Amostragem Estratificada (AE): parte 1

Prof. Caio Azevedo

- A amostragem estratificada (AE ou A_3) 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 conhecidas 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:

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

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

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

Cont.

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

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

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

- 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): $\overline{Y}_{es} = \frac{3\overline{Y}_A + 5\overline{Y}_B}{8}$.

Cont.

- 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, $\mathcal{V}_{AE}(Y_{es}) = \frac{9}{64}\mathcal{V}_{A_1}(\overline{Y}_A) + \frac{25}{64}\mathcal{V}_{A_1}(\overline{Y}_B) = 2,4.$
- Assim, temos que, $EPA = \frac{\mathcal{V}_{AE}(\overline{Y}_{es})}{\mathcal{V}_{A_1}(Y)} = \frac{2,6}{6,0} = 0,40$. Portanto, é mais apropriado utilizar o plano AE (com AAS_c dentro de cada estrato) do que o plano AAS_c.

Observações

- O resultado (estimativa, intervalo de confianca, teste de hipótese) será melhor 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 (com/sem reposição).

Exemplo (estratificação inapropriada)

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

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

- 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, obtemos $\mathcal{E}_{AE}(\overline{Y}_{es}) = \frac{4\mathcal{E}_{A_1}(\overline{Y}_1)+4\mathcal{E}_{A_1}(\overline{Y}_2)}{8} = \frac{4\mu_A+4\mu_B}{8} = \mu$, do mesmo modo.
- Contudo, $\mathcal{V}_{AE}(Y_{es}) = \frac{16}{64}\mathcal{V}_{A_1}(\overline{Y}_1) + \frac{16}{64}\mathcal{V}_{A_1}(\overline{Y}_2) = 5,86$.
- Assim, $EPA = \frac{\mathcal{V}_{AE}(\overline{Y}_{es})}{\mathcal{V}_{A_1}(Y)} = \frac{5{,}86}{6{,}00} = 0,98$. Logo, diferentemente do caso anterior, não há ganho em se utilizar AE em detrimento à AAS_c.

Estrutura da AE

- Divisão da população em subpopulações bem definidas (estratos).
- De cada estrato retira-se uma amostra, usualmente independente (sob AAS_c ou AAS_s) entre os estratos.
- Em cada amostra usam-se estimadores convenientes para estimar os parâmetros de cada estrato.
- Constroi-se, para o parâmetro de interesse, um estimador, combinando-se os estimadores de cada estrato, determinando-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.,

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

Notações e relações úteis

- Para cada estrato h, temos

$$
\begin{aligned}
\mathcal{U}_h &= \{(h,1),(h,2),...,(h,N_h)\} \\
&\Big\{\text{(índice do estrato;} \\
&\text{índice da unidade populacional, dentro do estrato)}\Big\}
\end{aligned} \tag{1}
$$

- 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 $\boldsymbol{d} = (y_{11}, ..., y_{1N_1}, ..., y_{hi}, ...y_{HN_H})^T$.
- N_h : tamanho do estrato h.
- y_{hi} : valor da variável de interesse da unidade amostral i no estrato h, $i = 1, ..., N_h$.
- Y_{hi} : variável aleatória que representa o i-ésimo elemento sorteado no estrato h, $i = 1, ..., n_h$.
- $\tau_h = \sum_{i=1}^{N_h} y_{hi}$: 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.

Notações e relações úteis (cont.)

- $s_h^2 = \frac{1}{N_h - 1}\sum_{i=1}^{N_h}(y_{hi} - \mu_h)^2$: 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, do livro - texto).
- Podemos escrever $\sigma^2 = \sigma_d^2 + \sigma_e^2$, em que $\sigma_d^2 = \sum_{h=1}^{H} W_h\sigma_h^2$ e $\sigma_e^2 = \sum_{h=1}^{H} W_h(\mu_h-\mu)^2$.
- $s^2 = \frac{1}{N-1}\sum_{h=1}^{H}\sum_{i=1}^{N_h}(y_{hi}-\mu)^2 =$ $\sum_{h=1}^{H}\frac{N_h-1}{N-1}s_h^2 + \sum_{h=1}^{H}\frac{N_h}{N-1}(\mu_h-\mu)^2 = s_{d'}^2 + s_{e'}^2$, $s_{d'}^2 = \sum_{h=1}^{H}\frac{N_h-1}{N-1}s_h^2$, $s_{e'}^2 = \sum_{h=1}^{H}\frac{N_h}{N-1}(\mu_h-\mu)^2$.

Notações e relações úteis (cont.)

- Para estratos (população) relativamente grandes $s^2 \approx \sigma^2$.
- Os elementos constantes nas decomposições (acima descritas) das variâncias correspondem a variâncias dentro de cada estrato e entre os estratos.

Cenário geral para estimação de parâmetros

- Considere os seguintes elementos:
  - Uma população estratificada como na seção anterior.
  - De cada estrato sorteia-se, de forma independente entre os estratos, uma amostra de tamanho n_h, podendo ou não ter sido usado o mesmo plano amostral dentro de cada estrato.
  - Vamos considerar que dentro de cada estrato, em princípio, sorteia-se as respectivas amostras via AAS_c ou AAS_s.

Estimação da média

- Assuma o seguinte estimador $\widehat{\mu}_{es} = \sum_{h=1}^{H} W_h \widehat{\mu}_h$, lembrando que $W_h = \frac{N_h}{N}$ e $\widehat{\mu}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} Y_{hi}$, $h = 1, 2, .., H$.
- Notação AE_i: amostragem estratificada, utilizando o plano $A_i, i = 1, 2$, dentro de cada estrato.
- Propriedades do estimador:
  - Esperança

$$
\begin{aligned}
\mathcal{E}_{AE_i}\left(\widehat{\mu}_{es}\right) &= \mathcal{E}_{AE_i}\left(\sum_{h=1}^{H} W_h \widehat{\mu}_h\right) = \sum_{h=1}^{H} W_h \mathcal{E}_{A_i}\left(\widehat{\mu}_h\right) \\
&= \sum_{h=1}^{H} W_h \mu_h = \mu
\end{aligned}
$$

Estimação da média

- Também, temos que ((*) pela independência entre os estimadores da média de cada estrato)

$$\mathcal{V}_{AE_i}(\widehat{\mu}_{es}) = \mathcal{V}_{AE_i}\left(\sum_{h=1}^{H} W_h \widehat{\mu}_h\right) \underbrace{=}_{(*)} \sum_{h=1}^{H} W_h^2 \mathcal{V}_{A_i}\left(\widehat{\mu}_h\right)$$

- Logo, $\mathcal{V}_{AE_1}(\widehat{\mu}_{es}) = \sum_{h=1}^{H} W_h^2 \mathcal{V}_{A_1}\left(\widehat{\mu}_h\right) = \sum_{h=1}^{H} W_h^2 \frac{\sigma_h^2}{n_h}$ (ou seja, sob AAS_c dentro de cada estrato).

Estimação da média

- Portanto, $\mathcal{V}_{AE_i}(\widehat{\mu}_{es}) = \sum_{h=1}^{H} W_h^2 \mathcal{V}_{A_2}(\widehat{\mu}_h) = \sum_{h=1}^{H} W_h^2 (1 - f_h) \frac{s_h^2}{n_h}$, $f_h = \frac{n_h}{N_h}$, (ou seja, sob AAS_s dentro de cada estrato).
- Por argumentos análogos àqueles apresentados para a AAS, temos que selecionar as amostras sob AAS_s em cada estrato, é melhor do que fazê-lo via AAS_c.

Estimação da média

- Assim, estimadores não viciados (e consistentes) para a variância do estimador, sob cada um dos planos amostrais, são dados, respectivamente, por:

$$
\begin{aligned}
\mathcal{V}_{AE_1}(\widehat{\mu}_{es}) &= \sum_{h=1}^{H} W_h^2 \frac{\widehat{\sigma}_h^2}{n_h} \ (AAS_c) \\
\mathcal{V}_{AE_2}(\widehat{\mu}_{es}) &= \sum_{h=1}^{H} W_h^2 (1 - f_h) \frac{\widehat{s}_h^2}{n_h} \ (AAS_s)
\end{aligned}
$$

em que $\widehat{\sigma}_h^2 = \widehat{s}_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (Y_{hi} - \widehat{\mu}_h)^2$ e as respectivas estimativas são calculadas da forma usual.

Alocação da amostra pelos estratos

- O processo de distribuição das n unidades (amostra) pelos estratos chama-se alocação da amostra.
- Tal distribuição é muito importante, pois ela pode ajudar a melhorar a precisão do procedimento amostral (inferência) como veremos a seguir.

Alocação da amostra pelos estratos

- Considere, novamente, o exemplo anterior, agora com a seguinte estratificação:

$$\mathcal{U}_1 = \{2, 4, 7\}, \text{ com } \boldsymbol{d}_1 = (17, 5, 19)$$

$$\mathcal{U}_2 = \{1, 3, 5, 6, 8\}, \text{ com } \boldsymbol{d}_2 = (13, 6, 10, 12, 6)$$

com os seguintes parâmetros populacionais, $\mu_1 \approx 13,7$, $s_1^2 \approx 57,3$, $\mu_2 = 9,4$, $s_2^2 = 10,8$

Alocação da amostra pelos estratos

- Considere duas situações:
  1. (AL_1) : Em ambos os estratos usou-se AAS_s com $n_1 = 1$ e $n_2 = 2$ (n = 3).
  2. (AL_2): Análogo à AL_1, com $n_1 = 2$ e $n_2 = 1$ (n=3).
- Dessa forma, temos que

$$\mathcal{V}_{AL_1}(\widehat{\mu})_{es} = \left(\frac{3}{8}\right)^2 \left(1 - \frac{1}{3}\right) \frac{57,3}{1} + \left(\frac{5}{8}\right)^2 \left(1 - \frac{2}{5}\right) \frac{10,8}{2} \approx 6,64$$

$$\mathcal{V}_{AL_2}(\widehat{\mu})_{es} = \left(\frac{3}{8}\right)^2 \left(1 - \frac{2}{3}\right) \frac{57,3}{2} + \left(\frac{5}{8}\right)^2 \left(1 - \frac{1}{5}\right) \frac{10,8}{1} \approx 4,72$$

Alocação da amostra pelos estratos

- Comparando-se as variâncias (efeito de alocação), temos que:

$$\frac{\mathcal{V}_{AL_1}(\widehat{\mu}_{es})}{\mathcal{V}_{AL_2}(\widehat{\mu}_{es})} \approx \frac{6,64}{4,72} \approx 1,41$$

- Portanto, a segunda alocação reduz a variância e, portanto, faz-se mister alocar as unidades amostrais da forma mais apropriada possível

Alocação da amostra pelos estratos

- Fatores a serem considerados no processo de alocação
  - Quanto maior for a variância do estrato, maior deve ser o respectivo tamanho amostral (n_h), consequentemente, $(f_h = \frac{n_h}{N_h})$.
  - O tamanho da amostra de cada estrato, entretanto, deve ser balanceado com o tamanho do estrato, ou seja $W_h = \frac{N_h}{N}$.
  - Os desenvolvimento serão feitos supondo-se AAS_c dentro de cada estrato. Exercício: repetir os procedimentos sob AAS_s dentro de cada estrato.

Alocação da amostra pelos estratos

- Tipos de alocação:
  - Alocação proporcional (AP): considera o tamanho de cada estrato assim, de estratos maiores, serão selecionadas amostras de maior tamanho.
  - Alocação uniforme (AU): considera-se (aproximadamente) o mesmo tamanho de amostra em cada estrato.
  - Alocação ótima de Neyman (AON): busca-se otimizar uma função objetivo que considera restrições de interesse.

Alocação proporcional

- A AP consiste em determinar o tamanho da amostra da seguinte forma

$$n_h = nW_h = n\frac{N_h}{N}, f_h = \frac{nW_h}{N_h} = \frac{n}{N}$$

- Este procedimento, as vezes, é chamado de "amostragem representativa".
- Usaremos a nomenclatura Amostragem Estratificada Proprocional (AEpr).

Alocação proporcional

- Sob AEpr a variância do estimador da média, sob seleção via A_1 e A_2 em cada estrato, ficam, respectivamente, iguais à:

$$\mathcal{V}_{AE_1(pr)}(\widehat{\mu}_{es}) = \sum_{h=1}^{H} W_h^2 \frac{\sigma_h^2}{nW_h} = \sum_{h=1}^{H} W_h \frac{\sigma_h^2}{n} = \frac{\sigma_d^2}{n}$$

$$\mathcal{V}_{AE_2(pr)}(\widehat{\mu}_{es}) = \sum_{h=1}^{H} W_h^2 (1 - f_h) \frac{s_h^2}{nW_h} = \sum_{h=1}^{H} W_h (1 - f_h) \frac{s_h^2}{n}$$

Alocação proporcional

- Com respectivos estimadores dados por:

$$\widehat{\mathcal{V}}_{AE_1(pr)}(\widehat{\mu}_{es}) = \sum_{h=1}^{H} W_h \frac{\widehat{\sigma}_h^2}{n} = \frac{\widehat{\sigma}_d^2}{n}, \widehat{\sigma}_d^2 = \sum_{h=1}^{H} W_h \widehat{\sigma}_h^2$$

$$\widehat{\mathcal{V}}_{AE_2(pr)}(\widehat{\mu}_{es}) = \sum_{h=1}^{H} W_h (1 - f_h) \frac{\widehat{s}_h^2}{n}$$

Alocação uniforme

- A AU consiste em determinar o tamanho da amostra da seguinte forma

$$n_h = \frac{n}{H} = k, f_h = \frac{k}{N_h}$$

- Usaremos a nomenclatura Amostragem Estratificada Uniforme (AEun).

Alocação proporcional

- Sob AEun a variância do estimador da média, sob seleção via A_1 e A_2 em cada estrato, ficam, respectivamente, iguais à:

$$\mathcal{V}_{AE_1(un)}(\widehat{\mu}_{es}) = \sum_{h=1}^{H} W_h^2 \frac{\sigma_h^2}{k}$$

$$\mathcal{V}_{AE_2(un)}(\widehat{\mu}_{es}) = \sum_{h=1}^{H} W_h^2 (1 - f_h) \frac{s_h^2}{k}$$

Alocação proporcional

- Com respectivos estimadores dados por:

$$\widehat{\mathcal{V}}_{AE_1(pr)}(\widehat{\mu}_{es}) = \sum_{h=1}^{H} W_h^2 \frac{\widehat{\sigma}_h^2}{k}$$

$$\widehat{\mathcal{V}}_{AE_2(pr)}(\widehat{\mu}_{es}) = \sum_{h=1}^{H} W_h^2 (1 - f_h) \frac{\widehat{s}_h^2}{k}$$

Alocação ótima de Neyman

- A AON consiste em minimizar alguma função custo de interesse.
- Considere, assim:

$$C = c_0 + \sum_{h=1}^{H} c_h n_h \rightarrow C' = C - c_0 = \sum_{h=1}^{H} c_h n_h \qquad (2)$$

em que c_0 é um custo inicial, c_h é o custo por unidade observada no estrato h e C' é um custo variável.

- Definamos, ainda, a notação $V_{es} = \mathcal{V}_{AE_1}(\widehat{\mu}_{es}) = \sum_{h=1}^{H} W_h^2 \frac{\sigma_h^2}{n_h}$
- Exercício: repetir dos desenvolvimentos para seleção AAS_s em cada estrato.

Alocação ótima de Neyman

- O problema consiste em minimizar V_{es} para C' fixo, ou minimizar C' para um V_{es} fixo, em n_h, $h = 1, 2, ..., H$.
- Portanto, o problema consiste em minimizar (é possível mostrar isso através da utilização de multiplicadores de Lagrange)

$$V_{es}C' = \left(\sum_{h=1}^{H} W_h^2 \frac{\sigma_h^2}{n_h}\right)\left(\sum_{h=1}^{H} c_h n_h\right) \tag{3}$$

- Contudo, o produto acima nos remete a desigualdade de Cauchy-Schwarz.

Alocação ótima de Neyman

- A desigualdade acima, por outro lado, nos diz que

$$\left(\sum_{h=1}^{H} a_h^2\right)\left(\sum_{h=1}^{H} b_h^2\right) \geq \left(\sum_{h=1}^{H} a_h b_h\right)^2$$

e que a igualdade só obtida se $\frac{b_h}{a_h} = k, \forall h$

- Fazendo $a_h = \frac{W_h \sigma_h}{\sqrt{n_h}}$ e $b_h = \sqrt{c_h n_h}$ em (3), obtem-se que (3) é mínimo, quando

$$\frac{b_h}{a_h} = \frac{\sqrt{c_h n_h}}{\frac{W_h \sigma_h}{\sqrt{n_h}}} = \frac{n_h \sqrt{c_h}}{W_h \sigma_h} = k$$

em que k é uma constante.

Alocação ótima de Neyman

- Tem-se, então, que (3) é mínimo quando

$$n_h = k\frac{W_h\sigma_h}{\sqrt{c_h}}. \tag{4}$$

- Contudo, como $\sum_{h=1}^{H} n_h = n$, tem-se, de (4), que

$$k = \frac{n}{\sum_{h=1}^{H} W_h\sigma_h/\sqrt{c_h}}. \tag{5}$$

Alocação ótima de Neyman

- Logo, de (5) em (4), vem que:

$$n_h = n\frac{W_h\sigma_h/\sqrt{c_h}}{\sum_{h=1}^{H} W_h\sigma_h/\sqrt{c_h}} = n\frac{N_h\sigma_h/\sqrt{c_h}}{\sum_{h=1}^{H} N_h\sigma_h/\sqrt{c_h}}$$

que é a forma de se alocar as unidades amostrais, ao longo dos estratos.

- Adicionalmente, para C' fixado, o tamanho ótimo da amostra é dado por (exercício):

$$n = C'\frac{N_h\sigma_h/\sqrt{c_h}}{\sum_{h=1}^{H} N_h\sigma_h/\sqrt{c_h}}$$

Alocação ótima de Neyman

- Para V_{es} fixado, o tamanho ótimo da amostra é dado por (exercício):

$$n = \frac{1}{V_{es}} \left(\sum_{h=1}^{H} W_h \sigma_h \sqrt{c_h} \right) \left(\sum_{h=1}^{H} \frac{W_h \sigma_h}{\sqrt{c_h}} \right)$$

- Para o caso em que o custo por unidade observada em todos os estratos é fixado em c, ou seja, $C' - c_o = nc$, a alocação ótima se reduz à

$$n_h = n \frac{N_h \sigma_h}{\sum_{h=1}^{H} N_h \sigma_h}, h = 1, 2, ..., H \tag{6}$$

Alocação ótima de Neyman

- Neste caso, V_{es} reduz-se a

$$V_{ot} = \frac{1}{n}\left(\sum_{h=1}^{H} W_h \sigma_h\right)^2 = \frac{\overline{\sigma}^2}{n}$$

- A alocação descrita em (6) é conhecida como alocação ótima de Neyman.

Efeito do planejamento

- Vamos comparar o desempenho do estimador para a média sob AE, em que as amostras em cada estrato são selecionadas segundo um plano AAS_c, com o plano AAS_c, individualmente.
- A prova, substituindo-se AAS_c por AAS_s, fica como exercício.
- Temos que $V_c = \mathcal{V}_{A_1}\left(\widehat{\mu}\right) = \frac{\sigma^2}{n}$, $V_{pr} = \frac{1}{n}\sum_{h=1}^{H} W_h\sigma_h^2 = \frac{\sigma_d^2}{n}$ e $V_{ot} = \frac{1}{n}\left(\sum_{h=1}^{H} W_h\sigma_h\right)^2 = \frac{\overline{\sigma}^2}{n}$
- Vamos provar que $V_{ot} \leq V_{pr} \leq V_c$.

Efeito do planejamento

- Temos que

$$
\begin{aligned}
\sigma^2 &= \frac{1}{N}\sum_{i=1}^{N}(Y_i-\mu)^2 = \frac{1}{N}\sum_{h=1}^{H}\sum_{i=1}^{N_h}(Y_{hi}-\mu)^2 \\
&= \frac{1}{N}\sum_{h=1}^{H}\sum_{i=1}^{N_h}\left([Y_{hi}-\mu_h]+[\mu_h-\mu]\right)^2 \\
&= \frac{1}{N}\sum_{h=1}^{H}\sum_{i=1}^{N_h}(Y_{hi}-\mu_h)^2 + \frac{1}{N}\sum_{h=1}^{H}N_h\left(\mu_h-\mu\right)^2 \\
&= \frac{1}{N}\sum_{h=1}^{H}\sum_{i=1}^{N_h}N_h\frac{(Y_{hi}-\mu_h)^2}{N_h} + \frac{1}{N}\sum_{h=1}^{H}N_h\left(\mu_h-\mu\right)^2 \\
&= \sum_{h=1}^{H}\sum_{i=1}^{N_h}W_h\sigma_h^2 + \sum_{h=1}^{H}W_h\left(\mu_h-\mu\right)^2 = \sigma_d^2+\sigma_e^2
\end{aligned}
$$

Efeito do planejamento

- Assim, do resultado anterior, temos que

$$
\begin{aligned}
\frac{\sigma^2}{n} &= \frac{\sigma_d^2}{n} + \frac{\sigma_e^2}{n} \rightarrow V_c = V_{pr} + \frac{\sigma_e^2}{n} \\
&\rightarrow V_c \geq V_{pr}
\end{aligned} \tag{7}
$$

pois $\frac{\sigma_e^2}{n} \geq 0$. Por outro lado, por construção $V_{ot} \leq V_{pr}$. Além disso, podemos provar (Exercício), que

$$
\begin{aligned}
V_{pr} - V_{ot} &= \frac{1}{n}\left\{\sum_{h=1}^{H} W_h\sigma_h^2 - \left(\sum_{h=1}^{H} W_h\sigma_h\right)^2\right\} \\
&= \frac{1}{n}\sum_{h=1}^{H} W_h\left(\sigma_h - \overline{\sigma}\right)^2 = \frac{\sigma_{dp}^2}{n}
\end{aligned} \tag{8}
$$

Efeito do planejamento

- Em que $\overline{\sigma} = \sum_{h=1}^{H} W_h \sigma_h$ que, juntamente com σ^2_{dp}, indica a variablidade entre os desvios-padrão dos estratos.
- Quando maior for a heterogeneidade dos dados pelos estraos, com mais ênfase recomenda-se o uso da alocação ótima.
- Portanto, de (8) em (7), temos que

$$V_c = V_{pr} + \frac{\sigma^2_e}{n} = V_{ot} + \frac{\sigma^2_e}{n} + \frac{\sigma^2_{dp}}{n}$$

Efeito do planejamento

- Sempre que os estratos tiverem médias distinstas (σ_e^2 grande), deve-se usar alocação proporcional, ou ótima.
- Se, além disso, também os desvios-padrão de cada estrato diferirem muito entre si (σ_{dp}^2 grande), recomenda-se a alocação ótima.

Alocação proporcional $\times$ AAS_c

- Temos que

$$
\begin{aligned}
EPA(AE_1(pr1)) &= \frac{\mathcal{V}_{AE_1(pr1)}(\widehat{\mu}_{es})}{\mathcal{V}_{A1}(\widehat{\mu})} = \frac{V_{pr}}{V_c} \\
&= \frac{\sigma^2/n - \sigma_e^2/n}{\sigma^2/n} = 1 - \frac{\sigma_e^2}{\sigma^2}
\end{aligned}
$$

- Assim, quanto maior for σ_e^2 (variabilidade entre os estratos), maior será o ganho da AEpr1 em relação à AAS_c. Ou seja, quanto melhor for o processo de estratificação, maior será tal ganho.
- Por outro lado, quanto maior for o tamanho da amostra (n) menor tenderá a ser o ganho da AEpr1 em relação à AAS_c.
- Neste caso, $AE_1(pr1)$ é sempre melhor que AAS_c.

Alocação ótima $\times$ AAS_c

- Temos que

$$
\begin{aligned}
EPA(AE_1(ot1)) &= \frac{\mathcal{V}_{AE_1(ot1)}(\widehat{\mu}_{es})}{\mathcal{V}_{A1}(\widehat{\mu})} = \frac{V_{ot}}{V_c} \\
&= \frac{\sigma^2/n - \sigma_e^2/n - \sigma_{dp}^2/n}{\sigma^2/n} = 1 - \frac{\sigma_e^2}{\sigma^2} - \frac{\sigma_{dp}^2}{\sigma^2}
\end{aligned}
$$

- Contudo, neste caso, o ganho em se fazer uma estratificação apropriada tender a ser maior, devido ao termo $-\frac{\sigma_{dp}^2}{\sigma^2}$.
- Por outro lado, quanto maior for o tamanho da amostra (n), menor tenderá a ser o ganho da $AE_1(ot)$ em relação à AAS_c.

Alocação ótima $\times$ AAS_c

- Contudo, neste caso, o ganho em se fazer uma estratificação apropriada tender a ser maior, devido ao termo $-\frac{\sigma^2_{dp}}{\sigma^2}$.
- Por outro lado, quanto maior for o tamanho da amostra (n), menor tenderá a ser o ganho da AEot1 em relação à AAS_c.

Observações

- Estes resultados indicam que, a não ser em circunstâncias bem específicas, a AE tende a levar à melhores resultados do que a AAS.
- Em geral, não é fácil medir o EPA[AE].
- Com efeito, temos que:

$$EP(AE_1) = \frac{\mathcal{V}_{AE_1}(\widehat{\mu}_{es})}{\mathcal{V}_{A_1}(\widehat{\mu})} = \frac{\sum_{h=1}^{H} W_h^2 \sigma_h^2 / n_h}{\sigma^2 / n} = \sum_{h=1}^{H} \frac{W_h}{\omega_h} \left(\frac{\sigma_h}{\sigma}\right)^2$$

  em que $\omega_h = n_h/n$.

- Em linhas gerais, se a estratificação for apropriada, $\sigma_h/\sigma < 1$ e, por este termo estar elevado ao quadrado, ele pode "anular" o efeito de $\omega_h = n_h/n$ (note que $n_h < n$).