

Técnicas de Amostragem - Aula 02

Definições e Notações Básicas II: Estatísticas, Distribuições Amostrais, Estimadores

Kaique Matias de Andrade Roberto

Ciências Atuariais

HECSA - Escola de Negócios

FIAM-FAAM-FMU

1. Conceitos que aprendemos em Aulas anteriores
2. Estatísticas e Distribuições Amostrais
3. Estimadores e Propriedades
4. Comentários Finais
5. Referências

Conceitos que aprendemos em Aulas anteriores

Conceitos que aprendemos em Aulas anteriores

Na aula passada nós lidamos com uma parte das principais definições, nomenclaturas e terminologias da Teoria de Amostragem, no caso:

- população;
- amostra;
- planejamento amostral.

Conceitos que aprendemos em Aulas anteriores

Vamos recapitular alguns exemplos/conceitos importantes da aula passada.

Exemplo 1.1

Considere a população formada por três domicílios \mathcal{U} e que estão sendo observadas as seguintes variáveis: nome (do chefe), sexo, idade, fumante ou não, renda bruta familiar e número de trabalhadores. Os dados estão na planilha "aula-02-exemplo".

Conceitos que aprendemos em Aulas anteriores

Vamos usar as notações

Unidade	i
Nome do Chefe	A_i
Sexo	X_i
Idade	Y_i
Fumante	G_i
Renda Bruta Familiar	F_i
Número de Trabalhadores	T_i

Definição 1.2

Uma função $P(\mathbf{s})$ definida em $\mathcal{S}(\mathcal{U})$, satisfazendo

$$P(\mathbf{s}) \geq 0, \text{ para quaisquer } \mathbf{s} \in \mathcal{S}(\mathcal{U})$$

e tal que

$$\sum_{\mathbf{s} \in \mathcal{S}(\mathcal{U})} P(\mathbf{s}) = 1,$$

é chamado um **planejamento amostral ordenado**.

Exemplo 1.3

Considere $\mathcal{U} = \{1, 2, 3\}$ (vide planilha "aula-01-exemplo") e a seguinte regra de sorteio:

- i - sorteia-se com igual probabilidade um elemento de \mathcal{U} , e anota-se a unidade sorteada;
- ii - este elemento é devolvido à população e sorteia-se um segundo elemento do mesmo modo.

Este é o mesmo plano amostral do Plano A. Este plano é conhecido como **amostragem aleatória simples com reposição (AASc)**.

Plano A

$$P(11) = P(12) = P(13) = 1/9$$

$$P(21) = P(22) = P(23) = 1/9$$

$$P(31) = P(32) = P(33) = 1/9$$

$$P(s) = 0, \text{ para as demais } s \in \mathcal{S}.$$

Exemplo 1.4

Considere $\mathcal{U} = \{1, 2, 3\}$ (vide planilha "aula-01-exemplo") e a seguinte regra de sorteio:

- i - sorteia-se com igual probabilidade um elemento de \mathcal{U} , e anota-se a unidade sorteada;
- ii - este elemento é retirado da população e sorteia-se um segundo elemento do mesmo modo.

Este é o mesmo plano amostral do Plano A. Este plano é conhecido como **amostragem aleatória simples sem reposição (AASs)**.

Plano B

$$P(12) = P(13) = P(21) = P(23) = P(31) = P(32) = 1/6$$

$P(s) = 0$, para as demais $s \in \mathcal{S}$.

Estatísticas e Distribuições Amostrais

Como já foi discutido, o objetivo principal da amostragem é adquirir conhecimentos sobre variáveis (características) de interesse, e desse modo, é necessário caracterizar as variáveis de interesse também na amostra.

Assim, associado a cada unidade elementar i tem-se uma característica Y_i , que pode ser reunida na matriz (ou vetor) de dados populacionais D .

Agora, fixada uma amostra \mathbf{s} ,

$$\mathbf{s} = (k_1, k_2, \dots, k_n)$$

sabe-se que associado a cada elemento k_j tem-se um vetor de características \mathbf{Y}_{k_j} .

Definição 2.1

Chama-se de dados da amostra \mathbf{s} à matriz ou vetor de observações pertencentes a amostra, isto é,

$$\mathbf{d}_{\mathbf{s}} = (Y_{k_1}, Y_{k_2}, \dots, Y_{k_n}) = (Y_{k_j}, k_j \in \mathbf{s}).$$

Quando \mathbf{s} percorre todos os pontos possíveis de um plano amostral \mathcal{S}_A , tem-se associado um vetor aleatório que será representado por

$$\mathbf{d} = \mathbf{y} = (y_1, \dots, y_n)$$

onde y_i é a variável aleatória que indica os possíveis valores que podem ocorrer na i -ésima posição.

Quando as observações são multidimensionais, os dados da amostra passam a ser a matriz

$$\mathbf{d}_s = (\mathbf{Y}_{k_i}, i \in s),$$

e tem-se associado a matriz aleatória

$$\mathbf{d} = (\mathbf{y}_1, \dots, \mathbf{y}_n).$$

Definição 2.2

Qualquer característica numérica dos dados correspondentes a amostra \mathbf{s} é chamada estatística, ou seja, qualquer função $h(\mathbf{d}_s)$ que relaciona as observações da amostra \mathbf{s} .

Exemplo 2.3

Agora considere os dados na planilha "aula-02-exemplo" e a amostra $s = (12)$. Desse modo, tem-se para o vetor $\begin{pmatrix} F_i \\ T_i \end{pmatrix}$ a seguinte matriz de dados da amostra:

$$\begin{pmatrix} 12 & 30 \\ 1 & 3 \end{pmatrix}.$$

As médias

$$\bar{f} = \frac{12 + 30}{2} = 21 \text{ e } \bar{t} = \frac{1 + 3}{2} = 2,$$

ou a razão

$$r = \frac{12 + 30}{1 + 3} = 10,5$$

são exemplos de estatísticas calculadas na amostra $\mathbf{s} = (12)$.

Escolhido um plano amostral A , tem-se associado o par (S_A, P_A) dos respectivos pontos amostrais e suas probabilidades.

Fixada agora uma estatística $h(\mathbf{d}_s)$, quando \mathbf{s} percorre \mathcal{S}_A , ter-se-á associado uma variável aleatória $H(\mathbf{d}_s)$ associada ao par (\mathcal{S}_A, P_A) .

Considere também a notação

$$p_h = P_A(\mathbf{s} \in \mathcal{S}_A; H(\mathbf{d}_s) = h),$$

que denota a probabilidade sobre o conjunto de todas as amostras \mathbf{s} tais que $H(\mathbf{d}_s) = h$.

Conhecendo-se todos os valores de h e as suas respectivas probabilidades, tem-se bem identificada a (distribuição da) variável aleatória H .

Definição 2.4

A distribuição amostral de uma estatística $h(\mathbf{d}_s)$ segundo um plano amostral A , é a distribuição de probabilidades de $H(\mathbf{d}_s)$, definida sobre \mathcal{S}_A , com função de probabilidade dada por

$$p_h = P_A(\mathbf{s} \in \mathcal{S}_A; H(\mathbf{d}_s) = h) = P(h).$$

Exemplo 2.5

Para os dados na planilha "aula-02-exemplo" com dados amostrais

$$\mathbf{D} = \begin{pmatrix} F_i \\ T_i \end{pmatrix} = \begin{pmatrix} 12 & 30 & 18 \\ 1 & 3 & 2 \end{pmatrix}, i \in \mathcal{U},$$

considere a estatística $r = h(\mathbf{d}_s)$ como sendo a razão entre o total da renda familiar e o número de trabalhadores na amostra. Considere também os planos amostrais A e B . Temos as seguintes distribuições amostrais:

Figura 1: Plano amostral A (AASc)

$s:$	11	12	13	21	22	23	31	32	33
$P(s):$	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9
$h(d_s) = r:$	12	10,5	10	10,5	10	9,6	10	9,6	9

Figura 2: Distribuição amostral de r na AASc

$h:$	9	9,6	10	10,5	12
$p_h:$	1/9	2/9	3/9	2/9	1/9

Figura 3: Plano amostral B (AASs)

$\mathbf{s}:$	12	13	21	23	31	32
$P(\mathbf{s}):$	1/6	1/6	1/6	1/6	1/6	1/6
$h(\mathbf{d}_s) = r:$	10,5	10	10,5	9,6	10	9,6

Figura 4: Distribuição amostral de r na AASs

$h:$	9,6	10	10,5
$p_h:$	$1/3$	$1/3$	$1/3$

A distribuição amostral, e conceitos derivados são básicos para uso e avaliação inteligente dos procedimentos amostrais.

Eles serão usados aqui para avaliar as propriedades e vantagens de um plano amostral, e/ou estatística, sobre seus concorrentes.

Considere dados: um plano amostral A , uma estatística $h(\mathbf{d}_s)$, $\mathbf{s} \in \mathcal{S}_A$ e seja p_h a função de probabilidade correspondente ao plano amostral.

Definição 2.6

O **valor esperado (esperança)** da variável H será

$$E_A[H] = \sum_{s \in \mathcal{S}_A} h(\mathbf{d}_s) P_A(\mathbf{s}).$$

Definição 2.7

A **variância** da variável H será

$$\text{Var}_A[H] = \sum_{s \in \mathcal{S}_A} (h(\mathbf{d}_s) - E_A[H])^2 P_A(\mathbf{s}).$$

Definição 2.8

Quando houver duas estatísticas $H(\mathbf{d}_s)$ e $G(\mathbf{d}_s)$, a **covariância e correlação** são respectivamente

$$\text{Cov}_A(H, G) = \sum_{s \in \mathcal{S}_A} (h(\mathbf{d}_s) - E_A[H])(g(\mathbf{d}_s) - E_A[G])P_A(\mathbf{s});$$

$$\text{Corr}_A(H, G) = \frac{\text{Cov}_A(H, G)}{\sqrt{\text{Var}_A[H]}\sqrt{\text{Var}_A[G]}}.$$

Exemplo 2.9

Considere os dados na planilha "aula-02-exemplo" com dados amostrais

$$\mathbf{D} = \begin{pmatrix} F_i \\ T_i \end{pmatrix} = \begin{pmatrix} 12 & 30 & 18 \\ 1 & 3 & 2 \end{pmatrix}, i \in \mathcal{U}$$

e plano amostral AASc. Agora considere as estatísticas r e \bar{f} . Vamos calcular

$$E_{AASc}[r], E_{AASc}[\bar{f}], \text{Var}_{AASc}[r], \\ \text{Var}_{AASc}[\bar{f}], \text{Cov}_{AASc}[r, \bar{f}] \text{ e } \text{Corr}_{AASc}[r, \bar{f}].$$

Definido um plano amostral A , as variáveis $f_i(\mathbf{s})$ e $\delta_i(\mathbf{s})$ também passam a possuir uma distribuição de probabilidades associada, que indicaremos por $f_i(A)$ e $\delta_i(A)$.

Exemplo 2.10

Considere o plano amostral A (com os mesmo dados). Para cada amostra, tem-se associado as variáveis $f_1, f_2, f_3, \delta_1, \delta_2, \delta_3$ cujos valores e respectivas probabilidades são dados por:

Figura 5: Distribuições amostrais de $f_1, f_2, f_3, \delta_1, \delta_2, \delta_3$ na AASc

s:	11	12	13	21	22	23	31	32	33
$P(\mathbf{s}):$	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9
$f_1:$	2	1	1	1	0	0	1	0	0
$f_2:$	0	1	0	1	2	1	0	1	0
$f_3:$	0	0	1	0	0	1	1	1	2
$\delta_1:$	1	1	1	1	0	0	1	0	0
$\delta_2:$	0	1	0	1	1	1	0	1	0
$\delta_3:$	0	0	1	0	0	1	1	1	1

Figura 6: Distribuição de f_1 na AASc

$h(\mathbf{d}_s) = f_1:$	0	1	2
$p_h:$	4/9	4/9	1/9

Figura 7: Distribuição de δ_1 na AASc

$h(\mathbf{d}_s) = \delta_1:$	0	1
$p_h:$	4/9	5/9

Vamos verificar que f_2 e f_3 tem a mesma distribuição que f_1 , δ_2 e δ_3 tem a mesma distribuição que δ_1 , e

$$E_A[f_1] = 2/3 \text{ e } E_A[\delta_1] = 5/9.$$

Estimadores e Propriedades

O objetivo principal da amostragem é produzir estimadores para parâmetros populacionais desconhecidos.

Isso é feito escolhendo-se uma estatística que tenha propriedades convenientes em relação ao parâmetro populacional.

Quando associa-se uma estatística com a expressão que irá "estimar" o parâmetro populacional ela recebe o nome de **estimador**. O valor numérico do estimador, para dada amostra, chama-se **estimativa**.

Simbolicamente, o objetivo é estimar um parâmetro populacional $\theta(\mathbf{D})$. Isto será feito através de uma estatística obtida a partir dos dados amostrais \mathbf{d}_s , o estimador que será representado por $\hat{\theta}(\mathbf{d}_s)$.

Como já foi discutido, as propriedades de um estimador dependem da sua distribuição amostral, e as principais qualidades procuradas em amostragem são: pequenos vieses (vícios) e pequenas variâncias.

Definição 3.1

Um estimador é dito **não-viciado** segundo um plano amostral A se

$$E_A[\hat{\theta}] = \theta.$$

Definição 3.2

O **viés** de um estimador $\hat{\theta}(\mathbf{d}_s)$ segundo um plano amostral A , é dado por

$$B_A[\hat{\theta}] = E_A[\hat{\theta} - \theta] = E_A[\hat{\theta}] - \theta;$$

e o **erro quadrático médio** por

$$\text{EQM}_A[\theta] = E_A[\hat{\theta} - \theta]^2.$$

Com essas definições verifica-se que

$$\text{EQM}_A[\theta] = \text{Var}_A[\theta] + B_A^2[\hat{\theta}].$$

Observe que para uma amostra particular \mathbf{s} , a diferença $\hat{\theta}(\mathbf{s}) - \theta$ mostra o desvio entre o valor estimado e o valor que se desejaria conhecer, ou seja, o erro cometido pelo uso da amostra e do estimador $\hat{\theta}$ para estimar a quantidade de interesse (parâmetro) θ .

Esse desvio é usualmente conhecido por **erro amostral**. Para dada amostra, o erro amostral só pode ser calculado, na situação improvável de θ ser conhecido.

Por isso, a estratégia da avaliação da amostragem não é julgar o resultado particular de uma amostra, mas do plano amostral. Em outras palavras, queremos avaliar as propriedades do estimador sob a ótica de um plano amostral A .

Exemplo 3.3

Usando os dados da planilha "aula-02-exemplo" (com as notações adotadas até então) temos

$$E_{AASc}(r) \cong 10.13 \text{ e } \text{Var}_{AASc}(r) \cong 0.6289.$$

Suponha que o parâmetro de interesse seja a renda média por trabalhador, R , ou seja,

$$R = \frac{12 + 30 + 18}{1 + 3 + 2} = \frac{60}{6} = 10.$$

Observa-se então que r é um estimador viesado para R , pois

$$E_{AASc}(r) \cong 10.13 \neq 10 = R.$$

O vício é dado por

$$B_{AASc}(r) \cong 10.13 - 10 = 0.13,$$

de modo que

$$EMQ_{AASc}(r) \cong 0.6289 + 0.13^2 = 0.6458.$$

Exemplo 3.4

Com os mesmos dados do Exemplo anterior, suponha agora que o parâmetro de interesse seja a renda média familiar $\mu_F = 20$. Observe que

$$E_{AASc}(\bar{f}) = 20 \text{ e } \text{Var}_{AASc}(\bar{f}) = 28.$$

Isso implica que \bar{f} não é viciado para μ_F , ou seja, $B_{AASc}(\bar{f}) = 0$, de modo que

$$\text{EMQ}_{AASc}(\bar{f}) = \text{Var}_{AASc}(\bar{f}) = 28.$$

Comentários Finais

Em resumo, na aula de hoje nós finalizamos a descrição das nomenclaturas, ou seja, falamos de:

- estatísticas;
- distribuições amostrais;
- estimadores e suas propriedades.

Nas próximas aulas nós vamos focar na Amostragem Aleatória Simples com Reposição.

ATIVIDADE PARA ENTREGAR (E COMPOR A NOTA N1)

Resolva em grupos de até 4 integrantes os Exercícios 2.1-2.7.

Referências



