Amostragem por Conglomerados e Amostragem em Múltiplos Estágios

JOÃO LUÍS F. BATISTA
Outubro de 2006

Notas para a disciplina LCF-764 Métodos de Amostragem em Levantamentos Florestais do Programa de Pós-Graduação em Recursos Florestais, ESALQ, Universidade de São Paulo.

1 Motivação

- Na *amostragem por conglomerados*, as unidades amostrais são chamadas de *elementos* sendo que a aleatorização é realizada para um conjunto ou conglomerado de unidades amostrais.
- Já na *amostragem em múltiplos estágios*, a população alvo é sofre subdivisões em várias escalas, chamadas de *estágios*. As unidade amostrais presentes na escala final (último estágio) são selecionadas a partir de um sistema de aleatorização que envolve todos os estágio.
- Tanto a amostragem por conglomerados como a amostragem em múltiplos estágios são utilizadas em levantamentos voltados para grandes áreas, com no caso de *levantamentos regionais*, ou quando o orçamento ou tempo disponível para o levantamento são muito limitantes. Nesses dois casos, os delineamentos básicos, como a amostragem aleatória simples com ou sem estratificação, se tornam de implementação muito difícil (grandes áreas) ou de eficiência muito limitada.

- Conglomerados e múltiplos estágios têm estrutura teórica de delineamento bastante semelhante, sendo mais fácil entender a amostragem em múltiplos estágios depois de aprendido a amostragem por conglomerados.

2 *Amostragem por Conglomerados*

- Na amostragem por conglomerados, as unidades amostrais (*elementos*) são espacialmente agrupadas formando um conglomerado, os quais são locados na população alvo segundo um delineamento amostral definido, o que implica que os conglomerados é que são aleatorizados e não as unidades amostrais.
- Considerando o mesmo tamanho de amostra, isto é, mesmo número de unidades amostrais ou elementos, a amostragem por conglomerados é geralmente *menos eficiente* que a amostragem aleatória simples, mas ela implica em menores custos, uma vez que várias unidades amostrais estão próximas.
- Por tanto a razão da utilização da amostragem por conglomerados é quase que invariavelmente a questão de custo.

Exemplo: Levantamento Regional de Fragmentos Florestais

Suponha que se deseja fazer um diagnóstico do estado de conservação dos fragmentos numa região de 6 milhões de hectares, sendo que os fragmentos totalizam 700 mil hectares dentro dessa região.

Considerando qualquer tipo de delineamento amostral, o custo de deslocamento entre unidades amostrais será provavelmente o custo majoritário no levantamento.

Se utilizarmos algum tipo de delineamento amostral básico, será inevitável que ocorra como frequência a situação em que a equipe de campo ao alcançar um fragmento instalará uma única unidade amostral no fragmento, tendo que se deslocar para o próximo fragmento.

O custo de um levantamento assim será muito alto e a maior parte do tempo da equipe de campo será gasto em deslocamento.

Na amostragem por conglomerado, a equipe ao alcançar um fragmento, instala um conglomerado ou um conjunto de unidades amostrais, melhorando a relação entre custo de deslocamento e custo de instalação e medição de unidades amostrais.

2.1 Estrutura de Conglomerado

Os elementos do conglomerado (unidades amostrais) seguem um arranjo sistemático dentro de cada conglomerado. Os vários arranjos resultam em várias estruturas de conglomerado (figura 1).

Conglomerados Tradicionais: Os conglomerados tradicionais são compostos de elementos (parcelas ou pontos amostrais) com arranjo espacial definido.

Parcelas em Faixa: As parcelas em faixa com largura fixa e comprimento variável são conglomerados onde cada elemento é composto por uma parcela quadrada (comprimento igual a larguar), e o tamanho do conglomerado é variável dependendo do comprimento da faixa.

A idéia de conglomerado também se apresenta quando as árvores dentro de uma unidade amostral são tomadas como elementos. Para estimativa de atributos das árvores individuais, as unidades amostrais devem ser consideradas como conglomerados:

- estimativas relacionadas a medidas tomadas em árvores individuais: DAP médio, altura média, DAP máximo, etc.;
- estimativa de proporções de árvores na população: taxa de árvores mortas, porcentagem de falhas, porcentagem de árvores bifurcadas, etc.

Discutiremos alguns estimadores para amostragem por conglomerados na amostragem aleatória simples.

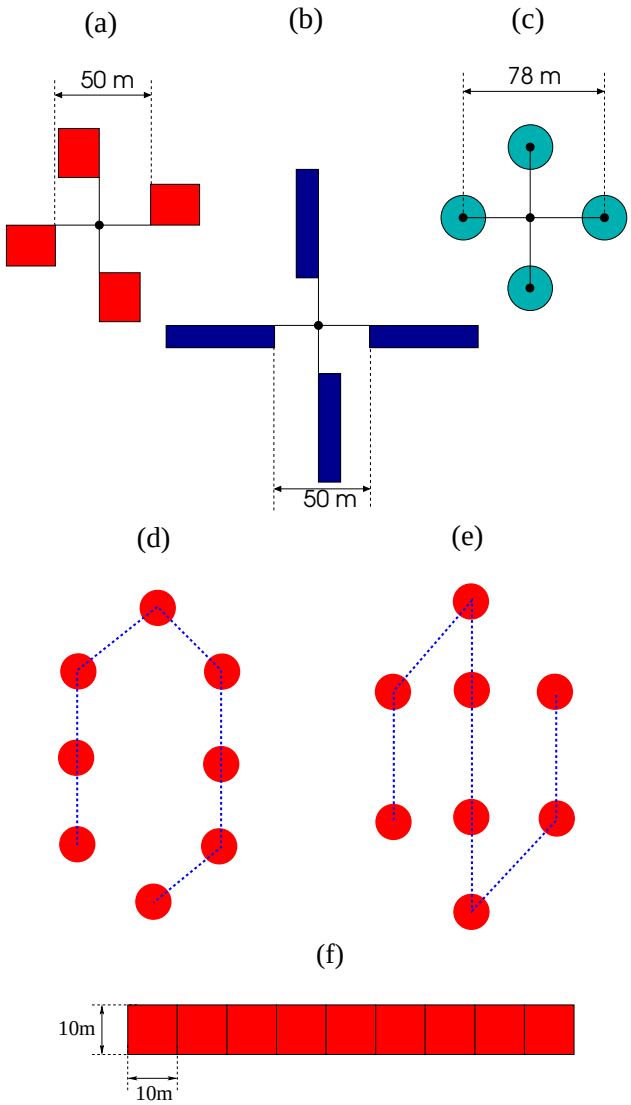

Figura 1: Exemplos de conglomerados utilizados em levantamentos florestais: (a), (b) e (c) conglomerados em cruz; (c) e (d) conglomerados em formato hexagonal, (f) conglomerado em faixa.

2.2 Conglomerados de Mesmo Tamanho

- Quando os conglomerados tem o mesmo tamanho é possível utilizar um estimador não-viciado.
- Estimador do total do conglomerado:

$$\widehat{\tau} = \frac{N}{n} \sum_{i=1}^{n} y_i = N\,\widehat{\mu}$$

onde:

$y_i = \sum_{j=1}^{m} y_{ij}$ – total do conglomerado i;

m – tamanho dos conglomerados;

$\widehat{\mu} = (1/n) \sum_{i=1}^{n} y_i$ – estimador da média dos conglomerados;

N – tamanho da população, isto é, número de conglomerados na população.

- A variância da estimativa do total é

$$\widehat{\mathbf{Var}}\{\widehat{\tau}\} = N^2 \frac{S_u^2}{n} \left(1 - \frac{n}{N}\right)$$

onde

$$S_u^2 = \frac{\sum_{i=1}^{n} (y_i - \widehat{\mu})^2}{n-1}$$

- O estimador da média dos conglomerados tem variância

$$\widehat{\mathbf{Var}}\{\widehat{\mu}\} = \frac{S_u^2}{n} \left(1 - \frac{n}{N}\right)$$

- Essencialmente, os estimadores nesse caso são os mesmos da amostragem aleatória simples, tomando os conglomerado como unidades amostrais.

2.3 Conglomerados de Tamanhos Diferentes

- Quando o tamanho do conglomerado varia bastante, o seu total (y_i) estará altamente correlacionado como o seu tamanho (x_i), sendo eficiente o uso do estimador de razão.

- Estimador de razão para o total

$$\widehat{\tau}_R = \widehat{R}\, X = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} X$$

onde:

$y_i = \sum_{j=1}^{x_i} y_{ij}$ – total do conglomerado i;

x_i – tamanho do conglomerado i, isto é, número de elementos;

X – total de elementos na população.

- A variância do estimador do total

$$\begin{aligned} \widehat{\mathbf{Var}}\{\widehat{\tau}_R\} &= N^2 \frac{S_R^2}{n}\left(1 - \frac{n}{N}\right) \\ &= \frac{N(N-n)}{n(n-1)} \sum_{i=1}^{n} \left(y_i - \widehat{R} x_i\right)^2 \end{aligned}$$

Fazendo o ajuste de variância sugerido por Cochran (1977):

$$\widetilde{\mathbf{Var}}\{\widehat{\tau}_R\} = \left(\frac{\mu_X}{\widehat{\mu}_X}\right)^2 \widehat{\mathbf{Var}}\{\widehat{\tau}_R\}$$

onde:

$\mu_X = X/N$ – tamanho médio de conglomerado na população; e

$\widehat{\mu}_X = \sum_{i=1}^{n} x_i/n$ – tamanho médio de conglomerado na amostra.

2.4 Estimando Proporções ou Atributos de Árvores Individuais

- No caso de atributos de árvores individuais e proporções o nosso interesse não está no total, mas na própria razão populacional.
- Nesse caso, cada árvore é um *elemento* e o conjunto de árvores faz com que a unidade amostral (parcela, ponto, etc.) seja considerada o *conglomerado*.
- A observação y_{ij} é uma medida tomada em cada árvore individualmente, como diâmetro, altura, biomassa, etc. No caso de porporção de árvores

em determinada categoria (morta, bifurcada, doente, etc.), a observação se torna binária:

$$y_{ij} = \begin{cases} 1 & \text{se a condição for satisfeita,} \\ 0 & \text{se a condição } \textbf{não} \text{ for satisfeita.} \end{cases}$$

- Para qualquer variável observada nas árvores individualmente, o total por conglomerado será

$$y_i = \sum_{j=1}^{x_i} y_{ij}$$

onde x_i é o tamanho do conglomerado, isto é, o número de árvores na parcela i.

- O estimador da razão nos dará o atributo médio por árvore ou a proporção desejada:

$$\begin{aligned} \widehat{R} &= \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} \\ \widehat{\mathbf{Var}}\{\widehat{R}\} &= \frac{1}{\mu_X^2}\frac{1}{n}\left[\frac{\sum_{i=1}^{n}(y_i - \widehat{R}x_i)^2}{n-1}\right]\left(1 - \frac{n}{N}\right) \end{aligned}$$

- O parâmetro μ_X, número médio de elementos por conglomerado na população, em geral não é conhecido no caso de árvores individuais, sendo substituído pela informação amostral

$$\widehat{\mu}_X = \frac{\sum_{i=1}^{n} x_i}{n}$$

2.5 Tamanho de Conglomerado e Tamanho de Amostra

- A questão do tamanho de conglomerado é análoga a questão de tamanho de parcela, pois na prática a parcela pode ser vista como um conjunto de elementos-árvores.

- As mesmas abordagens apresentadas no estudo de tamanho de parcela são válidas no estudo de tamanho de conglomerados.

- No caso de conglomerados, o coeficiente de correlação intraclasse é chamado de *coeficiente de correlação intraconglomerado* (intracluster correlation coefficient), e permite considerar simultaneamente a influência do tamanho e do número de conglomerados na precisão da estimativa da média.

3 Amostragem em Dois Estágios

- A amostragem em múltiplos estágios é outro delineamento amostral utilizado em grandes áreas, como em levantamentos regionais ou nacionais.
- A grande dimensão espacial leva o delineamento amostral a considerar diferentes escalas de observação para que a estimativa global, na maior escala, tenha alta confiabilidade.

Exemplo: Levantamento Regional de Florestas Nativas

Num levantamento regional, deseja-se fazer um diagnóstico do estado de conservação das florestas nativas e também apresentar uma estimativa regional da biomassa dessas florestas.

Podemos pensar em três estágios de amostragem:

1º Estágio: imagens de satélite cobrem toda a região a ser estudada. Cada cena (imagem) define uma *unidade primária de amostragem*, que pode ser selecionada por amostragem aleatória simples ou amostragem estratificada.

2º Estágio: dentro de cada cena (unidade primária) podem ser identificados os fragmentos florestais que são definidos como *unidades secundárias de amostragem*. Os framentos podem ser amostrados com probabilidade proporcional ao tamanho.

3º Estágio: dentro de cada fragmento, as unidades amostrais (parcelas) que serão instaladas no campo são definidas como *unidades terciárias de amostragem*, as quais podem ser instaladas segundo uma amostragem sistemática.

Os três estágios juntamente com o delineamento amostral utilizado em cada estágio consiste numa amostragem em múltiplos estágios. Os estimadores nesse exemplo serão bastante complexo, requerendo um desenvolvimento específico.

- Apesar de não haver limites teóricos para o número de estágios a serem utilizados numa amostragem em múltiplos estágios, os estimadores se tornam rapidamente bastante complexos e complicados de se aplicar.
- A maioria dos levantamento florestais utiliza levantamentos em dois estágios.
- A amostragem em dois estágios difere dos delineamentos amostrais básicos pelo fato de que se toma *uma amostra das unidades primárias*, enquanto que na amostra aleatória simples ou na amostra estratificada *todas unidades primárias* devem ser amostradas.
- A amostragem em dois estágios se apresenta superior em relação à amostragem aleatória simples ou amostra aleatória estratificada quando a dimensão da área de estudo ou a limitação de tempo resultaria numa baixa intensidade amostral em todas unidades primárias.

Exemplo: Valoração de uma Floresta Plantada

Foi solicitada uma valoração expedita de uma floresta plantada de 3000 ha composta por 67 talhões. O tempo disponível para o levantamento é muito pequena para se realizar um inventário na forma tradicional.

O levantamento expedito pode ser realizado em dois estágios:

1º Estágio: a floresta é dividida em 67 talhões sendo que cada um é uma *unidade primária*. Nesse estágio pode se utilizar amostragem aleatória simples, se houver certa homogeneidade entre os talhões, ou amostragem estratificada, caso se conheça a heterogeneidade espacial da floresta.

2º Estágio: como unidades secundárias se utiliza as unidades amostrais que podem ser parcelas ou pontos, que seriam amostrados de forma sistemática nos talhões selecionados no primeiro estágio.

Além dos delineamentos básicos, outros delineamentos, como amostragem dupla no segundo estágio, poderiam ser combinados com a amostragem em dois estágios para aumentar a sua eficiência.

3.1 Conceito de Estágios

- Um *estágio* na amostragem em múltiplos estágios representa uma *escala* na qual *toda a população alvo* é subdividida.
- Cada unidade num estágio é subsequentemente subdividida em unidades no estágio seguinte, de modo a formar um sistema hierárquico de unidades, no qual um dado número de unidades num certo estágio compõem uma unidade no estágio imediatamente superior.
- Na perspectiva de subdividir a população em estágios, a amostragem por conglomerados poderia ser designada como *amostragem em um único estágio*. Esse único estágio seria o conglomerado, dentro do qual todos elementos são observados.

3.2 Estimador Não-Viciado

- Utilizando a amostragem aleatória simples, é possível utilizar um estimador não-viciado para se estimar o total populacional.
- Estimador do total na $i^{\text{ésima}}$ unidade primária:

$$\widehat{y}_i = M_i \frac{\sum_{j=1}^{m_i} y_{ij}}{m_i} = M_i \, \overline{y}_i$$

onde:

y_{ij} – valor observado na $j^{\text{ésima}}$ unidade secundária dentro da $i^{\text{ésima}}$ unidade primária;

M_i – tamanho da $i^{\text{ésima}}$ unidade primária;

m_i – tamanho da *amostra* na $i^{\text{ésima}}$ unidade primária;

- O estimador do total populacional fica:

$$\widehat{\tau} = N\,\frac{\sum_{i=1}^{n}\widehat{y}_i}{n} = N\,\widehat{\mu}_1$$

- Para se obter um estimador da variância do total é necessário considerar que temos duas escalas em que se observa a variabilidade dos valores:
  - Variância ***dentro*** das unidades primárias:

$$S_D^2 = \sum_{i=1}^{n} M_i^2 \frac{S_{Di}^2}{m_i}\left(1 - \frac{m_i}{M_i}\right)$$

    onde: a variância *dentro da $i^{ésima}$ unidade primária* é dada por

$$S_{Di}^2 = \frac{1}{m_i - 1}\sum_{j=1}^{m_i}(y_{ij} - \overline{y}_i)^2.$$

  - Variância ***entre*** as unidades primárias (totais):

$$S_E^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\widehat{y}_i - \widehat{\mu}_1)^2\left(1 - \frac{n}{N}\right).$$

  A variância do total é uma composição das variâncias entre e dentro de unidades primárias:

$$\widehat{\mathbf{Var}}\{\widehat{\tau}\} = N^2\frac{S_E^2}{n} + N\frac{S_D^2}{n}$$

3.3 Estimador de Razão

- Quando as unidades primárias variam muito de tamanho, o estimador não viciado se torna pouco eficiente pois essa variação no tamanho aumentará a sua variância. Nesse caso, o estimador de razão, embora enviesado, será mais eficiente.
- O estimador de razão para o total populacional é

$$\widehat{\tau}_R = \widehat{R}\,M = \frac{\sum_{i=1}^{n}\widehat{y}_i}{\sum_{i=1}^{n}M_i}\,M$$

  onde

M – número de *unidades secundárias* na população;

$\widehat{y}_i$ – total estimado da $i^{\text{ésima}}$ unidade primária.

- Novamente a variância do total também tem dois componentes: variância entre e dentro unidades primárias.

$$\widehat{\mathbf{Var}}\{\widehat{\tau}_R\} = \frac{N^2}{M^2 n}\left(S_E^{*2} + S_D^2\right)$$

Mas a variância entre unidades primárias deve ser obtida através do estimador de razão:

$$S_E^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}(\widehat{y}_i - \widehat{R}M_i)^2\left(1 - \frac{n}{N}\right)$$

3.4 Amostragem com Probabilidade Proporcional ao Tamanho

- É comum que as unidades primárias sejam selecionadas com probabilidade proporcional ao tamanho, uma vez que frequentemente a amostragem é realizada com base em área.

- Nesse caso, a probabilidade de uma unidade primária ser observada é diretamente proporcional ao seu tamanho (M_i), o que deve ser considerado no estimador:

$$\widehat{\tau}_P = \frac{M}{n}\sum_{i=1}^{n}\frac{\widehat{y}_i}{M_i} = \frac{M}{n}\sum_{i=1}^{n}\overline{y}_i$$

onde temos

  - a média das unidades secundária na $i^{\text{ésima}}$ unidade primária

$$\overline{y}_i = \frac{1}{m_i}\sum_{j=1}^{m_i} y_{ij};$$

  - o total estimado da $i^{\text{ésima}}$ unidade primária

$$\widehat{y}_i = M_i\,\overline{y}_i.$$

- A variância do estimador do total é, nesse caso, obtida a partir das diferenças das estimativas médias nas unidades secundárias:

$$\widehat{\mathbf{Var}}\{\widehat{\tau}_P\} = \frac{M^2}{n(n-1)} \sum_{i=1}^{n} (\overline{y}_i - \widehat{\mu}_P)^2$$

  sendo que

$$\widehat{\mu}_P = \frac{\widehat{\tau}_P}{M}$$

  é a média estimada por unidade secundária.

- A amostra das unidades primárias sendo proporcional ao seu tamanho, torna-se proporcional ao número de unidades secundárias. Assim, as unidades secundárias têm a probabilidade de seleção constante, de modo que a variância do total possa ser obtida com base nelas.

3.5 Estratificação

- A amostragem em dois estágios é frequentemente combinada com a amostragem estratificada, geralmente aplicando-se a amostragem em dois estágios dentro de cada estrato.

- A maior homogeneidade gerada pela amostragem estratificada aumenta a eficiência da amostragem em dois estágios, pois aumenta a homogeneidade entre as unidades primárias, de modo que a amostragem aleatória simples da unidades primárias possa ser eficiente.

- Os totais obtidos em cada estrato, pela amostragem em dois estágios, podem ser combinados para se obter uma estimativa geral da população alvo através dos estimadores e variâncias apresentados na amostragem estratificada.

Problemas e Exercícios

1. Num inventário florestal em floresta plantada de *Eucalyptus grandis* foram medidas as alturas de todas as árvores dominantes nas parcelas (dados na ta-

bela abaixo). Encontre a altura média das árvores dominantes e o respectivo intervalo de confiança de 95%.

Parcela	Altura das Árvores Dominantes (m)				
1	28.2	28.7	21.2	27.9	
2	22.8	27.7	27.3		
3	25.3	24.9			
4	24.2	28.7	25.2	21.6	27.5
5	23.0	23.3	22.6	21.7	
6	22.3	21.9	21.5	25.5	22.4
7	26.3	22.3	26.2	19.6	24.8
8	25.4	26.9	25.3	24.9	
9	23.1	26.4	25.4	28.5	25.0
10	25.1	25.7			
11	21.6	26.9	20.4	24.7	23.0
12	23.3	21.1	27.2		
13	23.1	29.2	19.8	24.0	24.0
14	24.3	27.8	20.2	29.8	21.0
15	19.8	21.5	24.4	19.0	
16	28.7	26.5	20.5	22.3	29.8
17	29.3	24.7	24.1	22.4	28.0
18	25.6	25.7	23.6		
19	22.8	21.7	25.4	23.2	23.6
20	25.0	27.8	23.4	24.3	23.3
21	27.1	26.0			
22	23.1	20.3	28.7	21.4	23.2
23	22.5	20.3	25.4	27.5	23.3
24	22.3	23.9			
25	26.7	27.7	24.2	22.1	28.9
26	22.9	22.6	24.1		
27	26.5	28.4	20.6		
28	22.2	22.8	20.8	26.0	
29	26.3	20.1	25.7		
30	20.0	25.9	25.8	24.0	
31	25.3	23.5			
32	24.3	20.9	25.6		

2. Para estimar a regeneração do palmiteiro (*Euterpe edulis*), decidiu-se realizar um levantamento utilizando conglomerados compostos de quatro elementos (parcelas circulares de 12 m^2). A tabela abaixo apresenta os dados observados. Encontre a densidade média (plântulas/ha) e o respectivo intervalo de confiança de 95%.

Conglomerado	Elemento	N. Plântulas	Conglomerado	Elemento	N. Plântulas
2002	1	1	2019	1	13
2002	2	8	2019	2	1
2002	3	9	2019	3	3
2002	4	11	2019	4	29
2008	1	63	2020	1	16
2008	2	24	2020	2	42
2008	3	63	2020	3	4
2008	4	37	2020	4	8
2009	1	0	2021	1	1
2009	2	0	2021	2	3
2009	3	1	2021	3	5
2009	4	2	2021	4	11
2010	1	9	2023	1	6
2010	2	28	2023	2	8
2010	3	5	2023	3	6
2010	4	12	2023	4	8
2014	1	11	2024	1	25
2014	2	189	2024	2	8
2014	3	43	2024	3	6
2014	4	65	2024	4	16
2018	1	60			
2018	2	2			
2018	3	0			
2018	4	8			

3. Num levantamento expedito de floresta plantada de *Eucalyptus saligna* com 6100 ha, uma Engenheira Florestal decidiu utilizar a amostragem em dois estágios pois a floresta era composta de 120 talhões de tamanhos aproximadamente iguais ($\approx\ 50ha$). Ela selecionou aleatóriamente 10 talhões, e instalou dentro destes um número variável de parcelas de 500 m^2 (dados na tabela abaixo).

Talhão	Área	Volume					
	(ha)	(m^3/ha)					
1	53.3	119	123	128	153	153	
2	53.5	166	170	172	181	185	211
3	52.5	189	190	197	198	202	263
4	46.2	203	204	208	218		
5	52.6	226	234	243	244	246	
6	48.8	250	251	258	260		
7	55.8	265	270	276	283	288	
8	50.1	291	295	302	303	303	
9	47.6	306	308	309	314		
10	52.4	318	319	321	333	338	318

Questões:

(a) Encontre o volume médio por talhão (m^3) e respectivo intervalo de confiança de 95% utilizando o estimador sem viés.

(b) Encontre o volume total da floresta e respectivo intervalo de confiança de 95% utilizando o estimador sem viés.

(c) Qual a fonte de variabilidade: **dentro** de talhão ou **entre** talhões se mostrou mais importante, no caso do estimador sem viés?

(d) Encontre o volume total da floresta e respectivo intervalo de confiança de 95% levando em consideração o tamanho dos talhões.

(e) Qual a fonte de variabilidade (**dentro** de talhão ou **entre** talhões) se mostrou mais importante no caso do estimador de razão?