dias do mês de junho de dois mil e dezoito, às onze horas, na Sala 19 do Centro de Ciências Exatas e da Natureza da Universidade Federal da Paraíba, reuniram-se os membros da Banca Examinadora constituída para avaliar o Trabalho de Conclusão Curso intitulado "Avaliação de erros de mensuração em métodos de estimação sob abordagem de cadastro duplo" de autoria de **Clarissa de Oliveira Cavalcanti**. A Banca Examinadora foi composta pelos professores: Prof. Dr. Hemílio Fernandes Campos Coelho (DE-UFPB, orientador), Profa. Dra. Gilmara Alves Cavalcanti (DE-UFPB, examinadora) e Profa. Dra. Maria Lídia Coco Terra (DE-UFPB, examinadora). Dando início aos trabalhos, o presidente da banca cumprimentou os presentes, comunicou aos mesmos a finalidade da reunião e passou à palavra à discente para que se fizesse, oralmente, a exposição de seu trabalho de monografia. Concluída a apresentação, a discente foi arguida pela Banca Examinadora que sugeriu algumas alterações até o dia 20 de junho de 2018, de acordo com a Resolução No. 02/2014 do Colegiado do Curso de Bacharelado em Estatística da UFPB. Uma vez entregue a versão final do Trabalho de Conclusão de Curso à Coordenação do Bacharelado em Estatística, com as alterações solicitadas pela Banca Examinadora dentro do prazo estabelecido, a discente será aprovada com nota (10,0 , que é a média aritmética das notas atribuídas pelos membros da Banca Examinadora.

João Pessoa, 08 de junho de 2018.

Prof. Hemílio Fernandes Campos Coelho (Orientador)

Profª. Gilmara Alves Cavalcanti (Examinadora)

Profª Maria Lídia Coco Terra (Examinadora)

Clarissa de Oliveira Cavalcanti (Discente)

Catalogação na publicação
Seção de Catalogação e Classificação

C376a Cavalcanti, Clarissa de Oliveira.
Avaliação de erros de mensuração em métodos de estimação sob a abordagem de Cadastro Duplo / Clarissa de Oliveira Cavalcanti. - João Pessoa, 2018.
73 f.

Orientação: Hemílio Fernandes Campos Coelho.
TCC (Especialização) - UFPB/CCEN.

1. Amostragem. 2. Estimador Horvitz Thompson. 3. Cadastro Duplo. 4. Estratégia de Hartley. 5. Estratégia de Bankier. 6. Bootstrap. I. Coelho, Hemílio Fernandes Campos. II. Título.

UFPB/CCEN

Este trabalho é dedicado à minha família e amigos, que, certamente, em muito me agregaram ao longo deste curso.

Agradecimentos

Agradeço ao Senhor Deus pelo dom da vida e por me permitir dar mais um passo, ao concluir esta graduação.

Aos meus pais, Marlene e Cavalcanti, obrigada pelos incentivo e amor incondicional. Minhas irmãs, Clênia e Clécia, por todo apoio e confiança. A meu noivo, Kaike, por estar comigo em todos os momentos e me apoiar em todos os caminhos e escolhas.

Meus amigos, Roberto, Kelfãnio, Manoel, Danilo, Lukas, Diogo, Adenice, André, que participaram de muitos momentos em minha graduação, pessoas especiais que levarei para vida toda. A Anny e Tainara que me acolheram inúmeras vezes, me ouviram e aconselharam.

Professor Hemílio, por ter participado desta trajetória, me incentivando, apoiando e, principalmente orientando. Agradeço pela oportunidade de ter sido sua aluna PIBIC, por toda experiência que obtive e pela confiança que depositou em mim.

Minha gratidão também não poderia de deixar de existir para com os atenciosos professores do Departamento de Estatística, em especial aos que tive o privilégio de cursar disciplinas. Agradeço a todos que, de alguma maneira regaram a semente do meu trabalho durante o percurso no curso: Professor João Agnaldo,pelo incentivo, espelho de caráter e confiança, à professora Ana Flavia por ser transmitir alegria e compartilhar vários momentos descontraídos ao seu lado, professor Neir por acreditar em mim nos primeiros períodos do curso e me confiar uma Iniciação Cientifica, Professor Marcelo, Tatiene, Maria Lidia, Tarciana, Luiz, Hemílio, Rodrigo, Eufrásio e Izabel, a vocês meus agradecimentos por terem participado desta trajetória, me incentivando, apoiando e, principalmente, orientando de maneira dura e terna, enriquecendo enormemente o conteúdo deste trabalho.

Agradeço a banca avaliadora deste trabalho pela disposição em contribuir com esta monografia, pelos pertinentes apontamentos que enriqueceram este estudo.

Aos amigos e familiares, sintam-se incluídos nesta seção de agradecimentos. Vocês são responsáveis por partilhar de momentos alegres e tristes, sempre com muito apoio.

”O homem paciente resiste até o momento oportuno,
e será recompensado com a alegria.
Até o momento certo, ele esconde o que pensa,
e muitos elogiarão a sua inteligência”
(Eclesiástico cap: 1 vs 19,20.)

RESUMO

A abordagem de Cadastro Duplo (Dual Frame) envolve um levantamento amostral onde dois cadastros, denotados por A e B, são utilizados com o objetivo de fornecer cobertura para uma única população-alvo. Esta abordagem tem sido utilizada com êxito na literatura em situações onde um único cadastro não consegue fornecer cobertura completa dessa população. Este plano de trabalho objetiva estudar estratégias para amenizar os efeitos do erro de mensuração, utilizando estimadores para totais populacionais baseados em literaturas conhecidas. Estas estimativas foram avaliadas sobre efeito de um e dois cadastros. No Cadastro Duplo, onde apenas o cadastro A está contaminado com erros de mensuração e B está livre, percebe-se que quando aumenta a contaminação o Bootstrap tem dificuldade de corrigir as estimativas, independente da distribuição utilizada. Para os testes foram utilizadas as distribuições Uniforme(0,1), Gamma(2,1) e Weibull(2,1) e os estimadores utilizados foram Horvitz Thompson, Hartley e Bankier.

PALAVRAS CHAVE: Amostragem, Estimador Horvitz Thompson, Cadastro Duplo, Estratégia de Hartley, Estratégia de Bankier, Bootstrap.

Conteúdo

1 Introdução **1**

2 Objetivos **3**
2.1 Gerais . . . 3
2.2 Específicos . . . 3

3 Material e Métodos **4**
3.1 Método Probabilístico de Amostragem . . . 4
3.2 Estimador Horvitz-Thompson (HT) . . . 4
3.3 Simulação de Monte Carlo . . . 6
3.4 Bootstrap . . . 6
3.4.1 Estimação Bootstrap do viés . . . 7
3.4.2 Estimação Bootstrap do desvio padrão . . . 7

4 A abordagem de Cadastro Duplo **8**
4.1 Estratégia de Hartley . . . 11
4.2 Estratégia de Bankier . . . 13

5 Erro de mensuração **15**

6 Resultados e Discussão **17**
6.1 Estimação de totais populacionais . . . 19
6.2 Estimação de totais populacionais sob a abordagem de Cadastro Duplo . . . 25
6.2.1 Estimativas através da Estratégia de Hartley . . . 26
6.2.2 Estimativas através da Estratégia de Bankier . . . 30
6.3 Comparação dos Estimadores . . . 35

7 Conclusões **37**
7.0.1 Sugestões para Trabalhos Futuros . . . 38

8 Anexos **41**

Lista de Figuras

4.1 ((a) Cadastro duplo fornecendo cobertura completa para a população-alvo; (b) Cadastro Duplo utilizado para melhorar a relação custo benefício) 9
4.2 Domínios a, b e ab induzidos pela abordagem de Cadastro Duplo 9
4.3 Quantidades amostrais geradas pela estratégia BK 13

6.1 Viés Relativo ao estimador Horvitz Thompson da distribuição Uniforme nos três tipos de erro de mensuração. 23
6.2 Viés Relativo ao estimador Horvitz Thompson da distribuição Gamma nos três tipos de erro de mensuração. 24
6.3 Viés Relativo ao estimador Horvitz Thompson da distribuição Weibull nos três tipos de erro de mensuração 25
6.4 Avaliação do Viés Relativo sob estratégia de Hartley na distribuição Gamma nos três tipos de erro de mensuração 29
6.5 Avaliação do Viés Relativo sob estratégia de Hartley na distribuição Uniforme nos três tipos de erro de mensuração 30
6.6 Viés Relativo da distribuição Gamma estimado através da estratégia de Bankier, sob efeito do Bootstrap nos três tipos de erros de mensuração 33
6.7 Viés Relativo da distribuição Uniforme estimado através da estratégia de Bankier, sob efeito do Bootstrap nos três tipos de erros de mensuração 34

Lista de Tabelas

4.1 Cenários possíveis em uma abordagem de Cadastro Duplo 10
4.2 Notação utilizada em abordagem de Cadastro Duplo 10

6.1 Desvio Padrão de todos os cenários simulados 18
6.2 Principais descritivas do estimador Horvitz-Thompson considerando a distribuição Uniforme. 20
6.3 Principais descritivas do estimador Horvitz-Thompson considerando a distribuição Gamma. 21
6.4 Principais descritivas do estimador Horvitz-Thompson considerando a distribuição Weibull. 22
6.5 Principais descritivas do estimador Hartley considerando a distribuição Gamma. 27
6.6 Principais descritivas do estimador Hartley considerando a distribuição Uniforme. 28
6.7 Principais descritivas do estimador Bankier considerando a distribuição Gamma. 31
6.8 Principais descritivas do estimador Bankier considerando a distribuição Uniforme. 32
6.9 Coeficiente de Variação dos estimadores utilizados na distribuição Gamma . . . 35
6.10 Coeficiente de Variação dos estimadores utilizados na distribuição Uniforme . . 36

Capítulo 1

Introdução

Um plano amostral probabilístico deve especificar o universo de investigação, as unidades amostrais, os critérios de estratificação, os procedimentos de sorteio das unidades amostrais, as probabilidades de inclusão, os estimadores e os respectivos erros amostrais. Sabendo assim, de que e de quem estamos falando e avaliando os desvios esperados para as estimativas (Bolfarine, 2005).

A estatística inferencial utilizada no processo de estimação necessita de planos amostrais a respeito de uma determinada população, pois garantem a seleção de elementos para compor a amostra através de um processo de aleatorização (Coelho, 2007). O uso inadequado de procedimentos amostrais podem levar a resultados inadequados e vícios, fugindo da realidade, ou seja, obtenção de amostras com resultados livres de vícios e confiáveis precisam ser estabelecidos para o uso cientifico dos processos amostrais e não são triviais (Medeiros, 2013). Para decidir qual o melhor plano amostral é necessário ter acesso a uma lista de elementos que compõe a população de interesse, o cadastro (Coelho, 2007). Na prática, geralmente, a realização de planos amostrais a tomando-se como base apenas um cadastro. Porém, outra metodologia que vem ganhando espaço na área de planejamento amostral é a abordagem de Cadastro Duplo.

É possível que todas as etapas referentes ao processo de seleção estejam corretas, porém os resultados obtidos podem estar errados não por conta de um erro amostral, mas sim de um erro não amostral como o erro de mensuração. Erro comum e difícil de tratar, que pode ocorrer em qualquer etapa da coleta de dados por meio de instrumentos inadequados até a identificação de duplicatas após a seleção dos elementos que irão compor a amostra, ocasionando problemas de estimação de médias, proporções e totais populacionais (Sarndal et al, 1992).

Nesta direção, a abordagem de Cadastro Duplo pode ser uma alternativa para minimizar os possíveis erros de mensuração. Os primeiros estudos referentes às técnicas de estimação sob a abordagem de cadastro duplo foram desenvolvidos por Hartley (1962). Desde então, vários outros autores passaram a dar grandes contribuições na área, entre eles Lund (1968), Fuller e Burmeister (1972) que apresentaram melhorias nos trabalhos desenvolvidos por Hartley (1962). Além disso, planos amostrais mais complexos foram desenvolvidos por Bankier (1986) e Kalton e Anderson (1986).

Dessa forma, é possível perceber que a abordagem de Cadastro Duplo vem se tornando uma ferramenta muito útil nas estratégias de estimação, além de possibilitar outras alternativas na estrutura de consideração das informações. A abordagem de Cadastro Duplo é definida como um levantamento amostral onde dois cadastros são utilizados para identificar elementos de uma única população-alvo. Nesta abordagem, amostras aleatórias independentes são selecionadas de cada cadastro, sem necessariamente ter que usar o mesmo esquema amostral em ambos (Coelho, 2011).

Dado o exposto, este projeto visa a continuidade do desenvolvimento de estratégias de estimação de efeitos sobre erros de mensuração em uma abordagem de Cadastro Duplo no contexto de métodos de estimação em populações finitas, buscando potencialidades de aplicação de planos amostrais para pesquisas reais que podem estar sujeitas a um ou mais erros de mensuração que podem ser considerados durante as etapas de seleção e observação dos dados.

Este Trabalho de Conclusão de Curso está dividida em 6 capítulos. Neste primeiro capítulo foi apresentada uma introdução a respeito da abordagem de Cadastro Duplo. No capítulo 1 são apresentados os objetivos gerais e específicos do trabalho. O capítulo 2 mostra os materiais e métodos utilizados o qual descreve os Métodos Probabilísticos, o Estimador Horvitz-Thompson, a Simulação Monte Carlo e Bootstrap. Já o capítulo 3 introduz os aspectos de estimação sob a abordagem de Cadastro Duplo, concentrando-se nos estimadores propostos por Hartley (1962) e Bankier (1989). O capítulo 4 traz considerações sobre a aplicação dos métodos da abordagem de um e dois cadastros, com erros de mensuração. O capítulo 5 apresenta os resultados dos estudos de simulação referentes a abordagem de Cadastro Duplo, onde é utilizado o método de Monte Carlo, tornando possível a comparação entre os estimadores propostos e descreve os possíveis benefícios da utilização de dois cadastros e da correção via Bootstrap. As considerações finais e propostas para trabalhos futuros são retratadas no capítulo 6. A linguagem de programação R constitue-se na plataforma computacional utilizada no desenvolvimento desta monografia.

Capítulo 2

Objetivos

2.1 Gerais

Apresentar métodos de estimação sob a abordagem de Cadastro Duplo adaptados à situação em que ocorrem erros de mensuração na variável resposta.

2.2 Específicos

1. Realizar um estudo comparativo de estimadores sob a abordagem de um cadastro e de Cadastro Duplo no contexto de Erros de Mensuração, bem como apresentar suas vantagens e desvantagens;

2. Apresentar recomendações técnicas sob erros de mensuração para estudos reais de acordo com o plano amostral probabilístico considerado.

3. Avaliar efeito do Bootstrap em dados contaminados com Erros de Mensuração.

Capítulo 3

Material e Métodos

Neste capítulo está descrito brevemente algumas definições necessárias para melhor compreensão deste trabalho.

3.1 Método Probabilístico de Amostragem

Nesse tipo de amostragem, os elementos da amostra são selecionados aleatoriamente e todos eles possuem probabilidade conhecida de serem escolhidos. A descrição de um plano amostral probabilístico deve especificar o universo de investigação, as unidades amostrais, os critérios de estratificação, os procedimentos de sorteio das unidades amostrais, as probabilidades de inclusão, os estimadores e os respectivos erros amostrais. Desse modo, saberemos do que e de quem estamos falando e avaliando os desvios esperados para as estimativas (Bolfarine, 2005). Alguns planos amostrais garantem uma seleção dos elementos de forma probabilística como: Amostragem Aleatória Simples (AAS), Amostragem Sistemática (AS) e Amostragem Estratificada (AE).

Porém, neste trabalho, utilizaremos apenas AAS sem reposição que é o método mais simples e mais importante para a seleção de uma amostra. Além de servir como um plano próprio, o seu procedimento é usado de modo repetido em procedimento de múltiplos estágios. Ele pode ser caracterizado através da definição operacional: de uma lista com *N* unidade elementares, sorteiam-se com igual probabilidade *n* unidades ($n < N$) (Bolfarine, 2005).

3.2 Estimador Horvitz-Thompson (HT)

O estimador Horvitz-Thompson (HT) é um estimador não viesado do total populacional que foi proposto para tratar de amostras retiradas sem reposição de um universo finito com probabilidades desiguais de seleção. No entanto, este estimador tem aplicação em qualquer plano amostral, com ou sem reposição.

Conforme Horvitz e Thompson (1952), quando estudamos uma população finita onde somos capazes de identificar seus elementos individualmente, podemos atribuir um vetor qualquer de probabilidades de seleção a esses elementos. Fazendo uma escolha adequada desse vetor de probabilidades é possível reduzir a variância de estimativas não viesadas se compararmos com aquelas obtidas utilizando-se probabilidades iguais de seleção.

De modo a entender esta estratégia, seja $U = 1, 2, 3...N$, o conjunto dos índices que identificam os elementos que compõem a população alvo, de tamanho N e $S = S_1, S_2, S_3, ...S_M$, todo o conjunto de todas as possíveis amostras de tamanho n, em que cada amostra S_i que pode

ser obtida de U_i com uma probabilidade denotada por $P(S_i)$, chamada de **probabilidade de seleção**.

Para cada elemento de U, temos também uma probabilidade de cada elemento ser incluído na amostra, ou seja, o vetor de probabilidades de inclusão de primeira ordem é dado dado por: $\bar{\Pi} = \pi_1, \pi_2, \pi_3, ..., \pi_N$. Cada uma das possíveis probabilidades de inclusão podem ser obtidas da seguinte forma:

$$\pi_k = \sum_{i=1}^{m(k)} P(S_i), \tag{3.1}$$

em que $m(k)$ representa o número de amostras que contém o elemento k. De modo semelhante, temos ainda que π_{kl} é chamado de **probabilidade de inclusão de segunda ordem**. Ou seja,

$$\pi_{kl} = P(k, l \in S) = P(I_k I_l = 1) = \sum_{i=1}^{m(k,l)} P(S_i) \tag{3.2}$$

em que $m(k, l)$ representa o número de amostras que contém os elementos k e l ao mesmo tempo. Ao todo, existem $\frac{N(N-1)}{2}$ probabilidades de inclusão de segunda ordem.

Considerando o total populacional como parâmetro de interesse, o estimador de Horvitz-Thompson, e sua respectiva variância, são dados pela seguinte expressão.

$$\hat{t}_\pi = \sum_{k \in S} \frac{y_k}{\pi_k}, \tag{3.3}$$

$$\text{Var}\left(\hat{t}_\pi\right) = \sum_{k \in U} \sum_{l \in U} (\pi_{kl} - \pi_k \pi_l) \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}. \tag{3.4}$$

onde y_k é um valor numérico observável.

Para o plano AAS, temos que:

$$P(S_i) = \frac{1}{\left(\frac{N}{n}\right)} \quad \text{e} \quad \pi_k = \frac{n}{N}. \tag{3.5}$$

Dessa forma,

$$\hat{t}_{\pi(AAS)} = \sum_{k \in S} \frac{y_k}{\left(\frac{n}{N}\right)} = N \times \sum_{k \in S} \frac{y_k}{n} = N\bar{y} \tag{3.6}$$

e tem como variância $N^2(1 - \pi_k)\frac{S^2_{y_U}}{n}$.

3.3 Simulação de Monte Carlo

A simulação de Monte Carlo é uma técnica matemática computadorizada que possibilita levar em conta o riso em análises quantitativas e tomadas de decisão. Analisando e resolvendo problemas que envolvem o uso de números aleatórios e probabilidades. Portanto, é um método de avaliação interativa de um modelo determinístico que utiliza números randomizados como entradas. Esse método é mais utilizado quando o modelo é complexo, ou não-linear, ou quando envolve um número razoável de parâmetros de incerteza (Lima et. al, 2008)

A rigor, a Simulação de Monte Carlo consiste em um experimento de amostragem, cujo objetivo principal está em estimar o comportamento de uma variável de resultado, que depende de outras variáveis aleatórias de forma automática, gerando uma grande quantidade de cenários, automaticamente (Nascimento, 2007, Matias Jr., 2006). A cada iteração, o resultado é armazenado e, ao final de todas as repetições, a sequência de resultados gerados é transformada em uma distribuição de frequência que possibilita calcular estatísticas descritivas.

Neste trabalho foi realizado um estudo via simulação Monte Carlo com o objetivo de estudar os Erros de Mensuração sob abordagem de Cadastro Duplo. Estudando através do estimador Horvitz Thompson, Hartley e Bankier a possibilidade de uma melhor estimativa quando utiliza-se a correção Bootstrap.

Neste trabalho o número de réplicas de Monte Carlo foi fixado em 1000 e todas as simulações foram realizadas utilizando a linguagem de programação `R`. A edição do texto foi realizada a através do sistema de tipografia LaTeX.

3.4 Bootstrap

O Método de Bootstrap foi introduzido em 1979 por Efron. Os métodos de Bootstrap são uma classe de métodos de Monte Carlo não-paramétricos que estimam a distribuição da população por reamostragem. A distribuição da população finita representada pela amostra pode ser encarada como uma pseudo-população, com características análogas às da verdadeira população.

Através da geração repetida de amostras aleatórias desta pseudo-população (reamostragem), a distribuição de amostragem de uma estatística pode ser estimada. O Bootstrap gera amostras aleatoriamente a partir da distribuição empírica da amostra.

Seja $x = (x_1, \ldots, x_n)$(eventualmente com repetições) uma amostra aleatória observada da fd $\mathrm{F}_X(\cdot)$. A função distribuição associada a X^* que atribui uniformemente

$$P(X^* = x) = \frac{1}{n} \tag{3.7}$$

é a chamada função distribuição empírica (fde), e denota-se por $F_n(\cdot)$.

O procedimento para estimar θ através do estimador $\hat{\theta}$, gerando amostras Bootstrap por reamostragem a partir de $x = (x_1, \cdots, x_n)$, é dado por:

Para cada réplica Bootstrap, indexada em $b = 1, 2, \cdots B$:

- Gerar amostra Bootstrap $x^{*(b)} = x_1^*, \cdots; x_n^*$ através da amostragem com reposição da amostra observada $x_1, \cdots, x_n$
- Calcular a b-ésima réplica $\hat{\theta}^{(b)}$ na amostra Bootstrap $x^{*(b)}$

A estimativa Bootstrap de $F_{\hat{\theta}}(\cdot)$ é a função distribuição empírica das réplicas $\theta^{(1)}, \cdots, \theta^{(n)}$ dada por:

$$F_n^*(x) = \frac{1}{B}\sum_{b=1}^{B} 1_{\{\hat{\theta}^{(b)} \leq x\}} \tag{3.8}$$

3.4.1 Estimação Bootstrap do viés

O viés de um estimador $\hat{\theta}$ de θ é definido como $vies(\hat{\theta})$ = $E[\theta] - \theta$: A estimação Bootstrap do viés usa as réplicas Bootstrap de θ para estimar a distribuição de amostragem de θ.

$$\hat{vies}^*(\hat{\theta}) = \bar{\hat{\theta}}^* - \theta \tag{3.9}$$

com

$$\bar{\hat{\theta}}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{(b)} \tag{3.10}$$

e

$$\hat{\theta} = \hat{\theta}(x) = \hat{\theta}(x_1, \cdots, x_n) \tag{3.11}$$

3.4.2 Estimação Bootstrap do desvio padrão

A estimação Bootstrap do desvio padrão de um estimador $\hat{\theta}$ é o desvio padrão empírico das réplicas Bootstrap $\hat{\theta}^{(1)}, \cdots, \hat{\theta}^{(B)}$.

$$\hat{dp}^*(\hat{\theta}) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}^{(b)} - \bar{\hat{\theta}}^{(*)})^2} \tag{3.12}$$

com

$$\bar{\hat{\theta}}^{(*)} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{(b)} \tag{3.13}$$

Todas as simulações foram realizadas utilizando mil réplicas de Monte Carlo e 500 réplicas de Bootstrap, totalizando assim meio milhão de réplicas por experimento.

Capítulo 4

A abordagem de Cadastro Duplo

A consideração de planos amostrais para se realizar inferência a respeito de uma determinada população de interesse é de extrema importância, pois garantem a seleção de elementos para compor a amostra através de um processo de aleatorização que permite utilizar a teoria da inferência estatística. Para decidir sobre que tipo de plano amostral deve ser utilizado, é necessário ter acesso a uma lista de elementos que compõem a população de interesse, a qual é chamada de cadastro (Coelho,2007).

Em estudos mais sofisticados, é possível que todas as etapas referentes à forma da seleção da amostra a partir de um cadastro estejam corretas, porém todos os resultados obtidos podem estar errados não por conta de um erro amostral gerado pelo plano amostral, e sim devido a um erro não amostral, que é o erro de cobertura, ou seja, elementos que fazem parte da população alvo podem não estar incluídos no cadastro que se tem disponível, e nessa direção, a abordagem de Cadastro Duplo é uma alternativa para minimizar o erro de cobertura. A abordagem de Cadastro Duplo é definida como um levantamento amostral onde dois cadastros, denotados por A e B, são utilizados para identificar elementos de uma única população-alvo. Nesta abordagem, amostras aleatórias independentes, denotadas por S_A e S_B, são obtidas de cada cadastro, respectivamente, através de planos amostrais com probabilidade $p(S_A)$ e $p(S_B)$ possivelmente diferentes.

Existem inúmeras situações onde é possível utilizar a abordagem de cadastro duplo. Por exemplo, quando um dos cadastros não tem um grau de cobertura desejável da população-alvo, porém existe um segundo cadastro, o qual, em conjunto com o primeiro, fornece cobertura completa. Outra situação pode ser exemplificada por um cadastro que cobre por completo a população-alvo, mas tem alto custo de seleção da amostra. Nesse caso, se houver outro cadastro disponível e de abrangência menor que o primeiro, é possível utilizar ambos os cadastros de modo a tornar vantajosa a relação custo-benefício de seleção de elementos. Estas situações são representadas pelas figuras a seguir.

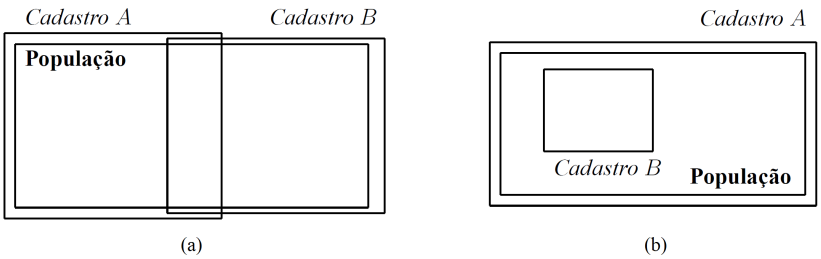

Figura 4.1: ((a) Cadastro duplo fornecendo cobertura completa para a população-alvo; (b) Cadastro Duplo utilizado para melhorar a relação custo benefício)

A abordagem de Cadastro Duplo é um caso particular da abordagem de cadastros múltiplos em que F cadastros são utilizados $(F \geq 2)$. Em geral, tal abordagem implica a existência de $2^F - 1$ domínios possíveis de serem identificados.

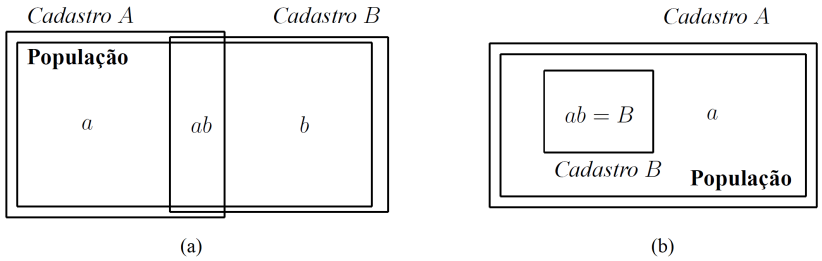

Figura 4.2: Domínios a, b e ab induzidos pela abordagem de Cadastro Duplo

Considere que estão disponíveis dois cadastros, A e B, fornecendo a cobertura para uma população-alvo. Para a implementação da abordagem de Cadastro Duplo, duas condições são necessárias:

1. Todos os elementos da população-alvo devem ser identificados pela união dos cadastros;
2. Pode-se verificar se qualquer elemento de um cadastro pertence ou não ao outro.

Denotando por $U = U_A \cup U_B$ o conjunto de elementos da população-alvo e considerando a notação de domínios representada na figura (4.2), Hartley (1962) faz uma descrição dos possíveis cenários gerados por um levantamento amostral realizado através da abordagem de Cadastro Duplo. Estes cenários dependem basicamente da disponibilidade de informações populacionais, como mostra a tabela (4.1) a seguir:

Tabela 4.1: Cenários possíveis em uma abordagem de Cadastro Duplo

Tipo de Informação Disponível	CENÁRIOS			
	1	**2**	**3**	**4**
Tamanho dos domínios e dos cadastros	Tamanhos dos domínios e cadastros conhecidos	Tamanhos dos domínios e cadastros conhecidos	Apenas os tamanhos dos cadastros conhecidos	Apenas as magnitudes relativas dos cadastros conhecidas
Possibilidade de alocação da amostra	Alocação de amostra aos domínios	Alocação de amostra aos cadastros	Alocação de amostra aos cadastros	Alocação de amostra aos cadastros

O cenário 1 permite que a amostra seja alocada a cada domínio, o que ilustra o caso de um plano sob estratificação. Os cenário 2, 3 e 4 admitem a alocação da amostra apenas aos cadastros.

No cenário 3, apenas os tamanhos populacionais nos cadastros são conhecidos, enquanto o cenário 4 apresenta como única informação disponível apenas o tamanho relativo dos cadastros, tornando possível apenas a estimação de médias populacionais. A Tabela 4.2 apresenta a notação utilizada nesta estratégia.

Tabela 4.2: Notação utilizada em abordagem de Cadastro Duplo

Quantidades de Interesse	CADASTRO		Domínio		
	A	B	a	b	ab
População	$\mathcal{A}$	$\mathcal{B}$	$\mathcal{A}$	$\mathcal{B}$	$A \cap B$
Tamanho da População	N_A	$N_{\mathcal{B}}$	N_a	N_b	N_{ab}
Total populacional	$t_{y\mathcal{A}}$	$t_{y\mathcal{B}}$	t_{ya}	t_{yb}	t_{yab}
Média populacional	$\mu_{y\mathcal{A}}$	$\mu_{y\mathcal{B}}$	μ_{ya}	μ_{yb}	μ_{yab}
Amostra	S_A	S_B	S_a	S_b	S_{ab}
Tamanho da amostra	$n_{\mathcal{A}}$	$n_{\mathcal{B}}$	n_a	n_b	$n_{ab(A)}$ $n_{ab(B)}$
Estimador do Total	$\widehat{t}_{y\mathcal{A}}$	$\widehat{t}_{y\mathcal{B}}$	$\widehat{t}_{ya}$	$\widehat{t}_{yb}$	$\hat{t}^A_{yab}$ $\hat{t}^B_{yab}$
Média amostral	$\overline{y}_A$	$\overline{y}_B$	$\tilde{y}_a$	$\tilde{y}_b$	$\tilde{y}^A_{ab}$ $\tilde{y}^B_{ab}$

Denota-se que $U_a = U_A \cup U_B^c, U_b = U_A^c \cup U_B$ e $U_{ab} = U_A \cup U_B$. As quantidades n'_{ab}, $\hat{t}'_{yab}$ e $\bar{y}'_{ab}$ são referentes a elementos na amostra que pertencem a U_{ab} e que foram obtidas do cadastro A. As quantidades n''_{ab}, $\hat{t}''_{yab}$ e $\bar{y}''_{ab}$ são definidas analogamente e estão relacionadas ao cadastro B.

Dessa forma, dentre outras relações identificáveis, tem-se que:

- $N_A = N_a + N_{ab}$;
- $N_B = N_b + N_{ab}$ e
- $N = N_a + N_b + N_{ab} = N_a + N_B = N_A + N_b$.

As quantidades na tabela (4.2), referentes aos totais populacionais, são definidas da seguinte forma:

$t_y = \sum_{k \in U} y_k$ é o total populacional da variável de interesse na população;

$t_{yA} = \sum_{k \in U_A} y_k$ é o total populacional da variável de interesse y, no cadastro A;

$t_{yB} = \sum_{k \in U_B} y_k$ é o total populacional da variável de interesse y, no cadastro$B$;

$t_{ya} = \sum_{k \in U_a} y_k$ é o total populacional da variável de interesse y, em U_a;

$t_{yb} = \sum_{k \in U_b} y_k$ é o total populacional da variável de interesse y, em U_b;

$t_{yab} = \sum_{k \in U_{ab}} y_k$ é o total populacional da variável de interesse y, em U_{ab}.

É possível observar que $t_y = t_{ya} + t_{yab} + t_{yb}$.

4.1 Estratégia de Hartley

O método de estimação desenvolvido por Hartley (1962) foi proposto com base em uma variável y_k para cada cadastro, e pode ser definida da seguinte forma:

$$y_k^* = \begin{cases} y_k, & \text{se } k \in a \text{ ou } k \in b \\ py_k, & \text{se } k \in ab \text{ e } k \in \mathcal{A} \\ (1-p)y_k, & \text{se } k \in ab \text{ e } k \in \mathcal{B}, \end{cases}$$

em que p é uma constante de ponderação para identificação dos elementos de cada cadastro na população, tal que $0 \leq p \leq 1$. Dessa forma , é possível reescrever o total populacional, t_y, da seguinte maneira:

$$\begin{aligned} t_y &= t_{ya} + t_{yb} + t_{yab} \\ &= t_{ya} + t_{yb} + (p + 1 - p)t_{yab} \\ &= t_{ya} + pt_{yab} + t_{yb} + (1-p)t_{yab} \\ &= \sum_{k \in \mathcal{A}} y_{kA}^* + \sum_{k \in \mathcal{B}} y_{kB}^* \\ &= t_{y\mathcal{A}}^* + t_{y\mathcal{B}}^*. \end{aligned} \tag{4.1}$$

Hartley enfatiza o fato de que t_y pode ser estimado utilizando estimadores para cada um dos domínios a, b e ab. É possível também representar t_y ou a média populacional, μ, da seguinte forma:

$$t_y = t_{yA} + t_{yB} - t_{yab} = t_{yA} + t_{yB} - (pt_{yab} + (1-p)t_{yab}) \tag{4.2}$$

$$\mu = N^{-1}t_y = N^{-1}\left(t_{yA} + t_{yB} - t_{yab}\right). \tag{4.3}$$

Considere o problema de obter um estimador para t_y, sob a situação do cenário 2, onde há disponibilidade de informação sobre os tamanhos populacionais de cada domínio (N_a, N_b e N_{ab}). Considere ainda que em cada cadastro foi aplicado um plano AAS, obtendo amostras denotadas por S_A e S_B, de tamanhos n_A e n_B, referindo-se aos cadastros A e B respectivamente. Nestas condições, o estimador proposto por Hartley (1962) assume a seguinte forma:

$$\hat{t}_{yH} = N_{a\tilde{y}a} + N_{ab}(p\tilde{y}'_{ab} + ((1-p)\tilde{y}''_{ab}) + N_{b\tilde{y}b}, \tag{4.4}$$

onde:

$\tilde{y}_a$ é a Média amostral do domínio a, a partir de elementos obtidos do cadastro A;

$\tilde{y}_b$ é a Média amostral do domínio b, a partir de elementos obtidos do cadastro B;

$\tilde{y}'_{ab}$ é a Média amostral do domínio ab, a partir de elementos obtidos do cadastro$A$;

$\tilde{y}''_{ab}$ é a Média amostral do domínio ab, a partir de elementos obtidos do cadastro B.

4.2 Estratégia de Bankier

A estratégia de estimação sob a abordagem de Cadastro Duplo adotada por Bankier, Lepkowski e Groves (1986), descrita aqui simplesmente por estratégia BK, atribui pesos, às unidades amostrais de A e B. A ideia dos autores foi adotar uma abordagem que envolvesse apenas quantidades amostrais referentes aos cadastros A e B, como mostra a figura a seguir.

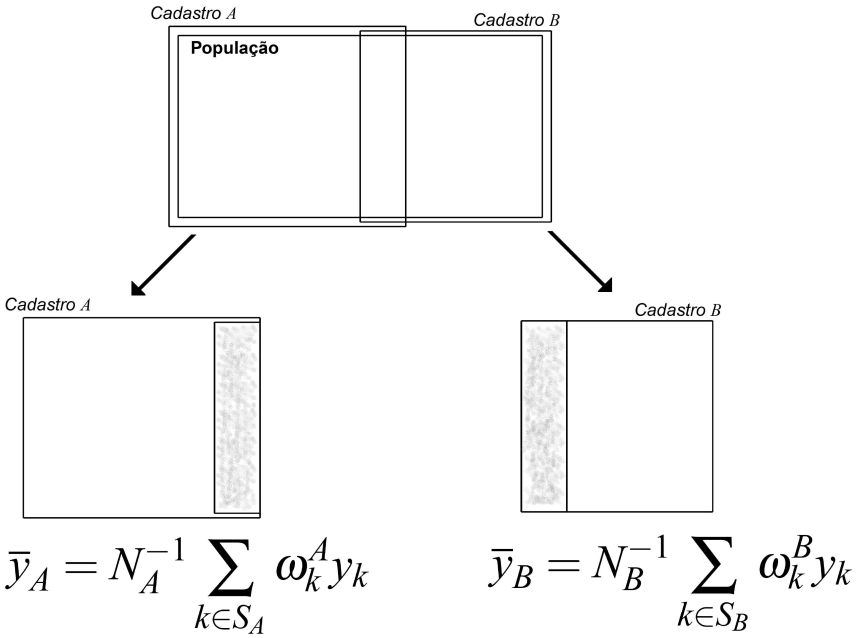

Figura 4.3: Quantidades amostrais geradas pela estratégia BK

Denotando por w_k^c o peso associado a um cadastro C, tem-se que:

$$w_k{}^c = \begin{cases} \frac{1}{\pi_k{}^A}, k \epsilon a \\ \frac{1}{\pi_k{}^B}, k \epsilon b \\ \frac{1}{\pi_k{}^A + \pi_k{}^B}, k \epsilon ab \end{cases} \tag{4.5}$$

Quando há o interesse em estimar por exemplo uma média populacional denotada por μ , tem se que o estimador BK é dado por:

$$\begin{aligned} \bar{Y}_{BK} &= N^{-1} \left(\sum_{k \epsilon S_A} w_k^A yk + \sum_{k \epsilon S_B} w_k^B yk \right) = w_A \frac{\sum_{k \epsilon S_A} w_k^A yk}{N_A} + w_B \frac{\sum_{k \epsilon S_B} w_k^B yk}{N_B} \\ &= N^{-1} \left(\sum_{k \epsilon a} w_k y_k + \sum_{k \epsilon b} w_k y_k + \sum_{k \epsilon ab} w_k y_k \right) = w_a \bar{y}_a + w_b \bar{y}_b + w_{ab} \bar{y}_{ab} \end{aligned} \tag{4.6}$$

com $w_A = \frac{N_A}{N}, W_B = \frac{N_B}{N}, w_a = \frac{N_a}{N}, w_b = \frac{N_b}{N}$ e $w_{ab} = \frac{N_{ab}}{N}$.

A variância de $\bar{y}BK$ é obtida a partir da expressão $Var(\bar{\theta}) = \mathbf{d}^T \sum \mathbf{d}$, onde:

$$d = \begin{bmatrix} 1 \\ 1 \end{bmatrix} e \sum_{BK} = \begin{bmatrix} \sigma^2_{yA} & 0 \\ 0 & \sigma^2_{yB} \end{bmatrix} \tag{4.7}$$

Na matriz $\sum_{BK}\ \sigma^2_{yA} = Var(\sum_{k\epsilon\mathcal{A}} w^A_k yk)$ e o $\sigma^2_{yB} = Var(\sum_{k\epsilon\mathcal{B}} w^B_k yk)$. É possível obter um estimador não-viesado para a variância do estimador de Bankier, substituindo em $\sum_{BK}$ as quantidades populacionais pelos seus respectivos estimadores de Horvitz-Thompson.

Bankier (1986) propôs ainda um estimador iterativo chamado *raking ratio estimator*, dado por :

$$\bar{y}_{BK}(r) = N^{-1}\left(\sum_{k\epsilon S_A} w^{(r)}_{ka} y_k + \sum_{k\epsilon S_B} w^{(r)}_{kb} y_k + \sum^{(r)}_{k\epsilon S_a b} w^{(r)}_{kab} y_k\right) \tag{4.8}$$

A ideia é que se façam r ajustes à constante W_k, onde

$$w^{(r)}_{ka} = \frac{N_a}{\widehat{N}^{A(r-1)}} w^{(r-1)}_{ka}, w^{(r)}_{kb} = w^{(r-1)}_{kb}, w^{(r)}_{kab} = \frac{N_a}{\widehat{N}^{A(r-1)}} w^{(r-1)}_{kab} \quad \text{para r} = 1, 3, 5, ... \tag{4.9}$$

$$w^{(r)}_{ka} = w^{(r-1)}_{ka}, w^{(r)}_{kb} = \frac{N_b}{\widehat{N}^{A(r-1)}_v} w^{(r-1)}_{kb}, w^{(r)}_{kab} = \frac{N_b}{\widehat{N}^{B(r-1)}_v} w^{(r-1)}_{kab} \quad \text{para r} = 2, 4, 6, ...$$

$$w^{(0)}_{ka} = \frac{1}{\pi^A_k}, \quad w^{(0)}_{kb} = \frac{1}{\pi^B_k}, \quad w^{(0)}_{kab} = \frac{1}{\pi^A_k + \pi^B_k}$$

$$\bar{N}_{A(r)} = w^{(r)}_{ka} N_a + w^{(r)}_{kab} N_{ab} \quad \text{e} \quad \bar{N}_{B(r)} = w^{(r)}_{kb} N_b + w^{(r)}_{kab} n_{ab}$$

O estimador $\bar{y}_{BK(r)}$ é obtido fazendo $\bar{y}_{BK(0)} = \bar{y}_{BK}$ e aplicando r ajustes com respeito aos domínios, alternadamente. Como para um determinado valor de r as expressões para a variância do estimador $\bar{y}_{BK(r)}$ são bastante complexas (Bankier, Lepkowski e Groves (1986); Brackstone e Rao (1979)), Skinner (1991) provou que quando o número de ajustes é suficientemente grande, ou seja, quando $r \to \infty$, tem-se que

$$\bar{y}_{BK(\infty)} = N^{-1}(N_A - \widetilde{N}_{ab})\bar{y}_a + \widetilde{N}_{ab}\bar{y}_{ab} + (N_B - \widetilde{N}_{ab})\bar{y}_b \tag{4.10}$$

Logo, a variância aproximada do estimador raking é dada por

$$AVar(\bar{y}_{BK(\infty)}) = \left(\frac{1}{\pi^A_k}\right)^{-1} N_a \sigma^2_a + \left(\frac{1}{\pi^B_k}\right)^{-1} N_b \sigma^2_b + (\pi^A_k + \pi^B_k)^{-1} N_{ab} \sigma^2_{ab} + (\mu_a + \mu_b - \mu_{ab})^2 \frac{N_{ab} N_a N_b}{n_A N_b + n_B N_a}(1 + \lambda^2), \tag{4.11}$$

onde μ_a, μ_b e μ_{ab} são as médias populacionaisnde cada domínio e $\sigma^2_a, \sigma^2_b, \sigma^2_{ab}$ são as variâncias populacionais de cada domínio, e

$$\lambda^2 = \frac{N_a N_b (N_{ab})^2 [(\pi^A_k)^2 N_A - (\pi^B_k)^2 N_B]^2}{N_A N_B \pi^A_K \pi^B_K (\pi^A_K + \pi^B_K)^2 (N^2_{ab} - N_A N_B)^2} \tag{4.12}$$

Capítulo 5

Erro de mensuração

O processo de estimação de médias, totais, ou proporções populacionais em diversas situações pode estar sujeito ao que se define como erro de mensuração na área de amostragem. Este tipo de erro pode ocorrer em qualquer etapa da fase de coleta de dados, que vai desde a medição por meio de um instrumento inadequado até a identificação de duplicatas após a seleção dos elementos que irão compor a amostra, algo bastante comum em uma abordagem de Cadastro Duplo.

Neste trabalho utilizaremos a suposição de que cada elemento pertencente aos conjuntos U_a ou U_b selecionado, para compor as amostras de um dos cadastros, esta associado a um valor numérico observável, denotado por y_k. Este valor será observado com exatidão, desde que este elemento seja selecionado e sujeito a observação (mensurado). Esta observação exata será definida como valor verdadeiro de medição da variável resposta para cada elemento da população de interesse nos cadastros$A$ eB.

Em uma abordagem de Cadastro Duplo, $\theta_{k(A)}$ e $\theta_{k(B)}$ denotarão os valores de medição para cada um dos elementos presentes nas amostras obtidas dos cadastros A e B, respectivamente. O esperado é que em todo levantamento amostral por meio de uma abordagem de cadastro duplo o procedimento de coleta seja realizado de modo que os valores verdadeiros de medição sejam observados. Porém este procedimento pode gerar erros, de modo que os valores mensurados nas amostras de cada cadastro terão uma distancia de afastamento do valor que seria observado caso não houvesse a presença de erros de mensuração no processo. Seja $d_{k(C)} = y_k - \theta_{k(C)}$, para todo K pertencente a U_C, em que C pertence $[A, B]$. Assim, $d_{K(C)}$ é denotada como a diferença entre o valor observável e o valor verdadeiro em cada uma das amostras obtidas dos cadastros$A$ e B. Esta diferença é chamada de erro de mensuração para o K-ésimo elemento do conjunto U_C.

Dado o exposto, este projeto visa a continuidade do desenvolvimento de estratégias de estimação de efeitos sob erros de mensuração em uma abordagem de Cadastro Duplo no contexto de métodos de estimação em populações finitas, buscando potencialidades de aplicação de planos amostrais para pesquisas reais que podem estar sujeitas a um ou mais erros de mensuração durante as etapas de seleção e observação de dados.

Estas considerações serão necessárias porque em diversas pesquisas por amostragem a forma de seleção da amostra a partir de um cadastro, mesmo que correta, podem conter resultados comprometidos, não por um erro amostral, gerado pelo plano amostral, ou por erro gerado na seleção dos elementos, mas devido a um ou mais erros não amostrais que podem resultar em erros de mensuração. Dessa forma, o presente projeto tem também a intenção de estender resultados já desenvolvidos e publicados na literatura sobre efeitos de erros de mensuração sob a ótica de apenas um cadastro, fornecidas por Mahalanobis(1946), Hansen et al. (1961), Hansen

et al. (1964), Bailar e Dalenius(1969) e Sãrndal et al(2003).

Para o cálculo e avaliação dos estimadores combinados, foram considerados o Viés, Erro Quadrático Médio, Viés Relativo, Desvio Padrão, Coeficiente de Variação e Intervalo de confiança dos estimadores do total populacional em cada cenário descrito na tabela (6.1).

O Viés de um estimador θ é dado por:

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

onde θ é o parâmetro de interesse e $\hat{\theta}$ é o estimador de θ. O viés estimado e o viés relativo estimado são obtidos a partir da estimativa de $E(\hat{\theta})$, via Monte Carlo.

O Erro Quadrático Médio é dado por:

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

O Viés Relativo é dado por:

$$VR(\hat{\theta}) = \frac{|E(\hat{\theta}) - \theta|}{\theta},$$

O Desvio Padrão é dado por:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} \qquad CV = \frac{\sigma}{\bar{x}}$$

x_i: valor na posição i no conjunto de dados
$\bar{x}$: média aritmética dos dados
n:quantidade de dados
CV: Coeficiente de Variação

Estimação por Intervalos de Confiança:

Também denominada Estimação Intervalar, consiste em determinar um intervalo numérico. A este intervalo está associada uma probabilidade (β) de que o mesmo contenha o valor efetivo do parâmetro estimado. A referida probabilidade é denominada "*Nível de Confiança*". O valor $(1 - \beta)$ também é conhecido como "*Margem de Erro*".

Níveis de Confiança e Valores Críticos:

A primeira providência a ser tomada na construção de um intervalo de confiança para estimar a média é verificar se o desvio padrão populacional (σ) é conhecido. Se a resposta for positiva, os valores críticos são tomados com base na Distribuição Normal e os limites do intervalo são calculados da seguinte forma:

$$(\bar{x} - Z_c\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + Z_c\frac{\sigma}{\sqrt{n}})$$

Foi utilizado nível de confiança (β) de 95% então seu valor crítico $(Z_c) = 1,96$.

Capítulo 6

Resultados e Discussão

O desempenho dos estimadores propostos foi avaliado através de um estudo de simulação de Monte Carlo, com número de réplicas fixado em 1000 e realizadas 500 reamostras através do Bootstrap. Para isto foram gerados cadastros de tamanho mil das distribuições de probabilidade (Uniforme(0,1), Gamma(2,1) e Weibull(2,1)) e acrescendo a cada um deles três tipos de erros, um erro pequeno, médio e grande, com distribuições Normal(0.001,0.01), Normal(0.1,0.01) e Normal(0.5,0.01), respectivamente.

Esses erros acrescidos simulam o comportamento do cadastro quando este contém Erros de Mensuração, porém o interesse do trabalho é estudar até que ponto esse tipo de erro pode ser corrigido. Dessa forma, os cadastros iniciais foram contaminados com os erros em diferentes proporções (25%, 50%, 75% e 100%).

Já no caso onde utilizamos Cadastro Duplo, apenas o cadastro A foi contaminado com os erros de mensuração e as distribuições utilizadas foram Uniforme(0,1) e Gamma(2,1). Para ambos os casos foi sorteado uma amostra de tamanho 100 para calcular as estimativas e uma amostra Bootstrap para avaliar a intensidade e o efeito do erro de mensuração nas diferentes situações simuladas, através de medidas descritivas como Viés, Erro Quadrático Médio(EQM) e Coeficiente de Variação (CV). Para o cálculo dos diversos intervalos de confiança foi utilizada uma confiança de 95%.

Na utilização de um Cadastro as estimativas foram obtidas através do estimador Horvitz Thompson, já para Cadastro Duplo estas foram obtidas através das estratégias de Hartley e Bankier.

A Tabela (6.1) ilustra todos os cenários simulados e seus respectivos desvios padrão, facilitando assim a visualização de todas as simulações realizadas neste trabalho. Nota-se que o desvio padrão tende a aumentar com o aumento da contaminação em todas as distribuições. É possível verificar que, para Gamma(2,1) por exemplo, a medida que os Erros de Mensuração aumentam seus desvios padrão também crescem. Para as demais estimativas, o comportamento observado foi o mesmo, o que evidencia que o nível de contaminação e tamanho dos Erros de Mensuração influenciam na estimação.

É importante ressaltar que a utilização do Bootstrap como tentativa de correção das estimativas, mostrou-se indiferente quando os dados estão contaminados com Erros de Mensuração. Os desvios padrão avaliados na tabela (6.1) são praticamente iguais com ou sem o método de reamostragem. Na distribuição uniforme(0,1) por exemplo, com contaminação de 25% e erro pequeno, os desvios são 26,876 e 26,864 sem e com correção Boostrap respectivamente.

Tabela 6.1: Desvio Padrão de todos os cenários simulados

Cenários										
		Horvitz Thompson			Hartley			Bankier		
Distrib.	Cont.	Erros de Mensuração								
		N(0.001,0.01)	N(0.1,0.01)	N(0.5,0.01)	N(0.001,0.01)	N(0.1,0.01)	N(0.5,0.01)	N(0.001,0.01)	N(0.1,0.01)	N(0.5,0.01)
Uniforme(0,1)	25%	26,876	28,893	35,101	33,888	36,417	35,434	47,219	0,721	59,546
	50%	27,144	28,723	36,193	36,192	34,976	35,955	47,637	51,966	62,677
	75%	28,118	29,792	34,522	37,184	37,085	35,833	47,314	48,772	61,984
	100%	26,724	27,585	27,618	35,301	35,462	35,306	46,317	50,128	62,948
Gamma(2,1)	25%	82,812	95,888	102,645	175,673	171,085	168,353	207,857	216,995	234,258
	50%	128,007	134,783	136,605	184,244	168,944	169,344	216,905	219,434	235,916
	75%	132,334	142,379	131,727	189,600	169,153	171,823	213,411	227,436	229,588
	100%	131,931	139,351	124,498	175,854	177,821	172,242	212,909	218,191	237,891
Weibull(2,1)	25%	36,451	37,345	48,452	-	-	-	-	-	-
	50%	45,129	44,011	52,503	-	-	-	-	-	-
	75%	44,189	46,608	57,544	-	-	-	-	-	
	100%	46,788	45,109	41,638	-	-	-	-	-	-
Bootstrap										
Uniforme(0,1)	25%	26,864	28,895	35,104	33,898	36,328	35,522	50,112	49,106	48,332
	50%	27,125	28,742	36,209	36,212	34,845	35,925	51,620	50,369	50,475
	75%	28,119	29,769	34,528	37,040	37,010	35,740	53,014	50,616	49,854
	100%	26,712	27,597	27,618	35,268	35,689	35,332	44,483	45,261	46,929
Gamma(2,1)	25%	82,810	95,932	102,711	196,384	190,720	187,242	222,702	217,839	215,731
	50%	128,038	134,818	136,672	203,366	183,330	182,937	223,718	222,932	216,049
	75%	132,315	142,457	131,727	209,190	186,512	191,496	223,718	222,932	216,049
	100%	131,934	139,288	124,489	195,059	195,913	187,979	216,847	220,105	207,057
Weibull(2,1)	25%	36,432	37,338	48,531	-	-	-	-	-	-
	50%	45,173	44,010	52,501	-	-	-	-	-	-
	75%	44,187	46,620	57,534	-	-	-	-	-	-
	100%	46,802	45,107	41,691	-	-	-	-	-	-

6.1 Estimação de totais populacionais

O Desempenho dos estimadores propostos sob o plano de Amostragem Aleatória Simples sob os cenários de contaminação, é apresentado a seguir pelas tabelas 6.2 a 6.4 e figuras 6.1 a 6.3. No estudo de Simulação o interesse era estimar o total populacional e para os intervalos de confiança construídos os valores dos parâmetros foram (505,27,2010,26 e 899,26) respectivamente para as distribuições Uniforme(0,1), Gamma(2,1) e Weibull(2,1).

A partir da distribuição Uniforme(0,1) (Tabela 6.2) nota-se que em todos os diferentes tipos de contaminação não há erros significantes, dado que praticamente, todos os intervalos contém o verdadeiro valor do parâmetro, com exceção do caso quando o erro é grande ou a contaminação é de 100%. É possível notar que o uso do método Bootstrap, implementado com intuito de corrigir o viés do estimador, não apresentou evidências de bom desempenho em situações nas quais os dados estão contaminados com Erro de Mensuração, resultando em vieses próximos quando comparados as estimativas sem a utilização do método.

Na (Tabela 6.3), com distribuição Gamma(2,1) observamos que a contaminação de 25% teve o pior comportamento, dado que tiveram altos EQM e não incluíram o verdadeiro valor do parâmetro no intervalo de confiança e o Bootstrap não conseguiu corrigir o problema. Na distribuição Weibull (Tabela 6.4) foi possível observar novamente que o uso do método Bootstrap para este cenário não foi capaz de fornecer uma estimativa precisa do parâmetro de interesse, distanciando-se do verdadeiro valor do parâmetro e consequentemente aumentando o Viés e o EQM.

Este método também se mostrou muito sensível a erros grandes (N(0.5,0.01)), onde suas estimativas e seus EQMs aumentam em proporções altas e seus intervalos não contem o verdadeiro valor do parâmetro, dando indícios de que o processo de reamostragem apenas aumentou o número de amostras contaminadas.

De modo geral, é possível observar que o estimador Horvitz Thompsom não obteve bons resultados, dado que a cobertura do intervalo de confiança foi pequena e a variabilidade demonstrada através do Coeficiente de Variação foi alta. Observando-se os diferentes cenários tem-se que o pior desempenho foi obtido quando utilizado a distribuição Gamma e o melhor com Uniforme, como mostram os valores de desvio padrão, viés relativo e EQM.

Tabela 6.2: Principais descritivas do estimador Horvitz-Thompson considerando a distribuição Uniforme.

	Distribuição Uniforme(0,1)								
Cont.	Erros de Mens.	Parâmetro(505,27)							
		Media	Vies	Vies. R	Desvio	EQM	CV	Lim. Inf.	Lim. Sup.
25%	$N(0.001,0.01)$	509,436	4,164	0,824	26,876	739,661	0,053	456,759	562,112*
	$N(0.1,0.01)$	524,248	18,977	3,756	28,893	1194,918	0,055	467,618	580,878*
	$N(0.5,0.01)$	633,291	128,020	25,337	35,101	17621,220	0,055	564,489	702,093
50%	$N(0.001,0.01)$	520,043	14,771	2,923	27,144	954,964	0,052	466,841	573,244*
	$N(0.1,0.01)$	555,364	50,092	9,914	28,723	3334,216	0,052	499,068	611,659*
	N(0.5,0.01)	751,266	245,994	48,686	36,193	61823	0,048	680,33	822,201
75%	$N(0.001,0.01)$	509,485	4,213	0,834	28,118	808,397	0,055	454,373	564,596*
	$N(0.1,0.01)$	563,086	57,814	11,442	29,792	4230,050	0,053	504,693	621,478*
	$N(0.5,0.01)$	869,528	364,257	72,091	34,522	133874,900	0,040	801,859	937,197
100%	$N(0.001,0.01)$	511,588	6,316	1,250	26,724	754,042	0,052	459,209	563,965*
	$N(0.1,0.01)$	588,075	82,803	16,388	27,585	7617,267	0,047	534,009	642,141
	$N(0.5,0.01)$	1003,799	498,528	98,665	27,618	249292,900	0,028	949,655	1057,943
25%**	$N(0.001,0.01)$	509,430	4,159	0,823	26,864	738,973	0,053	456,777	562,083*
	$N(0.1,0.01)$	524,222	18,951	3,751	28,895	1194,043	0,055	467,589	580,856*
	$N(0.5,0.01)$	633,310	128,038	25,34	35,104	17626	0,055	564,506	702,113
50%**	$N(0.001,0.01)$	520,054	14,782	2,926	27,125	954,247	0,052	466,89	573,217*
	$N(0.1,0.01)$	555,360	50,089	9,913	28,742	3334,992	0,052	499,025	611,695*
	$N(0.5,0.01)$	751,280	246,008	48,688	36,209	61831	0,048	680,311	822,248
75%**	$N(0.001,0.01)$	509,473	4,201	0,832	28,119	808,303	0,055	454,361	564,585*
	$N(0.1,0.01)$	563,070	57,798	11,439	29,769	4226,814	0,053	504,723	621,416*
	$N(0.5,0.01)$	869,533	364,262	72,092	34,528	133879	0,040	801,855	937,212
100%**	$N(0.001,0.01)$	511,590	6,318	1,250	26,712	753,471	0,052	459,233	563,945*
	$N(0.1,0.01)$	588,067	82,796	16,386	27,597	7616,764	0,047	533,974	642,160
	$N(0.5,0.01)$	1003,807	498,535	98,667	27,618	249299,900	0,028	949,692	1057,922

** Com correção Bootstrap.* Intervalo de confiança contém o verdadeiro valor do parâmetro.

Tabela 6.3: Principais descritivas do estimador Horvitz-Thompson considerando a distribuição Gamma.

	Distribuição Gamma(2,1)								
Cont.	Erros de Mens.	Parâmetro(2010,26)							
		Media	Vies	Vies. R	Desvio	EQM	CV	Lim. Inf.	Lim. Sup.
25%	N(0.001,0.01)	855,895	-1154,370	57,424	82,812	1339421,000	0,097	693,599	1018,192
	N(0.1,0.01)	917,864	-1092,400	54,341	95,888	1202530,000	0,104	729,912	1105,817
	N(0.5,0.01)	979,865	-1030,400	51,257	102,645	1072256,000	0,105	778,677	1181,053
50%	N(0.001,0.01)	1943,264	-66,998	3,333	128,007	20874,610	0,066	1692,371	2194,158*
	N(0.1,0.01)	2096,925	86,662	4,311	134,783	25676,790	0,064	1832,751	2361,100*
	N(0.5,0.01)	2226,813	216,550	10,772	136,605	65554,820	0,061	1959,066	2494,559*
75%	N(0.001,0.01)	1619,764	-390,499	19,425	132,334	170001,800	0,082	1360,387	1879,142
	N(0.1,0.01)	1748,039	-262,224	13,044	142,379	89033,270	0,081	1468,976	2027,102*
	N(0.5,0.01)	1903,625	-106,638	5,305	131,727	28723,550	0,069	1645,441	2161,808*
100%	N(0.001,0.01)	1978,656	-31,607	1,572	131,931	18404,800	0,067	1720,071	2237,240*
	N(0.1,0.01)	2123,725	113,462	5,644	139,351	32292,390	0,066	1850,598	2396,853*
	N(0.5,0.01)	2444,548	434,285	21,603	124,498	204103,200	0,051	2200,533	2688,563
25%**	N(0.001,0.01)	855,911	-1154,350	57,423	82,810	1339386,001	0,097	693,600	1018,222
	N(0.1,0.01)	917,834	-1092,430	54,343	95,932	1202604,020	0,105	729,815	1105,853
	N(0.5,0.01)	979,858	-1030,410	51,257	102,711	1072284,002	0,105	778,550	1181,165
50%**	N(0.001,0.01)	1943,218	-67,044	3,335	128,038	20888,550	0,066	1692,265	2194,172*
	N(0.1,0.01)	2096,877	86,615	4,309	134,818	25677,910	0,064	1832,635	2361,12*
	N(0.5,0.01)	2226,784	216,521	10,771	136,672	65560,680	0,061	1958,907	2494,662*
75%**	N(0.001,0.01)	1619,761	-390,502	19,425	132,315	169999,000	0,082	1360,425	1879,096
	N(0.1,0.01)	1748,096	-262,167	13,041	142,457	89025,490	0,081	1468,879	2027,313*
	N(0.5,0.01)	1903,640	-106,623	5,304	131,727	28720,360	0,069	1645,456	2161,824*
100%**	N(0.001,0.01)	1978,645	-31,618	1,573	131,934	18406,230	0,067	1720,054	2237,235*
	N(0.1,0.01)	2123,659	113,396	5,641	139,288	32259,850	0,066	1850,654	2396,664*
	N(0.5,0.01)	2444,566	434,303	21,604	124,489	204116,600	0,051	2200,568	2688,564

** Com correção Bootstrap.* Intervalo de confiança contém o verdadeiro valor do parâmetro.

Tabela 6.4: Principais descritivas do estimador Horvitz-Thompson considerando a distribuição Weibull.

	Distribuição Weibull								
Cont.	Erros de Mens.	Parâmetro(899,26)							
		Media	Vies	Vies. R	Desvio	EQM	CV	Lim. Inf.	Lim. Sup.
25%	$N(0.001,0.01)$	602,701	-296,924	33,005	36,451	89492,570	0,060	531,251	674,152
	$N(0.1,0.01)$	619,650	-279,975	31,121	37,345	79780,670	0,060	546,459	692,840
	$N(0.5,0.01)$	733,747	-165,878	18,439	48,452	29863,150	0,066	638,779	828,715
50%	$N(0.001,0.01)$	918,257	18,632	2,071	45,129	2383,794	0,049	829,804	1006,710*
	$N(0.1,0.01)$	925,712	26,087	2,900	44,011	2617,491	0,048	839,450	1011,973*
	$N(0.5,0.01)$	1152,337	252,712	28,091	52,503	66619,970	0,046	1049,434	1255,240
75%	$N(0.001,0.01)$	831	-68,625	7,628	44,189	6662,022	0,053	744,390	917,610*
	$N(0.1,0.01)$	857,891	-41,734	4,639	46,608	3914,071	0,054	766,539	949,242*
	$N(0.5,0.01)$	1162,959	263,334	29,272	57,544	72656,110	0,049	1050,171	1275,746
100%	$N(0.001,0.01)$	926,245	26,62	2,959	46,788	2897,743	0,050	834,541	1017,949*
	$N(0.1,0.01)$	983,665	84,040	9,342	45,109	9097,536	0,046	895,250	1072,079*
	$N(0.5,0.01)$	1389,993	490,369	54,508	41,638	242195,500	0,030	1308,371	1471,616
25%**	$N(0.001,0.01)$	602,700	-296,925	33,005	36,432	89491,710	0,060	531,300	674,099
	$N(0.1,0.01)$	619,645	-279,980	31,122	37,338	79782,960	0,060	546,460	692,830
	$N(0.5,0.01)$	733,740	-165,885	18,439	48,531	29873,050	0,066	638,618	828,863
50%**	$N(0.001,0.01)$	918,266	18,641	2,072	45,173	2388,057	0,049	829,728	1006,804*
	$N(0.1,0.01)$	925,686	26,061	2,897	44,010	2616,041	0,048	839,427	1011,945*
	$N(0.5,0.01)$	1152,330	252,705	28,090	52,501	66616,250	0,046	1049,429	1255,231
75%**	$N(0.001,0.01)$	831,015	-68,610	7,626	44,187	6659,839	0,053	744,408	917,622*
	$N(0.1,0.01)$	857,887	-41,738	4,640	46,620	3915,493	0,054	766,512	949,261*
	$N(0.5,0.01)$	1162,983	263,358	29,274	57,534	72667,640	0,049	1050,214	1275,752
100%**	$N(0.001,0.01)$	926,274	26,649	2,962	46,802	2900,557	0,051	834,543	1018,005*
	$N(0.1,0.01)$	983,665	84,040	9,342	45,107	9097,384	0,046	895,254	1072,076*
	N(0.5,0.01)	1389,993	490,368	54,508	41,691	242198,900	0,030	1308,271	1471,714

** Com correção Bootstrap.* Intervalo de confiança contém o verdadeiro valor do parâmetro.

Através do viés relativo (Figura 6.1) nota-se que a distribuição Uniforme tem padrão de aumento do viés a medida que o erro aumenta, nota-se também que a correção Bootstrap não amenizou este efeito. Já a Figura (Figura 6.2) mostra que a distribuição Gamma se comportou semelhante a Uniforme nas contaminações de 50% e 100%, entretanto as contaminações de 25% e 75% tiveram padrão descrente de viés a medida que aumenta-se o erro de mensuração.

Quando avaliamos através da distribuição Weibull (Figura 6.3), vemos que 50%,75% e 100% continuam, assim como a distribuição Gamma e Uniforme, com um padrão crescente de viés. Porém a contaminação de 25% apresenta-se diferente das demais, com aumento do Viés Relativo em erros pequenos e uma diminuição do Viés, quando o aumenta-se o Erro de Mensuração.

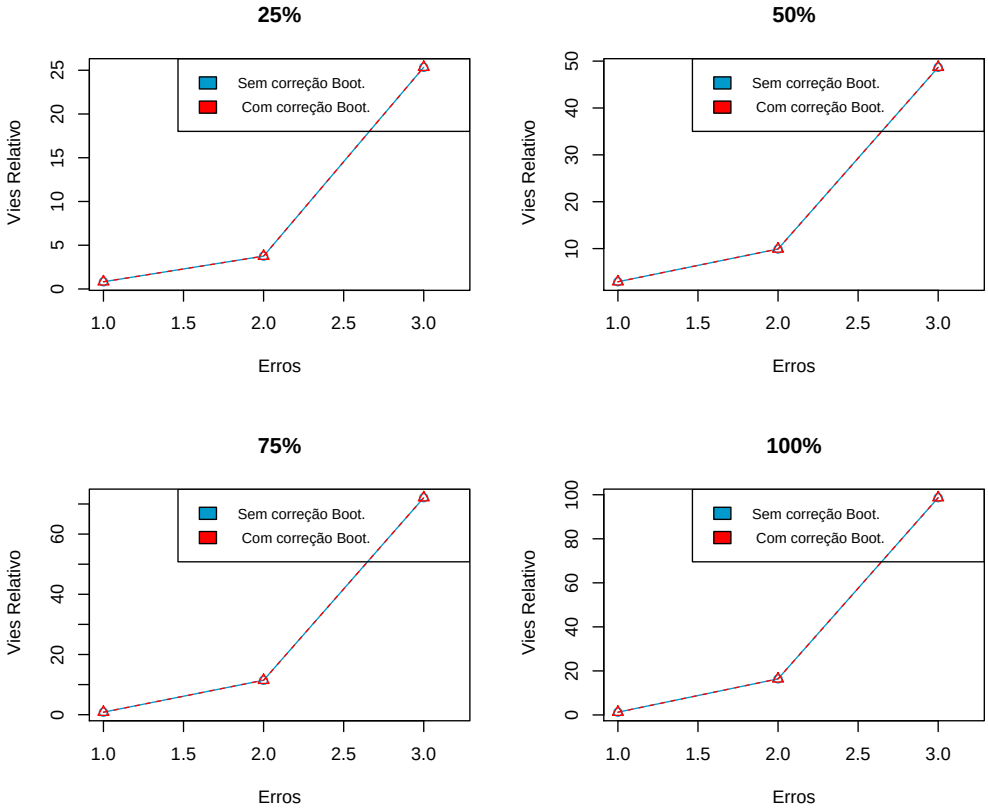

Figura 6.1: Viés Relativo ao estimador Horvitz Thompson da distribuição Uniforme nos três tipos de erro de mensuração.

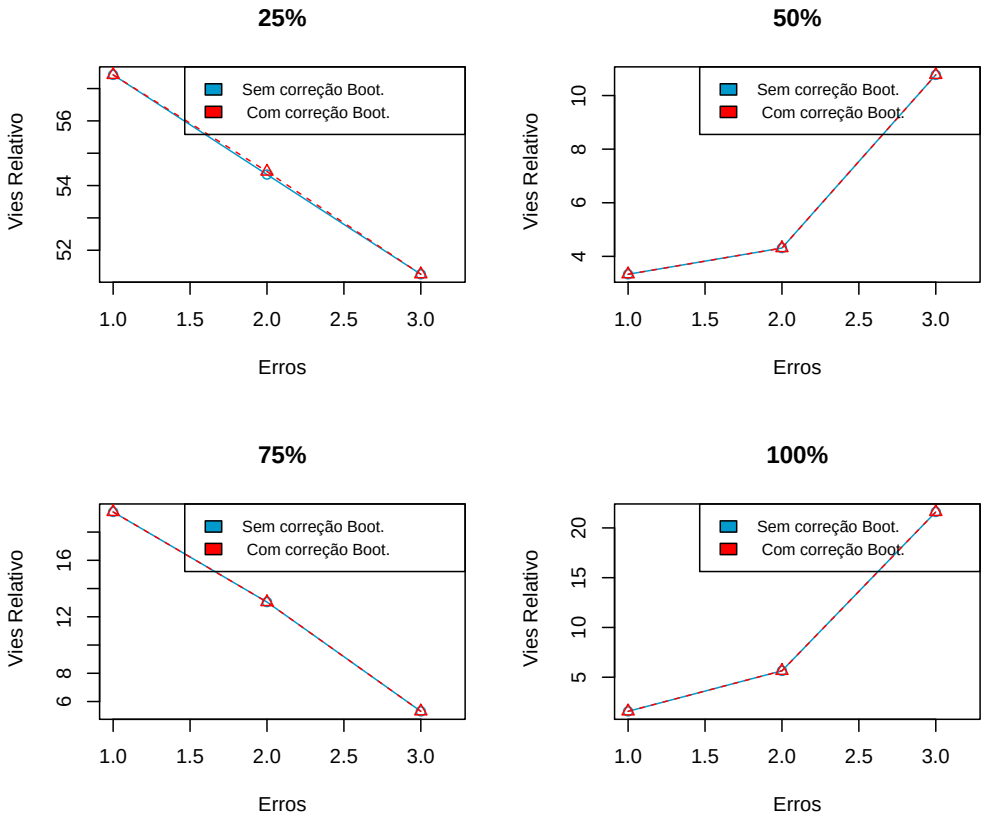

Figura 6.2: Viés Relativo ao estimador Horvitz Thompson da distribuição Gamma nos três tipos de erro de mensuração.

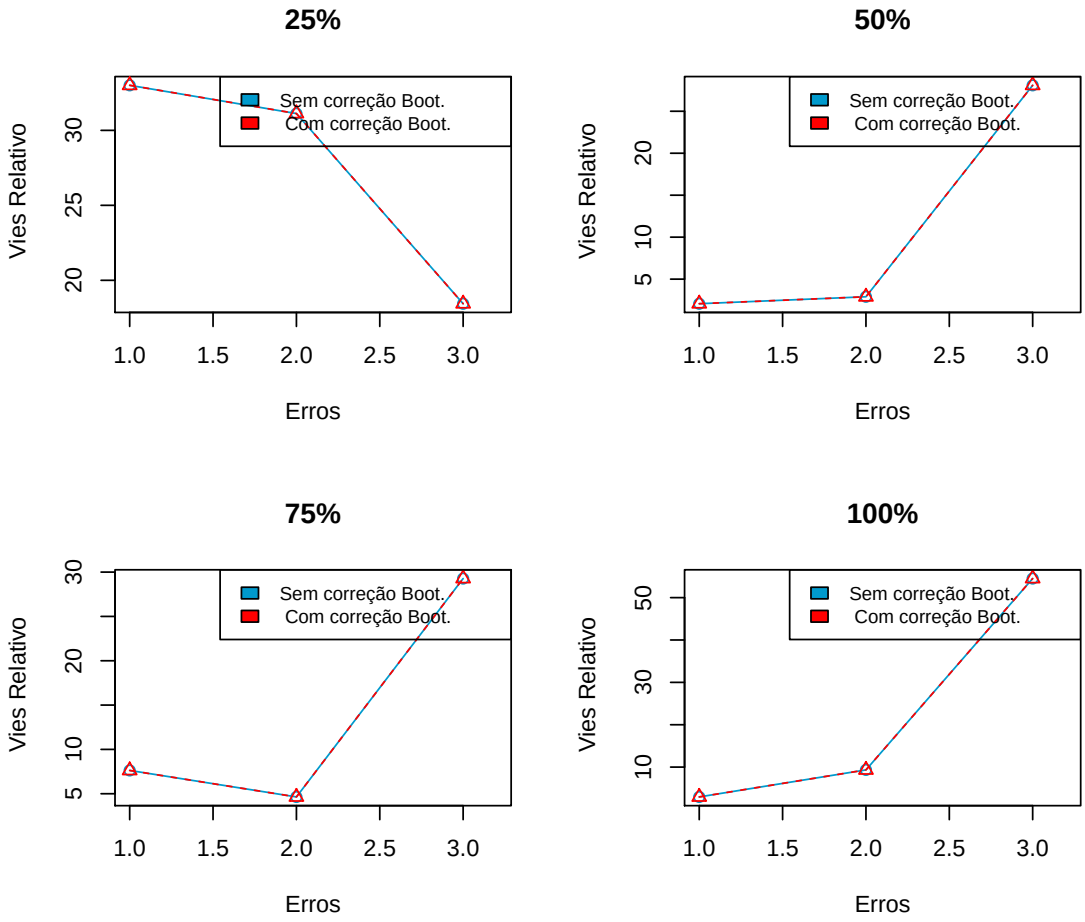

Figura 6.3: Viés Relativo ao estimador Horvitz Thompson da distribuição Weibull nos três tipos de erro de mensuração

Independente da distribuição utilizada, as estimativas com e sem a correção Bootstrap são muito próximas, porém vemos que o Viés Relativo para um único cadastro aumenta junto com o aumento da contaminação e o Bootstrap não consegue corrigir as estimativas no cenário estudado.

6.2 Estimação de totais populacionais sob a abordagem de Cadastro Duplo

Dado que as análises com o estimador Horvitz Thompson apresentaram dificuldades para corrigir os efeitos do Erro de Mensuração, foram utilizadas as distribuições Gamma(2,1) e Uniforme(0,1) sob efeito de Cadastro Duplo. Estas distribuições haviam apresentado o pior e o melhor cenário, respectivamente, na utilização de apenas um cadastro. Sob este aspecto, será avaliada a utilização de dois cadastros na estimação de forma geral e também com correção Bootstrap.

6.2.1 Estimativas através da Estratégia de Hartley

Quando utilizamos dois cadastros para calcular estimativas de Hartley, vemos que as estimativas foram mais próximas e os Vieses e EQMs diminuíram. Portanto, há indícios que o cadastro duplo mostra uma melhor estimativa, provavelmente por estar amostrando mais no cadastro B que não foi contaminado por erros de mensuração.

Ou seja, a contribuição do cadastro B melhora as estimativas, anulando um pouco o efeito do erro de mensuração inserido no cadastro A. Porém, o Bootstrap ainda não consegue corrigir as estimativas, apenas diminui os Vieses e Erros Quadráticos Médios. Quando consideramos o maior erro de mensuração inserido no cadastro A, não há melhoras significantes quando comparados com um único cadastro, pois as medidas descritivas de variação ainda são grandes e os intervalos de confiança ainda não contem o verdadeiro valor do parâmetro. As tabelas 6.5 e 6.6 mostram as estimativas obtidas através do estimador de Hartley. A distribuição Gamma, obteve bom desempenho para Erros de Mensuração pequenos e médios, onde o intervalo de confiança obteve o valor do parâmetro. Já com a distribuição Uniforme os desvios e Coeficientes de Variação são ainda menores, porém os únicos intervalos de confiança que contem o parâmetro são com erro de mensuração pequeno.

Tabela 6.5: Principais descritivas do estimador Hartley considerando a distribuição Gamma.

Cont.	Erros de Mens.	Distribuição Gamma Parâmetro(3687,87) Média	Vies	Vies. R	Desvio	EQM	CV	Lim. Inf.	Lim. Sup.
25%	N(0.001,0.01)	3729,510	41,643	1,129	175,673	32595,020	0,047	3385,200	4073,832*
	N(0.1,0.01)	3744,650	56,783	1,540	171,085	32494,380	0,046	3409,330	4079,980*
	N(0.5,0.01)	4105,140	417,274	11,315	168,353	202460,300	0,041	3775,170	4435,120
50%	N(0.001,0.01)	3748,290	60,415	1,638	184,244	37595,930	0,049	3387,170	4109,404*
	N(0.1,0.01)	3738,020	50,151	1,360	168,944	31057,040	0,045	3406,890	4069,151*
	N(0.5,0.01)	4159,600	471,731	12,791	169,344	251207,500	0,041	3827,690	4491,510
75%	N(0.001,0.01)	3765,060	77,186	2,093	189,600	41905,930	0,050	3393,440	4136,673*
	N(0.1,0.01)	3678,620	-9,253	0,251	169,153	28698,230	0,046	3347,080	4010,157*
	N(0.5,0.01)	4105,110	417,242	11,314	171,823	203614,000	0,042	3768,340	4441,880
100%	N(0.001,0.01)	3709,270	21,402	0,580	175,854	31382,680	0,047	3364,600	4053,946*
	N(0.1,0.01)	3711,580	23,707	0,643	177,821	32182,430	0,048	3363,050	4060,107*
	N(0.5,0.01)	4114,250	426,555	11,566	172,234	211613,600	0,042	3776,850	4452,000
25%**	N(0.001,0.01)	3725,420	37,546	1,018	196,384	39976,550	0,053	3340,500	4110,330*
	N(0.1,0.01)	3739,850	51,978	1,409	190,720	39075,690	0,051	3366,040	4113,659*
	N(0.5,0.01)	4105,470	417,604	11,324	187,242	209452,600	0,046	3738,480	4472,470
50%**	N(0.001,0.01)	3749,830	61,960	1,680	203,366	45196,580	0,054	3351,230	4148,427*
	N(0.1,0.01)	3732,660	44,790	1,215	183,330	35616,120	0,049	3373,330	4091,987*
	N(0.5,0.01)	4163,730	475,855	12,903	182,937	259903,900	0,044	3805,170	4522,280
75%**	N(0.001,0.01)	3763,640	75,773	2,055	209,190	49505,640	0,056	3353,610	4173,673*
	N(0.1,0.01)	3676,180	-11,687	0,317	186,512	34923,330	0,051	3310,620	4041,747*
	N(0.5,0.01)	4101,720	413,847	11,222	191,496	207940,300	0,047	3726,390	4477,050
100%**	N(0.001,0.01)	3712,260	24,386	0,661	195,059	38642,790	0,053	3329,940	4094,572*
	N(0.1,0.01)	3707,780	19,908	0,540	195,913	38778,350	0,053	3323,790	4071,768*
	N(0.5,0.01)	4119,100	431,230	11,693	187,979	221295,400	0,046	3750,660	4487,540

** Com correção Bootstrap.* Intervalo de confiança contém o verdadeiro valor do parâmetro.

Tabela 6.6: Principais descritivas do estimador Hartley considerando a distribuição Uniforme.

Distribuição Uniforme-Hartley									
		Parâmetro(894,705)							
Cont.	Erros de Mens.	Media	Vies	Vies. R	Desvio	EQM	CV	Lim. Inf.	Lim. Sup.
25%	$N(0.001,0.01)$	901,285	4,157	0,459	33,888	1165,648	0,038	834,865	967,704*
	$N(0.1,0.01)$	1018,680	113,238	12,506	36,417	1148,980	0,036	947,303	1090,056
	$N(0.5,0.01)$	1404,055	498,613	55,069	35,434	249870,700	0,025	1334,604	1473,506
50%	$N(0.001,0.01)$	902,245	3,197	0,353	36,192	1320,080	0,040	831,309	973,182*
	$N(0.1,0.01)$	998,829	93,387	10,314	34,976	9944,529	0,035	930,277	1067,382
	$N(0.5,0.01)$	1404,623	499,181	55,131	35,955	550474,200	0,026	1334,151	1475,094
75%	$N(0.001,0.01)$	900,189	5,253	0,580	37,184	1410,210	0,041	827,309	973,069*
	$N(0.1,0.01)$	1012,891	107,449	11,867	37,085	12920,640	0,037	940,205	1085,577
	$N(0.5,0.01)$	1404,596	499,154	55,128	35,833	250438,800	0,026	1334,363	1474,829
100%	$N(0.001,0.01)$	900,341	5,101	0,563	35,301	1272,165	0,039	831,152	969,531*
	$N(0.1,0.01)$	1016,168	110,726	12,229	35,462	13517,830	0,035	946,664	1085,673
	$N(0.5,0.01)$	1405,675	500,233	55,247	35,306	251479,600	0,025	1336,476	1474,874
25%**	$N(0.001,0.01)$	901,008	4,434	0,490	33,898	168,738	0,038	834,567	967,448*
	$N(0.1,0.01)$	1020,118	114,676	12,665	36,328	14470,310	0,036	948,914	1091,321
	$N(0.5,0.01)$	1408,055	502,614	55,510	35,522	253882,300	0,025	1338,432	1477,679
50%**	$N(0.001,0.01)$	901,885	3,557	0,393	36,212	1323,957	0,040	830,910	972,860*
	$N(0.1,0.01)$	999,877	94,435	10,430	34,845	1132,240	0,035	931,581	1068,174
	$N(0.5,0.01)$	1408,265	502,824	55,534	35,925	254122,200	0,026	1337,852	1478,679
75%**	$N(0.001,0.01)$	900,380	5,062	0,559	37,040	1397,610	0,041	827,780	972,979*
	$N(0.1,0.01)$	1013,977	108,536	11,987	37,010	13149,730	0,036	941,438	1086,517
	$N(0.5,0.01)$	1407,587	502,145	55,459	35,740	253427,100	0,025	1337,536	1477,638
100%**	$N(0.001,0.01)$	899,741	5,700	0,630	35,268	1276,315	0,039	830,616	968,866*
	$N(0.1,0.01)$	1017,331	111,889	12,357	35,689	13771,520	0,035	974,969	1086,693
	$N(0.5,0.01)$	1409,247	503,806	55,642	35,332	255068,400	0,025	1339,997	1478,498

** Com correção Bootstrap.* Intervalo de confiança contém o verdadeiro valor do parâmetro.

O gráfico abaixo (Figura 6.4) mostra que o Bootstrap não consegue corrigir as estimativas de totais populacionais, mantendo os Vieses praticamente iguais.

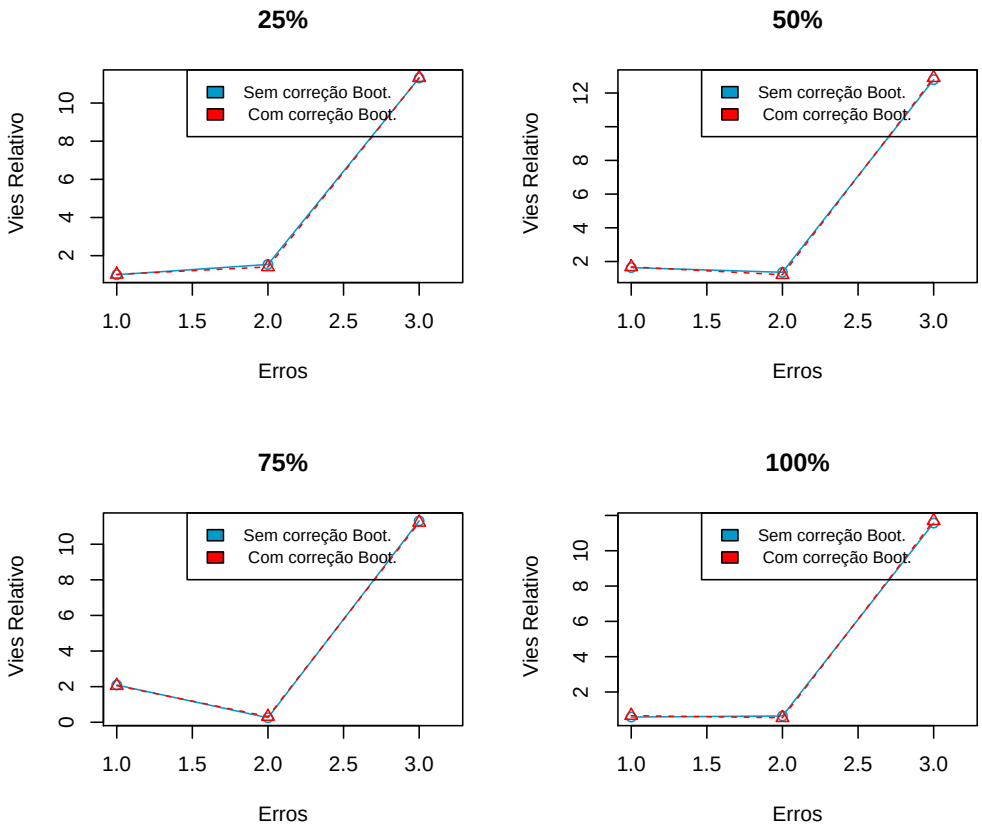

Figura 6.4: Avaliação do Viés Relativo sob estratégia de Hartley na distribuição Gamma nos três tipos de erro de mensuração

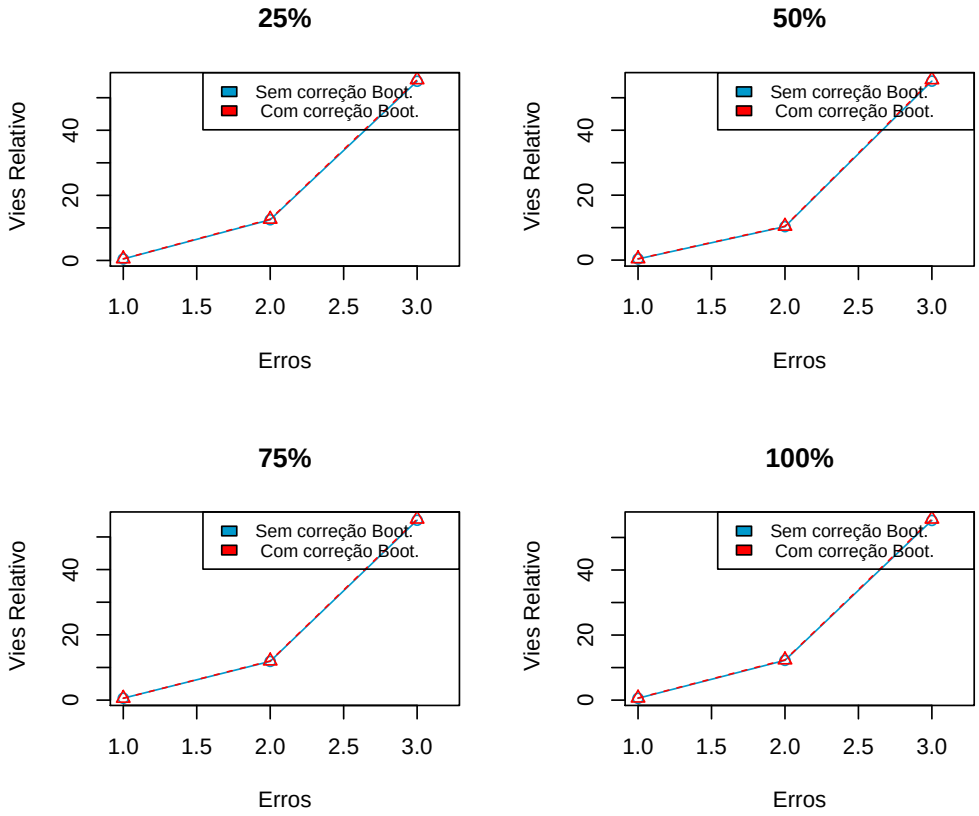

Figura 6.5: Avaliação do Viés Relativo sob estratégia de Hartley na distribuição Uniforme nos três tipos de erro de mensuração

6.2.2 Estimativas através da Estratégia de Bankier

A tabela 6.7 mostra o comportamento do estimador Bankier nas amostras geradas. Apenas nas contaminações de 25%, 50% e 75% com Erro de Mensuração $N(0.5,0.01)$ o intervalo de confiança conteve o verdadeiro valor do parâmetro. Além disso, nota-se que novamente o Bootstrap não conseguiu corrigir os dados, se mostrando método ineficaz para ser utilizado em dados com Erros de Mensuração.

Tabela 6.7: Principais descritivas do estimador Bankier considerando a distribuição Gamma.

	Distribuição Gamma								
Cont.	Erros de Mens.	Parâmetro(3677,041)							
		Media	Vies	Vies. R	Desvio	EQM	CV	Lim. Inf.	Lim. Sup.
25%	N(0.001,0.01)	2840,341	-836,699	22,755	207,857	743270,300	0,073	2432,941	3247,741
	N(0.1,0.01)	2951,052	-725,989	19,744	216,995	574146,400	0,074	2525,743	3376,361
	N(0.5,0.01)	3738,756	61,716	1,678	234,258	58685,380	0,063	3279,611	4197,901*
50%	N(0.001,0.01)	2845,525	-831,515	22,614	216,905	738465,700	0,076	2420,391	3270,660
	N(0.1,0.01)	3004,888	-672,153	18,280	219,434	499940,300	0,073	2574,797	3434,979
	N(0.5,0.01)	3655,586	-21,455	0,583	235,916	56116,700	0,065	3193,190	4117,982*
75%	N(0.001,0.01)	2831,969	-845,072	22,982	213,411	759690,900	0,075	2413,683	3250,255
	N(0.1,0.01)	3071,659	-605,381	16,464	227,436	418213,600	0,074	2625,885	3517,434
	N(0.5,0.01)	3722,130	45,089	1,226	229,588	54743,570	0,062	3272,138	4172,122*
100%	N(0.001,0.01)	2807,839	-869,202	23,639	212,909	800841,600	0,076	2390,537	3225,141
	N(0.1,0.01)	2994,868	-682,172	18,552	218,191	512966,400	0,073	2567,214	3422,523
	N(0.5,0.01)	3675,592	-1,449	0,039	237,891	56594,000	0,065	3209,326	4141,857*
25%**	N(0.001,0.01)	2896,278	-780,763	21,233	222,702	659186,300	0,077	2459,783	3332,773
	N(0.1,0.01)	2891,473	-785,568	21,364	217,839	664570,800	0,075	2464,509	3318,436
	N(0.5,0.01)	2897,652	-779,389	21,196	215,731	653986,500	0,074	2474,819	3320,485
50%**	N(0.001,0.01)	2900,091	-776,950	21,130	223,718	653700,600	0,077	2461,603	3338,579
	N(0.1,0.01)	2876,688	-800,352	21,766	222,932	690262,700	0,077	2439,741	3313,636
	N(0.5,0.01)	2889,194	-787,847	21,426	216,049	667379,100	0,075	2465,739	3312,650
75%**	N(0.001,0.01)	2900,091	-776,950	21,130	223,718	653700,600	0,077	2461,603	3338,579
	N(0.1,0.01)	2876,688	-800,352	21,766	222,932	690262,700	0,077	2439,741	3313,636
	N(0.5,0.01)	2889,194	-787,847	21,426	216,049	667379,100	0,075	2465,739	3312,650
100%**	N(0.001,0.01)	2884,886	-792,154	21,543	216,847	674530,600	0,075	2459,867	3309,906
	N(0.1,0.01)	2885,036	-792,005	21,539	220,105	675717,800	0,076	2453,630	3316,442
	N(0.5,0.01)	2871,839	-805,202	21,898	207,057	691222,200	0,072	2466,007	3277,671

** Com correção Bootstrap.* Intervalo de confiança contém o verdadeiro valor do parâmetro.

Tabela 6.8: Principais descritivas do estimador Bankier considerando a distribuição Uniforme.

Distribuição Uniforme-Bankier									
		Parâmetro(894,705)							
Cont.	Erros de Mens.	Media	Vies	Vies. R	Desvio	EQM	CV	Lim. Inf.	Lim. Sup.
25%	N(0.001,0.01)	733,112	161,593	18,061	47,219	28342,030	0,064	640,564	825,660
	N(0.1,0.01)	873,246	21,460	2,399	0,721	3033,158	0,058	773,832	972,659*
	N(0.5,0.01)	1433,797	539,092	60,254	59,546	294165,900	0,042	1317,087	1550,508
50%	N(0.001,0.01)	732,636	162,070	18,114	47,637	28535,840	0,065	639,268	826,004
	N(0.1,0.01)	879,229	15,476	1,730	51,966	2939,983	0,059	777,375	981,083*
	N(0.5,0.01)	1441,467	546,762	61,111	62,677	302877,200	0,043	1318,620	1564,315
75%	N(0.001,0.01)	740,095	154,610	17,281	47,314	26142,950	0,064	647,360	832,830
	N(0.1,0.01)	866,368	28,337	3,167	48,772	3181,685	0,056	770,775	961,961*
	N(0.5,0.01)	1430,231	535,526	59,855	61,984	290629,600	0,043	1308,742	1551,720
100%	N(0.001,0.01)	732,252	162,452	18,157	46,317	28536,190	0,063	641,471	823,034
	N(0.1,0.01)	867,252	27,453	3,068	50,128	3266,512	0,058	769,001	965,504*
	N(0.5,0.01)	1436,160	541,457	60,518	62,948	297137,800	0,044	1312,784	1559,540
25%**	N(0.001,0.01)	698,598	196,107	21,919	50,112	40969,180	0,072	600,378	796,819
	N(0.1,0.01)	710,527	184,178	20,585	49,106	36333,020	0,069	614,279	806,775
	N(0.5,0.01)	801,398	93,307	10,429	48,332	11042,200	0,060	706,667	896,130*
50%**	N(0.001,0.01)	694,335	200,371	22,395	51,620	42813,090	0,074	593,159	795,510
	N(0.1,0.01)	708,335	186,370	20,830	50,369	37270,890	0,071	609,612	807,058
	N(0.5,0.01)	798,119	96,586	10,795	50,475	11876,600	0,063	699,189	897,050*
75%**	N(0.001,0.01)	693,752	200,954	22,460	53,014	43192,910	0,076	589,845	797,658
	N(0.1,0.01)	713,021	181,685	20,307	50,616	35571,330	0,071	613,814	812,227
	N(0.5,0.01)	798,720	95,985	10,728	49,845	11697,650	0,062	701,024	896,417*
100%**	N(0.001,0.01)	707,555	187,149	20,917	44,483	37003,600	0,063	620,368	794,743
	N(0.1,0.01)	727,547	167,158	18,683	45,261	29990,380	0,062	638,836	816,259
	N(0.5,0.01)	805,904	88,801	9,925	46,929	10087,910	0,058	713,923	897,887*

** Com correção Bootstrap.* Intervalo de confiança contém o verdadeiro valor do parâmetro.

As figuras 6.6 e 6.7 mostram o comportamento do Viés Relativo nos diferentes níveis de erro e contaminação, das distribuições Gamma e uniforme, respectivamente. Na distribuição Gamma, o Bootstrap em todas as contaminações apresenta uma discreta melhora quando o erro tem distribuição $N(0.001,0.01)$, porém quando o erro aumenta o Bootstrap se distancia ainda mais do real valor, aumentando os Vieses, ou seja, para o aumento do erro de mensuração há um aumento do Viés Relativo que o Bootstrap não consegue corrigir.

Na distribuição Uniforme, o Bootstrap apresenta Vieses similares aos que foram estimados sem sua correção quando a contaminação é de 25%. Entretanto, quando a contaminação é de 50%, 75% e 100%, o método de reamostragem auxilia na estimação quando o erro de mensuração é grande, reduzindo os vieses. Considerando os demais casos (erros pequenos e médios) ele aumenta o Viés Relativo.

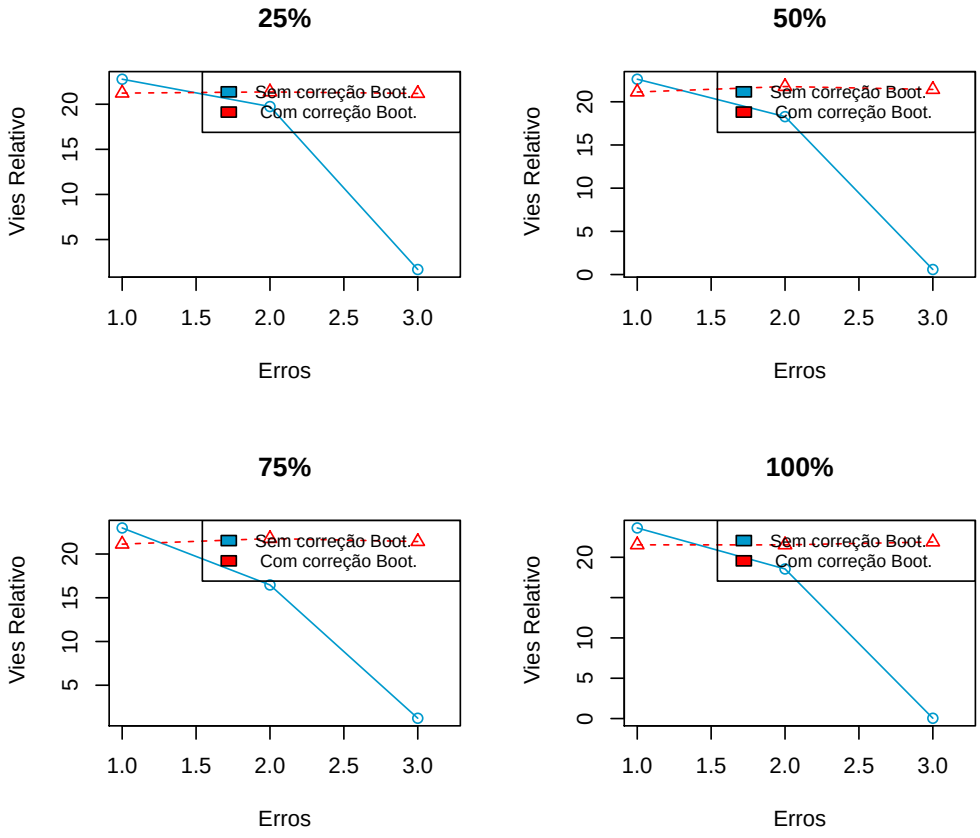

Figura 6.6: Viés Relativo da distribuição Gamma estimado através da estratégia de Bankier, sob efeito do Bootstrap nos três tipos de erros de mensuração

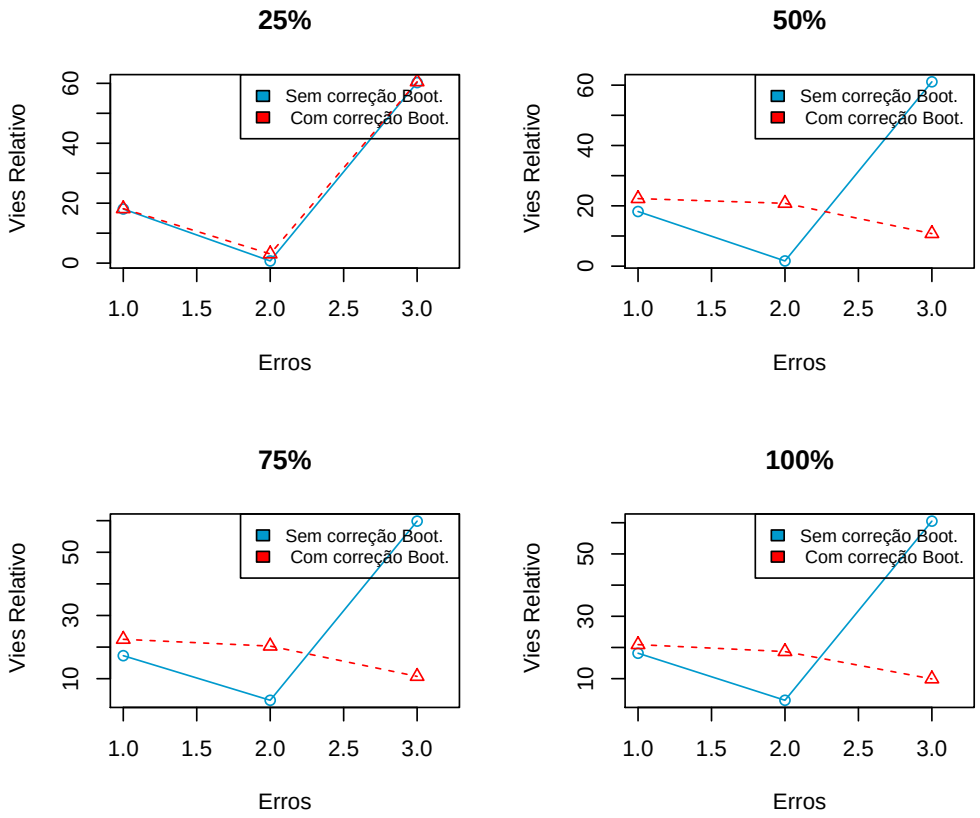

Figura 6.7: Viés Relativo da distribuição Uniforme estimado através da estratégia de Bankier, sob efeito do Bootstrap nos três tipos de erros de mensuração

6.3 Comparação dos Estimadores

Comparando os estimadores, Horvitz Thompson, Hartley e Bankier, quanto a sua variabilidade, nota-se que os estimadores de Cadastro Duplo apresentaram desempenho maior com coeficientes bem menores para distribuição Gamma. Além disso, há evidência de que os estimadores de Hartley são mais confiáveis para estimar o parâmetro populacional de interesse, dado que apresentou estimativas menores do Coeficiente de Variação, dentre os demais.

O método Bootstrap foi utilizado para que as amostras retiradas dele explicassem melhor as informações sobre o parâmetro de interesse. Porém, os resultados mostram que o coeficiente tende a aumentar com o uso da correção, apresentando coeficientes de variação maiores. Independente da distribuição utilizada, o Bootstrap não consegue corrigir as estimativas, nem quando utilizamos Cadastro Duplo. Entretanto, ele não sofre grande influência do tamanho da contaminação, nem de sua variabilidade.

Os estimadores de Hartley e Bankier se mostraram mais consistentes por não sofrerem grande influencia do tamanho da contaminação e de sua variabilidade. Já o estimador de cadastro simples utilizado sofre muita influência da situação a qual se encontra, refletindo assim na variabilidade de seus coeficientes.

Já quando utilizamos a distribuição Uniforme, os estimadores Hartley e Horvitz Thompsom tiveram melhor desempenho, porém Hartley apresentou coeficiente de variação bem menores e não sofreu influencia da contaminação. Neste contexto, o Bootstrap não conseguiu melhorar seu desempenho nem dos demais, mantendo os coeficientes praticamente iguais.

Dentre todos os estimadores utilizados o estimador de Hartley apresentou estimativas com os menores Coeficientes de Variação, tornando-se eficiente, independente da variabilidade do erro de mensuração ou da porcentagem de contaminação.

Tabela 6.9: Coeficiente de Variação dos estimadores utilizados na distribuição Gamma

Distribuição Gamma							
Cont.	**Erros de Mens.**	**HT**	**Hartley**	**Bankier**	**HT***	**Hartley***	**Bankier***
25%	N(0.001,0.01)	0,0968	0,0471	0,0780	0,0968	0,0527	0,0760
	N(0.1,0.01)	0,1045	0,0457	0,0750	0,1045	0,0510	0,0720
	N(0.5,0.01)	0,1048	0,0410	0,0640	0,1048	0,0456	0,0730
50%	N(0.001,0.01)	0,0659	0,0492	0,0720	0,0659	0,0542	0,0750
	N(0.1,0.01)	0,0643	0,0452	0,0740	0,0643	0,0491	0,0740
	N(0.5,0.01)	0,0613	0,0407	0,0630	0,0614	0,0439	0,0720
75%	N(0.001,0.01)	0,0817	0,0504	0,0710	0,0817	0,0556	0,0750
	N(0.1,0.01)	0,0815	0,0460	0,0720	0,0815	0,0507	0,0740
	N(0.5,0.01)	0,0692	0,0419	0,0670	0,0692	0,0467	0,0710
100%	N(0.001,0.01)	0,0667	0,0474	0,0760	0,0667	0,0525	0,0730
	N(0.1,0.01)	0,0656	0,0479	0,0700	0,0656	0,0528	0,0740
	N(0.5,0.01)	0,0509	0,0419	0,0650	0,0509	0,0456	0,0730

* Com correção Bootstrap.

Tabela 6.10: Coeficiente de Variação dos estimadores utilizados na distribuição Uniforme

Distribuição Uniforme							
Cont.	**Erros de Mens.**	**HT**	**Hartley**	**Bankier**	**HT***	**Hartley***	**Bankier***
25%	N(0.001,0.01)	0,0528	0,0376	0,0644	0,0527	0,0376	0,0717
	N(0.1,0.01)	0,0551	0,0357	0,0581	0,0551	0,0356	0,0691
	N(0.5,0.01)	0,0554	0,0252	0,0415	0,0554	0,0252	0,0603
50%	N(0.001,0.01)	0,0522	0,0401	0,0650	0,0522	0,0402	0,0743
	N(0.1,0.01)	0,0517	0,0350	0,0591	0,0518	0,0348	0,0711
	N(0.5,0.01)	0,0482	0,0256	0,0435	0,0482	0,0255	0,0632
75%	N(0.001,0.01)	0,0552	0,0413	0,0639	0,0552	0,0411	0,0764
	N(0.1,0.01)	0,0529	0,0366	0,0563	0,0529	0,0365	0,0710
	N(0.5,0.01)	0,0397	0,0255	0,0433	0,0397	0,0254	0,0624
100%	N(0.001,0.01)	0,0522	0,0392	0,0633	0,0522	0,0392	0,0629
	N(0.1,0.01)	0,0469	0,0349	0,0578	0,0469	0,0348	0,0622
	N(0.5,0.01)	0,0275	0,0251	0,0438	0,0275	0,0251	0,0582

* Com correção Bootstrap.

Capítulo 7

Conclusões

No presente plano de trabalho foram apresentadas estratégias de estimação para totais populacionais conhecidas na literatura como o estimador Horvitz Thompsom e novas versões destes estimadores baseadas no estimador de Hartley (1962) e Bankier (1989) sob efeito de Erros de Mensuração. Todas as principais propriedades estatísticas destes estimadores foram desenvolvidas de forma original e fornecerão contribuição adicional à teoria já existente.

Este trabalho apresentou estudo de simulação sob os estimadores para totais populacionais e avaliação do efeito sofrido por eles quando existiam Erros de Mensuração. A partir das tabelas, nota-se através de medidas descritivas que os estimadores funcionam melhor em pequenos Erros de Mensuração, e com contaminações pequenas seus Vieses Relativos são menores.

Para os cenários nos quais consideramos dois cadastros, onde apenas o cadastro$A$ está contaminado com erros de mensuração e B está livre, nota-se que há uma melhora nas estimativas, ou seja, a contribuição do cadastro B, anula um pouco o efeito do Erro de Mensuração inserido no cadastro A. Porém, o Bootstrap ainda não consegue corrigir as estimativas, apenas diminui os Vieses e Erros Quadráticos Médios. Quando consideramos o maior Erro de Mensuração inserido no cadastro A, não há melhoras significantes quando comparados com um único cadastro, pois as medidas descritivas de variação ainda são grandes e os intervalos de confiança ainda não contem o verdadeiro valor do parâmetro. Principalmente porque apesar de os Vieses Relativos e coeficientes de variação reduzirem quando passamos de um para dois cadastros, os intervalos de confiança para contaminação de 100% continuam sem conter o verdadeiro valor do parâmetro.

Foram realizados, neste trabalho, estudos via simulação a fim de avaliar o desempenho do delineamento Bootstrap em dados contaminados com Erros de Mensuração. Verificou-se, por meio dos resultados obtidos, que as estimativas com correção Bootstrap mostram-se similares às sem o método de reamostragem. Esta dificuldade de corrigir as estimativas ocorre independente da utilização de Cadastro Duplo. Estes resultados são contraditórios com as propriedades mencionadas do método Bootstrap, que é conhecido por apresentar resultados mais precisos. Porém, para realizar uma estimação através da utilização de Bootstrap é necessária a realização de um número muito grande de reamostragens e acredita-se que o método falha pois sua reamostragem apenas aumenta o número de informações contaminadas. Ou seja, a retirada de amostras de uma pseudo-amostra original, cria um cadastro com uma porcentagem maior de contaminação, tornando o método ineficiente para os casos estudados no trabalho .

Em relação a distribuição Gamma, quando comparamos todos os estimadores utilizados as melhores estimativas foram obtidas quando utilizados dois cadastros, uma vez que o Coeficiente de Variação foi reduzido. Além do resultado positivo na utilização do Cadastro Duplo, fica claro que entre os dois estimadores utilizados nesta situação, Hartley se comportou melhor, com

os menores Coeficientes de Variação. Já quando comparamos as estimativas apresentadas pela distribuição Uniforme, verifica-se que os melhores valores também foram obtidos com o estimador de Hartley, que além de obter as menores estimativas não sofre influência da porcentagem da contaminação. Portanto, em situações com dados contaminados por Erros de Mensuração as melhores estimativas para o total populacional serão obtidas por Cadastro Duplo através de estimativas de Hartley.

7.0.1 Sugestões para Trabalhos Futuros

Algumas sugestões para trabalhos futuros consistem em estudar o aumento do tamanho amostral e incluir a contaminação com Erro de Mensuração também no cadastro B, e então verificar o desempenho do estimador. Outra alternativa é variar também o método probabilístico de seleção da amostra.

Bibliografia

[1] Bankier (1986). Estimators based on several stratified samples with applications to multiple frame surveys. Journal of the American Statistical Association.

[2] Bailar,B.A. and Dalenius, T. (1969). Estimating the Response Variance Components of the U.S Bureau of the Census'Survey Model.

[3] Bolfarine, H. e Bussab, W. (2005). Elementos de Amostragem, volume 1. Blucher, 1 edition.

[4] Brackstone, G.J. and Rao, J.N.K. (1979). An Investigation of Raking Ratio Estimators. Sankhya, Series C.

[5] Coelho, H. F. C. (2007). A abordagem de cadastro duplo: Estimação assistida por modelos com aplicações em pesquisas agropecuárias. Master's thesis, Universidade Federal de Pernambuco, Departamento de Estatística.

[6] Coelho, H. F. C. (2011). Inferência sob planos amostrais de cadastro duplo. Tese (Doutorado). Programa de Pós-Graduação em Estatística, Universidade Federal de Pernambuco, Recife.

[7] Efron, B. (1979). Computer and the theory of statistics: thinking the unthinkable.

[8] Fuller,W. A. and Burmeister, L. F. (1972). Estimators for samples selected from two overlapping frames. Proceedings of social science section of The American Statistical Asociation.

[9] Hansen,M. H.,Hurwitz W.N., and Bershad, M. A. (1961). Mearurement Erros in Censuses and Surveys. Bulletin of the International Statistical Institute.

[10] Hansen,M. H.,Hurwitz W.N., and Pritzker, L. (1964). The Estimation and Interpenetration of Gross Differences and the Simple Response Variance. Contributions to Statistics.

[11] Hartley, H. O. (1962). Multiple frame surveys. Proceedings of the Social Stiatistics Association (ASA).

[12] Kalton, G., Anderson, D. (1986). Sampling Rare Populations. Journal of the Royal Statistical Society.

[13] Lepkowski, J.M. and R.M. Groves.(1986). A Mean Squared Error Model for Dual Frame, Mixed Mode Survey Design. Journal of the American Statistical Association.

[14] Lima, Claudio Ferreira, et al. (2008). "Comparison between analytical pyrolysis and nitrobenzene oxidation for determination of syringyl/guaiacyl ratio in Eucalyptus spp. lignin.

[15] Lohr, S. L. and Rao, J. (2000). Inference from dual frame surveys. Journal of the American Statistical Association.

[16] Lund, R. E. (1968). Estimators in multiple frame surveys. Proceedings ot the Social Statistics Section, American Statistical Association.

[17] Mahalanobis, P. C. (1946). On large-scale sample surveys. Philos. Trans. Roy. Soc. London Ser.

[18] Matias Jr., R. (2006). Análise quantitativa de risco baseada no método de Monte Carlo: Abordagem PMBOK. Congresso Brasileiro de Gerenciamento de Projetos.

[19] Medeiros, O. C. Estudo dos planos amostrais e estimadores para a aplicação no dimensionamento da popoluação canina de Rio Claro - SP. 2013. Dissertação (mestrado) - Universidade Estadual Paulista Júlio de Mesquita Filho, Instituto de Geociências e Ciências Exatas.

[20] Nascimento, C.A.D. (2007). Gerenciamento de Prazos: Uma revisão crítica das técnicas em uso em empreendimentos em regime de EPC. Dissertação de Mestrado.

[21] Sarndal C, Swensson B, Wretman J. (1992). Model assisted survey sampling. New York: Springer-Verlag.

[22] Skinner, C.J. (1991).On the efficiency of raking ratio estimation for multiple frame surveys, Journal of the American Statistical Association.

[23] Skinner, C. J. and Rao, J. N. (1996). Estimation in dual frame surveys with complex designs. Journal of the American Statistical Association.

Capítulo 8

Anexos

Rotina Computacional

Será exposto os casos de contaminação de 25% e 100%, como exemplificação da implementação para um e dois cadastros. Para Hartley foi demonstrado com a distribuição Gamma(2,0) e Bankier com a distribuição Uniforme(0,1).

```
#####################################################################
###############ERRO DE MENSURAÇÃO { 1 CADASTRO#######################
################# ################## HT#################################
#####################################################################
# BIBLIOTECAS
library(boot)
library(TeachingSampling)
library(SunterSampling)
library(samplingbook)
library(samplingVarEst)

#Criando o número de réplicas para o bootsrap
B=500
#Criando o número de réplicas para a simulação de Monte Carlo
R=1000
#Fixando o tamanho da amostra
n=100
# tamanho do cadastro
N=1000
# Fixando a Semente
set.seed(10)
# GERANDO OS DADOS
x=runif(N)
y=rgamma(N,2)
z=rweibull(N,2)
#Gerando Cadastro com erro de Mensuração
x1=runif(N)+rnorm(N,0.001,.01)
x2=runif(N)+rnorm(N,0.1,.01)
x3=runif(N)+rnorm(N,0.5,.01)
```

```
y1=rgamma(N,2)+rnorm(N,0.001,.01)
y2=rgamma(N,2)+rnorm(N,0.1,.01)
y3=rgamma(N,2)+rnorm(N,0.5,.01)
z1=rweibull(N,2)+rnorm(N,0.001,.01)
z2=rweibull(N,2)+rnorm(N,0.1,.01)
z3=rweibull(N,2)+rnorm(N,0.5,.01)
# Montando o banco de dados
dados=data.frame(x,y,z,x1,x2,x3,y1,y2,y3,z1,z2,z3)
dados
head(dados)
names(dados)
ttgeral=colSums(dados)

############################FUNÇÃO INTERESSANTE################################
stats=function(matriz,z,parametro)
{
c=ncol(matriz)
media=c()
vies=c()
viesr=c()
var=c()
desv=c()
ic1=c()
ic2=c()
eqm=c()

for(i in 1:1)
{
media[1]=mean(matriz[,1])
var[1]=var(matriz[,1])
desv[1]=sd(matriz[,1])
vies[1]=media[1]-parametro
viesr[1]=abs((media[1]-parametro)/parametro)*100
eqm[1]=var[1]+vies[1]^2
ic1[1]=media[1]-z*desv[1]
ic2[1]=media[1]+z*desv[1]
}
return(data.frame(media,vies,viesr,desv,eqm,ic1,ic2))
}
head(dados)
###########################################################################
######################### ####Estimação HT por AAS ########################
###########################################################################
######################### CONTAMINAÇÃO DE 100%#############################

tx=matrix(0,R,1)           # total de HT, com vetor coluna vazio
ty=matrix(0,R,1)           # total de HT, com vetor coluna vazio
tz=matrix(0,R,1)           # total de HT, com vetor coluna vazio
```

```
tx1=matrix(0,R,1)          # total de HT, com vetor coluna vazio
tx2=matrix(0,R,1)          # total de HT, com vetor coluna vazio
tx3=matrix(0,R,1)          # total de HT, com vetor coluna vazio
ty1=matrix(0,R,1)          # total de HT, com vetor coluna vazio
ty2=matrix(0,R,1)          # total de HT, com vetor coluna vazio
ty3=matrix(0,R,1)          # total de HT, com vetor coluna vazio
tz1=matrix(0,R,1)          # total de HT, com vetor coluna vazio
tz2=matrix(0,R,1)          # total de HT, com vetor coluna vazio
tz3=matrix(0,R,1)          # total de HT, com vetor coluna vazio

txboot=matrix(0,R,1)
tyboot=matrix(0,R,1)
tzboot=matrix(0,R,1)
tx1boot=matrix(0,R,1)
tx2boot=matrix(0,R,1)
tx3boot=matrix(0,R,1)
ty1boot=matrix(0,R,1)
ty2boot=matrix(0,R,1)
ty3boot=matrix(0,R,1)
tz1boot=matrix(0,R,1)
tz2boot=matrix(0,R,1)
tz3boot=matrix(0,R,1)

##########################  Amostragem Aleatoria Simples#### ############
for ( i in 1:R)
{
s=S.SI(N,n,e=runif(N))
amostra=dados[s,]
colnames(amostra)=c("x.","y.","z.","x1.","x2.","x3.","y1.","y2.","y3.",
"z1.","z2.","z3.")
names(amostra)

tx[i]=as.numeric(E.SI(N,n,amostra$x.)[,2][1])   #HT ( prim coluna, primei
ty[i]=as.numeric(E.SI(N,n,amostra$y.)[,2][1])   #HT ( prim coluna, primei
tz[i]=as.numeric(E.SI(N,n,amostra$z.)[,2][1])   #HT ( prim coluna, primei

tx1[i]=as.numeric(E.SI(N,n,amostra$x1.)[,2][1])   #HT ( prim coluna, prim
tx2[i]=as.numeric(E.SI(N,n,amostra$x2.)[,2][1])   #HT ( prim coluna, prim
tx3[i]=as.numeric(E.SI(N,n,amostra$x3.)[,2][1])   #HT ( prim coluna, prim

ty1[i]=as.numeric(E.SI(N,n,amostra$y1.)[,2][1])   #HT ( prim coluna, prim
ty2[i]=as.numeric(E.SI(N,n,amostra$y2.)[,2][1])   #HT ( prim coluna, prim

ty3[i]=as.numeric(E.SI(N,n,amostra$y3.)[,2][1])   #HT ( prim coluna, prim

tz1[i]=as.numeric(E.SI(N,n,amostra$z1.)[,2][1])   #HT ( prim coluna, prim
tz2[i]=as.numeric(E.SI(N,n,amostra$z2.)[,2][1])   #HT ( prim coluna, prim
tz3[i]=as.numeric(E.SI(N,n,amostra$z3.)[,2][1])   #HT ( prim coluna, prim
```

```
#Implementação do bootstrap para estimar t boot

tx.b1=matrix(0,B,1)
ty.b2=matrix(0,B,1)
tz.b3=matrix(0,B,1)

tx1.b4=matrix(0,B,1)
tx2.b5=matrix(0,B,1)
tx3.b6=matrix(0,B,1)
ty1.b7=matrix(0,B,1)
ty2.b8=matrix(0,B,1)
ty3.b9=matrix(0,B,1)
tz1.b10=matrix(0,B,1)
tz2.b11=matrix(0,B,1)
tz3.b12=matrix(0,B,1)


for (b in 1:B){
sb1 <- S.WR(length(amostra$x.),m)
sb2 <- S.WR(length(amostra$y.),m)
sb3 <- S.WR(length(amostra$z.),m)

sb4 <- S.WR(length(amostra$x1.),m)
sb5 <- S.WR(length(amostra$x2.),m)
sb6 <- S.WR(length(amostra$x3.),m)
sb7 <- S.WR(length(amostra$y1.),m)
sb8 <- S.WR(length(amostra$y2.),m)
sb9 <- S.WR(length(amostra$y3.),m)
sb10 <- S.WR(length(amostra$z1.),m)
sb11 <- S.WR(length(amostra$z2.),m)
sb12 <- S.WR(length(amostra$z3.),m)


tx.b1[b]=E.SI(N,m,amostra$x.[sb1])[1,2]
ty.b2[b]=E.SI(N,m,amostra$y.[sb2])[1,2]
tz.b3[b]=E.SI(N,m,amostra$z.[sb3])[1,2]

tx1.b4[b]=E.SI(N,m,amostra$x1.[sb4])[1,2]
tx2.b5[b]=E.SI(N,m,amostra$x2.[sb5])[1,2]
tx3.b6[b]=E.SI(N,m,amostra$x3.[sb6])[1,2]
ty1.b7[b]=E.SI(N,m,amostra$y1.[sb7])[1,2]
ty2.b8[b]=E.SI(N,m,amostra$y2.[sb8])[1,2]
ty3.b9[b]=E.SI(N,m,amostra$y3.[sb9])[1,2]
tz1.b10[b]=E.SI(N,m,amostra$z1.[sb10])[1,2]
tz2.b11[b]=E.SI(N,m,amostra$z2.[sb11])[1,2]
tz3.b12[b]=E.SI(N,m,amostra$z3.[sb12])[1,2]
```

```
}
txboot[i]=mean(tx.b1)
tyboot[i]=mean(ty.b2)
tzboot[i]=mean(tz.b3)

tx1boot[i]=mean(tx1.b4)
tx2boot[i]=mean(tx2.b5)
tx3boot[i]=mean(tx3.b6)
ty1boot[i]=mean(ty1.b7)
ty2boot[i]=mean(ty2.b8)
ty3boot[i]=mean(ty3.b9)
tz1boot[i]=mean(tz1.b10)
tz2boot[i]=mean(tz2.b11)
tz3boot[i]=mean(tz3.b12)

}
ttgeral
stats (tx,1.96,ttgeral[1])
stats (txboot,1.96,ttgeral[1])
stats (ty,1.96,ttgeral[2])
stats (tyboot,1.96,ttgeral[2])
stats (tz,1.96,ttgeral[3])
stats (tzboot,1.96,ttgeral[3])

stats (tx1,1.96,ttgeral[1])
stats (tx1boot,1.96,ttgeral[1])
stats (tx2,1.96,ttgeral[1])
stats (tx2boot,1.96,ttgeral[1])
stats (tx3,1.96,ttgeral[1])
stats (tx3boot,1.96,ttgeral[1])
stats (ty1,1.96,ttgeral[2])
stats (ty1boot,1.96,ttgeral[2])
stats (ty2,1.96,ttgeral[2])
stats (ty2boot,1.96,ttgeral[2])
stats (ty3,1.96,ttgeral[2])
stats (ty3boot,1.96,ttgeral[2])
stats (tz1,1.96,ttgeral[3])
stats (tz1boot,1.96,ttgeral[3])
stats (tz2,1.96,ttgeral[3])
stats (tz2boot,1.96,ttgeral[3])
stats (tz3,1.96,ttgeral[3])
stats (tz3boot,1.96,ttgeral[3])


#################### CADASTRO A COM CONTAMINAÇÃO DE 25%##################
aa=x
aa[c(S.SI(N,0.25*N))]=x1[c(S.SI(N,0.25*N))]
bb=x
```

```
bb[c(S.SI(N,0.25*N))]=x2[c(S.SI(N,0.25*N))]
cc=x
cc[c(S.SI(N,0.25*N))]=x3[c(S.SI(N,0.25*N))]
dd=x
dd[c(S.SI(N,0.25*N))]=y1[c(S.SI(N,0.25*N))]
ee=x
ee[c(S.SI(N,0.25*N))]=y2[c(S.SI(N,0.25*N))]
ff=x
ff[c(S.SI(N,0.25*N))]=y3[c(S.SI(N,0.25*N))]
gg=x
gg[c(S.SI(N,0.25*N))]=z1[c(S.SI(N,0.25*N))]
hh=x
hh[c(S.SI(N,0.25*N))]=z2[c(S.SI(N,0.25*N))]
ii=x
ii[c(S.SI(N,0.25*N))]=z3[c(S.SI(N,0.25*N))]


dados11=data.frame(aa,bb,cc,dd,ee,ff,gg,hh,ii)

ttgeral11=colSums(dados)
dados11
head(dados11)

taa=matrix(0,R,1)            # total de HT, com vetor coluna vazio
tbb=matrix(0,R,1)            # total de HT, com vetor coluna vazio
tzz=matrix(0,R,1)            # total de HT, com vetor coluna vazio
tcc=matrix(0,R,1)           # total de HT, com vetor coluna vazio
tdd=matrix(0,R,1)           # total de HT, com vetor coluna vazio
tee=matrix(0,R,1)           # total de HT, com vetor coluna vazio
tff=matrix(0,R,1)           # total de HT, com vetor coluna vazio
tgg=matrix(0,R,1)           # total de HT, com vetor coluna vazio
thh=matrix(0,R,1)           # total de HT, com vetor coluna vazio
tii=matrix(0,R,1)           # total de HT, com vetor coluna vazio

taaboot=matrix(0,R,1)
tbbboot=matrix(0,R,1)
tccboot=matrix(0,R,1)
tddboot=matrix(0,R,1)
teeboot=matrix(0,R,1)
tffboot=matrix(0,R,1)
tggboot=matrix(0,R,1)
thhboot=matrix(0,R,1)
tiiboot=matrix(0,R,1)

for ( i in 1:R)
{
amo3=S.SI(N,n,e=runif(N))
amostra122=dados11[amo3,]
```

```
colnames(amostra122)=c("a1.","b1.","c1.","d1.","e1.","f1.","g1.","h1.","i
names(amostra122)

taa[i]=as.numeric(E.SI(N,n,amostra122$a1.)[,2][1])   #HT ( prim coluna, p
tbb[i]=as.numeric(E.SI(N,n,amostra122$b1.)[,2][1])   #HT ( prim coluna, p
tcc[i]=as.numeric(E.SI(N,n,amostra122$c1.)[,2][1])   #HT ( prim coluna, p
tdd[i]=as.numeric(E.SI(N,n,amostra122$d1.)[,2][1])   #HT ( prim coluna, p
tee[i]=as.numeric(E.SI(N,n,amostra122$e1.)[,2][1])   #HT ( prim coluna, p
tff[i]=as.numeric(E.SI(N,n,amostra122$f1.)[,2][1])   #HT ( prim coluna, p
tgg[i]=as.numeric(E.SI(N,n,amostra122$g1.)[,2][1])   #HT ( prim coluna, p
thh[i]=as.numeric(E.SI(N,n,amostra122$h1.)[,2][1])   #HT ( prim coluna, p
tii[i]=as.numeric(E.SI(N,n,amostra122$i1.)[,2][1])   #HT ( prim coluna, p


#Implementação do bootstrap para estimar t  boot

taa.b1=matrix(0,B,1)
tbb.b2=matrix(0,B,1)
tcc.b3=matrix(0,B,1)
tdd.b4=matrix(0,B,1)
tee.b5=matrix(0,B,1)
tff.b6=matrix(0,B,1)
tgg.b7=matrix(0,B,1)
thh.b8=matrix(0,B,1)
tii.b9=matrix(0,B,1)

for (b in 1:B){
sb11 <- S.WR(length(amostra122$a1.),m)
sb22 <- S.WR(length(amostra122$b1.),m)
sb33 <- S.WR(length(amostra122$c1.),m)
sb44 <- S.WR(length(amostra122$d1.),m)
sb55 <- S.WR(length(amostra122$e1.),m)
sb66 <- S.WR(length(amostra122$f1.),m)
sb77 <- S.WR(length(amostra122$g1.),m)
sb88 <- S.WR(length(amostra122$h1.),m)
sb99 <- S.WR(length(amostra122$i1.),m)

taa.b1[b]=E.SI(N,m,amostra122$a1.[sb11])[1,2]
tbb.b2[b]=E.SI(N,m,amostra122$b1.[sb22])[1,2]
tcc.b3[b]=E.SI(N,m,amostra122$c1.[sb33])[1,2]
tdd.b4[b]=E.SI(N,m,amostra122$d1.[sb44])[1,2]
tee.b5[b]=E.SI(N,m,amostra122$e1.[sb55])[1,2]
tff.b6[b]=E.SI(N,m,amostra122$f1.[sb66])[1,2]
tgg.b7[b]=E.SI(N,m,amostra122$g1.[sb77])[1,2]
thh.b8[b]=E.SI(N,m,amostra122$h1.[sb88])[1,2]
tii.b9[b]=E.SI(N,m,amostra122$i1.[sb99])[1,2]

}
```

```
taaboot[i]=mean(taa.b1)
tbbboot[i]=mean(tbb.b2)
tccboot[i]=mean(tcc.b3)
tddboot[i]=mean(tdd.b4)
teeboot[i]=mean(tee.b5)
tffboot[i]=mean(tff.b6)
tggboot[i]=mean(tgg.b7)
thhboot[i]=mean(thh.b8)
tiiboot[i]=mean(tii.b9)

}

stats (taa,1.96,ttgeral11[1])
stats (taaboot,1.96,ttgeral11[1])
stats (tbb,1.96,ttgeral11[1])
stats (tbbboot,1.96,ttgeral11[1])
stats (tcc,1.96,ttgeral11[1])
stats (tccboot,1.96,ttgeral11[1])
stats (tdd,1.96,ttgeral11[2])
stats (tddboot,1.96,ttgeral11[2])
stats (tee,1.96,ttgeral11[2])
stats (teeboot,1.96,ttgeral11[2])
stats (tff,1.96,ttgeral11[2])
stats (tffboot,1.96,ttgeral11[2])
stats (tgg,1.96,ttgeral11[3])
stats (tggboot,1.96,ttgeral11[3])
stats (thh,1.96,ttgeral11[3])
stats (thhboot,1.96,ttgeral11[3])
stats (tii,1.96,ttgeral11[3])
stats (tiiboot,1.96,ttgeral11[3])
######################### Erro de Mensuração ##############################
######################### Cadastro duplo ##################################
############################GAMMA-HARTLEY#####################################
set.seed(10)
library(boot)
library(TeachingSampling)
library(SunterSampling)
library(samplingbook)
library(samplingVarEst)

#Criando o número de réplicas para o bootsrap
B=500
#Criando o número de réplicas para a simulação de Monte Carlo
R=1000
#Fixando o tamanho da amostra
n=100
# tamanho do cadastro
N=1000
```

```
Na=800
Nb=800
Nab=200

################## funções ####################

#FUNÇÃO INTERESSANTE
#
stats=function(matriz,z,parametro)
{
c=ncol(matriz)
media=c()
vies=c()
viesr=c()
var=c()
desv=c()
ic1=c()
ic2=c()
eqm=c()

for(i in 1:1)
{
media[1]=mean(matriz[,1])
var[1]=var(matriz[,1])
desv[1]=sd(matriz[,1])
vies[1]=media[1]-parametro
viesr[1]=abs((media[1]-parametro)/parametro)*100
eqm[1]=var[1]+vies[1]^2
ic1[1]=media[1]-z*desv[1]
ic2[1]=media[1]+z*desv[1]
}
return(data.frame(media,vies,viesr,desv,eqm,ic1,ic2))

}

amostra=function(N,n,dadoss)
{
s=S.SI(N,n,e=runif(N))
amostr1=dadoss[s,][,c(1,3)]
d=S.SI(N,n,e=runif(N))
amostr2=dadoss[d,][,c(2,3)]
amo=data.frame(amostr1,amostr2)
return(amo)
}

amostrawr=function(namostra,nbootstrap,dadossamostra)
{
s=S.WR(namostra,nbootstrap)
```

```
amostr1=dadossamostra[s,][,c(1,3)]
d=S.WR(namostra,nbootstrap)
amostr2=dadossamostra[d,][,c(2,4)]
amo=data.frame(amostr1,amostr2)
return(cbind(amo[,1],amo[,3],amo[,2],amo[,4]))
}


pest=function(Nab,amostra)
{
yabA=amostra[amostra[,2]==1,][,1]
yabB=amostra[amostra[,4]==1,][,3]
nabA=length(yabA)
nabB=length(yabB)
s2abA=var(yabA)
s2abB=var(yabB)
numerador=Nab^2*(1-nabB/Nab)*s2abB/nabB
denominador=Nab^2*(1-nabA/Nab)*s2abA/nabA+Nab^2*(1-nabB/Nab)*s2abB/nabB
p=numerador/denominador
return(p)
}

estimtotal=function(p,amostra,Na,Nb,Nab)
{
TabA=Nab*mean(amostra[amostra[,2]==1,][,1])
TabB=Nab*mean(amostra[amostra[,4]==1,][,3])
TaA=Na*mean(amostra[amostra[,2]==0,][,1])
TbB=Nb*mean(amostra[amostra[,4]==0,][,3])
Tn=TaA+ (p*TabA)+ (1-p)*TabB + TbB
return(Tn)
}

######################### Erro de Mensuração ##############################
######################### Cadastro duplo ##################################
###################################################################

#### gerando banco X

xb<-rgamma(Na,2)
xa<-rgamma(Nb,2)
xab<-rgamma(Nab,2)
tot=sum(c(xa,xb,xab))

Xb<-c(xb,xab)
Xb
Xa<-c(xa,xab)
Xa
Xi=ifelse(Xa==Xb,1,0)
```

```
banco=data.frame(Xa,Xb,Xi)
head(banco)

## como fazer pra determinar o numero de 1, e 0.

# com erro pequeno

xa1=rgamma(Na,2)+rnorm(Na,0.001,.01)
xb1=rgamma(Nb,2)
xab1=rgamma(Nab,2)+rnorm(Nab,0.001,.01)

Xb1<-c(xb1,xab1)
Xb1
Xa1<-c(xa1,xab1)
Xa1
Xi1=ifelse(Xa1==Xb1,1,0)
banco1=data.frame(Xa1,Xb1,Xi1)

head(banco1)
colSums(banco1)


# com erro medio

xa2=rgamma(Na,2)+rnorm(Na,0.1,.01)
xb2=rgamma(Nb,2)
xab2<-rgamma(Nab,2)+rnorm(Nab,0.1,.01)

Xb2<-c(xb2,xab2)
Xb2
Xa2<-c(xa2,xab2)
Xa2
Xi2=ifelse(Xa2==Xb2,1,0)
banco2=data.frame(Xa2,Xb2,Xi2)
banco2

head(banco2)
colSums(banco2)

# com erro grande

xa3=rgamma(Na,2)+rnorm(Na,0.5,.01)
xb3=rgamma(Nb,2)
xab3=rgamma(Nab,2)+rnorm(Nab,0.5,.01)

Xb3<-c(xb3,xab3)
Xb3
Xa3<-c(xa3,xab3)
```

```
Xa3
Xi3=ifelse(Xa3==Xb3,1,0)
banco3=data.frame(Xa3,Xb3,Xi3)
banco3

bancox.=data.frame(banco,banco1,banco2,banco3)
bancox.
head(bancox.)



tx=matrix(0,R,1)            # total de HT, com vetor coluna vazio
ty=matrix(0,R,1)            # total de HT, com vetor coluna vazio
tz=matrix(0,R,1)            # total de HT, com vetor coluna vazio
tw=matrix(0,R,1)          # total de HT, com vetor coluna vazio
xp=matrix(0,R,1)            #

tx.b1=matrix(0,R,1)             # total de HT, com vetor coluna vazio
ty.b2=matrix(0,R,1)             # total de HT, com vetor coluna vazio
tz.b3=matrix(0,R,1)             # total de HT, com vetor coluna vazio
tw.b4=matrix(0,R,1)           # total de HT, com vetor coluna vazio

txboot=matrix(0,R,1)
tyboot=matrix(0,R,1)
tzboot=matrix(0,R,1)
twboot=matrix(0,R,1)

############################  Amostragem Aleatoria Simples     ###########
#################### ########CONTAMINAÇÃO DE 100%#########################

for ( i in 1:R)
{
x.=amostra(N,n,banco)
xx=data.frame(c(x.$Xa,x.$Xb),c(x.$Xi,x.$Xi.1))
y.=amostra(N,n,banco1)
z.=amostra(N,n,banco2)
w.=amostra(N,n,banco3)

xp=pest(Nab,x.)
yp=pest(Nab,y.)
zp=pest(Nab,z.)
wp=pest(Nab,w.)

#NA=nn-sum(x.$Xi)
#NB=nn-sum(x.$Xi.1)

tx[i]=estimtotal(xp,x.,Na,Nb,Nab)
```

```
ty[i]=estimtotal(yp,y.,Na,Nb,Nab)
tz[i]=estimtotal(zp,z.,Na,Nb,Nab)
tw[i]=estimtotal(wp,w.,Na,Nb,Nab)

#Implementação do bootstrap para estimar t  boot

tx.b1=matrix(0,B,1)
ty.b2=matrix(0,B,1)
tz.b3=matrix(0,B,1)
tw.b4=matrix(0,B,1)



for (b in 1:B){
sb1 <- amostrawr(nrow(x.),500,x.)
sb2 <- amostrawr(nrow(y.),500,y.)
sb3 <- amostrawr(nrow(z.),500,z.)
sb4 <- amostrawr(nrow(w.),500,w.)
xpb=pest(Nab,sb1)
ypb=pest(Nab,sb2)
zpb=pest(Nab,sb3)
wpb=pest(Nab,sb4)

tx.b1[i]=estimtotal(xpb,sb1,Na,Nb,Nab)
ty.b2[i]=estimtotal(ypb,sb2,Na,Nb,Nab)
tz.b3[i]=estimtotal(zpb,sb3,Na,Nb,Nab)
tw.b4[i]=estimtotal(wpb,sb4,Na,Nb,Nab)

}
txboot[i]=mean(tx.b1)
tyboot[i]=mean(ty.b2)
tzboot[i]=mean(tz.b3)
twboot[i]=mean(tw.b4)
}

stats (na.omit(tx),1.96,tot)
stats (na.omit(txboot),1.96,tot)

stats (na.omit(ty),1.96,tot)
stats (na.omit(tyboot),1.96,tot)

stats (na.omit(tz),1.96,tot)
stats (na.omit(tzboot),1.96,tot)

stats (na.omit(tw),1.96,tot)
stats (na.omit(twboot),1.96,tot)

#########################################################################
```

```
#################################### 25%##################################
#########################################################################


ta=matrix(0,R,1)           # total de HT, com vetor coluna vazio
tb=matrix(0,R,1)           # total de HT, com vetor coluna vazio
tc=matrix(0,R,1)           # total de HT, com vetor coluna vazio

ta.b1=matrix(0,B,1)
tb.b2=matrix(0,B,1)
tc.b3=matrix(0,B,1)

taboot=matrix(0,R,1)
tbboot=matrix(0,R,1)
tcboot=matrix(0,R,1)


aa=banco1$Xa1
aa[c(S.SI(N,0.25*N))]=banco1$Xa1[c(S.SI(N,0.25*N))]
banco1.=data.frame(aa,banco1$Xb1,banco1$Xi1)

bb=banco2$Xa2
bb[c(S.SI(N,0.25*N))]=banco2$Xa2[c(S.SI(N,0.25*N))]
banco2.=data.frame(bb,banco2$Xb2,banco2$Xi2)

cc=banco3$Xa3
cc[c(S.SI(N,0.25*N))]=banco3$Xa3[c(S.SI(N,0.25*N))]
banco3.=data.frame(cc,banco3$Xb3,banco3$Xi3)

for ( i in 1:R)
{
a.=amostra(N,n,banco1.)
b.=amostra(N,n,banco2.)
c.=amostra(N,n,banco3.)

ap=pest(Nab,a.)
bp=pest(Nab,b.)
cp=pest(Nab,c.)


ta[i]=estimtotal(ap,a.,Na,Nb,Nab)
tb[i]=estimtotal(bp,b.,Na,Nb,Nab)
tc[i]=estimtotal(cp,c.,Na,Nb,Nab)

#Implementação do bootstrap para estimar t  boot

ta.b1=matrix(0,B,1)
tb.b2=matrix(0,B,1)
```

```
tc.b3=matrix(0,B,1)

for (b in 1:B){
sb1 <- amostrawr(nrow(a.),500,a.)
sb2 <- amostrawr(nrow(b.),500,b.)
sb3 <- amostrawr(nrow(c.),500,c.)

xpb=pest(Nab,sb1)
ypb=pest(Nab,sb2)
zpb=pest(Nab,sb3)
wpb=pest(Nab,sb4)

ta.b1[i]=estimtotal(xpb,sb1,Na,Nb,Nab)
tb.b2[i]=estimtotal(ypb,sb2,Na,Nb,Nab)
tc.b3[i]=estimtotal(zpb,sb3,Na,Nb,Nab)
}

taboot[i]=mean(ta.b1)
tbboot[i]=mean(tb.b2)
tcboot[i]=mean(tc.b3)
}

stats (na.omit(ta),1.96,tot)
stats (na.omit(tb),1.96,tot)
stats (na.omit(tc),1.96,tot)

stats (na.omit(taboot),1.96,tot)
stats (na.omit(tbboot),1.96,tot)
stats (na.omit(tcboot),1.96,tot)

#########################################################################
################################## Cadastro Duplo ########################
#####################################BANKIER#############################
#####################################UNIFORME############################
library(boot)
library(TeachingSampling)
library(SunterSampling)
library(samplingbook)
library(samplingVarEst)

#Criando o número de réplicas para o bootsrap
B=500
#Criando o número de réplicas para a simulação de Monte Carlo
R=1000
#Fixando o tamanho da amostra
n=100
# tamanho do cadastro
N=1000
```

```
##############################FUNÇÃO INTERESSANTE#########################
stats=function(matriz,z,parametro)
{
c=ncol(matriz)
media=c()
vies=c()
var=c()
viesr=c()
desv=c()
ic1=c()
ic2=c()
eqm=c()
cv=c()

for(i in 1:1)
{
media[1]=mean(matriz[,1])
var[1]=var(matriz[,1])
desv[1]=sd(matriz[,1])
vies[1]=media[1]-parametro
viesr[1]=abs((media[1]-parametro)/parametro)*100
cv[1]=desv[1]/media[1]
eqm[1]=var[1]+vies[1]^2
ic1[1]=media[1]-z*desv[1]
ic2[1]=media[1]+z*desv[1]
}
return(data.frame(media,vies,viesr, desv,eqm,cv,ic1,ic2))

}
#função para calcular amostra
amostra=function(N,n,dadoss)
{
s=S.SI(N,n,e=runif(N))
amostr1=dadoss[s,][,c(1,3)]
d=S.SI(N,n,e=runif(N))
amostr2=dadoss[d,][,c(2,3)]
amo=data.frame(amostr1,amostr2)
return(amo)
}
#########################################SIGLAS#################################
#bankier
#nA tamanho da amostra obtida do cadastro A
#N_A tamanho da população do cadastro A
#nb tamanho da amostra obtida no cadastro B
#NB tamanho da população do cadastro B
#sA amostra do cadastro A
#sB amostra do cad B
#dab indicador numerico do dominio ab
```

```
# FUNÇÃO PARA CALCULAR O TOTAL PARA O ESTIMADOR BANKIER
estbank=function(n_A, nB,N_A,NB,sA,sB)
{
domabA=subset(sA,sA[,2]==1)
doma=subset(sA,sA[,2]==0)
domabB=subset(sB,sB[,2]==1)
domb=subset(sB,sB[,2]==0)
fA=n_A/N_A
fB=nB/NB
somapondA=(1/fA)*sum(doma[,1])
somapondabA=(1/(fA+fB))*sum(domabA[,1])
soma1=somapondA+somapondabA
somapondB=(1/fB)*sum(doma[,1])
somapondabB=(1/(fA+fB))*sum(domabB[,1])
soma2=somapondB+somapondabB
est=soma1+soma2
return(est)
}

#funcao que calcula o estimador de HT para o caso do estimador de bankier
# essa funçao altera o E.SI do teachingSample
E.SImod=function(N_A,NB, n_A,nB,nd,y)
{
y<- as.data.frame(y)
val2=nB/NB
pik<- rep(1/(val1+val2),nd)
dk<- 1/pik
ty<-sum(y*dk)
return(ty)
}
# Fixando a Semente
set.seed(10)
Na=800
Nb=800
Nab=200
#### gerando banco para UNIFORME
xb<-runif(Na)
xa<-runif(Nb)
xab<-runif(Nab)
tot=sum(c(xa,xb,xab))

Xb<-c(xb,xab)
Xb
Xa<-c(xa,xab)
Xa
Xi=ifelse(Xa==Xb,1,0)
banco=data.frame(Xa,Xb,Xi)
head(banco)
```

```
# com erro pequeno

xa1=runif(Na)+rnorm(Na,0.001,.01)
xb1=runif(Nb)
xab1=runif(Nab)+rnorm(Nab,0.001,.01)

Xb1<-c(xb1,xab1)
Xb1
Xa1<-c(xa1,xab1)
Xa1
Xi1=ifelse(Xa1==Xb1,1,0)
banco1=data.frame(Xa1,Xb1,Xi1)

head(banco1)
colSums(banco1)


# com erro medio

xa2=runif(Na)+rnorm(Na,0.1,.01)
xb2=runif(Nb)
xab2<-runif(Nab)+rnorm(Nab,0.1,.01)

Xb2<-c(xb2,xab2)
Xb2
Xa2<-c(xa2,xab2)
Xa2
Xi2=ifelse(Xa2==Xb2,1,0)
banco2=data.frame(Xa2,Xb2,Xi2)
banco2

head(banco2)
colSums(banco2)

# com erro grande

xa3=runif(Na)+rnorm(Na,0.5,.01)
xb3=runif(Nb)
xab3=runif(Nab)+rnorm(Nab,0.5,.01)

Xb3<-c(xb3,xab3)
Xb3
Xa3<-c(xa3,xab3)
Xa3
Xi3=ifelse(Xa3==Xb3,1,0)
banco3=data.frame(Xa3,Xb3,Xi3)
banco3
```

```
bancox.=data.frame(banco,banco1,banco2,banco3)
bancox.
head(bancox.)

tx=matrix(0,R,1)           # total de HT, com vetor coluna vazio
ty=matrix(0,R,1)           # total de HT, com vetor coluna vazio
tz=matrix(0,R,1)           # total de HT, com vetor coluna vazio
tw=matrix(0,R,1)          # total de HT, com vetor coluna vazio

tx.b1=matrix(0,R,1)           # total de HT, com vetor coluna vazio
ty.b2=matrix(0,R,1)           # total de HT, com vetor coluna vazio
tz.b3=matrix(0,R,1)           # total de HT, com vetor coluna vazio
tw.b4=matrix(0,R,1)          # total de HT, com vetor coluna vazio

txboot=matrix(0,R,1)
tyboot=matrix(0,R,1)
tzboot=matrix(0,R,1)
twboot=matrix(0,R,1)
############ ######### Amostragem Aleatoria Simples  ######## ###########
############################ alterar de acordo com a pop ################
################## CONTAMINAÇÃO DE 100% NO CADASTRO A.###################
s=S.SI(N,nn,e=runif(N))
for ( i in 1:R)
{
x.=amostra(NN,nn,banco)
xx=data.frame(c(x.$Xa,x.$Xb),c(x.$Xi,x.$Xi.1))
y.=amostra(NN,nn,banco1)
z.=amostra(NN,nn,banco2)
w.=amostra(NN,nn,banco3)

sa1=x.[,c(1,2)]
sb1=x.[,c(3,4)]

sa1y=y.[,c(1,2)]
sb1y=y.[,c(3,4)]

sa1z=z.[,c(1,2)]
sb1z=z.[,c(3,4)]

sa1w=w.[,c(1,2)]
sb1w=w.[,c(3,4)]

tx[i]=estbank(nn,nn,Na,Nb,sa1,sb1)
ty[i]=estbank(nn,nn,Na,Nb,sa1y,sb1y)
tz[i]=estbank(nn,nn,Na,Nb,sa1z,sb1z)
tw[i]=estbank(nn,nn,Na,Nb,sa1w,sb1w)
```

```
#Implementação do bootstrap para estimar t  boot

tx.b1=matrix(0,B,1)
ty.b2=matrix(0,B,1)
tz.b3=matrix(0,B,1)
tw.b4=matrix(0,B,1)

#raz=round(NN/nn)
#b1=rep(x.,raz)

#  Nboot1=length(b1)

for (b in 1:B){
sb1 <- S.WR(nn,B)

amob1=x.[sb1,]
amob2=y.[sb1,]
amob3=z.[sb1,]
amob4=w.[sb1,]

s=S.SI(B,nn,e=runif(N))

amob1.= amob1[s,]

amobb=amob1.[,c(3,4)]
amoaa=amob1.[,c(1,2)]

tx.b1[b]=estbank(nn,nn,Na,Nb,amobb,amoaa)

amob2.= amob2[s,]

amobb2=amob2.[,c(3,4)]
amoaa2=amob2.[,c(1,2)]

ty.b2[b]=estbank(nn,nn,Na,Nb,amobb2,amoaa2)

amob3.= amob3[s,]

amobb3=amob3.[,c(3,4)]
amoaa3=amob3.[,c(1,2)]

tz.b3[b]=estbank(nn,nn,Na,Nb,amobb3,amoaa3)

amob4.= amob4[s,]

amobb4=amob4.[,c(3,4)]
amoaa4=amob4.[,c(1,2)]
```

```
tw.b4[b]=estbank(nn,nn,Na,Nb,amobb4,amoaa4)


}
txboot[i]=mean(tx.b1)
tyboot[i]=mean(ty.b2)
tzboot[i]=mean(tz.b3)
twboot[i]=mean(tw.b4)
}

stats (na.omit(tx),1.96,tot)
stats (na.omit(txboot),1.96,tot)
stats (na.omit(ty),1.96,tot)
stats (na.omit(tyboot),1.96,tot)
stats (na.omit(tz),1.96,tot)
stats (na.omit(tzboot),1.96,tot)
stats (na.omit(tw),1.96,tot)
stats (na.omit(twboot),1.96,tot)
##########################################################################
############## ###CONTAMINAÇÃO DE 25% NO CAD A############################
##########################################################################

ta=matrix(0,R,1)           # total de HT, com vetor coluna vazio
tb=matrix(0,R,1)           # total de HT, com vetor coluna vazio
tc=matrix(0,R,1)           # total de HT, com vetor coluna vazio

ta.b1=matrix(0,B,1)
tb.b2=matrix(0,B,1)
tc.b3=matrix(0,B,1)

taboot=matrix(0,R,1)
tbboot=matrix(0,R,1)
tcboot=matrix(0,R,1)


aa=banco1$Xa1
aa[c(S.SI(N,0.25*N))]=banco1$Xa1[c(S.SI(N,0.25*N))]
banco1.=data.frame(aa,banco1$Xb1,banco1$Xi1)

bb=banco2$Xa2
bb[c(S.SI(N,0.25*N))]=banco2$Xa2[c(S.SI(N,0.25*N))]
banco2.=data.frame(bb,banco2$Xb2,banco2$Xi2)

cc=banco3$Xa3
cc[c(S.SI(N,0.25*N))]=banco3$Xa3[c(S.SI(N,0.25*N))]
banco3.=data.frame(cc,banco3$Xb3,banco3$Xi3)

for ( i in 1:R)
```

```
{
a.=amostra(NN,nn,banco1.)
b.=amostra(NN,nn,banco2.)
c.=amostra(NN,nn,banco3.)

saa1=a.[,c(1,2)]
sbb1=a.[,c(3,4)]

saa1y=b.[,c(1,2)]
sbb1y=b.[,c(3,4)]

saa1z=c.[,c(1,2)]
sbb1z=c.[,c(3,4)]


ta[i]=estbank(nn,nn,Na,Nb,saa1,sbb1)
tb[i]=estbank(nn,nn,Na,Nb,saa1y,sbb1y)
tc[i]=estbank(nn,nn,Na,Nb,saa1z,sbb1z)

#Implementação do bootstrap para estimar t  boot

for (b in 1:B){
sb1 <- S.WR(nn,B)

amoa=a.[sb1,]
amob=b.[sb1,]
amoc=c.[sb1,]

amoa.= amoa[s,]

amoa1=amoa.[,c(3,4)]
amoa2=amoa.[,c(1,2)]

ta.b1[b]=estbank(nn,nn,Na,Nb,amoa1,amoa2)

amob.= amob[s,]

amob1=amob.[,c(3,4)]
amob2=amob.[,c(1,2)]

tb.b2[b]=estbank(nn,nn,Na,Nb,amob1,amob2)

amoc.= amoc[s,]

amoc1=amoc.[,c(3,4)]
amoc2=amoc.[,c(1,2)]

tc.b3[b]=estbank(nn,nn,Na,Nb,amoc1,amoc2)
```

```
}

taboot[i]=mean(ta.b1)
tbboot[i]=mean(tb.b2)
tcboot[i]=mean(tc.b3)
}

stats (na.omit(ta),1.96,tot)
stats (na.omit(tb),1.96,tot)
stats (na.omit(tc),1.96,tot)

stats (na.omit(taboot),1.96,tot)
stats (na.omit(tbboot),1.96,tot)
stats (na.omit(tcboot),1.96,tot)
```