Notas de Aula
02 - Definições e Notações Básicas

1. População

   População ou Universo é o conjunto $\mathcal{U}$ de todas as unidades elementares de interesse, indicado por $\mathcal{U} = \{1, \ldots, N\}$, em que N é o tamanho da população, fixo, finito e as vezes desconhecido.

   Elemento populacional (unidade elementar) é qualquer $i \in \mathcal{U}$.

   A cada elemento populacional está associada uma variável $Y_i, i \in \mathcal{U}$, ou um vetor de variáveis $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{ip}), i \in \mathcal{U}$, que indicam a(s) **característica(s) de interesse**.

   Parâmetros populacionais

   O parâmetro populacional é o vetor $\boldsymbol{D} = (Y_1, \ldots, Y_N)$, correspondente a todos os valores de uma única característica de interesse ou a matriz $\boldsymbol{D} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N)$,

   Uma função paramétrica populacional - também chamado de parâmetro populacional - é uma função numérica $\theta(\boldsymbol{D})$.

   Exercício 1: Considere a população formada por três domicílios $\mathcal{U} = \{1, 2, 3\}$, que estão sendo observadas as seguintes variáveis: nome (do chefe), sexo, idade, fumante ou não, renda bruta (mensal em s.m.) familiar e o número de trabalhadores. Vamos definir alguns parâmetros e funções paramétricas populacionais.

   Para o caso de uma única variável, os principais parâmetros são

   (a) Total populacional

   $$\theta(\boldsymbol{D}) = \tau = \sum_{i=1}^{N} Y_i.$$

   (b) Média populacional

   $$\theta(\boldsymbol{D}) = \mu = \overline{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i = \frac{\tau}{N}.$$

   (c) Variância populacional

   $$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \mu)^2 \quad \text{e} \quad S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \mu)^2.$$

   Para o caso de duas variáveis, os principais parâmetros são

   (a) Covariância populacional

   $$\sigma_{XY} = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu_X)(Y_i - \mu_Y) \quad \text{e} \quad S_{XY} = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \mu_X)(Y_i - \mu_Y).$$

   (b) Correlação populacional

   $$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}.$$

   (c) Razão populacional

   $$R = \frac{\tau_X}{\tau_Y} = \frac{\mu_X}{\mu_Y}.$$

(d) Razão média populacional

$$\overline{R} = \frac{1}{N}\sum_{i=1}^{N}\frac{Y_i}{X_i}.$$

Relações importantes

(a) Das definições de μ e τ segue

$$\tau = N\mu.$$

(b) Das definições de σ^2 e s^2 segue que

$$S^2 = \frac{N}{N-1}\sigma^2.$$

(c)

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} Y_i^2 - \mu^2.$$

(d)

$$\sigma_{XY} = \frac{1}{N}\sum_{i=1}^{N} X_i Y_i - \mu_X\mu_Y.$$

(e)

$$\sum_{i\neq j}^{N} Y_i Y_j = N^2\mu^2 - \sum_{i=1}^{N} Y_i^2.$$

2. Amostra

Considere uma população de tamanho fixo $\mathcal{U} = \{1, 2, \ldots, N\}$. Uma sequência qualquer de n unidades de $\mathcal{U}$ é denominada uma amostra ordenada de $\mathcal{U}$, isto é,

$$\boldsymbol{s} = (k_1, \ldots, k_n), k_i \in \mathcal{U}.$$

Exercício 2: Seja $\mathcal{U} = \{1, 2, 3\}$. Vamos definir algumas amostras ordenadas de $\mathcal{U}$.

Vamos considerar as seguintes funções:

(i) $f_i(\boldsymbol{s})$ = número de vezes que o $i-$ésimo elemento populacional aparece na amostra $\boldsymbol{s}$.

(ii) $\delta_i(\boldsymbol{s})$ = variável binária que indica a presença ou ausência do $i-$ésimo elemento populacional na amostra $\boldsymbol{s}$:

$$\delta_i(\boldsymbol{s}) = \begin{cases} 1, & \text{se } i \in \boldsymbol{s} \\ 0, & \text{caso contrário.} \end{cases}$$

A partir delas definimos o tamanho da amostra como

$$n(\boldsymbol{s}) = \sum_{i=1}^{N} f_i(\boldsymbol{s}),$$

e o tamanho efetivo da amostra como

$$\nu(\boldsymbol{s}) = \sum_{i=1}^{N} \delta_i(\boldsymbol{s}).$$

Vamos indicar por $S(\mathcal{U})$ o conjunto de todas as possíveis amostras ordenadas (espaço amostral) e $S_n(\mathcal{U})$ a subclasse de todas as amostras ordenadas de tamanho n.

Planejamento amostral

Um planejamento amostral ordenado é uma função $P(\boldsymbol{s})$ definida em $\mathcal{S}(\mathcal{U})$ que satisfazendo

$$P(\boldsymbol{s}) \geq 0, \forall \boldsymbol{s} \in S(\mathcal{U}),$$

e tal que

$$\sum_{\{\boldsymbol{s}:\boldsymbol{s} \in S\}} P(\boldsymbol{s}) = 1.$$

Em geral, na definição de um planejamento amostral, costuma-se restringir S à subclasse S_A, que contém as amostras $\boldsymbol{s}$ tais que $P(\boldsymbol{s}) > 0$.

Exercício 3: Considere os dados referentes ao Exemplo 1, em que $\mathcal{U} = \{1, 2, 3\}$ e

$$S(\mathcal{U}) = \{(1), (2), (3), (1, 1), (1, 2), (1, 3), \ldots, (2, 2, 1, 3, 2), \ldots\}.$$

Considere o plano amostral A: (1) Sorteamos um elemento de $\mathcal{U}$ com probabilidade proporcional ao número de trabalhadores; (2) Sem repormos o domicílio, sorteamos um segundo também com probabilidade proporcional ao número de trabalhadores. Apresente o plano amostral especificado.

Podemos construir outros?

Tipos de amostragem

(a) Amostragem Aleatória Simples (AAS) - selecionamos sequencialmente cada unidade amostral com igual probabilidade, de tal forma que cada amostra tenha a mesma chance de ser escolhida. A seleção pode ser feita com ou sem reposição.

(b) Amostragem Estratificada (AE) - A população é dividida em estratos e AAS é utilizada na seleção de uma amostra de cada estrato.

(c) Amostragem por Conglomerados (AC) - A população é dividida em subpopulações distintas. Alguns dos conglomerados são selecionados via AAS e todos os elementos do conglomerado são observados.

(d) Amostragem em dois estágios (A2E) - A população é dividida em subpopulações com na AE ou na AC. Num primeiro estágio algumas subpopulações são selecionadas usando a AAS. Num 2o estágio uma amostra de unidades experimentais é selecionada de cada subpopulação sorteado no 1o passo.

(e) Amostragem Sistemática (AS) - Quando existe disponível uma listagem de indivíduos da população, pode-se sortear, por exemplo, um nome entre os 10 primeiros indivíduos e então observar todo décimo indivíduo na lista a partir do primeiro indivíduo selecionado.

Estatísticas e distribuições amostrais

A cada elemento populacional tem-se associada a característica Y_i ou $\boldsymbol{Y}_i$. Resulta que fixada uma amostra $\boldsymbol{s} = (k_1, \ldots, k_n)$, a cada elemento k_j tem-se associada a característica Y_{k_j} ou $\boldsymbol{Y}_{k_j}$.

Os dados da amostra $\boldsymbol{s}$ são ou o vetor das observações pertencentes à amostra

$$\boldsymbol{d}_s = (Y_{k_1}, \ldots, Y_{k_n}), k_i \in \boldsymbol{s},$$

ou a matriz

$$\boldsymbol{d}_s = (\boldsymbol{Y}_{k_1}, \ldots, \boldsymbol{Y}_{k_n}), k_i \in \boldsymbol{s}.$$

Quando $\boldsymbol{s}$ percorre todas as amostras possíveis de um plano amostral S_A, tem-se associado um vetor aleatório $\boldsymbol{d} = \{y_1, \ldots, y_n\}$ ou uma matriz aleatória $\boldsymbol{d} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$.

Uma **estatística** é qualquer função numérica $h(\boldsymbol{d}_s)$ que relaciona as observações da amostra.

Escolhido o plano amostral A, temos associado o par (S_A, P_A). Ao definirmos uma estatística de interesse $h(\boldsymbol{d}_s)$, temos uma variável aleatória $H(\boldsymbol{d}_s)$ quando $\boldsymbol{s}$ percorre S_A.

A distribuição amostral de uma estatística $h(\boldsymbol{d}_s)$, segundo um plano amostral A, é a distribuição de probabilidade da variável aleatória $H(\boldsymbol{d}_s)$, definida sobre S_A, com função de probabilidade dada por

$$p_h = P_A(\boldsymbol{s} \in S_A; H(\boldsymbol{d}_s) = h).$$

Exercício 4: Considere os dados do Exemplo 1 e a estatística $r = h(\boldsymbol{d}_s)$ como sendo a razão entre o total da renda familiar e o número de trabalhadores na amostra. Considere o plano amostral A definido no Exemplo 3. Encontre a distribuição amostral de r.

Considere um plano amostral A, uma estatística $H(\boldsymbol{d}_s)$, $\boldsymbol{s} \in S_A$ e p_h a função de probabilidade correspondente ao plano amostral.

O valor esperado da variável será

$$E_A(H) = \sum_{\{\boldsymbol{s};\boldsymbol{s} \in S_A\}} h(\boldsymbol{d}_s) P_A(\boldsymbol{s}) = \sum_h h p_h.$$

A variância é dada por

$$Var_A(H) = \sum_{\{\boldsymbol{s};\boldsymbol{s} \in S_A\}} [h(\boldsymbol{d}_s) - E_A(H)]^2 P_A(\boldsymbol{s}).$$

A covariância entre H e G é dada por

$$Cov_A(H, G) = \sum_{\{\boldsymbol{s};\boldsymbol{s} \in S_A\}} [h(\boldsymbol{d}_s) - E_A(H)][g(\boldsymbol{d}_s) - E_A(G)] P_A(\boldsymbol{s}).$$

Exercício 5: Calcule $E_A(r)$ e $Var_A(r)$ com base no Exemplo 4.

Definido um plano amostral A, as variáveis $f_i(\boldsymbol{s})$ e $\delta_i(\boldsymbol{s})$ também possuem uma distribuição de probabilidade.

Exercício 6: Encontre as distribuições de $f_i(\boldsymbol{s})$ e $\delta_i(\boldsymbol{s})$, $i = 1, 2, 3$ para o plano amostral A definido no Exemplo 3. Calcule $E(f_i)$ e $E(\delta_i)$, $i = 1, 2, 3$.

Denotaremos por $\pi_i(A)$ a probabilidade do i-ésimo elemento de $\mathcal{U}$ pertencer a amostra segundo o planejamento A e $\pi_{ij}(A)$ a probabilidade do i-ésimo e j-ésimo elementos de $\mathcal{U}$ pertencerem à amostra. Deste modo:

$$\pi_i(A) = P_A(\delta_i = 1) = \sum_{\boldsymbol{s}:\boldsymbol{s} \subset i} P_A(\boldsymbol{s}),$$

e

$$\pi_{ij}(A) = \sum_{\boldsymbol{s}:\boldsymbol{s} \subset (i,j)} P_A(\boldsymbol{s}).$$

Exercício 7: Calcule π_i e π_{ij} para os dados do Exemplo 6.

Estimadores e suas propriedades

O principal objetivo da amostragem é produzir estimadores para parâmetros populacionais. Quando se associa uma estatística a expressão que irá "estimar" um parâmetro, ela recebe o nome de estimador. O valor numérico para cada amostra chama-se estimativa.

Um estimado é dito **não viciado** segundo o plano amostral A, se $E_A(\hat{\theta}) = \theta$. Caso ele seja viciado, o erro quadrático médio é dado por $EQM_A(\hat{\theta}) = Var_A(\hat{\theta}) + (E_A(\hat{\theta}) - \theta)$.