
Amostragem Estratificada

JOÃO LUÍS F. BATISTA
Setembor 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

- Raramente as populações naturais em geral, e as florestas em particular, são internamente homogêneas.
- Frequentemente, existem informações sobre a variação espacial das florestas (população alvo) de modo a permitir a sua sub-divisão em sub-populações internamente mais homogêneas.
- O processo de sub-divisão da população alvo em sub-populações mais homogêneas é chamado de *estratificação* e pode resultar num sensível ganho de precisão nas estimativas.
- Tradicionalmente a “*Amostragem Estratificada*” é definida como um tipo de delineamento amostral baseado na Amostragem Aleatória Simples.
- Mas a *estratificação* pode também ser interpretada como uma técnica amostral que pode ser consorciada a muitos tipos de delineamentos amostrais diferentes.

2 Estimativas por Estrato

- Na amostragem aleatória simples, uma única estimativa é obtida para toda a população alvo.
- Na amostragem estratificada (ou estratificação), a população alvo é subdividida em sub-populações internamente homogêneas chamadas *estratos*.
- Para cada estrato, é obtida uma estimativa que, dada a maior homogeneidade dos estratos, deve ser mais precisa que a estimativa da AAS para a população alvo.
- A estimativa para população alvo é obtida pela combinação das estimativas dos estratos.
- Notação utilizada:

$$\text{Tamanho da População: } N = \sum_{h=1}^L N_h$$

$$\text{Tamanho da Amostra: } n = \sum_{h=1}^L n_h$$

$$\text{Total da População: } \tau = \sum_{h=1}^L \tau_h$$

$$\text{Média da População: } \mu = \frac{\tau}{N}$$

onde:

h é o índice que indica o estrato;

L é o número total de estratos.

- Em cada estrato realiza-se uma amostragem *independente* dos demais estratos com as seguintes estimativas:

$$\text{Total Amostral: } \hat{\tau}_h$$

$$\text{Média Amostral: } \hat{\mu}_h$$

$$\text{Variância Amostral: } \hat{\sigma}_h^2$$

$$\text{Variância Estimada do Total: } \widehat{\text{Var}}\{\hat{\tau}_h\}$$

3 Combinação das Estimativas por Estrato

- Na amostragem estratificada utiliza-se sempre o mesmo delineamento amostral em todos os estratos e a sua implementação é totalmente independente em cada estrato.
- A maneira de combinar as estimativas de cada estrato para se obter uma estimativa global para a população alvo depende o delineamento utilizado em cada estrato.
- Apresentaremos duas formas de combinar as estimativas por estrato:
 1. forma geral que é válida para qualquer delineamento amostral; e
 2. forma específica quando se utiliza a AAS nos estratos.

3.1 Estimador para Qualquer Delineamento

A forma geral de combinação das estimativas por estrato se baseia na combinação das *estimativas dos totais*:

- Estimativa do total da população:

$$\hat{\tau}_{st} = \sum_{h=1}^L \hat{\tau}_h$$

- Estimativa da variância do total:

$$\widehat{\mathbf{Var}}\{\hat{\tau}_{st}\} = \sum_{h=1}^L \widehat{\mathbf{Var}}\{\hat{\tau}_h\}$$

- Estimativa da média da população:

$$\hat{\mu}_{st} = \frac{\hat{\tau}}{N}$$

- Variância da estimativa da média:

$$\widehat{\mathbf{Var}}\{\hat{\mu}_{st}\} = \frac{1}{N^2} \widehat{\mathbf{Var}}\{\hat{\tau}\}$$

3.2 Estimador para Amostragem Aleatória Simples

Na AAS, também se pode utilizar os estimadores baseados em totais, mas eles podem ser expressos em termos das estimativas das médias dos estratos:

- Estimativa do total por estrato:

$$\hat{\tau}_h = N_h \hat{\mu}_h$$

- Estimativa da média por estrato:

$$\hat{\mu}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}$$

- Estimativa do total da população:

$$\hat{\tau}_{st} = \sum_{h=1}^L \hat{\tau}_h \Rightarrow \hat{\tau}_{st} = \sum_{h=1}^L N_h \hat{\mu}_h$$

- Variância da estimativa do total da população:

$$\widehat{\mathbf{Var}}\{\hat{\tau}_{st}\} = \sum_{h=1}^L \widehat{\mathbf{Var}}\{\hat{\tau}_h\} \Rightarrow \widehat{\mathbf{Var}}\{\hat{\tau}_{st}\} = \sum_{h=1}^L N_h^2 \frac{\hat{\sigma}_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right)$$

- Estimativa da média:

$$\hat{\mu}_{st} = \frac{\hat{\tau}}{N} \Rightarrow \hat{\mu}_{st} = \frac{1}{N} \sum_{h=1}^L N_h \hat{\mu}_h$$

- Variância da estimativa da média:

$$\begin{aligned} \widehat{\mathbf{Var}}\{\hat{\mu}_{st}\} &= \frac{1}{N^2} \widehat{\mathbf{Var}}\{\hat{\tau}\} = \frac{1}{N^2} \sum_{h=1}^L N_h^2 \frac{\hat{\sigma}_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) \\ \Rightarrow \widehat{\mathbf{Var}}\{\hat{\mu}_{st}\} &= \sum_{h=1}^L \left(\frac{N_h}{N}\right)^2 \frac{\hat{\sigma}_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) \end{aligned}$$

- Tanto a estimativa da média como a estimativa da variância da média podem ser interpretadas como “*médias ponderadas*” das estimativas por estrato, tendo como “*peso*” o tamanho dos estratos (N_h), uma vez que:

$$N = \sum_{h=1}^L N_h.$$

3.3 Intervalo de Confiança

- Pelo Teorema Central do Limite, tanto a estimativa do total ($\hat{\tau}$) como da média (\bar{y}) no caso de grandes amostras teriam distribuição Gaussiana.
- Utilizando esse resultado podemos construir intervalos de confiança quando as amostras são “*suficientemente grandes em todos estratos*”.
- Intervalo de Confiança de $100(1 - \alpha)\%$:

$$\begin{aligned} \text{Estimativa do Total: } \hat{\tau}_{st} &\pm t(1 - \alpha/2; d) \sqrt{\widehat{\mathbf{Var}}\{\hat{\tau}_{st}\}} \\ \text{Estimativa da Média: } \hat{\mu}_{st} &\pm t(1 - \alpha/2; d) \sqrt{\widehat{\mathbf{Var}}\{\hat{\mu}_{st}\}} \end{aligned}$$

- O “*valor aproximado*” dos graus de liberdade (d) para distribuição t de Student é

$$d = \frac{\left(\sum_{h=1}^L a_h \hat{\sigma}_h^2\right)^2}{\sum_{h=1}^L (a_h \hat{\sigma}_h^2)^2 / (n_h - 1)}$$

onde $a_h = N_h (N_h - n_h) / n_h$.

3.4 Estimando Proporções

- Para se estimar uma proporção na população alvo através da amostragem estratificada toma-se em cada estrato uma AAS de observações, sendo que se define a seguinte variável aleatória:

$$y_{hi} = \begin{cases} 1 & \text{se sucesso,} \\ 0 & \text{se fracasso.} \end{cases}$$

onde o índice h ($h = 1, 2, \dots, L$) indica o estrato, e o índice i ($i = 1, 2, \dots, n_h$) indica as observações no estrato h .

- As estimativas por estrato são:

$$\begin{aligned} \hat{p}_h &= \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi} \\ \widehat{\mathbf{Var}}\{\hat{p}_h\} &= \frac{\hat{p}_h(1 - \hat{p}_h)}{n_h} \left(1 - \frac{n_h}{N_h}\right) \end{aligned}$$

- E a combinação dessas estimativas para se obter uma estimativa de proporção para população alvo como um todo segue o mesmo raciocínio da estimativa do total:

$$\begin{aligned}\hat{p}_{st} &= \frac{1}{N} \sum_{i=1}^L N_h \hat{p}_h \\ \mathbf{Var}\{\hat{p}_{st}\} &= \frac{1}{N^2} \sum_{i=1}^L N_h^2 \mathbf{Var}\{\hat{p}_h\}\end{aligned}$$

4 *Amostragem Estratificada e Amostragem Aleatória Simples*

- A amostragem estratificada é sempre melhor que a AAS?
- Se os estimadores dos totais (ou média) de cada estrato forem não viciados, o estimador da amostragem estratificada, também será não viciado:

$$E[\hat{\tau}_{st}] = E\left[\sum_{h=1}^L \hat{\tau}_h\right] = \sum_{h=1}^L E[\hat{\tau}_h] = \sum_{h=1}^L \tau_h = \tau$$

- Vejamos o que acontece com a variância da estimativa na amostragem estratificada (sem correção para populações finitas):

$$\mathbf{Var}\{\hat{\tau}_{st}\} = \mathbf{Var}\left\{\sum_{h=1}^L \hat{\tau}_h\right\} = \sum_{h=1}^L \mathbf{Var}\{\hat{\tau}_h\} = \sum_{h=1}^L N_h^2 \frac{\sigma_h^2}{n_h}$$

- No caso da AAS, a variância da estimativa do total é

$$\mathbf{Var}\{\tau\} = N^2 \frac{\sigma^2}{n}.$$

5 Tamanho da Amostra e Alocação das Unidades Amostrais

- Assumindo uma AAS em cada estrato, o tamanho da amostra para erro amostral aceitável E pode ser obtido por:

$$n^* = \frac{\sum_{h=1}^L N_h^2 \hat{\sigma}_h^2 / W_h}{N^2 E^2 / t^2 + \sum_{h=1}^L N_h \hat{\sigma}_h^2}$$

onde:

- N_h é o tamanho do estrato h ;
- $N = \sum N_h$ é o tamanho da população alvo;
- $\hat{\sigma}_h^2$ é a variância amostral no estrato h ;
- E é o erro amostral aceitável (em unidades da média);
- W_h é a proporção de unidades amostrais no estrato h .

- Quando a correção para populações finitas pode ser desprezada, a expressão simplifica para:

$$n^* = \frac{t^2 \sum_{h=1}^L N_h^2 \hat{\sigma}_h^2 / W_h}{N^2 E^2}$$

- Note que para encontrar o tamanho de amostral, é necessário primeiramente definir as proporções W_h , isto é, como as unidades amostrais serão distribuídas nos vários estratos.

5.1 Alocação Proporcional

- Uma primeira forma de alocar as várias unidades amostrais aos estratos é proporcionalmente ao tamanho dos estratos (N_h).

$$W_h = \frac{N_h}{N}$$

5.2 Alocação de Neyman

- Pode-se argumentar que nem sempre os maiores estratos são os que possuem maior variabilidade.

- Assim, pode-se considerar, além do tamanho, a variância populacional:

$$W_h = \frac{N_h \hat{\sigma}_h^2}{\sum_{h=1}^L N_h \hat{\sigma}_h^2}$$

5.3 Alocação Ótima

- Por fim, pode-se argumentar que além do tamanho dos estratos e das suas variabilidades, o custo do levantamento não é o mesmo em todos os estratos.
- Considerando que o esforço amostral deva ser *inversamente* proporcional à raiz quadrada do custo, a alocação ficaria:

$$W_h = \frac{N_h \hat{\sigma}_h^2 / \sqrt{C_h}}{\sum_{h=1}^L N_h \hat{\sigma}_h^2 / \sqrt{C_h}}$$

onde C_h é o custo (ou custo relativo) para se obter uma unidade amostral no estrato h .

6 Pós-Estratificação

- Frequentemente o tamanho dos estratos não é conhecido antes do levantamento de campo, mas pode ser estimado por procedimentos amostrais durante o levantamento de campo.
- Em algumas situações, os estratos são percebidos durante ou após o levantamento de campo e não previamente.
- Principalmente quando se utiliza a amostragem sistemática, é possível realizar uma estratificação “*a posteriori*” das unidades amostrais.
- Nesses casos, a pós-estratificação tende a gerar uma alocação das unidades amostrais próxima à alocação proporcional e isso pode ser utilizado para se “*estimar*” o tamanho dos estratos.

- Se na pós-estratificação o tamanho dos estratos (N_h) são *determinados* sem erro amostral ou com um erro muito pequeno, os estimadores apresentados acima podem ser utilizados.
- Se, no entanto, o tamanho dos estratos (N_h) é *estimado* com erro amostral o delineamento amostral não é mais a amostragem estratificada.

Nesse caso, devemos utilizar os estimadores da amostragem dupla.

7 Exercícios

1. A tabela 1 apresenta dados de um levantamento numa área de floresta ombrófila densa na Amazônia Oriental, onde se utilizou a amostragem estratificada. As parcelas tinha 5.000 m^2 e considera-se aceitável um erro amostral de 10% ao nível de significância de 5%.
 - (a) Encontre o intervalo de confiança de 95% para as médias de todas variáveis observadas.
 - (b) Encontre o tamanho adequado de amostra utilizando alocação proporcional.
 - (c) Encontre o tamanho adequado de amostra utilizando alocação de Neyman.
 - (d) Compare as variáveis medidas em termos dos resultados obtidos.
2. A tabela 2 apresenta dados de um Inventário Florestal em floresta plantada de *Eucalyptus*. Foi realizado uma amostragem aleatória simples em cada um dos seguintes estratos utilizando parcelas de 600 m^2 :
 - Estrato I: *E. grandis*, 1a. rotação: 350 ha;
 - Estrato II: *E. grandis*, 2a. rotação: 150 ha;
 - Estrato III: *E. saligna*, 1a. rotação: 450 ha;
 - Estrato IV: *E. saligna*, 2a. rotação: 50 ha.

Utilizando os estimadores da amostragem estratificada e considerando o erro amostral aceitável de 10% (nível de significância de 5%), encontre:

- (a) Encontre o intervalo de confiança de 95% para as médias de todas variáveis observadas.
- (b) Encontre o tamanho adequado de amostra utilizando alocação proporcional.
- (c) Encontre o tamanho adequado de amostra utilizando alocação de Neyman.
- (d) Compare as variáveis medidas em termos dos resultados obtidos.

Tabela 1: Dados de amostragem estratificada em floresta ombrófila densa na Amazônia Oriental utilizando parcelas de 5.000 m^2 .

| Estrato | Area do Estrato (ha) | Parcela | VARIÁVEIS | | | | |
|---------|-------------------------|---------|------------------------------|-------------------------------|---|---|---|
| | | | N. Árv. (<i>arv/ha</i>) | DAP médio (<i>cm</i>) | Altura Comercial Média (<i>m</i>) | Área Basal (m^2/ha) | Volume Comercial (m^3/ha) |
| A | 73 | 2 | 358 | 24.98 | 7.63 | 21.77 | 198.35 |
| A | 73 | 3 | 332 | 23.98 | 7.49 | 18.73 | 174.54 |
| A | 73 | 4 | 290 | 23.70 | 9.15 | 16.32 | 177.16 |
| B | 50 | 5 | 304 | 23.95 | 8.22 | 16.36 | 162.51 |
| B | 50 | 6 | 268 | 22.11 | 6.90 | 11.52 | 86.89 |
| C | 123 | 10 | 242 | 24.28 | 7.10 | 13.88 | 124.14 |
| C | 123 | 7 | 256 | 22.21 | 7.03 | 12.56 | 118.64 |
| C | 123 | 8 | 244 | 23.20 | 7.51 | 11.99 | 101.49 |
| C | 123 | 9 | 216 | 22.20 | 6.97 | 9.32 | 72.55 |
| D | 154 | 11 | 288 | 24.68 | 7.81 | 17.08 | 161.89 |
| D | 154 | 12 | 322 | 31.13 | 8.63 | 32.13 | 357.82 |
| D | 154 | 13 | 234 | 27.27 | 7.54 | 18.44 | 181.67 |
| D | 154 | 14 | 276 | 24.36 | 7.84 | 15.45 | 141.53 |

3. Foi realizado um levantamento num fragmento florestal de 47 ha , utilizando-se parcelas de 100 m^2 dipostas numa grade amostral quadrada (amostragem sistemática), cada parcela representando 1 ha . A tabela 3 apresenta os dados obtidos. Durante o levantamento, cada parcela foi classificada, por avaliação visual, numa das seguintes “Unidades Estruturais”:

BA: bambuzal; **CA:** capoeira alta;
CB: capoeira baxia; **FM:** floresta madura.

- Encontre os intervalos de confiança de 95% utilizando os estimadores da amostragem sistemática.
- Utilize as “unidades estruturais” para realizar uma *pós-estratificação* das parcelas e encontre os intervalos de confiança de 95% utilizando os estimadores da amostragem estratificada.
- Discuta os resultados.

Tabela 2: Dados de Inventário Florestal em floresta plantada de *Eucalyptus*, com parcelas de 600 m².

| Espécie | Rotação | Parcela | DAP médio (cm) | Altura média (m) | Falha (%) | Volume Comercial (m ³ /ha) |
|------------------|---------|---------|----------------------|------------------------|--------------|---|
| <i>E.grandis</i> | 1 | 1 | 8.89 | 17.09 | 5.93 | 126 |
| | 1 | 2 | 10.70 | 19.13 | 11.03 | 217 |
| | 1 | 3 | 9.85 | 16.58 | 10.14 | 133 |
| | 1 | 4 | 9.40 | 16.42 | 3.73 | 133 |
| | 1 | 5 | 10.04 | 15.67 | 17.42 | 119 |
| | 1 | 6 | 9.23 | 16.41 | 8.33 | 119 |
| | 1 | 7 | 11.33 | 21.55 | 6.87 | 231 |
| | 1 | 8 | 11.96 | 20.18 | 12.78 | 203 |
| | 1 | 9 | 10.96 | 18.35 | 15.15 | 168 |
| | 1 | 10 | 11.23 | 18.09 | 23.02 | 147 |
| <i>E.grandis</i> | 2 | 49 | 7.38 | 14.40 | 3.03 | 182 |
| | 2 | 50 | 8.57 | 15.95 | 14.75 | 208 |
| | 2 | 51 | 6.97 | 14.43 | 7.58 | 156 |
| | 2 | 52 | 8.21 | 14.74 | 13.64 | 247 |
| | 2 | 53 | 7.53 | 14.21 | 12.90 | 195 |
| | 2 | 54 | 8.70 | 16.19 | 18.75 | 247 |
| | 2 | 55 | 8.84 | 17.18 | 9.68 | 221 |
| | 2 | 56 | 8.29 | 13.87 | 25.81 | 156 |
| | 2 | 57 | 7.57 | 14.99 | 33.33 | 143 |
| | 2 | 58 | 8.20 | 15.78 | 24.24 | 156 |
| | 2 | 59 | 8.79 | 16.86 | 33.85 | 156 |
| <i>E.saligna</i> | 1 | 20 | 10.11 | 17.01 | 13.18 | 133 |
| | 1 | 31 | 9.60 | 15.01 | 6.72 | 119 |
| | 1 | 32 | 8.94 | 10.42 | 11.03 | 70 |
| | 1 | 33 | 11.42 | 17.00 | 26.67 | 126 |
| | 1 | 34 | 9.96 | 16.18 | 11.43 | 133 |
| | 1 | 35 | 11.44 | 18.77 | 11.85 | 168 |
| | 1 | 36 | 9.42 | 14.94 | 8.76 | 119 |
| | 1 | 37 | 10.45 | 15.98 | 20.00 | 133 |
| | 1 | 38 | 10.50 | 17.28 | 14.39 | 147 |
| | 1 | 39 | 10.02 | 15.28 | 9.77 | 105 |
| | 1 | 40 | 10.04 | 15.51 | 20.15 | 105 |
| <i>E.saligna</i> | 2 | 71 | 9.73 | 15.63 | 56.45 | 378 |
| | 2 | 72 | 11.33 | 15.97 | 60.61 | 476 |
| | 2 | 74 | 8.14 | 12.19 | 19.67 | 266 |
| | 2 | 75 | 9.26 | 15.38 | 25.00 | 364 |
| | 2 | 76 | 9.19 | 13.61 | 31.75 | 406 |
| | 2 | 77 | 9.21 | 13.55 | 46.97 | 392 |
| | 2 | 78 | 7.35 | 11.57 | 24.19 | 294 |
| | 2 | 79 | 8.64 | 14.17 | 25.00 | 238 |
| | 2 | 80 | 8.79 | 12.61 | 35.38 | 420 |
| | 2 | 81 | 9.53 | 14.22 | 45.45 | 476 |

Tabela 3: Dados de levantamento em fragmento florestal de 47 *ha* com parcelas de 100 *m*².

| Picada | Parcela | Unidade Estrutural | Árvores (<i>arv/ha</i>) | | | Área Basal (<i>m</i> ² / <i>ha</i>) |
|--------|---------|--------------------|---------------------------|----------|---------------|--|
| | | | Mortas | Ingresso | Sobreviventes | |
| A | 1 | CB | 100 | 0 | 100 | 1.16 |
| A | 2 | CB | 100 | 0 | 400 | 2.96 |
| A | 3 | CA | 100 | 0 | 1300 | 8.14 |
| A | 4 | CB | 0 | 0 | 400 | 2.13 |
| A | 5 | CA | 100 | 0 | 900 | 5.35 |
| B | 1 | CA | 0 | 100 | 600 | 22.64 |
| B | 2 | CA | 100 | 300 | 900 | 12.57 |
| B | 3 | CA | 0 | 200 | 400 | 4.80 |
| B | 4 | CA | 0 | 100 | 600 | 40.07 |
| B | 5 | CA | 300 | 100 | 700 | 13.74 |
| B | 6 | CA | 100 | 200 | 300 | 2.98 |
| C | 1 | CB | 0 | 100 | 300 | 1.61 |
| C | 2 | CB | 100 | 0 | 300 | 11.54 |
| C | 3 | CB | 100 | 0 | 600 | 4.96 |
| C | 4 | CA | 0 | 200 | 1000 | 26.50 |
| C | 5 | CA | 0 | 100 | 700 | 5.17 |
| C | 6 | CA | 0 | 600 | 600 | 3.91 |
| C | 7 | CA | 0 | 200 | 600 | 3.51 |
| D | 1 | CA | 0 | 0 | 600 | 5.15 |
| D | 2 | CA | 100 | 100 | 700 | 5.70 |
| D | 3 | CA | 100 | 100 | 1100 | 6.59 |
| D | 4 | CA | 0 | 0 | 600 | 3.26 |
| D | 5 | CA | 100 | 0 | 1500 | 9.81 |
| D | 6 | CA | 0 | 0 | 1500 | 12.45 |
| D | 7 | CA | 0 | 100 | 1500 | 24.00 |
| E | 1 | CA | 0 | 0 | 600 | 28.82 |
| E | 2 | CA | 0 | 0 | 1100 | 30.28 |
| E | 3 | CA | 0 | 0 | 1400 | 13.91 |
| E | 4 | CA | 0 | 0 | 1600 | 21.71 |
| E | 5 | CA | 200 | 300 | 1600 | 7.23 |
| E | 6 | CA | 0 | 100 | 700 | 3.11 |
| E | 7 | CA | 0 | 0 | 1400 | 14.99 |
| E | 8 | CA | 0 | 0 | 1100 | 14.52 |
| E | 9 | CB | 100 | 200 | 600 | 5.06 |
| E | 10 | CB | 100 | 0 | 900 | 14.32 |
| E | 11 | CB | 0 | 100 | 300 | 11.62 |
| F | 1 | BA | 0 | 100 | 900 | 24.99 |
| F | 2 | BA | 300 | 0 | 1300 | 22.49 |
| F | 3 | BA | 0 | 100 | 700 | 15.71 |
| F | 4 | BA | 100 | 100 | 800 | 20.81 |
| F | 5 | BA | 0 | 100 | 1000 | 7.21 |
| F | 6 | CA | 0 | 100 | 1300 | 25.25 |
| F | 7 | CA | 0 | 0 | 200 | 2.72 |
| F | 8 | FM | 100 | 200 | 600 | 13.75 |
| F | 9 | FM | 0 | 200 | 500 | 9.11 |
| F | 10 | FM | 100 | 0 | 800 | 15.49 |
| F | 11 | FM | 0 | 100 | 600 | 21.66 |