Introdução à Amostragem Estratificada (AE)

Prof. Caio Azevedo

19 de setembro de 2011

- A amostragem estratificada (AE) consiste:
  - Na divisão de uma população em grupos (chamados estratos).
  - Esta divisão é feita segundo alguma(s) característica(s) conhecida(s) na população sob estudo.
  - Em cada um desses estratos é selecionado uma amostra, essencialmente, segundo AAS com ou sem reposição, em proporções convenientes.
- Objetivos: produzir estimativas mais precisas, produzir estimativas para a população como um todo e para subpopulações, dentre outras.
- Em geral, quanto mais os elementos de cada estrato forem parecidos entre si e diferentes entre os estratos, maior será a precisão dos estimadores.

Exemplo

- Considere uma pesquisa feita em uma população com N = 8 domicílios, onde são conehcidas as variáveis renda domiciliar (Y) e local de domicílio (W), com os códigos A para a região alta e B para região baixa. Tem-se então:

$$\mathcal{U} = \{1, 2, 3, 4, 5, 6, 7, 8\}$$

$$\mathbf{d} = \begin{pmatrix} \mathbf{y}' \\ \mathbf{w}' \end{pmatrix} = \begin{pmatrix} 13 & 17 & 6 & 5 & 10 & 12 & 19 & 6 \\ B & A & B & B & B & A & A & B \end{pmatrix}$$

- Temos que $\mu = 11$ e $\sigma^2 = 24$.
- Sob AAS_c, tem-se que $\mathcal{V}_{A_1}\left(\overline{Y}\right) = \frac{24}{4} = 6$.

Cont.

- Usando-se a segunda variável para estratificar a população em dois estratos, pode-se construir as seguintes subpopulações:

$$\mathcal{U}_A = \{2, 6, 7\} \quad , \quad \mathbf{d}_A = (17, 12, 19)$$

$$\mathcal{U}_B = \{1, 3, 4, 5, 8\} \quad , \quad \mathbf{d}_B = (13, 6, 5, 10, 6)$$

- Nesse caso, temos que $\mu_A = 16$, $\sigma_A^2 = 8,7$, $\mu_B = 8$, $\sigma_B^2 = 9,2$.

Cont.

- Se sorteamos em cada estrato uma amostra de tamanho n = 2 (AAS_c), tem-se que.

$$\mathcal{V}_{A_1}(\overline{Y}_A) = \frac{8,7}{2} = 4,35\,;\ \mathcal{V}_{A_1}(\overline{Y}_B) = \frac{9,2}{2} = 4,60$$

- Com base em $\overline{Y}_A$ e $\overline{Y}_B$ é preciso construir um estimador para μ.
- Sugestão (média ponderada) $\overline{Y}_{es} = \frac{3\overline{Y}_A + 5\overline{Y}_B}{8}$.
- Nesse caso, temos que

$$\mathcal{E}_{AE}(\overline{Y}_{es}) = \frac{3\mathcal{E}_{A_1}(\overline{Y}_A) + 5\mathcal{E}_{A_1}(\overline{Y}_B)}{8} = \frac{3\mu_A + 5\mu_B}{8} = \mu.$$

- Além disso, $\mathcal{V}_{AE}(Y_{es}) = \frac{9}{64}\mathcal{V}_{A_1}(\overline{Y}_A) + \frac{25}{64}\mathcal{V}_{A_1}(\overline{Y}_B) = 2,4.$
- Nesse caso, $EPA = \frac{\mathcal{V}_{AE}(\overline{Y}_{es})}{\mathcal{V}_{A_1}(Y)} = \frac{2,6}{6,0} = 0,40.$

Observações

- O resultado (estimativa) será mais eficaz quanto maior for a habilidade do pesquisador em produzir estratos homogêneos.
- Se os elementos fossem todos idênticos, dentro de cada estrato, seria o ideal.
- A simples estratificação, por si só, não produz necessariamente estimativas mais eficientes do que a AAS.

Exemplo (estratificação inapropriada)

- Considere amesma população apresentada anteriormente, com a seguinte divisão em estratos.

$$\mathcal{U}_1 = \{1, 2, 3, 4\} \quad , \quad \mathbf{d}_1 = (13, 17, 6, 5)$$
$$\mathcal{U}_2 = \{5, 6, 7, 8\} \quad , \quad \mathbf{d}_2 = (10, 12, 19, 6)$$

- Nesse caso, temos que $\mu_1 = 10,25$, $\sigma_1^2 = 24,69$, $\mu_2 = 11,75$, $\sigma_2^2 = 22,19$.

Cont.

- Se sorteamos em cada estrato uma amostra de tamanho n = 2 (AAS_c), tem-se que.

$$\mathcal{V}_{A_1}(\overline{Y}_1) = \frac{24,69}{2} = 12,34\,;\ \mathcal{V}_{A_1}(\overline{Y}_2) = \frac{22,19}{2} = 11,09$$

- Nesse caso, temos que $\mathcal{E}_{AE}(\overline{Y}_{es}) = \frac{4\mathcal{E}_{A_1}(\overline{Y}_1)+4\mathcal{E}_{A_1}(\overline{Y}_2)}{8} = \frac{4\mu_A+4\mu_B}{8} = \mu$, do mesmo modo.
- Contudo, $\mathcal{V}_{AE}(Y_{es}) = \frac{16}{64}\mathcal{V}_{A_1}(\overline{Y}_1) + \frac{16}{64}\mathcal{V}_{A_1}(\overline{Y}_2) = 5,86$.
- Nesse caso, $EPA = \frac{\mathcal{V}_{AE}(\overline{Y}_{es})}{\mathcal{V}_{A_1}(Y)} = \frac{5,86}{6,0} = 0,98$.

Estrutura da AE

- Divisão da população em subpopulações bem definidas (estratos).
- De cada estrato retira-se uma amostra, usualmente independente.
- Em cada amostra, usam-se estimadores convenientes para os parâmetros de cada estrato.
- Monta-se para a população um estimador, combinando-se os estimadores de cada estrato, e determinam-se suas propriedades.

Notações e relações úteis

- População descrita por um sistema de referência $\mathcal{U} = \{1, 2, ..., N\}$.
- Existe uma partição (estratos) $\mathcal{U}_1 ..., \mathcal{U}_H$ de $\mathcal{U}$, i.e.,

$$\mathcal{U} = \cup_{h=1}^{H} \mathcal{U}_h \, ; \, \mathcal{U}_h \cap \mathcal{U}_{h'} = \emptyset, \forall h \neq h'$$

- Para cada estrato h, temos

$$\mathcal{U}_h = \{(h, 1), (h, 2), ..., (h, N_h)\}$$

- Para a população como um todo, temos

$$\mathcal{U} = \{(1, 1), ..., (1, N_1), ...(h, 1), ..., (h, i), ..., (h, N_h), ..., (H, 1), ..., (H, N_H)\}$$

Notações e relações úteis (cont.)

- Vetor de dados populacionais $\mathbf{d} = (y_{11}, ..., y_{1N_1}, ..., y_{hi}, ...y_{HN_H})$.
- N_h : tamanho do estrato h.
- $\tau_h = \sum_{i=1}^{N_h} y_{hi}$: total do estrato h.
- $\mu_h = \overline{y}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} y_{hi} = \frac{\tau_h}{N_h}$: média do estrato h.
- $s_h^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (y_{hi} - \mu_h)^2$: variância do estrato h.
- $\sigma_h^2 = \frac{1}{N_h} \sum_{i=1}^{N_h} (y_{hi} - \mu_h)^2$: variância do estrato h.
- $W_h = \frac{N_h}{N}$: peso (proporção) do estrato h, $\sum_{h=1}^{H} W_h = 1$.
- $\tau = \sum_{h=1}^{H} \tau_h = \sum_{h=1}^{H} \sum_{i=1}^{N_h} y_{hi} = \sum_{h=1}^{H} N_h \mu_h$: total populacional.
- $\mu = \overline{y} = \frac{\tau}{N} = \frac{1}{N} \sum_{h=1}^{H} \sum_{i=1}^{N_h} y_{hi} = \frac{1}{N} \sum_{h=1}^{H} N_h \mu_h = \sum_{h=1}^{H} W_h \mu_h$: média populacional.

Notações e relações úteis (cont.)

- $\sigma^2 : \frac{1}{N}\sum_{h=1}^{H}\sum_{i=1}^{N_h}(y_{hi}-\mu)^2 = \sum_{h=1}^{H} W_h\sigma_h^2 + \sum_{h=1}^{H} W_h(\mu_h-\mu)^2$ variância populacional (veja páginas 97 e 98).
- Podemos escrever $\sigma^2 = \sigma_d^2 + \sigma_e^2$, $\sigma_d^2 = W_h\sigma_h^2$ e $\sigma_e^2 = \sum_{h=1}^{H} W_h(\mu_h-\mu)^2$.
- $s^2 = \frac{1}{N-1}\sum_{h=1}^{H}\sum_{i=1}^{N_h}(y_{hi}-\mu)^2 = \sum_{h=1}^{H}\frac{N_h-1}{N-1}s_h^2 + \sum_{h=1}^{H}\frac{N_h}{N-1}(\mu_h-\mu)^2$.
- Para estratos (população) relativamente grandes $s^2 \approx \sigma^2$.