Amostragem por conglomerados em um único estágio (AC): parte 2

Prof. Caio Azevedo

Normalidade assintóticados estimadores

- Temos, para a e A suficientemente grandes, $i = 1, 2, 3$, $j = 1, 2$, que

$$\frac{\widehat{\mu}_{C_i} - \mu}{\text{EP}_{A_j}\left(\widehat{\mu}_{C_i}\right)} \xrightarrow[\substack{a \to \infty, \\ A - a \to \infty}]{D} N(0,1)$$

$$\frac{\widehat{\mu}_{C_i} - \mu}{\widehat{\text{EP}}_{A_j}\left(\widehat{\mu}_{C_i}\right)} \xrightarrow[\substack{a \to \infty, \\ A - a \to \infty}]{D} N(0,1)$$

- As notações relativas à Amostragem por Conglomerados (AC) seguem essas (aqui) e as de teoria assintótica essas (aqui e aqui).

Intervalos de confiança

- Conglomerados de tamanhos desiguais

$$\left[\widehat{\mu}_{C_i} - z_\gamma\sqrt{\widehat{\mathcal{V}}_{A_j}(\widehat{\mu}_{C_i})}; \widehat{\mu}_{C_i} + z_\gamma\sqrt{\widehat{\mathcal{V}}_{A_j}(\widehat{\mu}_{C_i})}\right]$$

- Erro da estimativa: $z_\gamma\sqrt{\widehat{\mathcal{V}}_{A_j}(\widehat{\mu}_{C_i})}$.

Intervalos de confiança

- Conglomerados de tamanhos iguais (AAS_c)

$$\left[\widehat{\mu}_C - z_{\gamma}\sqrt{\frac{1}{a(a-1)}\sum_{\alpha=1}^{a}(\widehat{\mu}_{\alpha} - \widehat{\mu}_C)^2}; \widehat{\mu}_C + z_{\gamma}\sqrt{\frac{1}{a(a-1)}\sum_{\alpha=1}^{a}(\widehat{\mu}_{\alpha} - \widehat{\mu}_C)^2}\right]$$

em que $\widehat{\mu}_C = \frac{1}{a}\sum_{\alpha=1}^{a}\widehat{\mu}_{\alpha}$.

- Erro da estimativa: $z_{\gamma}\sqrt{\frac{1}{a(a-1)}\sum_{\alpha=1}^{a}(\widehat{\mu}_{\alpha} - \widehat{\mu}_C)^2}$.

Cont.

- Os testes de hipóteses seguem, essencialmente, os mesmo procedimentos descritos nos slides 10 a 14 desses slides, em que a estatística do teste passa a ser

$$Z_t = \frac{\widehat{\mu}_{C_i} - \mu_0}{\widehat{\text{EP}}_{A_j}\left(\widehat{\mu}_{C_i}\right)}.$$

Determinação do tamanho da amostra

- O "tamanho da amostra" a ser determinado, nesse caso, é o número "a" de conglomerados a serem sorteados.
- Se os conglomerados tiverem tamanhos iguais, o tamanho da amostra real ($n = \sum_{\alpha=1}^{a} n_{\alpha}$) será constante, dado "a". Caso contrário, será uma variável aleatória.

Determinação do tamanho da amostra

- Supondo conglomerados de mesmo tamanho, e fixando a precisão (δ), temos que :

$$
\begin{aligned}
P(|\widehat{\mu}_C - \mu| \leq \delta) \geq \gamma \quad & \rightarrow \quad P\left(|Z| \leq \frac{\delta}{\sqrt{\sigma^2_{ec}/a}}\right) = \gamma \rightarrow \\
z_\gamma \quad & = \quad \frac{\delta}{\sqrt{\sigma^2_{ec}/a}} \rightarrow a = \frac{\sigma^2_{ec} z^2_\gamma}{\delta^2}
\end{aligned}
$$

em que $\sigma^2_{ec} = \dfrac{1}{A}\sum_{\alpha=1}^{A}(\mu_\alpha - \mu)^2$

- Se se fixar o erro da estimativa, obter-se-á a mesma fórmula para o tamanho da amostra.
- Exercício: obter o tamanho da amostra considerando conglomerados de tamanhos diferentes.

Determinação do tamanho da amostra

- Exercício: obter todas as expressões para as esperanças, variâncias, EQM, testes de hipótese, IC e tamanhos de amostra considerando que os conglomerados foram selecionados segundo um plano AAS_s.
- Exercício: repetir os procedimentos para a estimação do total populacional (τ), a partir de estimadores que você entender que são apropriados.

Estimação de proporções

- Todas as quantidades são como definidas aqui.
- Parâmetro de interesse: $p = \dfrac{\displaystyle\sum_{\alpha=1}^{A} \tau_{\alpha}}{\displaystyle\sum_{\alpha=1}^{A} B_{\alpha}}$.
- Como a proporção é uma média de variáveis binárias, os desenvolvimentos seguem àqueles apresentados para a estimação da média sob AC e para a proporção sob A_1 (veja aqui) e sob A_2 (exercício).

Estimação de proporções

- Estimador (razão): $\widehat{p}_{C_2} = \dfrac{\sum_{\alpha=1}^{a} \widehat{\tau}_{\alpha}}{\sum_{\alpha=1}^{a} b_{\alpha}} = \dfrac{\widehat{\tau}}{\widehat{\overline{B}}}$.
- Variância: $\mathcal{V}(\widehat{p}_{C_2}) \approx \dfrac{1}{aA\overline{B}^2} \sum_{\alpha=1}^{A} (\tau_{\alpha} - pB_{\alpha})^2$
- Estimador para a variância: $\widehat{\mathcal{V}}(\widehat{p}_{C_2}) \approx \dfrac{1}{a(a-1)\overline{b}^2} \sum_{\alpha=1}^{a} (\widehat{\tau}_{\alpha} - \widehat{p}_{C_2} b_{\alpha})^2$
- Os dois outros estimadores são:

$$\widehat{p}_{C_1} = \frac{\widehat{\tau}}{\overline{B}}; \widehat{p}_{C_3} = \frac{1}{a}\sum_{\alpha=1}^{a} \widehat{p}_{\alpha}$$

Cont.

- Suas variâncias são dadas, respectivamente, por:

$$\mathcal{V}(\widehat{p}_{C_1}) = \frac{1}{aA\overline{B}^2}\sum_{\alpha=1}^{A}(\tau_\alpha - p)^2 \, ; \mathcal{V}(\widehat{p}_{C_3}) = \frac{1}{aA}\sum_{\alpha=1}^{A}(p_\alpha - \overline{p})^2$$

- Estimadores são dados, respectivamente, por:

$$\widehat{\mathcal{V}}(\widehat{p}_{C_1}) = \frac{1}{a(a-1)\overline{B}^2}\sum_{\alpha=1}^{A}(\widehat{\tau}_\alpha - \widehat{p}_{C_1})^2 \, ; \widehat{\mathcal{V}}(\widehat{p}_{C_3}) = \frac{1}{a(a-1)}\sum_{\alpha=1}^{A}(\widehat{p}_\alpha - \widehat{p}_{C_3})^2$$

- Quando os conglomerados tiverem o mesmo tamanho, as expressões das variâncias (e de seus respectivos estimadores), simplificam-se. Veja o exercício 7.21 (Bolfarine & Bussab (2005)).

Normalidade assintótica

- Temos, para a e A suficientemente grandes, $i = 1, 2, 3$, $j = 1, 2$, que

$$\frac{\widehat{p}_{C_i} - p}{\text{EP}_{A_j}\left(\widehat{p}_{C_i}\right)} \xrightarrow[\substack{a \to \infty, \\ A - a \to \infty}]{D} N(0,1)$$

$$\frac{\widehat{p}_{C_i} - p}{\widehat{\text{EP}}_{A_j}\left(\widehat{p}_{C_i}\right)} \xrightarrow[\substack{a \to \infty, \\ A - a \to \infty}]{D} N(0,1)$$

Intervalos de confiança

- Conglomerados de tamanhos desiguais $j = 1, 2,\ i = 1, 2, 3$

$$\left[\widehat{p}_{C_i} - z_{\gamma}\sqrt{\widehat{\mathcal{V}}_{A_j}(\widehat{p}_{C_i})}; \widehat{p}_{C_i} + z_{\gamma}\sqrt{\widehat{\mathcal{V}}_{A_j}(\widehat{p}_{C_i})}\right]$$

- Erro da estimativa:$z_{\gamma}\sqrt{\widehat{\mathcal{V}}_{A_j}(\widehat{p}_{C_i})}$.
- Exercício: repetir os procedimentos considerando que os conglomerados são sorteados segundo um plano AAS_s e/ou quando estes tiverem o mesmo tamanho.

Intervalos de confiança

- Os testes de hipóteses seguem, essencialmente, os mesmo procedimentos descritos no slide 10 deste material, em que a estatística do teste passa a ser

$$Z_t = \frac{\widehat{p}_{C_i} - p_0}{\widehat{\mathrm{EP}}_{A_j}\left(\widehat{p}_{C_i}\right)}.$$

- Exercício: repetir os procedimentos considerando que os conglomerados são sorteados segundo um plano AAS_s e/ou quando os conglomerados tiverem o mesmo tamanho.

Amostragem sistemática (AS)

- Trata-se de um processo de seleção quase-aleatório (link).
- Considere uma população com N elementos em que $N = kn$, e $k \in \mathbb{Z}^+$ e que seus elementos estão ordenados de 1 a N (no sistema de referência).
- Uma unidade é selecionada aleatoriamente (segundo AAS) entre as k primeiras unidades do sistema de referência.
- As unidades seguintes, que farão parte da amostra, serão obtidas a partir da primeira unidade selecionada, em intervalos de comprimento k.

Exemplo

- Suponha uma população com $N = 1.000$ e $n = 200$. Portanto, $k = 5$.
- Ou seja a população está divididade em 200 grupos de 5 unidades populacionais.
- Suponha que uma unidade seja aleatoriamente selecionada entre as 5 primeiras e que esta tenha sido a de número 3.
- Portanto, em cada um dos 199 grupos restantes, será selecioada sempre a terceira unidade, levando, assim, a uma AS de n = 200 unidades.

Características

- Facilidade em sua execução (mecânica do sorteio).
- Bem menos sujeita a erros do entrevistador do que os outros métodos de amostragem vistos.
- Pode ou não levar à inferências mais precisas.
- Pode levar a inferências viesadas, dependendo de como a população esteja organizada no sistema de referência.

Características

- As estimativas das variâncias do estimadores dependem de como os elementos da população estão dispostos no sistema de referência (aleatório ou de forma tendenciosa).
- Essencialmente, a AS pode ser vista como uma caso particular da AC (embora, se a população esteja diposta de forma aleatória, algumas expressões podem ser próximas daquelas obtidas sob o plano AAS).
- Outros estimadores para as variâncias dos estimadores podem ser encontrados em Bolfarine e Bussab (2005) e Cochran (1977).

Relação com a AC

- Considere a população ordenada da seguinte forma:

$$\boldsymbol{d} = \left(y_1, ..., y_k, y_{k+1}, ..., y_{2k}, ..., y_{(n-1)k+1}, ..., y_{nk}\right)^t$$

que pode também ser representado através de uma matriz onde na linha α tem-se a $\alpha-$ésima amostra sistemática, enquanto que na coluna i tem-se o i-ésimo grupo(zona).

Relação com a AC

Amostra	grupo				média
	1	2	...	n	
1	y_1	y_{k+1}	$\cdots$	$y_{(n-1)k+1}$	μ_1
2	y_2	y_{k+2}	$\cdots$	$y_{(n-1)k+2}$	μ_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
k	y_k	y_{2k}	$\cdots$	y_{nk}	μ_k
média	$\mu_{.1}$	$\mu_{.2}$	$\cdots$	$\mu_{.n}$	μ

Relação com a AC

- Cada linha da matriz anterior corresponde a uma amostra sistemática de média μ_α, $\alpha = 1, 2, ..., k$.
- Na última coluna temos todas as médias das k amostras sistemáticas.
- Cada uma dessas amostras pode ser vista como um conglomerado (todos de tamanho n).
- Portanto a AS pode ser vista como uma AC em que $A = k$ conglomerados e destes, $a = 1$ é selecionado.

Estimador para a média (μ)

- $\widehat{\mu}_{sis} = \widehat{\mu}_{\alpha}$, a média da amostra sistemática (conglomerado) sorteada(o).
- Distribuição exata do estimador $\widehat{\mu}_{sis}$,

$\widehat{\mu}_{sis}$	μ_1	μ_2	$\ldots$	μ_α	$\ldots$	μ_k
$P(\widehat{\mu}_{sis} = r)$	$\frac{1}{k}$	$\frac{1}{k}$	$\ldots$	$\frac{1}{k}$	$\ldots$	$\frac{1}{k}$

- Portanto $\mathcal{E}(\widehat{\mu}_{sis}) = \dfrac{1}{k}\displaystyle\sum_{\alpha=1}^{k}\mu_k = \mu$ e $V_k = \mathcal{V}(\widehat{\mu}_{sis}) = \dfrac{1}{k}\displaystyle\sum_{\alpha=1}^{k}(\mu_\alpha - \mu)^2$.
- Como selecionamos apenas um único conglomerado $a = 1$, não é possível obter uma estimador não viciado de $\mathcal{V}(\widehat{\mu}_{sis})$.

Estimador para a média (μ)

- Na maioria dos casos, essa variância é estimada por:

$$\widehat{V}_k = \widehat{V}_s = \frac{1-f}{n(n-1)} \sum_{i=1}^{n} \left(Y_i - \widehat{\mu}_{sis}\right)^2 ,$$

em que $f = \frac{n}{N}$ e $(Y_1, ..., Y_n)^t$ são variáveis aleatórias que representam a amostra a ser sorteada.

- Tal estimador é adequado quando a AS é aproximadamente equivalente à *AAS* (elementos organizados de forma "aleatória").
- Contudo, se se tiver algum padrão na disposição dos elementos no sistema de referência, os resultados podem ser diferentes daquelas da *AAS*.

Estimador para a média (μ)

- Sob AS (lembrando que AS corresponde ao plano AC, sob certas restrições), temos que

$$V_k = \{1 + \rho_{int}(n-1)\} \frac{\sigma^2}{n}.$$

- em que $\rho_{int} = \dfrac{\sigma^2_{ec} - \sigma^2_{dc}/(n-1)}{\sigma^2}$.
- Como todos os conglomerados têm o mesmo tamanho e $B = n$ (tamanho de cada conglomerado) então

$$\sigma^2_{ec} = \frac{1}{k}\sum_{\alpha=1}^{k}(\mu_\alpha - \mu)^2,\ \sigma^2_{dc} = \frac{1}{k}\sum_{\alpha=1}^{k}\sigma^2_\alpha,\ \sigma^2_\alpha = \frac{1}{n}\sum_{i=1}^{n}(y_{\alpha i} - \mu_\alpha)^2$$

Exemplos

- (Exemplo 7.6, do livro Bolfarine &Bussab (2005)) Considere uma população onde $\boldsymbol{d} = (2, 6, 10, 8, 10, 12)$, $N = 6$ e dividida em $n = 2$ e $k = 3$ unidades cada, como segue:

	Grupo		
Amostra	1	2	Média
1	2	8	5
2	6	10	8
3	10	12	11

Exemplos

- Distribuição exata de $\widehat{\mu}_{sis}$

$\widehat{\mu}_{sis}$	5	8	11
$P(\widehat{\mu}_{sis} = r)$	1/ 3	1/ 3	1/ 3

- O que leva à $\mathcal{E}(\widehat{\mu}_{sis}) = 8$, $\mathcal{V}(\widehat{\mu}_{sis}) = 6$.

Exemplos

- Distribuição exata de $\widehat{\mu}$ (sob AAS_s para n $= 2$)

$\widehat{\mu}$	4	5	6	7	8	9	10	11
$P(\widehat{\mu} = r)$	$1/15$	$1/15$	$2/15$	$2/15$	$2/15$	$1/5$	$2/15$	$2/15$

- O que leva à $\mathcal{E}(\widehat{\mu}) = 8$, $\mathcal{V}(\widehat{\mu}) = 4,27$. Assim, nota-se que $\widehat{\mu}_{sis}$ é pior do que $\widehat{\mu}$. Com efeito, $\rho_{int} = 0,125 > 0$.

Exemplos

- Exemplo 7.7 (livro-texto). Considere uma população com $N = 40$ elementos distribuídos em $k = 10$ amostras sistemáticas de tamanho $n = 4$.
- Análise: aqui.
- Sob $AAS_s, n = 4$:

$$\mathcal{V}_{A_2}(\widehat{\mu}) = (1 - f)\frac{s^2}{n} = \left(1 - \frac{1}{10}\right)\frac{136,25}{4} = 30,66.$$

Exemplos

- Além disso, $\mathcal{V}(\widehat{\mu}_{sis}) = \frac{1}{k}\sum_{\alpha=1}^{k}(\mu_\alpha - \mu)^2 = 8,74$.
- Com efeito, neste caso $\rho_{int} = -0,217 < 0$
- Sob AE: cada um dos $H = 4$ grupos é um estrato de tamanho $N_h = 10$. Seleciona-se um elemento de cada estrato($n = \sum_h^4 1 = 4$). Assim:

$$
\begin{aligned}
\mathcal{V}(\widehat{\mu}_{es}) &= \sum_{h=1}^{4} W_h^2 (1 - f_h)\frac{s_h^2}{n_h} = \frac{1}{16}\left(1 - \frac{1}{10}\right)\sum_{h=1}^{4} s_h^2 \\
&= \left(1 - \frac{1}{10}\right) 53,94 \approx 3,03
\end{aligned}
$$