Amostragem Estratificada (AE): parte 3

Prof. Caio Azevedo

Estimação da proporção

- Mesma estrutura relativa à estimação da média sob AE, com algumas adaptações.
- 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, $i = 1, ..., N_h$, , em que $y_{hi} = 1$, se sucesso e 0, caso contrário.
- Y_{hi} : variável aleatória que representa o i-ésimo elemento sorteado no estrato h, $i = 1, ..., n_h$.
- $p_h = \frac{1}{N_h} \sum_{i=1}^{N_h} y_{hi}$: proporção do estrato h.

Notações e relações úteis (cont.)

- $\sigma_h^2 = \frac{1}{N_h}\sum_{i=1}^{N_h}(y_{hi} - p_h)^2 = p_h q_h$: variância do estrato h, $q_h = 1 - p_h$.
- $s_h^2 = \frac{1}{N_h - 1}\sum_{i=1}^{N_h}(y_{hi} - p_h)^2 = \frac{N_h}{N_h - 1}\sigma_h^2 = \frac{N_h}{N_h - 1} p_h q_h$: variância do estrato h.
- $W_h = \frac{N_h}{N}$: peso (proporção) do estrato h, $\sum_{h=1}^{H} W_h = 1$.
- $p = \frac{1}{N}\sum_{h=1}^{H}\sum_{i=1}^{N_h} y_{hi} = \frac{1}{N}\sum_{h=1}^{H} N_h p_h = \sum_{h=1}^{H} W_h p_h$: proporção populacional.

Notações e relações úteis (cont.)

- $\sigma^2 = \frac{1}{N}\sum_{h=1}^{H}\sum_{i=1}^{N_h}(y_{hi}-p)^2 = \sum_{h=1}^{H} W_h\sigma_h^2 + \sum_{h=1}^{H} W_h(p_h-p)^2$ $= = \sum_{h=1}^{H} W_h p_h q_h + \sum_{h=1}^{H} W_h(p_h-p)^2$.
- Podemos escrever $\sigma^2 = \sigma_d^2 + \sigma_e^2$, em que $\sigma_d^2 = \sum_{h=1}^{H} W_h p_h q_h$ e $\sigma_e^2 = \sum_{h=1}^{H} W_h(p_h-p)^2$.

Notações e relações úteis (cont.)

- $s^2 = \frac{1}{N-1}\sum_{h=1}^{H}\sum_{i=1}^{N_h}(y_{hi}-p)^2 = \sum_{h=1}^{H}\frac{N_h-1}{N-1}s_h^2 + \sum_{h=1}^{H}\frac{N_h}{N-1}(p_h-p)^2 = s_{d'}^2 + s_{e'}^2$,

$$s_{d'}^2 = \sum_{h=1}^{H}\frac{N_h-1}{N-1}s_h^2 = \sum_{h=1}^{H}\frac{N_h-1}{N-1}\frac{N_h}{N_h-1}\sigma_h^2 = \sum_{h=1}^{H}\frac{N}{N-1}\sigma_h^2 = \sum_{h=1}^{H}\frac{N_h}{N-1}p_h q_h,$$

$$s_{e'}^2 = \sum_{h=1}^{H}\frac{N_h}{N-1}(p_h-p)^2$$

.

Estimador proposto

- Basicamente, o estimador considerado, é aquele utilizado para a média, em que a variável resposta (observada) é binária.
- Ademais, os resultados obtidos para a estimação da proporção, sob AAS_c e AAS_s, serão utilizados.
- Estimador proposto: $\widehat{p}_{es} = \sum_{h=1}^{H} W_h \widehat{p}_h$, em que $\widehat{p}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} Y_{hi}$.
- Segue-se, então, que $\mathcal{E}_{AE_i}\left(\widehat{p}_{es}\right) = \mathcal{E}\left(\sum_{h=1}^{H} W_h \widehat{p}_h\right) = \sum_{h=1}^{H} W_h \mathcal{E}\left(\widehat{p}_h\right) = \sum_{h=1}^{H} W_h p_h = p$.

Estimador proposto

- De forma análoga,

$$\mathcal{V}_{AE_1}(\widehat{p}_{es}) = \sum_{h=1}^{H} W_h^2 \frac{\sigma_h^2}{n_h} = \sum_{h=1}^{H} W_h^2 \frac{p_h q_h}{n_h} \quad (1)$$

$$\begin{aligned} \mathcal{V}_{AE_2}(\widehat{p}_{es}) &= \sum_{h=1}^{H} W_h^2 (1 - f_h) \frac{s_h^2}{n_h} = \sum_{h=1}^{H} W_h^2 (1 - f_h) \frac{N_h}{N_h - 1} \frac{\sigma_h^2}{n_h} \\ &= \sum_{h=1}^{H} W_h^2 (1 - f_h) \frac{N_h}{N_h - 1} \frac{p_h q_h}{n_h} \\ &= \sum_{h=1}^{H} W_h^2 \frac{N_h - n_h}{N_h - 1} \frac{p_h q_h}{n_h} \quad (2) \end{aligned}$$

Propriedades do estimador

- Estimadores apropriados, para as respectivas variâncias do estimador, são dados por:

$$\widehat{\mathcal{V}}_{AE_1}(\widehat{p}_{es}) = \sum_{h=1}^{H} W_h^2 \frac{\widehat{\sigma}_h^2}{n_h} = \sum_{h=1}^{H} W_h^2 \frac{\widehat{p}_h \widehat{q}_h}{n_h - 1}$$

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

- em que $\widehat{q}_h = 1 - \widehat{p}_h$, $\widehat{\sigma}_h^2 = \widehat{s}_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (Y_{hi} - \widehat{p}_h)^2$

Propriedades do estimador

- No caso das alocações proporcional e uniforme, e AAS_c em cada estrato, temos

$$\mathcal{V}_{AE_1(pr)}(\widehat{p}_{es}) = \sum_{h=1}^{H} W_h^2 \frac{p_h q_h}{n W_h} = \frac{1}{n} \sum_{h=1}^{H} W_h p_h q_h$$

$$\mathcal{V}_{AE_1(un)}(\widehat{p}_{es}) = \sum_{h=1}^{H} W_h^2 \frac{p_h q_h}{n/H} = \frac{H}{n} \sum_{h=1}^{H} W_h^2 p_h q_h$$

Propriedades do estimador

- No caso das alocações proporcional e uniforme, e AAS_s em cada estrato, temos

$$\mathcal{V}_{AE_2(pr)}(\widehat{p}_{es}) = \sum_{h=1}^{H} W_h^2 (1 - f_h) \frac{N_h}{N_h - 1} \frac{p_h q_h}{n W_h} = \frac{1}{n} \sum_{h=1}^{H} W_h (1 - f_h) \frac{N_h}{N_h - 1} p_h q_h, \; f_h = \frac{n}{N}$$

$$\mathcal{V}_{AE_2(un)}(\widehat{p}_{es}) = \sum_{h=1}^{H} W_h^2 (1 - f_h) \frac{N_h}{N_h - 1} \frac{p_h q_h}{n/H} = \frac{H}{n} \sum_{h=1}^{H} W_h^2 (1 - f_h) \frac{N_h}{N_h - 1} p_h q_h, \; f_h = \frac{n}{H N_h}$$

Propriedades do estimador

- Estimadores apropriados, sob diferentes formas de alocação e AAS_c, são dadas por:

$$\widehat{\mathcal{V}}_{AE_1(pr)}(\widehat{p}_{es}) = \sum_{h=1}^{H} W_h^2 \frac{\widehat{p}_h \widehat{q}_h}{nW_h - 1} = \sum_{h=1}^{H} W_h \frac{\widehat{p}_h \widehat{q}_h}{n - W_H^{-1}} \quad (3)$$

$$\widehat{\mathcal{V}}_{AE_1(un)}(\widehat{p}_{es}) = \sum_{h=1}^{H} W_h^2 \frac{\widehat{p}_h \widehat{q}_h}{n/H - 1} = H \sum_{h=1}^{H} W_h^2 \frac{\widehat{p}_h \widehat{q}_h}{n - H} \quad (4)$$

Propriedades do estimador

- Estimadores apropriados, sob diferentes formas de alocação e AAS_s, são dadas por:

$$\widehat{\mathcal{V}}_{AE_2(pr)}(\widehat{p}_{es}) = \sum_{h=1}^{H} W_h \left(1 - \frac{n}{N}\right) \frac{\widehat{p}_h \widehat{q}_h}{n - W_h^{-1}} \quad (5)$$

$$\widehat{\mathcal{V}}_{AE_2}(\widehat{p}_{es}) = H \sum_{h=1}^{H} W_h^2 \left(1 - \frac{n}{HN_h}\right) \frac{\widehat{p}_h \widehat{q}_h}{n - H} \quad (6)$$

Estimativas

- $\widetilde{p} = \sum_{h=1}^{H} W_h \widetilde{p}_h$, em que $\widetilde{p}_h = \frac{1}{n_h} \sum_{i \in \boldsymbol{s}_h} y_{hi}$, $\boldsymbol{s}_h = \{k_{h1}, k_{h2}, ..., k_{hn_h}\}$.
- $\widetilde{\mathcal{V}}_{AE_1}(\widehat{p}_{es}) = \sum_{h=1}^{H} W_h^2 \frac{\widetilde{p}_h \widetilde{q}_h}{n_h - 1}$.
- $\widetilde{\mathcal{V}}_{AE_2}(\widehat{p}_{es}) = \sum_{h=1}^{H} (1 - f_h) W_h^2 \frac{\widetilde{p}_h \widetilde{q}_h}{n_h - 1}$.

Resultados assintóticos

- Novamente, temos que ter n_h e N_h, $h = 1, 2, ..., H$ suficientemente grandes.
- Os estimadores $\widehat{p}_{es}$, $(\widehat{\sigma}^2, \widehat{s}^2)$ são consistentes, respectivamente, para p e σ^2, s^2.
- Usando resultados anteriores, vem que (AAS_c e AAS_s, respectivamente):

$$\frac{\widehat{p}_{es} - p}{\sqrt{\sum_{h=1}^{H} W_h^2 p_h q_h / n_h}} \xrightarrow[\substack{n_h \to \infty, \\ N_h - n_h \to \infty, \\ h=1,2,\ldots,H}]{D} N(0,1)$$

$$\frac{\widehat{p} - p}{\sqrt{\sum_{h=1}^{H} (1 - f_h) W_h^2 \frac{N_h}{N_h - 1} p_h q_h / n_h}} \xrightarrow[\substack{n_h \to \infty, \\ N_h - n_h \to \infty, \\ h=1,2,\ldots,H}]{D} N(0,1)$$

Resultados assintóticos

- Além disso, sob AAS_c e AAS_s, respectivamente, temos

$$\frac{\widehat{p}_{es} - p}{\sqrt{\sum_{h=1}^{H} W_h^2 \widehat{p}_h \widehat{q}_h / (n_h - 1)}} \xrightarrow[\substack{n_h \to \infty, \\ N_h - n_h \to \infty, \\ h=1,2,\ldots,H}]{D} N(0,1)$$

$$\frac{\widehat{p} - p}{\sqrt{\sum_{h=1}^{H} (1 - f_h) W_h^2 \widehat{p}_h \widehat{q}_h / (n_h - 1)}} \xrightarrow[\substack{n_h \to \infty, \\ N_h - n_h \to \infty, \\ h=1,2,\ldots,H}]{D} N(0,1)$$

Intervalo de Confiança (AAS_c)

- Assim, dois intervalos de confiança (assintóticos) com coeficiente de confiança de aproximadamente γ, são dados por

$$IC(\mu,\gamma) \approx \left[\widehat{p}_{es} - z_{\gamma}\sqrt{\sum_{h=1}^{H} W_h^2 \frac{\widehat{p}_h \widehat{q}_h}{n_h - 1}}; \widehat{p}_{es} + z_{\gamma}\sqrt{\sum_{h=1}^{H} W_h^2 \frac{\widehat{p}_h \widehat{q}_h}{n_h - 1}}\right] \quad (7)$$

$$IC(\mu,\gamma) \approx \left[\widehat{p}_{es} - z_{\gamma}\sqrt{\sum_{h=1}^{H} W_h^2 \frac{1}{4(n_h - 1)}}; \widehat{p}_{es} + z_{\gamma}\sqrt{\sum_{h=1}^{H} W_h^2 \frac{1}{4(n_h - 1)}}\right] \quad (8)$$

em que $P(Z \leq z_{\gamma}) = \frac{1+\gamma}{2}$ e $Z \sim N(0,1)$.

Intervalo de Confiança (AAS_c)

- Erro da estimativa: $z_{\gamma}\sqrt{\sum_{h=1}^{H} W_h^2 \frac{\widehat{p}_h\widehat{q}_h}{n_h-1}}$ ou $z_{\gamma}\sqrt{\sum_{h=1}^{H} W_h^2 \frac{1}{4(n_h-1)}}$.
- O comprimento do intervalo (8) sempre será maior (ou igual) ao comprimento do intervalo (7).

Intervalo de Confiança (AAS_s)

- Assim, dois intervalos de confiança (assintóticos) com coeficiente de confiança de aproximadamente γ, são dados por

$$
\begin{aligned}
IC(\mu,\gamma) \quad \approx \quad & \left[\widehat{p}_{es} - z_{\gamma}\sqrt{\sum_{h=1}^{H} W_h^2 (1-f_h)\frac{\widehat{p}_h\widehat{q}_h}{n_h-1}};\right. \\
& \left.\widehat{p}_{es} + z_{\gamma}\sqrt{\sum_{h=1}^{H} W_h^2 (1-f_h)\frac{\widehat{p}_h\widehat{q}_h}{n_h-1}}\right] \qquad (9)
\end{aligned}
$$

Intervalo de Confiança (AAS_s)

- Assim, dois intervalos de confiança (assintóticos) com coeficiente de confiança de aproximadamente γ, são dados por

$$IC(\mu,\gamma) \approx \left[\widehat{p}_{es} - z_{\gamma}\sqrt{\sum_{h=1}^{H} W_h^2(1-f_h)\frac{1}{4(n_h-1)}};\right.$$
$$\left.\widehat{p}_{es} + z_{\gamma}\sqrt{\sum_{h=1}^{H} W_h^2(1-f_h)\frac{1}{4(n_h-1)}}\right] \qquad (10)$$

em que $P\left(Z \leq z_{\gamma}\right) = \frac{1+\gamma}{2}$ e $Z \sim N(0,1)$.

Intervalo de Confiança (AAS_s)

- Erro da estimativa: $z_\gamma \sqrt{\sum_{h=1}^{H} W_h^2 (1 - f_h) \frac{\widehat{p}_h \widehat{q}_h}{n_h - 1}}$ ou $z_\gamma \sqrt{\sum_{h=1}^{H} W_h^2 (1 - f_h) \frac{1}{4(n_h - 1)}}$.
- O comprimento do intervalo (10) sempre será maior (ou igual) ao comprimento do intervalo (9).

Testes de Hipótese (AAS_c)

- Hipóteses usuais (p_0 conhecido, $q_0 = 1 - p_0$)
  1. $H_0 : p = p_0$ vs $H_1 : p < p_0$.
  2. $H_0 : p = p_0$ vs $H_0 : p > p_0$.
  3. $H_0 : p = p_0$ vs $H_0 : p \neq p_0$.
- Estatística do teste $Z_t = \frac{\widehat{p} - p_0}{\sqrt{\sum_{h=1}^{H} W_h^2 \frac{\widehat{p}_h \widehat{q}_h}{n_h - 1}}}$.
- Sob H_0, vimos que $Z_t \approx N(0,1)$, para n e N-n suficientemente grandes.

Testes de Hipótese (AAS_c)

- Defina $z_t = \frac{\widetilde{p} - p_0}{\sqrt{\sum_{h=1}^{H} W_h^2 \frac{\widetilde{p}_h \widetilde{q}_h}{n_h - 1}}}$ o valor calculado da estatística do teste e z_c o(s) valor(es) crítico(s).
- Defina ainda $Z \sim N(0,1)$. Os procedimentos são análogos ao caso da média, com as devidas adaptações.

Testes de Hipótese (AAS_s)

- Hipóteses usuais (p_0 conhecido, $q_0 = 1 - p_0$)
  1. $H_0 : p = p_0$ vs $H_1 : p < p_0$.
  2. $H_0 : p = p_0$ vs $H_0 : p > p_0$.
  3. $H_0 : p = p_0$ vs $H_0 : p \neq p_0$.
- Estatística do teste $Z_t = \frac{\widehat{p} - p_0}{\sqrt{\sum_{h=1}^{H}(1-f_h)W_h^2 \frac{\widehat{p}_h \widehat{q}_h}{n_h - 1}}}$.
- Sob H_0, vimos que $Z_t \approx N(0,1)$, para n e N-n suficientemente grandes.

Testes de Hipótese (AAS_s)

- Defina $z_t = \frac{\widetilde{p} - p_0}{\sqrt{\sum_{h=1}^{H}(1-f_h)W_h^2 \frac{\widetilde{p}_h \widetilde{q}_h}{n_h - 1}}}$ o valor calculado da estatística do teste e z_c o(s) valor(es) crítico(s).
- Defina ainda $Z \sim N(0,1)$. Os procedimentos são análogos ao caso da média, com as devidas adaptações.

Determinação do tamanho da amostra

- Devido ao fato de que expressões dos estimadores das variâncias Equações (3), (4), (5) e (6), a obtenção analítica do tamanho da amostra, sob as alocações proporcinal e uniforme são bastante complicadas.
- Podemos utilizar as expressões da alocação ótima, ou seja, páginas 38 e 29 dos slides `http://www.ime.unicamp.br/~cnaber/aula_AE%20Amost%202S%202018.pdf`.

Determinação do tamanho da amostra

- Uma outra opção é utilizar as expressões das variâncias (1) e (2) e, posteriormente, considerar estimadores para $p_h, (q_h)$, por exemplo, valores relativos ao pior cenário $(p_h = q_h = \frac{1}{2}, h = 1, 2, ..., H)$.
- Nesse caso, basta substituir $\widehat{\sigma}_h^2$ e $\widehat{s}_h^2$, por $\widehat{p}_h\widehat{q}_h$, as expressões das páginas 15, 16, 18, 20, dos slides `http://www.ime.unicamp.br/~cnaber/aula_AE%20P2%20Amost%202S%202018.pdf`