Estimadores de Razão e Regressão

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*

- É comum que num levantamento ou inventário florestal exista uma variável principal, mas outras variáveis também sejam observadas.
- Essas outras variáveis podem ser utilizadas para se fazer uma estratificação da floresta, como foi visto na Amostragem Aleatória Estratificada.
- Mas essas variáveis podem ser utilizadas também para melhorara a *estimação* dos parâmetros da variável principal.
- No contexto da estimação, a variável principal é chamada de **variável de interesse** e as demais variáveis observadas de **variáveis auxiliares**.
- Um situação particularmente interessante é quando a variável de interesse é de medição difícil ou custosa, mas a variável auxiliar, que possui uma forte relação com a variável de interesse, é de medição fácil ou de baixo custo.
- Nessa situação, pode-se aumentar a precisão das estimativas relativas à variável de interesse utilizando-se informaçãoes sobre a variável auxiliar e sua relação com a variável de interesse.

- Veremos dois tipos de estimadores com essa finalidade: estimadores de razão e estimadores de regressão.

2 *Estimador de Razão*

2.1 Situações de Uso do Estimador de Razão

Situação I: num "*Inventário Pré-Corte*" de um talhão de floresta plantada de *Eucalyptus*, foram contadas todas as árvores. Nas parcelas de inventário foram medidas duas variáveis:

variável de interesse: volume de madeira na parcela (m^3);

variável auxiliar: número de árvores na parcela.

Situação II: num inventário florestal utilizou-se parcelas em faixa com largura fixa, mas com comprimento variável. A área da floresta é conhecida e deseja-se saber o volume de madeira da floresta. Nas parcelas de inventário foram medidas duas variáveis:

variável de interesse: volume de madeira na por unidade de área ($m^3\ ha^{-1}$);

variável auxiliar: área da parcela (ha).

Situação III: deseja-se saber a biomassa (massa seca) de madeira num pátio de fábrica. Todos os caminhões são pesados na entrada do pátio de modo que a massa verde (madeira + água) de madeira no pátio é conhecida. Alguns caminhões são amostrados e a sua biomassa (massa verde) é estimada através da determinação do teor de umidade das toras.

variável de interesse: biomassa (massa seca em kg);

variável auxiliar: massa verde (kg).

Situação IV: num levantamento das árvores das vias públicas numa cidade deseja-se estimar o número total de árvores. Definiu-se que as unidades amostrais serão os quarteirões e duas variáveis serão observadas:

variável de interesse: número de árvores ao longo das ruas do quarteirão;

variável auxiliar: comprimento das ruas do quarteirão.

- Nessas situações, o total da variável auxiliar é conhecido sem erro amostral, seja por informação cartográficao (situações II e IV), seja via censo (situações I e III).
- Às vezes, a **razão** entre variável de interesse e variável auxiliar pode ser o parâmetro de interesse.
  - Na situação IV, pode-se estar interessado no número de árvores por metro de rua.
  - Em fitossociologia, a *dominância relativa* (*densidade relativa*) de uma espécie é a razão da sua área basal (densidade) pela área basal (densidade) da floresta.

2.2 Notação dos Parâmetros

- Notação e parâmetros populacionais:

  N - tamanho da população;

  Y - variável de interesse;

  X - variável auxiliar;

  τ_Y - total da variável de interesse;

  μ_Y - média da variável de interesse;

  τ_X - total da variável auxiliar;

  μ_X - média da variável auxliar.

- A razão das variáveis define um novo parâmetro:

$$R = \frac{\mu_Y}{\mu_X} = \frac{N\ \mu_Y}{N\ \mu_X} = \frac{\tau_Y}{\tau_X}$$

  O significado do parâmetro da razão depende das variáveis Y e X, mas mesmo quando não há interesse direto na razão podemos concebê-la como um parâmetro que deve ser estimado para se chegar à estimativa de outros parâmetros de interesse (μ_Y ou τ_Y).

2.3 Estimador de Razão na AAS

- Na AAS, a razão populacional pode ser estimada pela razão amostral:

$$\widehat{R} = \frac{\widehat{\mu}_Y}{\widehat{\mu}_X} = \frac{(1/n)\ \sum_{i=1}^n y_i}{(1/n)\ \sum_{i=1}^n x_i} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i}$$

- A estimativa da variância da razão amostral é

$$\widehat{\textbf{\textit{Var}}}\{\widehat{R}\} = \frac{1}{\mu_X^2}\ \frac{s_u^2}{n}\ \left(1 - \frac{n}{N}\right)$$

  onde s_u^2 é a variância entre valores observados e os estimados pelo estimador de razão:

$$s_u^2 = \frac{\sum_{i=1}^n (y_i - \widehat{R}x_i)^2}{n-1} = \frac{\sum_{i=1} y_i^2 + \widehat{R}^2\ \sum_{i=1}^n x_i^2 - 2\ \widehat{R}\ \sum_{i=1}^n x_i y_i}{n-1}$$

- A partir da razão amostral estima-se os parâmetros de interesse:
  - Estimador da Média da Var. de Interesse:

$$\widehat{\mu}_R = \frac{\widehat{\mu}_Y}{\widehat{\mu}_X}\mu_X = \widehat{R}\mu_X = \widehat{R}\frac{\tau_X}{N}$$

    Note que o total (média) da variável auxiliar deve ser conhecido sem erro amostral.
  - Variância da média estimada:

$$\widehat{\textbf{\textit{Var}}}\{\widehat{\mu}_R\} = \frac{s_u^2}{n}\left(1 - \frac{n}{N}\right)$$

    Thompson (1992) argumenta que essa estimativa da variância tende a resultar em altos valores quando os valores amostrais de $\widehat{\mu}_X$ são altos e em baixos valores quando os valores amostrais de $\widehat{\mu}_X$ são baixos. O estimador ajustado para esse problema é:

$$\overline{\widehat{\textbf{\textit{Var}}}\{\widehat{\mu}_R\}} = \left(\frac{\mu_X}{\widehat{\mu}_X}\right)^2\ \widehat{\textbf{\textit{Var}}}\{\widehat{\mu}_R\}$$

  - Estimador do Total da Var. de Interesse:

$$\widehat{\tau}_R = \widehat{R}\ \tau_X$$

  - Variância do total estimado:

$$\widehat{\textbf{\textit{Var}}}\{\widehat{\tau}_R\} = \tau_X^2\ \frac{1}{\mu_X^2}\ \frac{s_u^2}{n}\left(1 - \frac{n}{N}\right) = \frac{s_u^2}{n}\ N\ (N-n)$$

2.4 Propriedades do Estimador de Razão

- Embora as médias amostrais da variável de interesse e da variável auxiliar sejam estimadores não viciados das respectivas médias populacionais, o estimador de razão **é viciado**.

- Tompson mostra que o viés pode ser obtida por:

$$\begin{aligned}
\boldsymbol{Cov}\{\widehat{\mu}_X, \widehat{R}\} &= \boldsymbol{E}\left\{\widehat{\mu}_X\ \widehat{R}\right\} - \boldsymbol{E}\left\{\widehat{\mu}_X\right\}\ \boldsymbol{E}\left\{\widehat{R}\right\} \\
&= \mu_Y - \mu_X\ \boldsymbol{E}\left\{\widehat{R}\right\} \\
\Rightarrow \boldsymbol{E}\left\{\widehat{R}\right\} &= \frac{\mu_Y - \boldsymbol{Cov}\{\widehat{\mu}_X, \widehat{R}\}}{\mu_X} = R - \frac{\boldsymbol{Cov}\{\widehat{\mu}_X, \widehat{R}\}}{\mu_X}
\end{aligned}$$

  Portanto, o estimador de razão **subestima** a razão populacional.

- Para se ter uma noção da magnitude desse viés, é necessário lembrar que:

$$\boldsymbol{Cov}\{\widehat{\mu}_X, \widehat{R}\} \leq \sqrt{\boldsymbol{Var}\{\widehat{\mu}_X\}\ \boldsymbol{Var}\{\widehat{R}\}}.$$

  Logo, o viés do estimador de razão está limitado por:

$$|\boldsymbol{E}\left\{\widehat{R}\right\} - R| \leq \frac{\sqrt{\boldsymbol{Var}\{\widehat{\mu}_X\}\ \boldsymbol{Var}\{\widehat{R}\}}}{\mu_X} \Longrightarrow \frac{|\boldsymbol{E}\left\{\widehat{R}\right\} - R|}{\sqrt{\boldsymbol{Var}\{\widehat{R}\}}} \leq \frac{\sqrt{\boldsymbol{Var}\{\widehat{\mu}_X\}}}{\mu_X}$$

  Conclui-se que o viés relativo ao erro padrão do estimador não é maior que o coeficiente de variação da média amostral de X ($\widehat{\mu}_X$).

- A situação ideal para uso do estimador de razão é quando:

  1. A relação entre Y e X é **linear**.
  2. A reta da relação Y-X passa pela **origem**.
  3. A variância ao redor da reta (ao redor de y_i) é proporcional a X, sendo **crescente**.

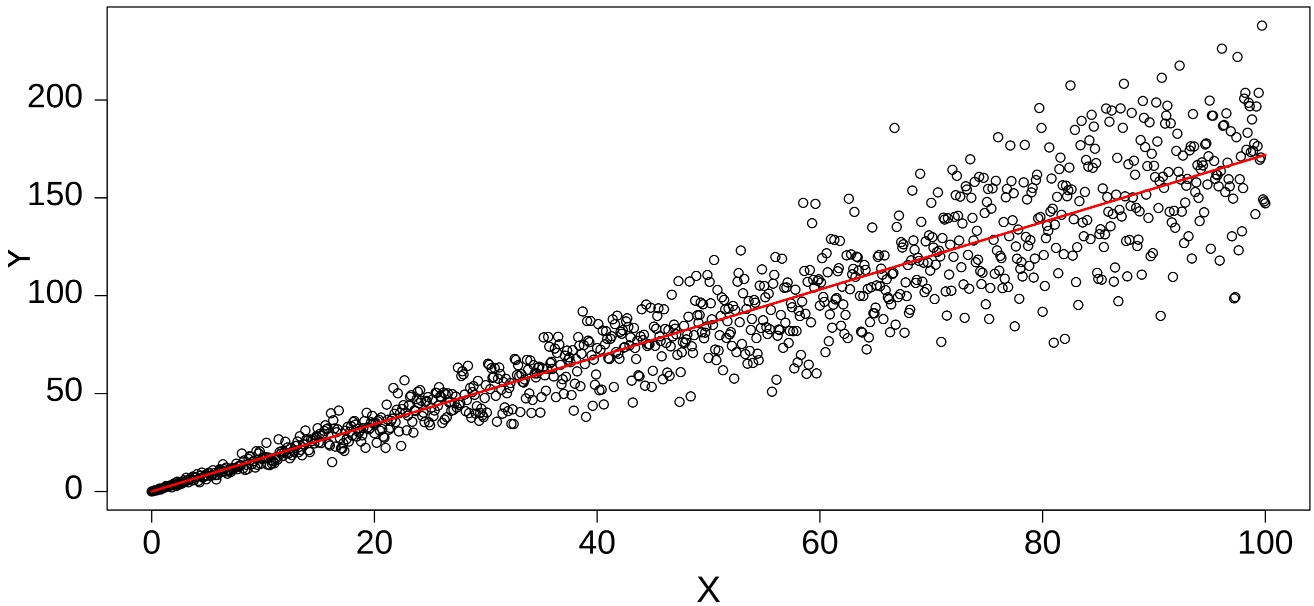

Figura 1: Relação entre a variável auxiliar (X) e variável de interesse (Y) ideal para aplicação do estimador de razão.

2.5 Tamanho de Amostra

E_R é o erro amostral aceitável na estimativa da razão populacional:

$$E_R = t\sqrt{\textbf{\textit{Var}}\{\widehat{R}\}} \quad \Rightarrow \quad \text{[com correção para pop. finita]} \qquad n^\star = \frac{t^2\ s_u^2\ N}{E_R^2\ \mu_X^2\ N + t^2\ s_u^2}$$

$$\Rightarrow \quad \text{[sem correção para pop. finita]} \qquad n^\star = \frac{t^2\ s_u^2}{E_R^2\ \mu_X^2}$$

E_μ é o erro amostral aceitável na estimativa da média populacional:

$$E_\mu = t\sqrt{\textbf{\textit{Var}}\{\widehat{\mu}_R\}} \quad \Rightarrow \quad \text{[com correção para pop. finita]} \qquad n^\star = \frac{t^2\ s_u^2\ N}{E_\mu^2\ N + t^2\ s_u^2}$$

$$\Rightarrow \quad \text{[sem correção para pop. finita]} \qquad n^\star = \frac{t^2\ s_u^2}{E_\mu^2}$$

E_τ é o erro amostral aceitável na estimativa do total populacional:

$$E_\tau = t\,\sqrt{\textbf{Var}\{\widehat{\tau}_R\}} \quad \Rightarrow \quad \text{[com correção para pop. finita]} \qquad n^\star = \frac{t^2\,N^2\,s_u^2}{E_\tau^2 + t^2\,N\,s_u^2}$$

$$\Rightarrow \quad \text{[sem correção para pop. finita]} \qquad n^\star = \frac{t^2\,N^2\,s_u^2}{E_\tau^2}$$

3 *Estimador de Regressão*

- O estimador de regressão é apropriado quando existe uma relação linear entre a variável de interesse (Y) e a variável auxiliar (X), mas essa relação não passa necessáriamente pela origem.
- O estimador de regressão pode ser utilizado inclusive quando a relação entre Y e X é negativamente correlacionada, isto é, á medida que X cresce, Y decresce.
- Assim como no estimador de razão, o estimador de regressão também requer o conhecimento da média populacional ou do total populacional da variável auxiliar.

3.1 Estimadores e Variâncias

- Mesmo na Amostragem Aleatória Simples, o estimador de regressão é *enviesado* (viciado), mas em geral o viés é despresível.
- O estimador de regressão para média populacional é

$$\widehat{\mu}_L = a + b\,\mu_X$$

onde:

$$\begin{aligned} a &= \widehat{\mu}_Y - b\,\widehat{\mu}_X \\ b &= \frac{\sum_{i=1}^n (x_i - \widehat{\mu}_X)(y_i - \widehat{\mu}_Y)}{\sum_{i=1}^n (x_i - \widehat{\mu}_X)^2} = \frac{\sum_{i=1}^n x_i y_i - \left[(\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)\right]/n}{\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2/n} \end{aligned}$$

- No estimador de regressão, a relação entre Y e X é definida por uma reta com inclinação dada pelo o valor de b e intercepto pelo valor de a.

  Os valores de a e b são encontrados ajustando-se a reta aos dados pelo método de quadrados mínimos.

- Se o valor de a for substituido na expressão original temos:

  $$\widehat{\mu}_L = \widehat{\mu}_Y + b\,(\mu_X - \widehat{\mu}_X)$$

  onde o estimador de regressão é expresso somente em função de b.

- A variância do estimador da média é

  $$\widehat{\textbf{\textit{Var}}}\{\widehat{\mu}_L\} = \frac{s^2_{Y \cdot X}}{n} \left(1 - \frac{n}{N}\right)$$

  onde $s^2_{Y \cdot X}$ é a porção da variância de Y que não pode ser explicada pela relação linear com X:

  $$\begin{aligned} s^2_{Y \cdot X} &= \frac{\sum_{i=1}^{n} (y_i - a - bx_i)^2}{n-2} \\ &= \frac{\sum_{i=1}^{n} (y_i - \widehat{\mu}_Y)^2 - b^2\ \sum_{i=1}^{n} (x_i - \widehat{\mu}_X)^2}{n-2} \end{aligned}$$

- Se existe uma forte relação linear entre Y e X o estimador de regressão é muito eficiente.

- O estimador do total populacional é

  $$\widehat{\tau}_L = N\,\widehat{\mu}_L = N\widehat{\mu}_Y + b(\tau_X - N\widehat{\mu}_X).$$

  com variância estimada

  $$\widehat{\textbf{\textit{Var}}}\{\widehat{\tau}_L\} = N^2\ \widehat{\textbf{\textit{Var}}}\{\widehat{\mu}_L\}$$

3.2 Tamanho de Amostra

Sendo E_μ o erro amostral aceitável para média, o tamanho de amostra é

Com correção para População Finita: $$n^\star = \frac{t^2\ s^2_{Y \cdot X}\ N}{N\ E^2_\mu + t^2\ s^2_{Y \cdot X}}$$

Sem correção para População Finita: $$n^\star = \frac{t^2\ s^2_{Y \cdot X}}{E^2_\mu}$$

Problemas e Exercícios

1. Para se realizar um levantamento num fragmento florestal de 23 ha utilizou-se parcelas em faixa com comprimento (e área variável). Os dados observados foram os seguintes:

Parcela	Área da Parcela	Árvores ($arv/parcela$) Mortas	Ingresso	Sobreviventes	Área Basal ($m^2/parcela$)
1	500	4	0	31	19.74
2	600	5	10	35	96.80
3	700	2	12	41	57.20
4	700	3	3	75	66.96
5	1100	4	7	113	165.57
6	1100	6	10	87	179.19

   Sabendo que o tamanho da população é $N = 230$, utilize o estimador de razão e encontre:

   (a) o intervalo de confiança de 95% para o **total** de árvores mortas, de ingresso e sobreviventes no fragmento.

   (b) o intervalo de confiança de 95% para a **média** da área basal (m^2/ha) do fragmento.

   (c) o tamanho de amostra necessário para erro amostral de 10% ao nível de probabilidade de 5% para cada uma das variáveis de interesse.

2. Para realizar o inventário pré-corte de um talhao de 23 ha de *Eucalyptus grandis* foi realizado a contagem de todas árvores no talhão, o que totalizou 30699. Após a medição das parcelas (parcelas de 300 m^2), obteve-se os seguintes dados:

Parcela	Árvores (arv/ha) Total	Com Cancro	Volume (m^3/ha)
2	1482	27	217
7	1552	40	231
8	1453	27	203
9	1414	21	168
10	1283	37	147

(a) Utilizando o estimador de razão, encontre o Intervalo de Confiança de 95% para:
   - volume total de madeira no talhão;
   - taxa de ocorrência de cancro.

(b) Utilizando o estimador convencional da Amostragem Aleatória Simples, encontre o Intervalo de Confiança de 95% para:
   - volume total de madeira no talhão;
   - taxa de ocorrência de cancro.

(c) Compare os intervalos de confiança.

3. Uma análise detalhada do inventário da questão anterior revelou que as parcelas, embora locadas aleatóriamente, possuiam uma cobertura espacial muito restrita. O talhão foi dividido em dois estratos e mais parcelas foram locadas e medidas, resultando nos seguintes dados:

Estrato	Área (ha)	Total de Árvores	Parcela	Árvores (arv/ha) Total	Árvores (arv/ha) Com Cancro	Volume (m^3/ha)
A	13	17352	1	1567	28	126
			3	1497	25	133
			4	1604	30	133
			5	1376	22	119
			6	1527	19	119
B	10	13347	2	1482	27	217
			7	1552	40	231
			8	1453	27	203
			9	1414	21	168
			10	1283	37	147

(a) Utilizando o estimador de razão, encontre o Intervalo de Confiança de 95% para:
   - volume total de madeira no talhão;
   - taxa de ocorrência de cancro.

(b) Utilizando o estimador convencional da Amostragem Aleatória Estratificada, encontre o Intervalo de Confiança de 95% para:
   - volume total de madeira no talhão;
   - taxa de ocorrência de cancro.

(c) Compare os intervalos de confiança.

4. Num levantamento de árvores de vias públicas num dado bairro de Piracicaba, utilizou-se como unidades amostrais o quarteirão e anotou-se as seguintes informações.

Quarteirão	Comprimento Total da Calçada (m)	Número de Árvores
1	444	29
1	451	22
1	408	17
1	524	28
1	202	31
1	270	27
1	354	22

Sabendo que o bairro é constituido por 95 quarteirões e as suas vias públicas totalizam 36.1 km, encontre:

- O número total de árvores de rua no bairro, com o seu respectivo Intervalo de Confiança de 95%.
- A distância média entre as árvores de rua do bairro, com o seu respectivo Intervalo de Confiança de 95%.