Amostragem por conglomerados em dois estágios (AC2E): uma introdução

Prof. Caio Azevedo

Motivação

- Na presença de conglomerados homogêneos (elementos dentro de cada conglomerado muito parecidos entre si) o uso da AC (completa) (selecionando-se todos os elementos de cada conglomerado) se torna menos indicada.
- Dado que os elementos são muito parecidos entre si, eles trarão a mesma informação, essencialmente.

Motivação

- Considere que os elementos dentro de cada conglomerado são todos iguais entre si. Basta selecionar um elemento dentro de cada conglomerado.
- Uma alternativa: selecionar conglomerados e, destes, selecionar um determinado número de elementos.
- Observações: as definições e notações são, essencialmente, as mesmas definidas para o plano AC (Parte 1, Parte 2).

Descrição do plano de Amostragem por conglomerados em dois estágios (AC2E)

- Após a população estar agrupada em A conglomerados, procede-se da seguinte forma:
  - Selecionam-se, aleatoriamente, no primeiro estágio, a conglomerados, segundo algum plano amostral.
  - De cada conglomerado selecionado, sorteiam-se b_α elementos, segundo o mesmo ou outro plano amostral.
  - Pode-se usar em cada estágio AAS_c ou AAS_s.
- Consideraremos, somente, AAS_c no dois estágios. Veja as referências, para outras combinações.
- Exercício: repetir os desenvolvimentos considerando AAS_s em cada estágio.

Exemplo

- Considere uma população agrupada em 3 conglomerados, como se segue (o mesmo visto aqui):

$$\mathcal{U} = \{(1), (2,3,4), (5,6)\} = \{C_1, C_2, C_3\}$$

em que $C_1 = \{1\}$, $C_2 = \{2,3,4\}$ e $C_3 = \{5,6\}$.

- O plano amostral adotado consiste em sortear dois conglomerados, sem reposição, e de cada conglomerado sorteado, sortear um elemento com igual probabilidade.
- Espaço amostral: construção em dois estágios.

Exemplo

- Espaço amostral em função dos conglomerados (primeiro estágio):

$$\begin{aligned} S_C(\mathcal{U}) &= \{C_1C_2, C_1C_3, C_2C_1, C_2C_3, C_3C_1, C_3C_2\} \\ &= \{\boldsymbol{s}_{c1}, \boldsymbol{s}_{c2}, \boldsymbol{s}_{c3}, \boldsymbol{s}_{c4}, \boldsymbol{s}_{c5}, \boldsymbol{s}_{c6}\}. \end{aligned}$$

- Cada ponto $(\boldsymbol{s}_{ci})$ tem probabilidade de $1/6$ de ser selecionado.
- Supondo, por exemplo, que o ponto C_1C_2 foi selecionado, então

$$S(C_1C_2) = \{12, 13, 14\} = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{s}_3\}.$$

- Condicionado-se a este subespaço $(S(C_1C_2))$, cada elemento $(\boldsymbol{s}_i)$ terá probabilidade de $1/3$ de ser selecionado aleatoriamente.

Exemplo

- Utilizando-se a probabilidade (marginal) de seleção do ponto $(C_1 C_2)$ e condicional de cada par $(\{12{,}13{,}14\})$, temos que a probabilidade de sorteio de cada um desses três elementos é

$$P(\boldsymbol{s}_i) = P(\boldsymbol{s}_i | \boldsymbol{s}_{c_i}) P(\boldsymbol{s}_{c_i}) = \frac{1}{3} \times \frac{1}{6} = \frac{1}{18}.$$

- Ao se considerar todas as possibilidades de pontos de $S_C(\mathcal{U})$ e de cada um dos pontos associados à cada um desses (pares de elementos), teremos bem caracterizado o plano amostral, que pode ser visto aqui, lembrando que $\boldsymbol{d} = (12, 7, 9, 14, 8, 10)^t$ e $\mu = 10$, $\sigma^2 = 34/6$, $N = 6$, $\overline{B} = 2$ e $A = 3$.

Inferência

- Considere o seguinte estimador $\widehat{\mu}_{2c} = \frac{1}{a}\sum_{\alpha=1}^{a}\widehat{\mu}_{2\alpha}$, em que a denota o número de conglomeradas selecionados no primeiro estágio e $\widehat{\mu}_{2\alpha}$ a média dos elementos selecionados, no segundo estágio, do conglomerado α.
- No exemplo em questão temos que $\widehat{\mu}_{2c} = \frac{\widehat{\mu}_{\boldsymbol{s}_{i1}} + \widehat{\mu}_{\boldsymbol{s}_{i2}}}{2}$, em que $\widehat{\mu}_{\boldsymbol{s}_{i1}}$ é a média (no caso o próprio valor) dos elementos do conglomerado $i, i = 1, 2$.
- Perceba que temos duas fontes de aleatorieade: 1 - seleção de conglomerados, 2 - seleção de elementos dentro dos conglomerados.

Inferência

- Com base nos resultados descritos aqui temos que a distribuição exata de $\widehat{\mu}_{2c}$ neste caso, é dada por:

$\widehat{\mu}_{2c}$	7,5	8,5	9,5	10	10,5	11	12	13
$P(\widehat{\mu}_{2c})$	1/18	1/ 9	1/ 6	1/ 6	1/ 9	2/ 9	1/18	1/9

Inferência

- O que nos leva à $\mathcal{E}(\widehat{\mu}_{2c}) \approx 10,33$ e $\mathcal{V}(\widehat{\mu}_{2c}) = 2$, portanto, $\widehat{\mu}_{2c}$ é um estimador viciado. Note que este estimador corresponde a média aritmética (simples) entre as médias dos elementos de cada conglomerado selecionado.
- Considere, agora, o seguinte estimador $\widehat{\mu}_{2c1} = \dfrac{1}{a}\displaystyle\sum_{\alpha=1}^{a} \dfrac{B_\alpha}{\overline{B}} \widehat{\mu}_\alpha$.

Inferência

- Fazendo os desenvolvimentos de modo semelhante ao caso anterior, temos que a respectiva distribuição de probabilidade (exata) é dada por:

$\widehat{\mu}_{2c1}$	7	8	8,25	9,25	9,75	10,25	10,75	11,75	13,5	14,5	15,5
$P(\widehat{\mu}_{2c1})$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{9}$	$\frac{1}{18}$	$\frac{1}{9}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{9}$	$\frac{1}{18}$	$\frac{1}{18}$

Inferência

- O que nos leva à $\mathcal{E}(\widehat{\mu}_{2c1}) = 10$ e $\mathcal{V}\left(\widehat{\mu}_{2c1}\right) \approx 6,92$.
- Isso nos sugere que o estimador $(\widehat{\mu}_{2c1})$ é não viciado.
- Mas, $EQM(\widehat{\mu}_{2c1}) = \mathcal{V}(\widehat{\mu}_{2c1}) = 6,92$ e

  $EQM(\widehat{\mu}_{2c}) = \mathcal{V}(\widehat{\mu}_{2c}) + B(\widehat{\mu}_{2c})^2 \approx 2 + 0,11 = 2,11$, ou seja,

$$EQM(\widehat{\mu}_{2c1}) > EQM(\widehat{\mu}_{2c}).$$

AC $\times$ AC2E

- Considere os dois estimadores: $\widehat{\mu}_{C_3} = \frac{1}{a}\sum_{\alpha=1}^{a} \mu_{\alpha}$ (AC) e $\widehat{\mu}_{2c} = \frac{1}{a}\sum_{\alpha=1}^{a} \widehat{\mu}_{2\alpha}$ (AC2E).
- No primeiro caso (AC) note que, dados os conglomerados sorteados, μ_{α}, $\alpha = 1, 2, ..., a$, **são não aleatórios**, uma vez que de cada um daqueles considerar-se-ão todos os elementos.
- No entanto, no segundo caso (AC2E), mesmo condicionado aos conglomerados sorteados, $\widehat{\mu}_{\alpha}$, $\alpha = 1, 2, ..., a$, **são aleatórios**, uma vez que de cada um daqueles sortear-se-ão em certo número de seus elementos.

Inferência

- Como a amostragem é feita em dois estágios, ou seja, primeiro selecionam-se os conglomerados e, depois, os elementos dentro de cada conglomerado, podemos calcular a esperança e a variância usando resultados de distribuições condicionais.
- Por exemplo, para o par C_1C_2, temos que $S_2(C_1C_2) = \{12, 13, 14\}$, de sorte que: $\widetilde{\mu}_{2c1}[12] = 8,25$, $\widetilde{\mu}_{2c1}[13] = 9,75$ e $\widetilde{\mu}_{2c1}[14] = 13,5$ (valores possíveis do estimador $\widehat{\mu}_{2c1}$).
- Usando-se o índice 2 para indicar o valor esperado condicional ao particular par de conglomerados, tem-se.

$$\mathcal{E}_2(\widehat{\mu}_{2c1}) \equiv \mathcal{E}(\widehat{\mu}_{2c1}|C_1C_2) = \frac{1}{3}(8,25 + 9,75 + 13,5) = \widetilde{\mu}_{cd}[C_1C_2].$$

Inferência

- Estendendo-se os desenvolvimentos para dos demais pares, teremos a seguinte distribuição condicional (note que estamos relaxando a notação apropriada para distribuições de probabilidade e, também, estamos condicionando nos pares de conglomerados $C_i C_j$):

$\boldsymbol{s}_{c_i}$:	$C_1 C_2$	$C_1 C_3$	$C_2 C_1$	$C_2 C_3$	$C_3 C_1$	$C_3 C_2$
$\mathcal{E}(\widehat{\mu}_{cd} \mid C_i C_j)$	10,5	7,5	10,5	12	7,5	12
$P(\mathcal{E}(\widehat{\mu}_{cd} = r \mid C_i C_j))$	1/ 6	1/ 6	1/ 6	1/ 6	1/ 6	1/ 6

Inferência

- Usando-se, agora, o índice 1 para indicar a esperança calculada no espaço amostral gerado pelos conglomerados, tem-se

$$\mathcal{E}_1(\widehat{\mu}_{2c1}) = \frac{1}{6}(10,5 + 7,5 + 10,5 + 12 + 7,5 + 12) = 10$$

- Esse modus operandi é bastante útil quando se tem vários níveis de seleção no plano amostral.

Notações

- As notações serão basicamente aquelas que foram utilizadas na AC e o planejamento amostral já foi descrito na página 4.
- Lembrando - média populacional: $\mu = \overline{y} = \dfrac{\tau}{N} = \dfrac{A\overline{\tau}}{A\overline{\overline{B}}} = \dfrac{\overline{\tau}}{\overline{\overline{B}}}$
- Estimador (utilizado): $\widehat{\mu}_{2c1} = \dfrac{1}{a}\sum_{\alpha=1}^{a}\dfrac{B_{\alpha}}{\overline{\overline{B}}}\widehat{\mu}_{\alpha}$, em que $\widehat{\mu}_{\alpha}$ é a média dos elementos selecionados do conglomerado α (que também fora selecionado).
- Para outros estimadores veja as referências, incluindo o livro Bolfarine & Bussab (2005).

Inferência

- Valor esperado e variância.
- Sejam X e Y duas variáveis aleatórias definidas no mesmo espaço de probabilidade. Então

$$\mathcal{E}_X(X) = \mathcal{E}_Y(\mathcal{E}_{X|y}(X|Y))$$

e

$$\mathcal{V}(X) = \mathcal{E}_Y(\mathcal{V}_{X|y}(X|Y)) + \mathcal{V}_Y(\mathcal{E}_{X|y}(X|Y)).$$

Inferência

- Vamos utilizar as notações definidas anteriormente, ou seja, $\mathcal{E}((.)|S_C(\mathcal{U})) \equiv \mathcal{E}_2((.))$ (entre os elementos, condicionado no espaço amostral gerado pelos conglomerados) e $\mathcal{E}_1(.)$ (esperança em função do espaço amostral gerado pelos conglomerados).
- Assim, temos que

$$\mathcal{E}(\widehat{\mu}_{2c1}) = \mathcal{E}_1(\mathcal{E}_2(\widehat{\mu}_{2c1})). \tag{1}$$

Inferência

- Dentro do $\alpha-$ésimo conglomerado (sorteado e com B_α unidades), os b_α elementos são selecionados segundo uma AAS_c. Assim, dessa parte, vem que:

$$\mathcal{E}_2(\widehat{\mu}_{2c1}) = \frac{1}{a}\sum_{\alpha=1}^{a}\frac{B_\alpha}{\overline{B}}\mathcal{E}_2(\widehat{\mu}_\alpha) = \frac{1}{a}\sum_{\alpha=1}^{a}\frac{B_\alpha}{\overline{B}}\mu_\alpha \tag{2}$$

- Atenção: Note, no entanto, que μ_α é uma variável aleatória que representa a média do $\alpha-$ésimo conglomerado sorteado, pois serão sorteados $b_\alpha \leq B_\alpha$ elementos, daquele.

Inferência

- Suponha uma população formada de A elementos (conglomerados) e defina $x_\alpha = \frac{B_\alpha}{\overline{B}}\mu_\alpha$, $\alpha = 1, 2, ..., A$.
- Assim, a média, sob uma amostra de a unidades (AAS_c), digamos, $\overline{X} = \frac{1}{a}\sum_{\alpha=1}^{a} X_\alpha = \frac{1}{a}\sum_{\alpha=1}^{a}\frac{B_\alpha}{\overline{B}}\mu_\alpha$, é tal que

$$\mathcal{E}_1(\overline{X}) = \frac{1}{A}\sum_{\alpha=1}^{A}\frac{B_\alpha}{\overline{B}}\mu_\alpha = \mu \tag{3}$$

(conforme visto aqui).

Inferência

- Logo, de (2) e (3) em (1), vem que:

$$\mathcal{E}(\widehat{\mu}_{2c1}) = \mathcal{E}_1(\overline{X}) = \mathcal{E}_1\left(\frac{1}{a}\sum_{\alpha=1}^{a}\frac{B_\alpha}{\overline{B}}\mu_\alpha\right) = \mu.$$

- Portanto, $\widehat{\mu}_{2c1}$ é um estimador não viciado para μ.
- Para a variância, temos que:

$$\mathcal{V}(\widehat{\mu}_{2c1}) = \mathcal{E}_1\left(\mathcal{V}_2(\widehat{\mu}_{2c1})\right) + \mathcal{V}_1\left(\mathcal{E}_2(\widehat{\mu}_{2c1})\right) \tag{4}$$

- Primeiramente, note, pelos mesmos argumentos anteriores, que

$$\begin{aligned}\mathcal{V}_1(\overline{X}) &= \mathcal{V}_1\left(\frac{1}{a}\sum_{\alpha=1}^{a}\frac{B_\alpha}{\overline{B}}\mu_\alpha\right) = \frac{\sigma_x^2}{a} = \frac{1}{aA}\sum_{\alpha=1}^{A}\left(\frac{B_\alpha}{\overline{B}}\mu_\alpha - \mu\right)^2 \\ &= \frac{\sigma_{ect}^2}{a} \qquad (5)\end{aligned}$$

Inferência

- Por outro lado, pelo mesmos argumentos que nos levaram à Equação (2) (sob AAS_c e lembrando que tomaremos uma amostra de b_α elementos do conglomerado α), temos que:

$$
\begin{aligned}
\mathcal{V}_2(\widehat{\mu}_{2c1}) &= \frac{1}{a^2}\sum_{\alpha=1}^{a}\left(\frac{B_\alpha}{\overline{B}}\right)^2 \mathcal{V}_2(\widehat{\mu}_\alpha) \\
&= \frac{1}{a^2}\sum_{\alpha=1}^{a}\left(\frac{B_\alpha}{\overline{B}}\right)^2 \frac{\sigma_\alpha^2}{b_\alpha} = \frac{1}{a}\sum_{\alpha=1}^{a} V_\alpha = \overline{V}
\end{aligned}
$$

em que $v_\alpha = \frac{1}{a}\left(\frac{B_\alpha}{\overline{B}}\right)^2 \frac{\sigma_\alpha^2}{b_\alpha}$ é uma variável aleatória auxiliar (semelhante à x_α).

Inferência

- Com efeito, utilizando-se raciocínio análogo àquele utilizado para X_α, vem que:

$$\begin{aligned}\mathcal{E}_1\left(\mathcal{V}_2(\widehat{\mu}_{2c1})\right) &= \mathcal{E}_1\left(\frac{1}{a^2}\sum_{\alpha=1}^{a}\left(\frac{B_\alpha}{\overline{B}}\right)^2\frac{\sigma_\alpha^2}{b_\alpha}\right) = \mathcal{E}_1\left(\overline{V}\right) = \overline{v} = \frac{1}{A}\sum_{\alpha=1}^{A} v_\alpha \\ &= \frac{1}{aA}\sum_{\alpha=1}^{A}\left(\frac{B_\alpha}{\overline{B}}\right)^2\frac{\sigma_\alpha^2}{b_\alpha} \qquad (6)\end{aligned}$$

em que $\overline{V} = \frac{1}{a}\sum_{\alpha=1}^{a} V_\alpha$ é a média de uma amostra de tamanho a de uma população com os valores $V_1, ..., V_A$.

Inferência

- Para a segunda parte, temos, de (2) e (5), que:

$$
\begin{aligned}
\mathcal{V}_1(\mathcal{E}_2(\widehat{\mu}_{2c1})) &= \mathcal{V}_1\left(\frac{1}{a}\sum_{\alpha=1}^{a}\frac{B_\alpha}{\overline{B}}\mu_\alpha\right) = \mathcal{V}_1\left(\overline{X}\right) = \frac{\sigma_x^2}{n} \\
&= \frac{1}{aA}\sum_{\alpha=1}^{A}\left(\frac{B_\alpha}{\overline{B}}\mu_\alpha - \mu\right)^2 \qquad (7)
\end{aligned}
$$

Inferência

- De (6) e (7), em (4), finalmente, temos que:

$$\mathcal{V}(\widehat{\mu}_{2c1}) = \frac{1}{aA}\sum_{\alpha=1}^{A}\left(\frac{B_\alpha}{\overline{B}}\mu_\alpha - \mu\right)^2 + \frac{1}{aA}\sum_{\alpha=1}^{A}\left(\frac{B_\alpha}{\overline{B}}\right)^2\frac{\sigma_\alpha^2}{b_\alpha}$$

ou seja, a variância do estimador é a soma da variância entre e a variância dentro dos conglomerados, diferentemente do mesmo estimador sob AC (um único estágio), cuja variância dependendia apenas da variância entre conglomerados.

Inferência

- Podemos, ainda, escrever a variância acima como

$$\mathcal{V}(\widehat{\mu}_{2c1}) = \frac{\sigma^2_{ect}}{a} + \frac{\sigma^2_{2dc}}{a\psi}$$

em que $\sigma^2_{2dc} = \frac{1}{A}\sum_{\alpha=1}^{A}\left(\frac{B_\alpha}{\overline{B}}\right)^2 \frac{\psi}{b_\alpha}\sigma^2_\alpha$ e $\psi = \frac{1}{A}\sum_{\alpha=1}^{A} b_\alpha$.

- Estimadores não viciados, para as duas variâncias acima, são dados, respectivamente, por
  - $\widehat{\sigma}^2_{2dc} = \frac{1}{a}\sum_{\alpha=1}^{a}\left(\frac{B_\alpha}{\overline{B}}\right)^2 \frac{\psi}{b_\alpha}\widehat{\sigma}^2_\alpha$, em que $\widehat{\sigma}^2_\alpha$ é a variância amostral do conglomerado $\alpha = 1, 2, ..., A$, dividida por $b_\alpha - 1$.
  - $\widehat{\sigma}^2_{ect} = \widehat{\sigma}^2_{2ect} - \frac{\widehat{\sigma}^2_{2dc}}{\psi}$, em que $\widehat{\sigma}^2_{2ect} = \frac{1}{a-1}\sum_{\alpha=1}^{a}\left(\frac{B_\alpha}{\overline{B}}\widehat{\mu}_\alpha - \widehat{\mu}_{2c1}\right)^2$

Intervalos de confiança, testes de hipótese e tamanho amostral

- Assim como no caso de AC, os resultados assintóticos usuais são válidos não somente quando a e A são suficientemente grandes, mas também quando b_α o foram, $\alpha = 1, 2, ..., A$.
- Tais resultados podem ser utilizados para consturir IC's, testes de hipóteses e determinar tamanhos de amostra de modo semelhante ao que foi feito para AC, veja aqui.

Coeficiente de correlação intraclasse e EPA

- Pesquisar sobre os dois outros estimadores (usuais) e como os três se comportam sob conglomerados de igual tamanho (populacional e amostralmente), bem como sob conglomerados de tamanhos diferentes.
- Sob conglomerados de mesmo tamanho ($B_\alpha = B$) (populacional) e ($b_\alpha = b$) (sorteado), $\forall \alpha$, temos que

$$
\begin{aligned}
(AC2E)EPA(\widehat{\mu}_{2c}) &= 1 + \rho_{int}(b-1) \\
(AC)EPA(\widehat{\mu}_{c}) &= 1 + \rho_{int}(B-1)
\end{aligned}
$$

Coeficiente de correlação intraclasse e EPA

- Em que $\widehat{\mu}_{2c}$ e $\widehat{\mu}_{c}$ são, respectivamente, o estimador (aos quais se igualam os três) sob os planos AC2E e AC, quando todos os conglomerados (populacionais e amostrais) possuem o mesmo tamanho.
- Note que, se $\rho_{int} > 0$, então:

$$\frac{\mathcal{V}_{AC2E_1}(\widehat{\mu}_{2c})}{\mathcal{V}_{AC_1}(\widehat{\mu}_{c})} = \frac{1+\rho_{int}(b-1)}{1+\rho_{int}(B-1)} \leq 1$$

- Pesquisar sobre como os resultados se apresentam no caso de $AC2E_2$.