

NELSON DE OLIVEIRA DOKI

DESENVOLVIMENTO DE METODOLOGIA ESTATÍSTICA PARA
ESTIMAÇÃO DE NÚMERO DE EQUIPAMENTOS DE USO
COMPARTILHADO

CURITIBA
31 DE JULHO DE 2002

NELSON DE OLIVEIRA DOKI

**DESENVOLVIMENTO DE METODOLOGIA ESTATÍSTICA PARA
ESTIMAÇÃO DE NÚMERO DE EQUIPAMENTOS DE USO
COMPARTILHADO**

Dissertação apresentada como requisito parcial à
obtenção do grau de Mestre em Métodos Numéricos
em Engenharia, Curso de Pós-Graduação em
Engenharia, Área de Concentração em Programação
Matemática, Universidade Federal do Paraná.
Orientador: Prof. Dr. Anselmo Chaves Neto

CURITIBA

31 de julho de 2002

DESENVOLVIMENTO DE METODOLOGIA ESTATÍSTICA PARA ESTIMAÇÃO DE
EQUIPAMENTOS DE USO COMPARTILHADO

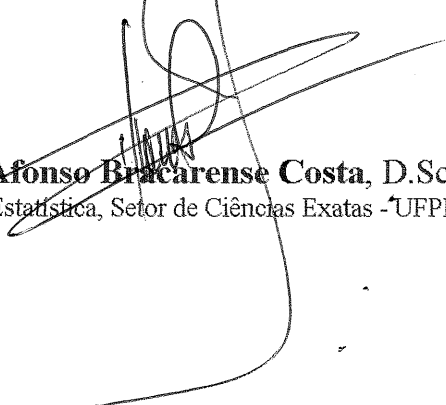
Dissertação aprovada como requisito parcial à obtenção do grau de Mestre
no Curso de Pós-Graduação em Métodos Numéricos em Engenharia, da
Universidade Federal do Paraná, pela Comissão formada pelos professores:



Orientador: Prof. Anselmo Chaves Neto, D. Sc.
Departamento de Estatística, Setor de Ciências Exatas - UFPR.



Prof. Marco Antonio Barbosa Cândido, D. Eng.
Programa de Pós-Graduação em Engenharia de Produção de Sistemas – PPGESP/PUC/PR



Prof. Paulo Afonso Bracarense Costa, D.Sc.
Departamento de Estatística, Setor de Ciências Exatas - UFPR.

Curitiba, 31 de Julho de 2002

NELSON DE OLIVEIRA DOKI

**DESENVOLVIMENTO DE METODOLOGIA ESTATÍSTICA PARA ESTIMAÇÃO DE
NÚMERO DE EQUIPAMENTOS DE USO COMPARTILHADO**

Dissertação aprovada como requisito parcial à obtenção do grau de Mestre no Curso de Pós-Graduação em Métodos Numéricos em Engenharia, da Universidade Federal do Paraná, pela Comissão formada pelos professores:

Orientador: **Prof. Anselmo Chaves Neto, D. Sc.**
**DEPARTAMENTO DE ESTATÍSTICA, SETOR DE CIÊNCIAS EXATAS -
UFPR.**

Prof. Marco Antonio Barbosa Cândido, D. Eng.
**PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE
PRODUÇÃO DE SISTEMAS – PPGESP/PUC/PR.**

Prof. Paulo Afonso Bracarense Costa, D.Sc.
**DEPARTAMENTO DE ESTATÍSTICA, SETOR DE CIÊNCIAS EXATAS -
UFPR.**

Curitiba, 31 de julho de 2002

AGRADECIMENTOS

Agradeço, em primeiro lugar, a Deus, pela existência e por todas as oportunidades que me foram concedidas na vida. Ao meu orientador e professor, Dr. Anselmo Chaves Neto, por sua orientação segura, paciência e amizade durante o desenvolvimento deste trabalho. Aos meus professores do curso de Pós Graduação CPGMNE e ao pessoal de apoio do curso, especialmente à Maristela e ao Eliseu. Aos diversos colegas, tanto de curso como de trabalho, aos quais, muito embora seus nomes não estejam explicitados aqui - para não correr o risco de esquecer algum - vão também os meus agradecimentos, por algum ou por diversos tipos de apoio e/ou de encorajamento em alguma ou em diversas fases do desenvolvimento deste trabalho. Aos professores do curso de Processamento de Dados do CEFET/PR, Unidade de Pato Branco, particularmente a Homero Francisco Bertol, Edilson Pontarolo, Éden Ricardo Dosciatti e José Donizeti de Lima, pelo auxílio que me prestaram em informática. Aos estudantes do curso de Processamento de Dados, Anderson Sauthier e Éder, pela ajuda nas diversas fases do desenvolvimento do programa computacional, elaborado para a obtenção dos resultados que fazem parte deste trabalho. Agradeço especialmente ao professor Jorge Jamhour, também do CEFET-PB, pela inestimável ajuda prestada na formatação desta dissertação. Finalmente desejo agradecer à direção do CEFET-PB e à Coordenação do Curso de Matemática pelo apoio irrestrito durante esse período de estudos.

DEDICATÓRIA

Dedico este trabalho à memória de Merchides que, por seu exemplo, me ensinou a amar os livros. À Nair, que nunca mediu esforços para que os filhos pudessem estudar. À memória da Vanessa que, muito embora com passagem meteórica por esta vida, deixou marcas indeléveis em minha lembrança. À Micheli, à Viviane e as suas extensões e ao Marcos, pelo que aprendi e pelo que continuo aprendendo com as personalidades e maneira de ser de cada um deles.

O êxito traz consigo os sacrifícios e esforços que não aparecem às vistas dos outros.

“Fins diferentes requerem meios diferentes e estruturas lógicas diferentes.”

“Muito embora as técnicas sejam importantes... saber quando usá-las e por que usá-las é ainda mais importante.”

“No fim das contas, não vale a pena a um estatístico enganar a si mesmo ou aos seus clientes.”

RESUMO

Existem muitas situações nas quais é necessário estimar-se uma quantidade em relação a outra, como por exemplo, a quantificação de um certo produto, o tamanho da população, os gastos públicos em um Estado, em um determinado ano considerando os dados disponíveis das diversas localidades que compõem a população total, em algum ano anterior. Neste trabalho, desenvolve-se uma metodologia estatística para estimar o número de equipamentos de uso compartilhado no Estado, aplicando aos métodos de estimação apropriados, técnicas de reamostragem, como exemplo de uma metodologia que pode ser aplicada em um número muito grande de situações. Agrupam-se as localidades semelhantes aplicando-se Análise de Agrupamento, dimensiona-se o tamanho da amostra de localidades aplicando-se a Teoria da Amostragem e aplica-se o estimador da razão para estimar o número de equipamentos de uso compartilhado. Finalmente, faz-se uma simulação para identificar o melhor dos estimadores, clássicos, Bootstrap ou Jackknife. Aplicam-se, também, as técnicas de computação intensiva, nas suas várias abordagens.

ABSTRACT

There are many situations in which it is necessary to estimate a quantity against other, as for example, the amount of a certain product, the size of the population, the public expenses in a State, at a certain year considering the available data from the different places which compose the total population, at any previous year. In this work, a statistical methodology is developed to estimate the number of equipments of shared use in the State, applying to the appropriate methods of estimation, techniques of resampling, as an example of a methodology that can be applied to a large number of situations. The similar localities are grouped by the application of Grouping Analysis, it is determined the sampling size of localities by applying the Sampling Theory and the ratio estimator is applied to estimate the number of equipments of shared use. Finally, it is held a simulation to identify the best estimator, classics, Bootstrap or Jackknife. Intensive computing techniques are also applied in their several approaches.

SUMÁRIO

RESUMO	vii
ABSTRACT	viii
ÍNDICE DE TABELAS.....	xii
ÍNDICE DE FIGURAS	xiii
ÍNDICE DE QUADROS	xv
ÍNDICE DE GRÁFICOS.....	xvi
INTRODUÇÃO	1
1.1. O PROBLEMA.....	1
1.2. OBJETIVOS	1
1.3. JUSTIFICATIVA.....	2
1.4. ESTRUTURA DO TRABALHO	2
2. REVISÃO BIBLIOGRÁFICA	3
2.1. A TEORIA DE AGRUPAMENTOS	3
2.1.1. <i>Introdução</i>	3
2.1.2. <i>Agrupamentos Hierárquicos</i>	4
2.1.3. <i>Agrupamentos Não Hierárquicos</i>	5
2.1.3.1. O Método das k-médias.....	6
2.2. A TEORIA DA AMOSTRAGEM	11
2.2.1. <i>Introdução</i>	11
2.2.2. <i>Tipos de Amostras</i>	12
2.2.2.1. Amostras Aleatórias Simples	13
2.2.2.2. Algumas propriedades das amostras aleatórias simples:	13
2.2.2.2.1- Esperanças e Variâncias de amostras aleatórias simples	14
2.2.2.3. Considerações sobre o dimensionamento da amostra	19
2.2.2.4. O tamanho da amostra	21
2.3. O ESTIMADOR DA RAZÃO	23
2.3.1- <i>Introdução</i>	23
2.3.2. <i>Esperança e Variância do Estimador da Razão</i>	23
2.3.3. <i>Estimador da Razão Separado</i>	26
2.3.4. <i>Estimador da Razão Combinado</i>	28
2.3.5. <i>Estimador da Razão Não Viciado</i>	29
2.3.6. <i>Intervalo de Confiança para os estimadores da Razão</i>	30
2.4. EXPANSÃO SIMPLES	31
2.5. REGRESSÃO LINEAR.....	33
2.5.1 <i>Introdução</i>	33
2.5.2 <i>Estrutura Probabilística</i>	33
2.5.3 <i>Análise da Variância da Regressão</i>	34

2.5.4 Coeficiente de Correlação Múltipla ao Quadrado (R^2)	36
2.6. TÉCNICAS DE REAMOSTRAGENS	37
2.6.1 Introdução.....	37
2.6.2 O Método Monte Carlo	37
2.6.2.1. Exigências de um procedimento Monte Carlo:.....	40
2.6.3. O Jackknife	41
2.6.3.1 Médias, Erro Padrão e Vícios.....	42
2.6.3.2. O Jackknife e Regressão Linear	43
2.6.4. O BOOTSTRAP	46
2.6.4.1. Introdução	46
2.6.4.2. O Procedimento Bootstrap	48
2.6.4.3. Melhor Estimativa Bootstrap do Vício (BBBE).....	49
2.6.4.4. Intervalo de Confiança Bootstrap, Acelerado com correção do Vício (Bca).....	51
2.6.4.5. Estimativa Jackknife do Vício Bootstrap (JAB – Jackknife After Bootstrap).....	52
2.6.4.6. Bootstrap e Regressão Linear.....	53
2.6.4.7. Redução de Vício em Amostras Bootstrap	55
2.6.4.8. O Bootstrap e o Estimador da Razão	60
3. MATERIAL E MÉTODO	62
3.1. INTRODUÇÃO.....	62
3.2. MATERIAL	62
3.3. METODOLOGIA.....	63
3.4. PROGRAMA COMPUTACIONAL	74
3.4.1. Introdução.....	74
4. RESULTADOS E DISCUSSÃO	78
4.1. DISCUSSÃO DOS RESULTADOS DAS SIMULAÇÕES, PARA CADA VARIÁVEL ANALISADA:.....	78
4.1.1. Discussão dos resultados das simulações para a variável NC_AES:....	79
4.1.2. Discussão dos resultados das simulações para a variável NC_COM:..	80
4.1.3. Discussão dos resultados das simulações para a variável NC_ILU:	82
4.1.4. Discussão dos resultados das simulações para a variável NC_IND:....	84
4.1.5. Discussão dos resultados das simulações para a variável NC_POD; ..	85
4.1.6. Discussão dos resultados das simulações para a variável NC_PRO;...	86
4.1.7. Discussão dos resultados das simulações para a variável NC_RES;....	87
4.1.8. Discussão dos resultados das simulações para a variável NC_RUR;...	88
4.1.9. Discussão dos resultados das simulações para a variável PR:	90
4.1.10. Discussão dos resultados das simulações para a variável PU;.....	91
5. CONCLUSÃO	93
REFERENCIAS BIBLIOGRÁFICA	95
ANEXOS	97

ANEXO 1 – EXEMPLO DE CÁLCULO DE ÁREA SIMPLES PELO MÉTODO MONTE CARLO	98
ANEXO 2 – AJUDA PARA OPERAÇÃO DO PROGRAMA EEC (ESTIMAÇÃO DE NÚMERO DE EQUIPAMENTOS DE USO COMPARTILHADOS).	99
ANEXO 3 –QUADROS DE ANÁLISES PRÉ-AMOSTRAIS.....	98
ANEXO 4 – RESULTADOS DAS ESTIMATIVAS CLÁSSICAS BOOTSTRAP E JACKKNIFE E SIMULAÇÕES MONTE CARLO.	113
ANEXO 5 – GRÁFICOS GERADOS EM PROGRAMA MICROSOFT EXCEL A PARTIR DOS RESULTADOS OBTIDOS COM O PROGRAMA EEC.....	116

ÍNDICE DE TABELAS

Tabela 01: RESULTADOS DAS ESTIMATIVAS: CLÁSSICAS, BOOTSTRAP, JACKKNIFE E SIMULAÇÕES MONTE CARLO	113
---	-----

ÍNDICE DE FIGURAS

Figura 01: Dendograma	5
Figura 02: FLUXOGRAMA – Método das k Médias	7
Figura 03: Curvas de Distribuições Normais para ilustração de Exatidão, Precisão e Vício, amostrais.	20
Figura 04: Relações entre erros e vícios das distribuições amostrais A, B, C e D, da Figura 03, acima.	20
Figura 05: Fluxograma Geral.....	77
Figura 06: Tela inicial do programa	100
Figura 07: Tela inicial do programa	100
Figura 08: Descrição das variáveis.....	101
Figura 09: Continuação descrição das variáveis	101
Figura 10: Continuação descrição das variáveis	102
Figura 11: Resultado do agrupamento.....	103
Figura 12: Resultado estratificação	103
Figura 13: Análise pré-amostal	104
Figura 14: Dados pós-amostragem	104
Figura 15: Continuação dados pós-amostragem	105
Figura 16: Resultado dos cálculos do agrupamento.....	105
Figura 17: Resultado dos cálculos estratificação	106
Figura 18: Demais telas	106
Figura 19: Estimação	107
Figura 20: Estimação (continuação).....	107
Figura 21: Tela de apresentação dos gráficos das estimativas do número de localidades estimadas dentro do intervalo do erro relativo percentual	

fixado pelo nível de confiança da solução. Dentro de cada grupo e na população.	108
Figura 22: Gráficos das estimativas do número de localidades estimadas dentro do intervalo do erro relativo percentual fixado pelo nível de confiança da solução. Dentro de cada estrato e na população.	108
Figura 23: Gráficos comparativos das estimativas populacionais de cada estimador.	109

ÍNDICE DE QUADROS

QUADRO 00: ANÁLISE PRÉ-AMOSTRAL, das variáveis NC_AES, NC_ILU, NC_PRO, NC_COM.....	110
QUADRO 01: ANÁLISE PRÉ-AMOSTRAL, da variável NC_AES.....	110
QUADRO 02: ANÁLISE PRÉ-AMOSTRAL, da variável NC_ILU.	110
QUADRO 03: ANÁLISE PRÉ-AMOSTRAL, da variável NC_POD.	110
QUADRO 04: ANÁLISE PRÉ-AMOSTRAL, da variável NC_PRO.	111
QUADRO 05: ANÁLISE PRÉ-AMOSTRAL, da variável NC_RES.	111
QUADRO 06: ANÁLISE PRÉ-AMOSTRAL, da variável NC_RUR.	111
QUADRO 07: ANÁLISE PRÉ-AMOSTRAL, da variável PU.....	112

ÍNDICE DE GRÁFICOS

Gráfico 00.01: Frequências dos pontos amostrais das amostras Jackknife, acima. .	45
Gráfico 00.02: Frequência dos pontos amostrais originais nas nove amostras Bootstrap Simples, acima.	58
Gráfico 00.03: Frequências dos pontos amostrais originais nas nove amostras Bootstrap Balanceado, acima.....	60
GRÁFICO 01.01: Médias dos erros relativos percentuais em valores absolutos de 25 rodadas para simulações de 75_76%, 75_80% e 75_90%, de compartilhados. Variável: NC_AES.....	117
GRÁFICO 01.02: Erro Padrão dos erros relativos percentuais em valores absolutos das estimativas populacionais das 75 rodadas de simulações. Variável: NC_AES.....	117
GRÁFICO 01.03: Amplitude de variação dos erros relativos percentuais em valores absolutos das estimativas populacionais das 75 simulações. Variável: NC_AES.....	117
GRÁFICO 01.04: Média do número de localidades estimadas com erro relativo percentual em valor absoluto menor ou igual a 5, para 25 simulações. Variável: NC_AES.....	118
GRÁFICO 01.05: Erro Padrão do número médio de localidades estimadas com erro relativo percentual em valor absoluto menor ou igual a 5, para as 75 simulações. Variável: NC_AES.....	118
GRÁFICO 01.06: Amplitude de variação do número de localidades estimadas com erros relativos percentuais em valores absolutos menores ou iguais a 5, para as 75 simulações. Variável: NC_AES.....	118
GRÁFICO 01.07: Cobertura do valor simulado das estimativas populacionais, em cada uma das 25 simulações. Variável: NC_AES.....	119
GRÁFICO 01.08: Erro Padrão do número de vezes que o intervalo cobre o valor simulado das estimativas populacionais, em 75 simulações. Variável: NC_AES.....	119
GRÁFICO 01.09: Amplitude de variação do número de vezes que o intervalo cobre o valor simulado, da estimativa populacional, para as 75 simulações. Variável: NC_AES.....	119

GRÁFICO 02.01: Médias dos erros relativos percentuais em valores absolutos, das estimativas populacionais, de 25 simulações, para 75_76%, 75_80% e 75_90% respectivamente. Variável: NC_COM.	120
GRÁFICO 02.02: Erro Padrão dos erros relativos percentuais em valores absolutos das estimativas populacionais de 75 simulações. Variável: NC_COM.	120
GRÁFICO 02.03: Amplitude de variação dos erros relativos percentuais em valores absolutos das estimativas populacionais das 75 simulações. Variável: NC_COM.	120
GRÁFICO 02.04: Média do número de localidades estimadas com erro relativo percentual em valor absoluto menor ou igual a 5, para as 25 simulações. Variável: NC_COM.	121
GRÁFICO 02.05: Erro Padrão do número de localidades estimadas com erro relativo percentual em valor absoluto menor ou igual a 5, para as 75 simulações. Variável: NC_COM.	121
GRÁFICO 02.06: Amplitude de variação do número de localidades estimadas com erro relativo percentual menor ou igual a 5. Variável: NC_COM.	121
GRÁFICO 02.07: Número de vezes que o intervalo definido pelos limites inferior e superior das estimativas de cada estimador, cobre o valor simulado populacional, em cada grupo de 25 rodadas de simulações, para as proporções de compartilhados: 75_76%, 75_80% e 75_90%. Variável: NC_COM.	122
GRÁFICO 02.08: Erro padrão do número máximo de vezes que o intervalo definido pelos limites inferior e superior das estimativas de cada estimador cobre o valor simulado populacional, em cada grupo de 25 rodadas de simulações, para as proporções de compartilhados: 75_76%, 75_80% e 75_90%. Variável: NC_COM.	122
GRÁFICO 02.09: Amplitude de variação dos números máximos de vezes que o intervalo definido pelos limites inferior e superior das estimativas de cada estimador, cobre o valor simulado populacional, em cada grupo de 25 rodadas de simulações, para as proporções de compartilhados: 75_76%, 75_80% e 75_90%. Variável: NC_COM.	122
GRÁFICO 03.01: Média dos erros relativos percentuais em valores absolutos das estimativas populacionais, de 25 simulações, para: 75_76%, 75_80% e 75_90%, respectivamente. Variável: NC_ILU.	123
GRÁFICO 03.02: Erro Padrão dos erros relativos percentuais em valores absolutos das estimativas populacionais, para as 75 simulações. Variável: NC_ILU.	123

GRÁFICO 03.03: Amplitude de variação dos erros relativos percentuais em valores absolutos das estimativas populacionais, para as 75 simulações. Variável: NC_ILU.....	123
GRÁFICO 03.04: Número de localidades estimadas com erro relativo percentual em valor absoluto menor ou igual a 5. Variável: NC_ILU.....	124
GRÁFICO 03.05: Erro Padrão do número de localidades estimativas com erro relativo percentual em valor absoluto menor ou igual a 5, para as 75 simulações. Variável: NC_ILU.	124
GRÁFICO 03.06: Amplitude de variação do número de localidades estimadas com erro relativo percentual em valor absoluto menor ou igual a 5. Variável: NC_ILU.....	124
GRÁFICO 03.07: Cobertura do valor simulado das estimativas populacionais, em cada uma das 25 simulações. Variável: NC_ILU.	125
GRÁFICO 03.08: Erro Padrão do número de vezes que o intervalo cobre os valores simulados das estimativas populacionais, em 75 simulações. Variável: NC_ILU.....	125
GRÁFICO 03.09: Amplitude de variação do número de vezes que o intervalo cobre o valor simulado da estimativa populacional, considerando-se os três grupos de simulações, 75_76%, 75_80% e 75_90% simultaneamente, cada um com 25 simulações. Variável: NC_ILU.....	125
GRÁFICO 03.10: Média dos erros relativos percentuais em valores absolutos das estimativas populacionais, de 25 simulações, para: 75_76%, 75_80% e 75_90%, respectivamente. Variável: NC_ILU, eliminado o primeiro estrato.	126
GRÁFICO 03.11: Erro Padrão dos erros relativos percentuais em valores absolutos das estimativas populacionais, para as 75 simulações. Variável: NC_ILU, eliminado o primeiro estrato.	126
GRÁFICO 03.12: Amplitude de variação dos erros relativos percentuais em valores absolutos das estimativas populacionais, para as 75 simulações. Variável: NC_ILU, eliminado o primeiro estrato.	126
GRÁFICO 03.13: Número médio de localidades estimadas com erro relativo percentual em valor absoluto menor ou igual a 5. Variável: NC_ILU, eliminado o primeiro estrato.	127
GRÁFICO 03.14: Erro Padrão do número médio de localidades estimadas com erro relativo percentual em valor absoluto menor ou igual a 5. Variável: NC_ILU, eliminado o primeiro estrato.	127

GRÁFICO 03.15: Amplitude de variação do número de localidades estimadas com erro relativo percentual em valor absoluto menor ou igual a 5. Variável: NC_ILU, eliminado o primeiro estrato.	127
GRÁFICO 03.16: Numero de vezes que o valor simulado populacional é coberto pelas estimativas populacionais, em dez rodadas. Variável: NC_ILU, eliminado o primeiro estrato.	128
GRÁFICO 03.17: Erro Padrão dos números máximos de vezes que o valor simulado populacional é coberto pelas estimativas populacionais. Variável: NC_ILU, eliminado o primeiro estrato.	128
GRÁFICO 03.18: Amplitude de variação do número de vezes que o valor simulado populacional é coberto pelas estimativas populacionais. Variável : NC_ILU, eliminado o primeiro estrato.	128
GRÁFICO 04.01: Médias dos erros relativos percentuais em valores absolutos, das estimativas populacionais, de 25 simulações, para: 75_76%, 75_8-% e 75_90%, respectivamente. Variável: NC_IND.	129
GRÁFICO 04.02: Erro Padrão dos erros relativos percentuais em valores absolutos das estimativas populacionais de 75 simulações. Variável: NC_IND.	129
GRÁFICO 04.03: Amplitude de variação dos erros relativos percentuais em valores absolutos das estimativas populacionais das 75 simulações. Variável: NC_IND.	129
GRÁFICO 04.04: Média do número de localidades estimadas com erro relativo menor ou igual a 5, para 20 simulações. Variável: NC_IND.	130
GRÁFICO 04.05: Erro Padrão do número médio de localidades estimadas com erro relativo percentual em valor absoluto menor ou igual a 5. Variável: NC_IND.	130
GRÁFICO 04.06: Amplitude de variação do número de localidades estimadas com erros relativos percentuais em valores absolutos menores ou iguais a 5. Variável: NC_IND.	130
GRÁFICO 04.07: Cobertura do valor simulado das estimativas populacionais, em cada uma das 20 simulações. Variável: NC_IND.	131
GRÁFICO 04.08: Erro Padrão do número de vezes que o intervalo cobre os valores simulados, das estimativas populacionais, em 60 simulações. Variável : NC_IND.	131
GRÁFICO 04.09: Amplitude de variação do número de vezes que o intervalo cobre os valores simulados das estimativas populacionais, para 60 simulações. Variável: NC_IND.	131

GRÁFICO 05.01: Médias dos erros relativos percentuais em valores absolutos das estimativas populacionais de 25 simulações, para: 75_76%, 75_80% e 75_90%, respectivamente. Variável: NC_POD.	132
GRÁFICO 05.02: Erro Padrão dos erros relativos percentuais em valores absolutos das estimativas populacionais de 75 simulações. Variável: NC_POD.	132
GRÁFICO 05.03: Amplitude de variação dos erros relativos percentuais em valores absolutos das estimativas populacionais das 75 simulações. Variável: NC_POD.	132
GRÁFICO 05.04: Número de localidades estimadas com erro relativo percentual em valor absoluto menor ou igual a 5. Variável: NC_POD.	133
GRÁFICO 05.05: Erro Padrão do número médio de localidades estimadas com erro relativo percentual menor ou igual a 5, para as 75 simulações. Variável: NC_POD.	133
GRÁFICO 05.06: Amplitude de variação do número de localidades estimadas com erro relativo percentual em valor absoluto menor ou igual a 5, em 75 simulações. Variável: NC_POD.	133
GRÁFICO 05.07: Cobertura do valor simulado das estimativas populacionais em valores absolutos, em cada uma das 21 simulações. Variável: NC_POD.	134
GRÁFICO 05.08: Erro Padrão do número de vezes que o intervalo cobre os valores simulados das estimativas populacionais, em 63 simulações. Variável: NC_POD.	134
GRÁFICO 05.09: Amplitude de variação do número de vezes que o intervalo cobre o valor simulado da estimativa populacional, para as 63 simulações. Variável: NC_POD.	134
GRÁFICO 06.01: Médias dos erros relativos percentuais em valores absolutos, das estimativas populacionais, para cada grupo de simulações: 75_76%, 75_80% e 75_90% respectivamente, cada uma com 25 simulações cada. Variável: NC_PRO.	135
GRÁFICO 06.02: Erro Padrão dos erros relativos percentuais em valor absoluto das estimativas populacionais, das 75 simulações, consideradas simultaneamente. Variável: NC_PRO.	135
GRÁFICO 06.03: Amplitude de variação dos erros relativos percentuais em valores absolutos das estimativas populacionais, das 75 simulações consideradas simultaneamente. Variável: NC_PRO.	135

GRÁFICO 06.04: Número médio de localidades estimadas com erro relativo percentual em valor absoluto menor ou igual a 5, para 25 simulações. Variável: NC_PRO.	136
GRÁFICO 06.5: Erro Padrão do número médio de localidades estimadas com erro relativo percentual em valor absoluto, menor ou igual a 5. Variável: NC_PRO.....	136
GRÁFICO 06.06: Amplitude de variação do número de localidades estimadas com erro relativo percentual em valor absoluto menor ou igual a 5, para as 75 simulações, consideradas simultaneamente. Variável : NC_PRO.	136
GRÁFICO 06.07: Cobertura dos valores simulados das estimativas populacionais, em cada grupo de simulação; 75_76%, 75_80% e 75_90%, respectivamente, com 25 simulações cada. Variável: NC_PRO.	137
GRÁFICO 06.08: Erro Padrão do número médio de vezes que o intervalo cobre os valores simulados das estimativas populacionais, em 63 simulações. Variável: NC_PRO.	137
GRÁFICO 06.09: Amplitude de variação do número de vezes que o intervalo cobre o valor simulado da estimativa populacional, para as 75 simulações. Considerou-se o mínimo e o máximo entre os três grupos de simulações. Variável: NC_PRO.	137
GRÁFICO 06.10: Médias dos erros relativos percentuais em valores absolutos das estimativas populacionais. Variável: NC_PRO, eliminado o primeiro estrato.	138
GRÁFICO 06.11: Erro Padrão da média de variação dos erros relativos percentuais em valores absolutos das estimativas populacionais. Variável: NC_PRO, eliminado o primeiro estrato.	138
GRÁFICO 06.12: Amplitude de variação dos erros relativos percentuais em valores absolutos das estimativas populacionais. Variável: NC_PRO, eliminado o primeiro estrato.	138
GRÁFICO 06.13: Número médio de localidades estimadas com erro relativo percentual em valor absoluto menor ou igual a 5. Variável: NC_PRO, eliminado o primeiro estrato.	139
GRÁFICO 06.14: Erro Padrão do número de localidades estimadas com erro relativo percentual em valor absoluto menor ou igual a 5. Variável: NC_PRO, eliminado o primeiro estrato.	139
GRÁFICO 06.15: Amplitude de variação do número médio de localidades estimadas com erro relativo percentual menor ou igual a 5. Variável: NC_PRO, eliminado o primeiro estrato.	139

GRÁFICO 06.16: Número de vezes que o valor simulado populacional é coberto pelas estimativas populacionais.Variável: NC_PRO, eliminado o primeiro estrato.	140
GRÁFICO 06.17: Erro Padrão do número de vezes que o valor simulado populacional é coberto pelas estimativas populacionais.Variável: NC_PRO, eliminado o primeiro estrato.	140
GRÁFICO 06.18: Amplitude de variação do número de vezes que o valor simulado populacional é coberto pelas estimativas populacionais.Variável: NC_PRO, eliminado o primeiro estrato.	140
GRÁFICO 07.01: Média dos erros relativos percentuais em valores absolutos das 25 simulações. Variável: NC_RES.....	141
GRÁFICO 07.02: Erro Padrão dos erros relativos percentuais em valores absolutos das 75 simulações. Variável: NC_RES.	141
GRÁFICO 07.03: Amplitude de variação dos erros relativos percentuais em valores absolutos das 75 simulações. Variável: NC_RES.....	141
GRÁFICO 07.04: Número médio de localidades estimadas em 25 simulações com erro relativo percentual em valor absoluto menor ou igual a 5. Variável: NC_RES.....	142
GRÁFICO 07.05: Erro Padrão do número de localidades estimadas com erro relativo percentual em valor absoluto menor ou igual a 5, para as 75 simulações. Variável: NC_RES.....	142
GRÁFICO 07.06: Amplitude de variação do número de localidades estimadas em 75 simulações com erro relativo percentual menor ou igual a 5. Variável: NC_RES.....	142
GRÁFICO 07.07: Cobertura do valor simulado das estimativas populacionais, em cada uma das 25 simulações. Variável: NC_RES.....	143
GRÁFICO 07.08: Erro Padrão do número médio de vezes que o intervalo cobre os valores simulados das estimativas populacionais, em 75 simulações. Variável: NC_RES.....	143
GRÁFICO 07.09: Amplitude de variação do número de vezes que o intervalo cobre o valor simulado, das estimativas populacionais, em 75 simulações. Variável: NC_RES.....	143
GRÁFICO 08.01: Média dos erros relativos percentuais em valores absolutos das 25 simulações. Variável: NC_RUR.....	144
GRÁFICO 08.02: Erro Padrão dos erros relativos percentuais em valores absolutos das estimativas populacionais das 75 simulações. Variável: NC_RUR.....	144

GRÁFICO 08.03: Amplitude de variação dos erros relativos percentuais em valores absolutos das estimativas populacionais das 75 simulações. Variável: NC_RUR.	144
GRÁFICO 08.04: Número médio de localidades estimadas em 75 simulações com erro relativo percentual em valor absoluto menor ou igual a 5. Variável: NC_RUR.	145
GRÁFICO 08.05: Erro Padrão do número de localidades estimadas em 75 simulações com erro relativos percentual em valor absoluto menor ou igual a 5. Variável: NC_RUR.	145
GRÁFICO 08.06: Amplitude de variação do número de localidades estimadas com erro relativo percentual menor ou igual a 5. Variável: NC_RUR.	145
GRÁFICO 08.07: Cobertura do valor simulado das estimativas populacionais, em cada uma das 25 simulações. Variável: NC_RUR.	146
GRÁFICO 08.08: Erro Padrão do número médio de vezes que o intervalo cobre os valores simulados das estimativas populacionais, em 75 simulações. Variável: NC_RUR.	146
GRÁFICO 08.09: Amplitude de variação do número de vezes que o intervalo cobre os valores simulados das estimativas populacionais, para as 75 simulações. Variável: NC_RUR.	146
GRÁFICO 09.01: Média dos erros relativos percentuais das estimativas populacionais de 25 simulações. Variável : PR.	147
GRÁFICO 09.02: Erro Padrão dos erros relativos percentuais em valores absolutos das estimativas populacionais das 75 simulações. Variável: PR.	147
GRÁFICO 09.03: Amplitude de variação dos erros relativos percentuais em valores absolutos das estimativas populacionais de 75 simulações. Variável: PR.	147
GRÁFICO 09.04: Número médio de localidades estimadas em 25 simulações com erro relativo percentual menor ou igual a 5. Variável: PR.	148
GRÁFICO 09.05: Erro Padrão do número médio de localidades estimadas com erro relativo percentual em valor absoluto menor ou igual a 5. Variável: PR.	148
GRÁFICO 09.06: Amplitude de variação do número de localidades estimadas com erro relativo percentual menor ou igual a 5, em 75 simulações. Variável: PR.	148
GRÁFICO 09.07: Cobertura dos valores simulados das estimativas populacionais, em cada grupo de simulação: 75_76%, 75_80% e 75_90%, respectivamente, com 25 simulações cada. Variável: PR.	149

GRÁFICO 09.08: Erro Padrão do número médio de vezes que o intervalo cobre os valores simulados das estimativas populacionais, em 75 simulações. Variável: PR.....	149
GRÁFICO 09.09: Amplitude de variação do número de vezes que o intervalo cobre os valores simulados das estimativas populacionais, em 75 simulações. Variável: PR.....	149
GRÁFICO 10.01: Média dos erros relativos percentuais em valores absolutos das estimativas populacionais de 25 simulações. Variável: PU.....	150
GRÁFICO 10.02: Erro Padrão dos erros relativos percentuais em valores absolutos das estimativas populacionais de 75 simulações. Variável: PR.....	150
GRÁFICO 10.03: Amplitude de variação dos erros relativos percentuais em valores absolutos das estimativas populacionais de 25 simulações. Variável: PU.....	150
GRÁFICO 10.04: Número médio de localidades estimadas em 75 simulações com erro relativo percentual em valor absoluto menor ou igual a 5. Variável: PU.....	151
GRÁFICO 10.05: Erro Padrão do número médio de localidades estimadas com erro relativo percentual em valor absoluto menor ou igual a 5, para as 75 simulações. Variável: PU.....	151
GRÁFICO 10.06: Amplitude de variação do número de localidades estimadas com erro relativo em valor absoluto menor ou igual a 5. Variável: PU.	151
GRÁFICO 10.07: Cobertura do valor simulado das estimativas populacionais, em cada uma das 25 simulações. Variável: PR.....	152
GRÁFICO 10.08: Erro Padrão do número médio de vezes que o intervalo cobre os valores simulados das estimativas populacionais, em 75 simulações. Variável: PU.....	152
GRÁFICO 10.09: Amplitude de variação do número de vezes que intervalo cobre os valores simulados das estimativas populacionais, para as 75 simulações. Variável: PU.....	152

INTRODUÇÃO

1.1. O Problema

O mundo atual, ancorado na ideologia neoliberal e na globalização, tem passado por várias modificações. Fusões de grandes empresas privadas acontecem rotineiramente com o objetivo de unificar departamentos e, conseqüentemente, reduzir custos. Mas o contrário também pode ocorrer e empresas gigantes são divididas, como tem acontecido, por exemplo, com grandes conglomerados nas áreas de telecomunicações, de produção e distribuição de energia, etc. Nesse caso, o motivo para esse fato é a privatização de grandes empresas estatais. Em tal contexto, pode ocorrer que algum equipamento urbano, ou não, que era utilizado em comodato por várias empresas públicas ao mesmo tempo, tenha o seu uso definido e cobrado pela empresa implantadora do equipamento. Como um exemplo de equipamentos de uso compartilhado, tem-se os postes das companhias de eletrificação, que em geral a elas pertencem, e que são utilizadas por outras empresas, como as empresas de telefonia, de TV a cabo, etc. Então, como se faz para estimar a quantidade de equipamentos que, durante muitos anos, foram construídos e utilizados por duas ou mais empresas? É interessante observar que a empresa implantadora do equipamento tem ciência da quantidade existente, porém não sabe exatamente quantos equipamentos estão em uso por outras empresas. Esse é um problema atual e de solução procurada por várias empresas nacionais.

1.2. Objetivos

O objetivo principal deste trabalho é desenvolver uma metodologia para estimação do número de equipamentos de uso compartilhado, em uma população de equipamentos da mesma espécie com tamanho N conhecido. De forma

secundária, tem-se também como objetivos a execução de uma simulação para avaliação da eficiência da metodologia desenvolvida e a construção de um programa executivo, tanto para a metodologia quanto para a simulação.

1.3. Justificativa

O presente estudo justifica-se pela demanda atual de metodologias para a quantificação de equipamentos de uso compartilhado tendo em vista a privatização de muitas empresas estatais, as quais, muitas vezes, usavam equipamentos públicos em comum. Mas, em função da situação de novas administrações, há necessidade de se identificar os equipamentos de uso compartilhado com o objetivo de estipular a conseqüente remuneração. Assim, a criação de metodologias com a finalidade de estimar a quantidade destes equipamentos é muito importante.

1.4. Estrutura do Trabalho

Desenvolve-se o assunto da seguinte maneira: no Capítulo 2, é feita uma revisão de literatura abordando livros e artigos sobre os seguintes assuntos: a Teoria de Agrupamentos, a Teoria da Amostragem, o Estimador da Razão, a Expansão Simples e a Regressão Linear, e duas técnicas estatísticas computacionalmente intensivas: o “Jackknife” e o “Bootstrap”. No Capítulo 3, descreve-se o material e o método, ou seja, aborda-se a metodologia de avaliação construída, a simulação e o programa. Finalmente, no Capítulo 4, faz-se a discussão dos resultados e no Capítulo 5, a conclusão.

2. REVISÃO BIBLIOGRÁFICA

2.1. A Teoria de Agrupamentos

2.1.1. Introdução

A resposta para a pergunta, “o que é um agrupamento?”, não tem uma resposta fácil. Existem muitas idéias intuitivas, freqüentemente conflitantes, sobre o que se constitui um “agrupamento”, mas poucas definições formais (CORMACK, 1971, p.329). Assim, o termo, permanece como um conceito indefinido. Segundo CORMACK, que fez uma revisão crítica sobre o assunto, duas idéias básicas estão envolvidas na conceituação de agrupamento, quais sejam: coesão interna entre os elementos do grupo (condição de minimalidade) e isolamento externo (condição de maximalidade) (CORMACK, 1971, p. 329).

Para ilustrar a dificuldade de se definir agrupamento JOHNSON (1998, 726 - 728) ilustra a situação com as possíveis maneiras de se agrupar as 16 cartas distintas de um baralho e mostra que existe uma única maneira para se agrupar as 16 cartas em um único grupo, 32.767 de particioná-las em dois grupos de tamanhos variados e 7.141.686 maneiras de agrupa-las em três grupos, também de tamanhos variáveis. E, de modo geral, o número de maneiras de alocar n objetos em k grupos distintos, é um número de Stirling, de segunda ordem dado por: $(1/k!) \sum_{j=0}^k (-1)^{k-j} \binom{k}{j} j^n$. De sorte que, adicionando esses números para $k=1, 2, \dots, n$ grupos, pode-se obter o número total de possíveis maneiras de alocar os n objetos dentro de grupos distintos (JOHNSON, 1998, p. 727). Algumas tentativas para definir agrupamento, por que, quando e como usá-lo, têm sido feitas, mas sem manifestações por parte dos usuários dos métodos de agrupamentos (CORMACK, 1971, p. 330).

A maioria dos esforços no sentido de produzir uma estrutura simples de agrupamento requer a noção de medida de “proximidade” ou de “similaridade”, e há muita subjetividade envolvida na escolha de uma medida de “similaridade”.

Considerações importantes incluem a natureza das variáveis (discreta e contínua) as escalas de medição (nominal, ordinal, intervalo, razão), além de conhecimento do assunto em questão por parte do investigador (JOHNSON, 1998, p. 728). Assim, quando itens (unidades ou casos) são agrupados, a medida de “proximidade” é usualmente indicada por algum tipo de distância. Por outro lado, variáveis são usualmente agrupadas com base nos coeficientes de correlação ou com medidas de associação (JOHNSON, 1998, p.728). O coeficiente de correlação é um bom indicador do grau de proximidade entre duas variáveis aleatórias ou vetores.

2.1.2. Agrupamentos Hierárquicos

As técnicas de agrupamento hierárquico trabalham com uma série de sucessivas fusões ou com uma série de sucessivas divisões. A primeira dessas técnicas é chamada de Método Hierárquico Aglomerativo, que considera inicialmente tantos grupos quanto o número de itens existentes na população e, através de uma seqüência de fusões, nas quais os itens mais “semelhantes” são agrupados primeiro, prossegue-se até que o último item seja incluído.

A segunda, chamada de Método Hierárquico Divisivo, trabalha no sentido inverso, ou seja, considera todos os indivíduos pertencentes a um único grupo e inicia separando este em dois subgrupos, de forma que os itens de cada um dos subgrupos esteja “longe” dos outros itens do outro subgrupo. Esses por sua vez, são divididos até que haja tantos grupos quantos itens iniciais, formando, assim, cada item um grupo. Os resultados dos métodos aglomerativo ou divisivo podem ser representados em um diagrama bidimensional, conhecido como dendograma, e algumas técnicas de agrupamento hierárquico podem ser usadas para agrupar tanto itens quanto variáveis (JOHNSON, 1998, p. 738-739).

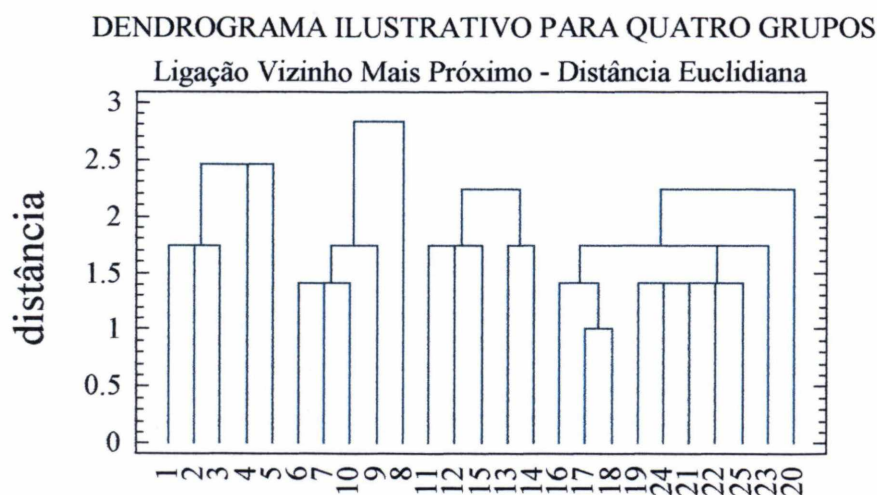


Figura 01: Dendrograma

Na figura 01 o eixo horizontal contém os itens agrupados segundo a distância representada no eixo vertical. Como pode ser observado no dendrograma, os itens 17 e 18 estão agrupados juntos a uma distância de uma unidade a partir da origem. Outros itens estão agrupados a uma distância maior. A distância entre um grupo e outro é a medida de dissimilaridade entre esses grupos. Maior distância, significa maior dissimilaridade ou isolamento entre os itens.

2.1.3. Agrupamentos Não Hierárquicos

As técnicas de Agrupamento Não Hierárquico são mais indicadas para agrupar itens (unidades ou casos) do que variáveis, em uma coleção de k grupos. Quando itens são agrupados, a proximidade é indicada por algum tipo de distância e, por outro lado, variáveis são geralmente agrupadas com base nos coeficientes de correlação ou por medidas semelhantes de associação (JOHNSON, 1998, p.724 - 728).

Neste trabalho, o interesse é agrupar itens (unidades ou casos). Para esse fim, utiliza-se o método mais popular de agrupamento Não Hierárquico, chamado k -médias, cujo algoritmo é descrito a seguir, e como medida de similaridade ou distância, usa-se a distância Euclidiana.

2.1.3.1. O Método das k-médias

O agrupamento de N elementos (itens ou casos) de uma população pelo método das k médias pode ser descrito pela sequência de passos que constitui o algoritmo de agrupamento.

Passo 1 - Reparte-se os N elementos (itens ou casos) da população em k grupos arbitrários e calculam-se os k centróides desses k grupos.

Passo 2 - Para cada um dos k grupos criados arbitrariamente, calcula-se a distância Euclidiana entre cada um dos N elementos da população e o centróide do grupo. Aloca-se o elemento ao grupo do qual esteja mais próximo, podendo esse elemento, pertencer ou não ao grupo inicial.

Passo 3 - Recalcula-se os novos centróides dos grupos formados tanto para aquele que recebeu um novo elemento, como para aquele que perdeu o referido elemento.

Repetem-se os Passos 2 e 3, até que todos os k grupos e todos os N elementos da população tenham sido verificados e que estejam satisfeitas as condições gerais de minimalidade entre os itens internos a cada grupo e as de maximalidade entre os itens de grupos distintos. Ilustra-se o algoritmo (HARTIGAN; WONG, 1973, p. 100-108) do método com o Fluxograma abaixo, cujos símbolos significam:

k o número de grupos fixado;

$Z_r(m)$ o centro do grupo r , na iteração m ;

$x_r, r = 1, 2, \dots, k$ o item inicial atribuído ao grupo r ;

$S_t(m)$ o grupo t , na iteração m ;

$N_t, t = 1, 2, \dots, m$ o número de itens do grupo $S_t(m)$;

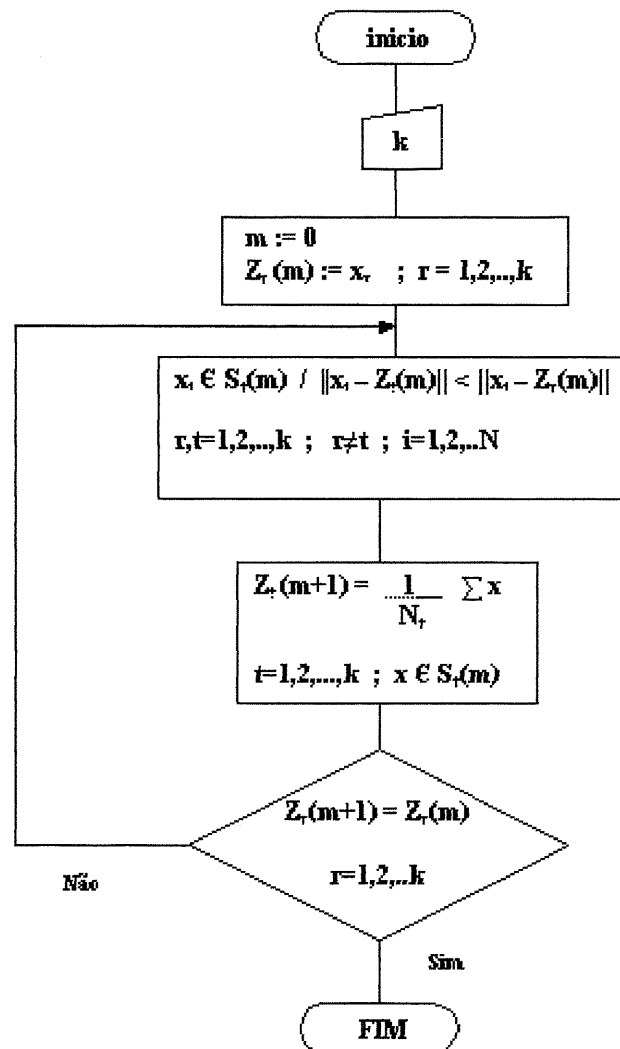


Figura 02: FLUXOGRAMA – Método das k Médias

Exemplo - Agrupamento pelo método das k-Médias (JOHNSON, 1998, p. 754 – 757, Exemplo, 12.2)

Suponha-se haver quatro itens a serem agrupados, A, B, C e D, e que cada item seja constituído de duas variáveis, X_1 e X_2 , cujas medidas para os respectivos itens estão dadas na tabela abaixo.

Item	Observações	
	x_1	x_2
A	5	3
B	-1	1
C	1	-2
D	-3	-2

O objetivo é dividir esses itens em $k = 2$ grupos tais que os itens dentro de um grupo estejam mais próximos um do outro do que estejam dos itens que pertencem ao outro grupo.

Seguindo os passos acima descritos para o procedimento das k-médias, tem-se :

Passo 1 – Escolhendo arbitrariamente (AB) e (CD), como grupos iniciais, determinam-se os seus respectivos centróides:

GRUPO	Coordenadas do Centróide	
	\bar{x}_1	\bar{x}_2
(AB)	$\frac{5+(-1)}{2} = 2$	$\frac{3+1}{2} = 2$
(CD)	$\frac{1+(-3)}{2} = -1$	$\frac{-2+(-2)}{2} = -2$

Passo 2 – Cálculo das distâncias Euclidianas, dos itens A, B, C e D, aos centróides iniciais, definidos pelos grupos (AB) e (CD);

$$d^2 (A, (AB)) = (5-2)^2 + (3-2)^2 = 10$$

$$d^2 (A, (CD)) = (5-(-1))^2 + (3-(-2))^2 = 61$$

$$d^2 (B, (AB)) = (-1-2)^2 + (1-2)^2 = 10$$

$$d^2 (B, (CD)) = (-1-(-1))^2 + (1-(-2))^2 = 9$$

Sendo a menor distância Euclidiana a distância do item B ao grupo inicial (CD), o item B deve ser alocado a este grupo, ficando o novo grupo constituído pelos itens B, C e D, formando o grupo (BCD), e o grupo (AB) perde o item B, ficando o grupo constituído, apenas, pelo item A, grupo (A).

Passo 3 – Recalculando os novos centróides;

GRUPOS	Coordenadas dos Centróides	
	\bar{x}_1	\bar{x}_2
(A)	5	3
(BCD)	$\frac{-1+1+(-3)}{3} = -1$	$\frac{1+(-2)+(-2)}{3} = -1$

Calculando as distâncias Euclidianas, para os itens A, B, C e D, para esses novos centróides, dos novos grupos recentemente formados, tem-se;

$$\begin{aligned}
d^2(A, (A)) &= (5-5)^2 + (3-3)^2 = 0 \\
d^2(B, (A)) &= (-1+5)^2 + (1+3)^2 = 32 \\
d^2(C, (A)) &= (1-5)^2 + (-2-3)^2 = 41 \\
d^2(D, (A)) &= (-3-5)^2 + (-2-3)^2 = 29 \\
d^2(A, (BCD)) &= (5-(-1))^2 + (3-(-1))^2 = 52 \\
d^2(B, (BCD)) &= (-1-(-1))^2 + (1-(-1))^2 = 4 \\
d^2(C, (BCD)) &= (1-(-1))^2 + (-2-(-1))^2 = 5 \\
d^2(D, (BCD)) &= (-3-(-1))^2 + (-2-(-1))^2 = 5 \\
d^2(B, C) &= (-1-1)^2 + (1-(-2))^2 = 13 \\
d^2(B, D) &= (-1-(-3))^2 + (1-(-2))^2 = 13 \\
d^2(D, D) &= (-3-(-3))^2 + (-2-(-2))^2 = 0
\end{aligned}$$

Verifica-se, dos resultados dos cálculos de distâncias acima, que o processo das k-médias, para esse exemplo, se encerra, pois a condição de coesão interna (ou condição de minimalidade) entre os elementos de cada grupo e a condição de isolamento externo (ou condição de maximalidade) entre itens de grupos distintos estão satisfeitas, com um grupo formado apenas pelo item A, grupo (A), e o outro grupo sendo formado pelos itens B, C e D , grupo (BCD).

2.2. A Teoria da Amostragem

2.2.1. Introdução

“Nossos conhecimentos, nossas atitudes e nossas ações baseiam-se, em grande parte, em exemplos. Isso é igualmente verdadeiro, quer na vida cotidiana, quer na pesquisa científica”.

Nos negócios e na ciência humanos, faltam-nos recursos para estudar mais que uma fração dos fenômenos que podem melhorar nossos conhecimentos.” (COCHRAN, 1965, p.13).

A teoria da amostragem trata dos métodos de seleção e observação de uma parte (uma amostra) da população inteira, para com ela fazer inferências sobre toda a população, em lugar de um senso completo. Algumas vantagens de uma inferência estatística baseada em uma amostra sobre um senso de toda a população são, entre outras, as seguintes:

- 1 – Economia. Uma vez que os dados são obtidos de uma pequena fração da população em lugar de obtê-los de toda a população, as despesas são menores do que a de um senso integral.
- 2 – Rapidez. Pelo mesmo motivo do item anterior, os dados podem ser coletados, sistematizados e sintetizados mais rapidamente, do que os de um senso completo. O que é de suma importância quando é urgente a necessidade das informações.
- 3 – Exeqüibilidade. Especialmente quando a observação é de caráter destrutivo, para a qual um senso completo seria impraticável. Ou mesmo no caso de uso de equipamentos caros e sofisticados e de pessoal altamente especializados.
- 4 – Qualidade e Exatidão. Em situações onde sejam necessários pessoal altamente treinado e equipamentos sofisticados, o que seria impraticável em um senso completo, pode ser perfeitamente exeqüível em uma pequena amostra da população inteira. Daí ser possível contar com as condições necessárias, tendo como resultado estimativas de maior qualidade e precisão do que um senso completo poderia fornecer.

Por outro lado, um censo completo pode ter vantagens especiais sobre uma amostra pequena, em outras situações, quais sejam:

- 1 – Dados para unidades pequenas podem ser obtidos com pouca dificuldade;
- 2 – Um censo completo é de maior aceitação por parte do público;
- 3 – A resposta e a compilação de dados se tornam mais seguras;
- 4 – Vícios de convergência podem ser mais facilmente verificados e corrigidos;
- 5 – Não se requer pessoal especializado na realização da tarefa de levantamento de dados (KISH, 1965, p. 18) e (COCHRAN, 1965, p. 13 - 15).

2.2.2. Tipos de Amostragens

A teoria da amostragem aborda diversas maneiras de retirar as amostras da população. Estas podem ser retiradas aleatoriamente, da população inteira como um todo, ou dividindo-se a população em estratos ou em grupos, e as amostras então serem retiradas de cada um desses estratos ou grupos, também de forma aleatória. Outra maneira de retirar as amostras, é fazê-lo de forma sistemática, que consiste em dividir o número total de unidades da população pelo número de unidades a ser amostrado, n ; em seguida, sorteia-se a primeira unidade que será retirada dentre a primeira unidade da população e o número obtido pela divisão do número total de unidades da população pelo número de unidades a ser amostrado, as demais unidades a serem amostradas serão, então, aquelas cujos números são o número da primeira unidade sorteada acrescido de 1, 2, 3, ..., $n-1$. Além disso, a amostragem sistemática pode também ser realizada dentro de estratos ou de grupos.

A forma de retirar as amostras aleatórias, tanto da população inteira quanto de estratos ou de grupos, também pode ser feita de duas formas diferentes: sem reposição da unidade sorteada ou com reposição da mesma.

Diversas formas de se dividir a amostra também são consideradas na teoria da amostragem como, por exemplo, amostra proporcional ao tamanho do estrato ou grupo a ser amostrado, etc. Neste trabalho, consideram-se, exclusivamente, amostras aleatórias simples, e os resultados se referem a esse tipo de amostragem.

2.2.2.1. Amostras Aleatórias Simples

Dá-se o nome de Amostra Aleatória Simples ao conjunto de observações de tamanho n , cujos elementos foram selecionados da população com probabilidades iguais de ocorrência. Os elementos dessas amostras são obtidos de todas as combinações amostrais possíveis, das N unidades populacionais tomadas n a n .

2.2.2.2. Algumas propriedades das amostras aleatórias simples:

Considera-se, no que segue, algumas propriedades das amostras aleatórias simples em uma população com N unidades populacionais. Para esse contexto tem-se a seguinte notação:

N	número total de unidades populacionais;
n	número total de unidades a ser amostrado na população;
x_i	quantidade da variável cujo valor é conhecido, presente na i -ésima unidade populacional;
y_i	quantidade da variável cujo valor total populacional se deseja estimar, presente na i -ésima unidade populacional;
$x = \sum_{i=1}^n x_i$	quantidade total amostral da variável cujo valor total populacional é conhecido;
$y = \sum_{i=1}^n y_i$	quantidade total amostral da variável cujo valor total populacional se deseja estimar;
$X = \sum_{i=1}^N x_i$	quantidade total populacional da variável cujo valor total populacional é conhecido;
$Y = \sum_{i=1}^N y_i$	quantidade total populacional da variável cujo valor total populacional se deseja estimar;
$f = n/N$	índice amostral, ou fração amostral;

$\bar{x} = (1/n) \sum_{i=1}^n x_i$	valor médio amostral da variável cujo valor total populacional é conhecido;
$\bar{y} = (1/n) \sum_{i=1}^n y_i$	valor médio amostral da variável cujo valor total populacional se deseja estimar;
$\bar{X} = (1/N) \sum_{i=1}^N x_i$	valor médio populacional da variável cujo valor total populacional é conhecido;
$\bar{Y} = (1/N) \sum_{i=1}^N y_i$	valor médio populacional da variável cujo valor total populacional se deseja estimar;
$\hat{X} = N \bar{x}$	estimador do valor total populacional da variável cujo valor total populacional é conhecido;
$\hat{Y} = \bar{y}$	estimador do valor médio populacional da variável cujo valor total populacional se deseja estimar;

2.2.2.2.1- Esperanças e Variâncias de amostras aleatórias simples

RESULTADO 2.2.2.2.1.1

A Esperança do valor médio amostral \bar{y} é um estimador não viciado do valor médio populacional \bar{Y} .

PROVA :

Se $E \bar{y}$ é não viciado, então $E \bar{y} = \frac{\sum \bar{y}}{C_n^N} = \frac{n!(N-n)!}{N!} \sum \left(\frac{y_1 + y_2 + \dots + y_n}{n} \right)$, sendo o somatório estendido a todas as combinações amostrais possíveis de tamanho n , retiradas das N unidades populacionais C_n^N . Mas, calcular o somatório $\sum (y_1 + y_2 + \dots + y_n)$ significa calcular em quantas amostras de tamanho n , obtidas das

N unidades populacionais, aparecerá um dado y_i . Fixado um y_i , restam (N-1) unidades populacionais de onde os demais (n-1) y_i serão retirados. Dessa forma, o número de amostras contendo um y_i fixo será dado por uma combinação de (N-1) y_i unidades amostrais tomadas (n-1) a (n-1), ou C_{n-1}^{N-1} . Assim, $\sum (y_1 + y_2 + \dots + y_n)$ calculado sobre todas as C_n^N possíveis combinações amostrais, dará

$$\sum (y_1 + y_2 + \dots + y_n) = \frac{(N-1)!}{(n-1)!(N-n)!} (y_1 + y_2 + \dots + y_N) \text{ que, substituído na última}$$

expressão para $E \bar{y}$, e fazendo-se as devidas simplificações, obtém-se:

$$E \bar{y} = (y_1 + y_2 + \dots + y_N) / N = \bar{Y} \Rightarrow E \bar{y} = \bar{Y}.$$

RESULTADO 2.2.2.2.1.1.1

$\hat{Y} = N \bar{y}$ é um estimador não viciado do valor total populacional, Y.

PROVA :

$$E \hat{Y} = N \frac{n!(N-n)! \cdot \sum \bar{y}}{N!} = \frac{n!(N-n)!}{(N-1)!} \frac{1}{n} \sum (y_1 + y_2 + \dots + y_n), \quad \text{com o somatório}$$

estendido à todas as C_n^N amostras. Mas, da demonstração do Resultado 2.2.2.2.1, $(y_1 + y_2 + \dots + y_N)$ que substituída na expressão acima, e fazendo as simplificações devidas, obtém-se: $E \hat{Y} = (y_1 + y_2 + \dots + y_N) = Y$.

Logo, $E \hat{Y} = Y \Rightarrow E \hat{Y}$ é um estimador não viciado do valor total populacional X.

RESULTADO 2.2.2.2.1.2

A variância do valor médio amostral \bar{x} de uma amostra aleatória simples é dada por:

$$V_{\bar{y}} = (1-f)(S^2 / n). \text{ Onde: } S^2 = \frac{1}{(N-1)} \sum_{i=1}^N (y_i - \bar{Y})^2.$$

PROVA:

$$[n(\bar{y} - \bar{Y})]^2 = n^2 (\bar{y} - \bar{Y})^2 = \left[\sum_{i=1}^n (y_i - \bar{Y}) \right]^2 \Rightarrow E[n^2 (\bar{y} - \bar{Y})^2] = E \left[\sum_{i=1}^n (y_i - \bar{Y}) \right]^2 \Rightarrow$$

$$n^2 E(\bar{y} - \bar{Y})^2 = E \left[\sum_{i=1}^n (y_i - \bar{Y}) \right]^2 = E \left[\sum_{i=1}^n (y_i - \bar{Y})^2 + 2 \sum_{i=1}^{n-1} (y_i - \bar{Y})(y_{i+1} - \bar{Y}) \right] \Rightarrow$$

$$\Rightarrow n^2 E(\bar{y} - \bar{Y})^2 = E \left[\sum_{i=1}^n (y_i - \bar{Y})^2 \right] + 2 E \left[\sum_{i=1}^{n-1} (y_i - \bar{Y})(y_{i+1} - \bar{Y}) \right]$$

Aplicando-se aos dois termos do lado direito da última equação o mesmo raciocínio aplicado na demonstração do Resultado 2.2.2.2.1, tem-se

$$E \left[\sum_{i=1}^n (y_i - \bar{Y})^2 \right] = \sum_{i=1}^n E(y_i - \bar{Y})^2 = \frac{\sum (y_i - \bar{Y})^2}{C_n^N} = \frac{n}{N} \sum_{i=1}^N (y_i - \bar{Y})^2 \quad (\text{a}), \text{ onde os limites}$$

não indicados para o somatório são todas as possíveis combinações, C_n^N .

$$E \left[\sum_{i=1}^{n-1} (y_i - \bar{Y})(y_{i+1} - \bar{Y}) \right] = \sum_{i=1}^{n-1} E(y_i - \bar{Y})(y_{i+1} - \bar{Y}) = \frac{\sum (y_i - \bar{Y})(y_{i+1} - \bar{Y})}{C_{n-1}^{N-1}}$$

onde os limites não especificados dos somatórios do lado direito se estendem a todas as combinações amostrais \mathbf{C}_{n-1}^{N-1} . Desenvolvendo os cálculos e simplificando, tem-se

$$E \left[\sum_{i=1}^{n-1} (y_i - \bar{Y})(y_{i+1} - \bar{Y}) \right] = \frac{n}{N} \frac{(n-1)}{(N-1)} \sum_{i=1}^{N-1} (y_i - \bar{Y})(y_{i+1} - \bar{Y}) \quad (b)$$

$$n^2 E \left[\sum_{i=1}^n (y_i - \bar{Y}) \right]^2 = \frac{n}{N} \left\{ \sum_{i=1}^N (y_i - \bar{Y})^2 + 2 \frac{(n-1)}{(N-1)} \sum_{i=1}^{N-1} (y_i - \bar{Y})(y_{i+1} - \bar{Y}) \right\}$$

Substituindo (a) e (b) na expressão de onde elas foram retiradas, e fazendo as simplificações, tem-se:

$$n^2 E \left[\sum_{i=1}^n (y_i - \bar{Y}) \right]^2 = \frac{n}{N} \left\{ \sum_{i=1}^N (y_i - \bar{Y})^2 + 2 \frac{(n-1)}{(N-1)} \sum_{i=1}^{N-1} (y_i - \bar{Y})(y_{i+1} - \bar{Y}) \right\}$$

completando o quadrado no último somatório, a última expressão pode ser escrita

$$\text{na forma: } n^2 E \left[\sum_{i=1}^n (y_i - \bar{Y}) \right]^2 = \frac{n}{N} \left\{ \left(1 - \frac{(n-1)}{(N-1)} \right) \sum_{i=1}^N (y_i - \bar{Y})^2 + \frac{(n-1)}{(N-1)} \left[\sum_{i=1}^N (y_i - \bar{Y}) \right]^2 \right\},$$

sendo que o último somatório dentro da chave é nulo, tem-se então que:

$$E \left[\sum_{i=1}^N (y_i - \bar{Y}) \right]^2 = \frac{(N-n)}{nN(N-1)} \sum_{i=1}^N (y_i - \bar{Y})^2 \Rightarrow E \left[\sum_{i=1}^N (y_i - \bar{Y}) \right]^2 = (1-f) \left(\frac{S^2}{n} \right) \Rightarrow$$

$$V_{\bar{y}} = (1-f) \left(\frac{S^2}{n} \right).$$

RESULTADO 2.2.2.2.1.3

Se x_i e y_i são duas quantidades presentes em todas as unidades da população, e \bar{x} e \bar{y} são os seus respectivos valores médios, obtidos por meio de uma amostragem aleatória simples, então a covariância de x e y será dada por:

$$E(\bar{y} - \bar{Y})(\bar{x} - \bar{X}) = \frac{N-n}{nN} \frac{1}{N-1} \sum_{i=1}^n (y_i - \bar{Y})(x_i - \bar{X})$$

PROVA:

Seja $z_i = x_i + y_i$ a quantidade presente em todas as unidades da população. Aplicando-se o Resultado 2.2.2.2.1.2 ao valor médio populacional dessa quantidade,

$$\text{Ou, seja: } \bar{Z} = \bar{X} + \bar{Y}, E(\bar{z} - \bar{Z})^2 = \frac{N-n}{nN} \frac{1}{N-1} \sum_{i=1}^N (z_i - \bar{Z})^2$$

$$E[(\bar{y} - \bar{Y}) + (\bar{x} - \bar{X})]^2 = \frac{N-n}{nN} \frac{1}{N-1} \sum_{i=1}^N [(y_i - \bar{Y}) + (x_i - \bar{X})]^2 \quad a)$$

$$\text{Mas, do Resultado 2.2.2.2.1.2: } E(\bar{y} - \bar{Y})^2 = \frac{N-n}{nN} \frac{1}{N-1} \sum_{i=1}^N (y_i - \bar{Y})^2$$

Essa expressão para a variância de \bar{y} , juntamente com a similar para \bar{x} ,

substituídas na expressão a) acima, após as simplificações, resulta em:

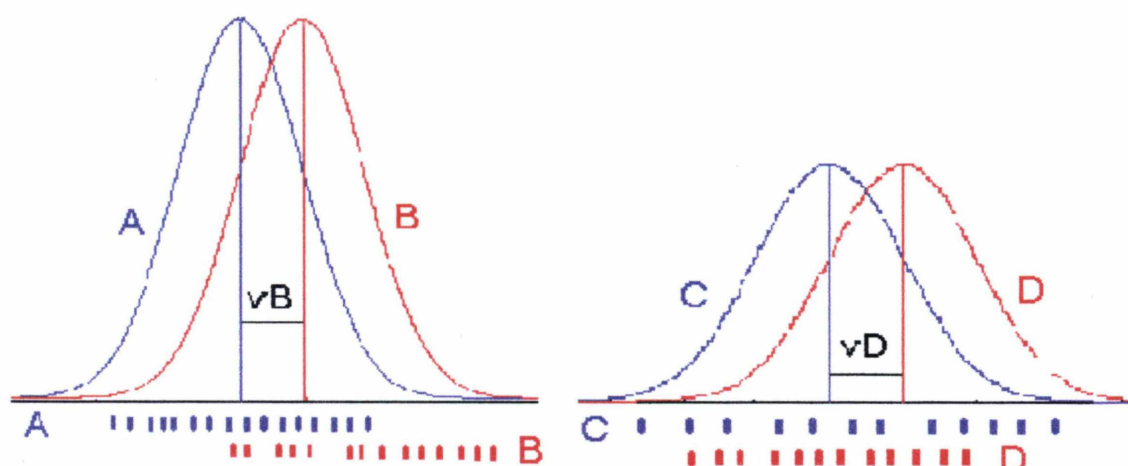
$$E(\bar{y} - \bar{Y})(\bar{x} - \bar{X}) = \frac{N-n}{nN} \frac{1}{N-1} \sum_{i=1}^N (y_i - \bar{Y})(x_i - \bar{X})$$

Observa-se que os Resultados acima, para as variâncias, são Resultados aplicáveis às populações infinitas. Para as populações finitas os mesmos devem ser acrescidos do fator de correção para populações finitas (cpf). Para a variância, em uma população finita, o cpf, é $(N-n)/N$. As vezes o c.p.f. é escrito na forma $(N-n)/N-1$, para simplificações, por alguns autores.

2.2.2.3. Considerações sobre o dimensionamento da amostra

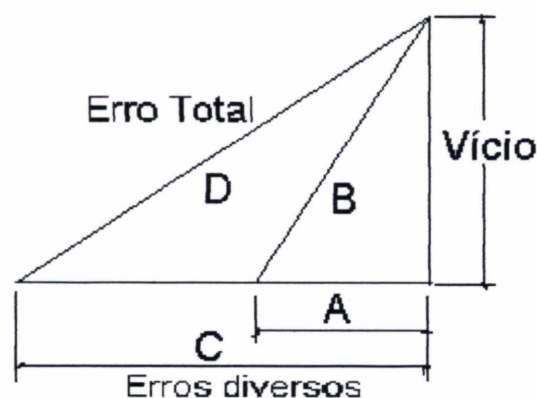
Para o dimensionamento da amostra a ser retirada da população, diversas considerações devem ser levadas em conta por parte do pesquisador. Entre essas considerações estão o nível de confiança da solução desejada, o erro relativo, a precisão, o vício e a exatidão, entre outras. Os termos exatidão e precisão são largamente usados para separar os efeitos dos vícios. Precisão, em geral, se refere a pequenos erros variáveis; as vezes o termo denota unicamente o inverso da variância amostral, mas em qualquer caso ele exclui o efeito do vício. A exatidão é uma referência à totalidade dos pequenos erros e inclui o efeito dos vícios. A exatidão é definida como sendo o inverso do erro total, incluindo os vícios e os demais erros pequenos. Assim, um planejamento amostral será exato se for preciso e tiver vícios nulos ou muito pequenos. Um planejamento amostral com um vício grande é ainda preciso se suas variáveis erro forem pequenas, mas não é exato. A seguir, lista-se algumas definições como dadas em KISH (1965, p. 24-25, 510):

Vício = $\mu - \bar{y}$	o vício diz respeito ao afastamento do valor médio amostral, \bar{y} , do valor médio populacional, μ ;
Erro relativo $\frac{Y_V - Y_E}{Y_V}$	o erro relativo é a razão entre a diferença do valor verdadeiro e o estimado, e o valor verdadeiro;
Precisão = $\frac{1}{\sigma_{\bar{y}}^2}$	a precisão é definida como o inverso da variância amostral;
Exatidão $\frac{ \mu - \bar{y} + \sum \text{erros} }{\sigma_{\bar{y}}^2}$	a exatidão é definida como a razão entre a soma dos vícios e dos pequenos erros de observação, e a variância,



As curvas normais representam as distribuições amostrais de quatro planejamentos amostrais. A altura das curvas mede a probabilidade de variação dos valores das estimativas. Os planejamentos B e D têm vícios grandes, enquanto que A e C parecem não viciados. Os planejamentos A e C são mais precisos, porque eles têm menores erros padrão que os planejamentos C e D. Dos quatro planejamentos, somente o planejamento A é exato, porque ambos, variável erro e vício são pequenos. Ao contrário, o planejamento D apresenta a maior distância média do valor verdadeiro [KISH, 1965, pág. 510]

Figura 03: Curvas de Distribuições Normais para ilustração de Exatidão, Precisão e Vício, amostrais.



A, erro total da distribuição dada pela curva normal, A.
 B, erro total da distribuição dada pela curva normal, B.
 C, erro total da distribuição dada pela curva normal, C.
 D, erro total da distribuição dada pela curva normal, D.

Figura 04: Relações entre erros e vícios das distribuições amostrais A, B, C e D, da Figura 03, acima.

2.2.2.4. O tamanho da amostra

Neste trabalho, os resultados foram obtidos a partir de amostras aleatórias simples e, para um nível de confiança da solução $(1 - \alpha)$, precisão d , erro relativo r , proporção inicial estimada em $\hat{\theta}_0$ e quantidade total de equipamentos na população, N , as amostras foram então determinadas a partir da seqüência de passos:

1 - Cálculo da quantidade de equipamentos a ser amostrada na população, n :

$$P\left(\left|\frac{\bar{Y} - \mu}{\mu}\right| \geq r\right) = \alpha \Rightarrow P(|\bar{Y} - \mu| \geq r\mu) = \alpha \Rightarrow d = r\mu = z\sigma_{\bar{Y}},$$

mas do RESULTADO 2.2.2.2.1.2, $\sigma_{\bar{Y}}^2 = (1-f)(S^2/n)$, ou seja,

$$\hat{\sigma}_{\bar{Y}}^2 = \frac{(N-n)\hat{\theta}_0(1-\hat{\theta}_0)}{(N-1)n}.$$

$$\text{Resulta daí que } n \geq \frac{Nz^2\hat{\theta}_0(1-\hat{\theta}_0)}{d^2(N-1) + z^2\hat{\theta}_0(1-\hat{\theta}_0)}$$

2 – Cálculo do número de localidades a ser amostrado por grupo ou estrato

2.1 - Fez-se uma alocação dos n equipamentos a serem amostrados na população, proporcional ao tamanho (quantidade de equipamentos) de cada grupo, ou estrato,

h , determinado pela expressão: $n_h = \frac{nN_h}{N}$

2.2 – Cálculo da quantidade de equipamentos a ser amostrada em cada grupo ou estrato, h:

$$P\left(\frac{|\bar{y}_h - \mu_h|}{\mu_h} \geq r\right) = \alpha \quad \Rightarrow \quad P(|\bar{Y} - \mu| \geq r\mu) = \alpha \quad \Rightarrow \quad d = r\mu = z\sigma_{\bar{Y}} \quad \Rightarrow$$

$$m_h \geq \frac{\left(\frac{z\sigma_h}{r\mu_h}\right)^2}{1 + \frac{1}{M_h} \left(\frac{z\sigma_h}{r\mu_h}\right)^2}$$

2.3. O Estimador da Razão

2.3.1-Introdução

Em muitas situações, numa população finita de tamanho N , duas quantidades X e Y variam muito de uma unidade populacional para outra; no entanto, a razão entre as duas quantidades, X e Y , $R = Y/X$, pode variar pouco de unidade para unidade. Nesse caso, a razão amostral poderá ser usada para se estimar o valor total populacional da quantidade Y desconhecida, presente em todas as N unidades populacionais, sendo conhecida a outra quantidade X .

O estimador da razão é uma técnica utilizada quando, em uma população, for necessário estimar a relação entre uma quantidade, cujo valor populacional é conhecido, e outra, com valor populacional desconhecido. A população deve ser composta de N unidades populacionais e, tanto a quantidade que se quer estimar quanto a outra, cujo valor populacional é conhecido, devem estar presentes em cada uma dessas N unidades. A técnica do estimador da razão foi utilizada pela primeira vez por Laplace, em 1802, para estimar a população da França (ENCYCLOPEDIA of Statistical Science, 1986 Vol.. 7, p. 639).

2.3.2. Esperança e Variância do Estimador da Razão

Sejam:

$R=Y/X$ a razão entre o valor total populacional Y , da variável que se deseja estimar, e o valor total populacional X , da variável cujo valor total populacional é conhecido;

$f = n/N$ o índice, ou fração amostral;

$\hat{Y} \approx (\bar{y}/\bar{x}) X$ a estimativa do valor total populacional da variável Y cujo valor total populacional se deseja estimar;

$\hat{R} = \bar{y}/\bar{x} = y/x$ o estimador da Razão $R = Y / X$.

Os resultados a seguir enunciados encontram-se provados em COCHRAN (1965).

RESULTADO 2.3.2.1

A variância aproximada do estimador da razão, $\hat{R} = \bar{y}/\bar{x}$, em uma amostra aleatória simples de tamanho n , suficientemente grande, com y_i e x_i medidos em cada unidade amostral, é dada por:

$$V(\hat{R}) \approx \frac{(1-f)}{(n\bar{X}^2)} \frac{\sum_{i=1}^N (y_i - R x_i)^2}{(N-1)}$$

onde R e f já foram definidos anteriormente.

RESULTADO 2.3.2.1.1

A variância aproximada do estimador \hat{Y} do valor total populacional, Y , é dada por:

$$V(\hat{Y}) \approx \left[(n^2(1-f))/n \right] \left[\sum_{i=1}^N (y_i - R x_i)^2 / (N-1) \right].$$

RESULTADO 2.3.2.1.2

A variância aproximada do estimador $\hat{\bar{Y}}$, do valor médio populacional \bar{Y} , é dada por:

$$V(\hat{\bar{Y}}) \approx ((1-f)/n) \left(\sum_{i=1}^N (y_i - R x_i)^2 / (N-1) \right)$$

Para uma população finita, o coeficiente de correlação ρ , entre x_i e y_i , é definido por

$$\rho = \frac{E(y_i - \bar{Y})(x_i - \bar{X})}{\sqrt{E(y_i - \bar{Y})^2} \sqrt{E(x_i - \bar{X})^2}} = \frac{\sum_{i=1}^N (y_i - \bar{Y})(x_i - \bar{X})}{(N-1)S_y S_x} \Rightarrow \rho = \frac{\sum_{i=1}^N (y_i - \bar{Y})(x_i - \bar{X})}{(N-1)S_Y S_X}$$

Usando ρ , S_Y e S_X , os RESULTADOS 2.3.2.1, 2.3.2.1.1 e 2.3.2.1.2 poderão ser escritos em termos de coeficientes de variação (erro padrão, dividido pela quantidade que está sendo estimada):

De $R = \bar{Y} / \bar{X} \Rightarrow \bar{Y} = R\bar{X}$ e podemos escrever $y_i - R x_i$ como:

$$(y_i - \bar{Y}) + (x_i - R\bar{X})$$

Essas considerações substituídas no Resultado 2.3.2.1 fornecem:

$$V(\hat{Y}) = \frac{N^2(1-f)}{n(N-1)} \sum_{i=1}^N [(y_i - \bar{Y}) - R(x_i - \bar{X})]^2 \Rightarrow$$

$$V(\hat{Y}) = \frac{N^2(1-f)}{n(N-1)} \left[\sum_{i=1}^N (y_i - \bar{Y})^2 + R^2 \sum_{i=1}^N (x_i - \bar{X})^2 - 2R \sum_{i=1}^N (y_i - \bar{Y})(x_i - \bar{X}) \right] \quad \text{ou seja,}$$

$$V(\hat{Y}) = \frac{N^2(1-f)}{n} [S_y^2 + R^2 S_x^2 - 2R\rho S_y S_x].$$

Pode-se também substituir R por $R = Y / X$ e, da expressão de ρ , tirar

$$\rho S_y S_x = \frac{1}{(N-1)} \sum_{i=1}^N (y_i - \bar{Y})(x_i - \bar{X}) = S_{yx} \quad \text{e substituir } \rho S_y S_x \text{ por } S_{yx} \text{ tem-se,}$$

$$\text{Assim, } \frac{V(\hat{Y})}{Y^2} = \frac{(1-f)}{n} [C_{yy} + C_{xx} - 2C_{yx}] \quad \text{onde, } C_{yy} = \frac{S_y^2}{Y^2}, C_{xx} = \frac{S_x^2}{X^2}, C_{yx} = \frac{S_{yx}}{YX},$$

são os quadrados dos coeficientes de variação de Y, X e de XY, respectivamente, e S_x , S_y e S_{yx} são as variâncias amostrais de x_i e y_i , respectivamente, e a covariância entre x_i e y_i . Da mesma forma, determinam-se as outras expressões semelhantes que seguem abaixo:

$$\frac{V(\hat{\bar{Y}})}{\bar{Y}^2} = \frac{(1-f)}{n} [C_{yy} + C_{xx} - 2C_{yx}] \quad \text{e} \quad \frac{V(\hat{R})}{R^2} = \frac{(1-f)}{n} [C_{yy} + C_{xx} - 2C_{yx}]$$

2.3.3. Estimador da Razão Separado

Quando se faz uma estratificação ou uma divisão da população em grupos, uma das maneiras de se usar o estimador da razão na estimação do total populacional é determinar-se as estimativas das razões em cada grupo ou estrato para, com elas, calcular as estimativas de cada grupo ou estrato e, daí, a estimativa populacional. Assim, dividindo-se a população em k grupos ou estratos, a estimativa separada de cada estrato ou de cada grupo será dada por :

$$\hat{Y}_{Rh} = \frac{\bar{y}_{ch}}{\bar{x}_{ah}} \quad \text{onde cada uma das quantidades se refere ao estrato ou grupo de ordem h sendo que:}$$

\bar{x}_{ah} é o valor da quantidade conhecida, no grupo ou estrato h;

\bar{y}_{ch} é o valor médio amostral da quantidade sob estimação, Y_h , no estrato ou grupo h;

$\hat{R}_h = \frac{\bar{y}_{ch}}{\bar{x}_{ah}}$ é o estimador da Razão, R_h , do estrato ou grupo h e

$\hat{Y}_{RS} = \sum_{h=1}^k \hat{Y}_{Rh}$ é a estimativa populacional da quantidade desconhecida, Y;

RESULTADO 2.3.3.1

Se as grandezas amostrais, n_h , forem grandes em todos os estratos ou grupos, então a variância da estimativa populacional da quantidade sob estimação, Y , obtida pelo estimador da razão separado, é dada por:

$$V(\hat{Y}_{RS}) = \sum_{h=1}^k \frac{N_h(1-f_h)}{n_h} (S_{yh}^2 + R_h^2 S_{xh}^2 - 2R_h \rho_h S_{yh} S_{xh}), \text{ onde:}$$

$$R_h = \frac{Y_h}{X_h} \quad \text{é a razão no estrato ou grupo } h;$$

ρ_h é o coeficiente de correlação entre as quantidades amostradas, x_i e y_i , do estrato ou grupo h ;

$$f_h = \frac{n_h}{N_h} \quad \text{é o índice, ou fração amostral, do estrato ou grupo } h.$$

$$S_{xh} = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (x_i - \bar{x})^2 \quad \text{é a variância amostral de } x_i, \text{ a quantidade conhecida, amostrada no grupo ou estrato } h;$$

$$S_{yh} = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (y_i - \bar{y})^2 \quad \text{é a variância amostral de } y_i, \text{ a quantidade sob estimação, no grupo ou estrato } h;$$

N_h é o número total de localidades com valores conhecidos, x_i , no grupo ou estrato h ;

n_h é o número total de quantidades com valores conhecidos, x_i , amostradas, no grupo ou estrato h ;

2.3.4. Estimador da Razão Combinado

Chama-se de Razão Combinada, a razão da soma das estimativas de todos os estratos ou grupos entre a quantidade sob estimação e a quantidade conhecida.

Assim,

$\hat{Y}_E = \sum_{h=1}^k N_h \bar{y}_h$ é a estimativa populacional da variável sob estimação obtida por uma expansão das médias amostrais em cada grupo ou estrato, h, a toda a população;

$\hat{X}_E = \sum_{h=1}^k N_h \bar{x}_h$ é a estimativa populacional da quantidade conhecida obtida por expansão das médias amostrais em cada grupo ou estrato, h, a toda a população;

$\hat{R}_C = \frac{\hat{Y}_E}{\hat{X}_E}$ é o estimador da razão combinado;

$\hat{Y}_{RC} = \hat{R}_C X$ é a estimativa populacional da variável sob estimação, obtida com o estimador da razão combinado;

N_h é o número total de unidades da população, presentes no estrato ou grupo h;

$\bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}$ é o valor médio amostral, da variável sob estimação, no estrato ou grupo h;

n_h é o número de unidades amostradas no estrato ou grupo h;

y_{hi} é o valor amostral da variável que sendo estimada na unidade i, do estrato ou grupo h;

\bar{y}_h o valor médio da variável sob estimação, no estrato ou grupo h;

RESULTADO 2.3.4

Quando o tamanho da amostra total n for grande, a variância da estimativa populacional da quantidade sob estimação, obtida com o estimador da razão combinado, será dada por

$$V(\hat{Y}_{RC}) = \sum_{h=1}^k \frac{N_h^2 (1-f_h)}{n_h} (S_{yh}^2 + R_C^2 S_{xh}^2 - 2R_C \rho_{yh} S_{yh} S_{xh})$$

onde cada um dos símbolos já foi definido anteriormente.

2.3.5. Estimador da Razão Não Viciado

A razão não viciada é obtida a partir dos valores médios das razões, $r_i = y_i / x_i$:

$$\bar{r} = \frac{1}{n} \sum_{i=1}^n r_i = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{x_i}$$

$$\frac{1}{N} \sum_{i=1}^N r_i (x_i - \bar{X}) = \frac{1}{N} \sum_{i=1}^N \frac{y_i}{x_i} - \left(\frac{1}{N} \sum_{i=1}^N r_i \right) \bar{X} = \bar{Y} - \bar{X} E(r_i)$$

Mas, na amostragem aleatória simples, $E(\bar{r}) = E(r_i)$, e, portanto, o vício de \bar{r} , é:

$$vicio(\bar{r}) = E(\bar{r}) - R = -\frac{1}{\bar{X}N} \sum_{i=1}^N r_i (x_i - \bar{X})$$

Agora, pelo RESULTADO 2.2.2.2.3, uma estimativa não viciada de uma

expressão similar à do vício acima: $\frac{1}{N-1} \sum_{i=1}^N r_i (x_i - \bar{X})$ é dada por:

$\frac{1}{n-1} \sum_{i=1}^n r_i (x_i - \bar{x}) = \frac{n}{n-1} (\bar{y} - \bar{r} \bar{x})$. Assim, a estimativa de \bar{r} , com o vício corrigido,

será dada por $\bar{r}' = \bar{r} + \frac{n(N-1)}{(n-1)N\bar{x}} (\bar{y} - \bar{r} \bar{x})$, e a estimativa do valor populacional Y ,

dada pelo estimador da razão não viciado, será: $\hat{Y} = \bar{r}' X = \bar{r} X + \frac{n(N-1)}{(n-1)} (\bar{y} - \bar{r} \bar{x})$.

2.3.6. Intervalo de Confiança para os estimadores da Razão

Os limites inferior e superior do intervalo de confiança da estimativa populacional \hat{Y} , pelos estimadores da razão, separado e não viciado são determinados pela expressão:

$$\left[\hat{Y} - z_{1-\alpha} \sqrt{\frac{M_h^2 \left(1 - \frac{m_h}{M_h}\right)}{m_h} \left(\hat{\sigma}_{ch}^2 + \hat{\theta}_h^2 \hat{\sigma}_{ah}^2 - 2\hat{\theta}_h \hat{\rho}_h \hat{\sigma}_{ch} \hat{\sigma}_{ah}\right)}, \hat{Y} + z_{1-\alpha} \sqrt{\frac{M_h^2 \left(1 - \frac{m_h}{M_h}\right)}{m_h} \left(\hat{\sigma}_{ch}^2 + \hat{\theta}_h^2 \hat{\sigma}_{ah}^2 - 2\hat{\theta}_h \hat{\rho}_h \hat{\sigma}_{ch} \hat{\sigma}_{ah}\right)} \right],$$

onde:

M_h é o número de localidades do grupo ou estrato h;

m_h é o número de localidades amostradas no grupo ou estrato h;

$\hat{\sigma}_{ch}^2 = \frac{\sum_{i=1}^{m_h} (y_{ch} - \bar{y}_{ch})^2}{m_h - 1}$ é a variância da quantidade de equipamentos compartilhados amostrada no grupo ou estrato h;

$\hat{\sigma}_{ah}^2 = \frac{\sum_{i=1}^{m_h} (x_{ch} - \bar{x}_{ch})^2}{m_h - 1}$ é a variância da quantidade de equipamentos amostrada no grupo ou estrato h;

$\hat{\rho}_h = \frac{\sum_{i=1}^{m_h} (x_{ahi} - \bar{x}_{ah})(y_{chi} - \bar{y}_{ch})}{m_h \sigma_{ch} \sigma_{ah}}$ é a correlação entre os amostrados e compartilhados amostrados no estrato ou grupo h;

$\hat{\theta}_h = \frac{\bar{y}_{ch}}{\bar{x}_{ah}}$ é a estatística amostral que estima a razão,
 $\theta = \frac{Y}{X}$;

x_{ahi} é o valor da quantidade conhecida X, na localidade i, amostrada no estrato ou grupo h;

\bar{x}_{ah} é o valor médio da quantidade X, dos amostrados no estrato ou grupo h;

y_{ahi} é o valor amostral da quantidade desconhecida Y, na localidade i, amostrada no estrato ou grupo h;

\bar{y}_{ah} é o valor médio da quantidade desconhecida Y, dos amostrados no estrato ou grupo h;

X é a quantidade total, conhecida na população, ou a quantidade total de equipamentos na população;

Y é a quantidade total desconhecida na população, ou seja, a quantidade total de equipamentos compartilhados, na população.

2.4. Expansão Simples

O método de estimação da Expansão Simples estima o valor médio populacional. Sua estimativa é dada pela expressão:

$\hat{\bar{Y}} = \sum_{h=1}^k \bar{y}_{ch}$ que é a estimativa do valor médio populacional, da quantidade populacional sob estimação, Y, pelo método da Expansão Simples, onde:

$\bar{y}_h = \frac{1}{m_h} \sum_{i=1}^{m_h} y_{chi}$ é o valor médio da quantidade populacional sob estimação, Y, no estrato ou grupo h;

m_h é o número de localidades amostradas no estrato ou grupo h;

y_{chi} é o valor da quantidade populacional sob estimação, Y, na localidade i, do estrato ou grupo h;

2.5 Regressão Linear

2.5.1 Introdução

A análise de regressão é uma técnica estatística usada para investigar e modelar o relacionamento entre variáveis. O estatístico inglês Galton estudando o tamanho de ervilhas, usou a palavra *regressão* pela primeira vez. As aplicações dessa técnica são numerosas e ocorrem em quase todos os Variáveis científicos, podendo a Análise de Regressão ser considerada como a mais usada das técnicas estatísticas. Aqui apresenta-se de forma resumida uma parte dessa técnica.

2.5.2 Estrutura Probabilística

Seja o modelo $\underline{Y} = X\underline{\beta} + \underline{\varepsilon}$ tal que \underline{Y} e $\underline{\varepsilon}$ têm dimensão n (número de observações), $\underline{\beta}$ tem dimensão p (número de parâmetros do modelo) e X (matriz do modelo) é de ordem $n \times p$. São admitidas para o modelo as seguintes suposições:

- 1ª) o vetor de erros $\underline{\varepsilon}^t = [\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n]$ é um vetor aleatório, ou seja, seus componentes ε_i , $i = 1, 2, \dots, n$ são variáveis aleatórias;
- 2ª) a esperança de cada componente de $\underline{\varepsilon}$ é zero, ou seja, $E(\underline{\varepsilon}) = \underline{0}$
- 3ª) os componentes do vetor $\underline{\varepsilon}$ não são correlacionados, ou seja, $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$ para $i \neq j$ e possuem variância constante, σ^2 . Assim, a matriz de Covariância de $\underline{\varepsilon}$ é a matriz diagonal $\sigma^2 I_n$, onde I_n é a matriz identidade de ordem n , $V(\underline{\varepsilon}) = \sigma^2 I_n$.

O modelo acima, com as três suposições anteriores, é conhecido como Modelo Linear de Gauss Markov e o Teorema de Gauss-Markov garante que, sob as três suposições e com $X^t X$ não singular, o estimador não viciado uniformemente de mínima variância (UMVU) do vetor $\underline{\beta}$ é $\hat{\underline{\beta}} = (X^t X)^{-1} X^t \underline{Y}$

Além das três suposições já citadas sobre o modelo, a seguinte também é comum:

4ª) a distribuição de ε_i $i=1,2, \dots, n$ é a Normal (Gaussiana).

Admitida essa última suposição, tem-se o modelo de Gauss-Markov Normal.

2.5.3 Análise da Variância da Regressão

A Análise da Variância é uma das técnicas estatísticas cujas bases foram lançadas por Fisher. Essa técnica é geralmente usada para verificar se o ajuste feito é bom. Inicialmente vai-se considerar o modelo com um suporte (uma covariável) apenas

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad i = 1, 2, \dots, n$$

Partindo-se da identidade $y_i - \bar{y} = \hat{y}_i - \bar{y} + y_i - \hat{y}_i$, elevando-se ao quadrado e somando-se para $\forall i$, tem-se:

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

É comum construir-se um quadro que resume as informações da Análise da Variância:

QUADRO DA ANÁLISE DE VARIÂNCIA

Modelo $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, com $p = 2$ parâmetros

Fonte de variação	Soma de quadrados	G.L.	Quadrado médio	F
Regressão	$SQ_{\text{Regr}} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	$p-1 = 1$	$\frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{p-1}$	$\frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 / (p-1)}{\sum_{i=1}^n (y_i - \hat{y}_i)^2 / n-p}$
Residual	$SQR = \sum_{i=1}^n (y_i - \hat{y}_i)^2$	$n-p = n-2$	$\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-p}$	
Total	$SQT = \sum_{i=1}^n (y_i - \bar{y})^2$	$n-1$		

O estimador UMVU da variância do erro ε_i , $V(\varepsilon_i) = \sigma^2$, é o quadrado médio dos resíduos

$$\hat{V}(\varepsilon_i) = \hat{\sigma}^2 = s^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-p}$$

O teste que se faz com a estatística F do quadro acima é o da hipótese $H_0: \beta_1 = 0$, ou seja, verifica-se se existe relação linear entre as variáveis Y e X.

A generalização para $p > 2$ parâmetros pode ser feita imediatamente:

QUADRO DA ANÁLISE DE VARIÂNCIA

Modelo $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_{p-1} X_{p-1i} + \varepsilon_i$

Fonte de variação	Soma de quadrados	G.L.	Quadrado médio	F
Regressão	$SQ_{Regr} = \hat{\beta}' X' Y - n\bar{y}^2$	p-1	$\frac{SQ_{Regr}}{p-1}$	$\frac{SQ_{Regr}}{p-1} / \frac{SQR}{n-p}$
Residual	$SQR = Y' Y - \hat{\beta}' X' Y$	n-p	$\frac{SQR}{n-p}$	
Total	$SQT = Y' Y - n\bar{y}^2$	n-1		

O teste feito com a estatística F acima é o da hipótese nula $H_0: \beta_1 = \beta_2 = \dots = \beta_{p-1} = 0$ ou seja, se existe regressão dos X's para Y, ou melhor, se existe regressão linear entre a variável resposta Y e as covariáveis X_i $i = 1, 2, \dots, p-1$.

2.5.4 Coeficiente de Correlação Múltipla ao Quadrado (R^2)

Para se medir a adequação do ajuste compara-se a *Soma de Quadrados da Regressão* com a *Soma de Quadrados Total* e tem-se o coeficiente de determinação ou correlação múltipla ao quadrado,

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad 0 < R^2 < 1$$

Quando o ajuste é bom, o modelo explica boa parte da variação total e, conseqüentemente, o valor de R^2 é próximo de 1. Em caso contrário, com um modelo pobre, o valor de R^2 é pequeno (CHAVES NETO, 1999).

2.6. Técnicas de Reamostragens

2.6.1 Introdução

A maioria dos métodos estatísticos de uso corrente hoje, foram desenvolvidos entre 1800 e 1930 (DIACONIS; EFRON, 1983 p. 116), época em que a computação de muitos dados era lenta e extremamente cara pois não existiam as calculadoras eletrônicas ou os computadores modernos de hoje.

A teoria estatística elementar, que é ordinariamente ensinada nas escolas, é uma versão simples de uma teoria complicada que foi desenvolvida a fim de evitar uma grande quantidade de cálculos numéricos. Com o advento da computação rápida e barata, novos métodos estatísticos foram desenvolvidos ou se tornaram disponíveis. Entre esses métodos encontram-se o de Monte Carlo, conhecido há mais tempo, o Jackknife, desenvolvido a partir do fim da década de 1940, e o Bootstrap, desenvolvido independentemente por Efron (CHAVES NETO, 1991) em 1977. O método Monte Carlo é usado geralmente para simulação e os outros métodos, Jackknife e Bootstrap, são técnicas estatísticas baseadas em computação intensiva.

Esses métodos empregam técnicas de reamostragem, ou seja, eles se valem de um conjunto de dados originado de observações ou de dados hipotéticos gerados por algum mecanismo como, por exemplo, lançamento de dados, retirada de carta de um baralho, lançamento de uma moeda, roleta russa, etc., para produzir novos dados hipotéticos para, com eles, fazer-se inferências estatísticas, ou seja, as técnicas de reamostragem são técnicas de simulação de novos dados hipotéticos a partir de dados originais.

2.6.2 O Método Monte Carlo

O método Monte Carlo é um método numérico de resolução de problema matemático por reamostragem aleatória. Como técnica numérica universal, o método Monte Carlo pôde ser aplicado de forma plena somente com o advento do computador e seu Variável de aplicação é expandido a cada nova geração de computadores (SOBOL, 1994, p.1). Os problemas tratados pelo método Monte

Carlo são de dois tipos: probabilísticos e determinísticos, segundo estejam eles relacionados ou não diretamente com o comportamento e resultados de processos aleatórios.

A primeira publicação de que se tem notícia sobre o Método Monte Carlo data de 1949, no *Journal of American Statistical Association* (HAMMERSLEY; HANDSCOMB, 1964), por N. Metropolis e S. Ulam. Sabe-se, no entanto, que o fundamento teórico do método já era conhecido bem antes dessa data, e que certos problemas de estatística foram algumas vezes resolvidos por esse método. Como a sua aplicação repetitiva e mecânica é enfadonha, o uso intensivo do método só foi possível com o advento da computação rápida e barata.

A utilização real do método Monte Carlo como uma ferramenta de pesquisa provém do trabalho sobre a bomba atômica durante a Segunda Guerra Mundial. No entanto, o desenvolvimento sistemático dessas idéias provém do trabalho de Harris e Herman Kahn, em 1948. Ainda nessa década, os autovalores da equação de Shöredinger foram estimados pelo método Monte Carlo. No entanto, as idéias modernas do método já eram conhecidas e foram, por exemplo, utilizadas por Lord Kelvin, sessenta anos antes dessa publicação, em um artigo onde discutia a solução da equação de Boltzman. Kelvin, porém, estava mais preocupado com os resultados em si do que com o método propriamente dito, de forma que o mérito da descoberta do método é atribuído a Ulan e a Fermi, não só por terem redescoberto o método, mas também por garantirem aos seus colegas cientistas, as possibilidades e potencialidades das aplicações do método na física (HAMMERSLEY; HANDSCOMB, 1964).

A partir dessa época, seguiu-se um intenso estudo do método Monte Carlo, principalmente na década de 1950. Paradoxalmente, isso foi o suficiente para pôr o método em descrédito. Havia uma tendência de se tentar resolver com ele todo o tipo de problema, sem dar atenção a quais desses problemas o método podia ser aplicado eficientemente e quais o método manipulava ineficientemente. Como consequência, o método caiu em descrédito. Nos últimos anos, o método Monte Carlo está voltando à tona. Principalmente devido ao melhor conhecimento a respeito de para quais problemas ele é a melhor, e as vezes a única técnica de

solução disponível. Tais problemas têm crescido em número, principalmente pelo incremento de técnicas de redução da variância recentemente descobertas, onde o método prévio era ineficiente, e parcialmente, porque o método tende a florescer em problemas que envolvem complicações práticas encontradas mais e mais freqüentemente na matemática aplicada e na pesquisa operacional (HAMMERSLEY; HANDSCOMB, 1964, p.9).

Uma das principais forças da matemática teórica está relacionada com a abstração e a generalidade: alguém pode escrever expressões simbólicas ou equações teóricas que abstraem a essência de um problema e revela a sua estrutura subjacente. Contudo, essa mesma força carrega consigo uma fraqueza inerente: quanto mais geral e formal é a sua linguagem, menos disponível está a teoria de prover uma solução numérica em uma aplicação particular. A idéia por detrás da aplicação do Método Monte Carlo em problemas determinísticos é explorar a força da matemática teórica e, ao mesmo tempo, evitar a sua fraqueza inerente, substituindo teoria por ‘experimento’ sempre que ela falhar. Especificamente, suponha haver um problema determinístico que possa ser formulado em linguagem teórica mas que não se possa resolvê-lo, por meios teóricos. Sendo determinístico, tal problema não tem nenhuma associação direta com processos aleatórios; mas, quando a teoria expõe a sua estrutura subjacente, talvez seja possível que essa estrutura ou expressão formal também descreva um processo aleatório aparentemente relacionado, e então pode-se resolver o problema determinístico numericamente, por meio de uma simulação Monte Carlo do problema probabilístico concomitante. Por exemplo, um problema da teoria eletromagnética pode requerer a solução da equação de Laplace sujeita a certas condições de contorno que frustra os métodos analíticos padrões. Agora, a equação de Laplace ocorre largamente nos estudos de partículas que se dispersam aleatoriamente em uma barreira de absorção. Assim, é possível resolver o problema eletromagnético realizando um experimento, no qual as partículas são guiadas por meio de números aleatórios até que elas sejam absorvidas sobre barreiras especialmente escolhidas para representar as condições de contorno prescritas (HAMMERSLEY; HANDSCOMB, 1964, p. 4-5).

O método Monte Carlo sempre provê uma solução logicamente aceitável. Porém, mais especificamente, é com relação aos testes de hipóteses, estatísticos, que o método Monte Carlo baseado sobre uma lógica de aleatorização, tem propriedades que os estatísticos estão julgando agora atrativas, porque elas são mais robustas que os testes paramétricos tradicionais (SIMON; ATKINSON;SEKOVAS, 1976).

2.6.2.1. Exigências de um procedimento Monte Carlo:

O procedimento Monte Carlo exige as seguintes condições:

- 1ª – As distribuições e os parâmetros do processo devem ser especificados. Eles poderão variar durante as diferentes rodadas do experimento, mas devem ser especificados a cada passo;
- 2ª – O problema não deveria ser propenso a uma solução rápida pelos métodos numéricos ou determinísticos usuais. Isto é, ele pode ser solucionável por meio de métodos analíticos, mas a solução é mais econômica pelo método Monte Carlo;
- 3ª – Requer-se um suprimento de números aleatórios (usualmente, uniformemente distribuídos).

Supondo que se deseje estimar o valor de uma quantidade desconhecida Z . Denotando a estimativa Monte Carlo de Z por z' , fazendo-se N ensaios para obter z' . Então, a estimativa Monte Carlo de Z , será dada por (1), e uma estimativa da variância de Z por (2), respectivamente,

$$z' = \frac{1}{N} \sum_{i=1}^N z_i \quad (1) \quad \text{e} \quad V = \left(\frac{1}{N} \right) \sum_{i=1}^N (z_i - z')^2 \quad (2)$$

Na execução do método Monte Carlo, aparentemente, o valor esperado de Z e sua variância podem ter seus erros reduzidos. A variância é função inversa do número de ensaios, e pode ser reduzida aumentando-se o número desses, o que poderá não ser viável economicamente, quando o número de ensaios for grande. Existe hoje contudo, um número considerável de técnicas desenvolvidas com o intuito de reduzir a variância, das quais mencionaremos algumas:

- 1-Amostragem dentro da região de Importância (Importance Sampling);
 - 2-Uso de Valores Esperados (Combinação de Métodos Analíticos e Probabilísticos);
 - 3-Correlação e Regressão;
 - 4-Amostragem sistemática;
 - 5-Amostragem Estratificada;
 - 6-Outras.
- (LACHENBRUCH, 1965 p. 10-12)

Um exemplo de estimativa de uma área simples por Monte Carlo pode ser visualizada no Anexo 01.

2.6.3. O Jackknife

O “Jackknife” é uma técnica de reamostragem computacionalmente intensiva, desenvolvida por Quenoüille (MILLER, 1974) e apresentada à comunidade científica em sua primeira versão em 1949. Consistia em um método para reduzir o vício de um estimador de correlação serial, com base na divisão da amostra original em duas semi-amostras. Posteriormente, em 1956, Quenoüille (MILLER, 1974) apresentou a versão generalizada e aperfeiçoada do método, onde a amostra original de tamanho n é dividida em g sub-amostras de tamanho h . Desse modo, tem-se $n = gh$. Fixando-se $h = 1$, a amostra original passa a ter n grupos com uma unidade observacional cada. Daí em diante, o método passou a ser aplicado de forma abrangente em vários problemas. (CHAVES NETO, 1991, p.21).

A partir de uma amostra aleatória de tamanho n , $\underline{X} = (X_1, X_2, \dots, X_n)$, de variáveis independentes e identicamente distribuídas (i.i.d.), com função de verossimilhança $f(\underline{X}, \theta)$, dessa amostra aleatória dependente do parâmetro θ , o procedimento “Jackknife”, na sua versão generalizada, pode ser descrito pela sequência de passos :

- 1) Da amostra original \underline{X} , toma-se uma outra amostra de tamanho $n-1$, $\underline{X}_{(-i)} = (X_1, X_2, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$, deixando fora o i -ésimo elemento amostral, X_i ;
- 2) Calculam-se as estatísticas $\hat{\theta} = T_n(\underline{X})$ e $\hat{\theta}_{(-i)} = T_{n-1}(\underline{X}_{(-i)})$, como estimativas de θ e de $\theta_{(-i)}$, respectivamente;
- 3) Calculam-se os valores $\tilde{\theta}_{(i)} = n\hat{\theta} - (n-1)\hat{\theta}_{(-i)}$, $n = 1, 2, \dots, n$, chamados pseudovalores;
- 4) Determina-se a estimativa “Jackknife” de θ , $\tilde{\theta}$ definida por $\tilde{\theta} = \sum_{i=1}^n \tilde{\theta}_{(i)} / n$, a média aritmética dos pseudovalores;
- 5) Determina-se o estimativa “Jackknife” do erro padrão de $\tilde{\theta}$ por

$$ep_{jack} = \left[\frac{1}{n(n-1)} \sum_{i=1}^n (\tilde{\theta}_{(i)} - \tilde{\theta})^2 \right]^{1/2} \quad (\text{CHAVES NETO, 1991}).$$

Embora os pseudovalores sejam intrigantes, não está claro serem eles a melhor maneira de se pensar o “Jackknife” (EFRON; TIBSHIRANI 1993 p.145)

2.6.3.1 Médias, Erro Padrão e Vícios

Suponha-se ter um vetor amostral \underline{X} , composto por quantidades aleatórias, x_i , $i=1, 2, \dots, n$, independentes e identicamente distribuídas (i.i.d.), de uma distribuição de probabilidades F desconhecida, isto é, se $x_i \sim F$, $i=1, 2, \dots, n$, i.i.d., seja θ , um parâmetro de interesse, $\theta = \theta(F)$, esperança, correlação, etc., o qual estima-se pela estatística $\hat{\theta} = \theta(\hat{F})$, onde \hat{F} é uma distribuição empírica de probabilidades, obtida

colocando-se massa $1/n$, em cada valor amostral, x_i , $i=1, 2, \dots, n$, da amostra original. O método “Jackknife” consiste em remover sucessivamente da amostra original um valor amostral x_i e recalculando $\hat{\theta}$. Removendo-se um valor amostral da amostra original, obtém-se uma distribuição empírica de probabilidades diferente $\hat{F}_{(-i)}$ com massa $\frac{1}{n-1}$, em cada valor amostral restante x_j , $j \neq i$ e o correspondente valor recalculado da estatística, e $\underline{X}_{(-i)} = (x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_n)$ chamada a i -ésima amostra “Jackknife” do vetor amostral \underline{X} original. Tem-se dessa forma $\hat{\theta} = \frac{1}{n} \sum_{i=1}^n x_i$, $\hat{\theta}_{(-i)} = \frac{1}{n-1} \sum_{j \neq i} x_j$ e $\hat{\theta}_{(.)} = \frac{1}{n} \sum_{i=1}^n \hat{\theta}_{(-i)}$ como as estimativas de θ de $\theta_{(-i)}$ e de $\hat{\theta}$ respectivamente. A estimativa “Jackknife” do erro padrão da estimativa “Jackknife” $\hat{\theta}_{(.)}$, de θ é definida como sendo: $ep_{jack} = \left[\frac{1}{n(n-1)} \sum_{i=1}^n (\tilde{\theta}_{(i)} - \tilde{\theta})^2 \right]^{1/2}$ e a estimativa “Jackknife” do vício de $\hat{\theta}_{(.)}$, por $\hat{b}_{jack}(\hat{\theta}, \hat{\theta}_{(.)}) = (n-1)(\hat{\theta}_{(-i)} - \hat{\theta}_{(.)})$ e, para a estimativa do vício da variância, $\hat{b}_{jack}(\hat{\sigma}^2, \hat{\sigma}_{jack}^2) = -\frac{\hat{\sigma}^2}{n}$ (EFRON; TIBSHIRANI, 1993, p.151).

2.6.3.2. O Jackknife e Regressão Linear

Seja $\underline{Y} = \underline{X} \underline{\beta} + \underline{\varepsilon}$, onde \underline{Y} é o vetor das respostas, \underline{X} é a matriz do modelo, de ordem $n \times p$, $\underline{\beta}$ é um vetor desconhecido, o vetor dos parâmetros do modelo de regressão linear, e $\underline{\varepsilon}$ é o vetor dos resíduos. A estimativa “Jackknife” de $\underline{\beta}$, $\hat{\underline{\beta}}$ é calculada com base nos n pseudovalores, $\hat{\underline{\beta}}_{(-i)}$, $\hat{\underline{\beta}}_{(-i)} = (\underline{X}_{(-i)}^t \underline{X}_{(-i)})^{-1} \underline{X}_{(-i)}^t \underline{Y}_{(-i)}$, onde $\underline{X}_{(-i)}$ é a matriz obtida da matriz do modelo \underline{X} , eliminando-se a i -ésima linha, e $\underline{X}_{(-i)}^t$ é a transposta da matriz $\underline{X}_{(-i)}$, e $\underline{Y}_{(-i)}$ é o vetor obtido do vetor das respostas, eliminando-se a i -ésima linha, ou seja:

$$\underline{Y}_{(-i)} = \begin{bmatrix} Y_1 \\ Y_2 \\ \cdot \\ \cdot \\ Y_{i-1} \\ Y_{i+1} \\ \cdot \\ \cdot \\ Y_n \end{bmatrix}, \quad \underline{\varepsilon}_{(-i)} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \varepsilon_{i-1} \\ \varepsilon_{i+1} \\ \cdot \\ \cdot \\ \varepsilon_n \end{bmatrix}, \quad \underline{X}_{(-i)} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & X_{(i-1)} \\ 1 & X_{(i+1)} \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & X_n \end{bmatrix}, \quad \underline{\beta}_{(-i)} = \begin{bmatrix} \beta_{0(-i)} \\ \beta_{1(-i)} \end{bmatrix}$$

As estimativas “Jackknife” são dadas, então, por :

$$\hat{\underline{\beta}} = \frac{1}{n-1} \sum_{i=1}^n \hat{\beta}_{(-i)}, \quad \hat{\underline{\varepsilon}} = \frac{1}{n-1} \sum_{i=1}^n \hat{\varepsilon}_{(-i)}, \quad \text{cov}_{jack}(\hat{\underline{\beta}}) = \frac{n-1}{n} \sum_{i=1}^n [\hat{\beta}_{(-i)} - \hat{\underline{\beta}}]^* [\hat{\beta}_{(-i)} - \hat{\underline{\beta}}]$$

$$\hat{\sigma}_{jack}(\hat{y}_i) = \frac{n-1}{n} \sum_{i=1}^n \hat{\varepsilon}_{(-i)}^2 \quad (\text{EFRON, 1982, p. 18-19}).$$

Exemplo: Amostras Jackknife

Seja $\underline{X} = (94, 19, 16, 38, 99, 14, 23)$, um vetor amostral. Para esse vetor amostral, vamos listar nove amostras Jackknife e as médias de cada amostra, A1, A2, ...A7, MA1, MA2, ...MA7, MO a média da amostra original, FA é a frequência de cada ponto amostral em cada amostra, FT é a frequência total, em todas as amostras de cada ponto amostral, MT é a média total das médias, MA1, MA2, ...MA7 e, $i = 1, 2, \dots, 7$ é o índice do ponto amostral da amostra original.

Amostra original

MO = média da amostra original

i	1	2	3	4	5	6	7	Médias
u	94	19	16	38	99	14	23	MO=43,28

Amostras Jackknife,

MA(i) = média(i) da amostra Jackknife, A(i).

A1	19	19	16	38	99	14	23	MA1= 38
A2	94	94	16	38	99	14	23	MA2= 63
A3	94	94	19	38	99	14	23	MA3= 63,5
A4	94	94	19	16	99	14	23	MA4= 59,83
A5	94	94	19	16	38	14	23	MA5= 49,66
A6	94	94	19	16	38	99	23	MA6= 63,83
A7	94	94	19	16	38	99	14	MA7= 62,33
MT = 57,16								

Frequência de cada ponto amostral da amostra original em cada uma das amostras Jackknife, acima;

FA1	00	01	01	01	01	01	01
FA2	01	00	01	01	01	01	01
FA3	01	01	00	01	01	01	01
FA4	01	01	01	00	01	01	01
FA5	01	01	01	01	00	01	01
FA6	01	01	01	01	01	00	01
FA7	01	01	01	01	01	01	00

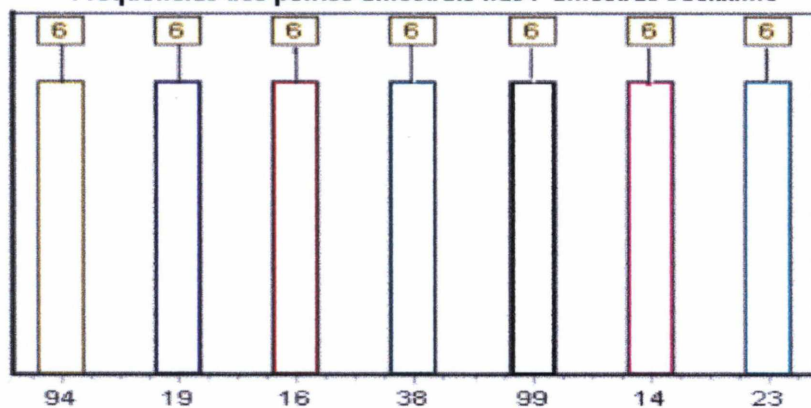
Frequência total de cada ponto amostral original, nas sete amostras Jackknife.

FT	06	06	06	06	06	06	06
----	----	----	----	----	----	----	----

Erro padrão Jack = 22,21

Gráfico 00.01: Frequências dos pontos amostrais das amostras Jackknife, acima.

Frequências dos pontos amostrais nas 7 amostras Jackknife



O número na base de cada retângulo é o valor amostral do vetor de amostras originais. No topo de cada retângulo, está o número de vezes que o ponto amostral aparece em todas as sete amostras Jackknife, acima.

2.6.4. O Bootstrap

2.6.4.1. Introdução

O método Bootstrap é também uma técnica de reamostragem computacionalmente intensiva, que tem suas raízes no Jackknife (YUNG, 1994, p. 393). Desenvolvida por Efron, em 1977, tem vantagens sobre os métodos estatísticos convencionais, tais como a de não depender da hipótese de que os dados provenham de uma distribuição normal, bem como a de não ser necessário o uso de fórmulas complicadas ou mesmo inexistentes para a solução exata e ainda a vantagem de tratar de amostras pequenas. O método Bootstrap substitui essas fórmulas por cálculos simples, porém volumosos, realizados pelo computador (SWANEPOEL, 1990, p.2).

A idéia subjacente ao Bootstrap é reamostrar um número grande de vezes a amostra original, para com elas obter as estatísticas e precisões desejadas.

Foram desenvolvidas diversas formas de aplicar o método Bootstrap, segundo as reamostragens sejam feitas com ou sem reposição, e o tamanho da amostra seja ou não igual ao da amostra original (RAO; KATZOFF, 1996). Há também diversos tipos de Bootstrap: paramétrico, não paramétrico e Bayseniano. No Bootstrap paramétrico, a forma da distribuição é conhecida e no Bootstrap não paramétrico e no Bayseniano, não. No caso não paramétrico, a função de distribuição F é uma distribuição onde é colocada massa $1/n$ em cada ponto amostral, no tipo Bayseniano, a distribuição é semelhante ao do caso não paramétrico, com a diferença de substituir o tamanho dos saltos da função de distribuição empírica pelas diferenças de $n-1$ variáveis aleatórias, independentes e identicamente distribuídas, segundo uma distribuição uniforme no intervalo $(0,1)$ (SALINAS, 1998, p.7-8).

Todo o método está calcado sobre o princípio chamado “plug-in”, que associa o valor de um parâmetro, cuja distribuição é desconhecida, à estatística amostral de uma distribuição empírica, obtida a partir da amostra disponível, colocando

probabilidade $1/n$ em cada um desses pontos amostrais dessa amostra (EFRON; TIBSHIRANI, 1993).

A parte difícil do procedimento Bootstrap refere-se ao efetivo cálculo da sua distribuição (EFRON, 1979, p. 1-26). Para esse cálculo, três métodos são possíveis:

- 1- Cálculo direto teórico, através da obtenção das amostras Bootstrap e de uma distribuição empírica, onde é atribuída massa $1/n$ a cada ponto amostral.
- 2- Aproximação Monte Carlo à distribuição Bootstrap, acima.
- 3- O método da expansão em séries de Taylor para a obtenção da média e da variância aproximadas da distribuição Bootstrap.

Encerrando esta introdução sobre o Bootstrap, deseja-se salientar primeiramente que tanto o Jackknife quanto o Bootstrap não são uma panacéia e não são remédios para um inadequado tamanho de amostra. No caso não paramétrico, a distribuição amostral deve estar, em algum sentido, próximo da distribuição populacional para que se obtenha inferências acuradas. No caso paramétrico somente o parâmetro estimador necessita estar próximo ao parâmetro populacional para que a inferência seja acurada. Isso porque, quanto menor for a amostra, maiores serão as flutuações da distribuição amostral. Os métodos não paramétricos, que são sensíveis a uma grande variedade de tais flutuações, serão mais afetados pelos menores tamanhos das amostras do que os métodos paramétricos se as hipóteses dos métodos paramétricos forem válidas (SARLE, 1995 p. 3). Em segundo lugar salienta-se que o Bootstrap tem os seus entusiastas e os seus críticos, como citado em PETERSON (1991 p. 3), por exemplo: “O trabalho de Efron, tem influenciado tremendamente as rotinas do pensamento estatístico, ...é uma técnica muito poderosa (o Bootstrap) - a maior contribuição à teoria e à prática da estatística” e, “Nem todos estão enamorados com estas técnicas [de reamostragens]. Os céticos argumentam que ‘você está tentando obter algo a partir do nada. Você usa os mesmos números vez após vez até ter uma resposta que não poderia obter de outra maneira. A fim de fazer isto, você deve assumir algo, e você poderá viver o suficiente para arrepender-se daquilo que assumiu’ Stephen E. Fienberg - York University in Toronto”. Por último as palavras finais de um revisor crítico dos métodos Bootstrap: “O esforço da

realização de Efron para o uso da idéia de reamostragem, especialmente quando combinado com o poder da computação moderna, tem de fato provocado um impacto significativo no pensamento estatístico, mas relativamente pouco, na prática estatística. O Bootstrap não paramétrico tem sido usado regularmente para suplementar a maioria das análises estatísticas padrão em problemas simples de estimação do erro, mas a história termina por aí” (YUNG, 1994).

2.6.4.2. O Procedimento Bootstrap

Sejam $\underline{X} = (X_1, X_2, \dots, X_n)$ uma amostra aleatória (a.a.), de tamanho n , proveniente de uma distribuição de probabilidades F desconhecida onde cada x_i , $i = 1, 2, \dots, n$ é independente e identicamente distribuído (i.i.d.) em F , $x_i \sim F$, $i = 1, 2, \dots, n$, θ é um parâmetro e $t(\underline{X})$ o seu estimador. O procedimento Bootstrap de reamostragens com reposição toma amostras com o mesmo tamanho, n , da amostra original \underline{x} , e pode ser descrito através dos seguintes passos :

1º) O estimador de máxima verossimilhança não paramétrico de F , $\hat{F}(\underline{X})$, é a distribuição empírica, onde cada valor amostral x_i , $i = 1, 2, \dots, n$, da amostra original, tem probabilidade $1/n$, $\hat{F}(\underline{X}) = \sum_{i=1}^n I(x_i \leq \underline{X}) (1/n)$, onde $I(x_i \leq \underline{X})$ é a função indicadora;

2º) Da amostra $\underline{X} = (X_1, X_2, \dots, X_n)$, toma-se B (muito grande) amostras de mesmo tamanho n e, com reposição, da amostra original: $\underline{X}_1^* = (X_{11}^*, X_{12}^*, \dots, X_{1n}^*)$, $\underline{X}_2^* = (X_{21}^*, X_{22}^*, \dots, X_{2n}^*)$, ..., $\underline{X}_B^* = (X_{B1}^*, X_{B2}^*, \dots, X_{Bn}^*)$;

3º) Calcula-se as B estatísticas Bootstrap $T_b^* = T^*(\underline{x}_b^*)$, $b = 1, 2, \dots, B$ e, a partir do conjunto de valores $T^*(\underline{x}_b^*)$, $b = 1, 2, \dots, B$, calcula-se a estatística $\bar{T}_{boot}^* = \frac{1}{B} \sum_{b=1}^B T_b^*$, que é uma estimativa Bootstrap do verdadeiro parâmetro estimado pela estatística $T(\underline{X}, F)$.

O conjunto de valores “Bootstrap” $T^*(\underline{X}_b)$, $b = 1, 2, \dots, B$ estima a verdadeira distribuição amostral da estatística $T(\underline{X}, F)$. E, denotando o número de vezes com que cada ponto amostral x_i , da amostra original aparece em cada uma das B , amostras Bootstrap, por N_i^* , ou seja, $N_i^* = \# \{X_i^* = X_i\}$, tem-se que, $\sum_{i=1}^n N_i^* = n$, e daí que $N^* = \{N_1^*, N_2^*, \dots, N_B^*\}$, tem distribuição multinomial. Também, em correspondência às frequências relativas, $P_i^* = N_i^* / n$, assumindo valores no conjunto $\{0/n, 1/n, 2/n, \dots, n/n\}$, $N_i^* \sim b(n, 1/n)$ e, P_i^* , tem média e variância dadas pelo resultado 2.6.4.2.1 (CHAVES NETO, 1991):

RESULTADO 2.6.4.2.1

Seja $P_i^* = N_i^* / n$, a proporção de vezes em que o ponto amostral X_i foi selecionado na amostra Bootstrap \underline{X}_b . Então, a variável aleatória P_i^* possui distribuição amostral com média $1/n$ e variância $(n-1)/n^3$.

Prova:

Como $N_i^* \sim b(n, 1/n)$, tem-se que :

$$E(P_i^*) = E(N_i^* / n) = (1/n) E(N_i^*) = 1/n, \text{ e}$$

$$V(P_i^*) = V(N_i^* / n) = (1/n^2) V(N_i^*) = (n-1)/n^3.$$

2.6.4.3. Melhor Estimativa Bootstrap do Vício (BBBE)

Sejam:

$$\underline{X} = (X_1, X_2, \dots, X_n)$$

o vetor da amostra original;

$\hat{\theta}_{inf}$	o 1000α - ésimo percentil Bootstrap;
$\hat{\theta}_{sup}$	o $1000(1-\alpha)$ - ésimo percentil Bootstrap;
$(\hat{\theta}_{inf}, \hat{\theta}_{sup}) = (\hat{\theta}^{*(\alpha)}, \hat{\theta}^{*(1-\alpha)})$	o intervalo de confiança BCa;
$\theta^{*(\alpha)}$	o extremo inferior do intervalo de confiança;
$\theta^{*(1-\alpha)}$	o extremo superior do intervalo de confiança;
$\underline{X}_1 = (X_{11}^*, X_{21}^*, \dots, X_{n1}^*)$	o primeiro vetor das amostras Bootstrap;
$P_{j1}^* = \# \{X_{i1}^* = X_j\} / k, j = 1, 2, \dots, n$	a proporção do número de vezes que cada valor amostral x_j da amostra original aparece na primeira amostra Bootstrap;
$\underline{X}_2 = (X_{12}^*, X_{22}^*, \dots, X_{n2}^*)$	o segundo vetor das amostras Bootstrap;
$P_{j2}^* = \# \{X_{i2}^* = X_j\} / k, j = 1, 2, \dots, n$	a proporção do número de vezes que cada valor amostral x_j da amostra original aparece na segunda amostra Bootstrap;
$\underline{X}_B = (X_{1B}^*, X_{2B}^*, \dots, X_{nB}^*)$	a B - ésima amostra Bootstrap;
$P_{jB}^* = \# \{X_{iB}^* = X_j\} / k, j = 1, 2, \dots, n$	a proporção do número de vezes que cada valor amostral x_j da amostra original aparece na B - ésima amostra Bootstrap;
$\underline{P}_1 = (P_{11}^*, P_{21}^*, \dots, P_{n1}^*)$	a proporção do número de vezes que o valor amostral x_j aparece na primeira amostra Bootstrap, $j = 1, 2, \dots, n$;
$\underline{P}_2 = (P_{12}^*, P_{22}^*, \dots, P_{n2}^*)$	a proporção do número de vezes que o valor amostral x_j aparece na segunda amostra Bootstrap, $j = 1, 2, \dots, n$;
$\underline{P}_B = (P_{1B}^*, P_{2B}^*, \dots, P_{nB}^*)$	a proporção do número de vezes que o valor amostral x_j aparece na B-ésima amostra Bootstrap, $j = 1, 2, \dots, n$;

$$\bar{P}^* = \frac{1}{B} \sum_{b=1}^B \underline{P}_b^*$$

o vetor média das proporções de ocorrência de cada valor amostral x_j , que aparecem nas B , amostras Bootstrap;

$$\bar{x}_b^* = \frac{1}{n} \sum_{i=1}^n x_i^*$$

o valor médio de cada amostra Bootstrap \underline{X}_b^* , $b=1,2,\dots,B$;

$$\bar{x}^* = \frac{1}{B} \sum_{b=1}^B \bar{x}_b^*$$

o valor médio Bootstrap, dos B valores médios \bar{x}_b^* .

A melhor estimativa Bootstrap do vício (bbbe) de \bar{x}^* é dado por

$$bbbe(\bar{x}^*) = \bar{x}^* - \bar{P}^* * \underline{X}'$$

2.6.4.4. Intervalo de Confiança Bootstrap, Acelerado com correção do Vício (Bca)

Seja $\hat{\theta}^*$ o 100α -ésimo, percentil das B estatísticas Bootstrap $\hat{\theta}_1^*, \hat{\theta}_2^*, \dots, \hat{\theta}_B^*$. O intervalo de convergência $1-2\alpha$, é obtido diretamente dos percentis.

Método percentil : $(\hat{\theta}_{\inf}, \hat{\theta}_{\sup}) = (\hat{\theta}^{*(\alpha)}, \hat{\theta}^{*(1-\alpha)})$.

Por exemplo, se $B=2000$ e $\alpha=0.05$, então, $(\hat{\theta}^{*(0.05)}, \hat{\theta}^{*(0.95)})$ é o intervalo entre o 100-ésimo e o 1900-ésimo valores ordenados dos $B=2000$ números, $\hat{\theta}_1^*, \hat{\theta}_2^*, \dots, \hat{\theta}_B^*$.

2.6.4.5. Estimativa Jackknife do Vício Bootstrap (JAB – Jackknife After Bootstrap)

Sejam $\underline{X}_{(-i)}^* = (X_1, X_2, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$ um dos B vetores de amostras Bootstrap que não contém o valor amostral original x_i , $i=1,2,\dots,n$, e que se tenha $B_{(i)}$ vetores sem o i -ésimo valor amostral x_i $i=1,2,\dots,B$ e $C_i = \{b / x_i \notin \underline{X}_b^* \mid i=1,2,\dots,n\}$ o conjunto dos índices das amostras Bootstrap \underline{X}_b^* , que não contém o valor amostral x_i , para cada i , $i=1,2,\dots,n$. Para cada $b \in C_i$, calculam-se

$$\bar{x}_b^* = \frac{1}{n} \sum_{i=1}^n x_{bi}, \quad \forall b \in C_i \quad \text{e} \quad \bar{x}_{B(i)}^* = \frac{1}{B_{(i)}} \sum_{b \in C_i} \bar{x}_b^* \quad \text{para } i=1,2,\dots,n \quad \text{e para cada } B_{(i)},$$

calculam-se $ep_{B(i)}$ e o erro padrão estimado Bootstrap das n estimativas, $ep_{B(.)}$.

$$ep_{B(.)} = \frac{1}{n} \sum_{i=1}^n ep_{B(i)}$$

E a estimativa Jackknife da variância de $ep_{B(i)}$

$$\text{var}_{jack}(ep_{B(.)}) = \left[\left(\frac{n-1}{n} \right) \sum_{i=1}^n \left(ep_{B(i)} - ep_{B(.)} \right)^2 \right]$$

$$\text{e o respectivo erro padrão } ep_{jack}(ep_{B(.)}) = \left[\left(\frac{n-1}{n} \right) \sum_{i=1}^n \left(ep_{B(i)} - ep_{B(.)} \right)^2 \right]^{1/2}.$$

2.6.4.6. Bootstrap e Regressão Linear

A aplicação do método Bootstrap, ao modelo de regressão linear, pode ser descrito pela seguinte sequência de passos (CHAVES NETO, 1985, p. 7-10):

1 – Seja, $\underline{Y} = X\underline{\beta} + \underline{\varepsilon}$ o modelo de regressão linear, sendo \underline{Y} o vetor de respostas, $\underline{\varepsilon}$ o vetor dos erros ou resíduos, ambos com dimensão n , $\underline{\beta}$ o vetor dos coeficientes de regressão com dimensão $p+1$, e X a matriz do modelo de ordem $(n, p+1)$ assumida não aleatória;

2 – Os resíduos ou erros $\underline{\varepsilon}$ são assumidos terem distribuição F , desconhecida, centrada em zero, de forma que:

$$\varepsilon_i \sim_{iid} F \quad \text{com} \quad E_F(\varepsilon_i) = 0;$$

3 – Com a matriz do modelo X especificada, faz-se um ajuste ao modelo através dos mínimos quadrados (por exemplo) obtendo-se as estimativas de $\underline{\beta}$, $\hat{\underline{\beta}}$:

$$\hat{\underline{\beta}} = (X'X)^{-1} X' \underline{Y};$$

4 – É determinado o vetor de resíduos:

$$\underline{e} = \underline{Y} - X \hat{\underline{\beta}}$$

5 – É estabelecida uma distribuição empírica de probabilidades \hat{F} para os resíduos corrigidos pela média deste modo, \hat{F} associa a probabilidade $1/n$ para cada resíduo e_i : $e_i \sim_{iid} \hat{F}$;

6 – É tomada uma amostra Bootstrap, \underline{e}^* de \hat{F} :

$$\underline{e}^* = (e_1, e_2, \dots, e_n);$$

7 – É construído um vetor de pseudodados \underline{Y}^* :

$$\underline{Y}^* = X \hat{\underline{\beta}} + \underline{e}^*;$$

8 – É feito um ajuste aos pseudo-dados e a estimativa Bootstrap $\hat{\underline{\beta}}^*$ de $\hat{\underline{\beta}}$ é determinada:

$$\hat{\underline{\beta}}^* = (X^t X)^{-1} X^t \underline{Y}^*;$$

9 – A distribuição Bootstrap de $\hat{\underline{\beta}}^*$ é uma aproximação da distribuição amostral de $\hat{\underline{\beta}}$.

RESULTADO 2.5.4.6

O estimador de mínimos quadrados ordinários do vetor de parâmetros $\hat{\underline{\beta}}$ no modelo $\underline{Y} = X \underline{\beta} + \underline{\varepsilon}$, $\hat{\underline{\beta}}^* = (X^t X)^{-1} X^t \underline{Y}^*$, tem média e variância Bootstrap dadas por

$$E_*(\hat{\underline{\beta}}^*) = \hat{\underline{\beta}} \quad \text{e} \quad V(\hat{\underline{\beta}}^*) = \hat{\sigma}^2 (X^t X)^{-1}$$

Prova:

Como os e_i são independentes com média zero e variância $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$ e

assim,

$$E_*(\underline{\hat{\beta}}^*) = E_*((X^t X)^{-1} X^t \underline{Y}^*) = (X^t X)^{-1} X^t E_*(\underline{Y}^*) = (X^t X)^{-1} X^t E_*(X \underline{\hat{\beta}} + \underline{e}^*) =$$

$$E_*(\underline{\hat{\beta}}^*) = (X^t X)^{-1} X^t (X \underline{\hat{\beta}} + 0) = (X^t X)^{-1} X^t \underline{Y}^* = E_*(\underline{\hat{\beta}}^*) = \underline{\hat{\beta}}$$

$$E_*(\underline{\hat{\beta}}^*) = \underline{\hat{\beta}}$$

$$V_*(\underline{\hat{\beta}}^*) = V[(X^t X)^{-1} X^t Y] = (X^t X)^{-1} X^t V(Y) X (X^t X)^{-1} = (X^t X)^{-1} X^t V(X \underline{\beta} + \underline{\varepsilon})$$

$$V_*(\underline{\hat{\beta}}^*) = (X^t X)^{-1} X^t V(X \underline{\beta} + \underline{\varepsilon}) X (X^t X)^{-1} = (X^t X)^{-1} X^t \sigma^2 X (X^t X)^{-1}$$

$$V_*(\underline{\hat{\beta}}^*) = \hat{\sigma}^2 (X^t X)^{-1}$$

.6.4.7. Redução de Vício em Amostras Bootstrap

O método Bootstrap é uma técnica de avaliação de propriedades dos procedimentos estatísticos em amostras repetidas. Em geral, estimativas numéricas incluem erros não só devido ao processo de simulação como também àqueles provenientes dos dados amostrais, os quais podem desempenhar um papel importante em todo o processo Bootstrap. Logo, é importante que se considere procedimentos que conduzam à redução da variância na simulação Bootstrap. Entre as técnicas desenvolvidas com esse fim, está a que considera o balanceamento das amostras Bootstrap, que será aqui descrita. O valor amostral da amostra original da estatística T será designado como T_{obs} , e aquelas relacionadas às B amostras

Bootstrap, por T_b^* , $b = 1, 2, \dots, B$. Uma aproximação normal para a estatística T é estimada ter Vício* e variância V^* dados por:

$$\text{Vício}^* = \bar{T}^* - T_{\text{obs}} = B^{-1} \sum_{b=1}^B T_b^* - T_{\text{obs}} \text{ e } V^* = \sum_{b=1}^B (T_b^* - \bar{T}^*)^2 / (B-1) \text{ respectivamente}$$

e as probabilidades $G(d) = P(T \leq \theta + d)$ podem ser estimadas diretamente pelas proporções simuladas $G^*(d) = B^{-1} \# \{i : T_i^* \leq T_{\text{obs}} + d\}$. Como o processo Bootstrap simula amostras aleatórias da função de distribuição empírica \hat{F} , isto é, $\hat{F}(x^*) = n^{-1} \# \{i : x_i \leq x^*\}$, pode-se, em princípio, estimar-se o vício e a variância de T , por $E_O^*(T) - T_{\text{obs}}$ e $V^* - \text{var}_O^*(T)$, e estimar-se $G(d)$ por $G^*(d) = P_O^*(T \leq T_{\text{obs}} + d)$, onde o asterisco denota amostra proveniente de \hat{F} , e o subscrito O , denota procedimento Bootstrap ordinário. As diferenças $E_O^*(T) - T_{\text{obs}}$ e $V^* - \text{var}_O^*(T)$, etc., denotam erros de simulação. Naturalmente que tais erros podem, em princípio, ser reduzidos aumentando-se B , mas essa pode não ser a melhor solução em um problema complicado. Os vícios, obtidos nesses processos, têm um comportamento randômico, pois são dependentes das frequências P_i^* com que os pontos amostrais x_i $i=1, 2, \dots, n$ aparecem nas B amostras Bootstrap. No caso, por exemplo, de $E(T^*) - E(T)$, o vício é da ordem de $B^{-1/2}$ (DAVIDSON; HINKLEY; SCHETCHTMAN, 1986), um erro de primeira ordem que pode ser removido, forçando $\sum_{b=1}^B P_{bi}^* = B$, isto é, fazendo-se que cada ponto amostral x_i , da amostra original seja igualmente freqüente no conjunto completo das B amostras Bootstrap, condição essa chamada de balanceamento amostral, ou permutação. Na prática, uma amostra Bootstrap balanceada ou uma “permutação” é obtida copiando-se B vezes os n pontos amostrais da amostra original, um ao lado do outro, formando uma seqüência de nB , pontos amostrais; a seguir, realiza-se uma permutação aleatória de toda a seqüência, e da nova seqüência obtida, a partir do primeiro ponto amostral, retira-se seqüencialmente as B amostras Bootstrap de tamanho n cada uma. Essas amostras, colocadas em forma de uma matriz $B \times n$, cujas B linhas terão, cada uma delas, $\sum_{i=1}^n P_{bi}^* = n$, e cada uma de suas colunas terão $\sum_{b=1}^B P_{bi}^* = B$.

(GRAHAM; HINKLEY; JOHN; SHI, 1990)

Exemplo: Amostras não Balanceadas (Bootstrap Simples)

Seja $\underline{X} = (94, 19, 16, 38, 99, 14, 23)$, um vetor amostral. Para esse vetor, vamos listar nove amostras, Bootstrap Simples e as médias de cada amostra, A_1, A_2, \dots, A_9 , MA_1, MA_2, \dots, MA_9 ; MO a média da amostra original, FA é a frequência de cada ponto amostral em cada amostra, FT é a frequência total, em todas as amostras de cada ponto amostral, MT é a média total das médias, MA_1, MA_2, \dots, MA_9 e $i = 1, 2, \dots, 7$, é o índice do ponto amostral da amostra original.

Amostra original								MO = média da amostra original
i	1	2	3	4	5	6	7	Médias
u	94	19	16	38	99	14	23	MO = 43,28

Amostras Bootstrap Simples								MA(i) = médias das amostras Bootstrap, A(i)
A1	14	38	99	38	38	23	38	MA1 = 41,14
A2	99	14	14	16	16	94	38	MA2 = 41,57
A3	99	19	19	14	19	99	14	MA3 = 40,42
A4	23	23	23	94	99	16	14	MA4 = 41,71
A5	19	16	19	19	14	19	94	MA5 = 28,57
A6	19	23	19	14	94	16	99	MA6 = 40,57
A7	99	19	99	38	14	14	94	MA7 = 53,85
A8	99	19	99	14	19	38	99	MA8 = 55,28
A9	14	23	38	23	23	38	94	MA9 = 36,14
								MT = 42,14

Frequência amostral de cada ponto da amostra original, em cada amostra Bootstrap Simples;

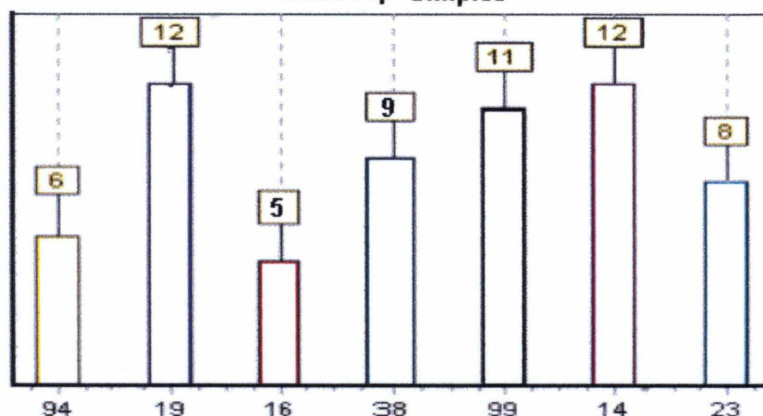
FA1	00	00	00	04	01	01	01
FA2	01	00	02	01	01	02	00
FA3	00	03	00	00	02	02	00
FA4	01	00	01	00	01	01	03
FA5	01	04	01	00	00	01	00
FA6	01	02	01	00	01	01	01
FA7	01	01	00	01	02	02	00
FA8	00	02	00	01	03	01	00
FA9	01	00	00	02	00	01	03

Frequência total de cada ponto amostral da amostra original nas nove amostras Bootstrap

FT 06 12 05 09 11 12 08

Gráfico 00.02: Frequência dos pontos amostrais originais nas nove amostras Bootstrap Simples, acima.

Frequências dos pontos amostrais nas 9 amostras Bootstrap Simples



Os números no topo dos retângulos referem-se ao número de vezes que cada ponto amostral da amostra original aparece nas nove amostras Bootstrap Simples acima.

Exemplo: Amostras, Bootstrap Balanceado

Seja $\underline{X} = (94, 19, 16, 38, 99, 14, 23)$, um vetor amostral. Para esse vetor, vamos listar nove amostras, Bootstrap Balanceado e as médias de cada amostra, A_1, A_2, \dots, A_9 , MA_1, MA_2, \dots, MA_9 ; MO é a média da amostra original, FA é a frequência de cada ponto amostral em cada amostra, FT é a frequência total, em todas as amostras de cada ponto amostral, MT é a média total das médias, MA_1, MA_2, \dots, MA_9 e, $i = 1, 2, \dots, 7$, é o índice, do ponto amostral da amostra original.

Amostra original

MO = média da amostra original

i	1	2	3	4	5	6	7	Médias
u	94	19	16	38	99	14	23	MO=43,28

Amostras Bootstrap Balanceado

A1	16	99	38	99	16	38	14	MA1=45,71
A2	23	38	94	14	16	19	23	MA2=32,42
A3	94	23	94	14	14	94	94	MA3=61
A4	99	38	38	99	14	19	16	MA4=46,14
A5	23	38	99	14	94	14	16	MA5=42,57
A6	23	38	14	16	23	14	23	MA6=21,57
A7	99	19	38	94	19	16	94	MA7=54,14
A8	38	16	99	23	94	23	99	MA8=56
A9	19	19	19	99	19	16	19	MA9=30
								MT =43,28

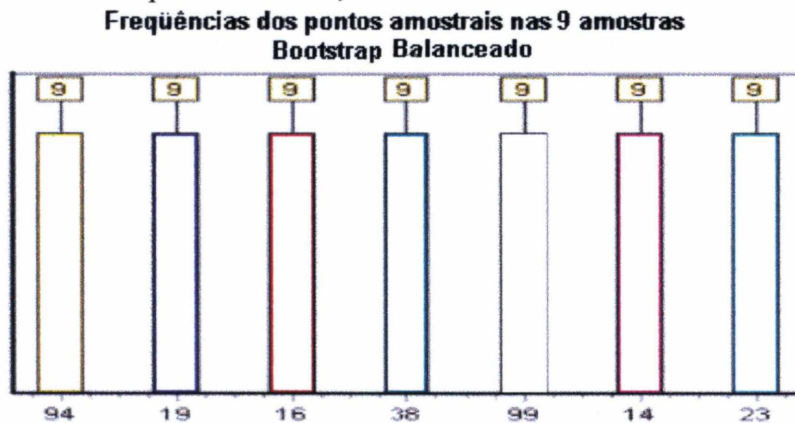
Frequência de cada ponto da amostra original em cada amostra Bootstrap Balanceado

FA1	00	00	02	02	02	01	00
FA2	01	01	01	01	00	01	02
FA3	04	00	00	00	00	02	01
FA4	00	01	01	02	02	01	00
FA5	01	00	01	01	01	02	01
FA6	00	00	01	01	00	02	03
FA7	02	02	01	01	01	00	00
FA8	01	00	01	01	02	00	02
FA9	00	05	01	00	01	00	00

Frequência total de cada ponto da amostra original nas nove amostras Bootstrap Balanceado;

FT	09	09	09	09	09	09	09
----	----	----	----	----	----	----	----

Gráfico 00.03: Frequências dos pontos amostrais originais nas nove amostras Bootstrap Balanceado, acima.



Os números no topo dos retângulos, indicam o número de vezes que o ponto amostral, cujo valor é indicado nas bases dos retângulos, aparece nas nove amostras Bootstrap Balanceado

2.6.4.8. O Bootstrap e o Estimador da Razão

Na amostragem estratificada, de estágio único ou de estágios múltiplos, as amostras são retiradas do seu estrato sem reposição. Diversos procedimentos Bootstrap para avaliar o vício, a estabilidade, a banda de confiança e a faixa de probabilidades para o estimador da razão foram propostos. Listam-se a seguir, as conclusões sobre um estudo comparativo destes procedimentos dados em RAO e KATZOFF (1996), a partir de resultados empíricos.

- 1- Como esperado, o vício e o Erro Quadrático Médio (MSE) de todas os estimadores da variância decrescem com o aumento do tamanho da amostra, n ;
- 2- Todos os procedimentos subestimam o $MSE(\hat{\bar{Y}}_R)$;
- 3- A Linearização e os métodos Bootstrap subestimam o $MSE(\hat{\bar{Y}}_R)$, mais do que o Jackknife;
- 4- Para os procedimentos Bootstrap, a subestimação é relativamente maior para amostragem sem reposição com seu tamanho ótimo, $n(2-f)$;

5- As diferenças na estimação dos erros quadráticos médios são desprezíveis, para os diferentes procedimentos e as variâncias desses estimadores são muito maiores de que a dos seus vícios quadráticos;

6- Para os procedimentos Jackknife e Bootstrap, a faixa das probabilidades e o intervalo de confiança não diferem significativamente, especialmente quando a fração amostral (n/N) excede aos cinco por cento. Nesse caso, as faixas das probabilidades para tais procedimentos estão próximas de noventa por cento para a taxa nominal de noventa e cinco por cento, sendo que, para o estimador clássico, a faixa de probabilidades é de oitenta e cinco por cento .

3. MATERIAL E MÉTODO

3.1. Introdução

Os estimadores descritos no capítulo 2 foram aplicados a dez variáveis, duas das quais pertencentes ao Bancos de Dados fornecidos pelo IPARDES, áreas Urbana e Rural, designadas respectivamente de PU e PR, e a oito das variáveis pertencentes a um Banco de Dados disponível no Laboratório de Estatística da UFPR, áreas Residenciais, Comerciais, Industriais, Públicas, Urbanas e Rurais. Estas últimas foram nomeadas com o prefixo NC e com o pós-fixos AES, ILU, IND, COM, POD, PRO, RES e RUR, respectivamente.

A simulação para o Banco de Dados fornecido pelo IPARDES abrangeu a quantidade de trezentos e setenta e um Municípios. Para o Banco de Dados disponível no Laboratório de Estatística da UFPR, a abrangência variou entre trezentos e seis e trezentos e vinte Municípios, pois foi escolhido para simulação o ano de mil novecentos e noventa.

3.2. Material

Para a simulação, foi construído o programa computacional EEC. Esse programa é composto de todos os estimadores descritos no capítulo 2, e de simulações Jackknife e Bootstrap, Simples e Balanceado, de alguns desses estimadores, nomeadamente dos estimadores da razão, separado(RhS) e não viciado (RhNV) e da Regressão Linear (RL). O programa, ainda, faz aproximações Monte Carlo das simulações Jackknife e Bootstrap. Para a Regressão Linear considerou-se dois casos, a saber:

1. Para agrupamentos com mais de dois grupos, os coeficientes de regressão linear foram determinados tomando-se os valores médios das amostras de

cada grupo. Com os coeficientes de regressão assim determinados, foram feitas as estimativas populacional e de cada localidade separadamente;

2. Para cada grupo ou estrato com mais de duas localidades amostradas, foram determinados os seus coeficientes de regressão linear, e a estimativa da regressão foi feita então para cada localidade do respectivo grupo. A soma das estimativas das localidades estimadas dessa forma aparece como estimativa RLGE, em “Gráficos 2”, como estimativa populacional, também.

3.3. Metodologia

a) Com o objetivo de se verificar que tipo de agrupamento forneceria a menor variância populacional, considerou-se para as variáveis NC_AES, NC_ILU e NC_PRO, do banco de dados disponível no laboratório de Estatística da UFPR, os agrupamentos utilizando apenas a variável de interesse e os agrupamentos envolvendo outras variáveis além daquela de interesse. Quando se faz agrupamentos das localidades, com mais variáveis além da variável de interesse, as variâncias populacionais se poderão tornar muito maiores do que aquela para a variável de interesse, produzindo amostra mínima do tamanho da própria quantidade existente na população, como são os casos para as variáveis NC_AES, NC_ILU e NC_PRO, vistos no **Quadro 00, ANEXO 3**. Assim, para a verificação do comportamento de cada estimador, para variáveis analisadas, todos os resultados aqui descritos foram obtidos fazendo-se agrupamentos unicamente com a variável de interesse.

b) Na busca do melhor método para se estimar a quantidade de equipamentos de uso compartilhado, fez-se simulações para dez das variáveis constantes nos dois conjuntos de dados acima mencionados: aquele fornecido pelo IPARDES e o disponível no laboratório de estatística da UFPR. Variou-se a quantidade de grupos dentro da amplitude delimitada no programa e dentro das limitações impostas pela estrutura dos dados ao método das k-Médias, distância Euclidiana. Consideraram-se proporções de equipamentos compartilhados de 75 e 76%, 75 e 80% e 75 e 90%

entre os grupos ou estratos nos quais a população inteira foi dividida. Nesse procedimento, os grupos ou estratos ímpares receberam a proporção de 75% enquanto que os grupos ou estratos pares receberam 76% em uma rodada, 80% em uma outra e, finalmente, 90% na última. O objetivo era verificar o comportamento de cada estimador em situações onde as variações nas proporções de compartilhados entre as localidades de grupos ou estratos diferentes fossem ‘pequena’ (da ordem de um por cento), ‘média’, (da ordem de cinco por cento) e ‘grande’ (da ordem de quinze por cento). Considerou-se para isso:

- 1 - os erros relativos percentuais em valor absoluto das estimativas de cada estimador, em relação ao valor simulado do número de equipamentos compartilhados na população;
- 2 - o número de localidades estimadas, com erro relativo percentual em valor absoluto, menor ou igual a cinco em relação ao valor simulado para as mesmas;
- 3 - a cobertura do valor simulado do número de equipamentos compartilhados na população, pelos limites (aproximados, em alguns casos) inferior e superior das estimativas de cada estimador.

c) Identificação e descrição do melhor método;

C₁) Como pode ser visto em cada um dos ‘Quadros’ que expõe cada ‘Análise - Pré Amostral’ de algumas das variáveis estudadas, o método de agrupamento das k-médias, em alguns casos, pode fornecer uma quantidade bastante grande de localidades a serem amostradas, em relação à quantidade de localidades existentes na população. Nesses casos, fez-se então estratificações com o objetivo de se obter um número menor de localidades a ser amostrado. São exemplos disso os casos das variáveis, NC_AES, NC_ILU e NC_PRO, para os quais o método de agrupamento das k-Médias agrupa todas as localidades da população em apenas dois grupos, os dois primeiros, e em três grupos, o último deles. Nessa circunstância, as quantidades de localidades a serem amostradas, são, respectivamente: 111, 97 e 153, ou seja, mais de trinta e até cinquenta por cento das localidades da população que, no caso, são 318, 318 e 306, respectivamente (para o ano de 1990, usado para

as simulações). As estratificações, para esses casos, forneceram 58, 65 e 62 localidades a serem amostradas para as respectivas variáveis, baixando o número de localidades a ser amostrado para cerca de vinte por cento do número total de localidades nas respectivas populações, dessas variáveis.

Verificou-se, também, o comportamento desses estimadores quando do aumento do número de grupos e/ou estratos, nos quais a população inteira foi dividida, para se verificar a qualidade dos mesmos em relação à diminuição do número de localidades a ser amostrado.

C₂) Com base nos resultados obtidos das simulações, verifica-se que cada um dos estimadores tem características que o classifica como ‘melhor estimador’, em determinadas situações, para algumas das variáveis estudadas. É exemplo disso o estimador RLGE, a regressão linear dentro do grupo ou estrato que, em geral, não fornece boas estimativas populacionais para a maioria das variáveis estudadas. Isso ocorre principalmente porque as suas estimativas devem ser feitas para agrupamentos ou estratificações com mais do que duas localidades dentro do grupo ou estrato mas, em diversos desses agrupamentos ou estratificações, muitos dos grupos ou estratos possuem menos do que três localidades amostradas. Esse estimador, no entanto, fornece o maior número de localidades estimadas dentro do intervalo de erro relativo percentual, para a variável NC_ILU, que todos os demais estimadores em todos os casos das simulações, estimando, praticamente todas as localidades da população, com erro relativo percentual em valor absoluto menor ou igual a 5, muito embora os erros relativos de suas estimativas populacionais tenham valores maiores. Outros exemplos são os dos estimadores RL, RLBB e RLBS para a variável NC_PRO, para a qual esses estimadores fornecem os menores erros relativos percentuais em valor absoluto das estimativas populacionais, mas as suas estimativas dos valores das localidades dentro do intervalo de erro relativo percentual estão entre as menores quantidades de localidades estimadas dentro do referido intervalo.

C₃) Assim, como critério para a decisão de ‘melhor’ estimador, estabeleceu-se que o estimador deveria fornecer, em primeiro lugar, o menor erro relativo percentual em valor absoluto das estimativas populacionais quando as diferenças de proporções do número de equipamentos compartilhados, nos grupos ou estratos, fossem maiores ou, então mantivessem a tendência de ser o ‘melhor estimador’, quando essa diferença aumenta entre os grupos ou estratos. Em segundo lugar, colocou-se como critério de ‘melhor estimador’ aquele segundo o qual o estimador deveria estimar o maior número de localidades dentro do intervalo de um erro relativo percentual fixado e manter ou acentuar essa tendência quando a diferença de proporção do número de compartilhados aumente de um grupo ou estrato para o outro. Como um terceiro critério de ‘melhor estimador’, considerou-se, ainda, a cobertura do valor populacional simulado do número de equipamentos de uso compartilhado, pelos limites inferior e superior, das estimativas desses estimadores, quando elas são calculadas, ainda que de maneira aproximada, como são os casos dos estimadores da razão, separado (RhS) e não Viciado (RhNV) e as suas simulações Bootstrap e Jackknife, RhSBB, RhSBS, RhSJK, RhNVBB, RhNVBS e RhNVJK. E, isso, para o menor número encontrado de localidades a ser amostrado na população.

C₄) Considerando-se os critérios acima estabelecidos para a identificação de ‘melhor’ estimador, e os resultados das simulações realizadas para as dez variáveis dos dois conjuntos de dados disponíveis (do IPARDES e daquele disponível no laboratório de Estatística da UFPR), verificou-se que o estimador da razão separado, RhS, em oito das dez variáveis estudadas, classificou-se como o ‘melhor estimador’. Nos outros dois casos, onde o estimador da razão separado, RhS, deixou de fornecer os menores erros relativos percentuais em valor absoluto, verifica-se que o motivo para tal reside na diferença entre os arredondamentos parciais e totais, quando os valores parciais das localidades estimadas são muito pequenos e multi-modais, como os das casas das unidades. Para se comprovar esse fato, onde o estimador da razão separado, RhS, não forneceu os menores erros relativos percentuais em valor absoluto, retirou-se os estratos que continham

equipamentos cujos valores eram da ordem das unidades. Para as variáveis NC_ILU e NC_PRO retiraram-se os estratos que continham localidades com apenas um, dois ou três equipamentos. Os resultados dessas restrições, podem ser vistos nos gráficos correspondentes a estas variáveis, com as respectivas restrições: Gráficos 03.10 a 03.18 e Gráficos 06.10 a 06.18, **ANEXO 5**. A partir desses resultados, pode-se concluir que o estimador da razão separado, RhS, pode ser considerado o estimador universal para esse tipo de estimação, se a ressalva de se excluir as localidades com um número de equipamentos da ordem das unidades com modas grandes.

C₅) O estimador da razão separado, RhS;

Quando se faz uma estratificação ou a divisão da população em grupos, uma das maneiras de se usar o estimador da razão na estimação do total populacional é determinar-se as estimativas das razões em cada grupo ou estrato, para com elas calcular as estimativas de cada um desses grupos ou estratos, e daí a estimativa populacional. Assim, dividindo-se a população em k grupos ou estratos, a estimativa separada de cada estrato ou de cada grupo será dada por :

$$\hat{Y}_{Rh} = \frac{\bar{y}_h}{\bar{x}_h} X_h \quad \text{onde cada uma das quantidades, se refere ao estrato ou grupo de ordem } h ; \text{ sendo que:}$$

X_h é o valor da quantidade conhecida, no grupo ou estrato h;

\bar{x}_h é o valor médio amostral da quantidade conhecida, no estrato ou grupo h;

\bar{y}_h é o valor médio amostral da quantidade sob estimação,

$\hat{R}_h = \frac{\bar{y}_h}{\bar{x}_h}$ é o estimador da Razão do estrato ou grupo h;

$\hat{Y}_{RS} = \sum_{h=1}^k \hat{Y}_{Rh}$ é a estimativa populacional da quantidade Y, desconhecida.

RESULTADO 2.3.3.1

Se as grandezas amostrais, n_h , forem grandes em todos os estratos ou grupos, então a variância da estimativa populacional da quantidade sob estimação, Y, obtida pelo estimador da razão separado, é dada por:

$$V(\hat{Y}_{RS}) = \sum_{h=1}^k \frac{N_h(1-f_h)}{n_h} (S_{yh}^2 + R_h^2 S_{xh}^2 - 2R_h \rho_h S_{yh} S_{xh}) \quad , \text{ onde:}$$

$R_h = \frac{Y_h}{X_h}$ é a razão no estrato ou grupo h;

ρ_h é o coeficiente de correlação entre as quantidades amostradas, x_i e y_i , do estrato ou grupo h;

$f_h = \frac{n_h}{N_h}$ é o índice, ou fração amostral, do estrato ou grupo h;

S_{xh} é a variância amostral de x_i , a quantidade conhecida, no grupo ou estrato h;

S_{yh} é a variância amostral de y_i , a quantidade sob estimação, no grupo ou estrato, h;

N_h é o número total de quantidades com valores conhecidos, x_i , no grupo ou estrato, h;

n_h é o número total de quantidades com valores conhecidos, x_i , amostradas, no grupo ou estrato h;

d) Roteiro do procedimento para a obtenção dos resultados:

Os resultados deste trabalho foram obtidos seguindo-se os passos descritos de maneira resumida conforme segue:

1 – Cálculo do tamanho total da amostra n , na população, ou seja, a quantidade de equipamentos a ser amostrado, na população, com o uso da expressão:

$$n \geq \frac{Nz^2\hat{\theta}_0(1-\hat{\theta}_0)}{d^2(N-1) + z^2\hat{\theta}_0(1-\hat{\theta}_0)}$$

N é o número total de equipamentos constantes na população;

Z é o escore padrão, obtido a partir do nível de confiança da solução fixado;

θ_0 é a proporção inicial do número de equipamentos compartilhados na população, fixada em 0.5;

2 – Cálculo do tamanho da amostra, em cada grupo ou estrato h :

Para o número de equipamentos a ser amostrado em cada grupo ou estrato, fez-se uma alocação dos n equipamentos a serem amostrados na população, determinados no item anterior, proporcionais ao tamanho do estrato ou grupo, ou seja, proporcionais à quantidade de equipamentos contidos em cada um desses grupos ou estratos, dados por:

$$n_h = \frac{nN_h}{N}$$

sendo:

n_h o número de equipamentos a ser amostrado no grupo ou estrato h ;

N_h o número de equipamentos do estrato ou grupo h ;

N o número total de equipamentos da população;

n o número total de equipamentos a ser amostrado na população;

3 – Cálculo da quantidade de localidades a ser amostrada, em cada grupo ou estrato h:

A quantidade de localidades a ser amostrada em cada grupo ou estrato, m_h , foi determinada por:

$$m_h \geq \frac{\left(\frac{z\sigma_h}{r\mu_h} \right)^2}{1 + \frac{1}{M_h} \left(\frac{z\sigma_h}{r\mu_h} \right)^2} \quad \text{onde:}$$

r é o erro relativo do número médio da estimativa de equipamentos no grupo ou estrato h, fixado em 0.10;

μ_h é o valor médio da quantidade conhecida x_h , no grupo ou estrato h, ou seja, da quantidade de equipamentos no estrato ou grupo h;

x_h é o valor médio amostral da quantidade conhecida, x_h , no grupo ou estrato h;

d é a precisão do erro relativo do número médio de equipamentos do estrato ou grupo h, fixada em $d = 0.05$;

M_h é o número total de localidades no estrato ou grupo h;

m_h é o número de localidades amostradas no estrato ou grupo h;

σ_h é o desvio padrão da quantidade x_h , conhecida, no estrato ou grupo h, ou seja, a quantidade de equipamentos no estrato ou grupo h;

4 – Cálculo do número de equipamentos a ser amostrado em cada localidade a ser amostrada, no estrato ou grupo h:

Uma vez determinado o número de localidades a ser amostrado, m_h , dentro do grupo ou estrato, h, esse número é então dividido pela soma do número de equipamentos dessas localidades e a quantidade a ser amostrada em cada uma delas é encontrada através de uma distribuição proporcional à quantidade de cada uma delas, ou seja:

$$n_{hi} = N_{hi} \frac{\sum_{i=1}^{m_h} n_{hi}}{\sum_{i=1}^{m_h} N_{hi}}, \text{ onde,}$$

n_{hi} é o número total de equipamentos a ser amostrado na localidade i , do estrato ou grupo h ;

N_{hi} é o número total de equipamentos na localidade amostrada i , do estrato ou grupo h ;

5 – Estimação o número de equipamentos compartilhados:

Para as simulações dentro de cada grupo ou estrato nos quais a população foi dividida, nos estratos de ordem ímpar, simulou-se uma proporção de 75% como sendo a proporção de compartilhados. Para os estratos ou grupos de ordem par, essa proporção foi simulada, em uma primeira rodada, em 76%, na segunda rodada, em 80% e, na terceira rodada, em 90%. A partir dessas simulações, calculou-se o total de compartilhados na população como sendo a soma da quantidade total do estrato ou grupo multiplicada pela respectiva proporção simulada do grupo ou estrato, arredondado para o número inteiro mais próximo. Para cada localidade, foi calculada também a quantidade de compartilhados, como o produto da quantidade de equipamentos da localidade pela proporção simulada do grupo ou estrato do qual a localidade faz parte, arredondada para o inteiro mais próximo.

6 – Cálculo das estimativas populacionais, a partir dos dados amostrais simulados:

Com as quantidades de equipamentos em cada localidade amostrada, rodou-se o programa três vezes, simulando-se proporções diferentes para a quantidade de equipamentos compartilhados. Com uma mesma amostra, na primeira rodada

simulou-se uma proporção de 75% de compartilhados para os estratos ou grupos ímpares e de 76% para os pares. Numa segunda rodada, simulou-se para a referida amostra 80% para os estratos ou grupos pares, enquanto os ímpares continuaram com 75%. Numa terceira rodada, a mesma amostra recebeu uma proporção de 90% de compartilhados nos estratos ou grupos pares. Com essas proporções dentro de cada estrato ou grupo, determinou-se o valor populacional simulado para o número total de compartilhados na população. A partir das proporções fixadas dentro de cada estrato ou grupo para as localidades neles amostradas, procedeu-se então às estimações de compartilhados a partir desses dados, com cada um dos estimadores descritos na Revisão Bibliográfica. Quais sejam:

- a- Estimador da razão separado, RhS;
- b- Estimador da razão não viciado, RhNV;
- c- Estimador da razão combinado, RhCB;
- d- Estimador da expansão simples, EXPS;
- e- Estimador da regressão linear dentro dos grupos ou estratos, RLGE;
- f- Estimador da razão separado, Bootstrap Balanceado, RhSBB;
- g- Estimador da razão separado, Bootstrap Simples, RhSBS;
- h- Estimador da razão separado, Jackknife, RhSJK;
- i- Estimador da razão não viciado, Bootstrap Balanceado, RhNVBB;
- j- Estimador da razão não viciado, Bootstrap Simples, RhNVBS;
- k- Estimador da razão não viciado, Jackknife, RhNVJK;
- l- Estimador da regressão linear, RL. Para esse estimador, tomou-se as médias das quantidades amostradas e as médias simuladas de compartilhados, em cada grupo ou estrato;
- m- Estimador da regressão linear, Bootstrap Balanceado, RLBB;
- n- Estimador da regressão linear, Bootstrap Simples, RLBS;
- o- Estimador da regressão linear, Jackknife, RLJK;
- p- Estimador da razão separado, Bootstrap Balanceado, RhSBBMC;
- q- Estimador da razão separado, Bootstrap Simples, RhSBSMC;
- r- Estimador da razão separado, Jackknife, RhSJKMC;
- s- Estimador da razão não viciado, Bootstrap Balanceado, RhNVBBMC;

- t- Estimador da razão não viciado, Bootstrap Simples, RhNVBSMC;
- u- Estimador da razão não viciado, Jackknife, RhNVJKMC;
- v- Estimador da regressão linear, Bootstrap Balanceado, RLBBMC;
- w- Estimador da regressão linear, Bootstrap Simples, RLBSMC;
- x- Estimador da regressão linear, Jackknife, RLJKMC;

As estimativas dos itens o até w são as simulações Monte Carlo para esses respectivos estimadores, com exceção, é claro, do estimador da expansão simples, EXPS, do estimador da razão combinado, RhCB, e do estimador da regressão linear, RLGE, dentro dos grupos ou estratos.

7 – Cálculo dos erros relativos percentuais em valor absoluto, das estimativas populacionais:

Com a estimativa populacional, determinou-se de cada estimador o erro relativo percentual em valor absoluto entre essas estimativas e o valor simulado populacional, para o número de compartilhados. Isso para, a partir desses valores, determinar-se o estimador que forneceria o menor erro em todas ou na maioria das simulações, da estimativa populacional.

8 – Cálculo do número de localidades estimadas com erro relativo percentual em valor absoluto menor ou igual a 5:

Com o valor simulado para cada localidade a partir da proporção simulada de compartilhados para o estrato ou grupo na qual ela se insere, e com o valor estimado para a mesma por cada um dos estimadores, calculou-se o erro relativo percentual em valor absoluto e, então, contou-se o número dessas localidades cujas estimativas tinham erro relativo percentual em valor absoluto menor ou igual a cinco. O objetivo era verificar qual estimador se comportava melhor estimando as localidades individualmente.

9 – Cálculo do número de vezes que o valor simulado populacional é coberto pelo intervalo dado pelos extremos inferior e superior da estimativa de cada estimador:

Foram calculados os extremos inferior e superior das estimativas do valor populacional e verificado se o intervalo dado por esses extremos cobriam o valor populacional simulado.

10 – Simulação Monte Carlo para a avaliação de cada estimador:

Foram realizadas simulações Monte Carlo para alguns estimadores para, com elas, verificar-se a eficiência de cada estimador.

3.4. Programa Computacional

3.4.1. Introdução

O programa computacional EEC foi elaborado visando ao seguinte:

1-Fazer simulações para se identificar o possível melhor estimador, para os dados sob consideração;

2-Agrupar as localidades constantes da população, em grupos fixados pelo usuário dentro dos limites de um até dezessete grupos. Observa-se que, sendo o agrupamento efetuado pelo método das k-Médias, pode ocorrer que o número de grupos fixado não seja atingido, pois isso é dependente da estrutura do Banco de Dados sob análise, imposta ao método de agrupamento das k-Médias, como pode

ser visto nas ‘**Análises Pré Amostrais**’, ANEXO 3, para algumas das variáveis analisadas, onde os agrupamentos têm sempre menos do que dezessete grupos. Os resultados dos agrupamentos podem ser visualizados em ‘**Grupo/Estrato**’, **Figura 6**, ANEXO 2. Aí, podem ser vistas as quantidades mínimas de equipamentos a ser amostrada na população, a quantidade de localidades a ser amostrada na mesma, a que grupo pertence cada localidade, a quantidade de equipamentos de cada uma delas e quantos, no mínimo, de seus equipamentos devem ser amostrados para que a estimativa atinja o nível de confiança desejado.

3-Estratificar as localidades constantes da população, em uma amplitude de um a dezessete estratos. Após fixar o número de estratos e clicar o botão ‘**OK**’, ver **Figura 5**, ANEXO 2, onde aparecerão as variáveis para preenchimento dos limites inferior e superior de cada estrato. Esses limites podem ser retirados da tabela visualmente, bastando para isso utilizar o comando ‘**Ordenar**’, e os valores da variável são ordenados crescentemente. Preenchidos os valores da variável com os limites dos estratos, clicando-se o botão ‘**Estratificar**’, ver **Figura 5**, ANEXO 2, é feita a estratificação e as localidades pertencentes a cada estrato, as quantidades de equipamentos de cada uma delas, aquelas que devem ser amostradas e o número mínimo de seus equipamentos a serem amostrados aparecerão em ‘**Grupos/Estratos**’, rolando a barra de rolagem para baixo, ver **Figura 5**, ANEXO 2;

4-Fazer amostragem Aleatória Simples ou Sistemática (aproximada, em geral) dentro de cada grupo ou estrato, selecionado em ‘**Grupos/Estratos**’, ‘**Amostragem Proporcional ao Tamanho**’, ver **Figura 5**, ANEXO 2;

5-Fazer estimativas do total populacional, com cada um dos estimadores descritos no capítulo 2. Os resultados dessas estimativas são mostrados em ‘**Resultados das Estimativas Clássicas, Jackknife e Bootstrap e Simulações Monte Carlo**’, ver **Figura 11 e 12** e, também, em ‘**Gráficos 2**’, ver **Figura 17**, em ‘**Resultados das Estimativas Populacionais**’:

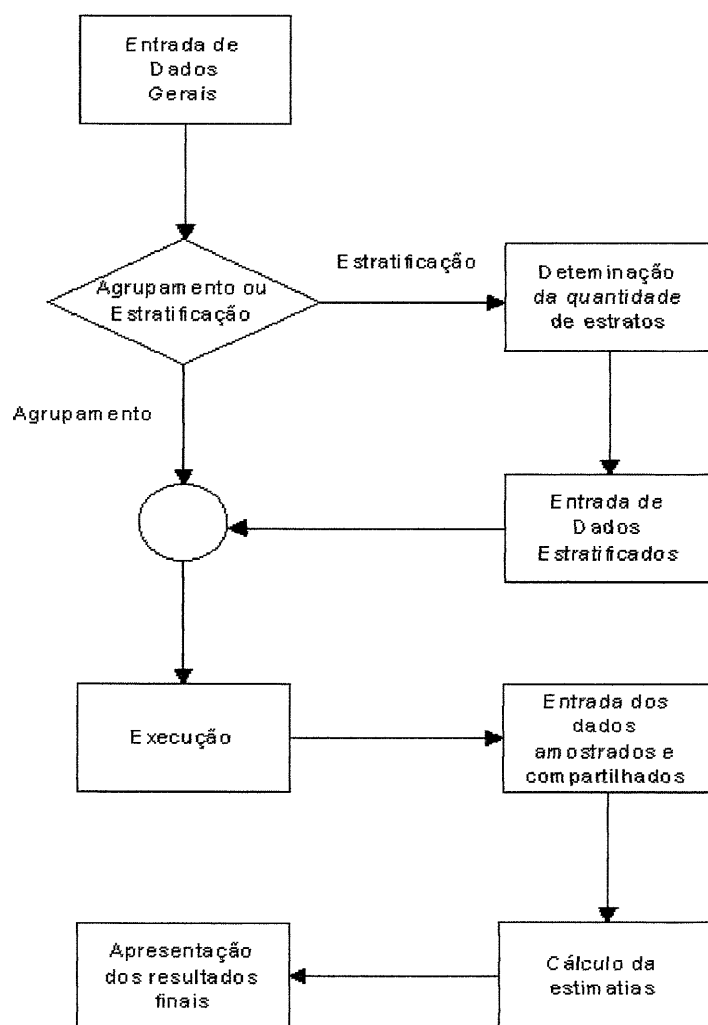
6-Fazer estimativas de cada uma das localidades constantes da população, com todos os estimadores, e indicar quais dos estimadores foram os que fizeram estimativas entre os limites de erro relativo percentual fixado pelo nível de

confiança estabelecido e, também, indicar qual foi o menor erro relativo percentual para cada localidade estimada, que pode ser visto em **‘Estimação’**, **Figura 14, ANEXO 2**. Em **‘Gráficos 1’**, é dada a quantidade de localidades estimadas dentro do erro relativo percentual absoluto, por cada estimador, dentro de cada grupo e/ou estrato e também em toda a população, ver **Figuras 16 e 17, ANEXO 2**;

7-Realizar simulações Bootstrap Simples e Balanceado, e Jackknife das estimativas dos estimadores da razão, separado (RhS) e não viciado (RhNV) e da Regressão Linear (RL). Isso é feito automaticamente. Os resultados podem ser vistos em **‘Resultados das Estimativas Clássicas, Jackknife Bootstrap e Simulações Monte Carlo’**, **Figura 12 e 13**;

8-Realizar aproximações Monte Carlo das simulações Bootstrap e Jackknife. Essas aproximações são realizadas a comando do usuário, após a realização das estimativas, isto é, após a inserção dos dados para a simulação em **‘Dados pós-amostragem’**, onde se pode entrar com as percentagens para cada grupo ou estrato aí contido, após a estratificação ou agrupamento e, então, no Campo ao lado, chamado de **‘Simulação’**, clicar em **‘Simular Compartilhados’** e depois o botão **‘Calcular’**. Após isso, as simulações ou aproximações Monte Carlo podem ser feitas, no Campo **‘Aproximações Monte Carlo’**, onde o usuário poderá determinar o número de replicações desejado, variando entre cinquenta e quinhentas replicações. Tais replicações podem ser feitas uma, duas, ou mais de cada vez, sendo que a cada valor fixado de replicações, o tempo aproximado para a execução da simulação aparecerá indicado como **‘Tempo Estimado’**. Os resultados dessas aproximações Monte Carlo podem ser vistas na parte final de **‘Resultados das Estimativas, Clássicas, Jackknife e Bootstrap’**, na página **‘Simulação Bootstrap’**, rolando-se a barra de rolagem. Ali, aparecerão os resultados dessas aproximações, precedidas pela indicação do número de replicações solicitadas. Alguns desses resultados das aproximações Monte Carlo podem ser vistas também em **‘Gráficos2’**, **Figura 18**. Nesse caso, só o resultado da última iteração é que aparece em **‘Gráficos 2’**;

FLUXOGRAMA GERAL

**Figura 05:** Fluxograma Geral

4. RESULTADOS E DISCUSSÃO

4.1. Discussão dos resultados das simulações, para cada variável analisada:

Para cada uma das variáveis analisadas, consideraram-se proporções de compartilhados, de 75% nos grupos ou estratos ímpares e 76%, 80% ou 90% nos pares, em rodadas diferentes, com a mesma amostra e o mesmo número de rodadas para cada uma das diferentes proporções de compartilhados com amostras distintas. De todas as rodadas, para cada grupo de proporções diferentes, 75 e 76%, 75 e 80%, 75 e 90%, determinou-se, média, erro padrão e amplitude de variação para os erros relativos percentuais em valores absolutos das estimativas populacionais e para o número de localidades estimadas. Para o número de vezes que as estimativas do estimador cobre o valor simulado da estimativa populacional, calculou-se o número de vezes de cobertura e as amplitudes de variação para cada uma dessas diferenças de proporção.

Considerou-se também o conjunto de todas as rodadas, com todas as diferenças de proporção acima mencionadas, e determinou-se, então, as médias, os erros padrão e as amplitudes de variação para as coberturas, para os erros relativos percentuais em valor absoluto e para o número de localidades estimadas.

Os resultados dessas simulações para cada uma das dez variáveis estudadas foram tabelados e então transformados em gráficos, que podem ser vistos nos anexos.

Além disso, fez-se também uma verificação do comportamento de cada estimador, para todos os agrupamentos permitidos pela estrutura dos dados em consideração e, em alguns casos, diversas estratificações, principalmente quando o número de grupos era pequeno.

4.1.1. Discussão dos resultados das simulações para a variável NC_AES:

Para a variável NC_AES, realizaram-se 25 rodadas de simulações, para as diferentes proporções de compartilhados entre os seis estratos nos quais a população inteira foi dividida com o menor número de localidades a ser amostrado na população, que no caso foi de sessenta e dois. Como resultado de todas as simulações, nas 25 rodadas e no conjunto de todas elas, ou seja, nas 75 rodadas, verificou-se que:

1 – Os erros relativos percentuais em valores absoluto, das estimativas populacionais, o estimador da razão separado, RhS, tem os seus valores decrescentes, com o aumento das diferenças de proporções de compartilhados, entre os estratos, atingindo o valor de cerca de um, para a maior diferença de proporção.

2 – Consideradas todas as 75 rodadas de simulações, com todas as suas diferenças nas proporções, o estimador da razão separado, RhS, juntamente com seus similares Bootstrap balanceado e simples e Jackknife, mantêm as menores médias, erro padrão e amplitude de variação de erro relativo percentual em valores absolutos. Os estimadores da razão acima mencionados são seguidos de perto em seus resultados pelos estimadores da regressão linear, RL, e de seus similares, Bootstrap balanceado RLBB, Bootstrap simples RLBS e Jackknife RLJK, o mesmo acontecendo com a simulação Monte Carlo RLJKMC, da regressão linear Jackknife.

As diferenças entre as estimativas do estimadores da razão separado, RhS e as da regressão linear, RL, estão no fato que, nas estimativas destas últimas, os erros relativos percentuais crescem com o aumento na diferença de proporção entre os compartilhados, enquanto que em relação ao estimador da razão acontece o contrário.

3- Quanto ao número de localidades estimadas, o maior número delas fica por conta da regressão linear dentro do estrato RLGE, que tem o maior número de localidades estimadas, aumentando esse número com o aumento na diferença de proporção de

compartilhados, com menor erro padrão e amplitude de variação dessas quantidades. Gráficos 01.04 a 01.06, Tabelas 01.04 a 01.06

4 – Com relação à cobertura do valor simulado da estimativa populacional, os maiores número de vezes estão a cargo dos estimadores da razão separado, RhS e não viciado, RhNV, e de seus similares Jackknife, RhSJK e RhNVJK. As estimativas desses estimadores cobrem o valor simulado das estimativas populacional em praticamente todas as simulações, com erro padrão e amplitude de variação desse referido número, nulos ou quase nulos.

Em resumo, o estimador da razão separado, RhS, é o melhor estimador das estimativas populacionais para a variável NC_AES, pois as suas estimativas têm o menor erro relativo percentual em valor absoluto quando comparadas às estimativas de todos os demais estimadores, porque suas estimativas cobrem o valor simulado em praticamente todas as simulações com erro padrão e amplitude de variação da ordem de zero. O estimador da razão separado, RhS, estima também um número significativamente grande de localidades, com erro padrão e amplitude de variação desse número um pouco menor, apenas, que o do estimador RLGE.

4.1.2. Discussão dos resultados das simulações para a variável NC_COM:

Para a variável NC_COM, foram realizadas 25 rodadas de simulações para as diferentes proporções de compartilhados entre os seis estratos nos quais a população inteira foi dividida com o menor número de localidades a ser amostrado na população entre os treze grupos, nos quais a população inteira foi dividida, Esse número foi de cinquenta e cinco. Como resultado de todas as simulações, nas 25 rodadas e no conjunto de todas elas, ou seja, nas 75 rodadas, verificou-se que:

1 – Os erros relativos percentuais em valor absoluto das estimativas dos estimadores da razão, separado RhS e não viciado, RhNV, têm valores próximos de zero. Os similares Bootstrap do estimador da razão separado, RhSBB e RhSBS, e Jackknife, RhSJK, têm os valores de erro relativo percentual da ordem de meio, diminuindo este valor com o aumento da diferença de compartilhados. As

estimativas dos similares Bootstrap e Jackknife, do estimador da razão não viciado, RhNVBB, RhNVBS e RhNVJK, respectivamente, têm erros da ordem de um que diminuem para cerca de meio quando as diferenças de proporção de compartilhados atinge o maior valor simulado.

Os erros padrão das estimativas dos estimadores da razão, separado e não viciado, bem como as amplitudes de variação dos erros relativos percentuais desses estimadores estão próximos de zero. Essas quantidades também estão próximas de zero para os similares Bootstrap e Jackknife, dos estimadores da razão, porém um pouco maiores do que as dos seus representantes clássicos.

2 – O maior número de localidades estimadas pertence ao estimador da razão separado, RhS, o qual estima todas as localidades, em todos os casos de diferença de proporção entre o número de compartilhados. Ao estimador da razão separado segue-se o estimador da razão não viciado, RhNV, que estima quase todas as localidades, praticamente não variando a quantidade das mesmas quando da variação das diferenças de proporção entre o número de compartilhados. Quanto ao número de localidades estimadas, erro padrão e amplitude de variação do número das mesmas, esses estimadores da razão são seguidos pelo estimador da regressão, RLGE, que também tem ótimos resultados.

3 – Quanto ao número de vezes que o valor simulado da estimativa populacional é coberto, os maiores números de vezes estão por conta dos estimadores da razão separado e não viciado, RhS e RhNV, e de seus similares Jackknife, RhSJK e RhNVJK, respectivamente.

Em resumo, podemos dizer que ao estimador da razão restam algumas pequenas vantagens em relação aos demais que o seguem, especialmente àquele que o segue mais de perto, o não viciado RhNV. Assim, o estimador da razão separado, RhS, por uma leve diferença, caracteriza-se como melhor estimador, para a variável NC_COM.

4.1.3. Discussão dos resultados das simulações para a variável NC_ILU:

Para a variável NC_ILU, realizaram-se 25 rodadas de simulações, para as diferenças de proporção entre os cinco estratos, nos quais a população inteira foi dividida, com o menor número de localidades a ser amostrado na população que, no caso foi de sessenta e nove. Como resultado de todas as simulações, nas 25 rodadas e no conjunto de todas elas, ou seja, nas 75 rodadas, verificou-se que:

1 – Os menores erros relativos percentuais em valor absoluto estão por conta das estimativas Bootstrap e Jackknife, RhSBB, RhSBS e RhSJK, do estimador da razão separado, RhS. As estimativas do erro relativo percentual desses estimadores têm os menores erros, as menores amplitudes de variação e os menores erros padrão, quando comparadas a todas as estimativas de todos os demais estimadores. Suas estimativas se mantêm quase que inalteradas com o aumento das diferenças de proporção do número de compartilhados. Os erros padrão do erro relativo percentual das estimativas populacionais desses estimadores, ficam em torno de um, e seus erros padrão também ficam em torno de um, enquanto que as suas amplitudes de variação estão em torno de seis.

2 – Quanto ao número de localidades estimadas, o maior número delas fica por conta da regressão linear RLGE, que estima praticamente todas as localidades, independentemente de as diferenças entre o número de compartilhados ser menor ou maior. O número de localidades estimadas pelo estimador da regressão linear, RLGE, tem o menor erro padrão e a menor amplitude de variação dentre todos os demais estimadores.

Ao número de localidades estimadas pelo estimador RLGE, seguem as dos estimadores da razão separado, Bootstrap e Jackknife, RhSBB, RhSBS e RhSJK respectivamente, os quais têm os números de localidades estimadas comparáveis aos do estimador RLGE, quando a diferença do número de compartilhados é da ordem de um ou de cinco por cento, diminuindo quando tal diferença atinge os quinze por cento. Mesmo assim, os estimadores RhSBB, RhSBS e RhSJK têm os maiores números de localidades estimadas que os demais estimadores, que não o RLGE.

3 – Quanto à cobertura do valor populacional simulado, os maiores números de vezes pertencem aos estimadores da razão clássicos, separado e não viciado, RhS e RhNV, respectivamente, juntamente com os seus representantes Jackknife, RhSJK e RhNVJK. Todos eles cobrem o valor simulado em praticamente todas as situações, não importando se as diferenças entre o número de compartilhados é pequena ou grande.

Em resumo, o estimador da razão separado Jackknife, RhSJK, é o estimador que tem o menor erro relativo percentual em suas estimativas do valor simulado, populacional, menor amplitude de variação desses erros e menor erro padrão dos mesmos. Com relação ao número de localidades estimadas, é o estimador que estima o maior número delas, depois do estimador RLGE. Além disso, suas estimativas populacionais cobrem o valor simulado o maior número de vezes, com amplitude de variação e erro padrão desse número de vezes praticamente nulos.

Assim, o estimador da razão separado Jackknife poderia, ser considerado o melhor estimador para a variável NC_ILU. Observa-se no entanto, que o estimador clássico da razão separado, RhS, deixa de ser o melhor estimador, de um lado por causa dos erros de arredondamentos parciais, feitos sobre as proporções do número de compartilhados de cada localidade. Por serem pequenas as quantidades de equipamentos dessas localidades, os arredondamentos das quantidades de compartilhados, destes valores, causam erros relativos percentuais elevados. Fez-se outras estimativas para esta variável, eliminando-se o estrato que continha as menores quantidades de equipamentos e o estrato contendo as localidades que continham de um a três equipamentos, e o resultado mostra que o estimador da razão separado, clássico, RhS, mantém a sua liderança como ‘melhor estimador’.

Por todas as considerações acima, o estimador da razão separado clássico, RhS, deve ser considerado o melhor estimador para esta variável também, juntamente com o estimador da razão separado Jackknife, RhSJK.

4.1.4. Discussão dos resultados das simulações para a variável NC_IND:

Para a variável NC_IND, realizaram-se 20 rodadas de simulações para as diferentes proporções de compartilhados entre os catorze estratos, nos quais a população inteira foi dividida, com o menor número de localidades a ser amostrado na população que, no caso, foi de sessenta e sete. Como resultado de todas as simulações, nas 20 rodadas e no conjunto de todas elas, ou seja, nas 60 rodadas, verificou-se que:

1 – Os estimadores da razão separado, RhS, e não viciado, RhNV, apresentam erros relativos percentuais das estimativas populacional em relação ao valor simulado próximos de zero, o mesmo acontecendo com o erro padrão e a amplitude de variação dos erros relativos percentuais, com uma leve vantagem para o estimador separado. Esses valores de erro relativo percentual se mantêm para todas as diferenças de proporção de compartilhados.

2 – Quanto ao número de localidades estimadas, o maior número delas é do estimador da razão separado; seus erros padrão e a amplitude de variação desses números são menores do que as de todos os outros estimadores, inclusive os do estimador não viciado, que o segue em segundo lugar. Em terceiro lugar, tem-se o estimador da regressão linear, RLGE, com um número de localidades um pouco menor, porém com um erro padrão e amplitude de variação menores que os do estimador não viciado, RhNV, porém um pouco maiores do que as do estimador separado, RhS.

3 – O número de vezes que o valor simulado populacional é coberto pelos intervalos das estimativas de cada estimador encontra os seus maiores valores para os estimadores da razão separado e não viciado, juntamente com os seus similares Jackknife, RhSJK e RhNVJK. As estimativas desses estimadores cobrem o valor populacional simulado em praticamente cem por cento das vezes. Seus erros padrão e amplitudes de variação dessa quantidade estão bem próximo a zero.

Em resumo, o estimador da razão separado, RhS, deve ser considerado o melhor estimador para a variável NC_IND.

4.1.5. Discussão dos resultados das simulações para a variável NC_POD;

Para a variável NC_POD, realizaram-se 21 rodadas de simulações para as diferentes proporções de compartilhados entre os oito grupos, nos quais a população inteira foi dividida, com o menor número de localidades a ser amostrado na população, que no caso foi de quarenta e três. Como resultado de todas as simulações, nas 21 rodadas e no conjunto de todas elas, ou seja, nas 63 rodadas, verificou-se que:

1 – Os menores erros relativos percentuais em valor absoluto, pertencem aos estimadores da razão separado, RhS, não viciado, RhNV, e combinado, RhCB, com uma leve vantagem para o estimador separado, RhS. Os erros padrão e amplitude de variação dos erros relativos percentuais desses estimadores são bastante pequenos, ainda com leve vantagem para o estimador separado. Os estimadores Bootstrap e Jackknife, dos estimadores da razão separado RhSBB, RhSBS e RhSJK, assim como os do não viciado, RhNVBB, RhNVBS e RhNVJK, também têm, aproximadamente, as mesmas características de seus similares clássicos, acima mencionadas.

2 – Quanto ao número de localidades estimadas, o maior número delas é a do estimador da razão separado, RhS, seguido pelo estimador da razão não viciado, RhNV e da regressão linear, RLGE, que apresentam valores muito próximos os do estimadores da razão separado e não viciado. O erro padrão do número de localidades estimadas do estimador da regressão linear, RLGE, é da mesma ordem que a do estimador da razão separado, RhS, ambas menores do que aquela do estimador não viciado, RhNV. O mesmo acontece com as amplitudes de variação do número de localidades estimadas por esses estimadores.

3 – Sobre a cobertura do valor populacional simulado, as maiores quantidades de vezes são dos estimadores da razão separado, RhS, e não viciado, RhNV, juntamente com seus similares Jackknife, RhSJK e RhNVJK, com os mesmos números de quantidades que as dos seus representantes clássicos. Todos esses

estimadores apresentam amplitude de variação e erro padrão, da quantidade de vezes, da ordem de zero.

Em conclusão, podemos afirmar que, por todas essas considerações feitas sobre os resultados das simulações, o estimador da razão separado RhS, por suas características e por suas, ainda que pequenas, diferenças em relação àqueles estimadores que o seguem, ser ele o melhor estimador para a variável NC_POD.

4.1.6. Discussão dos resultados das simulações para a variável NC_PRO;

Para a variável NC_PRO, realizaram-se 25 rodadas de simulações, para as diferentes proporções de compartilhados entre os oito grupos, nos quais a população inteira foi dividida, com o menor número de localidades a ser amostrado na população que, no caso, foi de cinquenta e seis. Como resultado de todas as simulações, nas 25 rodadas e no conjunto de todas elas, ou seja, nas 75 rodadas, verificou-se que:

- 1 – Quanto aos erros relativos percentuais em valor absoluto das estimativas populacionais os estimadores da regressão linear, RL, e seus representantes Bootstrap e Jackknife, RLBB, RLBS e RLJK, juntamente com as simulações Monte Carlo RLJKMC, do estimador da regressão linear Jackknife, RLJK, e do estimador da razão separado Bootstrap balanceado, RhSBB, têm os menores erros. O menor erro padrão e menor amplitude de variação dos erros relativos percentuais são do estimador da razão separado Bootstrap balanceado, RhSBB.
- 2 – Quanto ao número de localidades estimadas, o maior número pertence ao estimador da razão separado e ao estimador da regressão linear, RLGE. Ambos têm essas quantidades variando muito pouco com a variação nas diferenças de compartilhados. Ambos os estimadores têm erro padrão e amplitudes de variação do número de localidades estimadas, pequenos, com uma leve vantagem para o estimador da razão separado, RhS.
- 3 – Quanto à cobertura do valor populacional simulado, o estimador da razão separado, RhS, e o estimador da razão não viciado, RhNV, juntamente com os seus similares Jackknife, RhSJK e RhNVJK, possuem os maiores números de vezes. As

estimativas desses estimadores cobrem o valor simulado em, praticamente, cem por cento das vezes. Seus erros padrão e amplitudes de variação são praticamente nulos.

Em resumo, para essa variável o quadro parece um pouco confuso quanto à decisão sobre qual seria o melhor estimador, quando consideradas separadamente as características desses estimadores. Contudo, ao considerar-se em conjunto, tais características parecem indicar o estimador da razão separado Jackknife, RhSJK, como o melhor estimador para a variável NC_PRO. Pois, por exemplo, as estimativas do estimador da razão separado Bootstrap têm erros relativos percentuais um pouco menores do que os correspondentes, do estimador da razão separado Jackknife. Quanto às outras características, ambos têm diferenças pequenas, porém quanto às coberturas do valor populacional simulado, o estimador da razão separado Jackknife cobre cem por cento das vezes esse valor, enquanto que o correspondente estimador Bootstrap tem uma grande amplitude de variação no número de vezes que cobre o valor simulado.

4.1.7. Discussão dos resultados das simulações para a variável NC_RES;

Para a variável NC_RES, foram realizadas 25 rodadas de simulações para as diferentes proporções de compartilhados entre os catorze grupos, nos quais a população inteira foi dividida, com o menor número de localidades a ser amostrado na população que, no caso, foi de sessenta. Como resultado de todas as simulações, nas 25 rodadas e no conjunto de todas elas, ou seja, nas 75 rodadas, verificou-se que:

- 1 – Os estimadores da razão, separado RhS e não viciado RhNV, tem os menores erros relativos percentuais de suas estimativas populacionais em relação ao valor populacional simulado. Os erros relativos percentuais, seus erros padrão e amplitudes de variação são todos pequenos, próximos de zero.
- 2 – As maiores quantidades de localidades estimadas também são dos estimadores da razão separado, RhS, e não viciado, RhNV. Seus erros padrão e amplitude de variação do número de localidades estimadas têm valores próximos de zero, com

pequena vantagem para o estimador da razão separado. Observa-se nesse caso, que o estimador da regressão linear, RLGE, não estima sequer uma localidade, com erro relativo percentual em valor absoluto menor ou igual a 5, apesar de o estimador fazer estimativas em doze dos catorze grupos. O estimador da regressão linear, RLGE, não só não estima bem as localidades como também tem enormes erros relativos percentuais (maiores do que noventa por cento) nas suas estimativas populacionais. Tais resultados do estimador RLGE o caracterizam como um péssimo estimador para a variável NC_RES.

3 – Quanto à cobertura do valor populacional simulado, os estimadores da razão separado e não viciado, juntamente com seus similares Bootstrap e Jackknife, cobrem o valor simulado cem por cento das vezes, com exceção do estimador da razão Bootstrap balanceado que, nas setenta e cinco simulações, cobriu setenta e quatro vezes o valor simulado populacional.

Em resumo, das diversas considerações acima, dois estimadores podem ser considerados melhor estimador para a variável NC_RES, mas, ainda neste caso, uma diferença bastante pequena no número de localidades estimadas caracteriza o estimador da razão separado, RhS, como o melhor estimador para essa variável.

4.1.8. Discussão dos resultados das simulações para a variável NC_RUR;

Para a variável NC_RUR, realizaram-se 25 rodadas de simulações para as diferentes proporções de compartilhados entre os catorze grupos, nos quais a população inteira foi dividida, com o menor número de localidades a ser amostrado na população que, no caso, foi de sessenta. Como resultado de todas as simulações, nas 25 rodadas e no conjunto de todas elas, ou seja, nas 75 rodadas, verificou-se que:

1 – Os estimadores da razão separado, RhS, e não viciado, RhNV, tem os menores erros relativos nas estimativas do valor populacional simulado. Independentemente de serem ou não pequenas as diferenças entre o número de compartilhados, esses erros se mantêm próximos de zero, bem como suas amplitudes e erros padrão. Em todas as situações há uma pequena vantagem para o estimador da razão separado.

Observa-se que na maioria dos estimadores e na maior parte das simulações há erros relativos percentuais em valores absolutos menores do que um e que o estimador da razão combinado, RhCB, tem esse erro, em média, menor do que 0.2, e erro padrão e amplitude menores do que um.

2 – Quanto ao número de localidades estimadas, com erro relativo percentual em valor absoluto, menor ou igual a 5, os estimadores da razão separado, RhS, e não viciado, RhNV, estimam quase que cem por cento das localidades na maioria das simulações, com erros padrão e amplitudes de variação em torno de um. Observa-se, aqui, algo interessante com relação ao estimador da regressão linear, RLGE, esse estimador estima em todas as situações, independentemente de as diferenças no número de compartilhados ser pequena ou grande, um número constante de localidades, quarenta e nove. Ao se observar as quantidades de localidades dos grupos com mais de duas localidades a serem amostradas, verifica-se que os cinco grupos com mais do que duas localidades amostradas possuem, em conjunto, setenta e seis localidades. Delas, somente quarenta e nove são estimadas, com erro relativo percentual em valor absoluto menor ou igual a cinco, pelo estimador da regressão linear, RLGE.

3 – Com relação à cobertura do valor populacional simulado, os estimadores da razão, separado RhS, e não viciado RhNV, bem como os seus similares Jackknife, cobrem o valor simulado em cem por cento das vezes.

Em resumo, os estimadores da razão separado, RhS, e não viciado, RhNV, são ótimos estimadores para a variável NC_RUR, não só das estimativas populacionais, como também das estimativas das localidades. Esse saldo é favorável no que diz respeito ao erro relativo percentual em valor absoluto, para o estimador da razão separado, pois coloca-o como o melhor estimador, embora a diferença, na realidade, seja bastante pequena, da ordem da segunda ou terceira casa decimal.

4.1.9. Discussão dos resultados das simulações para a variável PR:

Para a variável PR, foram realizadas 25 rodadas de simulações para as diferentes proporções de compartilhados entre os catorze grupos, nos quais a população inteira foi dividida, com o menor número de localidades a ser amostrado na população que, no caso, foi de sessenta. Como resultado de todas as simulações, nas 25 rodadas e no conjunto de todas elas, ou seja, nas 75 rodadas, verificou-se que:

1 - Os menores erros relativos percentuais em valores absolutos pertencem aos estimadores da razão separado, RhS , e não viciado, $RhNV$, sendo praticamente nulos e pouco ou quase nada variando com as diferenças nas proporções de compartilhados. As estimativas Bootstrap e Jackknife, desses estimadores, $RhSBB$, $RhSBS$, $RhSJK$, $RhNVBB$, $RhNVBS$ e $RhNVJK$, respectivamente, têm erros maiores que os seus representantes clássicos, porém, em média menores do que 0.4. O estimador da razão combinado, assim como os demais estimadores já mencionados, também têm erros relativos percentuais pequenos, em média menores do que 0.4. Com relação aos estimadores da regressão linear, RL , e seus similares Bootstrap e Jackknife, seus erros relativos percentuais em valores absoluto, menores do que um. As melhores estimativas, porém, em todos os aspectos, no que diz respeito ao valor do erro relativo percentual, seus erros padrão e suas amplitudes de variação, ainda são dos estimadores clássicos, da razão separado, RhS e, não viciado, $RhNV$.

2-Quanto ao número de localidades estimadas, os estimadores da razão separado e não viciado, têm a totalidade deles, independentemente se as diferenças nas proporções de compartilhados sejam grandes ou pequenas. Nota-se aqui que o estimador da regressão linear, $RLGE$, não estima sequer uma localidade, com erro relativo percentual em valor absoluto menor ou igual a 5, muito embora faça estimativas em sete dos treze grupos nos quais a população foi dividida, caracterizando-se dessa forma, como um péssimo estimador para as localidades.

3-O valor simulado populacional do número de compartilhados é coberto praticamente em cem por cento das vezes pelas estimativas dos estimadores da razão separado, RhS e não viciado, RhNV, e por seus respectivos representantes Jackknife, RhSJK e RhNVJK.

Em resumo, uma pequena vantagem para o estimador da razão separado, RhS, em relação ao estimador não viciado, RhNV, o coloca como o melhor estimador para a variável PR

4.1.10. Discussão dos resultados das simulações para a variável PU;

Para a variável PU, realizaram-se 25 rodadas de simulações, para as diferentes proporções de compartilhados entre os catorze grupos, nos quais a população inteira foi dividida, com o menor número de localidades a ser amostrado na população que, no caso, foi de sessenta. Como resultado de todas as simulações, nas 25 rodadas e no conjunto de todas elas, ou seja, nas 75 rodadas, verificou-se que:

1 – Os estimadores da razão separado, RhS e não viciado, RhNV, têm os menores erros relativos percentuais em valores absolutos das estimativas populacionais em relação ao valor simulado populacional, seguidos dos seus representantes Bootstrap e Jackknife, RhSBB, RhSBS, RhSJK, RhNVBB, RhNVBS, e RhNVJK, respectivamente, que têm erros relativos percentuais, em valor absolutos menores que meio. O estimador da razão combinado RhCB também tem erros relativos percentuais, em valor absoluto, menores que meio. Já os estimadores da razão separado e não viciado têm os erros relativos percentuais, os erros padrão e a amplitude de suas variações menores do que os dos demais estimadores acima mencionados.

2 – Quanto ao número de localidades estimadas com erro relativo menor do que 5, os estimadores da razão separado e não viciado têm a quase totalidade delas, em todos os casos de variação nas diferenças do número de compartilhados. Aqui também, como no caso da variável PR, o estimador da regressão linear, RLGE, não estima sequer uma localidade com erro relativo percentual em valor absoluto menor

do que 5, apesar de fazer estimativas de treze dos dezesseis grupos nos quais a população está dividida, o que o caracteriza como um mau estimador das localidades para essa variável.

3 – Quanto à cobertura do valor simulado populacional, as estimativas dos estimadores separado, RhS e não viciado, $RhNV$, bem como as de seus representantes Jackknife, $RhSJK$ e $RhNVJK$, cobrem o valor simulado populacional em praticamente cem por cento das vezes, com erros padrão e amplitudes de variação desse número de vezes praticamente nulos.

Em resumo, na caracterização de ‘melhor estimador’ praticamente não há diferença entre os estimadores da razão separado, RhS , e não viciado, $RhNV$, clássicos, a não ser uma diferença muito pequena no número de localidades estimadas, a favor do estimador da razão separado, RhS .

5. CONCLUSÃO

Para a obtenção dos resultados deste trabalho, consideraram-se dez variáveis com características diferentes, pertencentes a dois bancos de dados distintos. Para algumas das variáveis estudadas, o método de agrupamento das k-Médias forneceu um número de localidades a ser amostrado bem maior do que para algumas estratificações das mesmas variáveis.

Fez-se uma comparação entre as estimativas de diversos estimadores clássicos e não clássicos (computacionalmente intensivos). Simulou-se uma quantidade de compartilhados na faixa de 75 a 90%. Realizaram-se, então simulações Monte Carlo para uma avaliação desses estimadores. Como resultado, verificou-se que o estimador da razão separado clássico é o melhor estimador para esse tipo de problema, na faixa considerada para o número de compartilhados, quando a quantidade de equipamentos não for simultaneamente multimodal e com valores da ordem das unidades.

Como não se fez um estudo para as proporções de compartilhados menores do que 75%, e tendo em vista a possibilidade da existência concomitante de valores da ordem das unidades e multimodais, na variável a ser estimada, sugere-se que:

- 1-Faça-se uma ordenação das quantidades de equipamentos da variável sob estimação;
- 2-Verificar-se a existência de valores da ordem das unidades multimodais, para as quantidades de equipamentos das localidades. Caso existam quantidades pequenas de equipamentos, com modas grandes, e se a amplitude de variação da quantidade de equipamentos na população não for grande, deve-se, além de agrupamentos, fazer-se também estratificações para a verificação da quantidade de equipamentos a ser amostrado. Possivelmente esses valores pequenos de quantidades de equipamentos tenham que ser filtrados (eliminados) das análises se as suas modas forem grandes e a amplitude de variação dessa quantidade for pequena na população.

3-Realize-se uma verificação do número máximo de agrupamentos possíveis para a variável sob estimação, para se identificar o menor número de localidades a ser amostrado na população, para a situação. Caso o número de grupos obtidos seja pequeno e a quantidade de localidades a ser amostrada seja grande, deve-se fazer também estratificações para comparação; caso essa comparação forneça menor número de localidades a ser amostrado, deve-se trabalhar então com as estratificações para a realização das estimativas;

4-Simule-se diferenças de proporção, variando as quantidades do número de compartilhados, entre os diversos grupos ou estratos nos quais a população inteira foi dividida, a partir de alguma estimativa inicial, a partir de uma amostra piloto ou de informações outras.

5-Proceda-se então, às simulações Monte Carlo para a verificação do comportamento de cada estimador, sob as diversas diferenças de compartilhados, próximas às obtidas no item 4, acima, para então decidir o estimador a ser empregado para a estimação desejada.

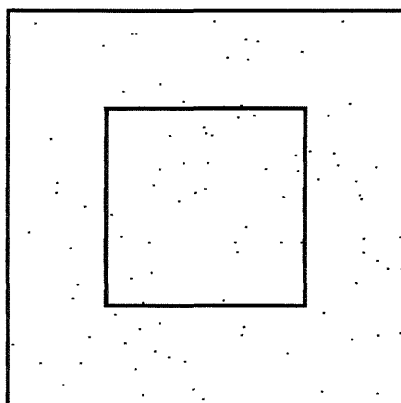
REFERENCIAS BIBLIOGRÁFICA

1. CHAVES NETO, A.. **Bootstrap e Regressão. Uma Opção aos Mínimos Quadrados Generalizados.** Dissertação de Mestrado, Instituto de Matemática, Estatística e Ciência da Computação -UNICAMP - Campinas, São Paulo, 1985.
2. _____.. **Bootstrap em Séries Temporais.** Tese de Doutorado, Departamento de Engenharia Elétrica - Grupo de Sistemas, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 1991.
3. _____..**Probabilidade e Estatística – Notas de Aulas.** UFPR, 87 páginas, 1999.
4. COCHRAN, W. G.. **Técnicas de Amostragem,** Ed. Fundo de Cultura S.A, Rio de Janeiro, 1965.
5. CORMACK, R.M.. A Review of Classification. **Journal of the Royal Statistical Society,** vol. 134, n. 3,321-367, 1971.
6. DAVIDSON, A.C.; HINKLEY, D.V.; SCHECTMAN, E..Efficient bootstrap simulation, **Biometrika,** vol. 73, n.3, p. 555-566, 1986.
7. DIACONIS, P.; EFRON, B.. Computer-Intensive Methods in Statistics. **Scientific American,** vol. 248, 116-130, 1983.
8. EFRON, B.. **The Jackknife, the Bootstrap and Other Resampling Plans,** Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania, USA, 1982
9. EFRON, B. & TIBISHIRANI, R.. 1993; **An Introduction to the Bootstrap,** CHAPMAN & HALL, New York, USA.
10. EFRON, B.. Bootstrap Methods; Another Look at the Jackknife. **The Annals of Statistics,** vol. 7, n. 1, 1-26, 1979.
11. **ENCYCLOPEDIA of Statistical Science,** Vol.. 7, p. 639, New York, USA, John Willey, 1986.
12. GRAHAM, R.L.; HINKLEY, D. V.; JHON, P. W. M.; SHI, S.. Balanced Design of Bootstrap Simulations. **Journal of the Royal Statistics Soc. B,** 52, n. 1, p. 185 – 202.
13. HAMMERSLEY, J. M.; HANDSCOMB, D. C.. 1964-**Monte Carlo Methods,** John Willey, New York, USA.

14. HARTIGAN, J. A.; WONG, M.A. A.. k-Means Clustering Algorithm. **Applied Statistics**, 22, 126-130, 1973.
15. JOHNSON, R. A.; WICHERN, D. W.. **Applied Multivariate Statistical Analysis**, New Jersey, USA Prentice Hall, 1998.
16. KISH, L.. **Survey Sampling**. New York, USA, John Willey, 1965.
17. LACHENBRUCH, P. A.. **An Introduction to Monte Carlo Methods with Selected Applications**, System Development Corporation, Santa Monica, California, USA, 1965.
18. MILLER, R.. The Jackknife - a review. **Biometrika**, 61, 1-15, 1974.
19. PETERSON, I. Pick a sample. **Science News**. July 27, 1991.
20. RAO, P.S.R.; KATZOFF, M.J.. Bootstrap for Finite Populations. **Commun. Statist.- SIMULA**, 25(4), 979- 994, 1996.
21. SALINAS, D.T.P.. **Bootstrap não-paramétrico aplicado a dados incompletos**. Dissertação de Mestrado- IME- USP- São Paulo-SP, 1996.
22. SARLE, W.. Bootstrap Confidence Intervals.- saswss@hotmailng.unx.sas.com (warrensarle).
23. SIMON, Julian L.; ATKINSON, David T.; SHEKOVAS, Carolyn..Probability and Statistics: Experimental Results of a Radically Different Teaching Method. **American Mathematical Monthly**, vol. 83, n. 9, 1976.
24. SOBOL, Ilia M.. **A Primer for the Monte Carlo Method**. CRC Press, Florida, USA, 1994.
25. SWANEPOEL, J.W.H..A Review of Bootstrap Methods. **Journal of the South African Statistics**, vol. 24, p. 1-34, 1990.
26. YUNG, G.A.. Bootstrap: More then a Stab on the Dark?- **Statistical Science**. Vol. 9, n. 3, 382-415, 1994.

ANEXOS

ANEXO 1 – Exemplo de cálculo de área simples pelo método Monte Carlo



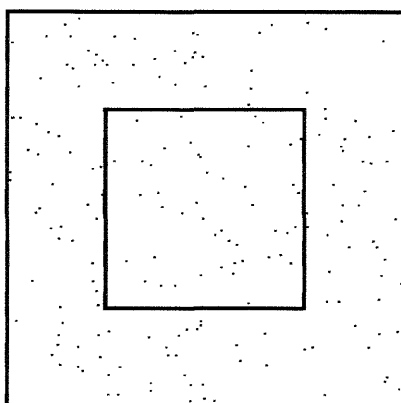
Para pontos = 100

A1 = 0,29	A6 = 0,25
A2 = 0,29	A7 = 0,24
A3 = 0,28	A8 = 0,19
A4 = 0,27	A9 = 0,32
A5 = 0,29	A10 = 0,32

Média = 0,274

Vício = 0,024

Erro P= 0,0392



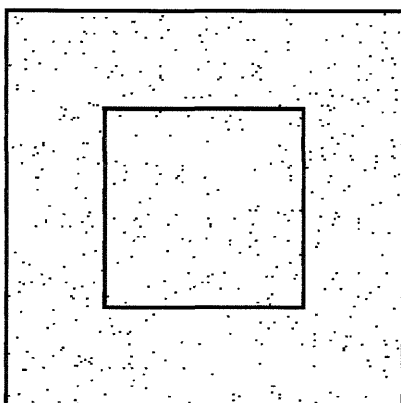
Para pontos = 200

A1 = 0,26	A6 = 0,23
A2 = 0,225	A7 = 0,26
A3 = 0,225	A8 = 0,22
A4 = 0,24	A9 = 0,2
A5 = 0,245	A10 = 0,255

Média = 0,236

Vício = -0,014

Erro P= 0,0195



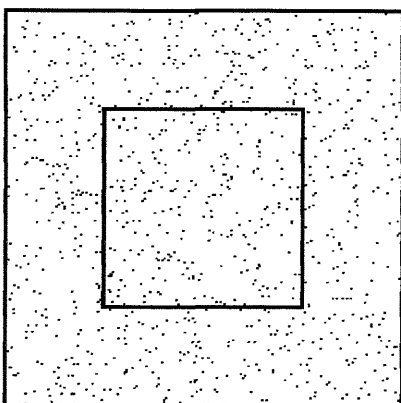
Para pontos = 500

A1 = 0,256	A6 = 0,258
A2 = 0,228	A7 = 0,228
A3 = 0,22	A8 = 0,264
A4 = 0,248	A9 = 0,27
A5 = 0,258	A10 = 0,228

Média = 0,2458

Vício = -4,1999

Erro P= 0,0180



Para pontos = 1000

A1 = 0,25	A6 = 0,254
A2 = 0,238	A7 = 0,262
A3 = 0,268	A8 = 0,248
A4 = 0,221	A9 = 0,228
A5 = 0,233	A10 = 0,262

Média = 0,2464

Vício = -3,5999

Erro P= 0,0158

ANEXO 2 – Ajuda para operação do programa EEC (Estimação de Número de Equipamentos de Uso Compartilhados).

1 – Apresentação

Este programa foi desenvolvido para fazer estimativas do número de equipamentos compartilhados, usando Agrupamento ou Estratificação. O agrupamento é automático, enquanto que a estratificação é um processo mecânico, de escolha dos limites inferior e superior de cada estrato. Para agrupar os itens, o método empregado é o método das k Médias.

O que segue, é uma sequência de passos, a ser seguida, para a utilização do programa, para a obtenção do número de equipamentos compartilhados.

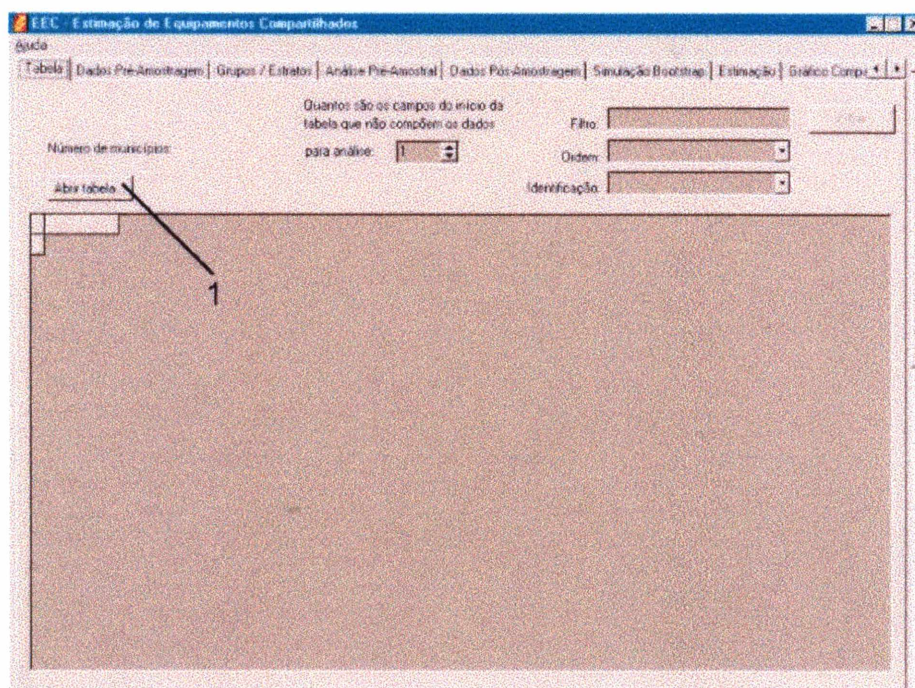
2 – Requisitos

A tabela a ser analisada deve ser em formato DBF. Este formato é o tipo de estrutura de dados do Dbase. Consegue-se este formato de dados, por exemplo, exportando uma planilha em forma de tabela do Excel (Salvar como...).

3 – Especificações

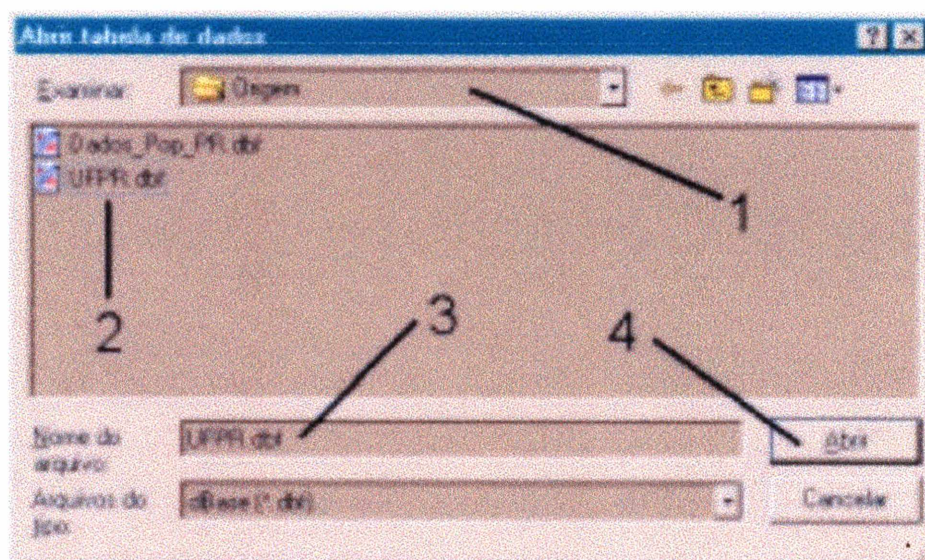
- Tabela com no máximo 10 variáveis para cálculo e com no máximo 5 variáveis de cabeçalho. Total de 15da variáveis;
- Número máximo de registros filtrados na tabela é 500;
- Nível de confiança da solução de 51 a 99%;
- Numero de grupos de 1 a 17;
- Número de estratos de 1 a 17;
- Número máximo de localidades (ou itens) por grupo/estrato é 500;

4 – Funcionamento



1- Abre a tabela.

Figura 06: Tela inicial do programa



- 1- Seleciona a pasta desejada;
- 2- Seleciona o arquivo desejável;
- 3- Arquivo selecionado;
- 4- Executa a abertura do arquivo.

Figura 07: Tela inicial do programa

EEC - Estimção de Equipamentos Compartilhados

Tabela | Dados Pré-Amostragem | Grupos / Estratos | Análise Pré-Amostragem | Dados Pós-Amostragem | Simulação Bootstrap | Estimção | Gráfico Comparativo

Quantos são os campos do início da tabela que não contém os dados para análise: 1

Filtro: Filtar

Ordem: Identificação:

Número de municípios: 3224

Arquivo tabela: C:\EEC\Origem\UFPR.dbf

CODIGO	MUNICIPIO	ANO	NC_RES	NC_COM	NC_IND	NC_RUR	NC_POD	NC_ILU	NC_AES
1	ABATIA	1990	890	159	10	393	27	1	4
1	ABATIA	1991	932	162	11	357	30	1	5
1	ABATIA	1992	1049	163	10	361	30	1	5
1	ABATIA	1993	1203	159	9	363	30	1	5
1	ABATIA	1994	1264	163	10	374	31	1	5
1	ABATIA	1995	1333	169	8	369	34	1	5
1	ABATIA	1996	1344	163	9	371	32	1	5
1	ABATIA	1997	1394	161	10	366	32	2	5
1	ABATIA	1998	1426	162	12	379	32	2	5
2	ADRIANOPOLIS	1990	735	98	14	214	47	4	6
2	ADRIANOPOLIS	1991	729	102	14	236	52	4	6
2	ADRIANOPOLIS	1992	727	100	19	240	53	4	6
2	ADRIANOPOLIS	1993	948	102	13	245	47	4	6
2	ADRIANOPOLIS	1994	996	103	13	259	50	4	6
2	ADRIANOPOLIS	1995	1013	101	13	264	50	4	6
2	ADRIANOPOLIS	1996	936	98	9	269	49	4	6
2	ADRIANOPOLIS	1997	940	90	9	268	49	4	6
2	ADRIANOPOLIS	1998	961	95	8	299	49	4	6
3	AGUADOS DO SUL	1990	506	74	10	452	32	1	4
3	AGUADOS DO SUL	1991	541	76	13	480	33	1	4

- 1- Número de registros filtrados da tabela;
- 2-da variável ANO ainda não filtrado;
- 3- Indicativo de cada variável;
- 4- Caminho do arquivo de dados;
- 5-da variável NC_RES ainda não ordenado crescentemente;

Figura 08: Descrição das variáveis

EEC - Estimção de Equipamentos Compartilhados

Tabela | Dados Pré-Amostragem | Grupos / Estratos | Análise Pré-Amostragem | Dados Pós-Amostragem | Simulação Bootstrap | Estimção | Gráfico Comparativo

Quantos são os campos do início da tabela que não contém os dados para análise: 1

Filtro: ANO = 1990 Filtar

Ordem: NC_RES Identificação: MUNICIPIO

Número de municípios: 320

Arquivo tabela: C:\EEC\Origem\UFPR.dbf

CODIGO	MUNICIPIO	ANO	NC_RES	NC_COM	NC_IND	NC_RUR	NC_POD	NC_ILU	NC_AES
68	EL VIEJO	1990	125	47	2	1159	43	4	0
244	RIO NEGRO	1990	150	25	7	637	20	1	1
44	CAMPO BONITO	1990	156	42	7	541	14	1	1
181	VALIANCAIAI	1990	170	12	5	72	16	1	1
136	JO. OLINDA	1990	214	13	1	30	22	1	1
95	GODOY MOREIRA	1990	244	77	4	306	18	1	3
101	GUAFORIMA	1990	268	29	5	121	25	3	1
206	PAULA FREITAS	1990	273	40	13	495	44	4	4
174	MIRADOR	1990	278	22	7	40	19	3	3
266	STO.A. PARAISO	1990	297	28	3	176	67	2	2
314	UNIFLOR	1990	305	29	2	161	23	1	2
290	SULINA	1990	306	59	7	260	19	3	1
207	PAULO FRONTIN	1990	329	69	18	748	49	2	4
70	CORUMBATAI SUL	1990	329	61	7	497	15	1	2
14	ANTONIO QUINTO	1990	336	58	13	623	26	3	5
223	PORTO VITORIA	1990	355	37	19	289	26	2	2
259	SANTA INES	1990	365	38	6	73	31	2	4
222	PORTO RICO	1990	369	40	6	143	34	2	5
285	S PEDRO PARANA	1990	383	41	6	222	40	2	4
36	CAFEARA	1990	390	26	3	90	21	1	1

- 1-das variáveis do início do arquivo;
- 2-da variável ANO filtrado. Quando o filtro está vazio, a tabela não está filtrada;
- 3-da variável NC_RES ordenado crescentemente;
- 4- Identificação dos municípios nos resultados;
- 5- Executa o filtro indicado.

Figura 09: Continuação descrição das variáveis

Dados para Agrupamento e Estratificação

1- Nível de confiança da solução: 95 1

2- Estimativa Inicial da Razão: 0.5 2

3- Erro Relativo (t): 0.1 3

4- Precisão da Estimativa (d): 0.005 4

Agrupamento (não hierárquico, médias, distância Euclidiana)

1- Número de Grupos: 10 5

2- Selecione os demais campos:

☒ NC_PES 6

☐ NC_COM

☐ NC_IND

☐ NC_PUJ

☐ NC_POD

Agrupar 8

Estratificação

1- Número de Estratos: 5 9

OK 10

Estrato	100	até	500
Estrato 1	100	até	500
Estrato 2	500	até	5000
Estrato 3	5000	até	30000
Estrato 4	30000	até	70000
Estrato 5	70000	até	400000

11

Estratificar 12

Amostragem proporcional ao tamanho do grupo/estrato

☒ Aleatória 7

☐ Sistemática

- 1- Nível de confiança da solução, vai de 51 a 99;
- 2- Entrar com a Estimativa Inicial da Razão;
- 3- Entrar com o Erro Relativo;
- 4- Entrar com a Precisão da Estimativa;
- 5- Números de Grupos, vai 1 a 17;
- 6- Já vem selecionada a variável de Ordem. Esta variável não pode ser desselecionada. Selecionar as demais variáveis para agrupamento caso haja interesse.
- 7- Selecionar entre amostragem Aleatória e Sistemática (aproximada).
- 8- Executa os cálculos para o agrupamento;
- 9- Número de estratos, vai de 1 a 17;
- 10- Cria as caixas de edição referente ao número de estratos.
- 11- Caixas de edição para entrar com os valores extremos dos estratos. Entrar com valores inteiros. O valor do limite inferior do estrato seguinte deve ser maior que o limite superior do estrato anterior.
- 12- Executa os cálculos para a estratificação. A estratificação e o agrupamento são independentes.

Figura 10: Continuação descrição das variáveis

EEC - Estimação de Equipamentos Compartilhados

Tela | Dados Pré-Amostragem | Grupos / Estratos | Análise Pré-Amostragem | Dados Pós-Amostragem | Simulação Bootstrap | Estimação | Gráfico Comparativo

Tamanho mínimo da amostra na População de eqtos: 37419

GRUPOS

Iterações: 51

Municípios a serem amostrados: 75

Grupo	Total Grupo	Amost. Total	Localidades	Amost. Mun.	Variação
4	160137	4157	9	5	7826979
Localidades	Total Loc.	Amostrados			
ARAPONGAS	14819				
CAMBE	15112				
PARANAVAI	15562				
CAMPO MOURAO	16227	716			
TOLEDO	17151	757			
UMUARAMA	18975	838			
APLICARANA	20461	903			
COLOMBO	21292				
PIRAQUARA	21348	943			

Exemplo de resultado do agrupamento.

Figura 11: Resultado do agrupamento

EEC - Estimação de Equipamentos Compartilhados

Tela | Dados Pré-Amostragem | Grupos / Estratos | Análise Pré-Amostragem | Dados Pós-Amostragem | Simulação Bootstrap | Estimação | Gráfico Comparativo

Tamanho mínimo da amostra na População de eqtos: 37419

ESTRATOS

Municípios a serem amostrados: 151

Estrato	Total Estrato	Amost. Total	Localidades	Amost. Mun.	Variação	Desv. P.
4	179770	4667	4	4	149424675	12224
Localidades	Total Loc.	Amostrados				
FOZ DO IGUAÇU	32347	840				
CASCATEL	37942	985				
PONTA GROSSA	49822	1293				
MARINGÁ	59667	1549				
Estrato	Total Estrato	Amost. Total	Localidades	Amost. Mun.	Variação	Desv. P.
5	433265	11247	2	2	31418476465	172253
Localidades	Total Loc.	Amostrados				
LONDRINA	91296	2370				
CIJURUPA	541969	8877				

Exemplo de resultado da estratificação.

Figura 12: Resultado estratificação

NC	S/A	G	Loc	Amostr. G/E	Amostr.	Variância
95 A	17	100	26	4	2	37419
95 A	17	13	32	6	10	37419
95 S	17	13	32	6	10	37419
95 A	1	7	220			37419
95 S	1	7	220			37419
95 A	1	4	12	156		37419
95 S	1	4	12	156		37419
95 A	1	3	7	10	124	37419
95 S	1	3	7	10	124	37419
95 A	1	3	6	10	21	37419
95 S	1	3	6	10	21	37419
95 A	1	3	3	5	9	37419
95 S	1	3	3	5	9	37419
95 A	1	3	3	5	8	37419
95 S	1	3	3	5	8	37419
95 A	1	3	3	5	8	37419
95 S	1	3	3	5	8	37419
95 A	1	3	3	5	8	37419
95 S	1	3	3	5	8	37419
95 A	1	3	3	5	8	37419
95 S	1	3	3	5	8	37419
95 A	1	3	3	5	8	37419
95 S	1	3	3	5	8	37419

- Análise Pré-Amostragem, tabela contendo os principais resultados para análise.

Figura 13: Análise pré-amostragem

Amostrados	Localidades	Compartilhados	% Estimado
6977	CURITIBA	6656	75
1293	PONTA GROSSA	970	75
1549	MARILIA	1162	75
2370	LAGEADO	1777	75
952	PARANAGUA	714	75
1081	FOZ DE IGUAÇUAS	788	75
1332	PARANAGUA	924	75
701	TOLEDO	536	75
722	JOINVILLE	542	75
796	COARÉ	597	75
949	PARANAGUA	712	75
980	APUCARANA	741	75

- 1- Caixa de edição com os valores dos amostrados;
- 2- Identificação dos municípios sorteados;
- 3- Caixa de edição com os valores dos compartilhados;
- 4- Caixa de edição com valores numéricos que vai de 20 a 300%;
- 5- Limpa todas as caixas de edição de Compartilhados, para possibilitar a execução do próximo item;
- 6- Preenche as caixas de edição de Compartilhados com a porcentagem de cada estrato % Estrato, caso todas as caixas de edição de compartilhados estejam vazias;
- 7- Quando se preenchem manualmente as caixas de edição de Compartilhados, é conveniente saber a que porcentagem corresponde estes valores, por estrato/grupo;
- 8- Executa todos os cálculos de análise para uma iteração, obtendo o tempo dos cálculos;

Figura 14: Dados pós-amostragem

EQUIPAMENTOS AMOSTRADOS E COMPARTILHADOS EM CADA LOCALIDADE

Amostrados	Localidades	Compartilhados	% Estimado
G/E 1			
1807	GUARUJÁ	6056	75
G/E 2			
1280	PONTA GROSSA	870	75
1549	PARANÁ	1162	
2390	LORETO	1777	
G/E 3			
952	PARANÁ	714	75
1060	RIO DO SUL	786	
1282	GUARUJÁ	924	
G/E 4			
701	PARANÁ	525	75
722	GUARUJÁ	542	
796	PARANÁ	597	
940	PARANÁ	712	
1060	PARANÁ	781	

Tempo Estimado: 00:19:45

- 1- Ao marcar a quantidade de iterações Monte Carlo, obtém-se o tempo estimado de duração dos cálculos;
- 2- Executa os cálculos, agora para as aproximações Monte Carlo.

Figura 15: Continuação dados pós-amostragem

DOS GRUPOS

Tempo de Processamento: 00:00:08

GRUPO: 10
CAMPO: - NC RES
AMOSTRAGEM ALÉATORIA

RESULTADOS DAS ESTIMATIVAS CLÁSSICAS BOOTSTRAP E JACKKNIFE

Média da variável de interesse	NC RES
Viz. tot. da variável de interesse	1441479
Viz. estimado de compartilhados	1061109
Estim. Expansão Simplex	80281
Estim. M. Combinado	1080854
Estim. Reg. Lin. Clássica (para K = 2)	1061133
Estim. Var. da estim. dos Comp.	0.17888260369941
Coef. de Variação (CV)	5.22903562341661E-2
Soma dos Resíduos	1.09423561307058E-12
Soma dos Comp. e Resíduos	3209.1820863093

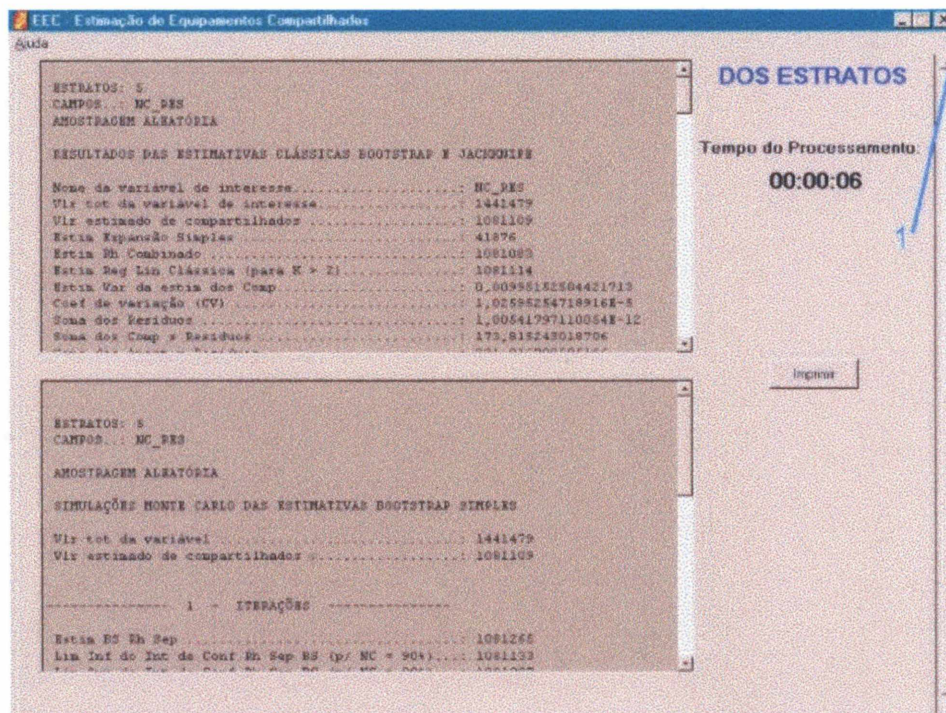
GRUPO: 10
CAMPO: - NC RES
AMOSTRAGEM ALÉATORIA

SIMULAÇÕES MONTE CARLO DAS ESTIMATIVAS BOOTSTRAP SIMPLES

Viz. tot. da variável	1441479
Viz. estimado de compartilhados	1061109
1 - ITERAÇÕES	
Estim. DO St. Sep.	1061333
Lim. Inf. do Int. de Conf. DO St. Sep. DO Int. NC = 3041	1061333

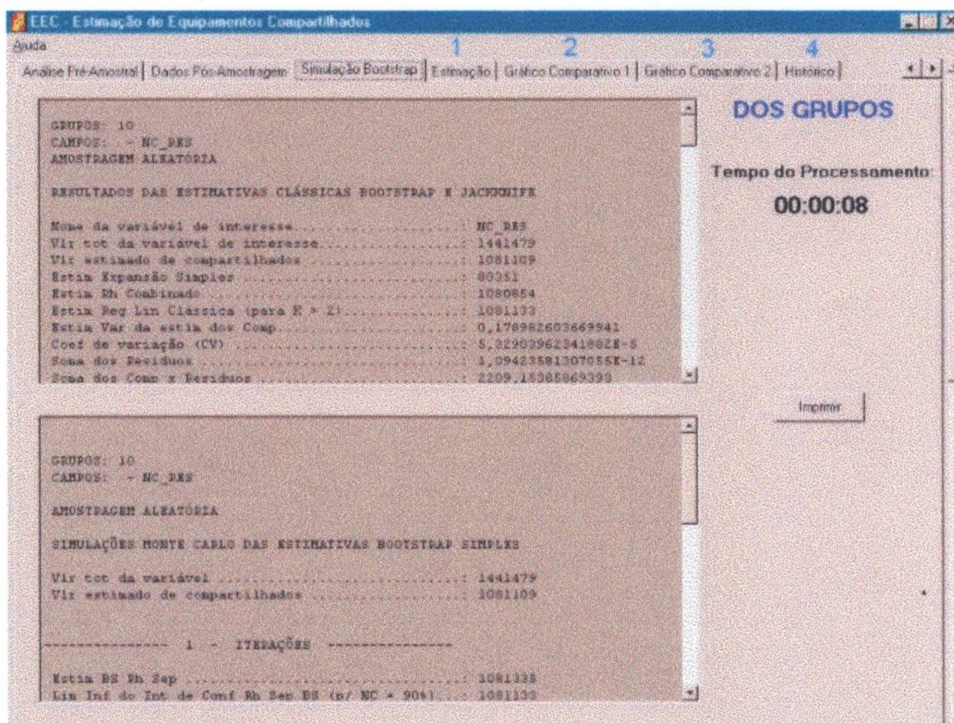
- Tela de resultados dos cálculos, para grupos;

Figura 16: Resultado dos cálculos do agrupamento.



- Exemplo de resultados dos cálculos, para estratos;
- 1- Sobe a tela e mostra o menu.

Figura 17: Resultado dos cálculos estratificação



- 1- Tela de Estimção, onde mostra as estimções para cada município e o melhor estimador;
- 2- Gráfico comparativo 1;
- 3- Gráfico comparativo 2;
- 4- Histórico das estimções, para posterior análise.

Figura 18: Demais telas

EEC - Estimaco de Equipamentos Compartilhados								
Tabela Dados Pr-Amostragem Grupos / Estratos Anlise Pr-Amostrai Dados Ps-Amostragem Simulao Bootstrap Estimaco Grfico Comp								
Grupo 1		CVAC:	0					
Localidades	TotLoc.	Est.Comp.	RhS	RhSBB	RhSJK	RhSBS	RhNV	F
CURITIBA	341969	256477	256486	256430	256430	256430	256486	2
Grupo 2		CVAC:	0,264168					
Localidades	TotLoc.	Est.Comp.	RhS	RhSBB	RhSJK	RhSBS	RhNV	F
LONDRINA	91296	68472	68472	68460	68460	68460	68476	6
MARINGA	59667	44750	44750	44742	44742	44742	44753	4
PONTA GROSSA	49822	37366	37366	37360	37360	37360	37369	3
Grupo 3		CVAC:	0,107502					

Figura 19: Estimaco

EEC - Estimaco de Equipamentos Compartilhados								
Tabela Dados Pr-Amostragem Grupos / Estratos Anlise Pr-Amostrai Dados Ps-Amostragem Simulao Bootstrap Estimaco Grfico Comp								
RhCB	MELHOR							ER 5
256386	- RhS - RhSBB - RhSJK - RhSBS - RhNV - RhNVBB - RhNVJK - RhNVBS - RL - RLBB - RLBS - RLJK - R 0							
RhCB	MELHOR							ER 5
68448	- RhS - RhSBB - RhSJK - RhSBS - RhNV - RhNVBB - RhNVJK - RhNVBS - RL - RLBB - RLBS - RLJK - R 0							
44734	- RhS - RhSBB - RhSJK - RhSBS - RhNV - RhNVBB - RhNVJK - RhNVBS - RL - RLBB - RLBS - RLJK - R 0							
37353	- RhS - RhSBB - RhSJK - RhSBS - RhNV - RhNVBB - RhNVJK - RhNVBS - RL - RLBB - RLBS - RLJK - R 0							

Figura 20: Estimaco (continuao)

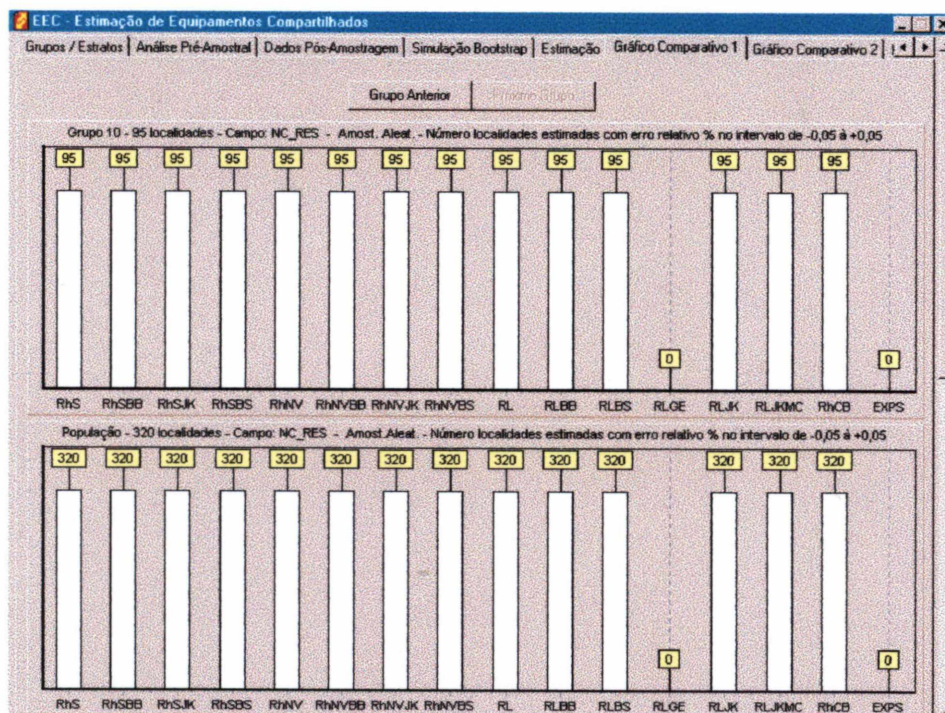


Figura 21: Tela de apresentação dos gráficos das estimativas do número de localidades estimadas dentro do intervalo do erro relativo percentual fixado pelo nível de confiança da solução. Dentro de cada grupo e na população.

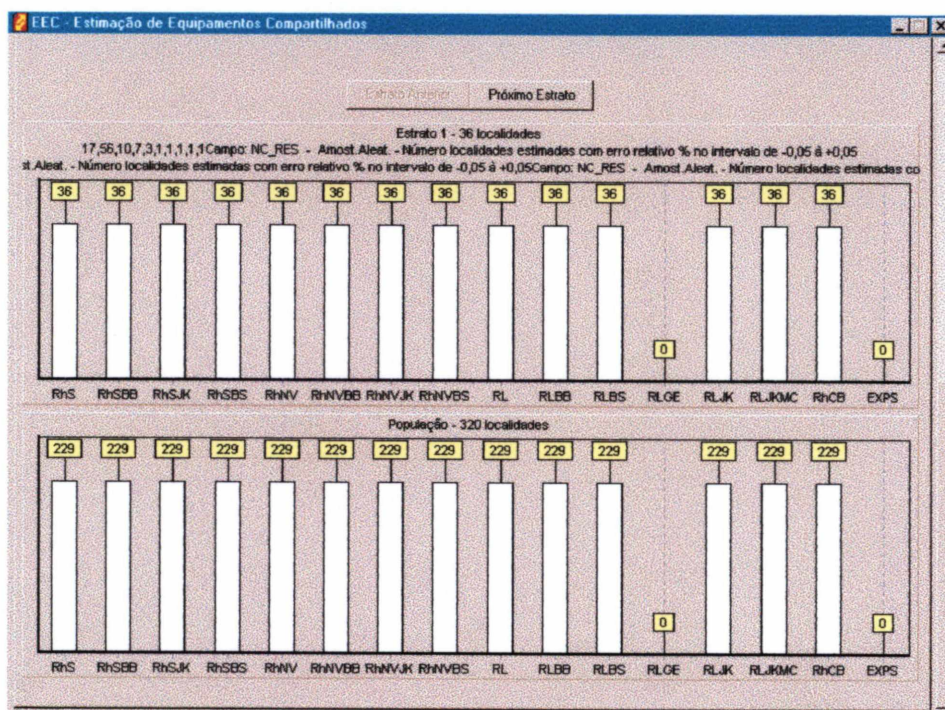


Figura 22: Gráficos das estimativas do número de localidades estimadas dentro do intervalo do erro relativo percentual fixado pelo nível de confiança da solução. Dentro de cada estrato e na população.

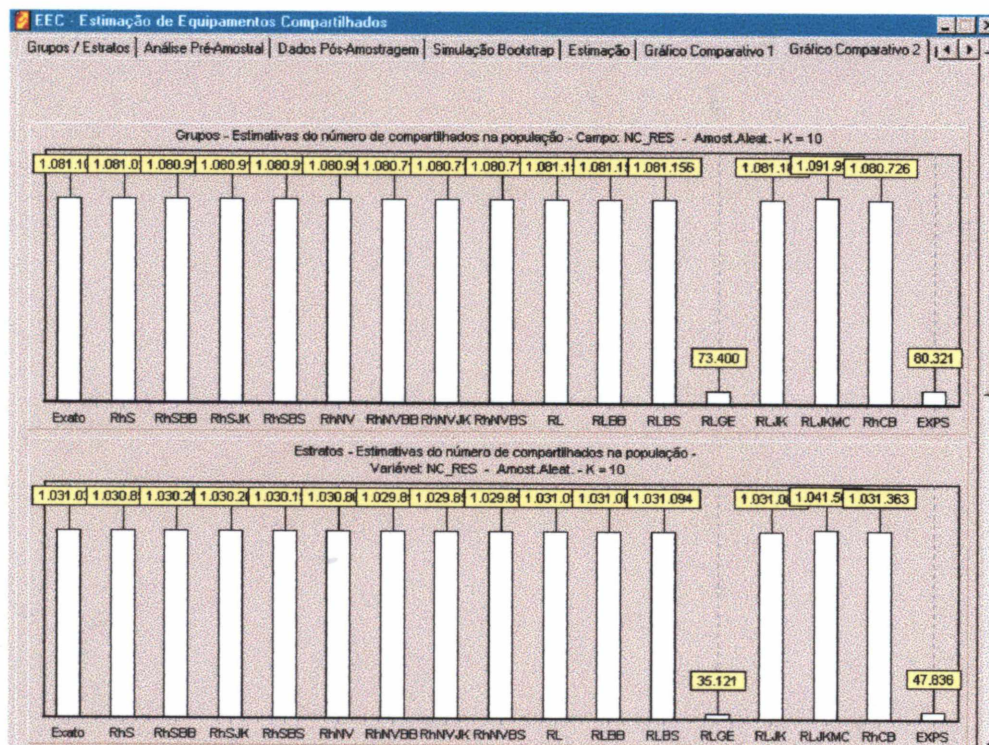


Figura 23: Gráficos comparativos das estimativas populacionais de cada estimador.

ANEXO 3 –Quadros de Análises Pré-Amostrais.

QUADRO 00: ANÁLISE PRÉ-AMOSTRAL, das variáveis NC_AES, NC_ILU, NC_PRO, NC_COM

QG/E	VARIÁVEIS	NC	S/A	QLP	QLAP	QLAGE	TOT-POP	AMOSTRA	VAR-POP.
E04	NC_ILU	95	A	318	65	49,9,4,3	1013	215	6
E06	NC_AES	95	A	318	64	36,9,8,2,4,5	1908	421	74
E06	NC_PRO	95	A	306	62	36,9,8,4,3,2	1296	668	10140
G02	NC_AES	95	A	318	111	77,34	1908	709	30
G02	NC_ILU	95	A	318	97	97,0	1013	390	7
G03	NC_PRO	95	A	306	153	1,7,145	1296	839	294
G12	NC_PRO,NC_COM	95	A	306	306	1,5,7,13,14,31,33,49,48,52,35,18	1296	1254	7343644
G13	NC_AES,NC_COM	95	A	318	317	1,5,7,13,15,32,33,48,46,49,32,31,5	1908	1818	7344022
G13	NC_ILU,NC_COM	95	A	318	318	1,5,7,13,15,32,33,48,48,51,37,23,5	1013	985	7343726

QUADRO 01: ANÁLISE PRÉ-AMOSTRAL, da variável NC_AES.

QG/E	VARIÁVEIS	NC	S/A	QLP	QLAP	QLAGE	TOT-POP	AMOSTRA	VAR-POP.
G01	NC_AES	95	A	318	145	145	1908	921	25
G01	NC_AES	95	S	318	145	145	1908	829	25
G02	NC_AES	95	A	318	111	77,34	1908	702	30
G02	NC_AES	95	S	318	111	77,34	1908	731	30
E04	NC_AES	95	A	318	72	46,15,7,4	1908	564	114
E04	NC_AES	95	S	318	72	46,15,7,4	1908	559	114
E05	NC_AES	95	A	318	67	36,9,15,2,5	1908	471	75
E05	NC_AES	95	S	318	67	36,9,15,2,5	1908	462	75
E06	NC_AES	95	A	318	64	36,9,8,2,4,5	1908	443	74
E06	NC_AES	95	S	318	64	36,9,8,2,4,5	1908	437	74
E07	NC_AES	95	A	318	58	36,9,3,2,2,1,5	1908	380	72
E07	NC_AES	95	S	318	58	36,9,3,2,2,1,5	1908	372	72

Quantidade de grupos QG/E, da variável para agrupamento, no caso a NC_AES, Nível de Confiança pretendido da solução NC, tipo de amostragem A = aleatória Simples, S = Sistemática, Quantidade de Localidades na População QLP, Quantidade de Localidades a serem amostradas na população QLAP, Quantidade de Localidades a serem amostradas no Grupo ou Estrato QLAGE, quantidade de equipamentos, total na população Tot Pop., quantidade de equipamentos a serem amostrados na população AMOSTRA, Variância Populacional, VARIÂNCIA POP.

QUADRO 02: ANÁLISE PRÉ-AMOSTRAL, da variável NC_ILU.

QG/E	VARIÁVEIS	NC	S/A	QLP	QLAP	QLAGE	TOT-POP	AMOSTRA	VAR-POP
G01	NC_ILU	95	A	320	152	152	1018	454	7
G02	NC_ILU	95	A	320	97	97,0	1018	361	7
E04	NC_ILU	95	A	320	65	50,9,3,3	998	203	6

Quantidade de grupos QG/E, da variável para agrupamento, sendo o primeiro deles o da variável de interesse, no caso a NC_ILU, Nível de Confiança pretendido da solução NC, tipo de amostragem A = aleatória Simples, S = Sistemática, Quantidade de Localidades na População QLP, Quantidade de Localidades a serem amostradas na população QLAP, Quantidade de Localidades a serem amostradas no Grupo ou Estrato QLAGE, quantidade de equipamentos, total na população Tot Pop., quantidade de equipamentos a serem amostrados na população AMOSTRA, Variância Populacional, VARIÂNCIA POP.

QUADRO 03: ANÁLISE PRÉ-AMOSTRAL, da variável NC_POD.

QG/E	VARIÁVEIS	NC	S/A	QLP	QLAP	QLAGE	TOT-POP	AMOSTRA	VAR-POP.
G01	NC_POD	95	A	320	236	236	20261	13265	9358
G02	NC_POD	95	A	320	139	3,136	20261	9309	346399
G03	NC_POD	95	A	320	97	1,21,75	20261	8021	10782
G04	NC_POD	95	A	320	78	1,3,23,51	20261	6531	6319
G05	NC_POD	95	A	320	59	1,3,13,16,26	20261	5780	6119
G06	NC_POD	95	A	320	50	1,3,6,8,10,22	20261	4958	5775
G07	NC_POD	95	A	320	45	1,3,6,8,6,8,13	20261	4829	5744
G08	NC_POD	95	A	320	43	1,3,6,7,6,6,5,9	20261	4577	5731
E14	NC_POD	95	A	320	29	9,4,3,1,1,1,1,1,0,2,1,1,3,1	20261	3271	5082

Quantidade de grupos QG/E, da variável para agrupamento, sendo o primeiro deles o da variável de interesse, no caso a NC_POD, Nível de Confiança pretendido da solução NC, tipo de amostragem A = aleatória Simples, S = Sistemática, Quantidade de Localidades na População QLP, Quantidade de Localidades a serem amostradas na população QLAP, Quantidade de Localidades a serem amostradas no Grupo ou Estrato QLAGE, quantidade de equipamentos, total na população Tot Pop., quantidade de equipamentos a serem amostrados na população AMOSTRA, Variância Populacional, VARIÂNCIA POP.

QUADRO 04: ANÁLISE PRÉ-AMOSTRAL da variável NC_PRO.

QG/E	VARIÁVEIS	NC	S/A	QLP	QLAP	QLAGE	TOT-POP	AMOSTRA	VAR-POP
G01	NC_PRO	95	A	306	287	287	1296	1174	187
G01	NC_PRO	95	S	306	287	287	1296	1243	187
G02	NC_PRO	95	A	306	238	1,237	1296	1066	44
G02	NC_PRO	95	S	306	238	1,237	1296	992	44
G03	NC_PRO	95	A	306	153	1,7,145	1296	854	294
G03	NC_PRO	95	S	306	153	1,7,145	1296	867	294
E07	NC_PRO	95	A	306	62	36,9,8,3,2,3,1	1296	686	261
E07	NC_PRO	95	S	306	62	36,9,8,3,2,3,1	1296	677	261

Quantidade de grupos QG/E, da variável para agrupamento, no caso a NC_PRO, Nível de Confiança pretendido da solução NC, tipo de amostragem A = aleatória Simples, S = Sistemática, Quantidade de Localidades na População QLP, Quantidade de Localidades a serem amostradas na população QLAP, Quantidade de Localidades a serem amostradas no Grupo ou Estrato QLAGE, quantidade de equipamentos, total na população Total Pop., quantidade de equipamentos a serem amostrados na população AMOSTRA, Variância Populacional, VARIÂNCIA POP.

QUADRO 05: ANÁLISE PRÉ-AMOSTRAL da variável NC_RES.

QG/E	VARIÁVEIS	NC	S/A	QLP	QLAP	QLAGE	TOT-POP	AMOSTRA	VAR-POP.
G01	NC_RES	95	A	320	308	308	1441479	37419	421509851
G02	NC_RES	95	A	320	277	1,276	1441479	37419	63592574
G03	NC_RES	95	A	320	228	1,7,220	1441479	37419	548952440
G04	NC_RES	95	A	320	178	1,4,17,156	1441479	37419	570356044
G05	NC_RES	95	A	320	153	1,3,7,18,124	1441479	37419	528869486
G06	NC_RES	95	A	320	123	1,3,6,10,21,82	1441479	37418	525674829
G07	NC_RES	95	A	320	110	1,3,3,5,9,23,66	1441479	37418	511504736
G08	NC_RES	95	A	320	91	1,3,3,5,8,9,14,48	1441479	37419	511029764
G09	NC_RES	95	A	320	84	1,3,3,5,8,9,9,14,32	1441479	37418	510944044
G10	NC_RES	95	A	320	75	1,3,3,5,8,9,6,6,7,27	1441479	37419	510898829
G11	NC_RES	95	A	320	71	1,3,3,5,8,9,6,6,6,18	1441479	37419	510888696
G12	NC_RES	95	A	320	67	1,3,3,5,8,9,4,4,3,6,17	1441479	37419	510849097
G13	NC_RES	95	A	320	62	1,3,3,5,8,9,4,4,3,2,3,4,13	1441479	37419	510843947
G14	NC_RES	95	A	320	60	1,3,3,5,8,9,4,4,3,2,3,3,5,7	1441479	37419	510841510

Quantidade de grupos QG/E, da variável para agrupamento, sendo o primeiro deles o da variável de interesse, no caso a NC_RES, Nível de Confiança pretendido da solução NC, tipo de amostragem A = aleatória Simples, S = Sistemática, Quantidade de Localidades na População QLP, Quantidade de Localidades a serem amostradas na população QLAP, Quantidade de Localidades a serem amostradas no Grupo ou Estrato QLAGE, quantidade de equipamentos, total na população Total Pop., quantidade de equipamentos a serem amostrados na população AMOSTRA, Variância Populacional, VARIÂNCIA POP.

QUADRO 06: ANÁLISE PRÉ-AMOSTRAL da variável NC_RUR.

QG/E	VARIÁVEIS	NC	S/A	QLP	QLAP	QLAGE	TOT-POP	AMOSTRA	VAR-POP.
G01	NC_RUR	95	A	306	147	147	227674	32870	409162
G02	NC_RUR	95	A	306	109	26,83	227674	32870	589657
G03	NC_RUR	95	A	306	91	13,23,55	227674	32871	569487
G04	NC_RUR	95	A	306	75	8,10,11,46	227674	30122	539201
G05	NC_RUR	95	A	306	69	3,6,8,11,41	227674	31071	444043
G06	NC_RUR	95	A	306	64	3,5,6,8,12,30	227674	29355	427376
G07	NC_RUR	95	A	306	57	3,2,3,4,6,11,28	227674	26696	372367
G08	NC_RUR	95	A	306	50	3,2,3,4,4,5,8,21	227674	25394	368997
G09	NC_RUR	95	A	306	46	3,2,2,2,3,3,5,6,20	227674	22345	359463
G10	NC_RUR	95	A	306	44	3,2,2,2,3,3,5,5,7,12	227674	22117	358766
G11	NC_RUR	95	A	320	43	3,2,2,2,2,2,3,3,4,8,12	230033	19466	356651
G12	NC_RUR	95	A	320	41	3,2,2,2,2,2,2,2,3,5,7,9	230033	19142	355892
G13	NC_RUR	95	A	320	37	3,2,2,2,2,2,2,2,2,3,3,7,5	230033	18911	355294

Quantidade de grupos QG/E, da variável para agrupamento, sendo o primeiro deles o da variável de interesse, no caso a NC_RUR, Nível de Confiança pretendido da solução NC, tipo de amostragem A = aleatória Simples, S = Sistemática, Quantidade de Localidades na População QLP, Quantidade de Localidades a serem amostradas na população QLAP, Quantidade de Localidades a serem amostradas no Grupo ou Estrato QLAGE, quantidade de equipamentos, total na população Total Pop., quantidade de equipamentos a serem amostrados na população AMOSTRA, Variância Populacional, VARIÂNCIA POP.

QUADRO 07: ANÁLISE PRÉ-AMOSTRAL, da variável PU.

QG/E	VARIÁVELS	NC	S/A	QLP	QLAP	QLAGE	TOT-POP	AMOSTRA	VAR-POP
G01	PU	95	A	371	354	354	7011990	38207	7081242135
G02	PU	95	A	371	319	1,318	7011990	38207	1329111984
G03	PU	95	A	371	265	1,7,257	7011990	38208	7798889292
G04	PU	95	A	371	210	1,4,15,190	7011990	38207	6925459664
G05	PU	95	A	371	160	1,4,12,26,117	7011990	38208	6886972631
G06	PU	95	A	371	139	1,4,10,13,21,90	7011990	38208	6920059894
G07	PU	95	A	371	130	1,4,3,5,11,20,86	7011990	38207	6458505933
G08	PU	95	A	371	110	1,4,3,5,9,9,18,61	7011990	38208	6450999466
G09	PU	95	A	371	103	1,4,3,5,4,6,9,16,55	7011990	38207	6431342292
G10	PU	95	A	371	91	1,4,3,5,4,6,8,10,13,37	7011990	38207	6430128539
G11	PU	95	A	371	85	1,4,3,3,5,3,4,4,9,12,37	7011990	38207	6409094907
G12	PU	95	A	371	79	1,4,3,3,5,3,4,4,4,5,10,33	7011990	38207	6408097609
G13	PU	95	A	371	73	1,4,3,3,5,3,4,3,4,4,6,8,25	7011990	38207	6407852575
G14	PU	95	A	371	65	1,4,3,3,5,3,4,3,3,2,4,5,9,16	7011990	38207	6407564302
G15	PU	95	A	371	61	1,4,3,3,5,3,4,3,3,2,2,4,4,6,14	7011990	38207	6407412493
G16	PU	95	A	371	58	1,4,3,3,5,3,4,3,3,2,2,4,4,5,6,6	7011990	38208	6407389074

QUADRO 178 - Quantidade de grupos QG/E, da variável para agrupamento, sendo o primeiro deles o da variável de interesse, no caso a PU. Nível de Confiança pretendido da solução NC, tipo de amostragem A = aleatória Simples, S = Sistemática, Quantidade de Localidades na População QLP, Quantidade de Localidades a serem amostradas na população QLAP, Quantidade de Localidades a serem amostradas no Grupo ou Estrato QLAGE, quantidade de equipamentos, total na população Total Pop., quantidade de equipamentos a serem amostrados na população AMOSTRA, Variância Populacional, VARIÂNCIA POP.

ANEXO 4 – Resultados das Estimativas Clássicas Bootstrap e Jackknife e Simulações Monte Carlo.

Tabela 01: RESULTADOS DAS ESTIMATIVAS: CLÁSSICAS, BOOTSTRAP, JACKKNIFE E SIMULAÇÕES MONTE CARLO

GRUPOS: 13 da variável: NC_COM AMOSTRAGEM ALEATÓRIA

Nome da variável de interesse	NC_COM		
Valor total da variável de interesse	170611		
Valor estimado de compartilhados	140492		
Estim. Expansão Simples	75081		
Estim. Rh Combinado	141028		
Estim. RL Clássico	127798		
Estim. Var da estim dos Comp	2674,13519882947		
Coef. de variação (CV)	0,00914927199211961		
Soma dos Resíduos	-5,11590769747272E-13		
Soma dos Comp. x Resíduos	850326,87996353		
Soma dos Amost. x resíduos	905206,823288965		
BetaZeroChapeu / BetaUmChapeu	31,4747629612634		0,748878368474899
Estim. Var. dos BetaChap Clássica	7,99163136205742		2,04602280508302E-6
Estim. covar dos BetaChap Clássica	-0,00183048179715656		-0,00183048179715656
Limite Inf. Estim: Rh Sep clássico	41914		
Estim: Rh Sep clássico	140499		
Limite Sup. Estim: Rh Sep clássico	239084		
Estim: BB 1000 Rh Sep	140195		
Lim Inf do Int de Conf Rh Sep BB (p/ NC = 90%)	135823	107	0,796097116157148
Lim Sup do Int de Conf Rh Sep BB (p/ NC = 90%)	143630	808	0,841856398555827
Estimador Rh Sep BB	0,821725789490328		
BBBEMEDIA Rh Sep BB	2,31856634586025E-16		
Epb Rh Sep BB	0,0204042259231646		
Ep JAB do Ep de Rh Sep BB	0,00207458473572636		
Estim: BS 1000 Rh Sep	139913		
Lim Inf do Int de Conf Rh Sep BS (p/ NC = 90%)	135786	109	0,795879636091785
Lim Sup do Int de Conf Rh Sep BS (p/ NC = 90%)	142758	820	0,836742948944253
Estimador Rh Sep BS	0,820069265966988		
BBBEMEDIA Rh Sep BS	8,85305283943039E-16		
Epb Rh Sep BS	0,0198647286729994		
Ep JAB do Ep de Rh Sep BS	0,00258000734466597		
Estim JK por Rh Sep	140251		
Lim Inf do Int de Conf JK Sep (p/ NC = 95%)	116922		
Lim Sup do Int de Conf JK Sep (p/ NC = 95%)	163580		
Estimador Rh Sep JK	0,822052178239253		
Ep JK de Rh Sep	0,251541547822458		
Vício JK de Rh Sep	0		
Limite Inf. Estim: Rh NV clássico	41935		
Estim: Rh NV clássico	140507		
Limite Sup. Estim: Rh NV clássico	239079		
Estim: BB 1000 Rh NV	141377		
Lim Inf do Int de Conf Rh NV BB (p/ NC = 90%)	137132	102	0,803770709230109
Lim Sup do Int de Conf Rh NV BB (p/ NC = 90%)	143880	767	0,843319610485979
Estimador Rh NV BB	0,828648232340947		
BBBEMEDIA Rh NV BB	8,57007607241167E-16		
Epb de Rh NV BB	0,0207828720112511		
Ep JAB do Ep de Rh NV BB	0,00181619068226398		
Estim: BS 1000 Rh NV	141400		
Lim Inf do Int de Conf Rh NV BS (p/ NC = 90%)	137154	102	0,803899812030152
Lim Sup do Int de Conf Rh NV BS (p/ NC = 90%)	143886	767	0,843356669044192
Estimador Rh NV BS	0,828787878844616		
BBBEMEDIA RhNV BS	3,4857099845409E-16		
Epb de Rh NV BS	0,0205915990980505		
Ep JAB do Ep de Rh NV BS	0,00215878597203291		

Estim JK por Rh NV	141280	
Lim Inf do Int de Conf JK Sep (p/ NC = 95%)	117645	
Lim Sup do Int de Conf JK Sep (p/ NC = 95%)	164914	
Estimador JK Rh NV	0,828080921665692	
Ep JK de Rh NV	0,254833387778079	
Vício JK de Rh NV	0	
Estim Reg Lin BB (p/ K > 2)	126590	
BetaZeroChapeuAst / BetaUmChapeuAst	37,8437551428879	0,741759421283377
Estim var dos BetaChapAst BB	11,5510106245933	2,95729746392979E-6
Estim covar dos BetaChapAst BB	-0,00264575700869623	-0,00264575700869623
Estim Reg Lin G/E (p/ mAh[i] > 2, K = 12)	70856	
vBetaChapZero / vBetaChapUm G/E (h = 2)	0,533906419980009	0,16533543283732
vBetaChapZero / vBetaChapUm G/E (h = 3)	1,50482412189022	0,331875176648936
vBetaChapZero / vBetaChapUm G/E (h = 4)	1,86641608960492	0,428890181033115
vBetaChapZero / vBetaChapUm G/E (h = 5)	-2,38654065504217	0,487013198631261
vBetaChapZero / vBetaChapUm G/E (h = 6)	-7,20992366410607	0,916030534351174
vBetaChapZero / vBetaChapUm G/E (h = 7)	-0,783406820840729	0,752032154533731
vBetaChapZero / vBetaChapUm G/E (h = 8)	2,33510214911209	0,891085168479696
vBetaChapZero / vBetaChapUm G/E (h = 9)	2,3331058020477	0,736291240045503
vBetaChapZero / vBetaChapUm G/E (h = 10)	3,70901639342433	0,864754098360612
vBetaChapZero / vBetaChapUm G/E (h = 11)	-1,63496533795511	0,749891681109185
vBetaChapZero / vBetaChapUm G/E (h = 12)	-0,776632302405375	0,914089347079048
vBetaChapZero / vBetaChapUm G/E (h = 13)	0,370736086175937	0,13016157989228
Estim Reg Lin JK (G/E > 2)	127671	
BetaZeroChapeuJK / BetaUmChapeuJK	34,8260157441258	0,748113359103854
Estim var dos BetaChap JK	6,56332273143882E-28	1,60237371373018E-31
Estim covar dos BetaChap JK	1,02551917678732E-29	1,02551917678732E-29
Estim Reg Lin BS (p/ K > 2)	126693	
BetaZeroChapeuAst / BetaUmChapeuAst	37,3022737608267	0,742364662623633
Estim var dos BetaChapAst BS	11,2229571092763	2,87330898357857E-6
Estim covar dos BetaChapAst BS	-0,0025706164070999	-0,0025706164070999
Estim MC Reg Lin JK (G/E > 2)	127671	
BetaZeroChapeuJK / BetaUmChapeuJK	34,8260157441258	0,748113359103854
Estim var dos BetaChap JK	87,4599872082281	0,0403586513841468
Estim covar dos BetaChap JK	1,87876745069708	1,87876745069708
Estim MC RL Clássico	127800	
Estim RL Clássico	127798	

APROXIMAÇÕES MONTE CARLO DAS ESTIMATIVAS BOOTSTRAP BALANCEADO - 500 - ITERAÇÕES

Estim MC Rh Sep Clássico	140250	
Estim MC BB Rh Sep	140251	
Estim BB Rh Sep	140251	
Lim Inf do Int de Conf Rh Sep BB (p/ NC = 90%)	139944	
Lim Sup do Int de Conf Rh Sep BB (p/ NC = 90%)	140559	
Estimador Rh Sep BB	0,822052178239247	
Epb de Rh Sep	0,0205703820170956	
Estim MC Jack Rh Sep BB	140273	
Estim Jack Rh Sep BB	140251	
Estimador Jack Rh Sep BB	0,822052178239247	
Epb Jack de Rh Sep BB	0,0208423328724743	
Estim MC Rh NV Clássico	141281	
Estim MC BB Rh NV	141280	
Estim BB Rh NV	141280	
Lim Inf do Int de Conf MC Rh NV BB	140968	
Lim Sup do Int de Conf MC Rh NV BB	141592	
Estimador Rh NV BB	0,828080921665688	
Epb de Rh NV	0,0208667915628001	

Estim Jack Rh NV BB	141280	
Estim Jack Rh NV BS	141280	
Estimador de Jack Rh NV BS	0,828080921665688	
Epb Jack de Rh NV BS	0,0210576414606147	
Estim MC Reg Lin BB	127795	
Estim Reg Lin BS	127798	
BetaZeroChapeuAst / BetaUmChapeuAst	31,4777981565342	0,748874975882438
Estim var dos BetaChapAst BS	7,99163136205743	2,04602280508303E-6
Estim covar dos BetaChapAst BS	-0,00183048179715653	-0,00183048179715653
Estim MC Rh Clássico	140250	
Estim MC Rh Sep BS	140252	
Estim BS Rh Sep	140499	
Lim Inf do Int de Conf Rh Sep BS (p/ NC = 90%)	139949	
Lim Sup do Int de Conf Rh Sep BS (p/ NC = 90%)	140553	
Estimador Rh Sep BS	0,822053252098708	
Epb de Rh Sep	0	
Estim Jack Rh Sep BS	140256	
Estimador Jack Rh Sep BS	0,82208239207027	
Epb Jack de Rh Sep BS	0,0201254752030092	
Estim Jack Rh Sep MC	140273	
Estim MC Rh NV Clássico	141281	
Estim MC BS Rh NV	141278	
Estim BS Rh NV	141278	
Lim Inf do Int de Conf Rh NV BS (p/ NC = 90%)	140973	
Lim Sup do Int de Conf Rh NV BS (p/ NC = 90%)	141583	
Estimador Rh NV BS	0,828070544238577	
Epb de Rh NV	0,0204141119700864	
Estim MC Jack Rh NV BS	141300	
Estim Jack Rh NV BS	141283	
Estimador Jack Rh NV BS	0,828098985285107	
Epb Jack de Rh NV BS	0,0203919734018822	
Estim MC Reg Lin	127795	
Estim Reg Lin BS	127800	
BetaZeroChapeuAst / BetaUmChapeuAst	31,4489909768619	0,74889054358688
Estim var dos BetaChapAst BS	7,99163136205743	2,04602280508303E-6
Estim covar dos BetaChapAst BS	-0,00183048179715653	-0,00183048179715653
Estim MC Reg Lin Jack (G/E > 2).	127671	
Estim Reg Lin Jack (G/E > 2)	127691	
BetaZeroChapeuJK / BetaUmChapeuJK BS	34,8313744171127	0,748228471176211
Estim var dos BetaChap Jack BS	0,000344584514171088	1,59009470430129E-7
Estim covar dos BetaChap Jack BS	7,40217543136935E-6	7,40217543136935E-6

ANEXO 5 – Gráficos gerados em Programa Microsoft Excel a partir dos resultados obtidos com o Programa EEC.

Os resultados obtidos para cada uma das dez variáveis estudadas foram tabelados e então transformados em gráficos. Nas páginas seguintes apresentam-se os gráficos referentes aos resultados de cada uma das variáveis estudadas.

GRÁFICO 01.01: Médias dos erros relativos percentuais em valores absolutos de 25 rodadas para simulações de 75_76%, 75_80% e 75_90%, de compartilhados. Variável: NC_AES.

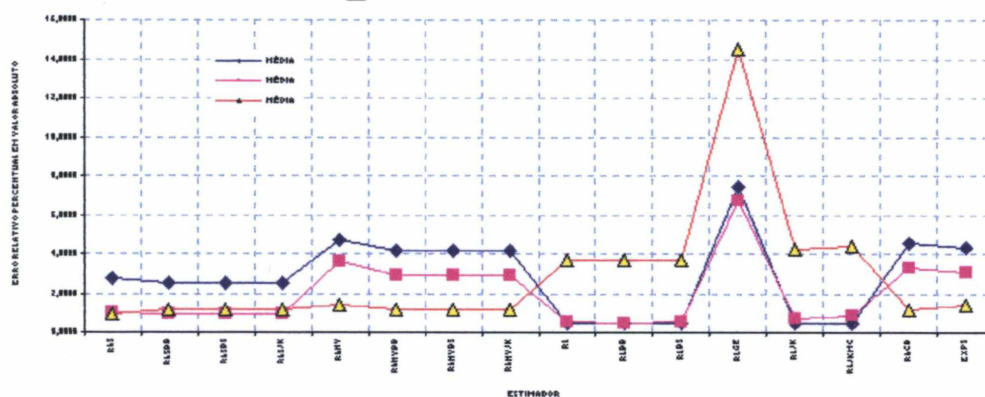


GRÁFICO 01.02: Erro Padrão dos erros relativos percentuais em valores absolutos das estimativas populacionais das 75 rodadas de simulações. Variável: NC_AES.

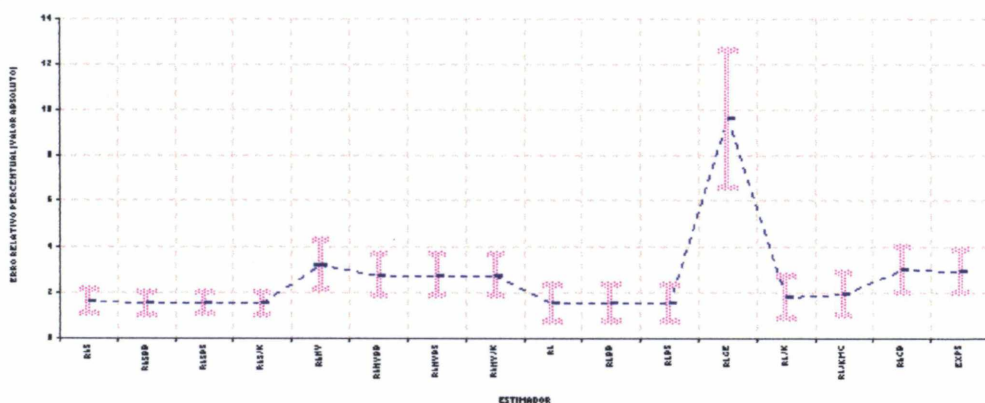


GRÁFICO 01.03: Amplitude de variação dos erros relativos percentuais em valores absolutos das estimativas populacionais das 75 simulações. Variável: NC_AES.

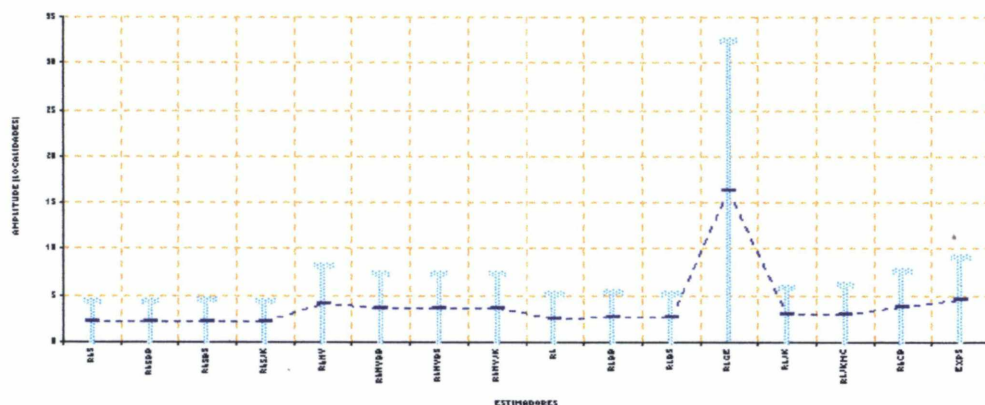


GRÁFICO 01.04: Média do número de localidades estimadas com erro relativo percentual em valor absoluto menor ou igual a 5, para 25 simulações. Variável: NC_AES.

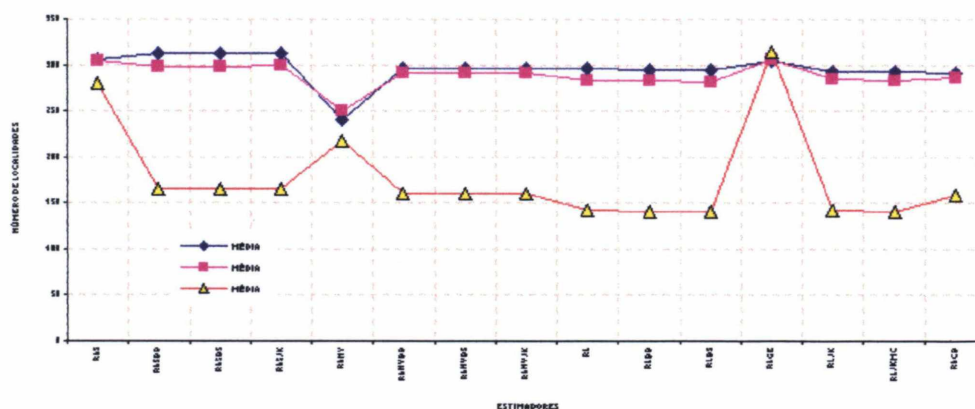


GRÁFICO 01.05: Erro Padrão do número médio de localidades estimadas com erro relativo percentual em valor absoluto menor ou igual a 5, para as 75 simulações. Variável: NC_AES.

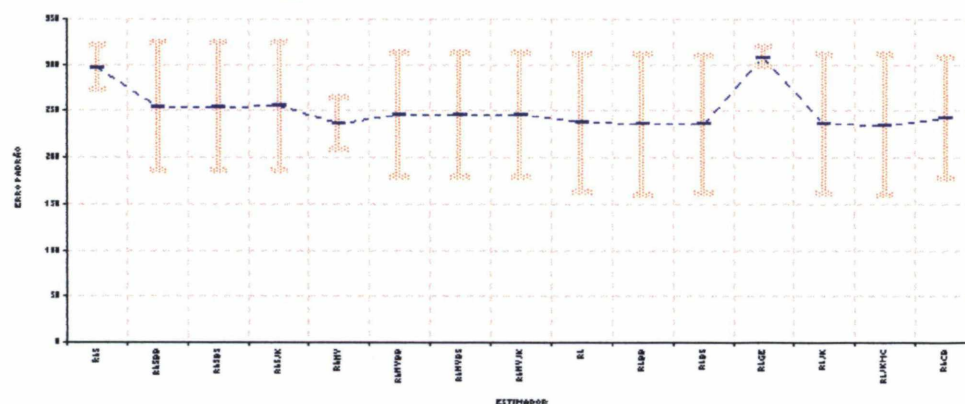


GRÁFICO 01.06: Amplitude de variação do número de localidades estimadas com erros relativos percentuais em valores absolutos menores ou iguais a 5, para as 75 simulações. Variável: NC_AES.

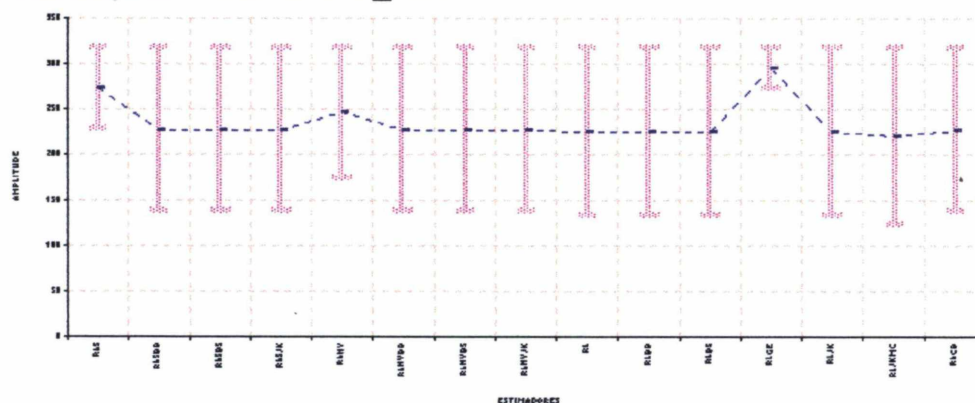


GRÁFICO 01.07: Cobertura do valor simulado das estimativas populacionais, em cada uma das 25 simulações. Variável: NC_AES.

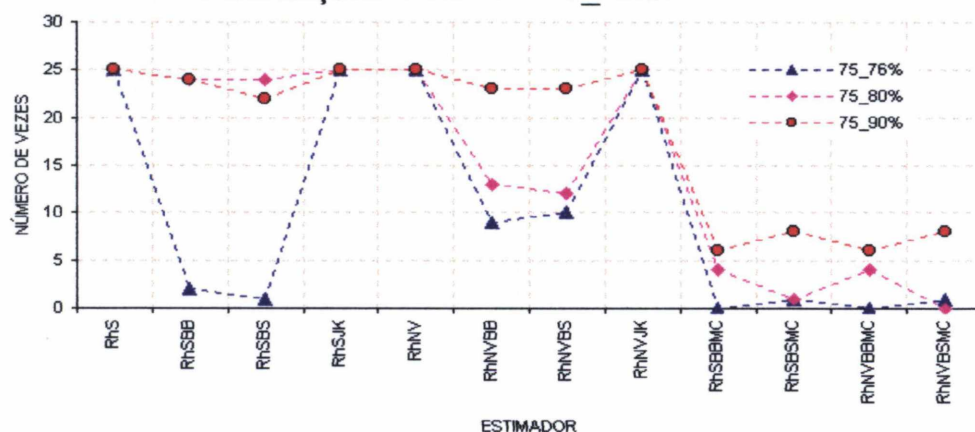


GRÁFICO 01.08: Erro Padrão do número de vezes que o intervalo cobre o valor simulado das estimativas populacionais, em 75 simulações. Variável: NC_AES.

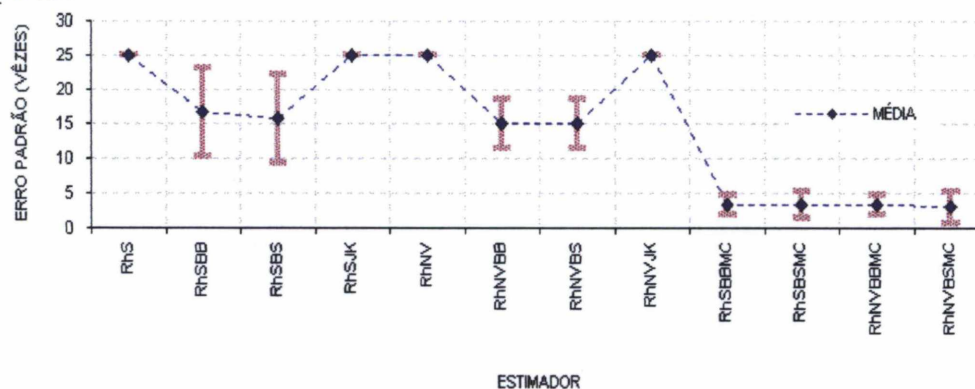


GRÁFICO 01.09: Amplitude de variação do número de vezes que o intervalo cobre o valor simulado, da estimativa populacional, para as 75 simulações. Variável: NC_AES.

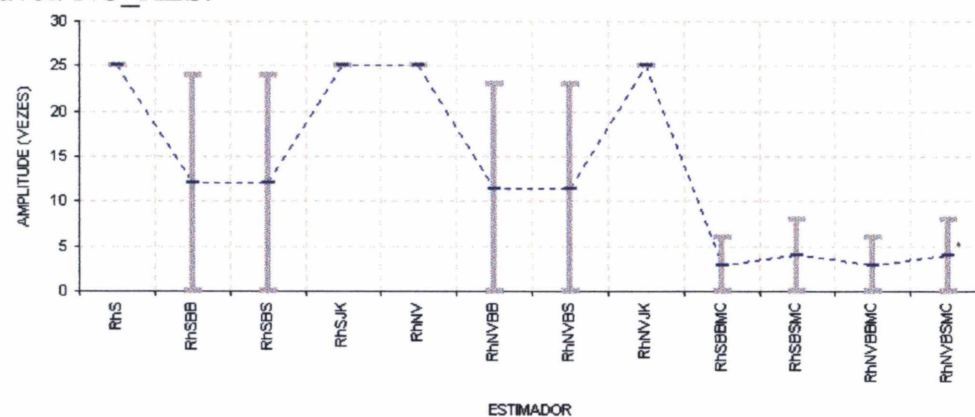


GRÁFICO 02.04: Média do número de localidades estimadas com erro relativo percentual em valor absoluto menor ou igual a 5, para as 25 simulações. Variável: NC_COM.

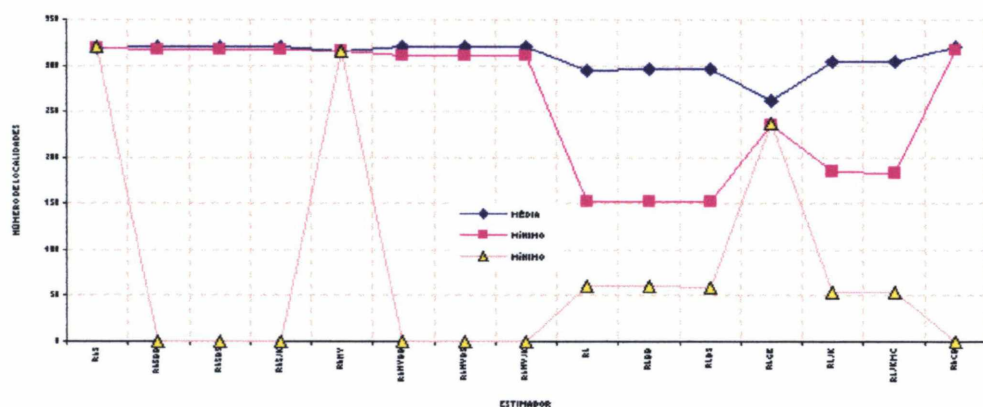


GRÁFICO 02.05: Erro Padrão do número de localidades estimadas com erro relativo percentual em valor absoluto menor ou igual a 5, para as 75 simulações. Variável: NC_COM.

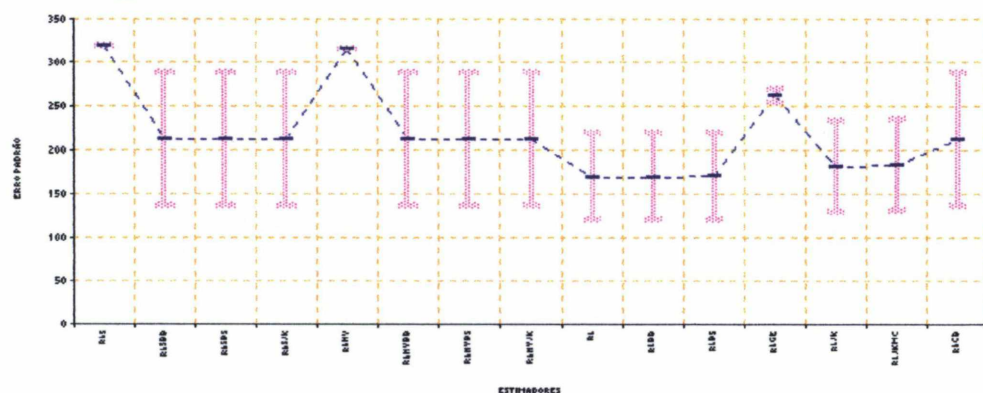


GRÁFICO 02.06: Amplitude de variação do número de localidades estimadas com erro relativo percentual menor ou igual a 5. Variável: NC_COM.

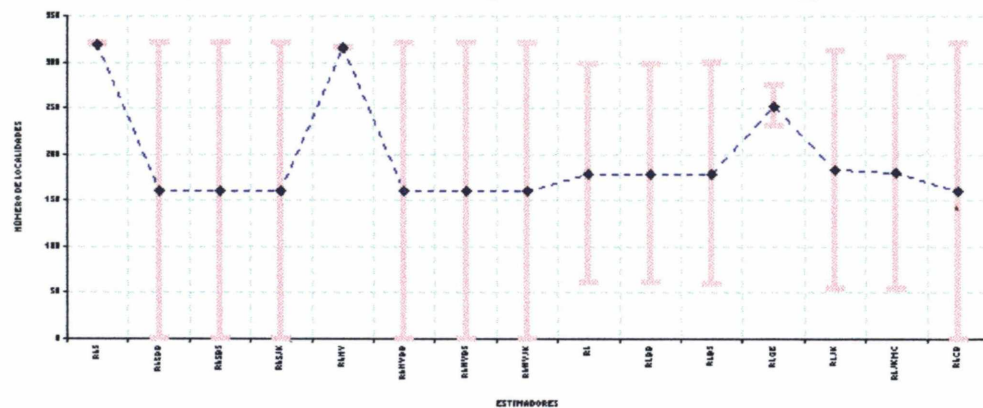


GRÁFICO 02.07: Número de vezes que o intervalo definido pelos limites inferior e superior das estimativas de cada estimador, cobre o valor simulado populacional, em cada grupo de 25 rodadas de simulações, para as proporções de compartilhados: 75_76%, 75_80% e 75_90%. Variável: NC_COM.

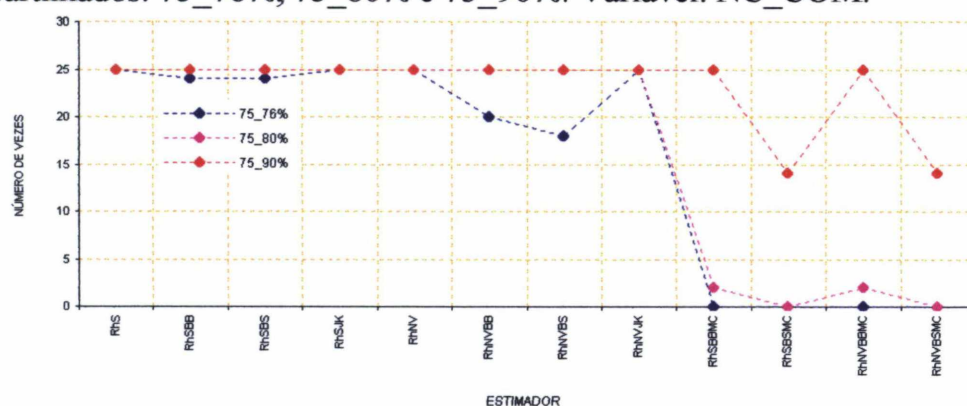


GRÁFICO 02.08: Erro padrão do número máximo de vezes que o intervalo definido pelos limites inferior e superior das estimativas de cada estimador cobre o valor simulado populacional, em cada grupo de 25 rodadas de simulações, para as proporções de compartilhados: 75_76%, 75_80% e 75_90%. Variável: NC_COM.

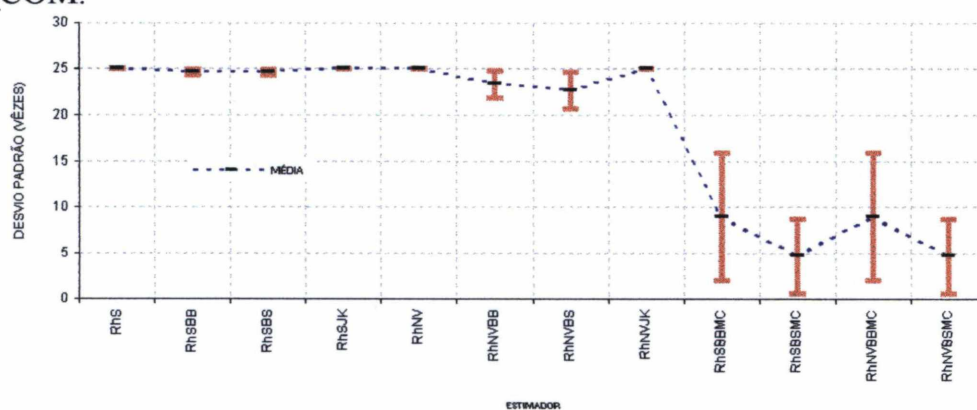


GRÁFICO 02.09: Amplitude de variação dos números máximos de vezes que o intervalo definido pelos limites inferior e superior das estimativas de cada estimador, cobre o valor simulado populacional, em cada grupo de 25 rodadas de simulações, para as proporções de compartilhados: 75_76%, 75_80% e 75_90%. Variável: NC_COM.

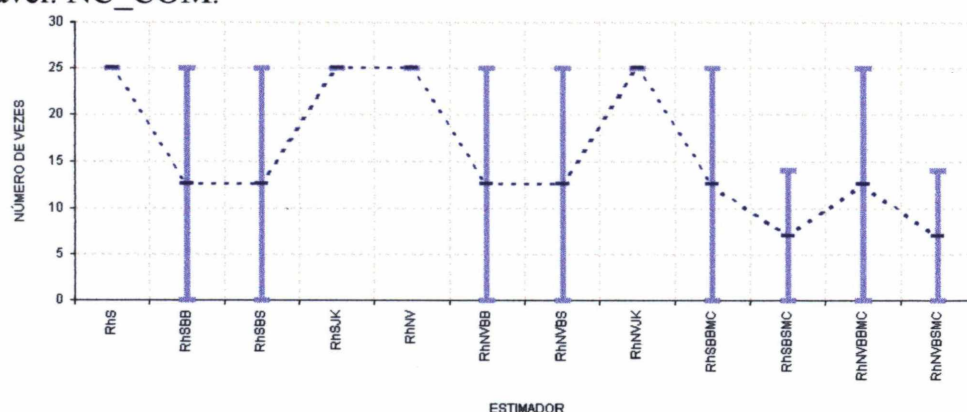


GRÁFICO 03.01: Média dos erros relativos percentuais em valores absolutos das estimativas populacionais, de 25 simulações, para: 75_76%, 75_80% e 75_90%, respectivamente. Variável: NC_ILU.

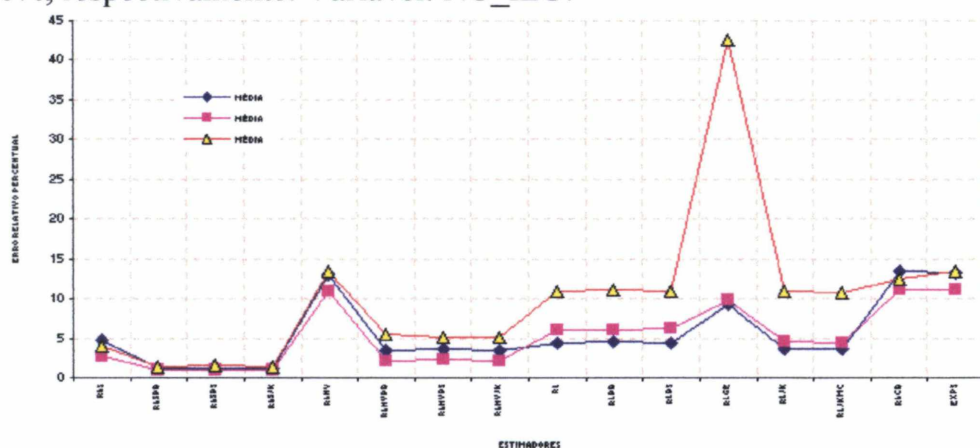


GRÁFICO 03.02: Erro Padrão dos erros relativos percentuais em valores absolutos das estimativas populacionais, para as 75 simulações. Variável: NC_ILU.

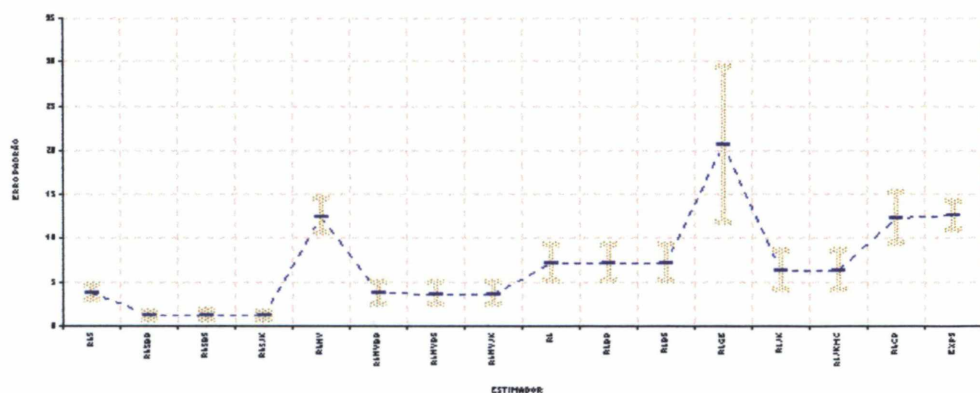


GRÁFICO 03.03: Amplitude de variação dos erros relativos percentuais em valores absolutos das estimativas populacionais, para as 75 simulações. Variável: NC_ILU.

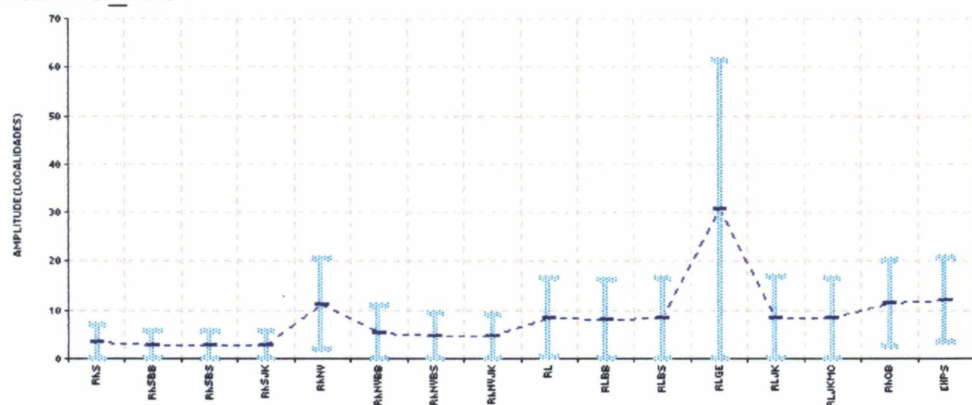


GRÁFICO 03.04: Número de localidades estimadas com erro relativo percentual em valor absoluto menor ou igual a 5. Variável: NC_ILU.

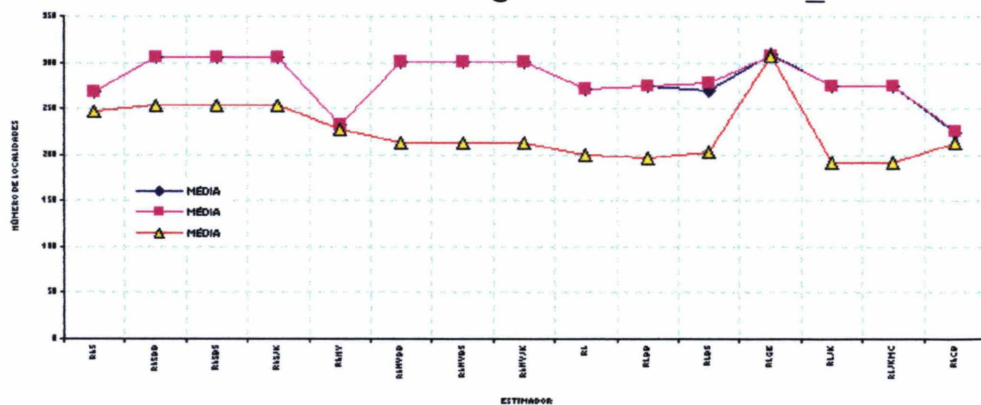


GRÁFICO 03.05: Erro Padrão do número de localidades estimativas com erro relativo percentual em valor absoluto menor ou igual a 5, para as 75 simulações. Variável: NC_ILU.

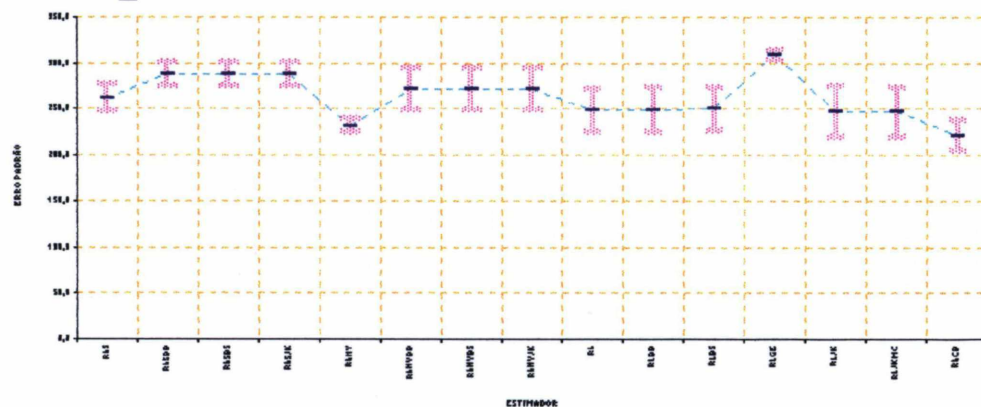


GRÁFICO 03.06: Amplitude de variação do número de localidades estimadas com erro relativo percentual em valor absoluto menor ou igual a 5. Variável: NC_ILU.

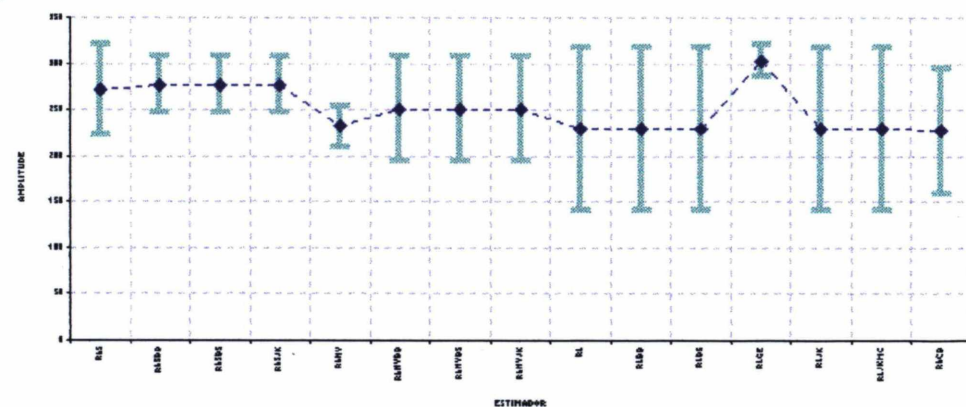


GRÁFICO 03.07: Cobertura do valor simulado das estimativas populacionais, em cada uma das 25 simulações. Variável: NC_ILU.

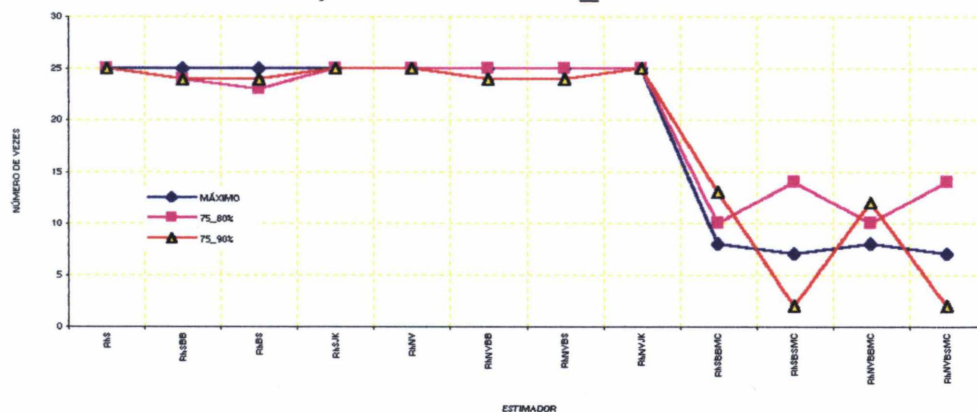


GRÁFICO 03.08: Erro Padrão do número de vezes que o intervalo cobre os valores simulados das estimativas populacionais, em 75 simulações. Variável: NC_ILU.

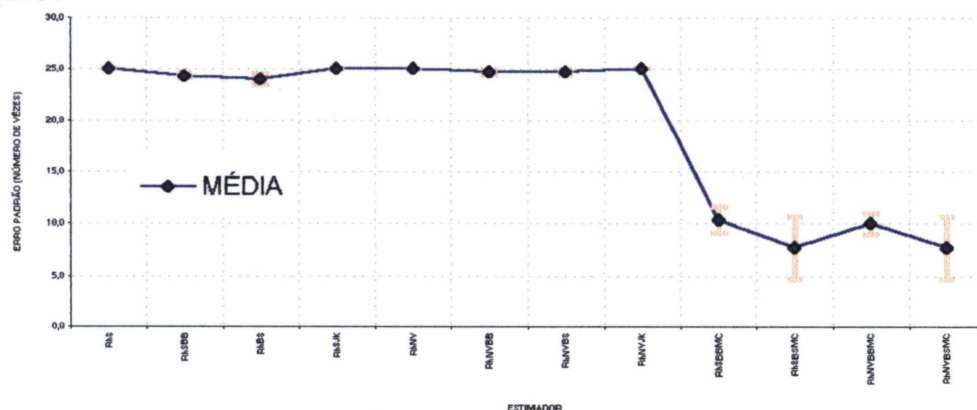


GRÁFICO 03.09: Amplitude de variação do número de vezes que o intervalo cobre o valor simulado da estimativa populacional, considerando-se os três grupos de simulações, 75_76%, 75_80% e 75_90% simultaneamente, cada um com 25 simulações. Variável: NC_ILU.

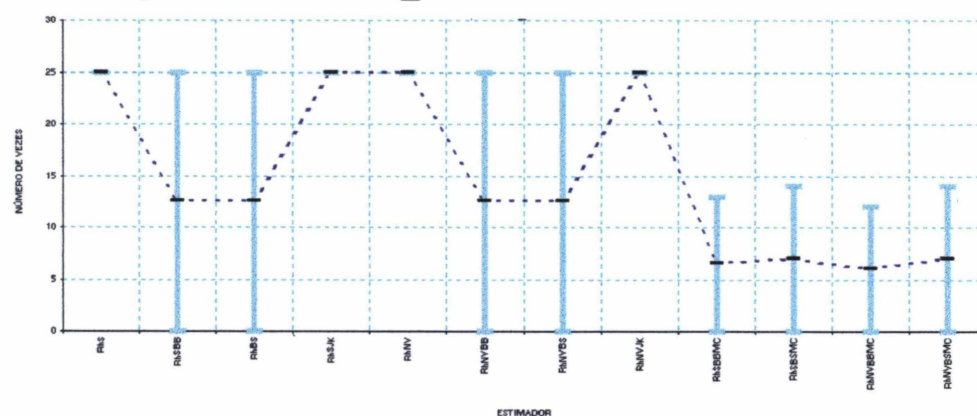


GRÁFICO 03.10: Média dos erros relativos percentuais em valores absolutos das estimativas populacionais, de 25 simulações, para: 75_76%, 75_80% e 75_90%, respectivamente. Variável: NC_ILU, eliminado o primeiro estrato.

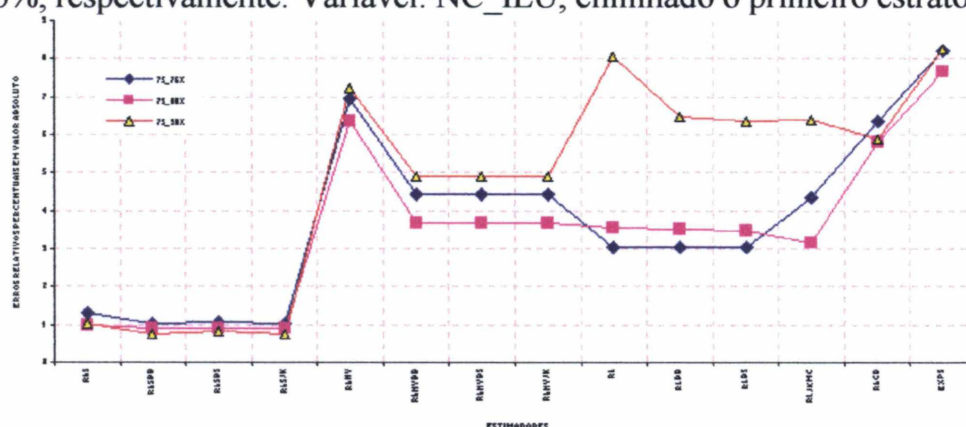


GRÁFICO 03.11: Erro Padrão dos erros relativos percentuais em valores absolutos das estimativas populacionais, para as 75 simulações. Variável: NC_ILU, eliminado o primeiro estrato.

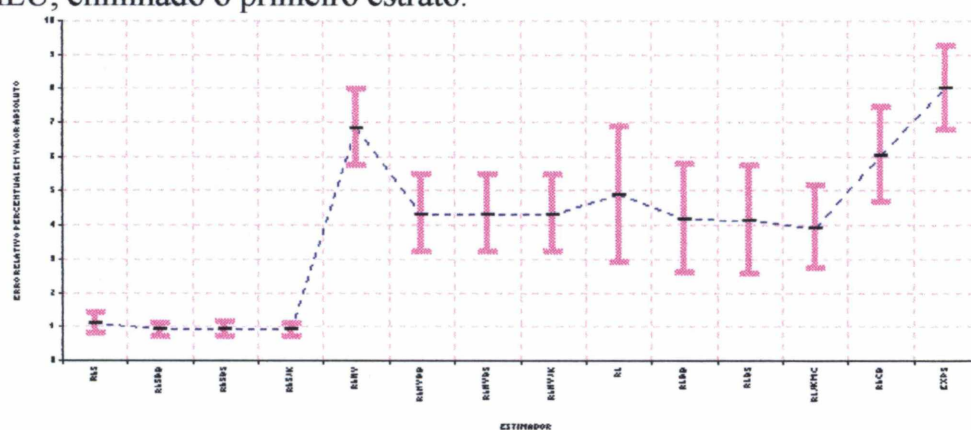


GRÁFICO 03.12: Amplitude de variação dos erros relativos percentuais em valores absolutos das estimativas populacionais, para as 75 simulações. Variável: NC_ILU, eliminado o primeiro estrato.

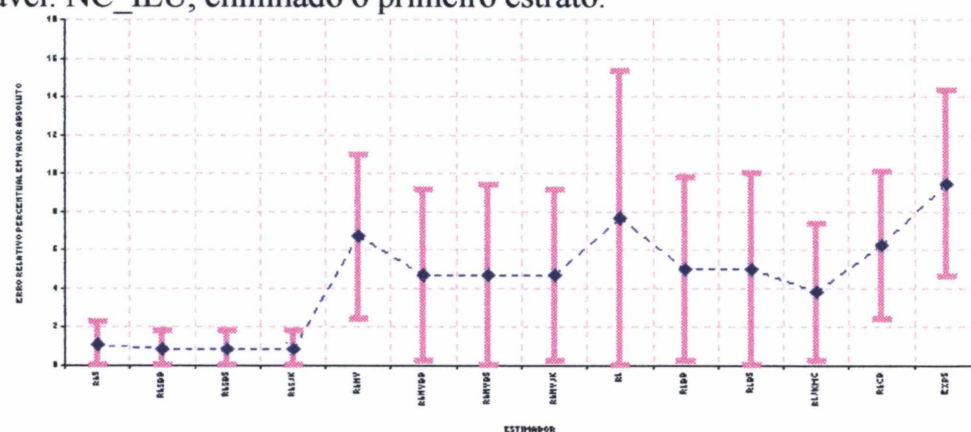


GRÁFICO 03.13: Número médio de localidades estimadas com erro relativo percentual em valor absoluto menor ou igual a 5. Variável: NC_ILU, eliminado o primeiro estrato.

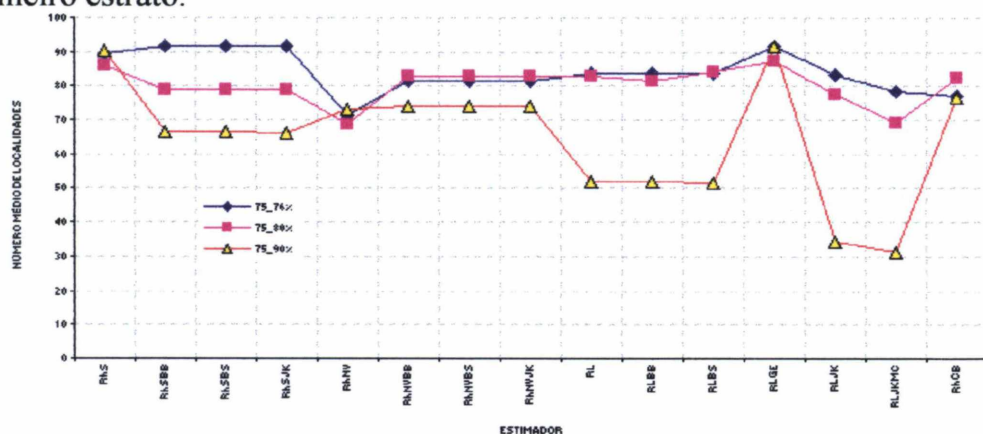


GRÁFICO 03.14: Erro Padrão do número médio de localidades estimadas com erro relativo percentual em valor absoluto menor ou igual a 5. Variável: NC_ILU, eliminado o primeiro estrato.

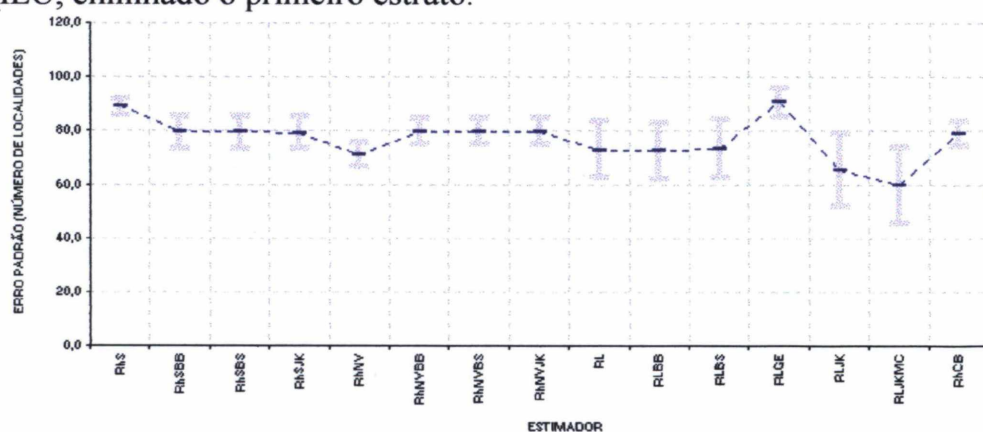


GRÁFICO 03.15: Amplitude de variação do número de localidades estimadas com erro relativo percentual em valor absoluto menor ou igual a 5. Variável: NC_ILU, eliminado o primeiro estrato.

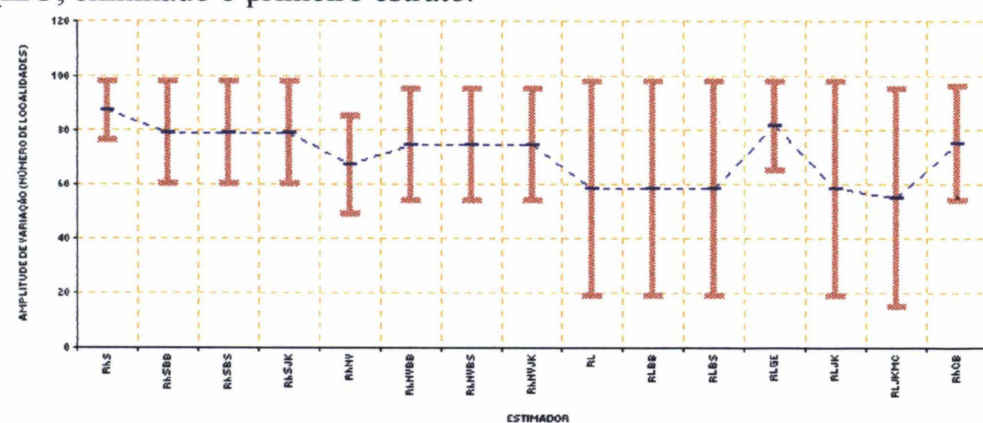


GRÁFICO 03.16: Numero de vezes que o valor simulado populacional é coberto pelas estimativas populacionais, em dez rodadas. Variável: NC_ILU, eliminado o primeiro estrato.

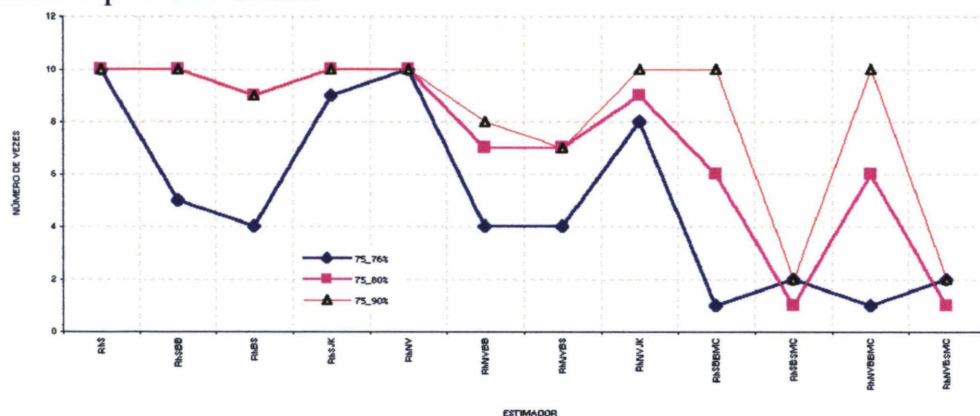


GRÁFICO 03.17: Erro Padrão dos números máximos de vezes que o valor simulado populacional é coberto pelas estimativas populacionais. Variável: NC_ILU, eliminado o primeiro estrato.

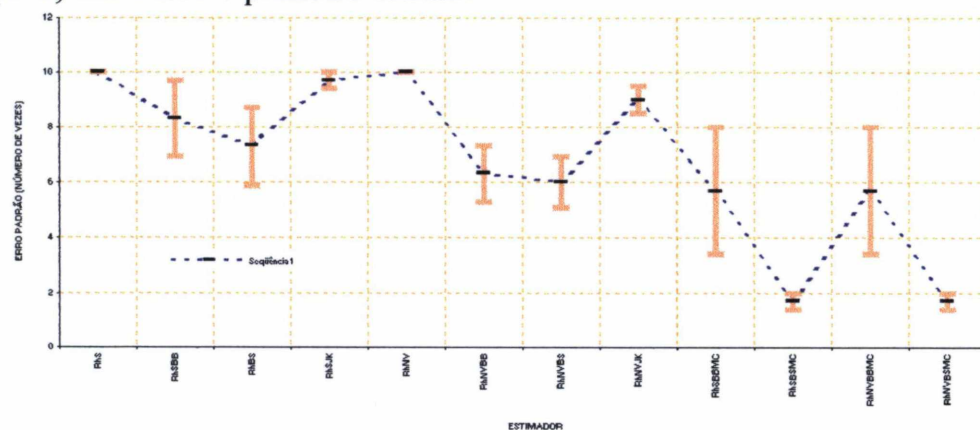


GRÁFICO 03.18: Amplitude de variação do número de vezes que o valor simulado populacional é coberto pelas estimativas populacionais. Variável : NC_ILU, eliminado o primeiro estrato.

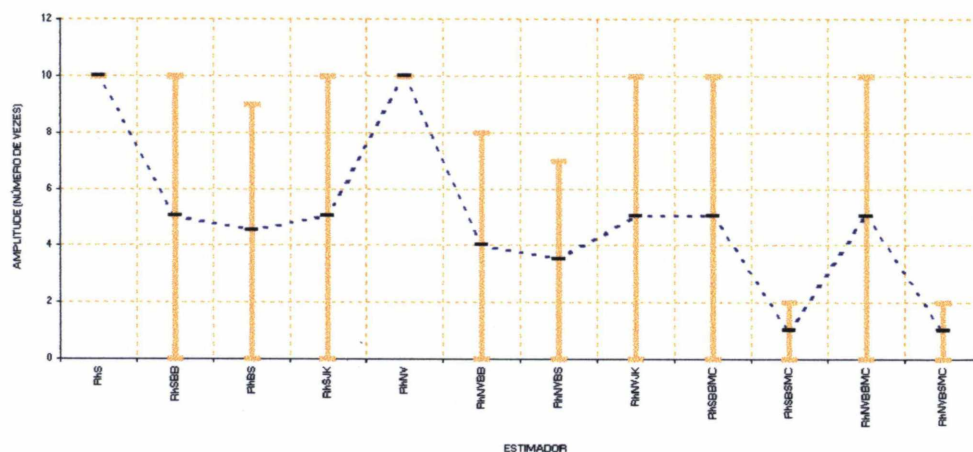


GRÁFICO 04.01: Médias dos erros relativos percentuais em valores absolutos, das estimativas populacionais, de 25 simulações, para: 75_76%, 75_8-% e 75_90%, respectivamente. Variável: NC_IND.

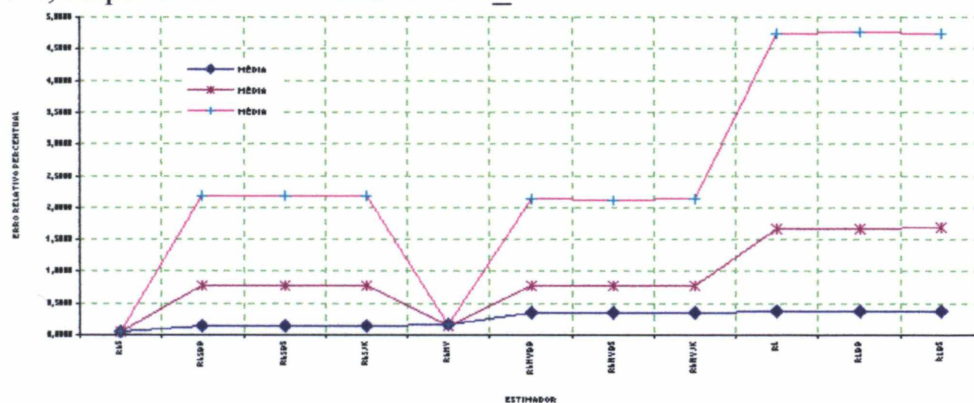


GRÁFICO 04.02: Erro Padrão dos erros relativos percentuais em valores absolutos das estimativas populacionais de 75 simulações. Variável: NC_IND.

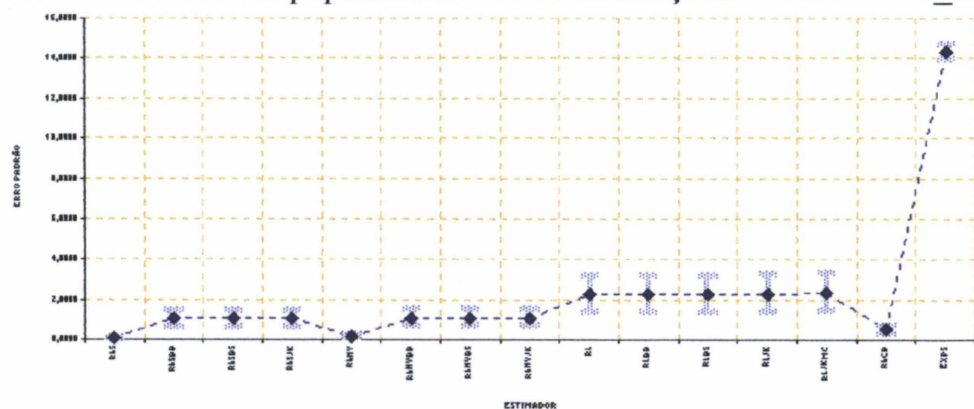


GRÁFICO 04.03: Amplitude de variação dos erros relativos percentuais em valores absolutos das estimativas populacionais das 75 simulações. Variável: NC_IND.

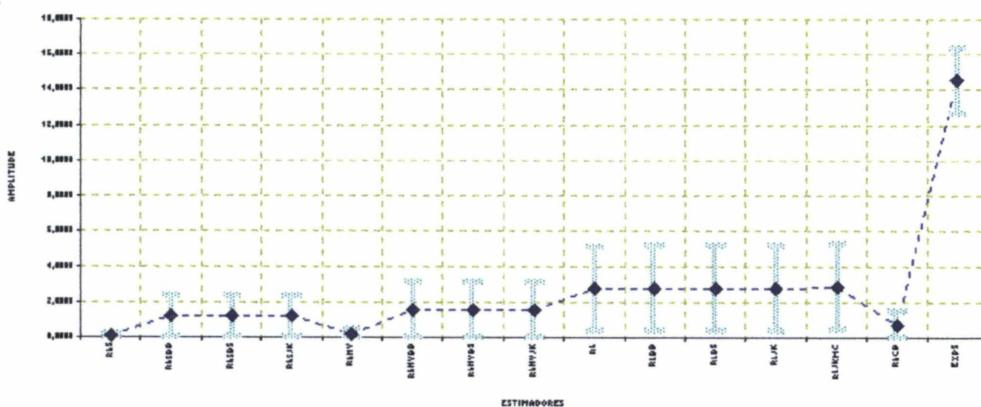


GRÁFICO 04.07: Cobertura do valor simulado das estimativas populacionais, em cada uma das 20 simulações. Variável: NC_IND.

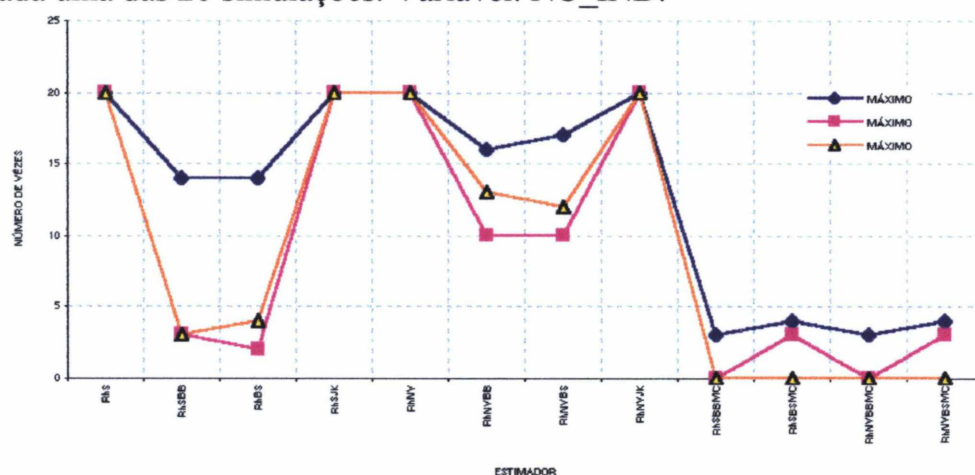


GRÁFICO 04.08: Erro Padrão do número de vezes que o intervalo cobre os valores simulados, das estimativas populacionais, em 60 simulações. Variável : NC_IND.

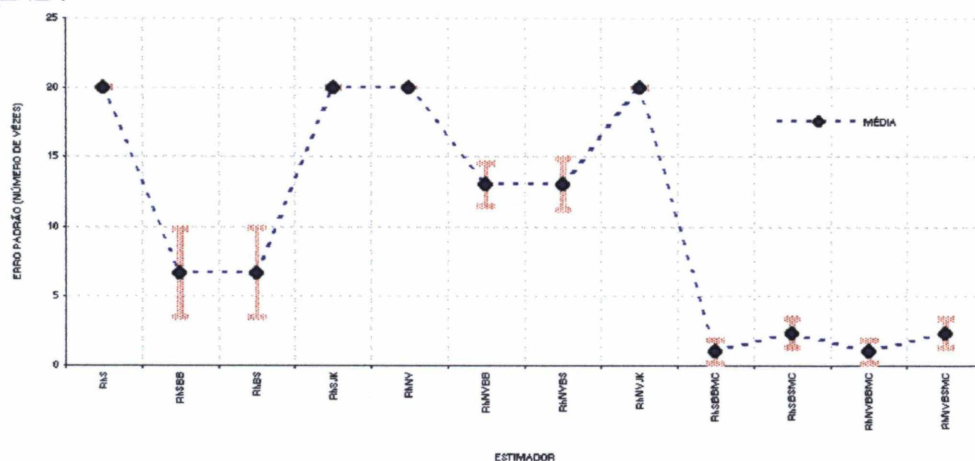


GRÁFICO 04.09: Amplitude de variação do número de vezes que o intervalo cobre os valores simulados das estimativas populacionais, para 60 simulações. Variável: NC_IND.

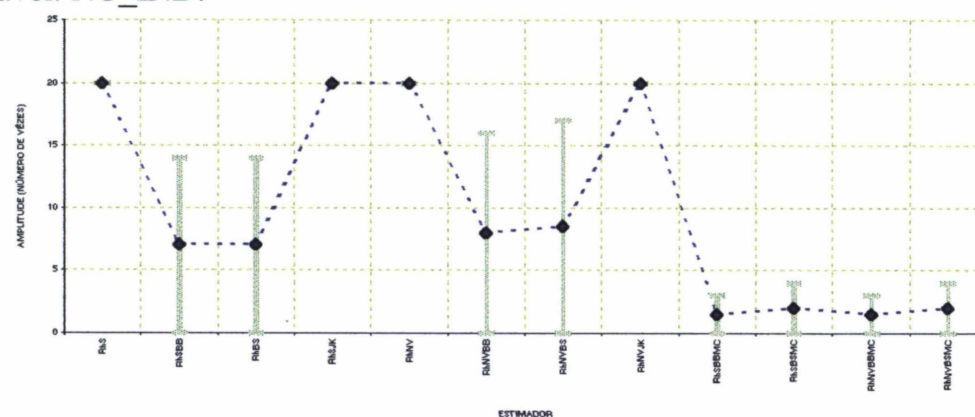


GRÁFICO 05.01: Médias dos erros relativos percentuais em valores absolutos das estimativas populacionais de 25 simulações, para: 75_76%, 75_80% e 75_90%, respectivamente. Variável: NC_POD.

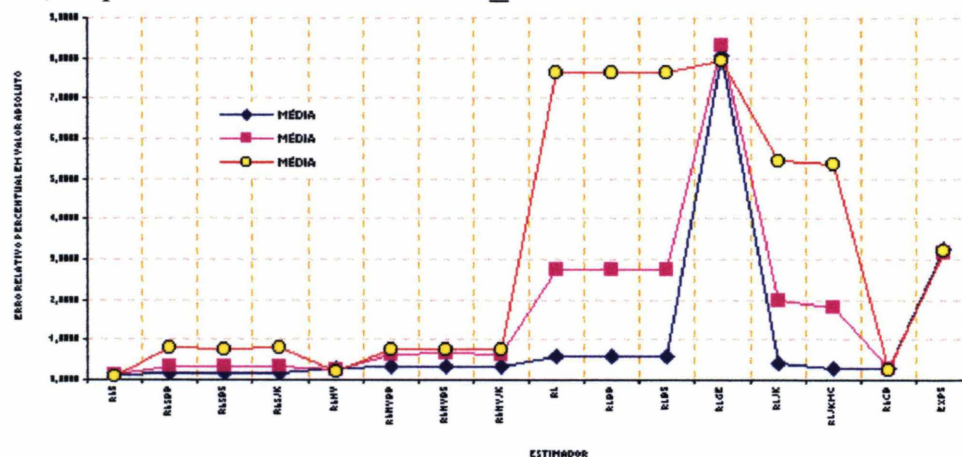


GRÁFICO 05.02: Erro Padrão dos erros relativos percentuais em valores absolutos das estimativas populacionais de 75 simulações. Variável: NC_POD.

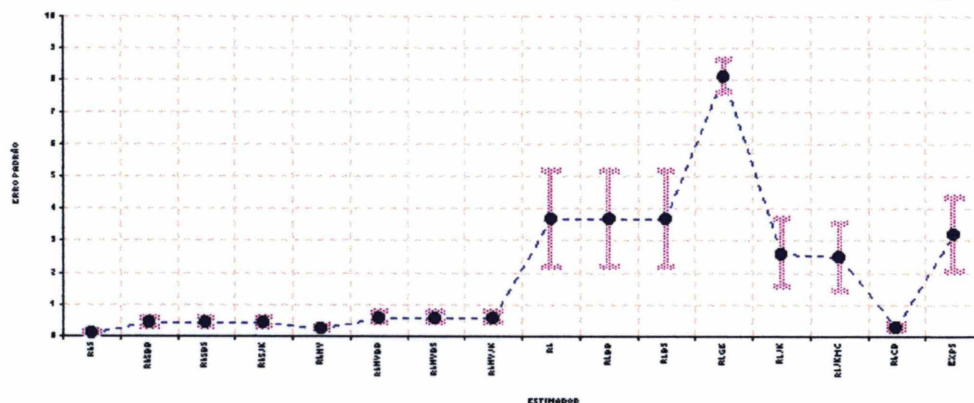


GRÁFICO 05.03: Amplitude de variação dos erros relativos percentuais em valores absolutos das estimativas populacionais das 75 simulações. Variável: NC_POD.

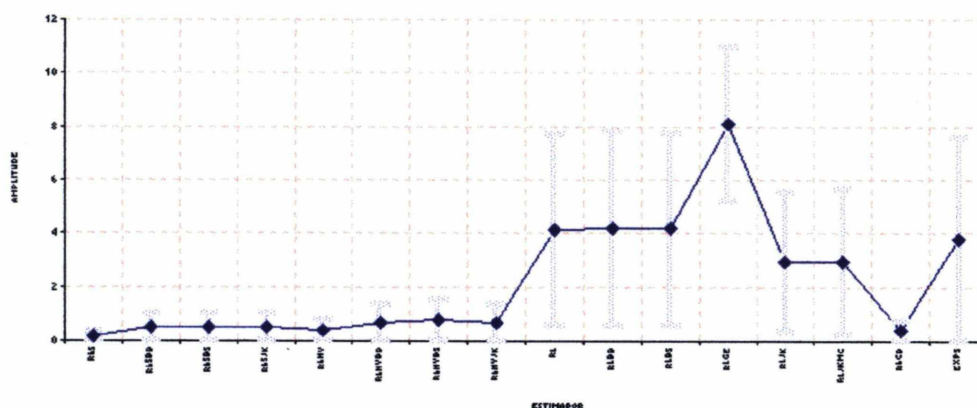


GRÁFICO 05.04: Número de localidades estimadas com erro relativo percentual em valor absoluto menor ou igual a 5. Variável: NC_POD.

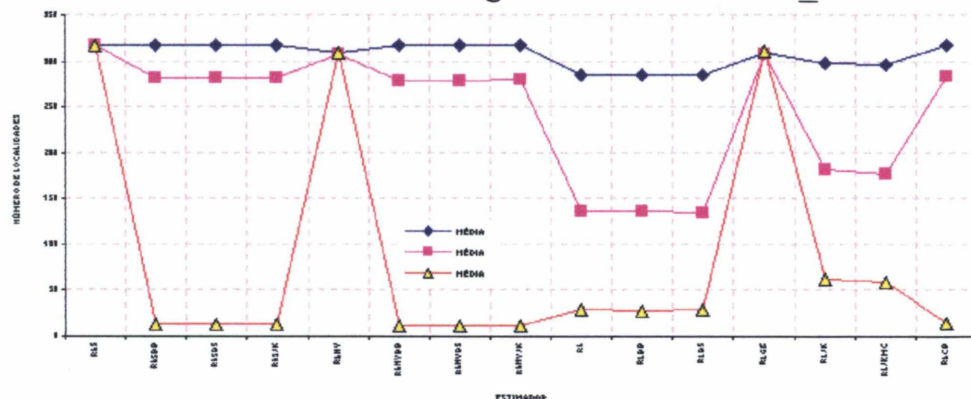


GRÁFICO 05.05: Erro Padrão do número médio de localidades estimadas com erro relativo percentual menor ou igual a 5, para as 75 simulações. Variável: NC_POD.

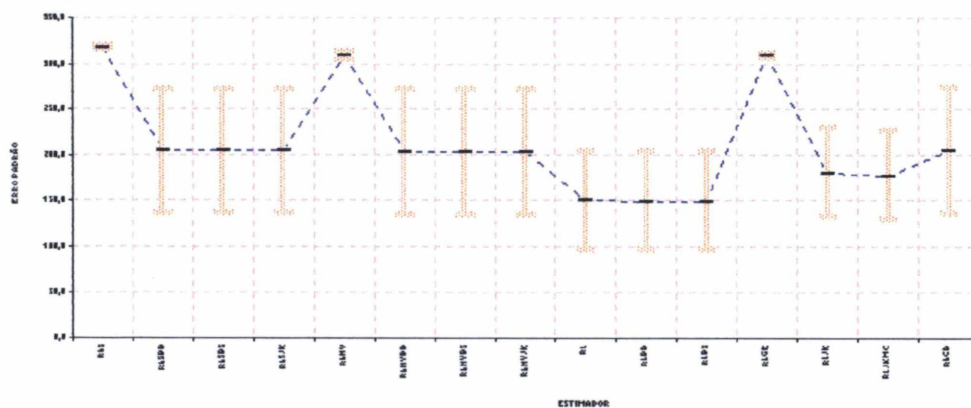


GRÁFICO 05.06: Amplitude de variação do número de localidades estimadas com erro relativo percentual em valor absoluto menor ou igual a 5, em 75 simulações. Variável: NC_POD.

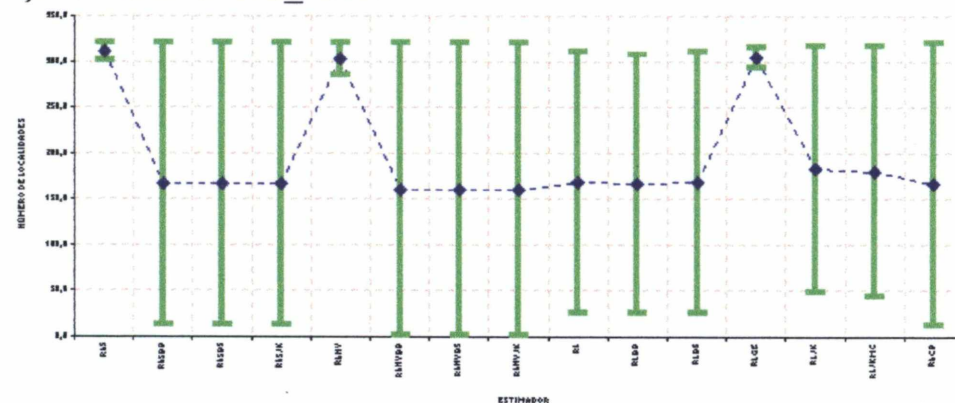


GRÁFICO 05.07: Cobertura do valor simulado das estimativas populacionais em valores absolutos, em cada uma das 21 simulações. Variável: NC_POD.

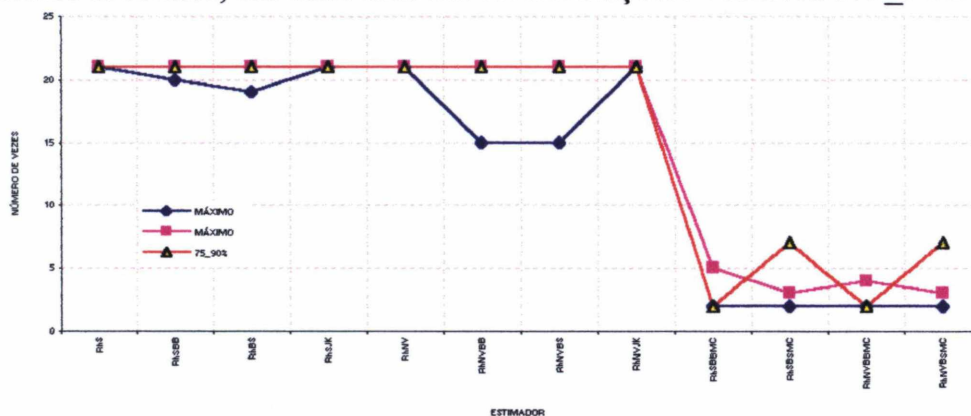


GRÁFICO 05.08: Erro Padrão do número de vezes que o intervalo cobre os valores simulados das estimativas populacionais, em 63 simulações. Variável: NC_POD.

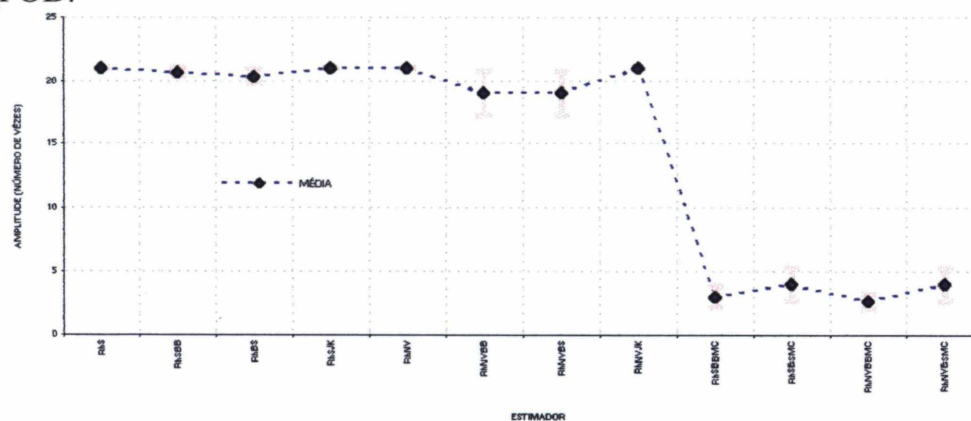


GRÁFICO 05.09: Amplitude de variação do número de vezes que o intervalo cobre o valor simulado da estimativa populacional, para as 63 simulações. Variável: NC_POD.

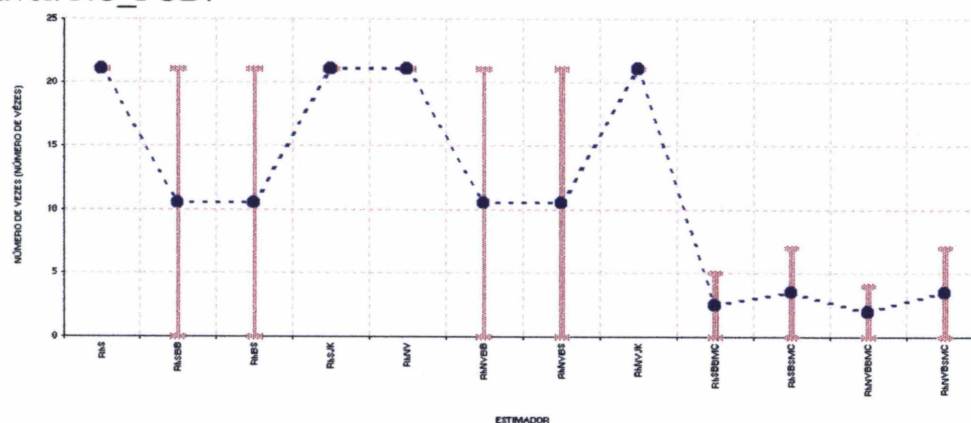


GRÁFICO 06.01: Médias dos erros relativos percentuais em valores absolutos, das estimativas populacionais, para cada grupo de simulações: 75_76%, 75_80% e 75_90% respectivamente, cada uma com 25 simulações cada. Variável: NC_PRO.

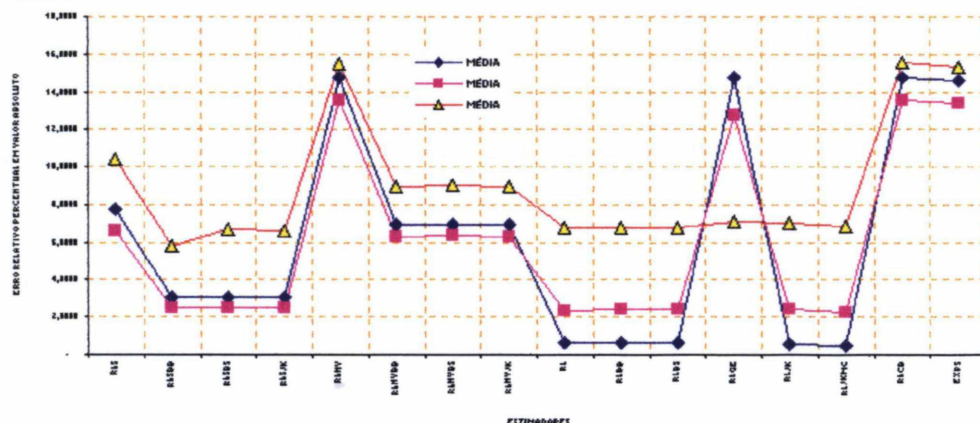


GRÁFICO 06.02: Erro Padrão dos erros relativos percentuais em valor absoluto das estimativas populacionais, das 75 simulações, consideradas simultaneamente. Variável: NC_PRO.

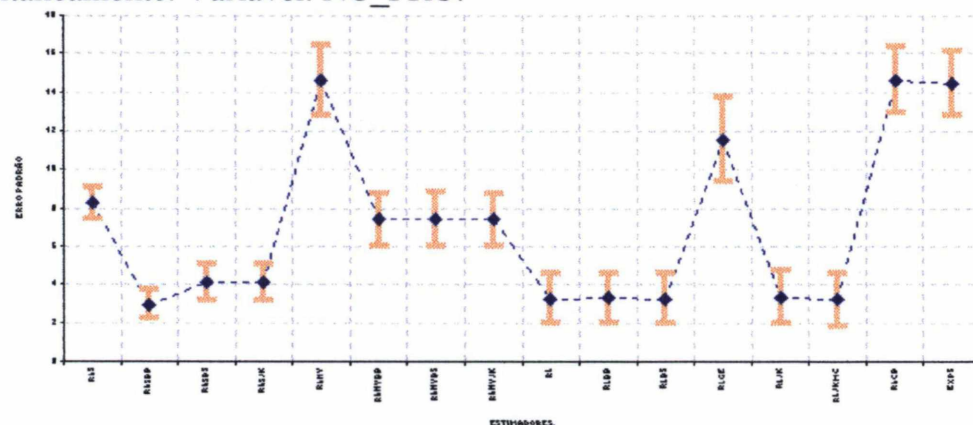


GRÁFICO 06.03: Amplitude de variação dos erros relativos percentuais em valores absolutos das estimativas populacionais, das 75 simulações consideradas simultaneamente. Variável: NC_PRO.

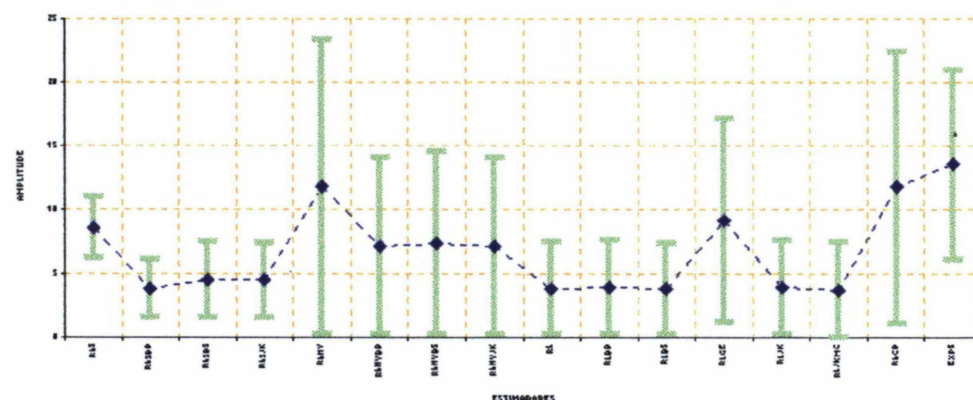


GRÁFICO 06.07: Cobertura dos valores simulados das estimativas populacionais, em cada grupo de simulação; 75_76%, 75_80% e 75_90%, respectivamente, com 25 simulações cada. Variável: NC_PRO.

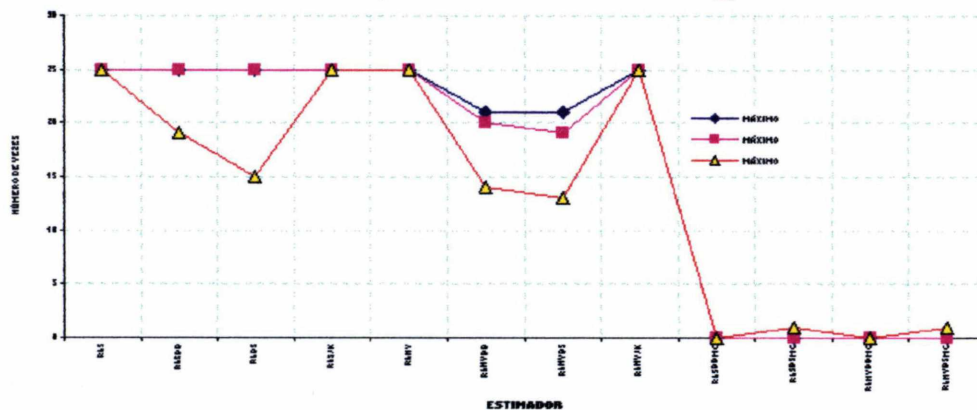


GRÁFICO 06.08: Erro Padrão do número médio de vezes que o intervalo cobre os valores simulados das estimativas populacionais, em 63 simulações. Variável: NC_PRO.

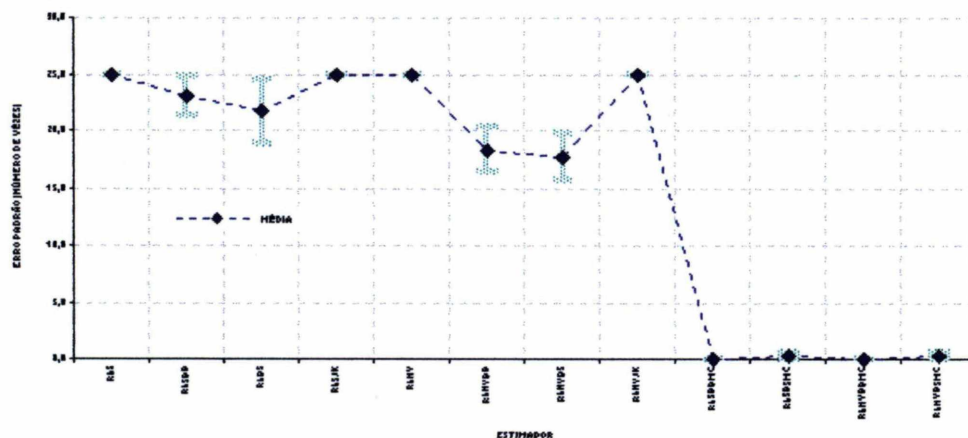


GRÁFICO 06.09: Amplitude de variação do número de vezes que o intervalo cobre o valor simulado da estimativa populacional, para as 75 simulações. Considerou-se o mínimo e o máximo entre os três grupos de simulações. Variável: NC_PRO.

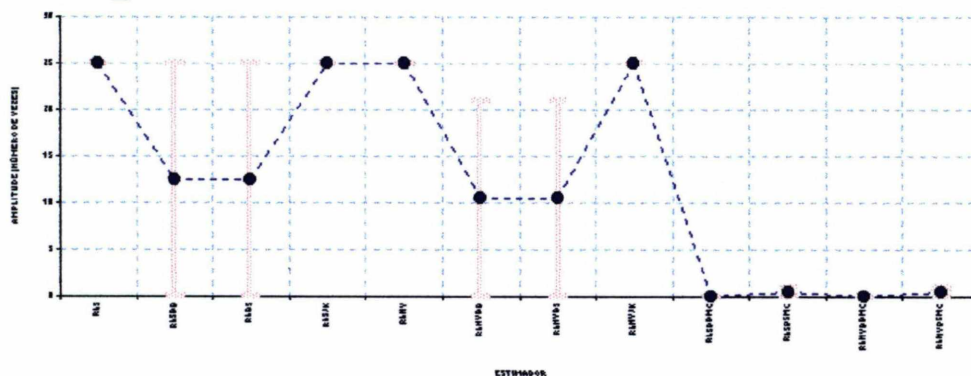


GRÁFICO 06.10: Médias dos erros relativos percentuais em valores absolutos das estimativas populacionais. Variável: NC_PRO, eliminado o primeiro estrato.

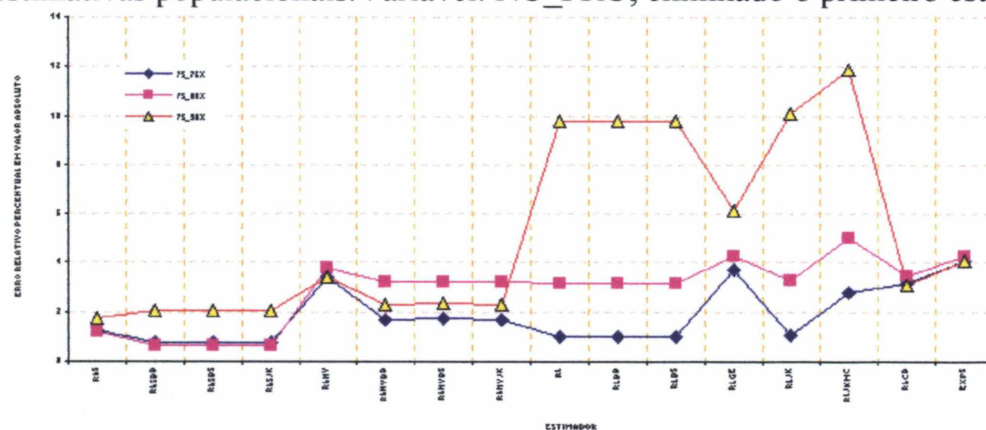


GRÁFICO 06.11: Erro Padrão da média de variação dos erros relativos percentuais em valores absolutos das estimativas populacionais. Variável: NC_PRO, eliminado o primeiro estrato.

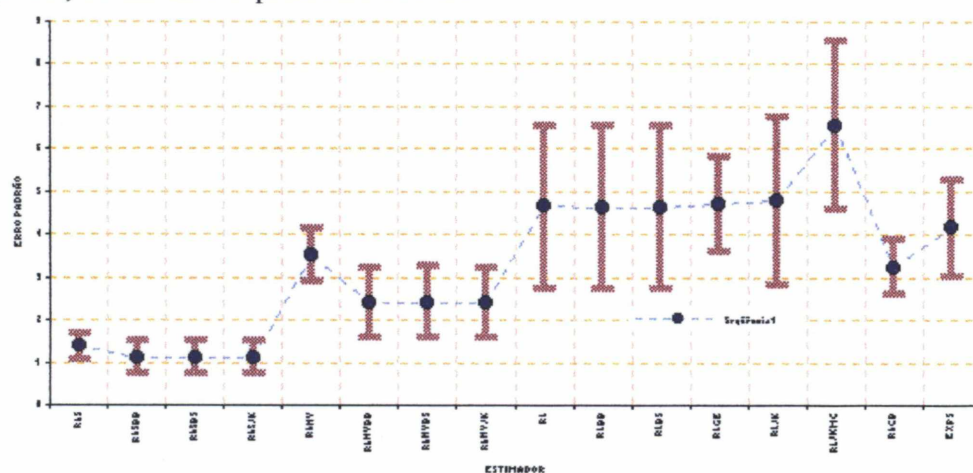


GRÁFICO 06.12: Amplitude de variação dos erros relativos percentuais em valores absolutos das estimativas populacionais. Variável: NC_PRO, eliminado o primeiro estrato.

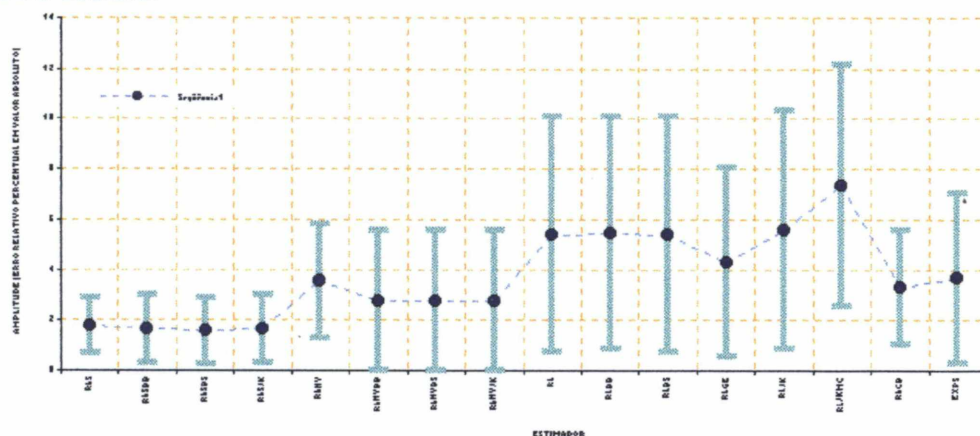


GRÁFICO 06.13: Número médio de localidades estimadas com erro relativo percentual em valor absoluto menor ou igual a 5. Variável: NC_PRO, eliminado o primeiro estrato.

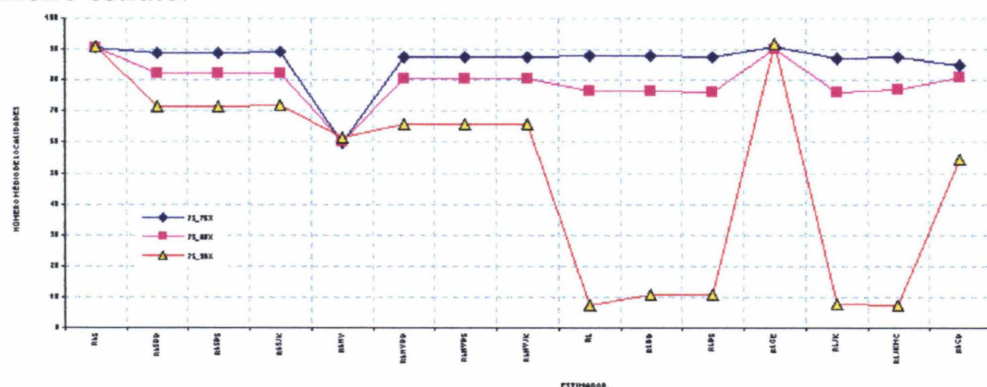


GRÁFICO 06.14: Erro Padrão do número de localidades estimadas com erro relativo percentual em valor absoluto menor ou igual a 5. Variável: NC_PRO, eliminado o primeiro estrato.

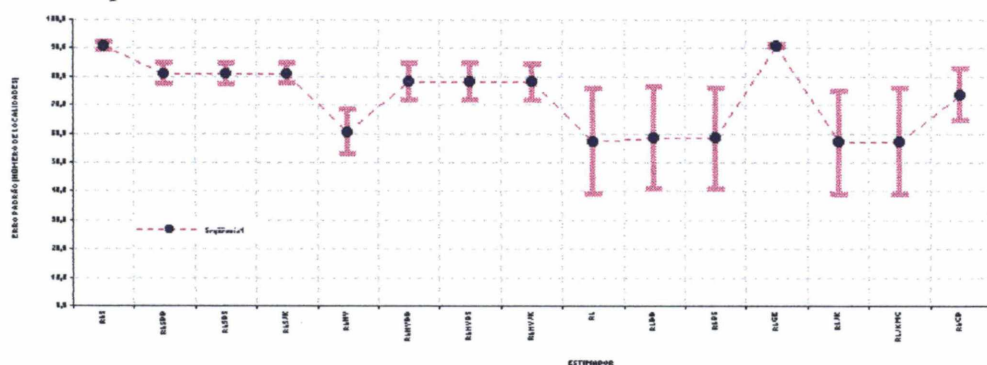


GRÁFICO 06.15: Amplitude de variação do número médio de localidades estimadas com erro relativo percentual menor ou igual a 5. Variável: NC_PRO, eliminado o primeiro estrato.

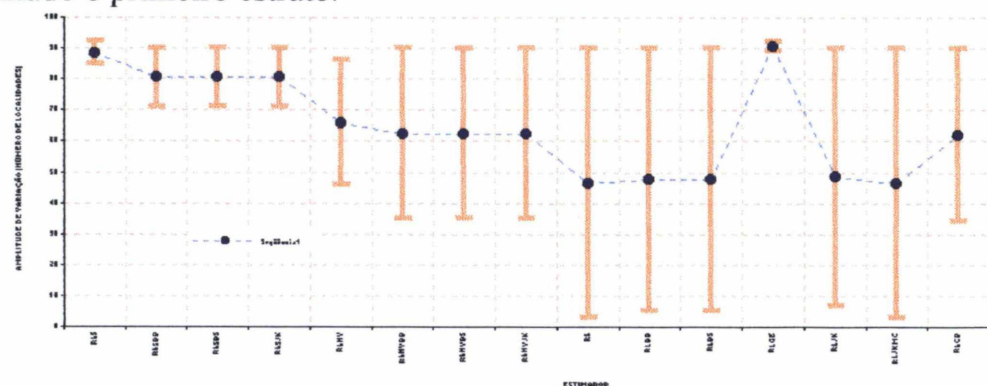


GRÁFICO 06.16: Número de vezes que o valor simulado populacional é coberto pelas estimativas populacionais. Variável: NC_PRO, eliminado o primeiro estrato.

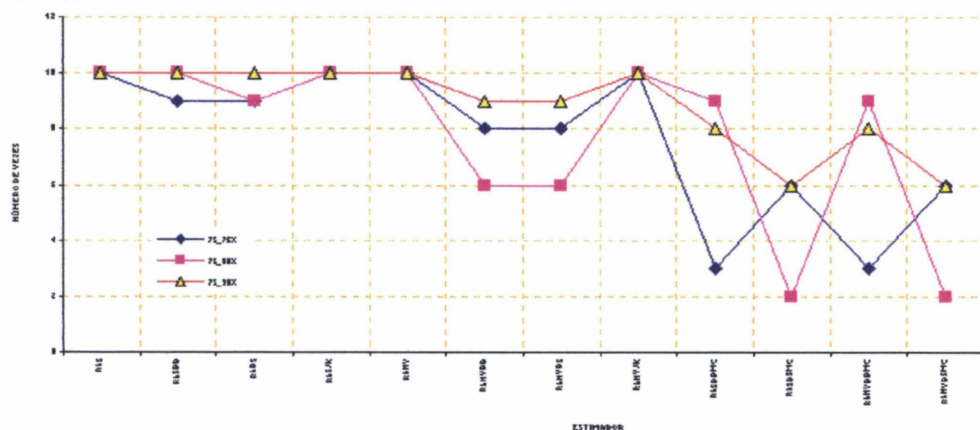


GRÁFICO 06.17: Erro Padrão do número de vezes que o valor simulado populacional é coberto pelas estimativas populacionais. Variável: NC_PRO, eliminado o primeiro estrato.

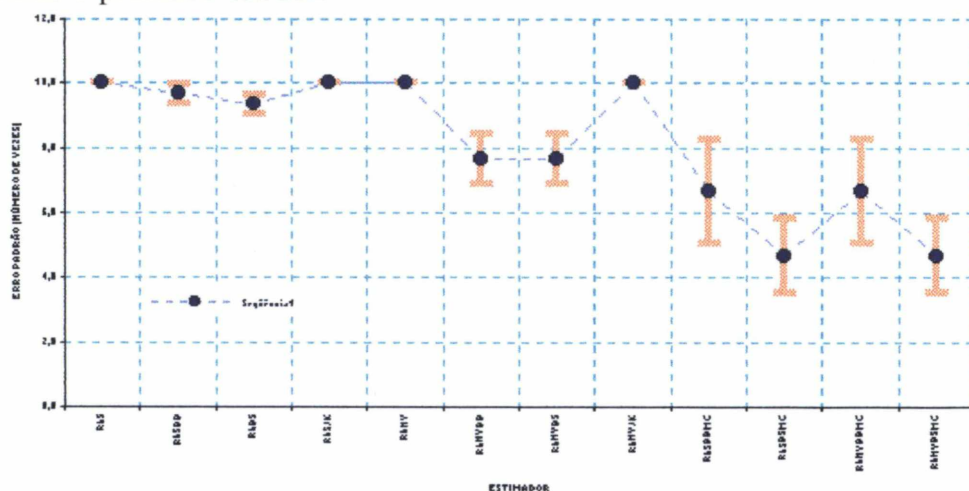


GRÁFICO 06.18: Amplitude de variação do número de vezes que o valor simulado populacional é coberto pelas estimativas populacionais. Variável: NC_PRO, eliminado o primeiro estrato.

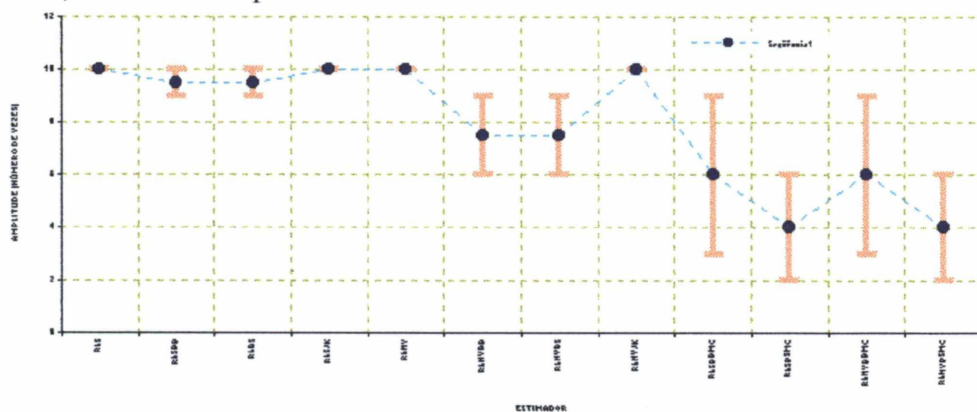


GRÁFICO 07.01: Média dos erros relativos percentuais em valores absolutos das 25 simulações. Variável: NC_RES.

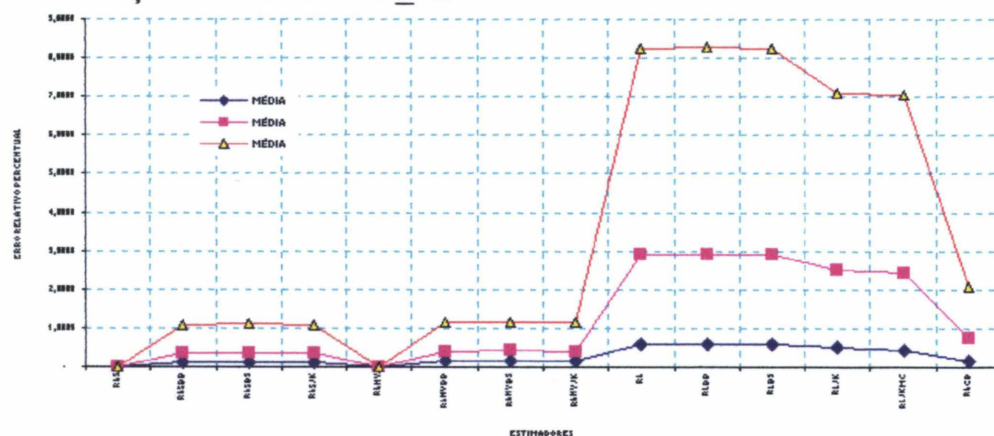


GRÁFICO 07.02: Erro Padrão dos erros relativos percentuais em valores absolutos das 75 simulações. Variável: NC_RES.

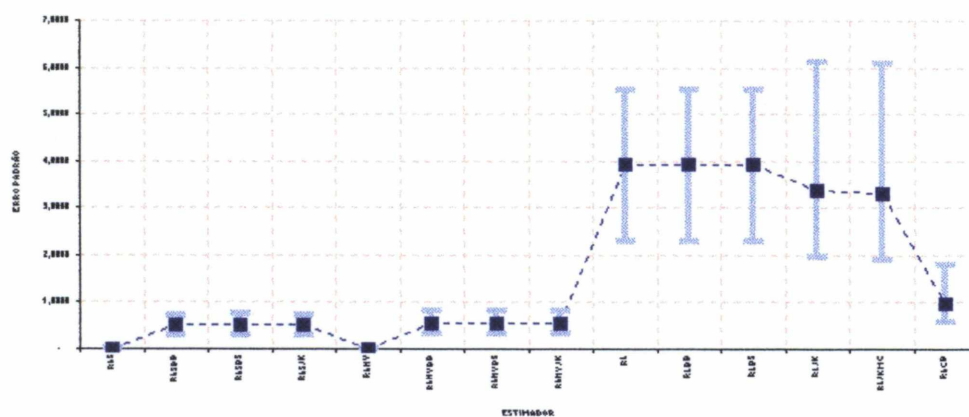


GRÁFICO 07.03: Amplitude de variação dos erros relativos percentuais em valores absolutos das 75 simulações. Variável: NC_RES.

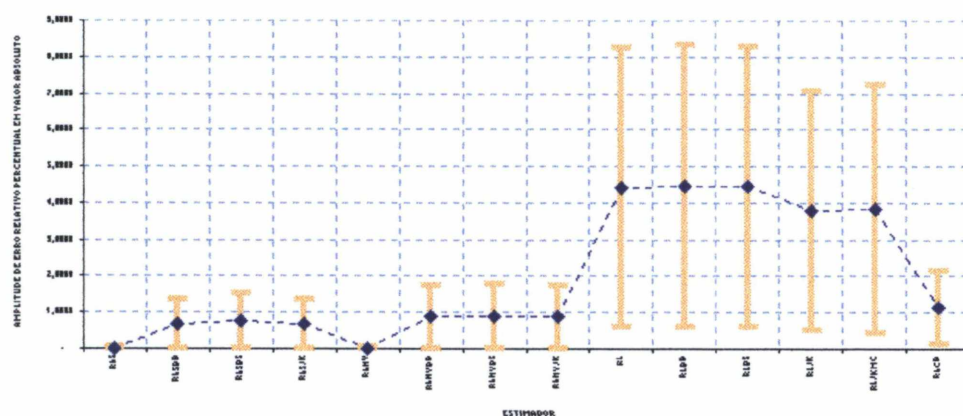


GRÁFICO 07.07: Cobertura do valor simulado das estimativas populacionais, em cada uma das 25 simulações. Variável: NC_RES.

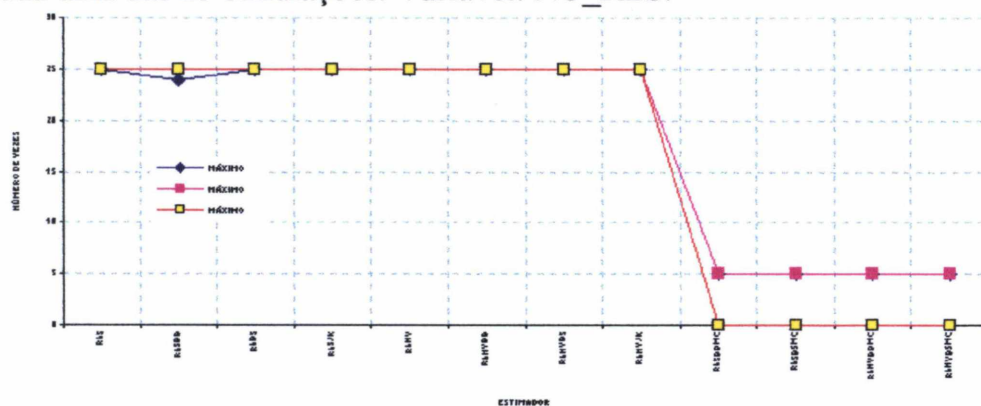


GRÁFICO 07.08: Erro Padrão do número médio de vezes que o intervalo cobre os valores simulados das estimativas populacionais, em 75 simulações. Variável: NC_RES.

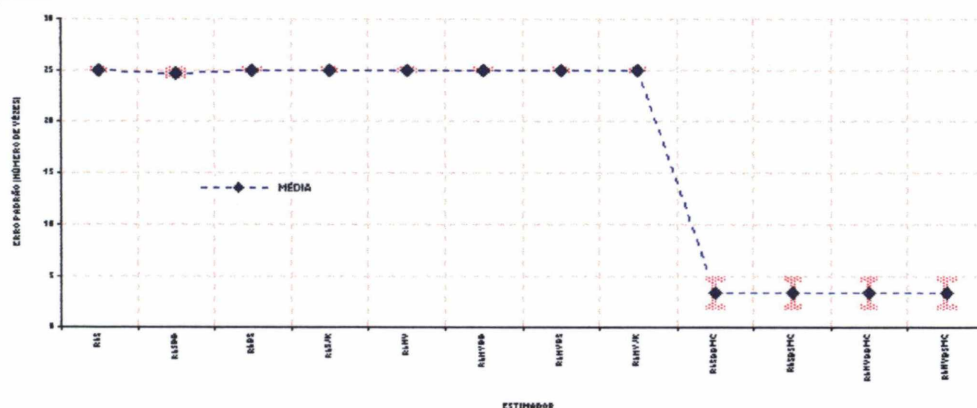


GRÁFICO 07.09: Amplitude de variação do número de vezes que o intervalo cobre o valor simulado, das estimativas populacionais, em 75 simulações. Variável: NC_RES.

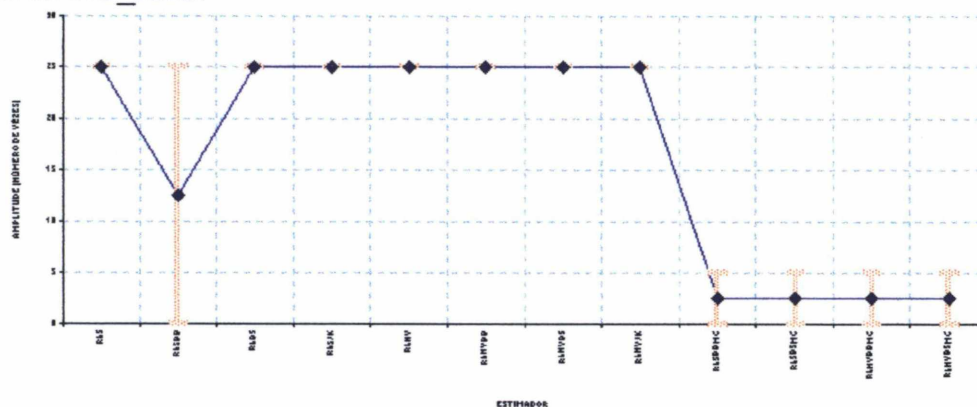


GRÁFICO 08.01: Média dos erros relativos percentuais em valores absolutos das 25 simulações. Variável: NC_RUR.

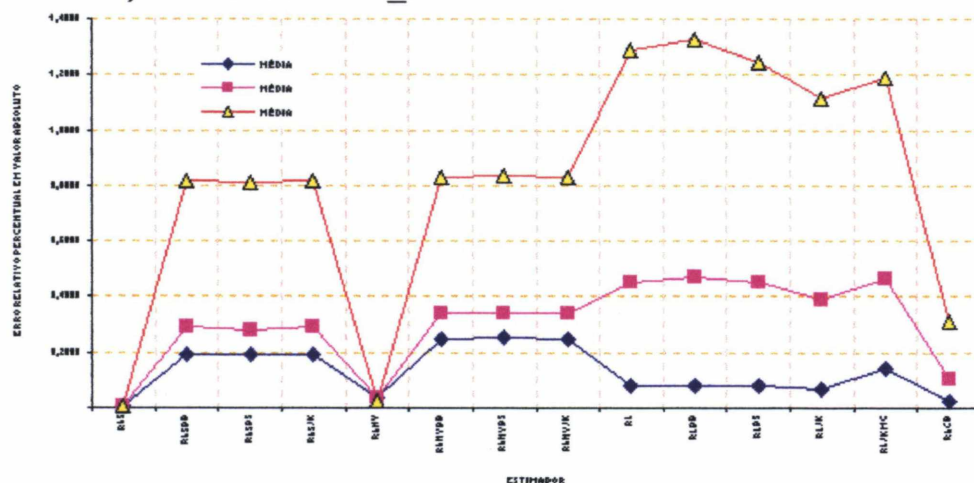


GRÁFICO 08.02: Erro Padrão dos erros relativos percentuais em valores absolutos das estimativas populacionais das 75 simulações. Variável: NC_RUR.

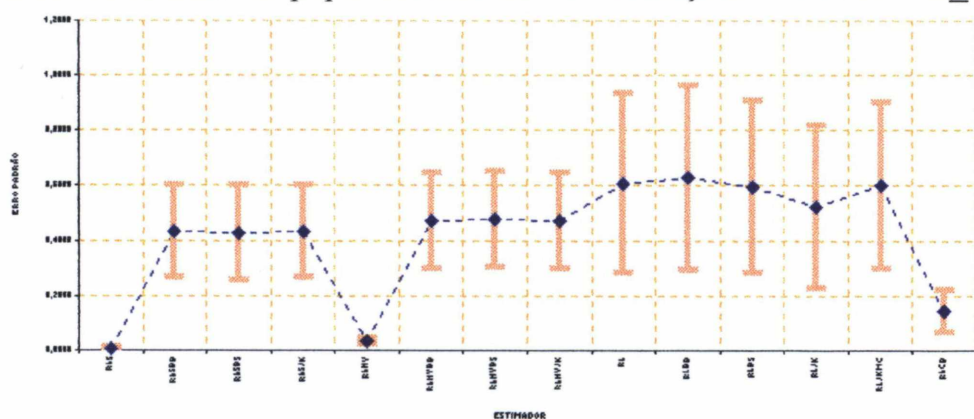


GRÁFICO 08.03: Amplitude de variação dos erros relativos percentuais em valores absolutos das estimativas populacionais das 75 simulações. Variável: NC_RUR.

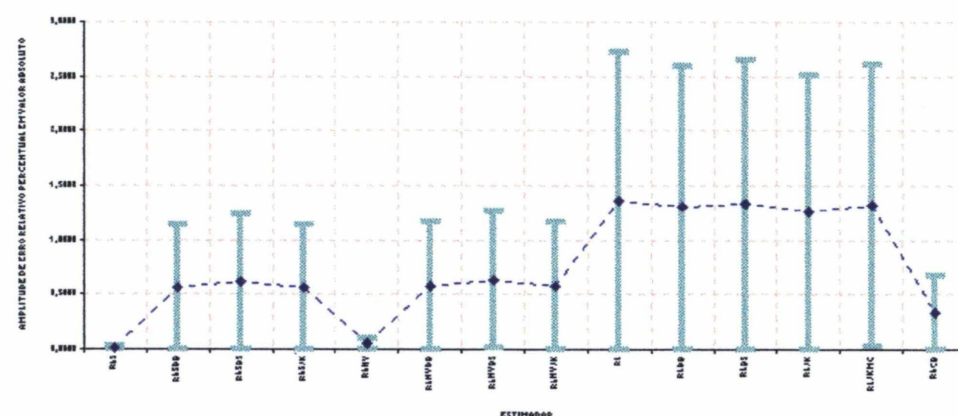


GRÁFICO 08.07: Cobertura do valor simulado das estimativas populacionais, em cada uma das 25 simulações. Variável: NC_RUR.

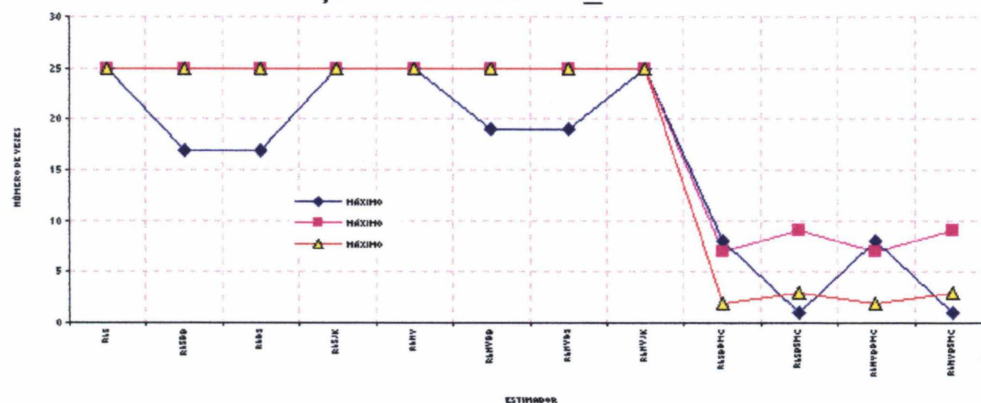


GRÁFICO 08.08: Erro Padrão do número médio de vezes que o intervalo cobre os valores simulados das estimativas populacionais, em 75 simulações. Variável: NC_RUR.

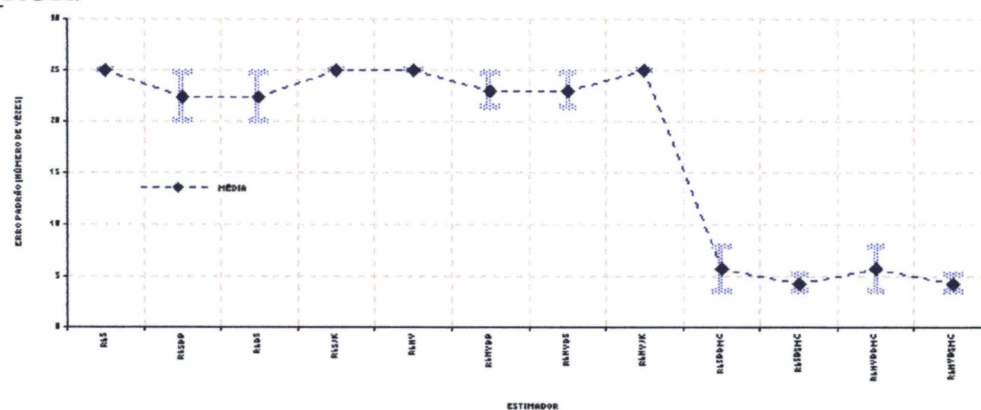


GRÁFICO 08.09: Amplitude de variação do número de vezes que o intervalo cobre os valores simulados das estimativas populacionais, para as 75 simulações. Variável: NC_RUR.

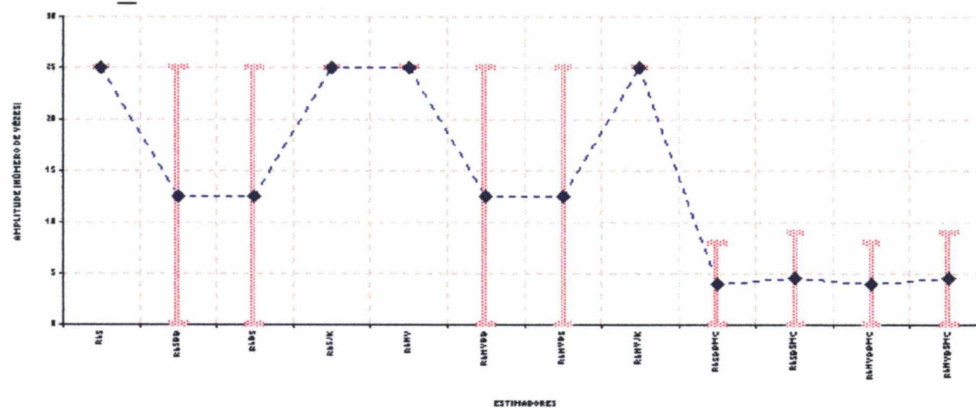


GRÁFICO 09.01: Média dos erros relativos percentuais das estimativas populacionais de 25 simulações. Variável : PR.

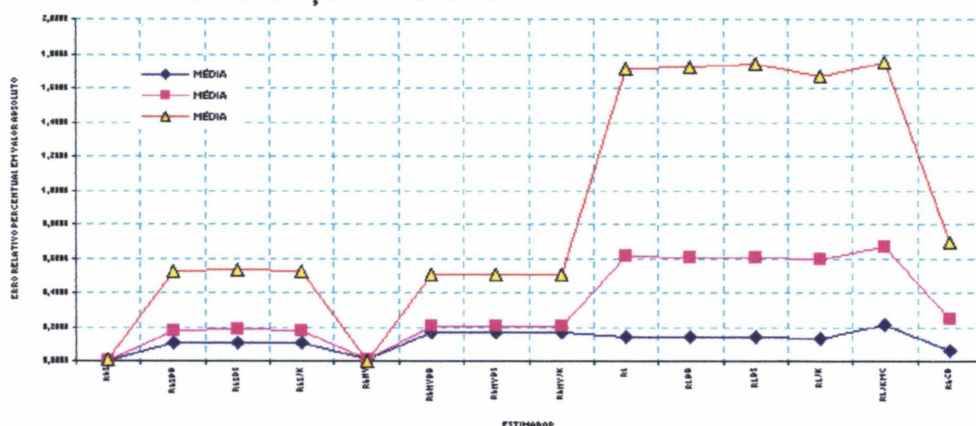


GRÁFICO 09.02: Erro Padrão dos erros relativos percentuais em valores absolutos das estimativas populacionais das 75 simulações. Variável: PR.

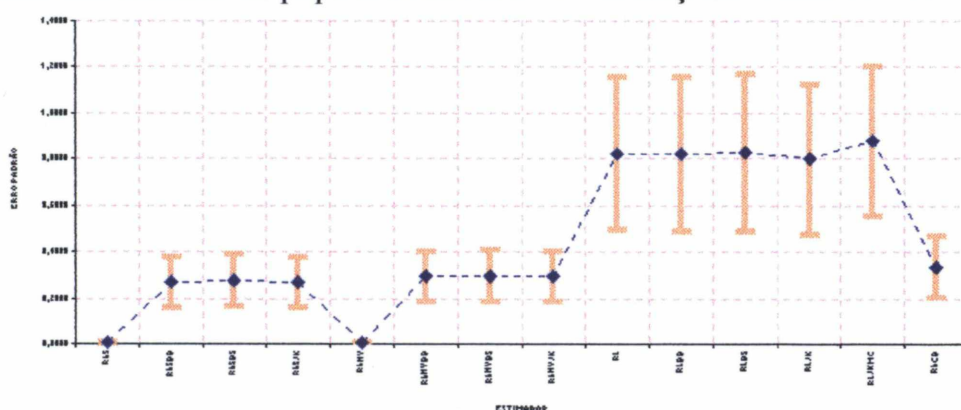


GRÁFICO 09.03: Amplitude de variação dos erros relativos percentuais em valores absolutos das estimativas populacionais de 75 simulações. Variável: PR.

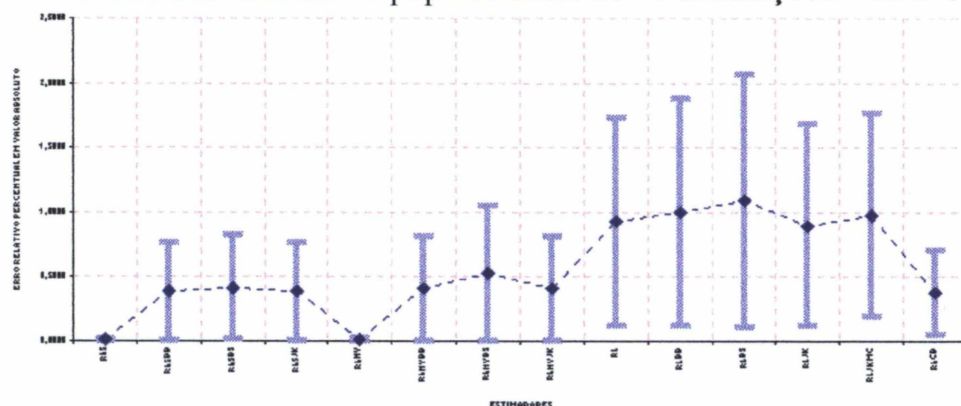


GRÁFICO 09.07: Cobertura dos valores simulados das estimativas populacionais, em cada grupo de simulação: 75_76%, 75_80% e 75_90%, respectivamente, com 25 simulações cada. Variável: PR.

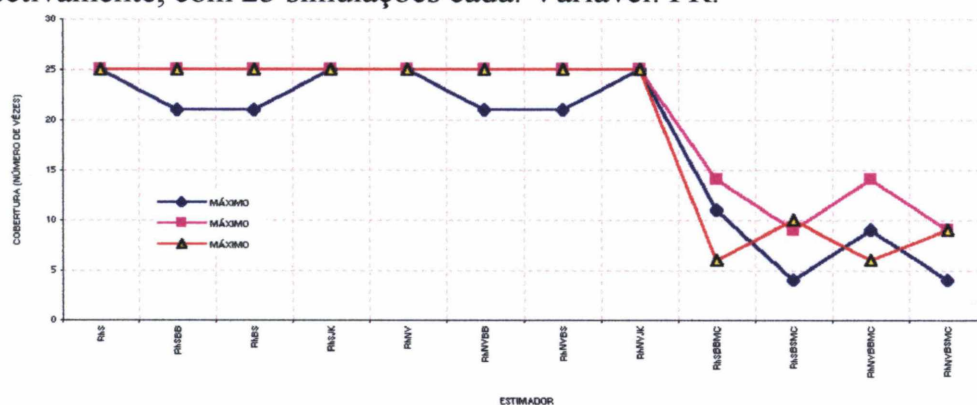


GRÁFICO 09.08: Erro Padrão do número médio de vezes que o intervalo cobre os valores simulados das estimativas populacionais, em 75 simulações. Variável: PR.

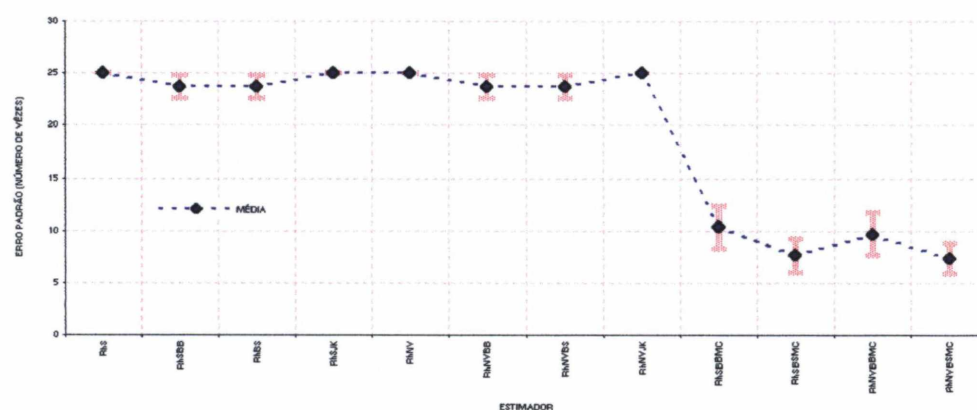


GRÁFICO 09.09: Amplitude de variação do número de vezes que o intervalo cobre os valores simulados das estimativas populacionais, em 75 simulações. Variável: PR.

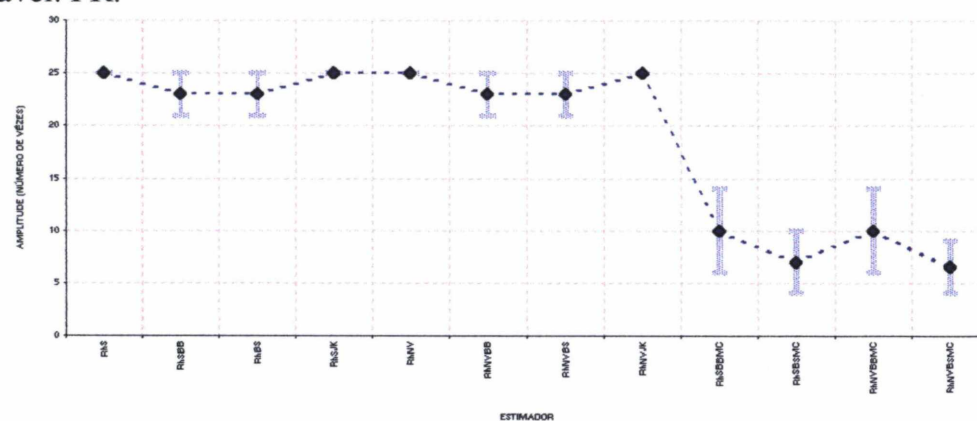


GRÁFICO 10.01: Média dos erros relativos percentuais em valores absolutos das estimativas populacionais de 25 simulações. Variável: PU.

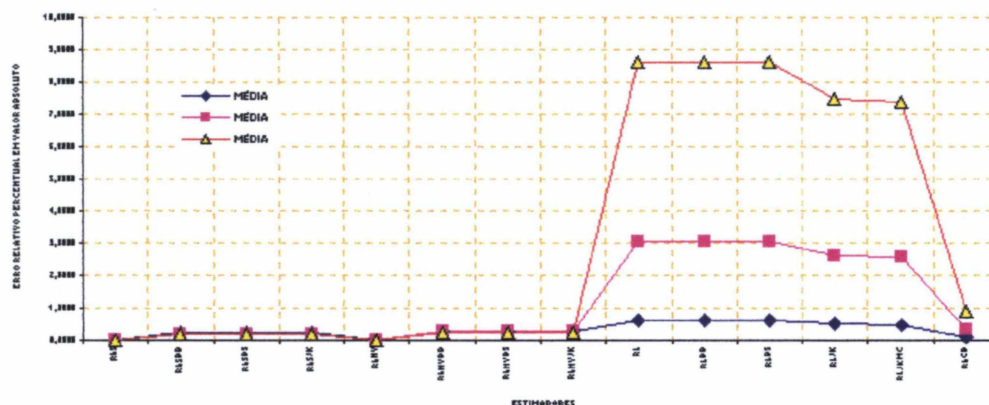


GRÁFICO 10.02: Erro Padrão dos erros relativos percentuais em valores absolutos das estimativas populacionais de 75 simulações. Variável: PR.

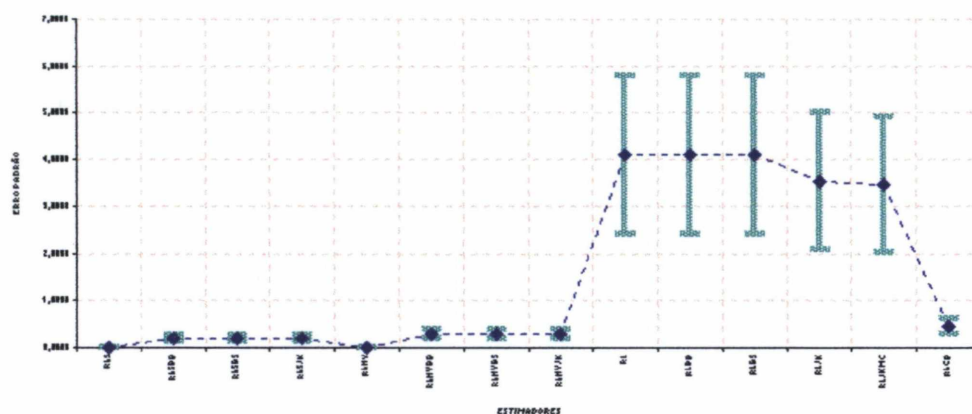


GRÁFICO 10.03: Amplitude de variação dos erros relativos percentuais em valores absolutos das estimativas populacionais de 25 simulações. Variável: PU.

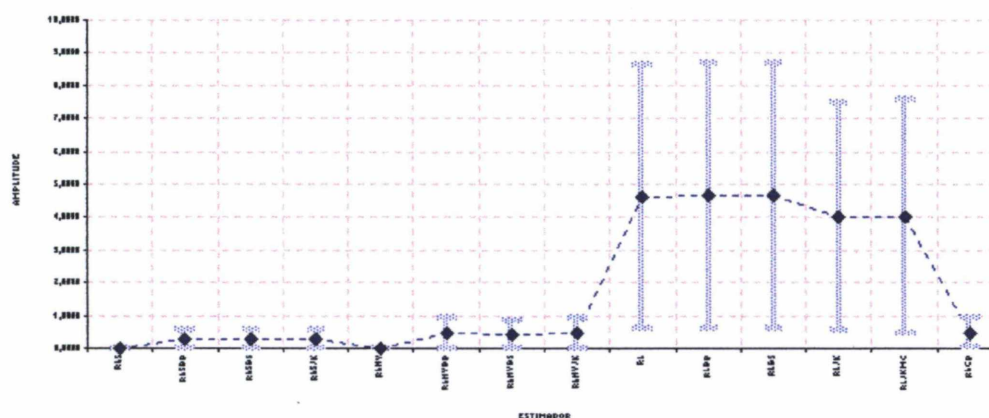


GRÁFICO 10.04: Número médio de localidades estimadas em 75 simulações com erro relativo percentual em valor absoluto menor ou igual a 5. Variável: PU.

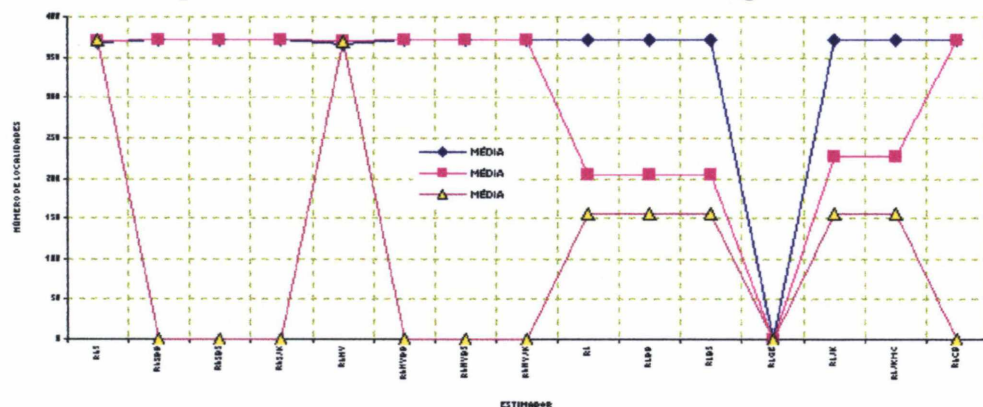


GRÁFICO 10.05: Erro Padrão do número médio de localidades estimadas com erro relativo percentual em valor absoluto menor ou igual a 5, para as 75 simulações. Variável: PU.

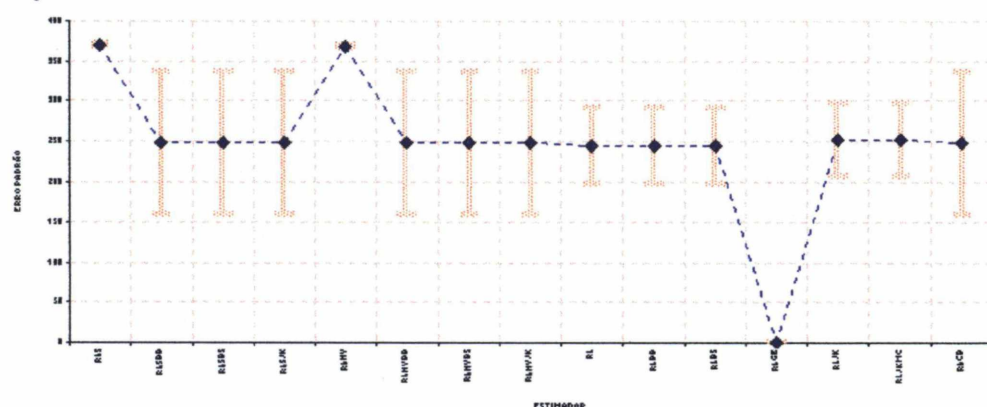


GRÁFICO 10.06: Amplitude de variação do número de localidades estimadas com erro relativo em valor absoluto menor ou igual a 5. Variável: PU.

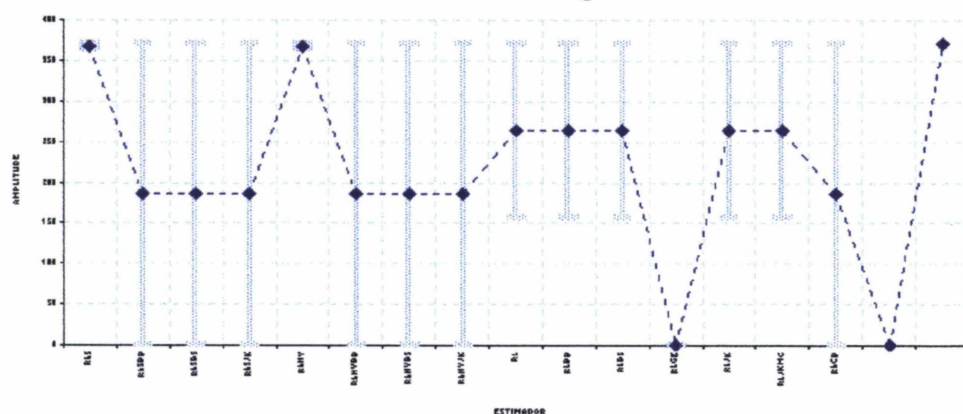


GRÁFICO 10.07: Cobertura do valor simulado das estimativas populacionais, em cada uma das 25 simulações. Variável: PR.

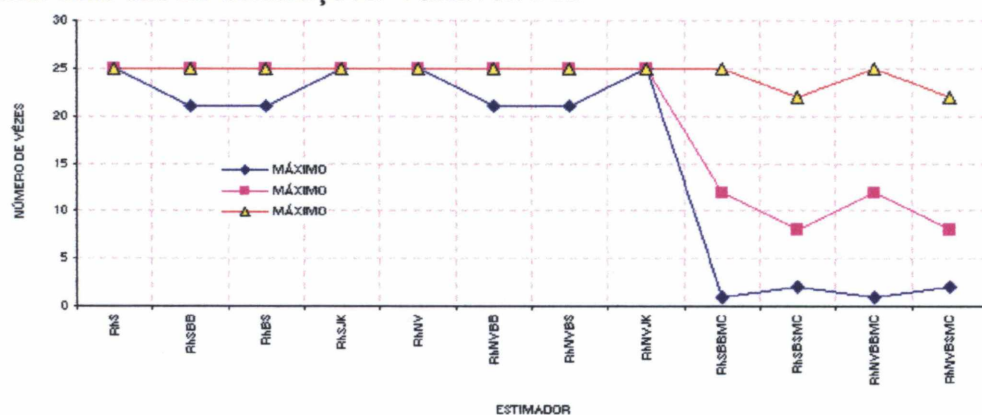


GRÁFICO 10.08: Erro Padrão do número médio de vezes que o intervalo cobre os valores simulados das estimativas populacionais, em 75 simulações. Variável: PU.

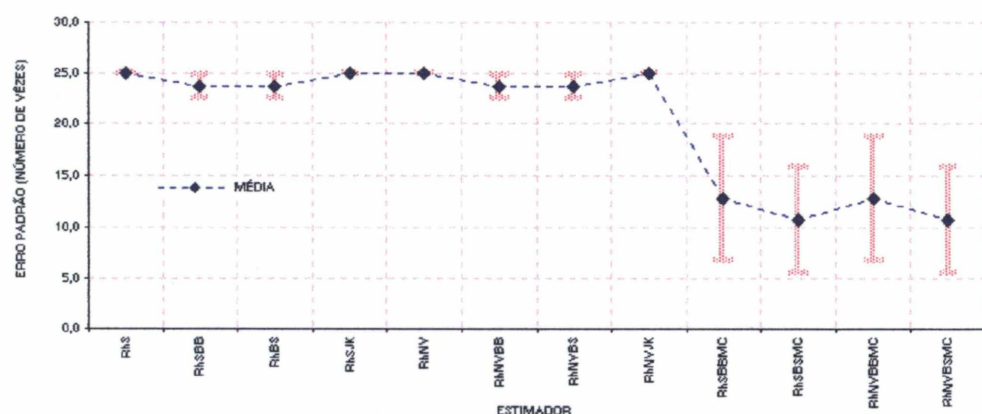


GRÁFICO 10.09: Amplitude de variação do número de vezes que intervalo cobre os valores simulados das estimativas populacionais, para as 75 simulações. Variável: PU.

