# Alguns Métodos de Amostragem para Populações Raras e Agrupadas

Luis Henrique Teixeira Alves Affonso

DISSERTAÇÃO APRESENTADA
AO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DA
UNIVERSIDADE DE SÃO PAULO
PARA
OBTENÇÃO DO GRAU
DE
MESTRE EM CIÊNCIAS

Área de Concentração: **Estatística**

Orientadora: **Profa. Dra. Lúcia Pereira Barroso**

São Paulo, fevereiro de 2008.

**Alguns métodos para populações raras e agrupadas**

Este exemplar corresponde a redação final da dissertação devidamente corrigida e defendida por Luis Henrique Teixeira Alves Affonso e aprovada pela comissão julgadora

São paulo, 11 de março de 2008.

Banca Examinadora:

- Profa. Dra. Lúcia Pereira Barroso (presidente) - IME/USP
- Prof. Dr. Paul G. Kinas - FURG
- Prof. Dr. Wilton O. Bussab - FGV/SP

*Aos meus pais Eudir e Marisa*

# Agradecimentos

Gostaria de agradecer a todos aqueles que de certa forma contribuiram para que eu tivesse condições de realizar esse trabalho, em especial o acompanhamento excelente da minha orientadora Lúcia, pois foi através dos questionamentos dela e da banca que o meu trabalho se consolidou.

Gostaria também de agradecer o professor Ademir por me incentivar a continuar estudando, Roberto que me orientou no começo de minha carreira. Seria muito complicado inciar o mestrado não fosse o apoio dos meus colegas da Nielsen, em especial Enzo que me abriu essa porta e Antonio Carlos com seu apoio incondicional.

Além disso queria agradecer também o apoio dos meus colegas do IME em especial Augusto cuja ajuda foi indispensável nos estudos de ingresso.

Nada disso teria sido possível se não fosse o suporte emocional da minha família, minha mãe pela ajuda com as pequenas coisas que fazem toda a diferença e meu pai que me deu um exemplo a seguir.

# Resumo

Em diversos levantamentos científicos, nos deparamos com a dificuldade de coletar os dados devido ao objeto em estudo ser de difícil observação, como por exemplo em estudos com indivíduos portadores de doenças raras, ou dotados de um comportamento evasivo, ou ainda indivíduos que distribuem-se de maneira geograficamente esparsa. Neste trabalho estudamos esquemas de amostragem voltados para populações raras com especial atenção às populações raras e agrupadas. Nos aprofundamos nas técnicas de amostragem por conglomerados adaptativos e amostragem seqüencial em dois estágios, fornecendo ao leitor subsídio teórico para entender os fundamentos das técnicas, bem como compreender a eficácia de seus estimadores apresentada em estudos de simulações. Em nossos estudos de simulação, mostramos que a técnica de amostragem seqüencial em dois estágios não apresenta perdas de eficiência quando o agrupamento dos elementos é menor. Entretanto, os estudos comparativos revelam que quando a população é rara e agrupada, a eficiência para a amostragem por conglomerados adaptativos é maior na maioria das parametrizações utilizadas. Ao final deste trabalho, fornecemos recomendações para as situações a respeito do conhecimento da raridade e agrupamento da população em estudo.

# Abstract

In many surveys we find hard observing individuals, like in rare diseases, elusive individuals or sparsely distributed individuals. This work is about sampling schemes for rare populations, more specifically rare and clustered, driving our attention to adaptive cluster sampling and two stage sequential sampling giving readers their theoretical basis and simulated efficiencies evaluation. In our simulation studies, we found that the efficiency of two-stage sequential sampling does not decrease when sample clustering is low. However, the comparison studies show that when sample is rare and clustered, adaptive cluster sampling in the majority of tested cases has better efficiency. At the end of this study, there are recommendations for each situation of knowing rarity and clustering of the population in study.

# Sumário

# Capítulo 1

# Introdução

Em diversos levantamentos científicos, nos deparamos com a dificuldade de coletar os dados devido ao objeto de estudo ser difícil de ser observado, por exemplo:

- Estudos de doenças raras
- Animais e plantas incomuns ou difíceis de capturar
- Indivíduos com perfis específicos ou difíceis de detectar
- Animais ou indivíduos que distribuem-se de maneira esparsa

Neste trabalho estudamos esquemas de amostragem voltados para populações raras. Kalton (2001) afirma que em geral as populações raras são uma fração da população total, como podemos ver por exemplo em estudos de doenças raras, em que o interesse se concentra em grupos específicos de sexo e idade. Segundo McDonald (2004), populações raras não são necessariamente aquelas que possuem poucos indivíduos e sim aquelas em que os indivíduos estão esparsamente distribuídos em grandes espaços. Podemos aplicar as abordagens aqui utilizadas a populações elusivas. A palavra elusivo significa algo que é difícil de capturar ou observar, dado um comportamento. Esse comportamento pode ser aquele de quem não gosta de se expor, que seja nômade ou tenha hábitos noturnos. As técnicas de populações raras em geral buscam maximizar a incorporação de indivíduos na amostra, portanto a estratégia de amostragem dos esquemas apresentados aqui valem tanto para populações raras quanto

elusivas. Kalton (2001) afirma que umas das questões chave para se escolher a técnica adequada para estudar as populações raras é a disponibilidade de um marco amostral. Caso um marco amostral adequado esteja disponível, podemos empregar os métodos conhecidos de amostragem e isso não será um problema. Contudo, na maioria das vezes essa informação não está disponível. São exemplos de técnicas para amostragem de populações raras (ou elusivas):

**Listas especiais**

Segundo Kalton & Sudman (1986), o uso de listas incompletas pode ser muito eficiente ao se amostrar populações raras. Através de uma amostra adicional é possível estimar o viés da lista e com isso buscar suplementar as informações da lista com o mínimo possível de unidades fora da lista pois seu custo para populações especiais é muito alto. A utilização de marcos múltiplos é interessante quando existem uma ou mais listas parciais sobre a população rara tais como registros de hospitais que podem prover informações de uma doença. Nesse caso, poderiam ser utilizados dois esquemas: fazer o levantamento no hospital e por outra parte realizar o procedimento de triagem que será explorado a seguir. Note que nesses casos podem haver duplicações dos indivíduos e portanto para esse tipo de procedimento é necessária uma abordagem específica.

**Triagem**

Segundo Kalton & Anderson (1986), a triagem é um método para identificar membros da população rara em uma população maior. Em geral é realizado um estudo com uma amostra grande tentando identificar os elementos da população rara, usualmente por telefone ou correio. Uma vez identificados os elementos da população rara, faz-se o levantamento completo com esses indivíduos, podendo-se inclusive modificar a abordagem para face a face.

**Estratificação desproporcional**

Segundo Kalton (2001), a estratificação desproporcional é utilizada quando é possível identificar os estratos com maior concentração da população rara. Conseguimos uma melhoria na

eficiência aumentando as frações amostrais nesses estratos. Veremos que as técnicas de amostragem por conglomerados adaptativos e amostragem seqüencial em dois estágios utilizam esse conceito.

### Amostragem por rede

A amostragem por rede é um recurso que aproveita algum tipo de ligação entre os indivíduos com o objetivo de minimizar o número de contatos necessários para identificar membros com o perfil raro (Kalton & Anderson (1986)). Essa ligação pode ser, por exemplo, uma relação próxima de parentesco ou até mesmo a vizinhança. A dificuldade está em que o contato selecionado deve fornecer informação acurada sobre os demais indivíduos. Além disso, as ligações entre os indivíduos devem ser claramente definidas para que seja possível calcular as probabilidades de seleção e conseqüentemente os pesos amostrais. Como um caso particular da amostragem por rede, Goodman (1961) formalizou a idéia da amostragem bola de neve. O procedimento é o seguinte: primeiro se toma uma amostra aleatória dos indivíduos de uma população finita. Cada indivíduo na amostra indica outros $k$ indivíduos distintos, como por exemplo, seus $k$-melhores amigos, $k$ pessoas a quem ele pede opiniões, etc. Os indivíduos que não foram selecionados na amostra mas foram indicados formam o segundo estágio e o procedimento continua até $s$ estágios. Os dados obtidos são colocados em uma matriz com valores 0 e 1, tal que na linha estejam os indivíduos que indicaram e na coluna os indivíduos indicados e assim é possível fazer inferências sobre os relacionamentos dentro da população.

### Estudos multipropósito

Os estudos multipropósito são uma alternativa para reduzir os custos de amostragem e são utilizados quando várias questões podem ser respondidas pela mesma amostra (Kish (1965)). Uma vantagem é que nessas pesquisas é possível identificar um número elevado de variáveis e caracterizar subclasses e suas combinações. Uma desvantagem é que os questionários longos podem gerar dificuldades na colaboração dos indivíduos respondentes.

**Amostragem em localidades**

A amostragem em localidades consiste na amostragem de pessoas que vão a locais específicos como bilbliotecas, museus, shopping centers e locais de votação. Usualmente a amostragem é conduzida tanto nas pessoas que entram no local como nas que saem. Nesse contexto existem duas unidades amostrais: visitas (as visitas que uma mesma pessoa pode realizar na localidade) e visitantes (as pessoas que frequentam um local) (Kalton (2001)).

**Amostragem adaptativa**

A amostragem adaptativa foi introduzida por Thompson (1990) como uma técnica eficiente para estimar populações raras e agrupadas. Essa técnica aproveita a idéia intuitiva de que se os elementos da população rara foram encontrados em uma área, as áreas de vizinhança têm maior probabilidade de possuírem elementos da população rara. Nos aprofundamos nesse tema no Capítulo 2.

**Amostragem seqüencial**

Segundo Kalton & Anderson (1986), a amostragem seqüencial é uma alternativa para obter estimativas razoavelmente acuradas com respeito à prevalência da população rara. Uma abordagem é realizar a seleção de uma amostra inicial suficientemente grande para ter um número desejado $n$ de elementos da população rara baseado em um chute otimista de prevalência. Ao realizarmos essa seleção, teremos um número $n'$ de elementos da população rara. Caso $n' < n$, tomamos uma amostra maior objetivando $(n - n')$ elementos raros com base na prevalência obtida na primeira amostragem. Para a amostragem seqüencial, existem outras abordagens, algumas similares à amostragem adaptativa, como a amostragem seqüencial em duas etapas, que é outra técnica voltada para populações raras e agrupadas e que exploramos no Capítulo 3.

Este trabalho tem como objetivos fornecer ao leitor subsídio teórico para entender os fundamentos das técnicas de amostragem por conglomerados adaptativos e amostragem seqüencial em dois estágios bem como compreender a eficácia de seus estimadores aplicados ao contexto das populações raras e agrupadas. No Capítulo 2, conceituamos a amostragem

por conglomerados adaptativos, passando por alguns casos particulares, tamanho amostral esperado e eficiência. Já no Capítulo 3, seguimos a estrutura proposta no capítulo anterior para explorar a técnica da amostragem seqüencial em dois estágios e no Capítulo 4, através de um estudo de simulações, exploramos o comportamento da eficiência dos estimadores das técnicas estabelecendo um comparativo. No Capítulo 5, elaboramos as conclusões do trabalho tendo em vista os resultados do Capítulo 4 e outros pontos do texto.

# Capítulo 2

# Amostragem por Conglomerados Adaptativos

Neste capítulo apresentamos diversos aspectos sobre o método de amostragem por conglomerados adaptativos tentando propiciar ao leitor os elementos fundamentais para sua utilização. A amostragem por conglomerados adaptativos introduzida por Thompson (1990) e Thompson & Seber (1996), é uma abordagem voltada para as populações raras e agrupadas. A técnica utiliza a informação dos valores observados para ter mais êxito na coleta de unidades da população rara, aumentando assim a eficiência do estimador. Isso se deve ao fato de que espera-se que é mais provável encontrar um elemento raro na vizinhança de outro elemento quando a população é agrupada. Essa metodologia foi utilizada com sucesso e é considerada uma das melhores técnicas para ser aplicada no contexto das populações raras e agrupadas (Thompson & Seber (1996); Christman (2000); Smith *et al.* (2004)).

## 2.1 Seleção Amostral

Considere uma partição da população em $N$ unidades, por exemplo um *grid* no plano, onde os elementos da população (representados pelos pontos na Figura 2.1) se distribuem em quadrados de lados iguais. Cada quadrado do *grid* (unidade) pode conter um ou mais elementos da população rara e, na maioria dos casos, nenhum.

O procedimento de seleção por conglomerados adaptativos passa pelas seguintes etapas:

1. Seleção de uma amostra inicial de $n_1$ unidades com uma metodologia probabilística, ou seja, uma amostra aleatória com probabilidades de seleção $\pi_i > 0, i = 1, 2, \ldots, N$ conhecidas previamente;

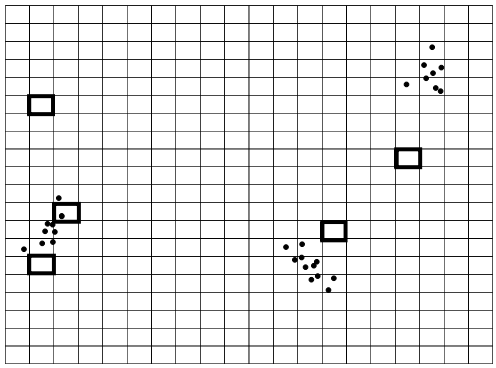

Figura 2.1: *Uma população rara no plano e uma seleção inicial ($n_1$=5)*

2. Verificação de quais unidades da seleção inicial satisfazem uma condição $C$, da forma $C = \{y_i | y_i > c\}$. A variável $y_i$ pode ser qualquer variável relativa à unidade $i$, como por exemplo uma contagem dos elementos, densidade populacional, a área poluída, biomassa (quantidade total de matéria viva) entre outras variáveis relativas à unidade $i$. No exemplo da Figura 2.2 adotamos a condição $y_i > 0$ onde $y_i$ corresponde à contagem de elementos na unidade $i$;

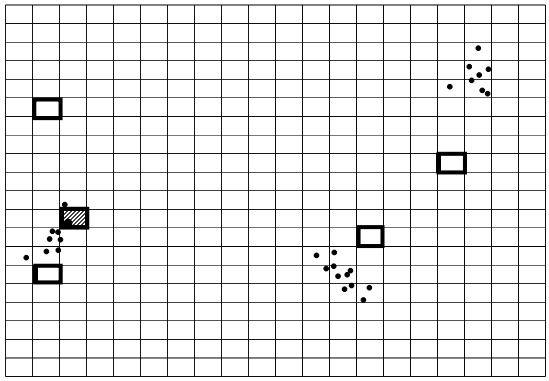

Figura 2.2: *Unidade que satisfaz $C$.*

3. Expansão nas ***unidades de vizinhança*** das unidades que satisfazem $C$.

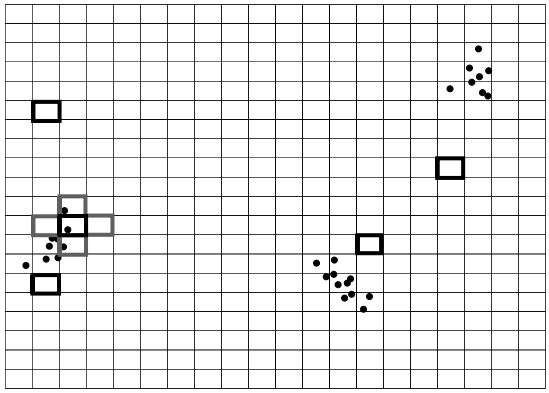

Figura 2.3: *Seleção da vizinhança em formato de cruz*

A vizinhança pode assumir várias formas e é um critério previamente estabelecido. Segundo Thompson & Seber (1996) a vizinhança não tem que ser contínua mas por definição ela deve ser simétrica, isto é, se a unidade $i$ está na vizinhança da unidade $j$, então $j$ deve estar na vizinhança de $i$.

No exemplo da Figura 2.1, consideramos uma amostra aleatória simples das unidades ($n_1 = 5$), ou seja, nessa seleção inicial, qualquer unidade (quadrado do *grid*) tem probabilidade de seleção igual. Na Figura 2.3, a vizinhança foi definida em formato de cruz, ou seja, para cada unidade satisfazendo $C$, são selecionadas as unidades imediatamente acima, abaixo, à esquerda e à direita.

4. O processo continua até que não haja mais nenhuma unidade satisfazendo a condição $C$ na vizinhança das unidades selecionadas.

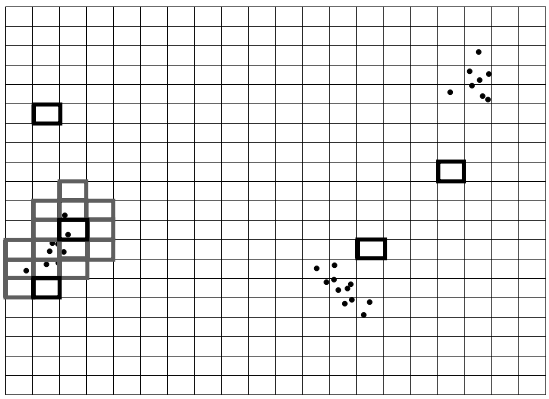

Figura 2.4: *Conglomerados formados após o processo de seleção*

Na amostragem por conglomerados adaptativos, conglomerado é um conjunto de unidades composto pela unidade da seleção inicial mais as unidades adicionadas adaptativamente. Note que quando uma unidade que não satisfaz $C$ é selecionada inicialmente, ela sozinha é um conglomerado.

No exemplo da Figura 2.4, temos $n_1 = 5$ conglomerados, entretanto existe uma sobreposição de uma unidade no maior conglomerado formado, na parte inferior, que além de ter sido adicionada pelo procedimento também é uma unidade da seleção inicial. Essa sobreposição fará com que consideremos uma mesma unidade duas vezes na amostra. Caso as duas unidades de seleção inicial satisfizessem a condição $C$, todo o conglomerado seria considerado duas vezes.

## 2.2 Estimadores

Antes de abordarmos os estimadores utilizados, necessitamos compreender algumas nomenclaturas utilizadas na concepção de amostragem por conglomerados adaptativos:

- ***Conglomerado:*** O conglomerado é o agrupamento formado pelas unidades durante o processo de seleção;

- ***Unidades de beirada:*** São unidades que não satisfazem a condição $C$, adicionadas pelo processo de seleção. Não estão incluídas as unidades que foram selecionadas inicialmente mas não satisfazem a condição $C$. Na Figura 2.5, as *unidades de beirada* correspondem à parte em cinza;

- ***Rede***: $A_i$ é o conjunto de unidades de um conglomerado formado pela unidade $i$ removendo-se as unidades de beirada. A seleção de qualquer unidade da rede $A_i$ leva à seleção da unidade $i$. Caso $i$ não satisfaça a condição $C$ ou não possua elementos em sua vizinhança que a satisfaçam, é formada uma rede de tamanho 1. As unidades de beirada são também redes de tamanho 1. Dessa forma os conglomerados podem ser decompostos em redes de tamanho 1 e as demais, o que é conveniente para o cálculo dos estimadores. Na Figura 2.5, a *rede* de tamanho maior que 1 corresponde à parte

hachurada. Na parte em cinza temos 10 unidades de beirada que equivalem a 10 redes de tamanho 1.

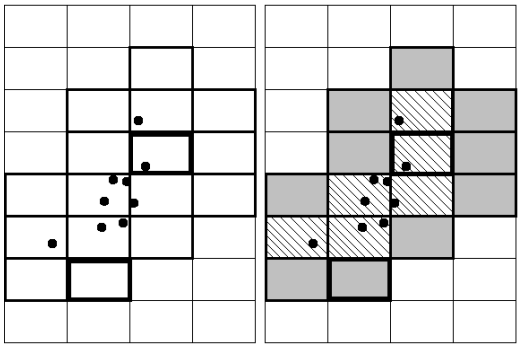

Figura 2.5: *Ilustração dos conceitos: o conglomerado e a sua divisão em unidades de beirada e rede*

### 2.2.1 Estimador do Tipo Horvitz-Thompson

Thompson (1990) apresentou um estimador para a média populacional que corresponde a uma modificação do estimador de Horvitz-Thompson. Para definirmos um estimador desse tipo, necessitamos primeiramente definir as probabilidades de seleção para cada unidade.

Nesse caso a probabilidade de seleção da unidade $i$ pode ser escrita com base na probabilidade da união dos dois eventos abaixo:

$E_1(i) = \{$*uma unidade da rede da qual i faz parte* ($A_i$)*, é selecionada na amostra inicial*$\}$

$E_2(i) = \{i$ *é uma unidade de beirada para algum conglomerado formado pelas unidades selecionadas*$\}$

Definimos:

$a_i$: número total de unidades em rede para os conglomerados em que $i$ é uma unidade de beirada

No exemplo da Figura 2.5, para as unidades em cinza $a_i = 6$, pois essas são unidades de beirada em relação a rede que está hachurada. Para as unidades hachuradas, que não são de

beirada, $a_i = 0$. O valor $a_i$ poderia ser maior que 6 caso houvesse outra rede em que uma unidade cinza fosse beirada.

$m_i$: número de unidades da rede de $i$ ($A_i$).

No exemplo da Figura 2.5, para as unidades em cinza $m_i = 1$ (assumindo que unidades de beirada são redes de tamanho 1) e para as unidades hachuradas $m_i = 6$. Note que ao final do processo de amostragem $m_i$ é uma quantidade conhecida, enquanto que $a_i$ pode ser maior do que o observado na amostra pois não temos controle se existe outra rede da qual $i$ seja unidade de beirada.

Para calcular a probabilidade de inclusão de uma unidade, notamos que o evento "seleção de uma unidade" é dado por $E_1(i) \bigcup E_2(i)$, cujo complementar é $E_1(i)^c \bigcap E_2(i)^c$.

Sendo assim, o número de possibilidades satisfazendo $E_1(i)^c \bigcap E_2(i)^c$ é $\binom{N-m_i-a_i}{n_1}$. Isso ocorre pois devemos ter combinações de tamanho $n_1$ que não contemplam nem as $a_i$ unidades que compõem a rede em que $i$ é unidade de beirada nem as $m_i$ unidades que fazem parte da rede de $i$. Assim sendo, a probabilidade de seleção de um elemento pode ser calculada pela fórmula

$$\pi_i = 1 - \left[ \frac{\binom{N-m_i-a_i}{n_1}}{\binom{N}{n_1}} \right]. \tag{2.1}$$

O estimador proposto por Horvitz & Thompson (1952) para a média populacional por unidade da variável de interesse $y$ é dado por

$$\hat{\mu}_{HT_0} = \frac{1}{N} \sum_{i=1}^{N} \frac{y_i I_i}{\pi_i}. \tag{2.2}$$

Na expressão (2.2), a função $I_i$ é uma função indicadora do evento $E_1(i) \bigcup E_2(i)$.

Nosso problema de encontrar um estimador não se encerra por aqui, uma vez que os valores de $a_i$ na maioria das vezes não são conhecidos. Isso se deve ao fato de que não possuímos todas as informações sobre quais redes possuem $i$ como unidade de beirada nem a priori e tampouco a posteriori. Os valores de $m_i$ são conhecidos pois ao final do processo de amostragem conheceremos todas as unidades que compõem a rede da qual $i$ faz parte. Isso inclui o caso em que $i$ é uma rede de tamanho 1.

Uma vez que não conhecemos $a_i$, podemos ***desconsiderá-los*** do processo de estimação. Isso significa que estaremos utilizando uma amostra de $n_1$ redes ao invés de uma amostra de

$n_1$ conglomerados. Lembre-se que não necessariamente teremos $n_1$ redes distintas. Assim, refazendo nossa abordagem chegaremos às probabilidades de seleção da rede da qual $i$ faz parte, dada por

$$\pi'_i = 1 - \left[ \frac{\binom{N-m_i}{n_1}}{\binom{N}{n_1}} \right]. \tag{2.3}$$

Sendo assim o estimador correspondente é

$$\hat{\mu}_{HT} = \frac{1}{N} \sum_{i=1}^{N} \frac{y_i I'_i}{\pi'_i}. \tag{2.4}$$

Aqui, a função indicadora $I'_i$ indica o evento $E_1(i)$.

**Propriedades do estimador**

Aqui mostramos as propriedades do estimador do tipo Horvitz-Thompson. Segundo Thompson & Seber (1996), para demonstrar as propriedades deste estimador é conveniente reescrevê-lo de forma que ao invés de $N$ unidades tenhamos $K$ redes distintas. Seja $x_k$ o número de unidades na rede $k$. Então podemos escrever

$$\hat{\mu}_{HT} = \frac{1}{N} \sum_{i=1}^{N} \frac{y_i I'_i}{\pi'_i} = \frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{x_k} \frac{y_i I'_i}{\pi'_i}. \tag{2.5}$$

Note que a probabilidade de seleção de uma unidade $i$ é igual na mesma rede $k$; isso quer dizer que dentro da mesma rede todas as unidades possuem uma probabilidade $\alpha_k$. Além disso, dentro da rede $k$, $I'_i$ é a mesma para todas as unidades. Podemos chamar a função indicadora da rede de $J_k$.

Dessa forma podemos escrever

$$\hat{\mu}_{HT} = \frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{x_k} \frac{y_i I'_i}{\pi'_i} = \frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{x_k} \frac{y_i I'_i}{\alpha_k} = \frac{1}{N} \sum_{k=1}^{K} \frac{y_k^* J_k}{\alpha_k} = \frac{1}{N} \sum_{k=1}^{\kappa} \frac{y_k^*}{\alpha_k}, \tag{2.6}$$

onde $y_k^* = \sum_{i=1}^{x_k} y_i$ e $\kappa$ é o número de redes selecionadas.

Podemos também calcular as probabilidades $\alpha_k$ de maneira análoga a (2.1), ou seja,

$$\alpha_k = 1 - \left[ \frac{\binom{N-x_k}{n_1}}{\binom{N}{n_1}} \right]. \tag{2.7}$$

**Vício**

Aqui calculamos a esperança do estimador proposto para verificar que ele é não-viesado.

$$\mathrm{E}\left(\hat{\mu}_{HT}\right) = \mathrm{E}\left(\frac{1}{N}\sum_{k=1}^{K}\frac{y_k^* J_k}{\alpha_k}\right) = \frac{1}{N}\sum_{k=1}^{K}\frac{y_k^*}{\alpha_k}\mathrm{E}\left(J_k\right) = \frac{1}{N}\sum_{k=1}^{K} y_k^* = \mu, \tag{2.8}$$

em que $\mu = \frac{1}{N}\sum_{i=1}^{N} y_i$.

**Variância**

Para podermos calcular a variância do estimador necessitamos calcular a probabilidade de se selecionar duas redes simultaneamente: $j$ e $k$. A probabilidade de que não selecionemos nem $k$ nem $j$ é dada por

$$\left[\frac{\binom{N-x_k-x_j}{n_1}}{\binom{N}{n_1}}\right]. \tag{2.9}$$

Definimos os eventos:

$E_k = \{a\ rede\ k\ é\ selecionada\}$ e $E_j = \{a\ rede\ j\ é\ selecionada\}$.

$$P(E_k^c \bigcap E_j^c) = \left[\frac{\binom{N-x_k-x_j}{n_1}}{\binom{N}{n_1}}\right]. \tag{2.10}$$

Sendo assim podemos aplicar um pouco de teoria de conjuntos para calcular a probabilidade das duas redes serem selecionadas, dada por

$$\begin{aligned}
\alpha_{jk} &= P(E_k \bigcap E_j) \\
&= -1 + P(E_k^c \bigcap E_j^c) + P(E_k) + P(E_j) \\
&= -1 + \left[\frac{\binom{N-x_k-x_j}{n_1}}{\binom{N}{n_1}}\right] + \left\{1 - \left[\frac{\binom{N-x_j}{n_1}}{\binom{N}{n_1}}\right]\right\} + \left\{1 - \left[\frac{\binom{N-x_k}{n_1}}{\binom{N}{n_1}}\right]\right\}.
\end{aligned} \tag{2.11}$$

Além de $\alpha_{jk}$ é necessário também calcular $\mathrm{Cov}\left(\frac{y_j^* J_j}{\alpha_j}, \frac{y_k^* J_k}{\alpha_k}\right)$. Nesse caso utilizamos a abordagem de Horvitz & Thompson (1952) em que as variáveis aleatórias são as funções indicadoras $J_j$ e $J_k$ que indicam os eventos $\mathrm{E}_j$ e $\mathrm{E}_k$.

$$\mathrm{Cov}\left(\frac{y_j^* J_j}{\alpha_j}, \frac{y_k^* J_k}{\alpha_k}\right) = \frac{y_j^* y_k^*}{\alpha_j \alpha_k}\left[\mathrm{E}(J_j J_k) - \mathrm{E}(J_j)\mathrm{E}(J_k)\right] = \frac{y_j^* y_k^*}{\alpha_j \alpha_k}\left(\alpha_{jk} - \alpha_j \alpha_k\right). \tag{2.12}$$

Sendo assim, basta aplicarmos (2.12) para obtermos a variância do estimador, dada por

$$\mathrm{Var}\left(\hat{\mu}_{HT}\right) = \mathrm{Var}\left(\frac{1}{N}\sum_{k=1}^{K}\frac{y_k^* J_k}{\alpha_k}\right) = \frac{1}{N^2}\sum_{k=1}^{K}\sum_{j=1}^{K}\frac{y_j^* y_k^*}{\alpha_j \alpha_k}\left(\alpha_{jk} - \alpha_j \alpha_k\right). \tag{2.13}$$

Segundo Horvitz & Thompson (1952), um estimador não viciado para (2.13) é dado por

$$\hat{\mathrm{Var}}\left(\hat{\mu}_{HT}\right) = \frac{1}{N^2}\sum_{k=1}^{K}\sum_{j=1}^{K}\frac{y_j^* y_k^*}{\alpha_j\alpha_k}\left(\alpha_{jk}-\alpha_j\alpha_k\right)J_j J_k = \frac{1}{N^2}\sum_{k=1}^{\kappa}\sum_{j=1}^{\kappa}\frac{y_j^* y_k^*}{\alpha_j\alpha_k}\left(\alpha_{jk}-\alpha_j\alpha_k\right). \quad (2.14)$$

Lembre-se que $\kappa$ é o número de redes selecionadas.

### 2.2.2 Estimador do Tipo Hansen-Hurwitz

Em uma amostragem com reposição de tamanho $n$ de uma população com $N$ unidades no universo e uma variável de interesse $y_i$ associada à $i$-ésima unidade, o estimador de Hansen & Hurwitz (1943) para $\mu$ é definido como

$$\hat{\mu}_{HH_0} = \frac{1}{nN}\sum_{j=1}^{n}\frac{y_i}{p_i}, \quad (2.15)$$

onde $p_i$ é a probabilidade de que a unidade $i$ seja selecionada quando apenas uma unidade é sorteada. O estimador em questão pode ser reescrito na forma

$$\hat{\mu}_{HH_0} = \frac{1}{nN}\sum_{i=1}^{N}\frac{y_i}{p_i}n_i, \quad (2.16)$$

onde $n_i$ é o número de vezes que a unidade $i$ é selecionada em $n$ retiradas.

A quantidade $n_i \sim Bin(n, p_i)$ pois em cada uma das $n$ retiradas $i$ pode ser selecionada com uma probabilidade $p_i$. Sendo assim, temos que $\mathrm{E}(n_i) = np_i$. Por isso podemos escrever

$$\hat{\mu}_{HH_0} = \frac{1}{nN}\sum_{i=1}^{N}\frac{y_i}{p_i}n_i = \frac{1}{N}\sum_{i=1}^{N}\frac{y_i}{\mathrm{E}(n_i)}n_i. \quad (2.17)$$

Modificando esse estimador para o contexto da amostragem por conglomerados adaptativos temos

$$\hat{\mu}_{HH} = \frac{1}{N}\sum_{i=1}^{N}\frac{y_i}{\mathrm{E}(f_i)}f_i, \quad (2.18)$$

em que $f_i$ é, analogamente, o número de vezes que a unidade $i$ é incorporada ao estimador, de acordo com o número de unidades da amostra inicial que intercepta a rede de $i$.

#### Propriedades do estimador

Nesta seção, estudamos o caso em que a amostragem inicial é sem reposição e portanto na expressão (2.18), $f_i$ tem uma distribuição hipergeométrica, pois de um total de $N$ unidades,

selecionamos $n_1$, em que $m_i$ são pertencentes a $A_i$ e $N-m_i$ não. Aplicando que $\mathrm{E}(f_i) = \frac{n_1 m_i}{N}$, temos

$$\hat{\mu}_{HH} = \frac{1}{n_1} \sum_{i=1}^{N} \frac{y_i f_i}{m_i}. \tag{2.19}$$

Note que dentro dessa rede existe um conjunto de unidades $u_1, \ldots, u_{m_i}$, para as quais existe um respectivo $y_{j(i)}$, $j = 1, 2, \ldots, m_i$.

Como as $n_1$ redes selecionadas não são necessariamente distintas, quando há a duplicação, algum $f_i > 1$. Quando uma unidade $i$ não foi selecionada, necessariamente $f_i = 0$ e portanto $\frac{y_i f_i}{m_i}$ será 0. Podemos então reconstruir o estimador (2.19) como uma soma de $n_1$ redes não necessariamente distintas. Sendo assim temos

$$\hat{\mu}_{HH} = \frac{1}{n_1} \sum_{i=1}^{n_1} \frac{1}{m_i} \sum_{j=1}^{m_i} y_{j(i)}. \tag{2.20}$$

Podemos ainda denominar $\frac{1}{m_i} \sum_{j=1}^{m_i} y_{j(i)}$ como $w_i$: a média de $y$ por unidade na rede formada também pela unidade $i$. E ainda podemos escrever a expressão (2.20) como $\bar{w}$, que indica uma média das médias de $y$ por unidade.

### Vício

Para mostrarmos que o estimador é não viciado, precisamos notar que a única variável aleatória que compõe o estimador em questão é $f_i$. A partir de então,

$$\mathrm{E}\left(\hat{\mu}_{HH}\right) = \mathrm{E}\left(\frac{1}{n_1} \sum_{i=1}^{N} \frac{y_i f_i}{m_i}\right) = \frac{1}{n_1} \sum_{i=1}^{N} \frac{y_i \mathrm{E}\left(f_i\right)}{m_i} = \frac{1}{n_1} \sum_{i=1}^{N} \frac{y_i}{m_i} \frac{n_1 m_i}{N} = \mu. \tag{2.21}$$

### Variância

Para efetuar o cálculo da variância é conveniente reescrever a expressão do estimador (2.20),

$$\hat{\mu}_{HH} = \frac{1}{n_1} \sum_{i=1}^{n_1} \frac{1}{m_i} \sum_{j=1}^{m_i} y_{j(i)} = \frac{1}{n_1} \sum_{i=1}^{n_1} w_i. \tag{2.22}$$

Segundo Thompson & Seber (1996), a expressão acima pode ser reconhecida como a média amostral da variável $w_i$ e portanto para obter a variância desse estimador basta usar a variância do estimador de amostra aleatória simples. Sendo assim temos

$$\mathrm{Var}\left(\hat{\mu}_{HH}\right) = \frac{N - n_1}{N n_1 \left(N - 1\right)} \sum_{i=1}^{N} \left(w_i - \mu\right)^2, \tag{2.23}$$

cujo estimador não viesado é dado por

$$\hat{\text{Var}}\left(\hat{\mu}_{HH}\right) = \frac{N - n_1}{N n_1 \left(N - 1\right)} \sum_{i=1}^{n_1} \left(w_i - \hat{\mu}_{HH}\right)^2 . \tag{2.24}$$

### 2.2.3 Amostragem Inicial com Reposição

Quando fazemos a amostragem inicial com reposição, existem algumas modificações nos estimadores que devemos fazer. Segundo Thompson & Seber (1996), a expressão de $\hat{\mu}_{HT}$ (2.6) assim como as expressões da esperança e variância também podem ser utilizadas. Entretanto, nesse caso as probabilidades $\alpha_k$(2.7) e $\alpha_{jk}$(2.14) devem ser recalculadas de maneira análoga à da Seção 2.2.1, ou seja,

$$\alpha_k = 1 - \left(1 - \frac{x_k}{N}\right)^{n_1} \tag{2.25}$$

$$\text{e } \alpha_{jk} = 1 - \left\{ \left(1 - \frac{x_j}{N}\right)^{n_1} + \left(1 - \frac{x_k}{N}\right)^{n_1} - \left(1 - \frac{x_j + x_k}{N}\right)^{n_1} \right\}. \tag{2.26}$$

Os autores ainda afirmam que o estimador Hansen-Hurwitz também pode ser utilizado, com expressões análogas às expressões da Seção 2.2.2, sendo que $f_i$ tem agora distribuição Binomial com número de ensaios igual a $n_1$ e probabilidade de sucesso $m_i/N$.

É importante destacar que aqui a probabilidade de sucesso é a probabilidade de encontrar uma unidade pertencente a uma rede que possui $m_i$ unidades em um total de $N$ unidades. Sendo assim, a esperança de $f_i$ também é igual a $\frac{n_1 m_i}{N}$, e assim não há alteração na esperança do estimador proposto inicialmente. Para calcularmos a variância, usamos o mesmo artifício apresentado anteriormente, escrevendo a expressão com base na variável $w_i$ (a média de $y$ por unidade na rede formada também pela unidade $i$). Tomamos então que

$$\text{Var}(\hat{\mu}_{HH}) = \text{Var}(\bar{w}) = \frac{\text{Var}(w)}{n_1} = \frac{1}{N n_1} \sum_{i=1}^{N} \left(w_i - \mu\right)^2 \tag{2.27}$$

e conseqüentemente seu estimador não viesado é

$$\hat{\text{Var}}\left(\hat{\mu}_{HH}\right) = \frac{1}{n_1(n_1 - 1)} \sum_{i=1}^{n_1} \left(w_i - \hat{\mu}_{HH}\right)^2 . \tag{2.28}$$

### 2.2.4 Amostra Inicial com Probabilidades Desiguais

Nas seções anteriores estudamos casos em que a seleção inicial é equiprovável, ou seja, as probabilidades de inclusão na amostra inicial são iguais. Note que o processo de amostragem

por conglomerados adaptativos como um todo não é equiprovável (salvo algum caso especial).

Queremos estudar o que ocorre quando temos uma seleção inicial das unidades com probabilidades desiguais, como por exemplo a amostragem proporcional ao tamanho ou outros procedimentos.

Thompson & Seber (1996) citam o exemplo abordado por Roesch Jr. (1993), sobre poluição de árvores. Nesse exemplo, cada árvore possui um tamanho, variável que é mensurada através do raio do tronco em uma altura basal.

Imagine que cada árvore seja rodeada por um círculo de raio proporcional ao tamanho da árvore. Seja $t_i$ a área do círculo associado à $i$-ésima árvore ($i = 1, 2, \ldots, N$).

Para realizarmos a seleção inicial, escolhemos pontos ao acaso no plano, com reposição. Note que se fizermos dessa maneira, as probabilidades de seleção das árvores serão distintas (se as árvores forem de diferentes tamanhos) calculadas por $t_i/A$, onde $A$ é a área da região de estudo.

Note que $t_i$ não poderá ser muito pequeno pois dessa forma a probabilidade de selecionar algum círculo seria pequena. Sendo assim, se $t_i$ é suficientemente grande, é natural imaginarmos que exista sobreposição entre os círculos de diferentes árvores e portanto dos $n_0$ pontos que selecionamos, são interceptados $n_1 \geq n_0$ círculos de árvores e conseqüentemente são selecionadas $n_1$ árvores. Interessante notar que nessa etapa, estamos utilizando as árvores como unidades e elementos simultaneamente.

Para aplicarmos o procedimento de amostragem por conglomerados adaptativos definimos a vizinhança de uma árvore. Para definir a vizinhança, tomamos um raio fixo $R$ e definimos um segundo círculo em torno da árvore. As árvores contidas nesse segundo círculo são verificadas e se o critério $C$ é satisfeito (no exemplo de Roesch se a árvore apresenta sinais relativos à poluição), para cada árvore em que o critério é satisfeito é definida uma vizinhança e assim por diante. Nesse contexto podemos também utilizar os estimadores apresentados anteriormente com algumas modificações conforme veremos a seguir.

### Modificações no estimador tipo Horvitz-Thompson

Temos $K$ redes de árvores sendo que $\kappa$ na amostra. As árvores da rede $k$ são incluídas na amostra final somente se a rede correspondente for interceptada na amostra inicial. Sendo

assim, podemos usar a expressão (2.6),

$$\hat{\mu}_{HT} = \frac{1}{N}\sum_{k=1}^{\kappa}\frac{y_k^*}{\alpha_k}. \tag{2.29}$$

No entanto as probabilidades $\alpha_k$ mudam. Nas expressões (2.25) e (2.26), as probabilidades se tornam a relação entre o tamanho da área de seleção das árvores na rede ($U_k^*$) e a área total. Sendo assim, as fórmulas para as probabilidades $\alpha$ se tornam

$$\alpha_k = 1 - \left(1 - \frac{U_k^*}{A}\right)^{n_0}. \tag{2.30}$$

Lembramos que $n_0$ é o número de pontos selecionados na amostragem inicial. Se definimos $U_{jk}^*$ como sendo a união das áreas $U_j^*$ e $U_k^*$ temos

$$\alpha_{jk} = 1 - \left\{\left(1 - \frac{U_j^*}{A}\right)^{n_0} + \left(1 - \frac{U_k^*}{A}\right)^{n_0} - \left(1 - \frac{U_{jk}^*}{A}\right)^{n_0}\right\}. \tag{2.31}$$

**Modificações no estimador tipo Hansen-Hurwitz**

O estimador do tipo Hansen-Hurwitz pode ser utilizado,

$$\hat{\mu}_{HH} = \frac{1}{N}\sum_{i=1}^{N}\frac{y_i}{\mathrm{E}(f_i)}f_i. \tag{2.32}$$

Aqui $f_i$ é o número de pontos aleatórios que interceptam a rede $A_i$ (rede definida pela árvore $i$). Seja $U_i$ a união das áreas de seleção dessa rede. Já que temos $n_0$ pontos independentes, a probabilidade de seleção de um elemento da rede $A_i$ é $U_i/A$ e portanto $f_i$ tem distribuição Binomial ($n_0$, $U_i/A$). Assim sendo, a esperança de $f_i$ é dada por $\frac{n_0 U_i}{A}$ e podemos desenvolver $\hat{\mu}_{HH}$ como

$$\hat{\mu}_{HH} = \frac{1}{N}\sum_{i=1}^{N}\frac{y_i}{\mathrm{E}(f_i)}f_i = \frac{A}{Nn_0}\sum_{i=1}^{N}\frac{y_i f_i}{U_i}. \tag{2.33}$$

Passando pela mesma abordagem que apresentamos na expressão (2.20), ou seja, escrever a expressão particionando em função das $n_1$ unidades iniciais, vemos que não é possível fazer da mesma forma. Isso ocorre pois os $n_0$ pontos aleatórios podem trazer mais de 1 elemento, podendo estes serem de redes distintas. Vamos ter que separar a soma para cada ponto inicial e somar na mesma rede $k$ tanto $U_k^*$ como os valores de $y_k^*$. Seja $K_h$ o número de redes selecionadas pelo ponto aleatório $h = (1, \ldots, n_0)$. Então

$$\hat{\mu}_{HH} = \frac{A}{Nn_0}\sum_{h=1}^{n_0}\sum_{k=1}^{K_h}\frac{y_k^*}{U_k^*}. \tag{2.34}$$

Para encontrar a variância desse estimador, podemos utilizar a mesma idéia da Seção 2.2.2 de reescrever o estimador como uma média amostral. Se chamarmos $w_h = \sum_{k=1}^{K_h} \frac{y_k^*}{U_k^*}$ então a expressão de $\hat{\mu}_{HH}$ fica: $\frac{A}{Nn_0} \sum_{h=1}^{n_0} w_h$. Aqui estamos tomando uma amostra aleatória simples com reposição dos valores $w_h$.

Podemos ainda reexpressar $w_h$ como sendo $w_h = \sum_{k=1}^{K} \frac{y_k^*}{U_k^*} J_k$ onde $J_k$ é a função indicadora do evento que a rede $k$ é selecionada pelo ponto $h$ (que tem probabilidade $U_k^*/A$). Podemos calcular a esperança e a variância de $w_h$ para obter posteriormente a esperança e a variância de $\hat{\mu}_{HH}$, ou seja,

$$\begin{aligned} \mathrm{E}(w_h) &= \sum_{k=1}^{K} \frac{y_k^*}{U_k^*} \mathrm{E}(J_k) = \sum_{k=1}^{K} \frac{y_k^*}{U_k^*} P(J_k = 1) \\ &= \sum_{k=1}^{K} \frac{y_k^*}{U_k^*} \frac{U_k^*}{A} = \frac{1}{A} \sum_{k=1}^{K} y_k^* = \frac{1}{A} \sum_{i=1}^{N} y_k^* = \frac{\tau}{A} \end{aligned} \tag{2.35}$$

e

$$\mathrm{Var}(w_h) = \sum_{j=1}^{K} \sum_{k=1}^{K} \frac{y_j^* y_k^*}{U_j^* U_k^*} \mathrm{Cov}[J_j, J_k]. \tag{2.36}$$

Sabendo que

$$\mathrm{E}[J_j J_k] = P([J_j = 1] \bigcap [J_k = 1]) = \frac{\Lambda_{jk}}{A}, \tag{2.37}$$

onde $\Lambda_{jk}$ é a área que corresponde à intersecção entre a $j$-ésima e a $k$-ésima redes e utilizando a definição de covariância pela combinação das esperanças, a variância de $w_h$ fica

$$\mathrm{Var}(w_h) = \sum_{j=1}^{K} \sum_{k=1}^{K} \frac{y_j^* y_k^*}{U_j^* U_k^*} \left( \frac{\Lambda_{jk}}{A} - \frac{U_j^* U_k^*}{A^2} \right). \tag{2.38}$$

Resgatando que $\hat{\mu}_{HH} = \frac{A}{Nn_0} \sum_{h=1}^{n_0} w_h = \frac{A}{N} \bar{w}$ podemos calcular a expressão para a esperança de $\hat{\mu}_{HH}$ como

$$\mathrm{E}(\hat{\mu}_{HH}) = \mathrm{E}\left( \frac{A}{Nn_0} \sum_{h=1}^{n_0} w_h \right) = \frac{A}{N} \mathrm{E}(\bar{w}) = \frac{A}{N} \frac{\tau}{A} = \frac{\tau}{N} = \mu. \tag{2.39}$$

Fazendo o mesmo para a variância, temos

$$\mathrm{Var}(\hat{\mu_{HH}}) = \mathrm{Var}\left( \frac{A}{N} \bar{w} \right) = \frac{A^2}{N^2} \mathrm{Var}(\bar{w}) = \frac{A^2}{N^2 n_0} \mathrm{Var}(w_h). \tag{2.40}$$

Segundo Thompson & Seber (1996), um estimador não viesado para essa variância é dado por

$$\hat{Var}(\hat{\mu_{HH}}) = \frac{A^2}{N^2} \hat{Var}(\bar{w}) = \frac{A^2}{N^2} \frac{{s_w}^2}{n_0} = \frac{A^2}{N^2 n_o (n_0 - 1)} \sum_{h=1}^{n_0} (w_h - \bar{w})^2. \tag{2.41}$$

### 2.2.5 Amostragem por Conglomerados Adaptativos em Duas Etapas

Segundo Salehi & Seber (1997), a amostragem por conglomerados adaptativos usual, dado o seu processo de amostragem, não possui uma teoria factível para a criação de pesquisas piloto para que seja possível desenhar um experimento dada uma eficiência ou custo e por isso propuseram a amostragem por conglomerados adaptativos em dois estágios. Aqui não iremos detalhar as metodologia do piloto e sim fornecer o subsídio teórico. A Figura 2.6, corresponde a uma área de estudo que foi subdividida em 8 unidades primárias (retângulos maiores) e 200 unidades secundárias (retângulos menores). Cada unidade secundária possui um número de indivíduos da população de marrecos da asa azul dado por Smith *et al.* (1995).

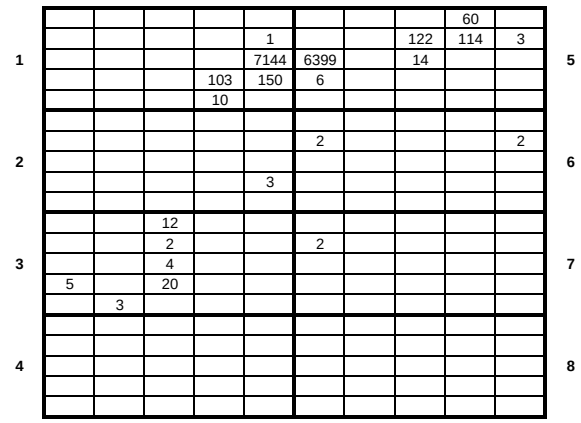


Figura 2.6: *Marrecos da asa azul*

A amostragem por conglomerados adaptativos em dois estágios foi desenvolvida em dois esquemas:

- **Esquema com sobreposição**

  O esquema com sobreposição constitui-se em, na primeira etapa, selecionar as unidades primárias por um sorteio aleatório e na segunda etapa aplicar a amostragem por conglomerados adaptativos, permitindo que o processo de amostragem adicione elementos de outras unidades primárias nos casos em que uma rede corta mais de uma unidade primária

- **Esquema sem sobreposição**

No esquema sem sobreposição, a adição das unidades secundárias, representadas pelos retângulos menores da Figura 2.6, é restrita à unidade primária, ou seja, mesmo que uma rede corte mais de uma unidade primária, na amostra só serão considerados os elementos da unidade primária selecionada.

Os autores Salehi & Seber (1997) revelam que não é claro qual dos esquemas citados é melhor pois enquanto o esquema com sobreposição seleciona toda a rede ele também gera uma amostra final esperada maior. Uma vez que é computacionalmente mais simples gerar o método sem sobreposição, iremos trabalhar com esse esquema em nossas simulações. Para esse esquema amostral, o estimador Horvitz-Thompson modificado é dado por

$$\hat{\mu}_{HT(2E)} = \frac{1}{N}\frac{M}{m}\sum_{i=1}^{m}\hat{\tau}_i, \tag{2.42}$$

onde $\hat{\tau}_i = N_i\hat{\mu}_{HT(i)}$, é o estimador Horvitz-Thompson para o total da unidade primária $i$ com $N_i$ o número de unidades secundárias na unidade primária $i$ e $\hat{\mu}_{HT(i)}$ o estimador da média de $y$ por unidade secundária na unidade primária $i$, cuja variância é dada por

$$\mathrm{Var}(\hat{\mu}_{HT(2E)}) = \frac{1}{N}M(M-m)\frac{\sigma_M^2}{m} + \frac{1}{N}\frac{M}{m}\sum_{i=1}^{m}\hat{V}_i, \tag{2.43}$$

com $\sigma_M^2 = \frac{\sum_{i=1}^{M}\left(\tau_i - \sum_{i=1}^{M}\frac{\tau_i}{M}\right)^2}{M-1}$. Segundo Salehi & Seber (1997), um estimador não viesado para (2.43) é dado por

$$\hat{\mathrm{V}}\mathrm{ar}(\hat{\mu}_{HT(2E)}) = \frac{1}{N}M(M-m)\frac{s_M^2}{m} + \frac{1}{N}\frac{M}{m}\sum_{i=1}^{m}\hat{V}_i, \tag{2.44}$$

com $s_M^2 = \frac{\sum_{i=1}^{m}\left(\hat{\tau}_i - \sum_{i=1}^{m}\frac{\hat{\tau}_i}{m}\right)^2}{m-1}$ e $\hat{V}_i$ a variância do estimador Horvitz-Thompson dentro da unidade primária $i$.

De maneira análoga, podemos obter as variâncias (2.43) e (2.44) do estimador de Hansen-Hurwitz, substituindo $\hat{\tau}_i$ na fórmula (2.42) por $\hat{\tau}_i = N_i\hat{\mu}_{HH(i)}$, sendo este o estimador Hansen-Hurwitz para o total da unidade primária $i$.

### 2.2.6 Comparação entre os Estimadores HT e HH

Os estimadores apresentados anteriormente são diferentes. O estimador de Hansen-Hurwitz é mais fácil de ser calculado. Entretanto, segundo Salehi (2003), a experiência de alguns pesquisadores indica que Horvitz-Thompson tem uma variância menor.
Se olhamos para os estimadores

$$\hat{\mu}_{HH} = \frac{1}{n_1}\sum_{i=1}^{n_1} w_i = \frac{1}{Nn_1}\sum_{i=1}^{n_1}\frac{{y_i}^*}{p_i}, \tag{2.45}$$

e

$$\hat{\mu}_{HT} = \frac{1}{N}\sum_{k=1}^{\kappa}\frac{y_k^*}{\alpha_k}, \tag{2.46}$$

em que $p_i = m_i/N$ e ${y_i}^* = \sum_{j=1}^{m_i} y_{j(i)}$, percebemos que se $p_i$ e $\alpha_k$ forem proporcionais a ${y_i}^*$ e ${y_k}^*$ respectivamente, suas estimativas serão constantes e portanto a variância do estimador $\hat{\mu}_{HH}$ será zero e a de $\hat{\mu}_{HT}$ será próxima de zero pois $\kappa$ é uma variável aleatória limitada a valores próximos a $n_1$ uma vez que não esperamos selecionar por muitas vezes as unidades da mesma rede na seleção inicial, isto é, selecionar uma mesma rede muitas vezes.

Segundo Sarndal *et al.* (1992), o estimador que segue mais fielmente a propriedade de proporcionalidade entre $y$ e a probabilidade de inclusão, no nosso caso, ${y_i}^* \propto p_i$ e ${y_k}^* \propto \alpha_k$, é mais eficiente.

Segundo Salehi (2003), nas redes criadas por condições do tipo $y_i > c$ , existe uma grande diferença entre os valores totais de $y$ nas redes com apenas uma unidade e as maiores. Isso se deve ao fato de que a maior parte das redes de tamanho 1 não possuem elementos suficientes para cumprir a condição $C$. Sendo assim, para satisfazer melhor a propriedade de propocionalidade o estimador deveria acompanhar esse salto entre as redes de uma unidade e as maiores.

Note que $p_i$ é uma função linear, que cresce conforme $m_i$. Observando $\alpha_k$ como função de $m_k$ notamos que $\alpha_k(m_k)$ é não-decrescente pois a diferença de primeira ordem é não negativa

como mostramos a seguir,

$$\begin{aligned}
\Delta\alpha_k(m_k) &= \alpha_k(m_k+1) - \alpha_k(m_k) \\
&= 1 - \left[\frac{\binom{N-m_k-1}{n_1}}{\binom{N}{n_1}}\right] - \left(1 - \left[\frac{\binom{N-m_k}{n_1}}{\binom{N}{n_1}}\right]\right) \\
&= \left[\frac{\binom{N-m_k}{n_1}}{\binom{N}{n_1}}\right] - \left[\frac{\binom{N-m_k-1}{n_1}}{\binom{N}{n_1}}\right] \\
&= \left[\frac{\frac{N-m_k}{N-m_k-n_1}\binom{N-m_k-1}{n_1} - \binom{N-m_k-1}{n_1}}{\binom{N}{n_1}}\right] \geq 0.
\end{aligned}$$

A taxa de crescimento vai diminuindo conforme aumenta $m_k$ pois a diferença de segunda ordem é não positiva como segue,

$$\begin{aligned}
\Delta^2\alpha_k(m_k) &= \Delta\alpha_k(m_k+1) - \Delta\alpha_k(m_k) \\
&= \left[\frac{\frac{N-m_k-1}{N-m_k-1-n_1}\binom{N-m_k-2}{n_1} - \binom{N-m_k-2}{n_1}}{\binom{N}{n_1}}\right] - \left[\frac{\frac{N-m_k}{N-m_k-n_1}\binom{N-m_k-1}{n_1} - \binom{N-m_k-1}{n_1}}{\binom{N}{n_1}}\right] \\
&= \left[\frac{A\binom{N-m_k-2}{n_1} - \binom{N-m_k-2}{n_1}}{\binom{N}{n_1}}\right] - \left[\frac{\frac{N-m_k}{N-m_k-n_1}A\binom{N-m_k-2}{n_1} - A\binom{N-m_k-2}{n_1}}{\binom{N}{n_1}}\right] \\
&= \frac{\left[A - 1 - A\frac{n_1}{N-m_k-n_1}\right]\binom{N-m_k-2}{n_1}}{\binom{N}{n_1}} \leq 0,
\end{aligned}$$

com $A = \frac{N-m_k-1}{N-m_k-1-n_1}$.

Isso demonstra que o maior salto em $\alpha_k(m_k)$ é de $m_k = 1$ para $m_k = 2$, o que nos leva a crer que o estimador $\hat{\mu}_{HT}$ respeita mais a proporcionalidade em relação a $y_k$ conforme comentamos anteriormente.

Pode-se tentar melhorar os estimadores empregando o teorema de Rao-Blackwell. Salehi (2003) mostra que não existe ganho para o estimador Horvitz-Thompson quando utilizamos esse método, entretanto para o estimador Hansen-Hurwitz existe uma melhoria das estimativas. A seguir apresentamos uma breve definição do teorema de Rao-Blackwell e a aplicação do mesmo neste contexto.

**Teorema de Rao-Blackwell**

Seja $T = T(D_0)$ um estimador de $\phi = \phi(\theta)$ e seja W suficiente para $\theta$. Então:

1. $T_W = \mathrm{E}[T|W] = \eta[W]$ é um estimador;

2. $\mathrm{E}[T_W] = \mathrm{E}[T]$;

3. $EQM[T_W] \leq EQM[T]$, com desigualdade estrita quando $P_\theta(T \neq T_W) > 0$.

Sendo assim, para melhorar um estimador, basta calcularmos a esperança deste condicionando-o a uma estatística suficiente.

**Usando Rao-Blackwell para melhorar $\hat{\mu}_{HT}$ e $\hat{\mu}_{HH}$**

Primeiramente temos de introduzir a notação utilizada. Retiramos uma amostra aleatória simples de $n_1$ unidades sem reposição. Em seguida, outras unidades são selecionadas conforme o procedimento de amostragem por conglomerados adaptativos.

Suponha que a amostra final ordenada seja $s_0 = \{i_1, i_2, \ldots, i_{n_1}\}$. Note que os rótulos podem se repetir à medida que existam unidades da seleção inicial que interceptam a mesma rede. Definimos também $d_0 = \{(i, y_i) : i \in s_0\}$ e $s_r = \{i_1, i_2, \ldots, i_\nu\}$ em que $\nu$ é o total de rótulos distintos em $s_0$ e ainda podemos definir $d_r = \{(i, y_i) : i \in s_r\}$.

Segundo Basu (1969), a estatística $d_r$ é suficiente e minimal para $\theta$, onde $\theta$ corresponde ao vetor dos valores $y$ para toda a população. Nesse caso, desejamos estimar uma função de $\theta$ que é dada por $\mu(\theta) = \mu$. E assim, para encontrarmos um estimador mais eficiente, basta aplicarmos o teorema de Rao-Blackwell.

Apresentamos a estatística $d_J$ que é composta por $d_r$ e por funções indicadoras $J_i$ que indicam se a unidade $i$ pertence à amostra, isto é, $d_J = \{(i, y_i, J_i) : i \in s_r\}$. A estatística $d_J$ pode ser escrita como uma função de $d_r$ e portanto é uma estatística suficiente para $\theta$ (Thompson & Seber (1996) - pg 38). Dessa forma, aplicamos o teorema de Rao-Blackwell acima com estatística $W = d_J$.

Podemos reescrever o estimador de Horvitz-Thompson como

$$\hat{\mu}_{HT} = \frac{1}{N} \sum_{k=1}^{\kappa} \frac{y_k^*}{\alpha_k} = \frac{1}{N} \sum_{i \in s_r} \frac{y_i}{\alpha_i'} J_i. \tag{2.47}$$

Isso porque as probabilidades $\alpha_i'$'s são iguais para uma mesma rede e sendo assim podemos

calcular $\mathrm{E}(\hat{\mu}_{HT} \mid d_J)$, ou seja,

$$\mathrm{E}(\hat{\mu}_{HT} \mid d_J) = \frac{1}{N} \sum_{i \in s_r} \frac{y_i}{\alpha_i'} \mathrm{E}(J_i \mid d_J). \tag{2.48}$$

Segundo Salehi (2003), podemos separar $s_r$ em um conjunto $s_a$ das unidades adicionadas adaptativamente e outro $s_1$ contendo as unidades restantes. Note que para o conjunto $s_a$, $\mathrm{E}(J_i \mid d_J) = 0$. Logo,

$$\mathrm{E}(\hat{\mu}_{HT} \mid d_J) = \frac{1}{N} \sum_{i \in s_r} \frac{y_i}{\alpha_i'} \mathrm{E}(J_i \mid d_J) = \frac{1}{N} \sum_{i \in s_1} \frac{y_i}{\alpha_i'} = \frac{1}{N} \sum_{k=1}^{\kappa} \frac{y_k^*}{\alpha_k} = \hat{\mu}_{HT}. \tag{2.49}$$

Dessa forma, vemos que não há ganho ao se aplicar Rao-Blackwell nesse caso.

Para o estimador Hansen-Hurwitz, usamos a forma da expressão (2.22) escrita de uma maneira um pouco diferente,

$$\hat{\mu}_{HH} = \frac{1}{n_1} \sum_{i=1}^{n_1} w_i = \frac{1}{n_1} \sum_{i=1}^{N} w_i I_i, \tag{2.50}$$

onde $I_i$ indica se $i$ faz parte da seleção inicial. Sendo assim, calculamos $\mathrm{E}(\hat{\mu}_{HH} \mid d_J)$:

$$\tilde{\mu}_{HH} = \mathrm{E}(\hat{\mu}_{HH} \mid d_J) = \frac{1}{n_1} \sum_{i=1}^{N} w_i \mathrm{E}(I_i \mid d_J) = \frac{1}{n_1} \sum_{i=1}^{N} w_i P(I_i = 1 \mid d_J). \tag{2.51}$$

A variável aleatória $I_i \mid d_J$ assume valor 1 somente se a unidade $i$ é selecionada inicialmente e faz parte do conjunto $d_j$. Vimos pela definição do Teorema de Rao-Blackwell que a desigualdade número 3 é estrita nos casos em que $P_\theta(T \neq T_W) > 0$, o que nos remete a $P_\theta(\hat{\mu}_{HH} \neq \tilde{\mu}_{HH}) > 0$. As variáveis aleatórias $I_i$ e $I_i \mid d_j$ são diferentes pois a probabilidade de que $i$ seja uma unidade inicial de seleção é maior quando se sabe que $i$ faz parte da amostra final.

Sendo assim, concluímos que a variância do estimador Hansen-Hurwitz melhorado por Rao-Blackwell é estritamente menor do que a do estimador original. Isso mostra que o estimador de Horvitz-Thompson carrega toda a informação da amostra, enquanto que o de Hansen-Hurvitz não.

## 2.3 Eficiência e Tamanho Amostral

### 2.3.1 Amostragem Adaptativa versus Amostragem Aleatória Simples

Para fazer uma comparação entre a amostragem adaptativa e a amostragem aleatória simples, utilizamos a relação entre as variâncias da amostragem por conglomerados adaptativos e da amostragem aleatória simples.

Para realizarmos essa comparação, primeiramente definimos uma notação a ser utilizada. Sejam:

- $K$: número de redes na população;
- $k(i)$: rótulo da rede que inclui a unidade $i$;
- $B_k(i)$: conjunto de unidades que fazem parte da rede $k(i)$;
- $m_{k(i)}$: número de unidades da rede $k(i)$ que contém a unidade $i$.

Definimos para cada $i$ a média $w_{k(i)}$ dos valores $y_i$ das unidades da rede que inclui $i$, isto é,

$$w_{k(i)} = \frac{1}{m_{k(i)}} \sum_{j \in B_k(i)} y_j. \tag{2.52}$$

Vimos anteriormente que o estimador Hansen-Hurwitz é dado por

$$\hat{\mu}_{HH} = \frac{1}{n_1} \sum_{i=1}^{n_1} w_{k(i)}, \tag{2.53}$$

cuja variância pode ser escrita da seguinte forma

$$\text{Var}\,(\hat{\mu}_{HH}) = \frac{N - n_1}{n_1 N(N-1)} \sum_{i=1}^{N} (w_{k(i)} - \mu)^2. \tag{2.54}$$

Segundo Thompson & Seber (1996), podemos construir uma nova forma para essa variância particionando a soma de quadrados total em soma de quadrados entre as redes e dentro das redes. Lembrando que $B_k(i)$ é o conjunto de unidades presentes na rede $k(i)$, temos que

$$\sum_{i=1}^{N} (y_i - \mu)^2 = \sum_{k=1}^{K} \sum_{i \in B_k(i)} (y_i - w_{k(i)})^2 + \sum_{k=1}^{K} \sum_{i \in B_k(i)} (w_{k(i)} - \mu)^2 \tag{2.55}$$

$$= \sum_{k=1}^{K} \sum_{i \in B_k(i)} (y_i - w_{k(i)})^2 + \sum_{i=1}^{N} (w_{k(i)} - \mu)^2. \tag{2.56}$$

Logo, substituindo na expressão (2.54) temos

$$\text{Var}\left(\hat{\mu}_{HH}\right) = \frac{N-n_1}{n_1 N(N-1)}\left[\sum_{i=1}^{N}(y_i-\mu)^2 - \sum_{k=1}^{K}\sum_{i\in B_k(i)}(y_i - w_{k(i)})^2\right]. \tag{2.57}$$

A variância para a amostragem aleatória simples sem reposição de tamanho $n$ é dada por:

$$\text{Var}\left(\bar{y}\right) = \frac{N-n}{nN(N-1)}\sum_{i=1}^{N}(y_i-\mu)^2. \tag{2.58}$$

Assim sendo, podemos calcular a eficiência relativa, dada por

$$\frac{\text{Var}\left(\hat{\mu}_{HH}; aca; n_1\right)}{\text{Var}\left(\bar{y}; aas; n\right)} = \frac{n}{n_1}\left(\frac{N-n_1}{N-n}\right)\left[1 - \frac{\sum_{k=1}^{K}\sum_{i\in B_k}(y_i - w_{k(i)})^2}{\sum_{i=1}^{N}(y_i-\mu)^2}\right]. \tag{2.59}$$

Ao observarmos a expressão acima, podemos perceber que um fator decisivo para uma maior eficiência relativa é a variabilidade dentro da rede.

### 2.3.2 Variabilidade Intra-rede como uma Vantagem em relação à AAS

Na expressão (2.59), observamos que o fator $1 - \frac{\sum_{k=1}^{K}\sum_{i\in B_k}(y_i - w_{k(i)})^2}{\sum_{i=1}^{N}(y_i-\mu)^2}$ será menor conforme seja maior a variabilidade intra-rede em relação à variabilidade total. É importante notar que esse efeito depende diretamente da condição $C$ e da definição da vizinhança. Segundo Salehi (2003), os estimadores da amostragem por conglomerados adaptativos não tomam em conta a variabilidade dentro das redes pois estes somam os valores de $y$ dentro de cada rede. O estimador da média por amostragem aleatória simples considera essa variabilidade. Dessa forma, quanto maior essa variabilidade, maior a vantagem em termos de eficiência relativa, ao usar amostragem por conglomerados adaptativos.

### 2.3.3 Tamanho Amostral e Custos

Segundo Thompson & Seber (1996), assim como a amostragem por conglomerados convencional, a amostragem por conglomerados adaptativos nos trás o benefício de selecionar as unidades dentro do conglomerado o que minimiza o tempo e os custos de deslocamento.

Observando-se a fórmula de eficiência (2.59), podemos notar o primeiro fator de eficiência relativa dado por

$$b(n_1, n, N) = \frac{n}{n_1}\frac{N - n_1}{N - n}. \tag{2.60}$$

Quando esse fator é pequeno, há uma vantagem da estratégia adaptativa e quando ele é grande, a amostragem convencional é mais eficiente. Analisando a função vemos que:

- Quando $n = n_1$ então $b = 1$;
- Se $n_1$ e $n$ crescem a uma razão constante, isto é, $n = a_1 k$ e $n_1 = a_2 k$ então $b$ é uma função crescente de $k$;
- Com $n$ e $n_1$ fixos, $b$ é uma função decrescente de $N$ se $n > n_1$ e crescente se $n < n_1$.

Em geral, queremos avaliar essa função quando $n \geq n_1$, o que resume as nossas conclusões no seguinte: a amostragem por conglomerados adaptativos será mais eficiente se $n$ não for muito maior do que $n_1$. Idealmente os valores de $n$ e $n_1$ para computar a função de eficiência relativa são baseados nas respectivas funções de custos tomando-se os maiores valores possíveis de serem atingidos dado um custo de tempo e dinheiro. Thompson & Seber (1996) propõe uma função linear de custo para a amostragem por conglomerados adaptativos que trataremos a seguir.

### Cálculo do custo e tamanho amostral esperado

Para a amostragem por conglomerados adaptativos, o tamanho amostral é um número aleatório, pois só paramos o procedimento quando o critério não for mais satisfeito. Sendo assim, há uma abordagem especial para esse tema. Seja $I_i$ a função indicadora de que a unidade foi selecionada. Então temos que

$$\nu = \sum_{i=1}^{N} I_i \tag{2.61}$$

é o número de unidades distintas selecionadas. Se denominamos $\pi_i$ a probabilidade de seleção da unidade $i$ então:

$$E[\nu] = \mathrm{E}\left(\sum_{i=1}^{N} I_i\right) = \sum_{i=1}^{N} E[I_i] = \sum_{i=1}^{N} \pi_i. \tag{2.62}$$

Note que para cada tipo de amostragem os valores de $\pi_i$ mudam. No caso da amostragem sem reposição utilizamos a fórmula $\pi_i = 1 - \left[\frac{\binom{N-m_i-a_i}{n_1}}{\binom{N}{n_1}}\right]$. Para a amostragem com reposição, $\pi_i = 1 - (1-p_i)^{n_1}$ onde $p_i = \frac{m_i+a_i}{N}$ .

Seja,

$$C_{T_a} = c_0 + c_1 n_1 + c_2(\nu - n_1), \tag{2.63}$$

onde $c_0$ é um custo fixo, $c_1$ é um custo por unidade selecionada aleatoriamente e $c_2$ é o custo por unidade selecionada adaptativamente. Sendo assim podemos calcular um custo total esperado dado por

$$E[C_{T_a}] = c_0 + c_1 n_1 + c_2(E[\nu] - n_1). \tag{2.64}$$

Analogamente podemos obter uma função de custo para a amostragem aleatória simples, dado por

$$E[C_{T_s}] = c_0 + c_1 m_c. \tag{2.65}$$

Aqui, $m_c$ corresponde ao tamanho amostral da amostragem aleatória simples. Igualando-se os custos esperados de (2.64) e (2.65) temos

$$m_c = \left(1 - \frac{c_2}{c_1}\right) n_1 + \frac{c_2}{c_1} E[\nu]. \tag{2.66}$$

Segundo Thompson & Seber (1996), em muitas pesquisas em populações naturais o custo por unidade adicionada adaptativamente ($c_2$) é menor que $c_1$ devido a questões de logística. Note que nesses casos $0 \leq \frac{c2}{c1} \leq 1$ e $m_c$ é uma média ponderada entre $n_1$ e $E[\nu]$ sendo que $E[\nu]$ satisfaz $n_1 \leq m_c \leq E[\nu]$. Uma vez que essa condição sobre $m_c$ é satisfeita e na amostragem aleatória simples a variância é uma função decrescente de $n$, então podemos estabelecer limites da eficiência relativa da amostragem por conglomerados adaptativos em relação à amostragem aleatória simples, dados por

$$\frac{\text{Var}\,(\hat{\mu}_{HH}; aca; n_1)}{\text{Var}\,(\bar{y}; aas; n_1)} \leq \frac{\text{Var}\,(\hat{\mu}_{HH}; aca; n_1)}{\text{Var}\,(\bar{y}; aas; m_c)} \leq \frac{\text{Var}\,(\hat{\mu}_{HH}; aca; n_1)}{\text{Var}\,(\bar{y}; aas; E[\nu])}. \tag{2.67}$$

Vimos que a variância do estimador de amostragem por conglomerados adaptativos é sempre menor que a variância do estimador por amostragem aleatória simples de tamanho $n_1$. Sendo assim, o lado esquerdo da expressão (2.67) é sempre menor ou igual a 1, favorecendo a amostragem por conglomerados adaptativos. Na ausência de informação de custos, o lado

direito da expressão (2.67) pode servir como uma avaliação conservadora dando um maior benefício para a amostragem aleatória simples. Essa avaliação pode ser tanto menor quanto maior que 1, dependendo da variância dentro das redes entre outros fatores a serem analisados. Deve-se notar que o tamanho final da amostra depende da amostra inicial. De maneira similiar existe a dependência de $n_1$ na quantidade

$$g(n_1) = \frac{E[\nu]}{n_1}\frac{N-n_1}{N-E[\nu]}, \tag{2.68}$$

que corresponde a $b(n_1, E[\nu])$ e compõe o limite superior da eficiência. Mesmo $E[\nu]$ sendo uma função crescente de $n_1$, a razão $\frac{E[\nu]}{n_1}$ é função decrescente de $n_1$. Mostramos esse fato a seguir. Conforme vimos anteriormente, para a amostragem adaptativa com seleção inicial por sorteio aleatório simples, a probabilidade de seleção para a inclusão do elemento $i$ na amostra é dada por $\pi_i = 1 - \left[\frac{\binom{N-m_i-a_i}{n_1}}{\binom{N}{n_1}}\right]$ onde $m_i$ é o número de unidades na rede que contêm a unidade $i$ e $a_i$ é o número de unidades em redes das quais $i$ é uma unidade de beirada. Substituindo $\pi_i$ na expressão (2.62) encontramos

$$\begin{aligned}
& E[\nu] = && (2.69)\\
&= \sum_{i=1}^{N} 1 - \left[\frac{\binom{N-m_i-a_i}{n_1}}{\binom{N}{n_1}}\right] && (2.70)\\
&= N - \sum_{i=1}^{N} \frac{(N-m_i-a_i)n_1!(N-n_1)}{n_1!(N-m_i-a_i-n_1)!N!} && (2.71)\\
&= N - \frac{1}{N!}\sum_{i=1}^{N}(N-m_i-a_i)!(N-n_1-m_i-a_i+1)\ldots(N-n_1). && (2.72)
\end{aligned}$$

Visualizando $E[\nu]$ como uma função de $n_1$, que aqui denominamos $f(n_1)$, temos que a derivada de $f(n_1)$ com relação a $n_1$ é dada por

$$\begin{aligned}
f'(n_1) &= -\frac{1}{N!}\sum_{i=1}^{N}(N-m_i-a_i)! \sum_{j=1}^{m_i+a_i}(-1)\prod_{k=1,k\neq j}^{m_i+a_i}(N-n_1+k) && (2.73)\\
&= \frac{1}{N!}\sum_{i=1}^{N}(N-m_i-a_i)! \sum_{j=1}^{m_i+a_i}\prod_{k=1,k\neq j}^{m_i+a_i}(N-n_1+k). && (2.74)
\end{aligned}$$

Dado que $n_1 < N-m_i-a_i$ todos os fatores são positivos e logo $f(n_1)$ é uma função crescente. A segunda derivada é dada por

$$f''(n_1) = \frac{1}{N!}\sum_{i=1}^{N}(N-m_i-a_i)! \sum_{j=1}^{m_i+a_i}\sum_{r=1,r\neq j}(-1)\prod_{k=1,k\neq j,k\neq r}^{m_i+a_i}(N-n_1+k). \tag{2.75}$$

Essa função é negativa e portanto a função $f(n_1)$ tem concavidade para baixo. Sendo assim, $f(n_1)/n_1$, o tamanho amostral esperado sobre o tamanho inicial, é uma função decrescente. Agora calculamos a derivada da função $g(n_1) = [f(n_1)/n_1]\{(N - n_1)/[N - f(n_1)]\}$ com relação a $n_1$:

$$\begin{aligned} g'(n_1) &= \frac{f'(n_1)(N-n_1)}{n_1[N-f(n_1)]} - \frac{Nf(n_1)}{n_1^2[N-f(n_1)]} - \frac{f(n_1)(N-n_1)[-f'(n_1)]}{n_1[N-f(n_1)]^2} \\ &= \{n_1 f'(n_1)(N-n_1) - f(n_1)[N-f(n_1)]\}\left\{\frac{N}{n_1^2[N-f(n_1)]^2}\right\}. \end{aligned}$$

Sendo assim, $g(n_1)$ decresce em $n_1$ somente se $n_1 f'(n_1)(N - n_1) - f(n_1)[N - f(n_1)] < 0$, ou seja,

$$f'(n_1) < \frac{f(n_1)[N-f(n_1)]}{n_1(N-n_1)}. \tag{2.76}$$

Então $g$ pode ser crescente ou decrescente dependendo da condição $C$, da população e dos tamanhos amostrais.

Segundo Thompson & Seber (1996), na amostragem por conglomerados adaptativos utilizando os estimadores apresentados, não há ganho de eficiência utilizando-se as unidades de beirada. Uma população muito espalhada terá muitas redes de tamanho 1 com nenhuma variância e portanto a seleção por conglomerados adaptativos implicaria em muitas unidades de beirada o que deixa uma vantagem nesse contexto para a amostragem convencional. Outro fato é que em alguns casos o custo de amostragem das unidades satisfazendo a condição $C$ poderá ser maior do que o custo de amostragem das unidades que não satisfazem $C$. Por exemplo em levantamentos de biomassa, a biomassa nula pode ser levantada rapidamente enquanto que a maior que zero toma mais tempo e tem maior custo por envolver uma cuidadosa coleta e análise laboratorial.

### Raridade da população

A raridade geográfica ou densidade por unidade de área é um outro fator que influencia na eficiência relativa da amostragem por conglomerados adaptativos em relação à amostragem aleatória simples. Isso quer dizer que uma população com comportamento de formação de grupos espalhados em um espaço geográfico grande tem uma raridade geográfica maior do que uma população de mesmo tamanho espremida em um espaço geográfico menor.

Assumindo por simplicidade a condição $C = \{y_i | y_i > 0\}$, isso quer dizer que para aumentar a raridade geográfica de uma população com características de agrupamento e tamanho fixas devemos aumentar $N$, o que decorre do imediato aumento de unidades sem nenhum elemento.

Sejam:

- $\tau = \sum_{n=1}^{N} y_i$ o total populacional;
- $A = \sum_{k=1}^{K} \sum_{i \in B_k} (y_i - w_{k(i)})^2$, a variância intra-rede;
- $B = \sum_{i=1}^{N} y_i^2$, soma de quadrados totais não corrigida;
- $T = \tau^2$;
- $\eta$, o número de unidades que satisfazem a condição mais as unidades de beirada na população.

Assumimos os fatores acima fixos pois estamos fixando o padrão de agregação da população. Essa idéia está ilustrada pela Figura 2.7.

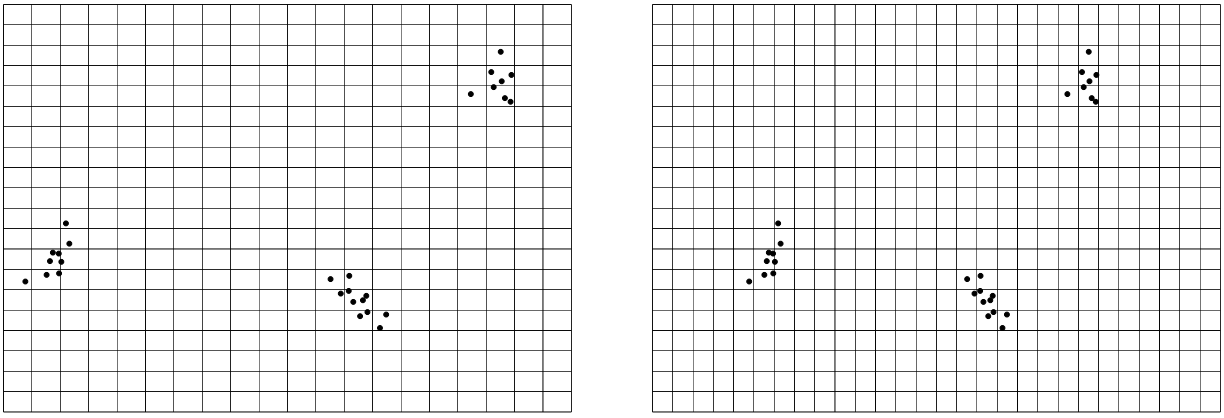

Figura 2.7: *População distribuída em* 400 *e* 560 *unidades respectivamente.*

Como o tamanho amostral inicial também é fixado, o tamanho amostral final esperado $E[\nu]$ é uma função de N, $f(N)$. Definimos $g(N) = [f(N)/n_1]\{(N - n_1)/[N - f(N)]\}$ e $k(N) = 1 - \{A/[B - (T/N)]\}$. Dessa forma temos que a eficiência relativa da amostragem por conglomerados adaptativos com tamanho final $E(\nu)$ em relação a amostragem aleatória simples é dada por $h(N) = g(N)k(N)$, através da expressão (2.59).

De maneira similar à expressão (2.72), $E[\nu]$ pode ser escrita como uma função de $N$ da seguinte forma

$$f(N) = N - \sum_{i=1}^{N} \frac{(N - n_1 - m_i - a_i + 1)}{(N - m_i - a_i + 1)} \cdots \frac{(N - n_1)}{N}. \tag{2.77}$$

Dado que o número de unidades que satisfazem a condição e as unidades de beirada é fixo em $\eta$, o tamanho da rede para todas as outras unidades é $m_i = 1$ pela definição de rede e o número de unidades satisfazendo a condição é $a_i = 0$.

Sendo assim, substituindo na fórmula acima temos que:

$$\begin{aligned} f(N) &= N - (N - \eta)\frac{(N - n_1)}{N} - \sum_{i=1}^{\eta} \frac{N - n_1 - m_i - a_i + 1}{N - m_i - a_i + 1} \cdots \frac{N - n_1}{N} \\ &= n_1 + \eta - \frac{n_1\eta}{N} - \sum_{i=1}^{\eta} \frac{N - n_1 - m_i - a_i + 1}{N - m_i - a_i + 1} \cdots \frac{N - n_1}{N}. \end{aligned}$$

A soma acima contém $\eta$ termos sendo que cada um deles é composto por fatores que se aproximam de 1 quando $N$ vai para o infinito. Como a parcela $\frac{n_1\eta}{N}$ tende a zero, então temos que:

$$\lim_{N\to\infty} f(N) = n_1 + \eta - \eta = n_1. \tag{2.78}$$

Isso significa que, com o aumento da raridade geográfica, o tamanho amostral final é o da própria amostra inicial.

Além disso, com $N$ tendendo ao infinito, a função $g(N) = [f(N)/n_1]\{(N - n_1)/[N - f(N)]\}$ tende a 1 e a função $k(N) = 1 - \{A/[B - (T/N)]\}$ tende a $1 - A/B$, o que implica em $\lim_{N\to\infty} h(N) = 1 - A/B$.

Sendo assim, quando a variância dentro da rede é positiva, as populações geograficamente raras são amostradas melhor utilizando um desenho adaptativo ao invés de um convecional.

**Fatores que influenciam a eficiência**

Como vimos, a eficiência relativa da amostragem por conglomerados adaptativos e seus respectivos custos dependem de alguns fatores. Segundo Thompson & Seber (1996), os fatores que favorecem a eficiência da amostragem por conglomerados adaptativos são:

1. a variância dentro da rede corresponde a uma alta proporção da variância total da população;

2. a área de estudo é grande em relação a população de estudo, ou seja, trata-se de uma população rara;

3. o tamanho amostral final não é muito maior do que o tamanho amostral inicial;

4. o custo de observar unidades dentro dos conlgomerados ou redes é menor devido a questões logísticas;

5. o custo de observar unidades satisfazendo a condição imposta é maior do que o custo de se observar unidades que não satisfazem a condição;

6. em alguns casos podem ser utilizadas variáveis auxiliares de baixo custo para estabelecer as condições de amostragem adicional;

7. as análises acima realizadas são para o estimador Hansen-Hurwitz. Maiores eficiências podem ser obtidas usando outros estimadores como por exemplo Horwitz-Thompson.

## 2.4 Fórmulas Úteis

Aqui, listamos as fórmulas dos estimadores da média populacional e dos estimadores de suas variâncias referentes a amostragem por conglomerados adaptativos:

O estimador do tipo Horvitz-Thompson é dado por $\hat{\mu}_{HT} = \frac{1}{N}\sum_{k=1}^{\kappa}\frac{y_k^*}{\alpha_k}$, com variância estimada por $\hat{\text{Var}}(\hat{\mu}_{HT}) = \frac{1}{N^2}\sum_{k=1}^{\kappa}\sum_{j=1}^{\kappa}\frac{y_j^* y_k^*}{\alpha_j\alpha_k}(\alpha_{jk} - \alpha_j\alpha_k)$.

O estimador do tipo Hansen-Hurwitz é dado por $\hat{\mu}_{HH} = \frac{1}{n_1}\sum_{i=1}^{n_1} w_i$, com variância estimada por $\hat{\text{Var}}(\hat{\mu}_{HH}) = \frac{N-n_1}{Nn_1(N-1)}\sum_{i=1}^{n_1}(w_i - \hat{\mu}_{HH})^2$, onde:

- $\mu$ é a média populacional da variável de interesse $y$, por exemplo, média do número de elementos por unidade;

- $N$ é o total de unidades (primárias) na população;

- $\kappa$ é o número de redes distintas selecionadas;

- $\alpha_k$ é a probabilidade de seleção da rede $k$;

- $y_k^*$ é a soma de $y$ na rede $k$;

- $n_1$ é o tamanho da amostra inicial;
- $w_i$ é a média de $y$ dentro da rede que contém a unidade $i$;

### 2.4.1 Exemplo de Cálculo

Considere a população fictícia representada pela Figura 2.8. A seleção inicial tamanho $n_1 = 4$ foi realizada por AAS e gerou 4 redes, sendo que 3 delas são distintas. As unidades de seleção inicial correspondem às unidades em preto, enquanto que as unidades hachuradas correspondem às unidades de beirada. Aqui assumimos a condição $C = 0$ e $y_i$ a contagem dos elementos na unidade $i$.

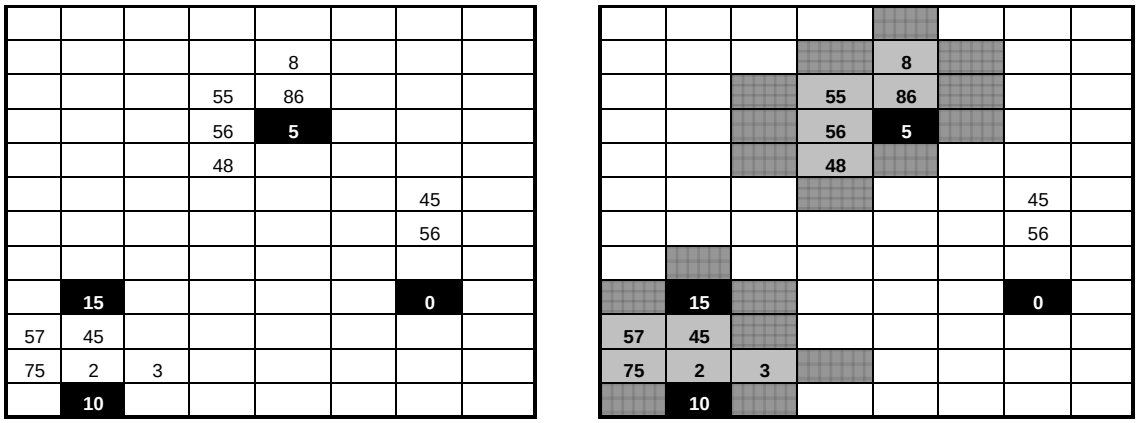


Figura 2.8: *Exemplo de seleção*

Para essa população, temos os parâmetros $\tau = 566$, $\mu = 5,9$ e $N = 96$. Para computar os estimadores, lembramos que $\alpha_k = 1 - \left[\frac{\binom{N-x_k}{n_1}}{\binom{N}{n_1}}\right]$ e $x_k$ é o número de unidades em uma rede. Substituímos os valores nas expressões dos estimadores HH e HT como segue:

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $x_1$ | 6 | $w_1$ | 43,0 | $\alpha_1$ | 0,231 | $y_1^*$ | 258 | $\frac{y_1^*}{\alpha_1}$ | 1117,76 | $\hat{\mu}_{HT}$ | 19,8 |
| $x_2$ | 7 | $w_2$ | 29,6 | $\alpha_2$ | 0,265 | $y_2^*$ | 207 | $\frac{y_2^*}{\alpha_2}$ | 781,12 | | |
| $x_3$ | 7 | $w_3$ | 29,6 | $\alpha_3$ | 0,265 | $y_3^*$ | 207 | $\frac{y_3^*}{\alpha_3}$ | 781,12 | $\hat{\mu}_{HH}$ | 25,5 |
| $x_4$ | 1 | $w_4$ | 0,0 | $\alpha_4$ | 0,042 | $y_4^*$ | 0 | $\frac{y_4^*}{\alpha_4}$ | 0 | | |

Importante observar que para o cálculo dos estimador HT, deve-se considerar as redes distintas portanto, deve-se considerar uma das redes 2 ou 3 no cálculo. O mesmo não ocorre para o estimador HH. Através dos resultados vimos que as estimativas dos estimadores HT e HH apresentaram um erro absoluto de 13,9 e 19,6 respectivamente.

# Capítulo 3

# Amostragem Seqüencial em Duas Etapas

A amostragem seqüencial em duas etapas foi proposta por Salehi & Smith (2005) como uma alternativa de desenho para populações raras e agrupadas. Segundo os autores, a proposta da amostragem seqüencial em duas etapas é aproveitar o agrupamento da população rara sem necessitar definir uma vizinhança como na amostragem por conglomerados adaptativos. Ainda segundo os autores, a amostragem por conglomerados adaptativos possui alguns fatores que limitam sua eficiência, que são:

- **Vizinhança e unidades de beirada**

  Nesse esquema amostral, o formato da vizinhança poderá levar a uma seleção exagerada de unidades de beirada. Segundo Salehi (1999) as unidades de beirada são elementos indesejáveis uma vez que não contribuem para a precisão das estimativas;

- **Simetria da vizinhança**

  A escolha da vizinhança pode ser complicada ou impossível pois é necessária a simetria que requer que se a unidade $i$ está na vizinhança da unidade $j$, então $j$ deve estar na vizinhança de $i$. Segundo Salehi & Smith (2005) isso pode ser impossível pois nem sempre a simetria existe. Um exemplo é o da relação de amizade, em que uma pessoa pode se considerar amiga da outra mas a outra não, logo não existe simetria e a amostragem por conglomerados adaptativos não funcionaria com esse tipo de vizinhança;

- **Problemas práticos na regra de parada**

  O procedimento da amostragem por conglomerados adaptativos é o de percorrer toda

a vizinhança até que a condição $C$ não seja mais satifeita. Segundo Salehi & Smith (2005), em algumas situações isso não é prático como por exemplo no caso de pesquisas de pesca em que grandes regiões são estudadas. Alguns estudiosos nessa área como Lo *et al.* (2001) e Hanselman *et al.* (2003) aplicaram a amostragem por conglomerados adaptativos restringindo a amostragem adaptativa por uma regra arbitrária, chegando em estimativas viesadas.

A amostragem seqüencial em dois estágios é uma abordagem que não gera unidades de beirada e nem requer a definição de vizinhança. O desenho difere dos propostos por Francis (1984), Jolly & Hampton (1990), Salehi & Seber (1997) e Christman (2003) que são desenhos de dois estágios ou estratificados que incorporam amostragem adaptativa ou seqüencial.

## 3.1 Seleção Amostral

O primeiro passo da seleção amostral é a escolha da partição da população que forma as unidades primárias de seleção. Essa escolha deve ser feita de acordo com o conhecimento prévio e não tem uma forma fechada de ser realizada. Segundo Salehi & Smith (2005), devemos tomar como guias as informações disponíveis e as restrições naturais mesmo que com freqüência uma informação precisa não esteja disponível.

Seja uma população de $N$ unidades secundárias, representadas pelos retângulos menores na Figura 3.1. Agrupamos essas unidades secundárias em $M$ unidades primárias de tamanho $N_i$. Seja $(i, j)$ a $j$-ésima unidade (secundária) da $i$-ésima unidade primária com o valor da variável de interesse $y_{ij}$ que pode ser por exemplo uma contagem, área poluída, biomassa (medida do total de material biológico vivo em uma área, geralmente medido por uma unidade de peso por metro quadrado), etc. Seja $\tau_i = \sum_{j=1}^{N_i} y_{ij}$ a soma dos valores de $y$ na unidade primária $i$ e $\tau = \sum_{i=1}^{M} \tau_i$ o total populacional da variável de interesse.

No primeiro estágio da amostragem, escolhemos uma amostra de tamanho $m$ das $M$ unidades primárias. Essa seleção pode ser qualquer uma com probabilidade de seleção da $i$-ésima unidade primária conhecida $\pi_i$ assim como a probabilidade de seleção de cada par $i$ e $i'$, $\pi_{ii'}$. No segundo estágio, dentro de cada unidade primária $i$ adotamos o seguinte procedimento:

- selecionamos uma amostra aleatória simples sem reposição de tamanho $n_{i1}$ de unidades secundárias na unidade primária $i$, e então $n_1 = \sum_{i=1}^{m} n_{i1}$ é o tamanho total da amostra inicial;

- se uma condição $C$ é satisfeita por qualquer unidade secundária dentro da respectiva unidade primária, é realizada a seleção de um número predeterminado de unidades secundárias adicionais $n_{i2}$, e então $n_2 = \sum_{i=1}^{m} n_{i2}$ é o número total de unidades adicionadas pela condição $C$. Sendo assim $n_2$ é uma variável aleatória.

Seja $l_i$ o número de unidades da unidade primária $i$ satisfazendo a condição $C$ na amostra final da unidade primária $i$. Note que quando $m = M$, temos uma amostra seqüencial estratificada. Na Figura 3.1, vemos um exemplo correspondente a uma população de marrecos da asa azul dado por Smith *et al.* (1995).

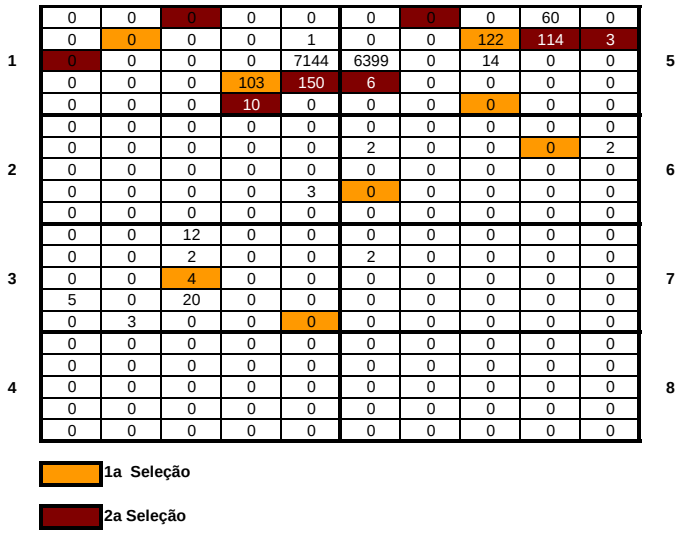


Figura 3.1: *População de marrecos da asa azul*

Nesse exemplo temos $M = 8$ unidades primárias (retângulos maiores) sendo que $m = 4$ foram selecionadas no primeiro estágio da amostra que foi sorteada de maneira aleatória.

No segundo estágio, foram selecionadas 2 unidades secundárias nas unidades primárias selecionadas que na Figura 3.1 correspondem aos quadrados na cor mais clara. Nas unidades primárias 1 e 5 a condição $C = y_i > 10$ foi satisfeita e portanto em cada uma delas foram selecionadas aleatoriamente e sem reposição mais 4 unidades primárias que correspondem aos quadrados mais escuros.

## 3.2 Estimadores

Salehi & Smith (2005) utilizam o estimador de Murthy (1957) para $\tau = \sum_{i=1}^{N} y_i$ sob esse desenho. Seja $I_{ij}$ uma função indicadora do evento "$j$ é selecionada na primeira extração da unidade primária $i$", com probabilidade $p_{ij}$ e $N_i$ o total de unidades secundárias na unidade primária $i$. Partindo do estimador de Raj (1956), podemos calcular

$$\hat{t_i} = \sum_{j=1}^{N_i} \frac{y_{ij}}{p_{ij}} I_{ij}. \tag{3.1}$$

Esse estimador é trivial não viciado para $\tau_i$. Se utilizarmos a melhoria por Rao-Blackwell, obteremos o estimador de Murthy. Seja $s_i$ o conjunto das unidades secundárias selecionadas dentro da unidade primária $i$, $P(s_i)$ a probabilidade de obter a amostra $s_i$ na unidade primária $i$ e $P(s_i \mid j)$ a probabilidade condicional na unidade primária $i$ de tomar a amostra dado que a $j$-ésima unidade secundária foi selecionada na primeira retirada. Então

$$\hat{\tau_i} = \mathrm{E}\left[\hat{t_i} \mid s_i\right] = \sum_{j=1}^{N_i} \frac{y_{ij}}{p_{ij}} P(j \mid s_i), \tag{3.2}$$

que aplicando o Teorema de Bayes, se torna

$$\hat{\tau_i} = \sum_{j=1}^{N_i} \frac{y_{ij}}{p_{ij}} P(s_i \mid j) p_{ij} \frac{1}{P(s_i)} = \sum_{j \in s_i} \frac{P(s_i \mid j)}{P(s_i)} y_{ij}. \tag{3.3}$$

Para avaliar $\frac{P(s_i|j)}{P(s_i)}$ é necessário desmembrar nos seguintes casos:

$$\frac{P(s_i \mid j)}{P(s_i)} = \begin{cases} \frac{N_i}{n_{i1}} & \text{se } n_{i2} = 0 \\ \frac{N_i}{n_{i1}+n_{i2}} & \text{se } n_{i2} > 0 \text{ e } l_i > n_{i2} \\ \frac{N_i (n_{i1}+n_{i2}-1)!}{(n_{i1}+n_{i2})! - n_{i2}!(n_{i1}+n_{i2}-l_i)!/(n_{i2}-l_i)!} & \text{se } n_{i2} > 0 \text{ e } l_i \leq n_{i2} \text{ e } j \text{ satisfaz } C \\ \frac{N_i\{(n_{i1}+n_{i2}-1)! - n_{i2}!(n_{i1}+n_{i2}-1-l_i)!/(n_{i2}-l_i)!\}}{(n_{i1}+n_{i2})! - n_{i2}!(n_{i1}+n_{i2}-l_i)!/(n_{i2}-l_i)!} & \text{se } n_{i2} > 0 \text{ e } l_i \leq n_{i2} \text{ e } j \text{ não satisfaz } C \end{cases}$$

Aqui, $l_i$ corresponde ao número de unidades secundárias na amostra da unidade primária $i$ que satisfazem a condição $C$. Para a estimação do total $\tau = \sum_{i=1}^{M} \tau_i$, como a seleção em cada unidade primária e a subamostragem é realizada independentemente, podemos aplicar a teoria do estimador proposto por Horvitz & Thompson (1952). Sendo assim, temos

$$\hat{\tau} = \sum_{i=1}^{m} \frac{\hat{\tau}_i}{\pi_i}. \tag{3.4}$$

Encontramos a seguir as propriedades de vício e variância para $\hat{\tau}_i$ e em seguida para $\hat{\tau}$.

**Vício**

Para o estimador de $\tau_i$, podemos notar que, pela aplicação do teorema de Rao-Blackwell o estimador proposto tem a mesma esperança de $\hat{t_i}$, ou seja, é não viciado para $\tau_i$.

$$\mathrm{E}(\hat{\tau}_i) = \mathrm{E}(\hat{t_i}) = \sum_{j=1}^{N_i} \frac{y_{ij}}{p_{ij}} E[I_{ij}] = \sum_{j=1}^{N_i} \frac{y_{ij}}{p_{ij}} p_{ij} = \tau_i. \tag{3.5}$$

Para o estimador de $\tau$, denotamos $\mathbf{s}^* = (s_1, s_2, s_3, \ldots, s_m)$ e dessa forma temos que

$$\mathrm{E}(\hat{\tau}) = \mathrm{E}(\mathrm{E}(\hat{\tau} \mid s^*)) = \mathrm{E}\left(\sum_{i=1}^{M} \frac{\mathrm{E}(\hat{\tau}_i) I_i}{\pi_i}\right) = \sum_{i=1}^{M} \tau_i = \tau. \tag{3.6}$$

Sendo assim, os estimadores são não viesados.

**Variância**

A variância de $\hat{\tau}_i$ é dada por

$$\mathrm{Var}(\hat{\tau}_i) = \sum_{j=1}^{N_i} \sum_{j<j'}^{N_i} \left(1 - \sum_{s_i \ni j,j'} \frac{P(s_i \mid j) P(s_i \mid j')}{P(s_i)}\right) \left(\frac{y_{ij}}{p_{ij}} - \frac{y_{ij'}}{p_{ij'}}\right)^2 p_{ij} p_{ij'}. \tag{3.7}$$

Como $p_{ij} = \frac{n_{i1}}{N_i}$ para todo $j = 1, 2, \ldots, N_i$,

$$\mathrm{Var}(\hat{\tau}_i) = \sum_{j=1}^{N_i} \sum_{j<j'}^{N_i} \left(1 - \sum_{s_i \ni j,j'} \frac{P(s_i \mid j) P(s_i \mid j')}{P(s_i)}\right) \left(y_{ij} - y_{ij'}\right)^2. \tag{3.8}$$

Seja $P(s_i \mid j, j')$ a probabilidade da amostra $s_i$ dado que as unidades secundárias $j$ e $j'$ foram escolhidas nas duas primeiras retiradas na unidade primária $i$. O estimador não viesado da variância é dado por

$$\hat{Var}(\hat{\tau}_i) = \sum_{j \in s_i} \sum_{j<j'} \left(\frac{P(s_i \mid j, j')}{P(s_i)} - \frac{P(s_i \mid j) P(s_i \mid j')}{P(s_i)^2}\right) \left(y_{ij} - y_{ij'}\right)^2. \tag{3.9}$$

Para calcularmos $P(s_i \mid j, j')$, é necessário desmembrar nos casos:

$$\frac{P(s_i \mid j, j')}{P(s_i)} = \begin{cases} \frac{N_i(N_i-1)}{n_{i1}(n_{i1}-1)} & \text{se } n_{i2} = 0 \\ \frac{N_i(N_i-1)}{(n_{i1}+n_{i2})(n_{i1}+n_{i2}-1)} & \text{se } n_{i2} > 0 \text{ e } l_i > n_{i2} \\ \frac{N_i(N_i-1)(n_{i1}+n_{i2}-2)!}{(n_{i1}+n_{i2})!-n_{i2}!(n_{i1}+n_{i2}-l_i)!/(n_{i2}-l_i)!} & \text{se } n_{i2} > 0 \text{ e } l_i \leq n_{i2} \text{ e } j \text{ ou } j' \text{ satisfaz } C \\ \frac{N_i\{(n_{i1}+n_{i2}-2)!-n_{i2}!(n_{i1}+n_{i2}-2-l_i)!/(n_{i2}-l_i)!\}}{(n_{i1}+n_{i2})!-n_{i2}(n_{i1}+n_{i2}-l_i)!/(n_{i2}-l_i)!} & \text{se } n_{i2}>0 \text{ e } l_i \leq n_{i2}, j \text{ e } j' \text{ não satisfazem } C \end{cases}$$

A variância do total $\hat{\tau}$ é dada por

$$\mathrm{Var}(\hat{\tau}) = \mathrm{Var}(\mathrm{E}(\hat{\tau} \mid \mathbf{s}^*)) + \mathrm{E}(\mathrm{Var}(\hat{\tau} \mid \mathbf{s}^*)). \tag{3.10}$$

Mas,

$$\mathrm{E}(\hat{\tau} \mid \mathbf{s}^*) = \sum_{i=1}^{M} \frac{\tau_i I_i}{\pi_i} \tag{3.11}$$

e

$$\mathrm{Var}(\hat{\tau} \mid \mathbf{s}^*) = \mathrm{Var}\left(\sum_{i=1}^{M} \frac{\hat{\tau}_i I_i}{\pi_i}\right),$$

sendo que $\mathrm{Cov}(\hat{\tau}_i, \hat{\tau}_{i'}) = 0$ uma vez que estes são estimadores do total para unidades primárias distintas. A variância passa a ser

$$\mathrm{Var}(\hat{\tau} \mid \mathbf{s}^*) = \sum_{i=1}^{M} \frac{\mathrm{Var}(\hat{\tau}_i) I_i^2}{\pi_i^2} = \sum_{i=1}^{M} \frac{\mathrm{Var}(\hat{\tau}_i) I_i}{\pi_i^2}. \tag{3.12}$$

Substituindo (3.11) e (3.12) em (3.10), obtemos

$$\begin{aligned} \mathrm{Var}(\hat{\tau}) &= \mathrm{Var}\left(\sum_{i=1}^{M} \frac{\tau_i I_i}{\pi_i}\right) + \mathrm{E}\left(\sum_{i=1}^{M} \frac{\mathrm{Var}(\hat{\tau}_i) I_i}{\pi_i^2}\right) \\ &= \sum_{i=1}^{M}\sum_{i'=1}^{M} \frac{\tau_i \tau_{i'}}{\pi_i \pi_{i'}}(\pi_{ii'} - \pi_i \pi_i') + \sum_{i=1}^{M} \frac{\mathrm{Var}(\hat{\tau}_i)}{\pi_i}, \end{aligned} \tag{3.13}$$

em que $\pi_{ii}$ deve ser interpretado por $\pi_{i.}$. Segundo Salehi (2003), um estimador não viesado para essa variância é dado por

$$\hat{\mathrm{Var}}(\hat{\tau}) = \sum_{i=1}^{m}\sum_{i'=1}^{m} \frac{\hat{\tau}_i \hat{\tau}_{i'}}{\pi_{ii'} \pi_i \pi_{i'}}(\pi_{ii'} - \pi_i \pi_i') + \sum_{i=1}^{M} \frac{\hat{Var}(\hat{\tau}_i)}{\pi_i}. \tag{3.14}$$

## 3.3 Eficiência e Tamanho Amostral

### 3.3.1 Amostragem Seqüencial em Dois estágios versus Amostragem Aleatória Simples

Segundo Kish (1965), o desenho amostral em duas etapas em geral não possui maiores ganhos em termos de variância, pois uma vez que há a restrição da seleção das unidades secundárias às unidades primárias selecionadas, a liberdade de espalhamento da amostra é menor. Por outro lado, como dentro de uma mesma unidade primária o custo de coleta é menor, temos uma redução de custos. Infelizmente até esse ponto não conseguimos estabelecer as relações teóricas como fizemos para a amostragem por conglomerados adaptativos (ver Seção 2.3). As simulações que realizaremos a seguir nos fornecerão alguns subsídios sobre essas relações.

### 3.3.2 Tamanho Amostral e Custos

Assim como na amostragem por conglomerados adaptativos, o tamanho amostral na amostragem seqüencial em dois estágios é uma variável aleatória pois a incorporação dos elementos adicionais na amostra dependem do cumprimento da condição $C$. O tamanho amostral é dado por

$$\nu^* = \sum_{i=1}^{M} n_{i1}K_i + n_{i2}J_iK_i, \tag{3.15}$$

onde $K_i$ indica a seleção da unidade primária e $J_i$ indica a presença de qualquer unidade que satisfaça a condição $C$ na unidade primária $i$ durante o primeiro estágio de seleção. Logo, o tamanho amostral esperado é

$$E[\nu^*] = \sum_{i=1}^{M} n_{i1}E[K_i] + n_{i2}E[J_i]E[K_i] = \frac{m}{M}\sum_{i=1}^{M} n_{i1} + n_{i2}\left(1 - \frac{\binom{N_i - L_i}{n_{i1}}}{\binom{N_i}{n_{i1}}}\right), \tag{3.16}$$

onde $L_i$ é o número de unidades secundárias que satisfazem $C$ na unidade primária $i$. Analogamente às expressões da amostragem por conglomerados adaptativos temos que

$$C_{T_s} = c_0 + c_1n_1 + c_2(\nu^* - n_1), \tag{3.17}$$

onde $c_0$ é um custo fixo e $c_1$ é um custo por unidade selecionada aleatoriamente e $c_2$ é o custo por unidade selecionada adaptativamente. Esse custo é limitado a $c_0 + c_1n_1 + c_2n_2$

pois corresponde ao caso em que para todas as unidades primárias existe ao menos uma unidade secundária satisfazendo a condição $C$. Sendo assim podemos calcular um custo total esperado dado por

$$E[C_{T_a}] = c_0 + c_1 n_1 + c_2(E[\nu^*] - n_1). \tag{3.18}$$

## 3.4 Fórmulas Úteis

Aqui mostramos algumas fórmulas para a amostragem seqüencial em duas etapas.

O estimador para a amostragem seqüencial em duas etapas é dado por $\hat{\tau} = \sum_{i=1}^{m} \frac{\hat{\tau}_i}{\pi_i}$, com $\hat{\tau}_i = \sum_{j \in s_i} \frac{P(s_i|j)}{P(s_i)} y_{ij}$ e

$$\frac{P(s_i \mid j)}{P(s_i)} = \begin{cases} \frac{N_i}{n_{i1}} & \text{se } n_{i2} = 0 \\ \frac{N_i}{n_{i1}+n_{i2}} & \text{se } n_{i2} > 0 \text{ e } l_i > n_{i2} \\ \frac{N_i(n_{i1}+n_{i2}-1)!}{(n_{i1}+n_{i2})! - n_{i2}!(n_{i1}+n_{i2}-l_i)!/(n_{i2}-l_i)!} & \text{se } n_{i2} > 0 \text{ e } l_i \leq n_{i2} \text{ e } j \text{ satisfaz } C \\ \frac{N_i\{(n_{i1}+n_{i2}-1)! - n_{i2}!(n_{i1}+n_{i2}-1-l_i)!/(n_{i2}-l_i)!\}}{(n_{i1}+n_{i2})! - n_{i2}!(n_{i1}+n_{i2}-l_i)!/(n_{i2}-l_i)!} & \text{se } n_{i2} > 0 \text{ e } l_i \leq n_{i2} \text{ e } j \text{ não satisfaz } C \end{cases}$$

O estimador da variância do total é dado por $\hat{\text{Var}}(\hat{\tau}) = \sum_{i=1}^{m} \sum_{i'=1}^{m} \frac{\hat{\tau}_i \hat{\tau}_{i'}}{\pi_{ii'} \pi_i \pi_{i'}} (\pi_{ii'} - \pi_i \pi'_i) + \sum_{i=1}^{M} \frac{\hat{Var}(\hat{\tau}_i)}{\pi_i}$ e o estimador da variância dentro de cada unidade primária é dada por $\hat{\text{Var}}(\hat{\tau}_i) =$

$\sum_{j\in s_i}\sum_{j<j'}\left(\frac{P(s_i|j,j')}{P(s_i)}-\frac{P(s_i|j)P(s_i|j')}{P(s_i)^2}\right)(y_{ij}-y_{ij'})^2$, onde:

$$\frac{P(s_i \mid j,j')}{P(s_i)}=\begin{cases}\frac{N_i(N_i-1)}{n_{i1}(n_{i1}-1)} & \text{se } n_{i2}=0\\ \frac{N_i(N_i-1)}{(n_{i1}+n_{i2})(n_{i1}+n_{i2}-1)} & \text{se } n_{i2}>0 \text{ e } l_i>n_{i2}\\ \frac{N_i(N_i-1)(n_{i1}+n_{i2}-2)!}{(n_{i1}+n_{i2})!-n_{i2}!(n_{i1}+n_{i2}-l_i)!/(n_{i2}-l_i)!} & \text{se } n_{i2}>0 \text{ e } l_i\leq n_{i2} \text{ e } j \text{ ou } j' \text{ satisfaz } C\\ \frac{N_i\{(n_{i1}+n_{i2}-2)!-n_{i2}!(n_{i1}+n_{i2}-2-l_i)!/(n_{i2}-l_i)!\}}{(n_{i1}+n_{i2})!-n_{i2}(n_{i1}+n_{i2}-l_i)!/(n_{i2}-l_i)!} & \text{se } n_{i2}>0 \text{ e } l_i\leq n_{i2}, j \text{ e } j' \text{ não satisfazem } C\end{cases}$$

e

- $\pi_i$ é a probabilidade de seleção da unidade primária $i$;
- $\tau$ é o total da variável de interesse y, ou seja, $N\mu$;
- $M$ é o número de unidades primárias na população;
- $m$ é o número de unidades primárias selecionadas;
- $N_i$ é o número de unidades secundárias dentro da unidade primária $i$;
- $n_{i1}$ é o tamanho amostral inicial na unidade primária $i$;
- $n_{i2}$ é o tamanho da amostra adicionada adaptativamente na unidade primária $i$;
- $l_i$ é o número de unidades secundárias da amostra satisfazendo a condição $C$.

### 3.4.1 Exemplo de Cálculo

Considerando-se a população do marreco da asa azul da figura 3.1, temos que das $M = 8$ unidades primárias, $m = 4$ foram selecionadas. A população possui $\tau = 14181$, $N = 200$ e $\mu = 70,905$. Além disso, $N_1 = N_2 = ... = N_8 = 25$. Na unidade primária 1, selecionamos as unidades secundárias com valores de $y_1 = \{0, 103, 0, 0, 10, 150\}$ e nas unidades primárias 3, 5 e 6 respectivamente $y_2 = \{4, 0\}$, $y_3 = \{122, 0, 0, 6, 114, 3\}$ e $y_4 = \{0, 0\}$.

Assumindo $C = 10$, podemos calcular o estimador $\hat{\tau}$ usando as expressões da seção anterior, nas quais substituimos os valores $l_1 = 2$, $l_2 = 0$, $l_3 = 2$ e $l_4 = 0$, o tamanho amostral inicial $n_{11} = ... = n_{41} = 2$ e o tamanho amostral adicional $n_{12} = 4$, $n_{22} = 0$, $n_{32} = 4$ e $n_{42} = 0$.

Usando esses valores é possível calcular $\frac{P(s_i|j)}{P(s_i)}$ que denominamos aqui $r_i$, então temos $r_1 = \{2,78; 6,94; 2,78; 2,78; 2,78; 6,94\}$, $r_2 = \{12,5; 12,5\}$, $r_3 = \{6,94; 2,78; 2,78; 2,78; 6,94; 2,78\}$ e $r_4 = \{12,5; 12,5\}$. Logo, para alcançarmos o estimador $\hat{\tau}_i = \sum_{j \in s_i} \frac{P(s_i|j)}{P(s_i)} y_{ij}$, calculamos os valores de $y_i r_i$: $y_1 r_1 = \{0; 715,28; 0; 0; 27,78; 1041,67\}$, $y_2 r_2 = \{50; 0\}$, $y_3 r_3 = \{847,22; 0; 0; 16,67; 791,67; 8,33\}$ e $y_4 r_4 = \{0; 0\}$. Finalmente, temos $\hat{\tau}_1 = 1784,72$, $\hat{\tau}_2 = 50$, $\hat{\tau}_3 = 1663,89$, $\hat{\tau}_4 = 0$, o que significa uma estimativa $\hat{\tau} = 6997,22$.

# Capítulo 4

# Simulações

Neste capítulo simulamos a amostragem de uma população verdadeira e duas artificiais com o intuito de compreender os efeitos referentes às alterações de parâmetros da amostragem por conglomerados adaptativos e da amostragem seqüencial em duas etapas. Também é objetivo deste capítulo realizar uma comparação da eficiência dos estimadores estudados, para identificar o método cuja eficiência relativa é maior.

## 4.1 Populações

Para realizar as simulações nos baseamos em duas populações: a população de marrecos da asa azul de Smith *et al.* (1995) e uma população artificial com os elementos mais espalhados nas unidades primárias. Usamos o aplicativo R para fazer as simulações.

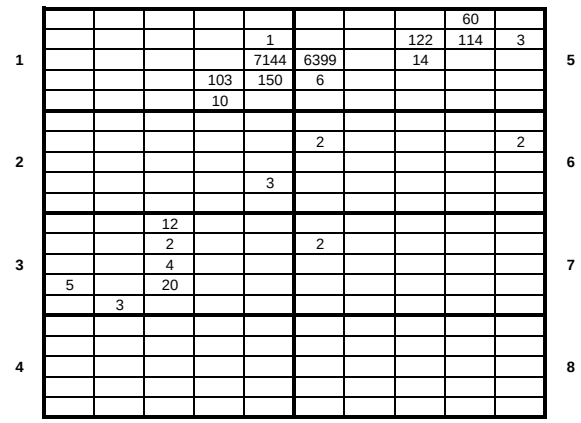


Figura 4.1: *Marrecos da asa azul*

A população dos marrecos da asa azul (Figura 4.1) constitui-se em uma população de elementos raros e extremamente agrupados. Os números que aparecem na figura correspondem ao número de marrecos por unidade. Duas unidades secundárias concentram cerca de 96% dos elementos da população.

A população artificial 1 (Figura 4.2) é uma modificação da população de marrecos da asa azul. Diminuimos o número de unidades nas duas maiores redes da população original, espalhando-as de maneira a gerar uma população menos rara e menos agrupada.

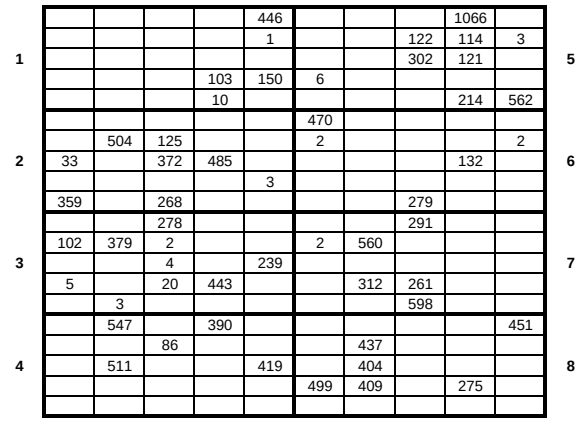


Figura 4.2: *População artificial 1 - menos mara e menos agrupada*

A população artificial 2 (Figura 4.3) constitui-se em uma modificação da população de marrecos da asa azul. Ao criar essa população, utilizamos o mesmo número de elementos da população original por unidade, espalhando-as no grid, ou seja, é tão rara quanto a população original mas é menos agrupada.

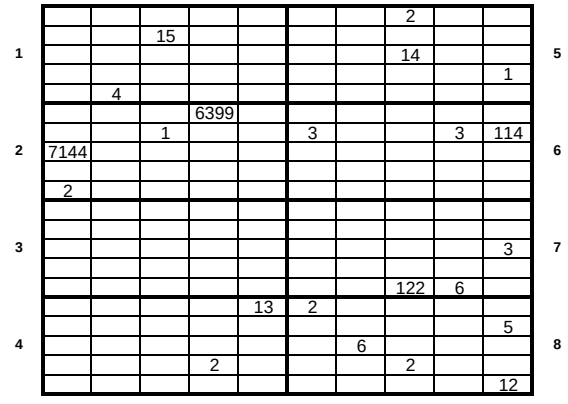


Figura 4.3: *População artificial 2 - rara e não agrupada*

Para elucidar as diferenças entre as populações apresentadas, utilizamos uma medida de

raridade da população e outra com respeito ao agrupamento dos elementos em unidades vizinhas. As densidade das três populações são as me Para avaliar a raridade da população, utilizamos a proporção de unidades contendo ao menos um elemento da população rara, $PV = \frac{1}{N}\sum_{i=1}^{N} \mathrm{I}\,(e_i > 0)$, onde $e_i$ é o número de elementos da população rara na unidade $i$. A variabilidade dentro das redes, (ver expressão 2.59) VIR $= \frac{\sum_{k=1}^{K}\sum_{i\in B_k}(y_i - w_{k(i)})^2}{\sum_{i=1}^{N}(y_i-\mu)^2}$, quando usamos a condição $C = 0$, pode ser considerada uma medida relacionada ao grau de agrupamento da população, uma vez que, caso não haja redes de tamanho maior que um, a variabilidade dentro das redes é 0. Caso todas as unidades estejam em um único bloco, ela corresponde a 1.

| **População** | **PV** | **VIR** |
|---|---|---|
| **Original** | 0.11 | 0.7114 |
| **Artificial 1** | 0.27 | 0.3246 |
| **Artificial 2** | 0.11 | 0.0002 |

Tabela 4.1: *PV e VIR para cada população*

Pelos resultados da Tabela 4.1, podemos perceber que as medidas são pertinentes às nossas premissas na criação das populações artificiais 1 e 2.

## 4.2 Resultados

Utilizamos as simulações para entender as diferenças de eficiência dos métodos de amostragem estudados. Para cada uma das populações, geramos 10.000 amostras para cada parametrização da amostragem por conglomerados adaptativos em um e dois estágios e a amostragem seqüencial em duas etapas. Para todas as simulações, utilizamos a vizinhança em formato de cruz, ou seja, consideramos vizinhos os elementos que dividem uma mesma linha de borda. Para as amostragens estudadas para populações raras utilizamos a variância amostral das 10.000 amostras para calcular a eficiência enquanto que para as amostras simples e em dois estágios utilizamos as expressões teóricas assim como no artigo Smith *et al.* (1995) (pg 780).

### Amostragem por conglomerados adaptativos

Para efetuar os cálculos de eficiência, cujos resultados estão disponíveis na Tabela 4.2, calculamos a variância amostral das 10.000 estimativas de $\hat{\tau}_{HT} = N\hat{\mu}_{HT}$ e $\hat{\tau}_{HH} = N\hat{\mu}_{HH}$

e comparamos com a variância do estimador para o total da amostra aleatória simples $\hat{\tau}_{aas} = \frac{1}{f}\sum_{i=1}^{N} y_i$, dada por

$$\mathrm{Var}(\hat{\tau}_{aas}) = \frac{1-f}{n}(S_y N)^2 \tag{4.1}$$

com $N = \sum_{i=1}^{M} N_i$, $n = \mathrm{E}(\nu)$, $f = \frac{n}{N}$ e $S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2$ (Sarndal *et al.* (1992)).

A eficiência do estimador $\hat{\tau}_{HT}$ é então dada pela relação $\mathrm{ef}(\hat{\tau}_{HT}) = \frac{\mathrm{Var}(\hat{\tau}_{aas})}{\mathrm{Var}(\hat{\tau}_{HT})}$ e para o estimador $\hat{\tau}_{HH}$ temos $\mathrm{ef}(\hat{\tau}_{HH}) = \frac{\mathrm{Var}(\hat{\tau}_{aas})}{\mathrm{Var}(\hat{\tau}_{HH})}$.

| | **Condição C = 0 (VIR=0.71)** | | | **Condição C = 5 (VIR=0.66)** | | | **Condição C = 10 (VIR=0.60)** | | |
|---|---|---|---|---|---|---|---|---|---|
| $\mathbf{n_1}$ | $\mathbf{E}(\nu)$ | $\mathbf{ef}(\hat{\tau}_{\mathbf{HT}})$ | $\mathbf{ef}(\hat{\tau}_{\mathbf{HH}})$ | $\mathbf{E}(\nu)$ | $\mathbf{ef}(\hat{\tau}_{\mathbf{HT}})$ | $\mathbf{ef}(\hat{\tau}_{\mathbf{HH}})$ | $\mathbf{E}(\nu)$ | $\mathbf{ef}(\hat{\tau}_{\mathbf{HT}})$ | $\mathbf{ef}(\hat{\tau}_{\mathbf{HH}})$ |
| **4** | 8.63 | 1.64 | 1.53 | 6.57 | 1.86 | 1.79 | 5.77 | 1.37 | 1.34 |
| **8** | 16.84 | 1.76 | 1.57 | 12.99 | 1.93 | 1.76 | 11.40 | 1.42 | 1.33 |
| **10** | 20.68 | 1.85 | 1.61 | 15.95 | 2.03 | 1.80 | 14.11 | 1.47 | 1.36 |
| **16** | 31.72 | 2.06 | 1.61 | 24.80 | 2.23 | 1.82 | 22.10 | 1.53 | 1.35 |
| **20** | 38.22 | 2.27 | 1.63 | 30.31 | 2.41 | 1.86 | 27.21 | 1.62 | 1.37 |
| **32** | 56.08 | 3.02 | 1.69 | 45.51 | 3.04 | 1.92 | 41.95 | 1.85 | 1.41 |
| **40** | 66.44 | 3.69 | 1.75 | 55.05 | 3.67 | 1.92 | 51.27 | 2.04 | 1.40 |
| **64** | 93.03 | 8.00 | 1.84 | 80.62 | 6.97 | 2.07 | 77.12 | 2.88 | 1.45 |

Tabela 4.2: *Eficiências relativas dos estimadores $\hat{\tau}_{HT}$ e $\hat{\tau}_{HH}$ - população original*

Pelos resultados da Tabela 4.2, nota-se que para a condição $C = 10$, perde-se eficiência nos estimadores HT e HH. Isso pode ser explicado por aquilo que comentamos na Seção 2.3.2 sobre a variabilidade intra-rede. Note que para cada condição $C$ a variabilidade intra-rede muda, pois as redes têm sua configuração alterada. Se tomarmos como exemplo $n_1$=16, obervamos que a adoção da condição $C = 5$ em relação à condição $C = 0$ gera um decréscimo da variabilidade intra-rede de 5 pontos percentuais e uma diminuição do tamanho da amostra esperada em $6,9$ unidades. Já a adoção da condição $C = 10$ em relação à condição $C = 5$ leva ao decréscimo da variabilidade intra-rede de 6 pontos percentuais com uma redução da amostra final de apenas $2,7$ unidades. Isso finalmente acarreta em uma eficiência relativa maior para a condição $C = 5$.

Como podemos notar pela Tabela 4.3, a amostragem por conglomerados adaptativos perde em eficiência para a amostra aleatória simples. Uma situação como essa foi encontrada no estudo dos mexilhões de água doce de Smith *et al.* (2003). Segundo o mesmo estudo, a técnica perde em eficiência para a população dos mexilhões de água doce mas aumenta o número de elementos amostrados da população rara. Podemos observar essa relação calculando a probabilidade de inclusão de um elemento da população rara na amostragem por

| | Condição C = 0 (VIR=0.325) | | | Condição C = 5 (VIR=0.194) | | | Condição C = 10 (VIR=0.191) | | |
|---|---|---|---|---|---|---|---|---|---|
| $\mathbf{n_1}$ | $\mathbf{E}(\nu)$ | $\mathbf{ef}(\hat{\tau}_{\mathbf{HT}})$ | $\mathbf{ef}(\hat{\tau}_{\mathbf{HH}})$ | $\mathbf{E}(\nu)$ | $\mathbf{ef}(\hat{\tau}_{\mathbf{HT}})$ | $\mathbf{ef}(\hat{\tau}_{\mathbf{HH}})$ | $\mathbf{E}(\nu)$ | $\mathbf{ef}(\hat{\tau}_{\mathbf{HT}})$ | $\mathbf{ef}(\hat{\tau}_{\mathbf{HH}})$ |
| **4** | 14.1 | 0.41 | 0.40 | 10.7 | 0.44 | 0.44 | 10.0 | 0.49 | 0.48 |
| **8** | 27.4 | 0.40 | 0.38 | 20.8 | 0.46 | 0.45 | 19.5 | 0.49 | 0.49 |
| **10** | 33.2 | 0.40 | 0.39 | 25.6 | 0.46 | 0.44 | 24.1 | 0.49 | 0.48 |
| **16** | 49.6 | 0.42 | 0.40 | 39.4 | 0.48 | 0.45 | 37.0 | 0.49 | 0.47 |
| **20** | 59.3 | 0.42 | 0.39 | 47.8 | 0.47 | 0.45 | 45.1 | 0.50 | 0.47 |
| **32** | 84.4 | 0.44 | 0.39 | 70.6 | 0.49 | 0.44 | 67.0 | 0.51 | 0.46 |
| **40** | 98.0 | 0.47 | 0.40 | 83.7 | 0.48 | 0.42 | 79.8 | 0.54 | 0.47 |
| **64** | 128.6 | 0.53 | 0.39 | 115.6 | 0.54 | 0.43 | 111.8 | 0.56 | 0.44 |

Tabela 4.3: *Eficiências relativas dos estimadores $\hat{\tau}_{HT}$ e $\hat{\tau}_{HH}$ - população artificial 1*

conglomerados adaptativos, comparando com a amostragem aleatória simples. Para avaliar no nosso contexto, geramos 10.000 amostras com tamanho inicial $n_1 = 20$. Ao final avaliamos a proporção de unidades com elementos da população rara dada pela relação entre o número total de unidades amostradas com algum elemento da população rara e o número total de unidades. Os resultados estão apresentados na Tabela 4.4.

| População | Condição C = 0 | Condição C = 5 | Condição C = 10 |
|---|---|---|---|
| **Original (PV=0.11)** | 0.202 | 0.210 | 0.205 |
| **Artificial 1 (PV=0.27)** | 0.269 | 0.263 | 0.267 |
| **Artificial 2 (PV=0.11)** | 0.097 | 0.105 | 0.106 |

Tabela 4.4: *Proporção de unidades com elementos raros*

Como podemos ver pela Tabela 4.4, para o caso da população artificial 1, o método não incrementa a proporção de unidades com elementos da população em estudo. Para a população original o método traz o dobro de unidades com os elementos da população rara.

| | Condição C = 0 (VIR=0.0002) | | | Condição C = 5 (VIR=0.0001) | | | Condição C = 10 (VIR=0) | | |
|---|---|---|---|---|---|---|---|---|---|
| $\mathbf{n_1}$ | $\mathbf{E}(\nu)$ | $\mathbf{ef}(\hat{\tau}_{\mathbf{HT}})$ | $\mathbf{ef}(\hat{\tau}_{\mathbf{HH}})$ | $\mathbf{E}(\nu)$ | $\mathbf{ef}(\hat{\tau}_{\mathbf{HT}})$ | $\mathbf{ef}(\hat{\tau}_{\mathbf{HH}})$ | $\mathbf{E}(\nu)$ | $\mathbf{ef}(\hat{\tau}_{\mathbf{HT}})$ | $\mathbf{ef}(\hat{\tau}_{\mathbf{HH}})$ |
| **4** | 5.9 | 0.64 | 0.64 | 5.0 | 0.80 | 0.80 | 4.7 | 0.86 | 0.86 |
| **8** | 11.7 | 0.63 | 0.63 | 9.9 | 0.79 | 0.79 | 9.4 | 0.83 | 0.83 |
| **10** | 14.5 | 0.71 | 0.71 | 12.3 | 0.80 | 0.80 | 11.7 | 0.83 | 0.83 |
| **16** | 23.0 | 0.65 | 0.65 | 19.6 | 0.82 | 0.82 | 18.6 | 0.86 | 0.86 |
| **20** | 28.5 | 0.67 | 0.67 | 24.3 | 0.81 | 0.81 | 23.2 | 0.85 | 0.85 |
| **32** | 44.6 | 0.69 | 0.69 | 38.5 | 0.84 | 0.84 | 36.7 | 0.84 | 0.84 |
| **40** | 54.6 | 0.67 | 0.67 | 47.6 | 0.80 | 0.80 | 45.6 | 0.84 | 0.84 |
| **64** | 83.4 | 0.65 | 0.65 | 74.1 | 0.80 | 0.80 | 71.7 | 0.84 | 0.84 |

Tabela 4.5: *Eficiências relativas dos estimadores $\hat{\tau}_{HT}$ e $\hat{\tau}_{HH}$ - população artificial 2*

Como podemos notar pela Tabela 4.5, a amostragem por conglomerados adaptativos

perde em eficiência para a amostragem aleatória simples. No entanto, para essa população, as eficiências são maiores do que as apresentadas pela população artificial 1. Considerando que a variabilidade intra-rede dessa última população é muito menor, o fator que mais pesou a favor de uma melhor eficiência foi a raridade populacional, a qual pode ser observada através da Tabela 4.1.

**Amostragem por conglomerados adaptativos em duas etapas**

Na amostragem por conglomerados adaptativos em duas etapas, utilizamos como base para cálculo da eficiência, a variância da amostragem em dois estágios por ter maior similaridade no método de seleção e por não apresentar diferenças relevantes de variância em relação à amostragem aleatória simples. Sendo assim, utilizamos a expressão de cálculo da variância da amostragem em dois estágios (Sarndal *et al.* (1992)),

$$\mathrm{Var}(\hat{\tau}_{aas2E}) = M^2\left(1-\frac{m}{M}\right)\frac{1}{m}\frac{1}{M-1}\sum_{i=1}^{M}(\tau_i-\bar{\tau})^2 + \frac{M}{m}\sum_{i=1}^{M}\mathrm{Var}(\hat{\tau_i}^*), \tag{4.2}$$

em que $\mathrm{Var}(\hat{\tau_i}^*)$ é a variância do estimador da amostra aleatória simples para $\tau_i$, que é o total de $y$ dentro da unidade primária $i$. Consideramos uma amostra de tamanho $m$ na primeira etapa e tamanho $\frac{E(\nu)}{m}$ para a segunda etapa, onde $N_i$ é o número de unidades secundárias na unidade primária $i$. Portanto, a eficiência do estimador $\hat{\tau}_{HT2E}$ é então dada pela relação $\mathrm{ef}(\hat{\tau}_{HT2E}) = \frac{\mathrm{Var}(\hat{\tau}_{aas2E})}{\mathrm{Var}(\hat{\tau}_{HT2E})}$ e para o estimador $\hat{\tau}_{HH2E}$ temos $\mathrm{ef}(\hat{\tau}_{HH2E}) = \frac{\mathrm{Var}(\hat{\tau}_{aas2E})}{\mathrm{Var}(\hat{\tau}_{HH2E})}$. Devido a algumas dificuldades de simulação, incluímos nos resultados os cálculos teóricos. Detalhamos esse problema no Apêndice A.

| | | **Condição C = 0** | | | **Condição C = 5** | | | **Condição C = 10** | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **m** | **$n_1$** | **E($\nu$)** | **ef($\hat{\tau}_{HT}$)** | **ef($\hat{\tau}_{HH}$)** | **E($\nu$)** | **ef($\hat{\tau}_{HT}$)** | **ef($\hat{\tau}_{HH}$)** | **E($\nu$)** | **ef($\hat{\tau}_{HT}$)** | **ef($\hat{\tau}_{HH}$)** |
| **2** | **2** | 7.2 | 1.61 | 1.55 | 5.6 | 1.92 | 1.86 | 5.3 | 1.17 | 1.16 |
| | **5** | 15.7 | 1.68 | 1.48 | 13.0 | 1.94 | 1.72 | 12.4 | 1.21 | 1.16 |
| | **8** | 22.4 | 1.68 | 1.42 | 19.4 | 1.88 | 1.59 | 18.9 | 1.24 | 1.16 |
| **4** | **2** | 14.4 | 1.65 | 1.58 | 11.2 | 1.97 | 1.90 | 10.6 | 1.17 | 1.16 |
| | **5** | 31.4 | 1.79 | 1.55 | 25.9 | 2.07 | 1.81 | 24.9 | 1.23 | 1.17 |
| | **8** | 44.8 | 1.86 | 1.50 | 38.9 | 2.09 | 1.70 | 37.8 | 1.27 | 1.18 |
| **8** | **2** | 28.8 | 1.74 | 1.65 | 22.4 | 2.10 | 2.01 | 21.2 | 1.19 | 1.17 |
| | **5** | 62.7 | 2.16 | 1.75 | 51.8 | 2.51 | 2.08 | 49.8 | 1.27 | 1.21 |
| | **8** | 89.5 | 2.75 | 1.86 | 77.7 | 3.09 | 2.15 | 75.6 | 1.37 | 1.24 |

Tabela 4.6: *Eficiências relativas dos estimadores $\hat{\tau}_{HT}$ e $\hat{\tau}_{HH}$ - população original*

Podemos observar pela Tabela 4.6 que os estimadores da amostragem por conglomerados adaptativos são eficientes para a população do marreco da asa azul. Da mesma forma que na amostragem por conglomerados adaptativos em um estágio, a amostragem por conglomerados adaptativos em dois estágios apresenta as maiores eficiências.

| | | **Condição C = 0** | | | **Condição C = 5** | | | **Condição C = 10** | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **m** | **$n_1$** | **E($\nu$)** | **ef($\hat{\tau}_{HT}$)** | **ef($\hat{\tau}_{HH}$)** | **E($\nu$)** | **ef($\hat{\tau}_{HT}$)** | **ef($\hat{\tau}_{HH}$)** | **E($\nu$)** | **ef($\hat{\tau}_{HT}$)** | **ef($\hat{\tau}_{HH}$)** |
| **2** | **2** | 11.2 | 0.48 | 0.46 | 9.0 | 0.54 | 0.52 | 8.8 | 0.55 | 0.53 |
| | **5** | 22.5 | 0.61 | 0.52 | 19.3 | 0.63 | 0.56 | 19.0 | 0.64 | 0.58 |
| | **8** | 30.0 | 0.74 | 0.59 | 27.1 | 0.73 | 0.61 | 26.6 | 0.75 | 0.63 |
| **4** | **2** | 22.5 | 0.47 | 0.45 | 17.9 | 0.53 | 0.51 | 17.6 | 0.54 | 0.52 |
| | **5** | 45.1 | 0.58 | 0.50 | 38.7 | 0.61 | 0.54 | 38.0 | 0.62 | 0.56 |
| | **8** | 60.0 | 0.71 | 0.55 | 54.1 | 0.70 | 0.58 | 53.3 | 0.72 | 0.59 |
| **8** | **2** | 45.0 | 0.45 | 0.43 | 35.9 | 0.51 | 0.50 | 35.2 | 0.52 | 0.51 |
| | **5** | 90.1 | 0.52 | 0.44 | 77.3 | 0.56 | 0.49 | 75.9 | 0.58 | 0.51 |
| | **8** | 119.9 | 0.61 | 0.45 | 108.2 | 0.63 | 0.50 | 106.6 | 0.65 | 0.51 |

Tabela 4.7: *Eficiências relativas dos estimadores $\hat{\tau}_{HT}$ e $\hat{\tau}_{HH}$ - população artificial 1*

De acordo com a Tabela 4.7, os estimadores da amostragem por conglomerados adaptativos não se mostram eficientes para a população artificial 1, de maneira análoga à amostragem por conglomerados adaptativos em um estágio.

| | | **Condição C = 0** | | | **Condição C = 5** | | | **Condição C = 10** | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **m** | **$n_1$** | **E($\nu$)** | **ef($\hat{\tau}_{HT}$)** | **ef($\hat{\tau}_{HH}$)** | **E($\nu$)** | **ef($\hat{\tau}_{HT}$)** | **ef($\hat{\tau}_{HH}$)** | **E($\nu$)** | **ef($\hat{\tau}_{HT}$)** | **ef($\hat{\tau}_{HH}$)** |
| **2** | **2** | 5.5 | 0.74 | 0.74 | 4.7 | 0.86 | 0.86 | 4.5 | 0.90 | 0.90 |
| | **5** | 13.1 | 0.79 | 0.79 | 11.4 | 0.89 | 0.89 | 11.0 | 0.92 | 0.92 |
| | **8** | 20.1 | 0.83 | 0.83 | 17.9 | 0.91 | 0.91 | 17.4 | 0.93 | 0.93 |
| **4** | **2** | 10.9 | 0.73 | 0.73 | 9.3 | 0.86 | 0.86 | 8.9 | 0.90 | 0.90 |
| | **5** | 26.2 | 0.76 | 0.76 | 22.9 | 0.88 | 0.88 | 22.0 | 0.91 | 0.91 |
| | **8** | 40.2 | 0.80 | 0.80 | 35.8 | 0.89 | 0.89 | 34.7 | 0.92 | 0.92 |
| **8** | **2** | 21.9 | 0.71 | 0.71 | 18.7 | 0.84 | 0.84 | 17.8 | 0.89 | 0.89 |
| | **5** | 52.4 | 0.70 | 0.70 | 45.7 | 0.84 | 0.84 | 44.0 | 0.89 | 0.89 |
| | **8** | 80.5 | 0.70 | 0.70 | 71.7 | 0.84 | 0.84 | 69.4 | 0.88 | 0.88 |

Tabela 4.8: *Eficiências relativas dos estimadores $\hat{\tau}_{HT}$ e $\hat{\tau}_{HH}$ - população artificial 2*

Segundo a Tabela 4.8 as eficiências dos estimadores $\hat{\tau}_{HT}$ e $\hat{\tau}_{HH}$ para a população artificial 2 são menores do que 1 quando comparadas com a amostra aleatória em dois estágios, mas são maiores do que as encontradas na população artificial 1.

### Amostragem seqüencial em dois estágios

Para a amostragem seqüencial em dois estágios, calculamos as eficiências relativas utilizando a variância das 10.000 amostras e a variância teórica da amostragem aleatória em dois estágios.

| | | **Condição C = 0** | | | | | | **Condição C = 5** | | | | | | **Condição C = 10** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $\mathbf{n_{i2}} = 5$ | | $\mathbf{n_{i2}} = 10$ | | $\mathbf{n_{i2}} = 15$ | | $\mathbf{n_{i2}} = 5$ | | $\mathbf{n_{i2}} = 10$ | | $\mathbf{n_{i2}} = 15$ | | $\mathbf{n_{i2}} = 5$ | | $\mathbf{n_{i2}} = 10$ | | $\mathbf{n_{i2}} = 15$ | |
| **m** | $\mathbf{n_{i1}}$ | **E**($\nu$) | **ef**($\hat{\tau}$) | **E**($\nu$) | **ef**($\hat{\tau}$) | **E**($\nu$) | **ef**($\hat{\tau}$) | **E**($\nu$) | **ef**($\hat{\tau}$) | **E**($\nu$) | **ef**($\hat{\tau}$) | **E**($\nu$) | **ef**($\hat{\tau}$) | **E**($\nu$) | **ef**($\hat{\tau}$) | **E**($\nu$) | **ef**($\hat{\tau}$) | **E**($\nu$) | **ef**($\hat{\tau}$) |
| **2** | **2** | 6.0 | 1.16 | 8.0 | 1.26 | 10.1 | 1.45 | 5.1 | 1.17 | 6.2 | 1.59 | 7.4 | 1.57 | 5.0 | 1.07 | 5.9 | 1.30 | 6.8 | 1.55 |
| | **5** | 13.9 | 1.17 | 17.7 | 1.35 | 21.8 | 1.50 | 12.2 | 1.27 | 14.5 | 1.46 | 16.6 | 1.77 | 11.9 | 1.22 | 13.9 | 1.39 | 15.9 | 1.50 |
| | **8** | 21.0 | 1.23 | 25.8 | 1.39 | 30.7 | 1.51 | 18.8 | 1.36 | 21.6 | 1.54 | 24.5 | 1.82 | 18.6 | 1.25 | 21.4 | 1.42 | 24.1 | 1.57 |
| **4** | **2** | 12.0 | 1.07 | 16.0 | 1.27 | 20.0 | 1.44 | 10.2 | 1.22 | 12.5 | 1.48 | 14.5 | 1.70 | 9.9 | 1.10 | 11.8 | 1.26 | 13.7 | 1.42 |
| | **5** | 27.7 | 1.24 | 35.5 | 1.39 | 43.3 | 1.61 | 24.4 | 1.27 | 28.9 | 1.59 | 33.4 | 1.84 | 23.9 | 1.23 | 27.7 | 1.45 | 31.8 | 1.58 |
| | **8** | 41.9 | 1.27 | 51.7 | 1.46 | 61.8 | 1.63 | 37.7 | 1.32 | 43.3 | 1.72 | 49.1 | 2.04 | 37.3 | 1.25 | 42.6 | 1.48 | 48.1 | 1.74 |
| **8** | **2** | 23.9 | 1.13 | 32.1 | 1.33 | 40.1 | 1.52 | 20.4 | 1.22 | 24.8 | 1.45 | 29.2 | 1.83 | 19.7 | 1.11 | 23.6 | 1.25 | 27.3 | 1.48 |
| | **5** | 55.5 | 1.27 | 70.8 | 1.60 | 86.1 | 2.06 | 48.8 | 1.37 | 57.7 | 1.77 | 66.3 | 2.39 | 47.9 | 1.26 | 55.7 | 1.51 | 63.7 | 1.84 |
| | **8** | 83.8 | 1.33 | 103.7 | 1.84 | 123.6 | 3.17 | 75.4 | 1.45 | 86.8 | 2.21 | 98.5 | 3.91 | 74.7 | 1.35 | 85.3 | 1.81 | 95.9 | 2.62 |

Tabela 4.9: *Eficiências relativas do estimador $\hat{\tau}$ - população original*

Na Tabela 4.9, vemos que o estimador da amostragem seqüencial em dois estágios mostrou-se eficiente para a população de marrecos da asa azul. Para essa população, aumentando-se $n_{i2}$ temos maiores eficiências.

| | | **Condição C = 0** | | | | | | **Condição C = 5** | | | | | | **Condição C = 10** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $\mathbf{n_{i2}} = 5$ | | $\mathbf{n_{i2}} = 10$ | | $\mathbf{n_{i2}} = 15$ | | $\mathbf{n_{i2}} = 5$ | | $\mathbf{n_{i2}} = 10$ | | $\mathbf{n_{i2}} = 15$ | | $\mathbf{n_{i2}} = 5$ | | $\mathbf{n_{i2}} = 10$ | | $\mathbf{n_{i2}} = 15$ | |
| **m** | $\mathbf{n_{i1}}$ | **E**($\nu$) | **ef**($\hat{\tau}$) | **E**($\nu$) | **ef**($\hat{\tau}$) | **E**($\nu$) | **ef**($\hat{\tau}$) | **E**($\nu$) | **ef**($\hat{\tau}$) | **E**($\nu$) | **ef**($\hat{\tau}$) | **E**($\nu$) | **ef**($\hat{\tau}$) | **E**($\nu$) | **ef**($\hat{\tau}$) | **E**($\nu$) | **ef**($\hat{\tau}$) | **E**($\nu$) | **ef**($\hat{\tau}$) |
| **2** | **2** | 8.7 | 0.62 | 13.5 | 0.43 | 18.1 | 0.34 | 8.0 | 0.63 | 11.9 | 0.45 | 15.9 | 0.34 | 7.8 | 0.62 | 11.6 | 0.44 | 15.6 | 0.33 |
| | **5** | 17.4 | 0.62 | 26.8 | 0.43 | 36.4 | 0.32 | 15.9 | 0.62 | 23.7 | 0.44 | 31.7 | 0.33 | 17.1 | 0.83 | 24.2 | 0.64 | 31.2 | 0.49 |
| | **8** | 34.7 | 0.60 | 53.6 | 0.40 | 72.6 | 0.28 | 31.9 | 0.60 | 47.4 | 0.41 | 63.5 | 0.29 | 24.7 | 0.93 | 33.3 | 0.79 | 42.1 | 0.66 |
| **4** | **2** | 18.1 | 0.83 | 26.1 | 0.67 | 34.2 | 0.53 | 17.3 | 0.82 | 24.5 | 0.63 | 31.8 | 0.49 | 15.6 | 0.61 | 23.2 | 0.44 | 30.7 | 0.32 |
| | **5** | 36.1 | 0.83 | 52.1 | 0.63 | 68.1 | 0.49 | 34.6 | 0.81 | 49.0 | 0.59 | 63.8 | 0.45 | 34.2 | 0.79 | 48.5 | 0.60 | 62.5 | 0.44 |
| | **8** | 72.1 | 0.80 | 104.4 | 0.56 | 136.5 | 0.35 | 69.0 | 0.78 | 98.3 | 0.55 | 127.4 | 0.33 | 49.4 | 0.92 | 66.8 | 0.75 | 84.1 | 0.60 |
| **8** | **2** | 25.4 | 0.91 | 34.6 | 0.82 | 44.0 | 0.74 | 24.9 | 0.91 | 33.8 | 0.80 | 42.6 | 0.67 | 31.2 | 0.60 | 46.3 | 0.41 | 61.3 | 0.30 |
| | **5** | 50.7 | 0.94 | 69.4 | 0.82 | 87.9 | 0.69 | 49.7 | 0.92 | 67.5 | 0.78 | 85.1 | 0.60 | 68.3 | 0.75 | 96.5 | 0.53 | 125.3 | 0.34 |
| | **8** | 101.4 | 0.91 | 138.7 | 0.72 | 175.9 | 0.37 | 99.4 | 0.90 | 134.8 | 0.66 | 170.4 | 0.34 | 98.8 | 0.88 | 133.7 | 0.65 | 168.4 | 0.34 |

Tabela 4.10: *Eficiências relativas do estimador $\hat{\tau}$ - população artificial 1*

Na Tabela 4.10, para a população artificial 1 o comportamento da amostragem seqüencial em duas etapas é muito distinto daquele comportamento da população original, pois quanto menor $n_{i2}$ maior eficiência.

| | | **Condição C = 0** | | | | | | **Condição C = 5** | | | | | | **Condição C = 10** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **$n_{i2}$** = 5 | | **$n_{i2}$** = 10 | | **$n_{i2}$** = 15 | | **$n_{i2}$** = 5 | | **$n_{i2}$** = 10 | | **$n_{i2}$** = 15 | | **$n_{i2}$** = 5 | | **$n_{i2}$** = 10 | | **$n_{i2}$** = 15 | |
| **m** | **$n_{i1}$** | **E($\nu$)** | **ef($\hat{\tau}$)** | **E($\nu$)** | **ef($\hat{\tau}$)** | **E($\nu$)** | **ef($\hat{\tau}$)** | **E($\nu$)** | **ef($\hat{\tau}$)** | **E($\nu$)** | **ef($\hat{\tau}$)** | **E($\nu$)** | **ef($\hat{\tau}$)** | **E($\nu$)** | **ef($\hat{\tau}$)** | **E($\nu$)** | **ef($\hat{\tau}$)** | **E($\nu$)** | **ef($\hat{\tau}$)** |
| **2** | **2** | 6.1 | 0.86 | 8.2 | 0.80 | 10.2 | 0.79 | 5.2 | 1.00 | 6.3 | 1.04 | 7.5 | 0.92 | 5.0 | 0.94 | 6.0 | 0.85 | 7.0 | 0.80 |
| | **5** | 14.4 | 0.96 | 18.9 | 0.91 | 23.4 | 0.88 | 12.6 | 1.07 | 15.5 | 1.06 | 18.2 | 1.02 | 12.3 | 0.98 | 14.8 | 0.92 | 17.0 | 0.89 |
| | **8** | 22.1 | 0.98 | 28.2 | 0.99 | 34.5 | 0.92 | 20.0 | 1.05 | 24.0 | 1.06 | 27.9 | 1.10 | 19.5 | 0.98 | 23.1 | 0.99 | 26.7 | 0.95 |
| **4** | **2** | 12.2 | 0.81 | 16.2 | 0.81 | 20.3 | 0.71 | 10.3 | 1.04 | 12.7 | 0.98 | 15.1 | 0.96 | 10.0 | 0.89 | 11.9 | 0.90 | 13.9 | 0.82 |
| | **5** | 29.0 | 0.95 | 38.0 | 0.87 | 46.9 | 0.83 | 25.4 | 1.07 | 30.7 | 1.09 | 36.0 | 1.07 | 24.7 | 1.00 | 29.4 | 0.93 | 33.9 | 0.90 |
| | **8** | 44.3 | 1.02 | 56.6 | 0.96 | 68.7 | 0.93 | 40.0 | 1.09 | 47.9 | 1.15 | 56.0 | 1.14 | 39.1 | 1.03 | 46.4 | 0.99 | 53.3 | 0.97 |
| **8** | **2** | 24.4 | 0.80 | 32.9 | 0.72 | 41.0 | 0.65 | 20.6 | 1.01 | 25.3 | 0.97 | 30.0 | 0.95 | 19.9 | 0.89 | 23.7 | 0.88 | 28.0 | 0.76 |
| | **5** | 57.7 | 0.93 | 75.7 | 0.82 | 93.7 | 0.72 | 50.8 | 1.11 | 61.2 | 1.14 | 72.4 | 1.12 | 49.5 | 0.97 | 58.9 | 0.91 | 68.0 | 0.84 |
| | **8** | 88.5 | 0.99 | 112.8 | 0.93 | 137.7 | 0.81 | 79.9 | 1.17 | 95.7 | 1.30 | 111.7 | 1.40 | 78.3 | 1.04 | 92.4 | 1.01 | 107.0 | 0.93 |

Tabela 4.11: *Eficiências relativas do estimador $\hat{\tau}$ - população artificial 2*

Para a população artificial 2, conforme apresentado na Tabela 4.11, não há ganhos de eficiência com relação à amostragem aleatória em dois estágios por utilizar o método seqüencial, mas por outro lado as eficiências aparentam ser maiores do que as apresentadas pela população artificial 1.

### Comparação: amostragem por conglomerados adaptativos e amostragem seqüencial

A Tabela 4.12 dispõe os resultados tanto da amostragem por conlgomerados adaptativos quanto da amostragem seqüencial em duas etapas. Os valores de $m$ e $C$ foram utilizados da mesma forma nos dois métodos. O parâmetro $n_1$ corresponde ao método da amostragem por conglomerados adaptativos em duas etapas assim como o valor esperado de amostra $E(\nu)$ e a eficiência do estimador Horvitz-Thompson ef$(\hat{\tau})_{HT}$. Variamos os parâmetros $n_{i1}$ e $n_{i2}$ para que o tamanho amostral esperado da amostragem seqüencial em dois estágios $E(\nu^*)$ se aproximasse ao máximo do valor esperado $E(\nu)$.

Pelos resultados da Tabela 4.12, vemos que o estimador do tipo Horvitz-Thompson para a amostragem por conglomerados adaptativos é mais eficiente em 19 dentre 27 casos. Notamos

| | | Condição C = 0 | | | | | | Condição C = 5 | | | | | | Condição C = 10 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **m** | **$n_1$** | **E($\nu$)** | **$n_{i1}$** | **$n_{i2}$** | **E($\nu^*$)** | **ef($\hat{\tau}_{HT}$)** | **ef($\hat{\tau}$)** | **E($\nu$)** | **$n_{i1}$** | **$n_{i2}$** | **E($\nu^*$)** | **ef($\hat{\tau}_{HT}$)** | **ef($\hat{\tau}$)** | **E($\nu$)** | **$n_{i1}$** | **$n_{i2}$** | **E($\nu^*$)** | **ef($\hat{\tau}_{HT}$)** | **ef($\hat{\tau}$)** |
| **2** | **2** | 7.2 | 2 | 8 | 7.1 | 1.61 | 1.23 | 5.6 | 2 | 7 | 5.6 | 1.92 | 1.25 | 5.3 | 2 | 7 | 5.3 | 1.17 | 1.17 |
| | **5** | 15.7 | 7 | 2 | 15.9 | 1.68 | 1.06 | 13.0 | 4 | 13 | 12.9 | 1.94 | 1.62 | 12.4 | 4 | 13 | 12.4 | 1.21 | 1.38 |
| | **8** | 22.4 | 6 | 12 | 22.4 | 1.68 | 1.38 | 19.4 | 8 | 6 | 19.4 | 1.88 | 1.33 | 18.9 | 7 | 10 | 19.0 | 1.24 | 1.40 |
| **4** | **2** | 14.4 | 2 | 8 | 14.4 | 1.65 | 1.18 | 11.2 | 2 | 7 | 11.1 | 1.97 | 1.30 | 10.6 | 2 | 7 | 10.6 | 1.17 | 1.22 |
| | **5** | 31.4 | 7 | 2 | 31.7 | 1.79 | 1.07 | 25.9 | 4 | 13 | 25.9 | 2.07 | 1.75 | 24.9 | 4 | 13 | 24.8 | 1.23 | 1.40 |
| | **8** | 44.8 | 7 | 9 | 44.7 | 1.86 | 1.38 | 38.9 | 8 | 6 | 38.9 | 2.09 | 1.42 | 37.8 | 7 | 10 | 38.0 | 1.27 | 1.43 |
| **8** | **2** | 28.8 | 2 | 8 | 28.9 | 1.74 | 1.24 | 22.4 | 2 | 7 | 22.2 | 2.10 | 1.30 | 21.2 | 2 | 7 | 21.3 | 1.19 | 1.14 |
| | **5** | 62.7 | 7 | 2 | 63.5 | 2.16 | 1.12 | 51.8 | 4 | 13 | 51.6 | 2.51 | 1.93 | 49.8 | 5 | 6 | 49.5 | 1.27 | 1.29 |
| | **8** | 89.5 | 7 | 9 | 89.5 | 2.75 | 1.66 | 77.7 | 8 | 6 | 77.7 | 3.09 | 1.59 | 75.6 | 7 | 10 | 75.8 | 1.37 | 1.69 |

Tabela 4.12: *Eficiências relativas dos estimadores $\hat{\tau}$ e $\hat{\tau}_{HT}$*

que os casos em que sua eficiência é menor se concentram na situação em que $C = 10$.

## 4.3 Discussão

Conforme podemos observar pelos resultados, resumidos na Tabela 4.13, as técnicas propostas são eficientes na situação específica de sua utilização, ou seja, para populações raras e agrupadas. As eficiências variam de acordo com o grau de agrupamento e raridade das populações e com a escolha da condição de adição de novos elementos uma vez que esta determina as redes. Pudemos também observar que no caso do método de amostragem seqüencial por conglomerados adaptativos, além de possuir um estimador eficiente nos casos em que a população é rara e agrupada ela não revelou perdas de eficiência para a população rara e não agrupada simulada. Esses resultados se somam ao observado por Salehi & Smith (2005), para a população *Lampsili cariosa*. A população artificial 1 revelou uma situação em que a amostragem seqüencial em duas etapas não é eficiente nos casos de uma população não rara e agrupada, ao contrário da população *Eliptio complanata* (Salehi & Smith (2005)), indicando que os resultados podem variar conforme a população e os parâmetros.

Analisando as comparações realizadas entre a amostragem por conglomerados adaptativos em dois estágios e seqüencial em dois estágios, apresentadas na Tabela 4.12, observamos que na proposta inicial de população rara e agrupada, apesar da melhor eficiência do primeiro método na maior parte dos casos, quando a variabilidade intra-rede descresce a amostragem

| | População | | |
|---|---|---|---|
| | **Original** | **Artificial 1** | **Artificial 2** |
| **Técnica** | **Rara e Agrupada** | **Não Rara e Menos Agrupada** | **Rara e Não Agrupada** |
| **Conglomerados Adaptativos** | Eficiente. Melhor C=5. | Não eficiente. Melhor C=10. | Não eficiente. Melhor C=10. |
| **Conglomerados Adaptativos 2 estagios** | Eficiente. Melhor C=5. | Não eficiente. Melhor C=10. | Não eficiente. Melhor C=10. |
| **Seqüencial em duas Etapas** | Eficiente. Melhor C=5. | Não eficiente. | Mesma eficiência. Melhor C=5. |

Tabela 4.13: *Quadro resumo - análise dos resultados das simulações individuais*

seqüencial tem melhor desempenho, o que indica que todas as comparações dependem das características da população em estudo. Nos casos de uma população rara mas com variabilidade intra-rede baixa, a amostragem seqüencial tem maiores eficiências, e o oposto acontece quando a variabilidade intra-rede é alta.

# Capítulo 5

# Conclusões

Os métodos para populações raras são úteis em muitas situações práticas. A maioria deles é resultado de uma série de suposições e algumas estão mais sujeitas a viés tendo em vista que não são metodologias probabilísticas. A amostragem por conlgomerados adaptativos em um e dois estágios e a amostragem seqüencial em duas etapas são alternativas eficientes para a estimação das populações raras e agrupadas e são probabilísticas pois existe a aleatorização e as probabilidades são conhecidas ao final do processo.

Como pudemos observar por nossos estudos de simulações, enquanto a amostragem por conglomerados adaptativos é eficiente apenas no caso em que a população possui ambas as características, rara e agrupada, a amostragem seqüencial em duas etapas não perde eficiência no caso de populações apenas raras. Por outro lado, na maioria dos casos de populações raras e agrupadas a eficiência da amostragem por conglomerados adaptativos foi maior, variando de acordo com a escolha da condição $C$.

Na prática, essas técnicas revelam-se úteis quando já se detém algum conhecimento prévio das populações, sendo necessário no mínimo uma percepção, por exemplo, se a população é rara como uma fração populacional ou fração de área coberta de $0,1\%$ ou grande como $20\%$. Ainda mais complexa é a percepção do grau de agrupamento da população em estudo.

Concluímos que se por um lado a amostragem por conglomerados adaptativos é mais eficiente nos casos em que a população é rara e agrupada, na maioria dos casos existe a dificuldade na determinação desses fatores, o que torna a técnica de amostragem seqüencial em duas etapas uma opção mais aplicável. Além disso, como já exploramos no Capítulo 3

há vantagens em sua implementação na prática como por exemplo não necessita de definição de vizinhança e possui a amostra final restrita a um valor conhecido.

Assim como Salehi & Smith (2005), recomendamos que, para populações em que o grau de agrupamento das unidades é desconhecido, escolham-se valores moderados e grandes para $n_{i1}$ e pequenos para $n_{i2}$ pois dessa maneira, se a população não for agrupada a amostra estará distribuída de maneira mais esparsa; se o for, então a eficiência será similiar a um desenho de duas etapas convencional. Já nos casos em que sabemos que a população é agrupada, recomendamos valores pequenos de $n_{i1}$ e maiores valores de $n_{i2}$ para aumentar as probabilidades de incorporação das unidades raras na amostra. A condição $C$ deve ser escolhida de acordo com a percepção do pequisador com respeito ao agrupamento das unidades. Por exemplo, no caso da população do marreco da asa azul, a escolha da condição $C = 1000$ levaria a eleger apenas 2 unidades para a amostra adicional, o que faria com que a eficiência dos estimadores ficasse prejudicada.

# Apêndice A

# Simulação da Amostragem por Conglomerados Adaptativos em dois Estágios

Para a geração da amostragem por conglomerados adaptativos em dois estágios, identificamos problemas no procedimento de simulação, muito embora as simulações da amostragem por conglomerados adaptativos tenham sido validadas com as referências e os cálculos das expressões teóricas segundo os exemplos. Na Tabela A.1, mostramos uma comparação entre os valores teóricos e simulados para a amostragem por conglomerados adaptativos em 1 estágio.

| | **Simulado** | | **Teórico** | |
|---|---|---|---|---|
| **n1** | **ef**($\hat{\tau}_{\mathbf{HT}}$) | **ef**($\hat{\tau}_{\mathbf{HH}}$) | **ef**($\hat{\tau}_{\mathbf{HT}}$) | **ef**($\hat{\tau}_{\mathbf{HH}}$) |
| **4** | 1.368 | 1.337 | 1.373 | 1.337 |
| **8** | 1.422 | 1.334 | 1.424 | 1.334 |
| **10** | 1.468 | 1.358 | 1.456 | 1.358 |
| **16** | 1.528 | 1.353 | 1.547 | 1.353 |

Tabela A.1: *Eficiências simuladas versus teóricas - população original* ($C = 10$)

Como vemos pela Tabela A.1, as diferenças são pequenas entre as eficiências simuladas e as teóricas. Embora a amostragem por conlgomerados adaptativos tenha apresentado resultados coerentes com os cálculos teóricos e as referências, para a simulação por conglomerados em dois estágios apresentou diferenças elevadas. Pela Tabela A.2 podemos verificar que as

| | | Teórico | | | Simulado | | |
|---|---|---|---|---|---|---|---|
| **m** | **$n_1$** | **E($\nu$)** | **ef($\hat{\tau}_{HT}$)** | **ef($\hat{\tau}_{HH}$)** | **E($\nu$)** | **ef($\hat{\tau}_{HT}$)** | **ef($\hat{\tau}_{HH}$)** |
| **4** | **2** | 10.6 | 1.17 | 1.16 | 11.2 | 2.18 | 1.99 |
| | **5** | 24.9 | 1.23 | 1.17 | 25.9 | 2.09 | 1.68 |
| | **8** | 37.8 | 1.27 | 1.18 | 39.1 | 1.94 | 1.67 |
| **8** | **2** | 21.2 | 1.19 | 1.17 | 22.1 | 2.18 | 2.09 |
| | **5** | 49.8 | 1.27 | 1.21 | 52.1 | 2.48 | 2.03 |
| | **8** | 75.6 | 1.37 | 1.24 | 77.9 | 3.18 | 2.35 |

Tabela A.2: *Eficiências simuladas versus teóricas - amostragem por conglomerados adaptativos em dois estágios*

diferenças são altas. Essas diferenças podem se dever a problemas de aleatorização no R de tal forma a gerar algum tipo de redução na variância do estimador. O programa está disponível no apêndice B.

# Apêndice B

# Programas

## B.1 Amostragem por conglomerados adaptativos

```
############################
############################
############################
##                        ##
##  DADOS DO MARRECO AZUL  ##
##                        ##
############################
############################
############################

#Limpa Tudo
rm(list = ls())
#Exemplo dos Marrecos da Asa Azul
teal<-array(0,c(20,10))
teal[1,9]<-60
teal[2,5]<-1
teal[2,8]<-122
teal[2,9]<-114
teal[2,10]<-3
teal[3,5]<-7144
teal[3,6]<-6399
teal[3,8]<-14
teal[4,4]<-103
teal[4,5]<-150
teal[4,6]<-6
teal[5,4]<-10
teal[7,6]<-2
teal[7,10]<-2
teal[9,5]<-3
teal[11,3]<-12
teal[12,3]<-2
teal[12,6]<-2
teal[13,3]<-4
teal[14,1]<-5
teal[14,3]<-20
```

```
teal[15,2]<-3



#tamanho total da população
N<-nrow(teal)*ncol(teal)



#até aqui insere as informações




################################################
################################################
################################################
##                                            ##
##  AMOSTRAGEM POR CONGLOMERADOS ADAPTATIVOS  ##
##                                            ##
################################################
################################################
################################################

ACS<-function(popul,n1,c)
{

#Cria a Matriz dos Indices da Matriz de Dados
indices<-array(1:(nrow(popul)*ncol(popul)),c(nrow(popul),ncol(popul)))
#tamanho total (número de unidades primárias)
d<-(nrow(popul)*ncol(popul))

#soma dos yk
soma<-integer(0)

#divisao
divisao<-integer(0)

#definição das prob alpha
alpha<-integer(0)

#Retira uma amostra desses indices de tamanho n1
amostrat<-sample(indices,n1)


#variavel incial
empilha<-array(NA,c(length(amostrat),d))


for (l in 1:length(amostrat))
{
amostra<-amostrat[l]

#determina a posicao do elemento da amostra

referencia<-amostra

#cria os vizinhos
```

```
#PASSO Número 0

vizinho<-function(x)
{
#determina a posição do elemento da referencia
lca<-array(0,c(length(x),2))
for (i in 1:length(x))
{lca[i,]<-which(indices==x[i], arr.ind=TRUE)}
linha<-lca[1,1]
coluna<-lca[1,2]
#zera os vizinhos
cima<-integer(0)
baixo<-integer(0)
esquerda<-integer(0)
direita<-integer(0)
#cria vizinhos
if (popul[x]>c)
{
if (linha!=1) {cima<-indices[lca[1,1]-1,lca[1,2]]}
if (linha!=nrow(popul)) {baixo<-indices[lca[1,1]+1,lca[1,2]]}
if (coluna!=1) {esquerda<-indices[lca[1,1],lca[1,2]-1]}
if (coluna!=ncol(popul)) {direita<-indices[lca[1,1],lca[1,2]+1]}
}
b<-integer(0)
b<-cbind(cima,baixo,esquerda,direita)
b
}
b<-vizinho(referencia)


#cria um vetor contendo os vetores que ja visitamos ate aqui
dados<-array(NA,c(d,1))

#cria um vetor indicando se os vizinhos daquele elemento já estão contemplados
explore<-array(0,c(d,1))

#coloca o elemento amostrado na matriz de dados
dados[referencia]<-popul[referencia]

#cria quem sera visitado no primeiro passo
visita<-amostra[which(popul[amostra]>c)]

#repeticao

while (length(visita)>0)
{
referencia<-min(visita)
b<-vizinho(referencia)
for (i in 1:d)
{if (length(b[which(b==i)])>0) dados[i]<-popul[i]}
#indica que os vizinhos da referencia ja foram explorados
explore[referencia]<-1
#elementos a serem explorados
visita<-which(dados>c)
explorados<-which(explore==1)

comuns<-array(0,c(length(visita),length(explorados)))
for (i in 1:length(visita))
for (j in 1:length(explorados))
{if (visita[i]==explorados[j]) comuns[i,j]<-i}
```

```
visita<-visita[-comuns]
}
empilha[l,]<-dados
}
#conglomerados distintos
conglomerados<-unique(empilha)
#redes distintas (ainda que não tenham nada)
redes<-array(NA,c(nrow(conglomerados),ncol(conglomerados)))

#redes de tamanho1

for (i in 1:nrow(conglomerados)){
for (j in 1:ncol(conglomerados)){
if ((is.na(conglomerados[i,j])==T|(conglomerados[i,j]<=c)))
redes[i,j]<-NA
else redes[i,j]<-conglomerados[i,j]

#mesmo que nao assuma a condicao a unidade da selecao incial
#tem de ser incorporada no estimador

if (length(which(is.na(conglomerados[i,])==F))==1)
redes[i,j]<-conglomerados[i,j]
}
}


################################################
#conglomerados não necessariamente distintos #
################################################
conglomerados<-empilha

#redes não distintas (ainda que não tenham nada)
redesr<-array(NA,c(nrow(conglomerados),ncol(conglomerados)))

#redes de tamanho1

for (i in 1:nrow(conglomerados)){
for (j in 1:ncol(conglomerados)){
if ((is.na(conglomerados[i,j])==T|(conglomerados[i,j]<=c)))
redesr[i,j]<-NA
else redesr[i,j]<-conglomerados[i,j]

#mesmo que nao assuma a condicao a unidade da selecao incial
#tem de ser incorporada no estimador

if (length(which(is.na(conglomerados[i,])==F))==1)
redesr[i,j]<-conglomerados[i,j]
}
}


#truque para tamanho amostral total
x<-integer(0)
for (i in 1:ncol(conglomerados))
{
x[i]<-1-min(is.na(conglomerados[,i]))
}

amtotal<-which(x==1)

w<-integer(0)
```

```
for (k in 1:nrow(redes))
{
rede<-redes[k,which(is.na(redes[k,])==F)]
#conta o tamanho da rede o minimo é um
lr<-max(1,length(rede))

#calcula a soma de yk
soma[k]<-sum(rede)

#calcula wk para HH
w[k]<-mean(rede)

#cálculo das probabilidades alpha
alpha[k]<-
 1-(choose((d-lr),length(amostrat))/choose(d,length(amostrat)))

divisao[k]<-soma[k]/alpha[k]
}

for (k in 1:nrow(redesr))
{
reder<-redesr[k,which(is.na(redesr[k,])==F)]

#calcula wk para HH
w[k]<-mean(reder)

}


HT<-sum(divisao)
HH<-N*mean(w)
size<-length(amtotal)
result<-cbind(HT,size,HH)
result
}


####################################
####################################
####################################
###                              ###
### PARTE REPETITIVA DA SIMULAÇÃO ###
###     ###
####################################
####################################
####################################



#Declaro os vetores que serão construidos a diante
HT<-integer(0)
tam<-integer(0)
HH<-integer(0)
DES<-integer(0)
tamDES<-integer(0)
aas<-integer(0)
junto<-0


for (n1 in c(4,8,10,16,20,32,40,64))
```

```
for (c in c(0,5,10))
{
{


for (r in 1:10000)
{
############################################################
### Parte que simula os estimadores referentes ao       ###
### Esquema de amostragem por conglomerados adaptativos ###
############################################################

acsobs<-ACS(teal,n1,c) #vetor com o valor dos estimadores e do tamanho amostral
HT[r]<-acsobs[1,1]     #valor estimado pelo Horvitz-Thompson
tam[r]<-acsobs[1,2]    #tamanho de amostra retirado
HH[r]<-acsobs[1,3]     #valor estimado pelo Hansen-Hurwitz
}

#####################################################
###   cálculo das Estatísticas dos Estimadores    ###
#####################################################


#Amostragem por conglomerados adaptativos

EHT<-mean(HT) #Média do estimador Horvitz-Thompson
VHT<-var(HT) #Variância do estimador Horvitz-Thompson

EHH<-mean(HH) #Média do estimador Hansen-Hurwitz
VHH<-var(HH) #Variância do estimador Hansen-Hurwitz

Eeta<-mean(tam) #tamanho amostral médio
estatisticas<-c(n1,c,EHT,VHT,EHH,VHH,Eeta)
junto<-rbind(junto,estatisticas)

}
}


write.table(junto,"c:\\acs_original.txt",sep="\t")
```

## B.2 Amostragem por conglomerados adaptativos em dois estágios

```
##############################
##############################
##############################
##                          ##
##  DADOS DO MARRECO AZUL   ##
##                          ##
##############################
##############################
```

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

#Limpa Tudo
rm(list = ls())
#Exemplo dos Marrecos da Asa Azul
teal<-array(0,c(20,10))
teal[1,9]<-60
teal[2,5]<-1
teal[2,8]<-122
teal[2,9]<-114
teal[2,10]<-3
teal[3,5]<-7144
teal[3,6]<-6399
teal[3,8]<-14
teal[4,4]<-103
teal[4,5]<-150
teal[4,6]<-6
teal[5,4]<-10
teal[7,6]<-2
teal[7,10]<-2
teal[9,5]<-3
teal[11,3]<-12
teal[12,3]<-2
teal[12,6]<-2
teal[13,3]<-4
teal[14,1]<-5
teal[14,3]<-20
teal[15,2]<-3

#tamanho total da população
N<-nrow(teal)*ncol(teal)



########################################
########################################
########################################
##                                    ##
##    CRIAÇÃO DAS UNIDADES PRIMÁRIAS   ##
##    ##
########################################
########################################
########################################

#cria uma variável chamada grupo
grupo<-array(NA,c(200,1))
dados<-cbind(which(teal>=0,arr.ind=T),teal[which(teal>=0,arr.ind=T)])



for (i in 1:nrow(dados))
{
if (dados[i,1]>=1 && dados[i,1]<=5 && dados[i,2]>=1 && dados[i,2]<=5) grupo[i]<-1
if (dados[i,1]>=6 && dados[i,1]<=10 && dados[i,2]>=1 && dados[i,2]<=5) grupo[i]<-2
if (dados[i,1]>=11 && dados[i,1]<=15 && dados[i,2]>=1 && dados[i,2]<=5) grupo[i]<-3
if (dados[i,1]>=16 && dados[i,1]<=25 && dados[i,2]>=1 && dados[i,2]<=5) grupo[i]<-4
if (dados[i,1]>=1 && dados[i,1]<=5 && dados[i,2]>=6 && dados[i,2]<=10) grupo[i]<-5
if (dados[i,1]>=6 && dados[i,1]<=10 && dados[i,2]>=6 && dados[i,2]<=10) grupo[i]<-6
if (dados[i,1]>=11 && dados[i,1]<=15 && dados[i,2]>=6 && dados[i,2]<=10) grupo[i]<-7
if (dados[i,1]>=16 && dados[i,1]<=25 && dados[i,2]>=6 && dados[i,2]<=10) grupo[i]<-8
}
```

```
#coloca ela em uma matiz dados2 junto com as outras informações
dados2<-cbind(dados,grupo)

#até aqui insere as informações

#o input do programa é a tabela com o formato da dados2

M<-8 #Número de unidades primárias na população


################################################
################################################
################################################
##                                            ##
##  AMOSTRAGEM POR CONGLOMERADOS ADAPTATIVOS  ##
##                                            ##
################################################
################################################
################################################

ACS<-function(popul,n1,c)
{

#tamanho em unidades
N<-nrow(popul)*ncol(popul)

#Cria a Matriz dos Indices da Matriz de Dados
indices<-array(1:(nrow(popul)*ncol(popul)),c(nrow(popul),ncol(popul)))
#tamanho total (número de unidades primárias)
d<-(nrow(popul)*ncol(popul))

#soma dos yk
soma<-integer(0)

#divisao
divisao<-integer(0)

#definição das prob alpha
alpha<-integer(0)

#Retira uma amostra desses indices de tamanho n1
amostrat<-sample(indices,n1)


#variavel incial
empilha<-array(NA,c(length(amostrat),d))


for (l in 1:length(amostrat))
{
amostra<-amostrat[l]

#determina a posicao do elemento da amostra

referencia<-amostra

#cria os vizinhos

#PASSO Número 0
```

```
vizinho<-function(x)
{
#determina a posição do elemento da referencia
lca<-array(0,c(length(x),2))
for (i in 1:length(x))
{lca[i,]<-which(indices==x[i], arr.ind=TRUE)}
linha<-lca[1,1]
coluna<-lca[1,2]
#zera os vizinhos
cima<-integer(0)
baixo<-integer(0)
esquerda<-integer(0)
direita<-integer(0)
#cria vizinhos
if (popul[x]>c)
{
if (linha!=1) {cima<-indices[lca[1,1]-1,lca[1,2]]}
if (linha!=nrow(popul)) {baixo<-indices[lca[1,1]+1,lca[1,2]]}
if (coluna!=1) {esquerda<-indices[lca[1,1],lca[1,2]-1]}
if (coluna!=ncol(popul)) {direita<-indices[lca[1,1],lca[1,2]+1]}
}
b<-integer(0)
b<-cbind(cima,baixo,esquerda,direita)
b
}
b<-vizinho(referencia)


#cria um vetor contendo os vetores que ja visitamos ate aqui
dados<-array(NA,c(d,1))

#cria um vetor indicando se os vizinhos daquele elemento já estão contemplados
explore<-array(0,c(d,1))

#coloca o elemento amostrado na matriz de dados
dados[referencia]<-popul[referencia]

#cria quem sera visitado no primeiro passo
visita<-amostra[which(popul[amostra]>c)]

#repeticao

while (length(visita)>0)
{
referencia<-min(visita)
b<-vizinho(referencia)
for (i in 1:d)
{if (length(b[which(b==i)])>0) dados[i]<-popul[i]}
#indica que os vizinhos da referencia ja foram explorados
explore[referencia]<-1
#elementos a serem explorados
visita<-which(dados>c)
explorados<-which(explore==1)

comuns<-array(0,c(length(visita),length(explorados)))
for (i in 1:length(visita))
for (j in 1:length(explorados))
{if (visita[i]==explorados[j]) comuns[i,j]<-i}
visita<-visita[-comuns]
}
```

```
empilha[l,]<-dados
}
#conglomerados distintos
conglomerados<-unique(empilha)
#redes distintas (ainda que não tenham nada)
redes<-array(NA,c(nrow(conglomerados),ncol(conglomerados)))

#redes de tamanho1

for (i in 1:nrow(conglomerados)){
for (j in 1:ncol(conglomerados)){
if ((is.na(conglomerados[i,j])==T|(conglomerados[i,j]<=c)))
redes[i,j]<-NA
else redes[i,j]<-conglomerados[i,j]

#mesmo que nao assuma a condicao a unidade da selecao incial
#tem de ser incorporada no estimador

if (length(which(is.na(conglomerados[i,])==F))==1)
redes[i,j]<-conglomerados[i,j]
}
}

################################################
#conglomerados não necessariamente distintos #
################################################
conglomerados<-empilha

#redes não distintas (ainda que não tenham nada)
redesr<-array(NA,c(nrow(conglomerados),ncol(conglomerados)))

#redes de tamanho1

for (i in 1:nrow(conglomerados)){
for (j in 1:ncol(conglomerados)){
if ((is.na(conglomerados[i,j])==T|(conglomerados[i,j]<=c)))
redesr[i,j]<-NA
else redesr[i,j]<-conglomerados[i,j]

#mesmo que nao assuma a condicao a unidade da selecao incial
#tem de ser incorporada no estimador

if (length(which(is.na(conglomerados[i,])==F))==1)
redesr[i,j]<-conglomerados[i,j]
}
}



#truque para tamanho amostral total
x<-integer(0)
for (i in 1:ncol(conglomerados))
{
x[i]<-1-min(is.na(conglomerados[,i]))
}

amtotal<-which(x==1)

w<-integer(0)

for (k in 1:nrow(redes))
```

```
{
rede<-redes[k,which(is.na(redes[k,])==F)]
#conta o tamanho da rede o minimo é um
lr<-max(1,length(rede))

#calcula a soma de yk
soma[k]<-sum(rede)

#cálculo das probabilidades alpha
alpha[k]<-
 1-(choose((d-lr),length(amostrat))/choose(d,length(amostrat)))

divisao[k]<-soma[k]/alpha[k]
}

for (k in 1:nrow(redesr))
{
reder<-redesr[k,which(is.na(redesr[k,])==F)]

#calcula wk para HH
w[k]<-mean(reder)

}


HT<-sum(divisao)
HH<-N*mean(w)
size<-length(amtotal)
result<-cbind(HT,size,HH)
result
}


################################################
################################################
################################################
##                                    ##
## AMOSTRAGEM POR CONGLOMERADOS ADAPTATIVOS ##
## EM DOIS ESTÁGIOS       ##
##   ##
################################################
################################################
################################################


ACS2EST<-function(m,n1,c)
{
indices<-integer(0)
#ordena aleatoriamente os indices
indices<-sample(1:M,m)

tau<-integer(0)
tam<-integer(0)
tauhh<-integer(0)

for (i in 1:length(indices))
{

#cria alguns elementos a serem utilizados posteriormente
psu<-integer(0)
```

```
#elemento da amostra de unidades primárias
psu<-subset(dados2,grupo==indices[i])

linhas<-max(psu[,1])-min(psu[,1])+1
colunas<-max(psu[,2])-min(psu[,2])+1


psuf<-array(integer(0),c(linhas,colunas))

#coloca a matriz psu na forma desejada
for (j in 1:nrow(psu))
{
r<-psu[j,1]-min(psu[,1])+1
c<-psu[j,2]-min(psu[,2])+1
v<-psu[j,3]

psuf[r,c]<-v
}


#total na psu
acsvec<-ACS(psuf,round(n1/m),c)
tau[i]<-acsvec[1,1] #computa o total para a unidade primaria
tam[i]<-acsvec[1,2] #tamanho amostral dentro da unidade primária
tauhh[i]<-acsvec[1,3] #computa o total HH para a unidade primaria
}

total<-sum(tau)*(M/m)
totalhh<-sum(tauhh)*(M/m)
tamanho<-sum(tam)

result<-cbind(total,tamanho,totalhh)
result
}




########################################
########################################
########################################
##                                    ##
## AMOSTRAGEM EM DOIS ESTÁGIOS SIMPLES ##
##    ##
########################################
########################################
########################################


#amostragem em dois estágios simples

aas2st<-function(m,n)
{
indices<-integer(0)
secundaria<-integer(0)

#ordena aleatoriamente os indices das unidades primárias de seleção
indices<-sample(1:M,m)

tau<-integer(0)
for (i in 1:length(indices))
```

```
{
#cria alguns elementos a serem utilizados posteriormente
psu<-integer(0)

#elemento da amostra de unidades primárias
psu<-subset(dados2,grupo==indices[i])

#amostra das unidades secundarias

secundaria<-sample(1:nrow(psu),round(n/m))

#valores
y<-psu[secundaria,3]

#total na psu
tau[i]<-(nrow(psu)/length(y))*sum(y)
}

est2<-sum(tau)*(M/m)
tam<-round(n/m)*m

cbind(est2,tam)
}

#####################################
#####################################
#####################################
###                               ###
### PARTE REPETITIVA DA SIMULAÇÃO ###
###     ###
#####################################
#####################################
#####################################



#Declaro os vetores que serão construidos a diante
HT<-integer(0)
tam<-integer(0)
HH<-integer(0)
junto<-0


for (m  in c(4,8))
for (n1 in c(2*m,5*m,8*m))
for (c in c(10))
{
{
{



for (r in 1:1000)
{
#############################################################################
### Parte que simula os estimadores referentes ao         ###
### Esquema de amostragem por conglomerados adaptativos em dois estágios ###
#############################################################################

acsobs<-ACS2EST(m,n1,c) #vetor com o valor dos estimadores e do tamanho amostral
HT[r]<-acsobs[1,1]     #valor estimado pelo Horvitz-Thompson
```

```
tam[r]<-acsobs[1,2]    #tamanho de amostra retirado
HH[r]<-acsobs[1,3]     #valor estimado pelo Hansen-Hurwitz
}

#####################################################
###   cálculo das Estatísticas dos Estimadores     ###
#####################################################


#Amostragem por conglomerados adaptativos

EHT<-mean(HT) #Média do estimador Horvitz-Thompson
VHT<-var(HT) #Variância do estimador Horvitz-Thompson

EHH<-mean(HH) #Média do estimador Hansen-Hurwitz
VHH<-var(HH) #Variância do estimador Hansen-Hurwitz

Eeta<-mean(tam) #tamanho amostral médio


estatisticas<-c(m,n1,c,EHT,VHT,EHH,VHH,Eeta)
junto<-rbind(junto,estatisticas)
}
}
}

write.table(junto,"c:\\acs2estagios_valid.txt",sep="\t")
```

## B.3 Amostragem seqüencial em dois estágios

```
#Limpa Tudo
rm(list = ls())
#Exemplo dos Marrecos da Asa Azul
teal<-array(0,c(20,10))
teal[1,9]<-60
teal[2,5]<-1
teal[2,8]<-122
teal[2,9]<-114
teal[2,10]<-3
teal[3,5]<-7144
teal[3,6]<-6399
teal[3,8]<-14
teal[4,4]<-103
teal[4,5]<-150
teal[4,6]<-6
teal[5,4]<-10
teal[7,6]<-2
teal[7,10]<-2
teal[9,5]<-3
teal[11,3]<-12
teal[12,3]<-2
teal[12,6]<-2
teal[13,3]<-4
teal[14,1]<-5
```

```
teal[14,3]<-20
teal[15,2]<-3

e<-c(1,0,2,10,1000)
exemplo<-array(e,c(1,5))
N<-nrow(teal)*ncol(teal)


#cria uma variável chamada grupo
grupo<-array(NA,c(200,1))
dados<-cbind(which(teal>=0,arr.ind=T),teal[which(teal>=0,arr.ind=T)])

for (i in 1:nrow(dados))
{
if (dados[i,1]>=1 && dados[i,1]<=5 && dados[i,2]>=1 && dados[i,2]<=5) grupo[i]<-1
if (dados[i,1]>=6 && dados[i,1]<=10 && dados[i,2]>=1 && dados[i,2]<=5) grupo[i]<-2
if (dados[i,1]>=11 && dados[i,1]<=15 && dados[i,2]>=1 && dados[i,2]<=5) grupo[i]<-3
if (dados[i,1]>=16 && dados[i,1]<=25 && dados[i,2]>=1 && dados[i,2]<=5) grupo[i]<-4
if (dados[i,1]>=1 && dados[i,1]<=5 && dados[i,2]>=6 && dados[i,2]<=10) grupo[i]<-5
if (dados[i,1]>=6 && dados[i,1]<=10 && dados[i,2]>=6 && dados[i,2]<=10) grupo[i]<-6
if (dados[i,1]>=11 && dados[i,1]<=15 && dados[i,2]>=6 && dados[i,2]<=10) grupo[i]<-7
if (dados[i,1]>=16 && dados[i,1]<=25 && dados[i,2]>=6 && dados[i,2]<=10) grupo[i]<-8
}


#coloca ela em uma matiz dados2 junto com as outras informações
dados2<-cbind(dados,grupo)

#até aqui insere as informações

#o input do programa é a tabela com o formato da dados2

M<-8 #Número de unidades primárias na população


tsa<-function(m,ni1,kni2,c)
{
indices<-integer(0)
#ordena aleatoriamente os indices
indices<-sample(1:M,m)

etai<-integer(0)
taui<-integer(0)
for (i in 1:length(indices))
{

#cria alguns elementos a serem utilizados posteriormente
matriz<-integer(0)
yi<-integer(0)
lambda<-integer(0)



#elemento da amostra de unidades primárias
matriz<-subset(dados2,grupo==indices[i])

#amostra das unidades secundarias
amostrasec<-sample(1:nrow(matriz),nrow(matriz))

#parte referente a amostra inicial
inicial<-amostrasec[1:ni1]
```

```
adicional<-integer(0)

#selecionar mais 4 unidades nas unidades primárias que a condicao é satisfeita
if (max(matriz[inicial,3])>c) adicional<-amostrasec[(ni1+1):(ni1+kni2)]

#indices amostrados

totali<-cbind(t(inicial),t(adicional))
yti<-cbind(t(matriz[inicial,3]),t(matriz[adicional,3]))

#calculo de l
li<-length(which(yti>c))
Ni<-nrow(matriz)

ni2<-length(adicional)

for (j in 1:length(yti))
{
#Calculo de P(si/j)/P(si) aqui chamado de lambda

if (ni2==0) lambda[j]<-Ni/ni1

else if (ni2>0 && li>ni2)
lambda[j]<-Ni/(ni1+ni2)

else if (ni2>0 && li<=ni2 && yti[j]>c)
{num<-Ni*factorial(ni1+ni2-1)
deno<-factorial(ni1+ni2)-(factorial(ni2)*factorial(ni1+ni2-li)/factorial(ni2-li))
lambda[j]<-num/deno}

else if (ni2>0 && li<=ni2 && yti[j]<=c)
{num<-Ni*(factorial(ni1+ni2-1)-(factorial(ni2)*factorial(ni1+ni2-1-li)/factorial(ni2-li)))
deno<-factorial(ni1+ni2)-(factorial(ni2)*factorial(ni1+ni2-li)/factorial(ni2-li))
lambda[j]<-num/deno}
}

#estimador de murthy para cada psu

taui[i]<-sum(lambda*t(yti))
etai[i]<-ni1+ni2
}

tau<-(M/m)*sum(taui[which(taui>0)])
eta<-sum(etai)

cbind(tau,eta)
}

aas2st<-function(m,n)
{
indices<-integer(0)
#ordena aleatoriamente os indices
indices<-sample(1:M,m)

tau<-integer(0)
for (i in 1:length(indices))
{

#cria alguns elementos a serem utilizados posteriormente
psu<-integer(0)
```

```
#elemento da amostra de unidades primárias
psu<-subset(dados2,grupo==indices[i])

#amostra das unidades secundarias
secundaria<-sample(1:nrow(psu),round(n/m))

#valores
y<-psu[secundaria,3]

#total na psu
tau[i]<-(nrow(psu)/length(y))*sum(y)
}

est2<-sum(tau)*(M/m)
tam<-round(n/m)*m

cbind(est2,tam)
}


m<-4 #Número de unidades primárias a serem selecionadas na amostra
ni1<-2 #Número de unidades secundárias em cada unidade primária para a primeira seleção
kni2<-7 #Número de unidades secundárias a serem selecionadas adicionalmente
c1<-10 #condição C amostragem seqüencial em dois estágios

total<-integer(0)
amostra<-integer(0)
eq<-integer(0)
eqaas<-integer(0)
junto<-integer(0)
#simulações

#variar amostra inicial n1 de 1 até 20
#variar condicao C de
#condição C 0,1,2,3,4,5,6,10,12,14,20
#variar elementos adicionais kni2 1,2,3,4,5,6,7,8,9,10
#variar m de 2 a 8



#for (m  in c(2,4,8))
#for (n1 in c(2*m,5*m,8*m))
#for (c in c(0,5,12))



##
#c(0,5,12)#
#c(2,5,8)#
#c(5,10,15)#

tst2<-integer(0)
st2<-integer(0)
aas<-integer(0)
total2st<-integer(0)
amostra2st<-integer(0)
st2esp<-integer(0)
st2espm<-integer(0)

for (ni1 in c(2,5,8))
```

```
for (c1 in c(10))
for (m in c(2,4,8))
for (kni2 in  c(2,5,8))
{
{
{
{

for (r in 1:10000)
{
tsss<-tsa(m,ni1,kni2,c1)
total[r]<-tsss[1,1]
amostra[r]<-tsss[1,2]
aas[r]<-N*(sum(sample(teal,amostra[r]))/amostra[r])
tst2<-aas2st(m,amostra[r])
total2st[r]<-tst2[1,1]
amostra2st[r]<-tst2[1,2]

eq[r]<-(total[r]-14181)**2
eqaas[r]<-(aas[r]-14181)**2
}


t1<-round(mean(amostra))
t2<-t1+1

aasesp<-integer(0)
aasespm<-integer(0)
st2esp<-integer(0)
st2espm<-integer(0)

for (r in 1:10000)
{
aasesp[r]<-N*(sum(sample(teal,t1))/t1)
aasespm[r]<-N*(sum(sample(teal,t2))/t2)
st2esp[r]<-aas2st(m,t1)[1,1]
st2espm[r]<-aas2st(m,t2)[1,1]
}



#estatísticas amostragem dois estágios seqüencial

stat<-c(m,ni1,kni2,c1,mean(amostra),mean(total),var(total),mean(amostra2st),mean(total2st),var(total2st),mean(eq),mean(aas),var(aas),mean(eqaas),var(aasesp),var(aasespm),
junto<-rbind(junto,stat)
}
}
}
}

write.table(junto,"c:\\seq2stmarrecoc10.txt",sep="\t")
```

# Referências Bibliográficas


Basu, D. 1969. Role of Sufficiency and likelihood principles in sample survey theory. *Sankyā A*, **31**, 441–454.

Christman, M. C. 2000. A Review of Quadrat-Based Sampling of Rare, Geographically Clustered Populations. *Journal of Agricultural, Biological and Environmental Statistics*, **5**, 168–201.

Christman, M. C. 2003. Adaptive Two-Stage One-per-Stratum Sampling. *Environmental and Ecological Statistics*, **10**, 43–60.

Francis, R. I. C. C. 1984. An Adaptive Strategy for Stratified Reom Trawl Surveys. *New Zeale Journal of Marine and Freshwater Research*, **18**, 59–71.

Goodman, L. A. 1961. Snowball Sampling. *The Annals of Mathematical Statistics*, **32**, 148–170.

Hanselman, D. H., Quinn II, T. J., Lunsford, C., & Clausen, D. 2003. Aplications in Adaptive Cluster Sampling Gulf of Alaska Rockfish. *Fisheries Bulletin*, **101**, 501–513.

Hansen, M. M., & Hurwitz, W. N. 1943. On The Theory of Sampling From Finite Populations. *Annals of Mathematical Statistics*, **14**, 333–362.

Horvitz, D. G., & Thompson, D. J. 1952. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.

Jolly, G. M., & Hampton, I. 1990. A Stratified Reom Transect Design for Acoustic Surveys of Fish Stocks. *Canadian Journal of Fisheries and Aquatic Sciences*, **47**, 1282–1291.

Kalton, G. 2001. Practical Methods for Sampling Rare and Mobile Populations. *Proceedings of the Annual Meeting of the American Statistical Association*, 5–9.

Kalton, G., & Anderson, D. W. 1986. Sampling Rare Populations. *Journal of Royal Statistical Society*, **149**, 65–82.

Kalton, G., & Sudman, S. 1986. New Developments in the Sampling of Special Populations. *Annual Review of Sociology*, **12**, 401–29.

Kish, L. 1965. *Survey Sampling*. New York: Wiley.

Lo, N. C. H., Griffith, D., & Hunter, J. R. 2001. Using Restricted Adaptive Cluster Sampling to estimate Pacific Hake Larval Abundance. *Biometrics*, **37**, 160–174.

McDonald, L. L. 2004. Sampling Rare Populations. *Sampling Rare or Elusive Species*, 11–42.

Murthy, M. N. 1957. Ordered and Unordered Estimators in Sampling Without Replacement. *Sankyā*, **18**, 379–390.

Raj, D. 1956. Some Estimators in Sampling With Varying Probabilities without replacement. *Journal of the American Statistical Association*, **51**, 269–284.

Roesch Jr., F. A. 1993. Adaptive cluster sampling for forest inventories. *Forest Science*, **39**, 655–669.

Salehi, M. M. 1999. Rao-Blackwell Versions of the Horvitz-Thompson and Hansen-Hurwitz in Adaptive Cluster Sampling. *Environmental and Ecological Statistics*, **6**, 183–195.

Salehi, M. M. 2003. Comparison between Hansen-Hurwitz and Horvitz-Thompson estimators for adaptive cluster sampling. *Environmental and Ecological Statistics*, **10**, 115–127.

Salehi, M. M., & Seber, G. A. F. 1997. Two-Stage Adaptive Cluster Sampling. *Biometrics*, **53**, 959–970.

Salehi, M. M., & Smith, D. R. 2005. Two-Stage Sequential Sampling: A Neighborhood-Free Adaptive Sampling Procedure. *Journal of Agricultural, Biological and Environmental Statistics*, **10**(1), 84–103.

Sarndal, C.E., Swensson, B., & Wretman, J. 1992. *Model Assisted Survey Sampling*. New York: Springer.

Smith, D. R., Conroy, M. J., & Brakhage, D. H. 1995. Efficiency of Adaptive Cluster Sampling for Estimation Density of Wintering Waterfowl. *Biometrics*, 777–788.

Smith, D. R., Villela, R. F., & Lemarié, D. P. 2003. Application of Adaptive Cluster Sampling to Low-density Populations of Freswater Mussels. *Environmental and Ecologial Statistics*, **10**(1), 7–15.

Smith, D. R., Brown, J. A., & Lo, N. C. H. 2004. Sampling Rare Populations. *in Sampling Rare or Elusive Species*.

Thompson, S. K. 1990. Adaptive Cluster Sampling. *Journal of the American Statistical Association*, **85**, 1050–1059.

Thompson, S. K., & Seber, G. A. F. 1996. *Adaptive Sampling*. New York: Wiley.