Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

IMPLEMENTAÇÃO E APLICAÇÃO DO ESTIMADOR HORVITZ THOMPSON NO SOFTWARE SAS

IGOR FERREIRA DO NASCIMENTO 08/31468

Brasília

2011

IGOR FERREIRA DO NASCIMENTO **08/31468**

IMPLEMENTAÇÃO E APLICAÇÃO DO ESTIMADOR HORVITZ THOMPSON NO SOFTWARE SAS

Relatório apresentado à disciplina Estágio Supervisionado II do curso de graduação em Estatística, Departamento de Estatística, Instituto de Exatas, Universidade de Brasília, como parte dos requisitos necessários para o grau de Bacharel em Estatística.

Orientador: Prof. Dr. Alan Ricardo da Silva

Brasília

2011

Dedicatória

À minha amada família e aos amigos sinceros e verdadeiros.

"Inteligência é capacidade que se tem de aceitar o que está ao redor."

William Faulkner

Agradecimentos

Agradeço, majoritariamente, a Deus que até aqui me ajudou.

À Joana Darc e Noé, meus pais, que com muito esforço e amor tornaram possível essa conquista. Ao carinho e cumplicidade dos meus amados e queridos irmãos.

Agradeço também aos meus avós, Maria Madela e Leôncio, e tios que contribuíram grandiosamente para minha criação e formação de caráter. Aos primos e amigos pela alegre presença e fundamental importância nos momentos de descontração.

Ao atencioso professor Alan, que me orientou com o suporte necessário para a conclusão desse trabalho. Aos bons professores do departamento de Estatística e Matemática da Universidade de Brasília que acreditaram no meu potencial.

Resumo

A amostragem para pequenas populações é vastamente encontrada em pesquisas científicas. Nessas situações, a estimativa do parâmetro populacional possue elevado erro padrão devido o tamanho da amostra. O estimador Horvitz Thompson é apropriado para o estudo de populações finitas com pequenas amostras. Para isso, utiliza informações auxiliares que permitem obter estimativas não viesadas e com pequenas variâncias.

Assim, o objetivo desse trabalho é desenvolver o algoritmo para o estimador Horvitz Thompson tanto em um contexto populacional, a fim de verificar as propriedades de esperança e variância, quanto em um contexto amostral. Para isso, será utilizado o ambiente **IML** do *software* **SAS** 9.2.

Os resultados evidenciaram que para o estudo de caso analisado, de fato o estimador de Horvitz Thompson é um estimador não viesado, bem como as variâncias amostrais. Além disso, constatou-se que é 10 vezes mais eficiente que no caso de uma $AAS_{(s)}$.

Ademais, construiu-se também o algoritmo no ambiente **R**.

Abstract

The sampling for small populations is widely found in scientific research. In these situations, the estimated population parameter possesses high standard error because the sample size. Horvitz Thompson estimator is suitable for the study of finite populations with small samples. It uses auxiliary information for obtaining unbiased estimates with small variances.

Thus, our objective is to develop the algorithm for the Horvitz Thompson estimator both in a population context, in order to check the properties of hope and variance, as in a sampling context. This will use the environment **IML** of *software* **SAS** 9.2.

The results showed that for the case study analyzed, in fact the Horvitz Thompson estimator is an unbiased estimator and the variances. Moreover, it was found that is 10 times more efficient than if a $AAS_{(s)}$.

Moreover, it is also built in the algorithm environment **R**.

Lista de Tabelas

3.1 Tamanho dos supermercado 15

3.2 Probabilidade conjunta de inclusão 15

3.3 Distribuição amostral do estimador HT 16

3.4 Distribuição amostral do estimador $AAS_{(s)}$ 16

Lista de Figuras

2.1 Representações gráficas dos estimadores: (a) não viesado e preciso (b) não viesado e impreciso (c) viesado e preciso (d) viesado e impreciso . 8

Sumário

Resumo **iv**

Abstract **v**

1 Introdução **1**

1.1 Objetivos 2

2 Amostragem **3**

2.1 Conceitos básicos 4

2.2 Com reposição 4

2.3 Sem reposição 5

2.3.1 Aproximações 6

2.4 Estimação: conceitos básicos 7

2.5 Estimador Hansen Hurwitz(HH) 9

3 Estimador HT **11**

3.1 Introdução 11

3.2 Estimadores HT 12

3.3 Problema dos supermercados(Lohr, 1999) 14

3.3.1 Horvitz Thompson 14

3.3.2 Aleatória simples 16
3.4 Generalização do estimador HT 17
3.4.1 Aleatória simples 17
3.4.2 Estratificada 18

4 Algoritmo desenvolvido **19**
4.1 Macro 19
4.1.1 %HT_general 19
4.1.2 %HT 21
4.2 Exemplos HT 21
4.2.1 HT n=2 21
4.2.2 $AAS_{(s)}$ n=2 através do estimador HT 23

5 Considerações finais **25**

Referências **26**

Apêndice **27**

A %HT_General - SAS **27**

B %HT - SAS **29**

C HTGeral - R **30**

Capítulo 1

Introdução

A escolha do melhor plano amostral depende diretamente da estrutura de como os indivíduos estão distribuídos na população e como podem ser coletados. O estimador Horvitz Thompson tem interessantes características de aplicação nas amostragens probabilísticas mais complexas. Pode ser utilizado quando os elementos amostrais possuem diferentes probabilidades de seleção e quando não há interesse na seleção com reposição. Devido a intensidade computacional necessária para cálculo desse estimador, ele é indicado para populações finitas e amostras pequenas. Conforme Lohr (1999), o estimador Horvitz Thompson pode ser considerado o caso geral da amostragem sem reposição.

Utilizando as propriedades de esperança e variância nesse estimador é possível mostrar que é não viesado e preciso. Esses resultados podem ser obtidos através da geração de todas as amostra possíveis.

1.1 Objetivos

O objetivo geral do trabalho é implementar o estimador Horvitz Thompson no software SAS. Os objetivos específicos são:

- Revisão das técnicas de amostragem;
- Conhecer e estudar o estimador;
- Mostrar o poder e aplicabilidade;
- conduzir sua implementação via **PROC IML** no *software* **SAS** 9.2.

Capítulo 2

Amostragem

A informação tem uma papel fundamental na atualidade, pois naturalmente é utilizada para consolidar tomadas de decisões em qualquer área do conhecimento e até em situações comuns do cotidiano. Com importância igual à maneira utilizar essas informações é como podem ser adquiridas e se estão disponíveis, direta ou indiretamente. O estudo de técnicas em amostragem tem por objetivo, baseado em um subconjunto da população, definir de que forma essa informação será coletada e o grau de incerteza associado a estimativa. As vantagens associadas a utilização da amostragem são claramente citadas pelas principais referências em amostragem, como Cochran (1977). Porém não é difícil percerber que, seja qual for o dispêndio para a coleta dos dados (tempo, dinheiro , etc.), a amostragem é substancialmente vantajosa. Por existir diversas formas de fazer amostragem, é necessário comparar as metodologias possíveis e decidir dentre essas qual é a mais econômica e/ou eficiente.

Em amostragem para populações com uma quantidade finita e completamente enumerável de elementos, a escolha de uma metodologia que utiliza informações auxiliares pode resultar em estimativas mais precisas.

2.1 Conceitos básicos

Comumente em amostragem associa-se a distribuição probabilística de seleção a uma uniforme discreta, em que todos os elementos tem igual probabilidade de serem selecionados. No entanto, esse modelo pode ser considerado como o modelo mais simples devido a falta de informação sobre a variável do estudo. Informações como, tamanho da população, de estratos e *cluster* incrementam o modelo de estimação. Essa idéia pode ser facilmente notada no processo de seleção estratificada, em que os estratos, na grande maioria dos casos, tem tamanhos diferentes e a inserção de um modelo probabilístico de seleção, que leva em consideração essa diferença de tamanhos, resultaria em estimativas mais apropriadas e precisas, isto é, tem-se um impacto na redução da variância do estimador.

No processo de seleção, a probabilidade de inclusão depende intrinsecamente da forma como os elementos são amostrados e do tamanho que cada unidade representa para a população. É possível selecionar os elemento com ou sem reposição e com os elementos ocupando ou não o mesmo tamanho populacional. Por isso é necessário entender um pouco de como a probabilidade rege o processo de seleção na amostragem.

2.2 Com reposição

Se existe diferença na probabilidade de seleção em um plano amostral com reposição tem-se o seguinte desenho amostral:

Seja Ψ_i a probabilidade do i-ésimo elemento ser escolhido na primeira retirada.

Em uma amostra de tamanho n, como cada elemento pode ser retirado mais de uma vez e sempre com a mesma probabilidade, as seleções são independentes. Com isso, seja X o número de vezes que o elemento i aparece na amostra é fácil perceber que $X \sim Bin(n, \Psi_i)$, pois cada uma das n seleções tem distribuição Bernoulli de parâmetro Ψ_i. Com isso, tem-se que a probabilidade de inclusão na amostrada é denotada por π_i é:

$$\pi_i = \sum_{k=1}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) \Psi_i^k (1 - \Psi_i)^{n-k} = 1 - (1 - \Psi_i)^n \tag{2.1}$$

A probabilidade de que quaisquer 2 elementos, i e j, estejam na amostra tem distribuição multinomial de parâmetros n,Ψ_i e Ψ_j:

$$\pi_{ij} = \sum_{k=1}^{n} \sum_{w=1}^{n-k} \left(\begin{array}{c} n \\ k\,,\, w \end{array} \right) \Psi_i^k \Psi_j^w (1 - \Psi_i - \Psi_j)^{n-k-w} \tag{2.2}$$

2.3 Sem reposição

Para a situação em que as unidades de observação ocupam proporções diferentes no conjunto populacional e em um processo de seleção sem reposição, os modelos probabilísticos associados às amostras tornam-se mais complicados. Nesse caso, os eventos não são independentes e por isso é necessário trabalhar com as probabilidades condicionais de seleção. Seja a unidade i com uma probabilidade de ser selecionada na primeira amostra Ψ_i, a probabilidade do elemento j ser retirado na segunda amostra é:

$$P(j/i) = \frac{\Psi_j}{1 - \Psi_i}$$

É fácil perceber, que quando as probabilidades Ψ_i e Ψ_j são diferentes, a probabilidade da amostra i e j ser realizada é diferente quando a ordem de seleção é

alterada:

$$\Psi_i \frac{\Psi_j}{1-\Psi_i} \neq \Psi_j \frac{\Psi_i}{1-\Psi_j} \tag{2.3}$$

Para uma amostra de tamanho 2, tem-se que a probabilidade de que i e j estejam na amostra, π_{ij}, é:

$$\pi_{ij} = P(i, j \in A) = P(i)P(j/i) + P(j)P(i/j) = \Psi_i \frac{\Psi_j}{1-\Psi_i} + \Psi_j \frac{\Psi_i}{1-\Psi_j} \tag{2.4}$$

A probabilidade de inclusão do elemento i na amostra é a soma das probabilidades conjuntas das amostras que contém esse elemento. É fácil notar que esse número é $n!C_n^N$ amostras. Utilizando as propriedades da variável aleatória Z_i com distribuição Bernoulli, 1 caso tenha sido selecionado e 0 caso contrário, é possível mostrar que, (Lohr, 1999):

$$(n-1)\pi_i = \sum_{i \neq j}^{N} \pi_{ij} \tag{2.5}$$

Na amostragem, a flexibilização da probabilidade de seleção torna os cálculos um pouco mais elaborados, além de problemas como a instabilidade da variância do estimador e o tempo de processamento.

2.3.1 Aproximações

O valor de Ψ_i é definido utilizando o conhecimento do pesquisador ou estudos que indiquem haver relação entre a variável de estudo e uma variável X. Caso essa relação se dê de forma direta, é possível definir:

$$\Psi_i = \frac{x_i}{\sum x_i} \tag{2.6}$$

Existem aproximações para a probabilidade de inclusão utilizando a relação $\pi_i \propto \Psi_i$, uma delas é $\pi_i = n\Psi_i$. No entando, se o tamanho da amostra é diferente da

população, tem-se que $\pi_i < 1$ e com isso, é possível obter um $\pi_i > 1$.

2.4 Estimação: conceitos básicos

As informações obtidas através da amostragem são as estatísticas, e os estimadores são funções dessas estatísticas. Elas servem como uma medida representativa, uma aproximação, do valor populacional que é chamado de parâmetro. É possível propor mais de uma função estatística e com isso é necessário estabelecer critérios para selecionar o melhor estimador.

Para o processo de escolha, algumas características dos estimadores servem como regra de decisão para o mais apropriado. São elas:

- viés;
- variância;
- erro quadrático médio(EQM);
- consistência.

Os três primeiros tópicos são medidas e a última é uma propriedade do estimador. Esses conceitos estão atrelados ao fato de que as estatísticas amostrais, obviamente, mudam de acordo com a amostra selecionada. Com isso, é desejável que o estimador erre pouco para cima (superestimação) e para baixo (subestimação) em relação ao valor populacional. O estimador que possui a média dos erros igual a zero é dito não viesado, isto é, em média, não erra o valor do parâmetro de interesse.

Além disso, mesmo que o estimador seja não viesado é importante possuir uma baixa disperção em entorno do valor alvo da população. Ou seja, deseja-se que os

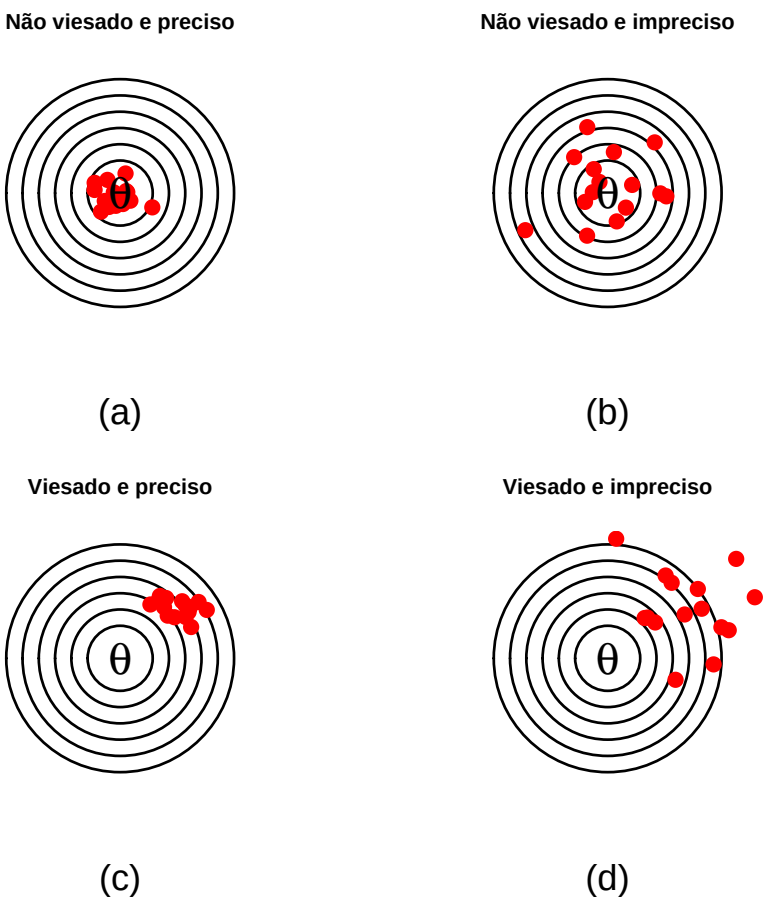

Figura 2.1: Representações gráficas dos estimadores: (a) não viesado e preciso (b) não viesado e impreciso (c) viesado e preciso (d) viesado e impreciso

valores do estimador em cada amostra tenha uma pequena variância entorno do valor alvo. Para ilustrar esses conceitos fundamentais para o processo de estimação, a Figura 2.1 compara os tipos de estimadores.

O erro quadrático médio (EQM) combina o viés com a variância do estimador e é o principal critério para escolhas do estimador.

$$EQM(\hat{\theta}) = VAR(\hat{\theta}) + (E(\hat{\theta}) - \theta)^2 \tag{2.7}$$

Seja $\hat{\theta}_n$ um estimador de θ para uma amostra de tamanho n, ele é dito consistente

se:

$$\lim_{n\to\infty} VAR(\hat{\theta}_n) \to 0 \tag{2.8}$$

2.5 Estimador Hansen Hurwitz(HH)

O estimador de Hansen and Hurwitz (1943) está inserido no contexto de população finitas em que o interesse é obter uma amostra de tamanho n de uma população de tamanho N através de uma amostra aleatória simples com reposição ($AAS_{(c)}$). Nesse processo de coleta, as seleções são independentes e, assim, a probabilidade do indivíduo k ser selecionado em quaisquer uma das n amostras é sempre p_k.

O estimador do total populacional $\sum_i^N y_i$ é:

$$\hat{t}_{pwr} = \frac{1}{n}\sum_i^n \frac{y_i}{p_i} \tag{2.9}$$

É possível perceber que o estimador HH utiliza a aproximação para a probabilidade de seleção apresentada na Seção 2.3.1. É também conhecido como *"p-expanded with replacement"* e é não viesado, (Hansen and Hurwitz, 1943). A variância desse estimador é dada por:

$$V(\hat{t}_{pwr}) = \frac{1}{n}\sum_i^n \left(\frac{y_i}{p_i} - \hat{t}_{pwr}\right)^2 p_i \tag{2.10}$$

A estimativa não viesada baseada nas informações amostrais é, (Särndal et al., 1992):

$$\widehat{V(\hat{t}_{pwr})} = \frac{1}{(n-1)n}\sum_i^n \left(\frac{y_i}{p_i} - \hat{t}_{pwr}\right)^2 \tag{2.11}$$

Os trabalhos sobre populações finitas continuaram ao longo da década, até que em 1953 surgiu um estimador capaz de dar estimativas não viesadas com probabili-

dade diferente de inclusão sem reposição, conhecido por Estimador Horvitz Thompson.

Capítulo 3

Estimador Horvitz Thompson(HT)

3.1 Introdução

Esse estimador foi proposto por Horvitz e Thompson em 1952 e é a generalização do estimador de Hansen Hurwitz. Pode-se realizar as estimativas para o caso em que não há interesse na amostragem com reposição. Utiliza as informações da distribuição não estruturada de probabilidade dos elementos populacionais. Essa informação minimiza a variância do estimador. O uso apropriado da variação probabilística de seleção permite ganhos em eficiência quando comparado com o plano amostral em que as unidades tem probabilidades iguais de seleção (Thompson, 1952). A técnica é apropriada para pequenas populações pois trata individualmente cada elemento da população. Sendo assim, para populações maiores surgem os problemas computacionais associados à estimação.

Thompson e Horvitz também discutiram os problemas associados a utilização de diferentes probabilidades de inclusão para os indivíduos, sendo uma delas a instabilidade da variância, caso em que a mesma pode ser negativa.

3.2 Estimadores HT

No estudo de amostragem direcionado a populações finitas, em que se tem um cadastro, é sensato imaginar que toda informação sobre a população alvo pode ser utilizada como uma fonte de qualidade e eficiência para as estimativas. O estimador HT faz uso dessas informações para tratar com especificidade cada elemento da população. Esse tratamento individual concede ao pesquisador mais confiança para a utilização dos resultados obtidos.

De posse das informações cadastrais, escolhe-se uma variável que esteja relacionada com a variável de interesse e, assim, cria-se o vetor de probabilidades de seleção como mostrado no inicio da Seção 2.3.1. Isto é, o estimador HT possibilita utilizar esse vetor de probabilidade associados aos elementos da população e com isso modela-se o plano amostral baseado na realidade dos dados. Através desses valores calcula-se a probabilidade de inclusão associada a cada elemento e conjunta de qualquer par desses elementos. Essas probabilidades são essenciais para o cálculo das estimativas.

O estimador HT para o total populacional t é dado por:

$$\hat{t}_\pi = \sum_{i=1}^{n} \frac{y_i}{\pi_i} \tag{3.1}$$

Sendo π_i a probabilidade do i-ésimo elemento estar na amostra. Fazendo Z ser uma variável indicadora do elemento na amostra, 1 quando amostrado e 0 caso contrário, tem-se que $P(Z_i = 1) = \pi_i$, (Lohr, 1999). Com isso, o estimador é não tendencioso para o total t.

$$E(\hat{t}_\pi) = E\left(\sum_i^N z_i \frac{y_i}{\pi_i}\right) = \sum_i^N \pi_i \frac{y_i}{\pi_i} = t \tag{3.2}$$

Intuitivamente, obtém-se a estimativa para a média populacional dividindo a Equação 3.1 por N. A variância associada ao estimador é dada por:

$$V(\hat{t}_\pi) = \sum_i^N \frac{1-\pi_i}{\pi_i} y_i^2 + \sum_i^N \sum_{j \neq i}^N \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j} y_i y_j \tag{3.3}$$

Sendo π_i a probabilidade inclusão do elemento i na amostra e π_{ij} a de que os elementos i e j estejam na amostra.

Como mostrado na Seção 2.3, π_i é obtido com a soma de todas as amostras que o elemento i está contido. Esse cálculo gera problemas de tempo de processamento quando a população é muito grande, visto que são necessárias ter as probabilidades de ocorrencia das $n!C_n^N$ amostras. Na Seção 4.1, o algoritmo para calcular essa probabilidade é mostrado e comentado.

É possível ter uma estimativa não viesada baseando-se nas informações amostrais (Thompson, 1952):

$$\widehat{V_1(\hat{t}_\pi)} = \sum_i^n \frac{1-\pi_i}{\pi_i^2} y_i^2 + \sum_i^n \sum_{j \neq i}^n \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_{ij} \pi_i \pi_j} y_i y_j \tag{3.4}$$

Esse estimador da variância populacional pode fornecer estimativas negativas. Um estimador dado por Yates and Grundy (1953), através de outra forma para a Equação 3.3, fornece estimativas não negativas e é dada por:

$$\widehat{V_2(\hat{t}_\pi)} = \sum_i^n \sum_{j>i}^n \frac{(\pi_i \pi_j - \pi_{ij})}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 \tag{3.5}$$

Pode-se ainda utilizar o estimador da variância populacional HH para estimativa intervalares conservativas do HT. Durbin (1953) mostrou que o viés desse estimador

para a variância HT é:

$$B(v) = E(v) - V(\hat{t}_\pi) = \frac{m}{m-1}[V(\hat{t}_{pwr}) - V(\hat{t}_\pi)] \tag{3.6}$$

Sendo v dado por 2.11.

Para o problema computacional proveniente do cálculo de π_i dado na Seção 2.3, segundo Cochran (1977), existem mais de 30 métodos alternativos dessa probabilidade de inclusão. Com relação a negatividade das estimativas de 3.3 existem mais estimadores propostos que também são estritamente positivos. No entanto, esse trabalho está focado apenas nas propriedades e conceitos do estimador HT conjuntamente com seu desenvolvimento e implementação no **SAS**.

Não obstante, o uso de um apropriado método de seleção permiti ganhos em eficiência em relação ao método com iguais probabilidades de inclusão .

3.3 Problema dos supermercados(Lohr, 1999)

Deseja-se estimar a quantidade total de vendas de uma pequena cidade. Essa cidade possui 4 supermercados e é proposto a estimação do total através de uma amostra de tamanho 2.

3.3.1 Horvitz Thompson

Supondo que existe uma correlação positiva entre o tamanho do supermercado e o número de vendas, utiliza-se a probabilidade de seleção primária igual a proporção da área do supermercado. É interessante notar que o cálculo da probabilidade de inclusão de cada unidade é feita a partir de uma variável *proxy* que tenha-se conhecimento.

A Tabela 3.1 mostra a probabilidade de seleção inicial Ψ_i de cada supermercado (proporcional ao tamanho), tamanho em m^2 e o valor total de venda em cada estabelecimento, y_i.

Tabela 3.1: Tamanho dos supermercado

Supermercado	Ψ_i	m^2	y_i
A	0,0625	100	11
B	0,1250	200	20
C	0,1875	300	24
D	0,6250	1.000	245
Total	1	1.600	300

O próximo passo é construir e calcular a probabilidade de inclusão para os 4 elemento e para todos os C_2^4 pares de elementos. Como visto na Seção 2.1, as probabilidades de inclusão são calculadas através das probabilidades de todas as possíveis amostras da população, isto é, os arranjos. O tempo de processamente dessa matriz fica mais difícil à medida que aumenta o tamanho da amostra.

Tabela 3.2: Probabilidade conjunta de inclusão

π_{ij}	A	B	C	D	π_i
A	0	0,0173	0,0269	0,1458	0,1900
B	0,0173	0	0,0556	0,2977	0,3705
C	0,0269	0,0556	0	0,4567	0,5393
D	0,1458	0,2977	0,4567	0	0,9002
π_i	0,1900	0,3705	0,5393	0,9002	2

Na marginal da Tabela 3.2 estão as probabilidades de inclusão de cada indivíduo e, a partir disso, é obtido a estimativa do total populacional.

Com a Tabela 3.3 nota-se que é um estimador não viciado, esperança 300, e apresenta uma variância de $4.383,56$. Os resultados foram obtidos multiplicando os valores da segunda e terceira colunas por suas respectivas probabilidades. Para

Tabela 3.3: Distribuição amostral do estimador HT

p	$\hat{t}$	$(\hat{t}-300)^2$
0,0172	111,8684	35.393,49
0,0269	102,39247	39.048,73
0,1458	330,05599	903,36
0,0556	98,48255	40.609,28
0,2976	326,14607	683,61
0,4567	316,67015	277,8
Esperança	300	4.383,56

calcular a variância do estimador de forma direta pela Equação 3.2 é necessário utlizar os valores encontrados na matriz de probabilidade conjunta (Tabela 3.2).

3.3.2 Aleatória simples

Como pode ser visto na Tabela 3.4, caso fosse utlizado o estimador de média simples da $AAS_{(s)}$, ainda que não viesado, teria uma variância 10 vezes maior que o estimador HT.

Tabela 3.4: Distribuição amostral do estimador $AAS_{(s)}$

p	$\hat{t}$	$(\hat{t}-300)^2$
0,1666	62	56.644
0,1666	70	52.900
0,1666	512	44.944
0,1666	88	44.944
0,1666	530	52.900
0,1666	538	56.644
Esperança	300	51.496

É possível introduzir ao estimador HT a falta de informação sobre a distribuição inicial de seleção, obtendo os mesmos valores encontrados para a $AAS_{(s)}$. Por tanto, além de muito mais preciso, o estimador HT é a generalização dos planos amostrais sem reposição (Lohr, 1999).

3.4 Generalização do estimador HT

Além do uso nos casos de probabilidades de seleção diferente, o estimador é o caso geral dos planos amostrais sem reposição. Isso pode ser visto através da fórmula da variância desses planos, pois são casos particulares da Equação 3.3.

3.4.1 Aleatória simples

Seja a Equação 3.3 para a variância proposta por Thompson (1952). Em uma amostragem aleatória simples, a probabilidade de um elemento pertencer a amostra é $\pi_i = \frac{n}{N}$, sendo n e N o tamanho da amostra e da populacão, respectivamente. Com isso tem-se:

$$V(\hat{t}_\pi) = \sum_i^N \frac{1 - \frac{n}{N}}{\frac{n}{N}} y_i^2 + \sum_i^N \sum_{j \neq i}^N \frac{\left[\frac{n(n-1)}{N(N-1)} - \frac{n^2}{N^2}\right]}{\frac{n^2}{N^2}} y_i y_j \tag{3.7}$$

$$V(\hat{t}_\pi) = \sum_i^N \frac{N-n}{n} y_i^2 + \sum_i^N \sum_{j \neq i}^N \frac{n-N}{(N-1)n} y_i y_j \tag{3.8}$$

Usando o seguinte resultado:

$$(N\bar{y})^2 = \left[\sum_i^N y_i\right]^2 = \sum_i^N y_i^2 + \sum_i^N \sum_{j \neq i}^N y_i y_j \tag{3.9}$$

tem-se:

$$V(\hat{t_\pi}) = \sum_i^N \frac{N-n}{n} y_i^2 - \frac{N-n}{(N-1)n}\left[(N\bar{y})^2 - \sum_i^N y_i^2\right] \tag{3.10}$$

Colocando em evidência o termo em comum e juntando os somatórios, tem-se:

$$V(\hat{t}_\pi) = \frac{N-n}{n}\left[\sum_i y_i^2 - \frac{1}{N-1}\left(N^2\bar{y}^2 - \sum_i y_i^2\right)\right] \tag{3.11}$$

Colocando a parte dentro do somatório no mesmo denominador e cancelando alguns termos, tem-se:

$$V(\hat{t}_\pi) = \frac{N-n}{n} N \left(\frac{\sum_i^N y_i^2 - N\bar{y}^2}{N-1} \right) \tag{3.12}$$

O último termo é S^2, variância populacional de y. Substituindo por tal, tem-se:

$$V(\hat{t}_\pi) = N^2 \frac{N-n}{N} \frac{S^2}{n} = V(\hat{t}) \tag{3.13}$$

Assim, atravéz do estimador HT, pode-se chegar ao resultado da $AAS_{(s)}$.

3.4.2 Estratificada

Para o caso estratificado, o procedimento é similar ao da amostra aleatória simples, apenas substituindo n por n_h. Com isso, a estimativa HT no estrato h é:

$$V(\hat{t}_{\pi h}) = N_h^2 (1 - f_h) \frac{S_h^2}{n_h} \tag{3.14}$$

Sendo f_h o fator de correção no estrato, $\frac{n_h}{N_h}$. Como o processo de seleção nos estratos se dá de forma independente, isto é, não há covariância entre os estratos, a variância do estimador é a soma das variâncias dentro de cada estrato. A estimativa da variância populacional é dada por:

$$V(\hat{t}_{\pi str}) = \sum_{h=1}^{H} N_h^2 (1 - f_h) \frac{S_h^2}{n_h} \tag{3.15}$$

Verifica-se assim a particuliaridade da amostragem estratificada com relação ao estimador Horvitz Thompson, (Lohr, 1999).

Capítulo 4

Algoritmo desenvolvido

O ambiente **IML** (***I**nteractive **M**atrix **L**anguage*) é uma poderosa plataforma de programação no **SAS** com uma linguagem flexível, dinâmica e iterativa semelhante a encontrada no **MATLAB** e **R**. Sua capacidade de criar funções e sub-rotinas combinada com a de manipulação e armazenamento de dados do **SAS**, torna esse módulo o ambiente perfeito para a implementação do estimador Horvitz Thompson.

4.1 Macro

Foram desenvolvidas as macros **%HT_general** e **%HT** no ambiente **SAS/IML**. A primeira tem por objetivo confirmar os resultados teóricos apresentados, demonstrando que tanto as estimativas do parâmetro quanto sua variância são não viesados. A segunda gera a estimativa HT para uma dada amostra.

4.1.1 %HT_general

Primeira parte da macro gera todas as permutações para um vetor de tamanho igual ao da amostra. Esse é o único processo do algoritmo que foi usado funções fora do **IML**, pois esse módulo do **SAS** não aceita recursividade.

```
/** generate all permutations of n elements, in order **/
data perm&n (drop=i);array a{&n};
do i = 1 to &n; a[i]=i; end; /** initialize **/
do i = 1 to fact(&n);
call allperm(i, of a[*]);
output;end;run;
```

O segundo passo do processo é gerar todas as combinação C_n^N e também C_2^n. Com essas duas matrizes e a desenvolvida no passo anterior é possível calcular os valores do estimador HT.

```
proc iml;
start combinacao(pop,tamamostra);
n=pop;m=tamamostra;vetor=1:n;e=0;
h=m;indice=1:h;vec=vetor[indice];
count=fact(n)/(fact(m)*fact(n-m));
matriz=vec;do i=1 to count-1;matriz=matriz||vec;
end;i=2;criterio=n-m+1;do k=1 to count;aa=indice[1];
if(aa^=criterio) then do;
if (e<n-h) then do;
h=1;e=indice[m];G=1;end;else do;e=indice[m-h];h=h+1;g=1:h;
end;indice[m-h+g]=e+g;vec=vetor[indice];matriz[,i]=vec;i=i+1;
end;end;return(matriz);finish combinacao;
```

Nesse ponto da programação são calculados os principais subsídios para o cálculo do estimador HT, a matriz de probabilidade conjunta de inclusão. A partir dela é possível calcular as probabilidas de inclusão para cada indivíduo através da Equação 2.5.

```
l1=comb(pop,n);l2=fact(N);l3=comb(n,2);
do i=1 to l1;vec1=comb[,i]‘;do j=1 to l2;
vec2=perm[j,];ind=vec1[vec2];pijk=1;soma=0;
do k=1 to n;pk=p[1,ind[k]]/(1-soma);soma=soma+p[1,ind[k]];
pijk=pijk*pk;end;ppp=ppp||pijk;do k=1 to l3;vec3=comb2[,k]‘;
ind=vec1[vec2[vec3]];pp[ind[1],ind[2]]=pp[ind[1],ind[2]]+pijk;
pp[ind[2],ind[1]]=pp[ind[2],ind[1]]+pijk;end;end;end;do i=1 to pop;
pi[i]=sum(Pp[,i])/(n-1);end;
```

Além disso, é possível comparar com os resultados da aleatória simples e notar a eficiência e precisão do estimador. A saída do programa compara os resultados obtidos pelo estimador HT com o $ASS_{(s)}$. Caso o plano amostral seja estratificado, todo

o procedimento é repetido para o número de estratos e os resultado são combinados como mostra a Seção 3.4.2. Por isso existe uma única macro que foi inteiramente desenvolvidar para o caso estratificado.

4.1.2 %HT

A única diferença do das duas macro é que o **%HT** fornece a estimativa apenas para a amostra fornecida.

4.2 Exemplos HT

A seguir serão apresentados alguns exemplos da saída do estimador HT. Para que a macro funcione é necessário informar a tabela, *data set*, variável de interesse, variável auxiliar e o tamanho amostral. O exemplo a seguir mostra como conseguir os resultados do estimador HT para o problema dos supermercados usando amostra de tamanho $n = 2$ e utilizando as informações amostrais de tamanho do supermercado e no caso da falta dessa informação. Os dados estão na Tabela 3.1.

4.2.1 HT n=2

O primeiro valor obtido é a matriz de probabilidade conjunta dos elementos da população. Esse mesmos valores podem ser encontrados na Tabela 3.2. Os valores da probabilidade inclusão são obtidos através de qualquer das marginais e a soma dessas probabilidades é o tamanho da amostra.

```
                          PP
                 0 0.0172619 0.0269231 0.1458333
         0.0172619         0 0.0556319  0.297619
         0.0269231 0.0556319         0 0.4567308
         0.1458333  0.297619 0.4567308         0

           PI                                    SUMPI
 0.1900183 0.3705128 0.5392857 0.9001832         2
```

A segunda parte do *output* do programa mostra todas as $n!C_n^N$ estimativas HT, as variâncias de 3.5 (Yates and Grundy, 1953), e as de 3.4 (Thompson, 1952), que podem ser negativas. Essas informações são mostradas no *output* apenas se o número de arranjos for menor que 50, pois essas e outras informações são armazenadas no *data set* chamado **estimativas** para posteriores análises.

```
                    Number of Samples:         12

                                HTS
            COL1       COL2       COL3       COL4       COL5       COL6
ROW1   111.8684   111.8684   102.3925   102.3925   330.0560   330.0560

                                HTS
            COL7       COL8       COL9      COL10      COL11      COL12
ROW1    98.4826    98.4826   326.1461   326.1461   316.6701   316.6701

                              VARHTC
            COL1       COL2       COL3       COL4       COL5       COL6
ROW1    47.0638    47.0638   502.8143   502.8143 7,939.7510 7,939.7510

                              VARHTC
            COL7       COL8       COL9      COL10      COL11      COL12
ROW1   232.7159   232.7159 5,744.0610 5,744.0610 3,259.7842 3,259.7842

                              VARHTN
            COL1       COL2       COL3       COL4       COL5       COL6
ROW1 -14691.483 -14691.483 -10832.071 -10832.071 4,659.3028 4,659.3028

                              VARHTN
            COL7       COL8       COL9      COL10      COL11      COL12
ROW1 -9705.1479 -9705.1479 5,682.8026 5,682.8026 6,782.8174 6,782.8174
```

Nesse ponto o programa informa qual é o parâmetro de interesse e a esperança das estimativas HT e suas variâncias. É possível analisar que os estimadores amostrais

são não viesados e comparar a eficiência sobre o plano amostral da aleatória simples.

```
_TOTAL_
    300

E_HTS    E_VHTC    E_VHTN
  300 4383.5622 4383.5622

   VARHT    VARSRS
4383.5622     51496

DEFF
8.51%
```

4.2.2 $AAS_{(s)}$ n=2 através do estimador HT

Como mostrado no capítulo anterior, todos os resultados da $AAS_{(s)}$ podem ser obtidos substituindo a variável auxiliar com valores iguais para todos os elementos da população. Com isso é possível ainda certificar-se da generalidade desse estimador.

```
                    PP
         0 0.1666667 0.1666667 0.1666667
0.1666667         0 0.1666667 0.1666667
0.1666667 0.1666667         0 0.1666667
0.1666667 0.1666667 0.1666667         0

  PI                                  SUMPI
 0.5       0.5       0.5       0.5       2


  Number of Samples:         12

                    HTS
          COL1      COL2      COL3      COL4      COL5      COL6
ROW1   62.0000   62.0000   70.0000   70.0000  512.0000  512.0000

                    HTS
          COL7      COL8      COL9     COL10     COL11     COL12
ROW1   88.0000   88.0000  530.0000  530.0000  538.0000  538.0000
```

VARHTC

	COL1	COL2	COL3	COL4	COL5	COL6
ROW1	162.0000	162.0000	338.0000	338.0000	109512.000	109512.000

VARHTC

	COL7	COL8	COL9	COL10	COL11	COL12
ROW1	32.0000	32.0000	101250.000	101250.000	97682.0000	97682.0000

VARHTN

	COL1	COL2	COL3	COL4	COL5	COL6
ROW1	162.0000	162.0000	338.0000	338.0000	109512.000	109512.000

VARHTN

	COL7	COL8	COL9	COL10	COL11	COL12
ROW1	32.0000	32.0000	101250.000	101250.000	97682.0000	97682.0000

TOTAL
300

E_HTS	E_VHTC	E_VHTN
300	51496	51496

VARHT	VARSRS
51496	51496

DEFF
100.00%

Capítulo 5

Considerações finais

Como apresentado na proposta de trabalho final, o algoritmo do estimador Horvitz Thompson foi desenvolvido no **SAS** para amostragem estratificada, mas todos os outros planos amostrais em 1 estágio podem ser obtidos. O desenvolvimento desse estimador além de reforcar os conceitos de estimador e a revisão bibliográfica da teoria amostral mostrou o poder do mesmo como importante ferramenta estatística na amostragem em pequenas populações.

Foram desenvolvidas a macro **%HT_general**, **%HT** ambos nos **SAS** e um algoritmo no **R**. A primeira confirma os resultados teórios de Thompson (1952), as estimativas do parâmetro e sua variância são não viesados. A segunda gera o resultado da estimativa HT para a amostra especificada e o procedimento no **R** apresenta apenas os resultados populacionais.

O estimador Horvitz Thompson tem estimativas mais precisas utilizando o poder computacional e as informações auxiliares. Constatou-se para o estudo de caso uma eficiência mais de 90% sobre $ASS_{(s)}$. Esse estimador é indicado para pequenos tamanho de amostras em populações de tamanho finito onde mostrou-se eficiente.

Referências Bibliográficas

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

Durbin, J. (1953). Some results in sampling theory when units are selected with unequal probabilities. Vol. 15, No. 2, pp. 262-269.

Hansen, M. H. & Hurwitz, W. N. (1943). On the theory of sampling from finite population. *The Annals of Mathematical Statistics*. pp. 333-362.

Lohr, S. L. (1999). *Sampling: Design and Analysis*. Duxbury Press.

SAS (2008). *SAS Institute Inc.* Cary, NC: SAS Institute Inc. Version 9.2.

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

Särndal, C. E., Swensson, B., & Wretman, J. (1992). *Model Assisted Survey Sampling*, (2th ed.). Springer Series in Statistics.

Thompson, D. G. H. . D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*. pp. 663-685.

Yates, F. & Grundy, P. M. (1953). Selection without replacement from within strata with probability proportinal to size. *Journal of the Royal Statistical Society*. pp. 253-261.

R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Apêndice A

%HT_General - SAS

```
%macro HT_general(tab=,var=,aux=,n=,str=);
/** generate all permutations of n elements, in order **/
data perm&n (drop=i);array a{&n};
do i=1 to &n; a[i]=i; end; /** initialize **/
do i=1 to fact(&n);call allperm(i, of a[*]);
output;end;run;

PROC IML;
/***COMBINATION FUNCTION - BEGIN***/;
start combination(pop,tamamostra);
N=pop;m=tamamostra;vetor=1:n;
e=0;h=m;indice=1:h;vec=vetor[indice];
count =fact(n)/(fact(m)*fact(n-m));
matriz=vec;do i=1 to count-1;
matriz=matriz||vec;end;
i=2;criterio=n-m+1;
do k=1 to count;aa=indice[1];
if (aa^=criterio) then
do;if(e<n-h) then
do;h=1;e=indice[m];g=1;
end;else
do;e=indice[m-h];
           h=h+1;g=1:h;
end;indice[m-h+G]=e+G;
        vec=vetor[indice];
        matriz[,i]=vec;
        i=i+1;end;end;
return(matriz);
finish combination;

/***** COMBINATION FUNCTION - END **********;

/***** inserting data and using comb/perm ****/
str=1;%if &str ^= %then %do;use &tab var{&str};
read all;tstr=nrow(&str);str=unique(&str);
%end;do nstr=1 to ncol(str);st=str[nstr];
%if &str ^= %then %do;print st;%end;
use &tab var{&var};%if &str ^= %then %do;
use &tab var{&var &str} where(&str=:st);%end;
read all into y;y=y[,1];use &tab var{&aux &str};
%if &str ^= %then %do;use &tab var{&aux &str} where(&str=:st);%end;
read all into trab;trab=trab[,1];use perm&n;
read all into perm;y=y‘;
trab=trab‘;pop=ncol(y);
n=&n;comb=combination(POP,n);
comb2=combination(n,2);

/******** joint probalility matrix  **********/
P=trab/trab[+];pi=J(1,pop,0);PP=J(pop,pop,0);
_Total_=y[+];l1=comb(pop,n);l2=fact(N);
l3=comb(n,2);
free ppp hts varhtc varhtn pij;
do i=1 to l1;vec1=comb[,i]‘;do j=1 to l2;
vec2=perm[j,];ind=vec1[vec2];pijk=1;
soma=0;do k=1 to n;pk=p[1,ind[k]]/(1-soma);
soma=soma+p[1,ind[k]];pijk=pijk*pk;
end;ppp=ppp||pijk;/*probability of selected sample*/
do k=1 to l3;vec3=comb2[,k]‘;
ind=vec1[vec2[vec3]];
pp[ind[1],ind[2]]=pp[ind[1],ind[2]]+pijk;
pp[ind[2],ind[1]]=pp[ind[2],ind[1]]+pijk;
end;end;end;
do i=1 to pop;pi[i]=sum(pp[,i])/(n-1);
end;

/************ HT Estimator ***********/;
um=j(n,1,1/n);
free varsrsa srs amostra_pop;

do i=1 to l1;vec1=comb[,i];
do j=1 to l2;vec2=perm[j,];
ind=vec1[vec2];
hts=hts || sum(y[ind]/pi[ind]);
srs=srs || (pop/n)*sum(y[ind]);
amostra_pop=amostra_pop||y[ind];
amos=y[ind];media=amos‘*um;
varm=(amos‘-media)*(amos‘-media)‘/(n-1);
varsrsa=varsrsa||(varm/n)*pop**2*(1-n/pop);

vp=0;do ii=1 to (n-1); do jj=(ii+1) to n;
vp=((pi[ind][ii]*pi[ind][jj]-pp[ind,ind][ii,jj])/
pp[ind,ind][ii,jj])*((y[ind][ii]/pi[ind][ii])
-(y[ind][jj]/pi[ind][jj]))**2+vp;
end;end;varhtc=varhtc||vp;
var1=0;do ii=1 to n;
var1=((1-pi[ind][ii])/(pi[ind][ii]**2))*y[ind][ii]**2+var1;
end;var2=0;do ii=1 to (n-1); do jj=(ii+1) to n;
var2=((pp[ind,ind][ii,jj]-(pi[ind][ii]*pi[ind][jj])
/(pp[ind,ind][ii,jj]*
pi[ind][ii]*pi[ind][jj]))*(y[ind][ii]*y[ind][jj])+var2;
end;end;varhtn=varhtn||var1+2*var2;end;end;

amostra_pop2=amostra_pop‘;

do i=1 to (pop-1);do j=(i+1) to pop;
pij= pij || PP[i,j];
end;end;

/******* Var HT **********/;
var1=0;do i=1 to pop;
var1=((1-pi[i])/pi[i])*(y[i]**2)+var1;
end;var2=0;cont=0;
do i=1 to (pop-1);do j=(i+1) to pop;
cont=cont+1;
var2=2*((pij[cont]-(pi[i]*pi[j]))/
(pi[i]*pi[j]))*(y[i]*y[j])+var2;
end;end;varHT=var1+var2;

/*average and variance of SRS */;
um=j(pop,1,1/pop);
media=y*um;
var=(y-media)*(y-media)‘/(pop-1);
varSRS=(var/n)*pop**2*(1-n/pop);
```

```
/*average and variance of HT*/
E_hts=hts*ppp‘;
E_vhtc=varhtc*ppp‘;
E_vhtn=varhtn*ppp‘;

print PP;
sumpi=pi[+];
print Pi sumpi;
print ’Number of Samples: ’ (l1*l2);
if l1*l2<100 then do;
print hts[format=comma10.4];
print varhtc[format=comma10.4];
print varhtn[format=comma10.4];
end;
print _Total_;

print E_hts E_vhtc E_vhtn;
print varHT varSRS;
deff=varHT/varSRS;
print deff[format=percent10.2];

strata=j(ncol(hts),1,st)‘;
if nstr=1 then do;
Estimates=amostra_pop2||hts‘||srs‘||ppp‘||varhtc‘||varhtn‘
||varsrsa‘||strata‘;
end;
else do;
Estimates=Estimates//(amostra_pop2||hts‘||srs‘||ppp‘||varhtc‘
||varhtn‘||varsrsa‘||strata‘);
end;
end;

%if &str ^= %then %do;

print "************  stratified estimates   ********************";
hts=Estimates[,ncol(amostra_pop2)+1];
ppp=Estimates[,ncol(amostra_pop2)+3];
varhtc=Estimates[,ncol(amostra_pop2)+4];
varhtn=Estimates[,ncol(amostra_pop2)+5];
E_hts=hts‘*ppp;
E_vhtc=varhtc‘*ppp;
E_vhtn=varhtn‘*ppp;
print E_hts E_vhtc E_vhtn;

varAE=0;use &tab var{&var};
read all into y;popt=nrow(y);
read all into y;close &tab;

do i=1 to ncol(str);st=str[i];
use &tab var{&var &str} where(&str=:st);
read all into y;y=y[,1];
pop=nrow(y);um=j(pop,1,1/pop);
media=y‘*um;var=(y-media)‘*(y-media)/(pop-1);

**********Stratified Sample**********************;
varAE=varAE+popt**2*(var/n)*(pop/popt)**2*(1-n/pop);
end;print varAE;use &tab var{&var};
read all into y;pop=nrow(y);um=j(pop,1,1/pop);
media=y‘*um;n=ncol(str)*n;var=(y-media)‘*(y-media)/(pop-1);
*SRS;varSRS=(var/n)*pop**2*(1-n/pop);
print varSRS;%end;
colnames="selection1":"selection&n"||{"hts" "srs"
"prob" "varhtc" "varhtn" "varsrs" "strata"};
create Estimates from Estimates[colname=colnames];
append from Estimates;quit;%mend HT_general;
```

Apêndice B

%HT - SAS

```
%macro HT(tab=,var=,str=,tab2=,aux=,n=,str2=);
/** generate all permutations of n elements, in order **/
data perm&n (drop=i);array a{&n};
do i = 1 to &n; a[i]=i; end; /** initialize **/
do i = 1 to fact(&n);call allperm(i, of a[*]);
output;end;run;

PROC IML;
/***COMBINATION FUNCTION - BEGIN***/;
start combination(pop,tamamostra);
N=pop;m=tamamostra;vetor=1:n;
e=0;h=m;indice=1:h;vec=vetor[indice];
count =fact(n)/(fact(m)*fact(n-m));
matriz=vec;do i=1 to count-1;
matriz=matriz||vec;end;
i=2;criterio=n-m+1;
do k=1 to count;aa=indice[1];
if (aa^=criterio) then
do;if(e<n-h) then
do;h=1;e=indice[m];g=1;
end;else
do;e=indice[m-h];
           h=h+1;g=1:h;
end;indice[m-h+G]=e+G;
        vec=vetor[indice];
        matriz[,i]=vec;
        i=i+1;end;end;
return(matriz);
finish combination;

/***** COMBINATION FUNCTION - END **********;

/***** inserting data and using comb/perm ****/
str=1;%if &str ^= %then %do;
use &tab2 var{&str};read all;
tstr=nrow(&str);str=unique(&str);
%end;do nstr=1 to ncol(str);
st=str[nstr];%if &str ^= %then %do;print st;
%end;use &tab var{&var};%if &str ^= %then %do;
use &tab var{&var &str} where(&str=:st);%end;
read all into y;y=y[,1];
use &tab2 var{&aux};%if &str ^= %then %do;
use &tab2 var{&aux &str} where(&str=:st);%end;
read all into trab;trab=trab[,1];use perm&n;
read all into perm;y=y`;trab=trab`;

pop=ncol(trab);n=&n;comb=combination(POP,n);
comb2=combination(n,2);

/******** joint probalility matrix  **********/
p=trab/trab[+];pi=j(1,pop,0);PP=j(pop,pop,0);
l1=comb(pop,n);l2=fact(n);l3=comb(n,2);
free ppp hts varhtc varhtn pij;
do i=1 to l1; vec1=comb[,i]`;
do j=1 to l2; vec2=perm[j,];
ind=vec1[vec2];pijk=1;soma=0;
do k=1 to n;pk=p[1,ind[k]]/(1-soma);
soma=soma+p[1,ind[k]];pijk=pijk*pk;
end;ppp=ppp||pijk;/*probability of selected sample*/
do k=1 to l3;vec3=comb2[,k]`;ind=vec1[vec2[vec3]];
pp[ind[1],ind[2]]=pp[ind[1],ind[2]]+pijk;
pp[ind[2],ind[1]]=pp[ind[2],ind[1]]+pijk;
end;end;end;
do i=1 to pop;pi[i]=sum(Pp[,i])/(n-1);end;

/**********************SAMPLE******************/
ind=1:n;ps=pi[ind];pps=pp[ind,ind];
ht=sum(y`/ps);
/****************** HT Variance 1 *************/
var1=0;do i=1 to n;
var1=((1-ps[i])/(ps[i]**2))*y[i]**2+var1;
end;var2=0;do i=1 to (n-1); do j=(i+1) to n;
var2=((pps[i,j]-(ps[i]*ps[j]))/(pps[i,j]*ps[i]*ps[j]))*
(y[i]*y[j])+var2;
end;end;varhtn=var1+2*var2;

/************ HT Variance 2 ****************************/
varhtc=0;do i=1 to (n-1); do j=(i+1) to n;
varhtc=((ps[i]*ps[j]-pps[i,j])/pps[i,j])*
((y[i]/ps[i])-(y[j]/ps[j]))**2+varhtc;
end;end;um=j(n,1,1/n);
media=y*um;vsrs=(y-media)*(y-media)`/(n-1);
varsrs=(vsrs/n)*pop**2*(1-n/pop);srs=(pop/n)*y[+];
print PP;sumpi=pi[+];print Pi sumpi;print y;
print ht srs;deffn=varhtn/varsrs;deffc=varhtc/varsrs;
print varsrs varhtn varhtc;if deffn>0 then do;
print deffn[format=percent10.2];end;
print deffc[format=percent10.2];

strata=j(ncol(ht),1,st)`;if nstr=1 then do;
Estimates=ht||ppp[1]||varhtc||varhtn||strata;
end;else do;Estimates=Estimates//(ht||ppp[1]||varhtc||varhtn||strata);
end;end;%if &str ^= %then %do;
print "*************stratified estimates***************";
ht=Estimates[,1];ppp=Estimates[,2];varhtc=Estimates[,3];
varhtn=Estimates[,4];ht_st=ht[+];vhtc_st=varhtc[+];
vhtn_st=varhtn[+];print ht_st vhtc_st vhtn_st;
varAE=0;use &tab2 var{&aux};read all into y;popt=nrow(y);
read all into y;close &tab;do i=1 to ncol(str);st=str[i];
use &tab var{&var &str} where(&str=:st);read all into y;
use &tab2 var{&aux &str} where(&str=:st);read all into w;
popw=nrow(w);y=y[,1];pop=nrow(y);um=j(pop,1,1/pop);
media=y`*um;var=(y-media)`*(y-media)/(pop-1);

/****************** AE **************************/
varAE=varAE+popt**2*(var/n)*(popw/popt)**2*(1-n/popw);
end;print varAE;use &tab var{&var};
read all into y;use &tab2 var{&aux};
read all into w;popt=nrow(w);pop=nrow(y);
um=j(pop,1,1/pop);media=y`*um;
n=ncol(str)*n;var=(y-media)`*(y-media)/(pop-1);
/****************** SRS**************************/
varSRS=(var/n)*pop**2*(1-n/popt);print varSRS;%end;
quit;%mend HT;
```

Apêndice C

HTGeral - R

```
troca=function(matriz,i,col,a){
nn=ncol(matriz)
nnn=factorial(nn)
left=matriz[i,nn-col+1]
right=matriz[i,nn-col+1+a]
matriz[i,nn-col+1]=right
matriz[i,nn-col+1+a]=left
if(i<=(nnn-1)){
if(a!=0 |col==2){
ant=matriz[i,]
matriz[i+1,]=ant
}
}
if(col>=3){
agora2=dentro(matriz,col-1,i)
matriz=agora2[[1]]
i=agora2[[2]]
}
list(matriz,i)
}
########Função que começa a recursividade
dentro=function(matriz,col,contador){
a=0
while(a<=col-1){
agora=troca(matriz,contador,col,a)
matriz=agora[[1]]
contador=agora[[2]]
nn=ncol(matriz)
nnn=factorial(nn)
if(col==2 & contador<=(nnn-1)){
contador=contador+1
}
a=a+1
}
if(a==col & contador!=(nnn)){
vec=matriz[contador,(nn-col+1):nn]
matriz[contador,(nn-col+1):nn]=sort(vec)
}
list(matriz,contador)
}
#####Função que agrega as duas funções e mostra as permutações
permuta=function(x){
n=length(x)
matriz=matrix(x,factorial(n),n,byrow=T)
matriz=dentro(matriz,n,1)
matriz
}
ht=function(y,trab,n,estrato=rep(1,length(y))){
#/******** matriz de probabilidade conjunta **********/
pop=length(y)
perm=permuta(1:n)[[1]]
comb=combn(pop,n)
comb2=combn(n,2)
p=trab/sum(trab)
pi=rep(0,pop)
pp=matrix(0,pop,pop)
Total=sum(y)
ppp=numeric()
l1=choose(pop,n)
l2=factorial(n)
l3=choose(n,2)
for(i in 1:l1){
vec1=comb[,i]
for(j in 1:l2){
vec2=perm[j,]
ind=vec1[vec2]
pijk=1
soma=0
for(k in 1:n){
pk=p[ind[k]]/(1-soma)
soma=soma+p[ind[k]]
pijk=pijk*pk
}
ppp=c(ppp,pijk)#/*Probabilidade de seleção da amostra*/

for(k in 1:l3){
vec3=comb2[,k]
ind=vec1[vec2[vec3]]
pp[ind[1],ind[2]]=pp[ind[1],ind[2]]+pijk
pp[ind[2],ind[1]]=pp[ind[2],ind[1]]+pijk
}

}

for(i in 1:pop){
pi[i]=sum(pp[,i])/(n-1)
}
pi
hts=numeric()
ass=numeric()
amostra_pop=numeric()
varassa=numeric()
varhtc=numeric()
varhtn=numeric()
pij=numeric()
for(i in 1:l1){
vec1=comb[,i]
for(j in 1:l2){
vec2=perm[j,]
ind=vec1[vec2]
hts=c(hts,sum(y[ind]/pi[ind]))
ass=c(ass,(pop/n)*sum(y[ind]))
amostra_pop=rbind(amostra_pop,y[ind])
amos=y[ind]
varm=var(amos)
varassa=c(varassa,((varm/n)*((pop)^2)*(1-n/pop)))
vp=0
for(ii in 1:(n-1)){
for(jj in (ii+1):n){
vp=((pi[ind][ii]*pi[ind][jj]-pp[ind,ind][ii,jj])/pp[ind,ind][ii,jj])*
((y[ind][ii]/pi[ind][ii])-(y[ind][jj]/pi[ind][jj]))^2+vp
}
}
varhtc=c(varhtc,vp)
var1=0
for(ii in 1:n){
```

```
var1=((1-pi[ind][ii])/(pi[ind][ii]^2))*(y[ind][ii]^2)+var1
}
var2=0
for(ii in 1:(n-1)){
for(jj in (ii+1):n){
var2=((pp[ind,ind][ii,jj]-(pi[ind][ii]*pi[ind][jj]))/
(pp[ind,ind][ii,jj]*pi[ind][ii]*
pi[ind][jj]))*(y[ind][ii]*y[ind][jj])+var2
}
}
varhtn=c(varhtn,var1+2*var2)
}
}
for(i in 1:(pop-1)){
for(j in (i+1):pop){
pij=c(pij,pp[i,j])
}
}
#Var HT
var1=0
for(i in 1:pop){
var1=((1-pi[i])/pi[i])*(y[i]^2)+var1
}
var2=0
cont=0
for(i in 1:(pop-1)){
for(j in (i+1):pop){
cont=cont+1
var2=2*((pij[cont]-(pi[i]*pi[j]))/(pi[i]*pi[j]))*(y[i]*y[j])+var2
}
}
varHT=var1+var2

#/*media e variancia e variancia da variancia ASS */;
varr=var(y)
varASS=(varr/n)*(pop^2)*(1-n/pop)
#/*variancia da variancia HT*/
E_hts=hts%*%ppp
E_vhtc=varhtc%*%ppp
E_vhtn=varhtn%*%ppp
sumpi=sum(pi)
deff=varHT/varASS
BancoEstimativas=as.data.frame(round(cbind(amostra_pop,hts,ass,ppp,varhtc,varhtn,varassa,estrato),2))
nomeamostra=numeric()
for(ll in 1:n){
nomeamostra=c(nomeamostra,paste("Amostra",ll,sep=""))
}
names(BancoEstimativas)[1:n]=nomeamostra
return(list(pp,pi,sumpi,Total,hts,varhtc,varhtn,E_hts,E_vhtc,E_vhtn,varHT,varASS,deff,BancoEstimativas))
}
htgeral=function(y,trab,amostra,estrato=rep(1,length(y))){
a=numeric()
banco=as.data.frame(cbind(y,trab,estrato))
nestratos=unique(estrato)
if(length(nestratos)>1){
for(k in nestratos){
banco1=subset(banco,estrato==k)
a=c(a,ht(banco1$y,banco1$trab,amostra[k],banco1$estrato))
}
ordem=rev(order(amostra))
pop=0
e.ht=0
e.vhtc=0
e.vhtn=0
e.vht=0
e.vass=0
nomes=names(a[[(ordem[1]-1)*14+14]])
bancoEstimativa=numeric()
for(k in ordem){
pop=pop+a[[(k-1)*14+4]]
e.ht=e.ht+a[[(k-1)*14+8]]
e.vhtc=e.vhtc+a[[(k-1)*14+9]]
e.vhtn=e.vhtn+a[[(k-1)*14+10]]
e.vht=e.vht+a[[(k-1)*14+11]]
e.vass=e.vass+a[[(k-1)*14+12]]
if(amostra[k]<max(amostra)){
aa=numeric()
for(kk in 1:(max(amostra)-amostra[k])){
aa=cbind(aa,rep(NA,length(a[[(k-1)*14+14]][,1])))
}
aaa=cbind(a[[(k-1)*14+14]][,(1:amostra[k])],aa,a[[(k-1)*14+14]][,-(1:amostra[k])])
names(aaa)=nomes
rm(aa)
}else{
aaa=a[[(k-1)*14+14]]
}
bancoEstimativa=rbind(bancoEstimativa,aaa)
rm(aaa)
}
deff=e.vht/e.vass
return(list(pop,e.ht,e.vhtc,e.vhtn,e.vht,e.vass,deff,bancoEstimativa))
}else{
ht(y,trab,amostra)
}
}
```