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

16/09 => NÃO HAVERÁ AULA!

Motivo: estarei no congresso EBL (encontro brasileiro da lógica).

19:00 - 21:50
19:10 - 21:30

18:45 - 22:10

Conteúdo

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. ← Probabilidades

Conceitos que aprendemos em Aulas anteriores

Vamos recapitular alguns exemplos/conceitos importantes da aula passada.

Conceitos que aprendemos em Aulas anteriores

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 $P(\boldsymbol{s})$ definida em $\mathcal{S}(\mathcal{U})$, satisfazendo

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

e tal que

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

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)**.

Conceitos que aprendemos em Aulas anteriores

Plano A

$$P(11) = P(12) = P(13) = 1/9$$
$$P(21) = P(22) = P(23) = 1/9$$
$$P(31) = P(32) = P(33) = 1/9$$
$$P(\boldsymbol{s}) = 0, \text{ para as demais } \boldsymbol{s} \in \mathcal{S}.$$

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)**.

Conceitos que aprendemos em Aulas anteriores

Plano B

$$P(12) = P(13) = P(21) = P(23) = P(31) = P(32) = 1/6$$
$$P(\boldsymbol{s}) = 0, \text{ para as demais } \boldsymbol{s} \in \mathcal{S}.$$

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 $\boldsymbol{Y}_i$, que pode ser reunida na matriz (ou vetor) de dados populacionais $\boldsymbol{D}$.

Agora, fixada uma amostra $\boldsymbol{s}$,

$$\boldsymbol{s} = (k_1, k_2, ..., k_n)$$

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

Definição 2.1

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

$$\boldsymbol{d}_s = (Y_{k_1}, Y_{k_2}, ..., Y_{k_n}) = (Y_{k_j},\ k_j \in \boldsymbol{s}).$$

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

$$\boldsymbol{d} = \boldsymbol{y} = (y_1, ..., y_n)$$

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

$$\boldsymbol{d_s} = (\boldsymbol{Y}_{k_i},\ i \in \boldsymbol{s}),$$

e tem-se associado a matriz aleatória

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

Definição 2.2

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

Exemplo 2.3

Agora considere os dados na planilha "aula-02-exemplo"e a amostra $\boldsymbol{s} = (12)$. 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

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

ou a razão

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

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

Escolhido um plano amostral A, tem-se associado o par $(\mathcal{S}_A, P_A)$ dos respectivos pontos amostrais e suas probabilidades.

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

Considere também a notação

$$p_h = P_A(\boldsymbol{s} \in \mathcal{S}_A;\, H(\boldsymbol{d_s}) = h),$$

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

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

$$p_h = P_A(\boldsymbol{s} \in \mathcal{S}_A;\, H(\boldsymbol{d_s}) = h) = P(h).$$

1 - Dados amostrais

2 - Identificar / Escolher a estatística (a conta que fazemos com a amostra).

3 - Nomeamos a estatística, digamos h

4 - Calculamos $h(s)$ para todas as amostras no plano amostral.

5 - Identifica os possíveis valores para $h(s)$

6 - Calculamos $P(h(s) = h_0)$ para todos os possíveis valores h_0.

Exemplo 2.5

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

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

considere a estatística $r = h(\boldsymbol{d_s})$ 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:

$\boxed{AASc}$ Plano A = AASc

$P(11) = P(12) = P(13) = P(21) = P(22) =$

$P(23) = P(31) = P(32) = P(33) = 1/9$

$P(s) = 0$ caso contrário

$$D = \begin{pmatrix} F_i \\ T_i \end{pmatrix} = \begin{pmatrix} 12 & 30 & 14 \\ 1 & 3 & 2 \end{pmatrix}$$

$$r = \frac{\sum F_i}{\sum T_i}$$

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
$r(s)$	12	10	10.5	10	9	9.6	10.5	9.6	10

$$d(11) = \begin{pmatrix} 12 & 12 \\ 1 & 1 \end{pmatrix}$$

$$r(11) = \frac{12+12}{1+1} = 12$$

$$d(12) = \begin{pmatrix} 12 & 18 \\ 1 & 2 \end{pmatrix}$$

$$r(12) = \frac{12+18}{1+2} = \frac{30}{3} = 10$$

$$d(13) = \begin{pmatrix} 12 & 30 \\ 1 & 3 \end{pmatrix}$$

$$r(13) = \frac{12+30}{1+3} = \frac{42}{4} = 10.5$$

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
Y(s)	12	10	10.5	10	9	9.6	10.5	9.6	10

$$P(A) = \frac{\text{nº de caso favoráveis}}{\text{nº de casos possíveis}}$$

$r(s)$	9	9.6	10	10.5	12
$P(r(s)=h)$	$1/9$	$2/9$	$3/9$	$2/9$	$1/9$

$$P(r(s)=9) = \frac{1}{9}$$

$$P(r(s)=9.6) = \frac{2}{9}$$

Figura 1: Plano amostral *A* (AASc)

$\mathbf{s}$:	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
$h(\mathbf{d_s}) = r$:	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

Variável Aleatória Discreta

Figura 4: Distribuição amostral de *r* na AASs

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

$$E_{AAS_s}[r] = 9.6 \cdot \frac{1}{3} + 10 \cdot \frac{1}{3} + 10.5 \cdot \frac{1}{3}$$

$$= \frac{9.6 + 10 + 10.5}{3} = \frac{30.1}{3} = \boxed{10.03}$$

Estatísticas e Distribuições Amostrais

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.

Estatísticas e Distribuições Amostrais

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

Definição 2.6

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

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

Definição 2.7

A **variância** da variável H será

$$\mathrm{Var}_A[H] = \sum_{s \in \mathcal{S}_A} (h(\boldsymbol{d_s}) - E_A[H])^2 P_A(\boldsymbol{s}).$$

Definição 2.8

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

$$\mathrm{Cov}_A(H,G) = \sum_{s \in \mathcal{S}_A} (h(\boldsymbol{d_s}) - E_A[H])(g(\boldsymbol{d_s}) - E_A[G]) P_A(\boldsymbol{s});$$

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

Exemplo 2.9

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

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

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

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

Definido um plano amostral A, as variáveis $f_i(\boldsymbol{s})$ e $\delta_i(\boldsymbol{s})$ 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

$\mathbf{s}$:	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
δ_1:	1	1	1	1	0	0	1	0	0
δ_2:	0	1	0	1	1	1	0	1	0
δ_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

Estimadores e Propriedades

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

Estimadores e Propriedades

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

Estimadores e Propriedades

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 $\theta(\boldsymbol{D})$. Isto será feito através de uma estatística obtida a partir dos dados amostrais $\boldsymbol{d_s}$, o estimador que será representado por $\hat{\theta}(\boldsymbol{d_s})$.

Estimadores e Propriedades

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 $\hat{\theta}(\boldsymbol{d_s})$ 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

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

Estimadores e Propriedades

Com essas definições verifica-se que

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

Estimadores e Propriedades

Observe que para uma amostra particular $\boldsymbol{s}$, a diferença $\hat{\theta}(\boldsymbol{s}) - \theta$ 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.

Estimadores e Propriedades

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.

"Filosofando"

- Temos um parâmetro populacional θ (média, variância, razão...) desconhecido/inacessível.

- realizamos uma amostragem dentro de um plano amostral A e para $s \in S_A$, calculamos a estimativa $\hat{\theta}(s)$.

Exemplo 3.3

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

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

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

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

Exemplo 3.4

Com os mesmos dados do Exemplo anterior, suponha agora que o parâmetro de interesse seja a renda média familiar $\mu_F = 10$. Observe que

$$E_{AASc}(\bar{f}) = 10 \text{ e } \mathrm{Var}_{AASc}(\bar{f}) = 28.$$

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

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

Comentários Finais

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.

Comentários Finais

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

Comentários Finais

ATIVIDADE PARA ENTREGAR (E COMPOR A NOTA N1)

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

12.2 AASs

$$D = \begin{pmatrix} 12 & 30 & 18 \\ 1 & 3 & 2 \end{pmatrix} = \begin{pmatrix} F_i \\ T_i \end{pmatrix}$$

↳ dados amostrais

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

↳ Plano amostral

$$r = \frac{\sum F_i}{\sum T_i}$$ ← estatística

s	(12)	(13)	(21)	(23)	(31)	(32)
$P(s)$	1/6	1/6	1/6	1/6	1/6	1/6
$r(s)$	10	10.5	10	9.6	10.5	9.6

$r(s)$	9.6	10	10.5
$P(r(s))$	1/3	1/3	1/3

$r(s)$	9.6	10	10.5
$P(r(s))$	$1/3$	$1/3$	$1/3$

distribuição amostral de r associada ao AAS_s

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

razão populacional

r (com AAS_s) é viciado p/ R?

$$E_{AAS_s}(r) = 9.6 \cdot \frac{1}{3} + 10 \cdot \frac{1}{3} + \frac{10.5}{3}$$

$$= \frac{9.6 + 10 + 10.5}{3} = 10.03$$

Como $E_{AAS_s}(r) \neq R$ r é viesado para R (com relação ao AAS$_s$).

$$B_{AAS_s}(r) = E_{AAS_s}(r) - R$$

$$= 10.03 - 10 = 0.03$$

$$EQM_{AAAs}(r) = Var_{AAS_s}(r) - E_{AASs}(r)^2$$

$$Var_{AAS_s}(r) = (9.6 - 10.03)^2 \cdot \frac{1}{3} +$$

$$(10 - 10.03)^2 \cdot \frac{1}{3} + (10.5 - 10.03)^2 \cdot \frac{1}{3}$$

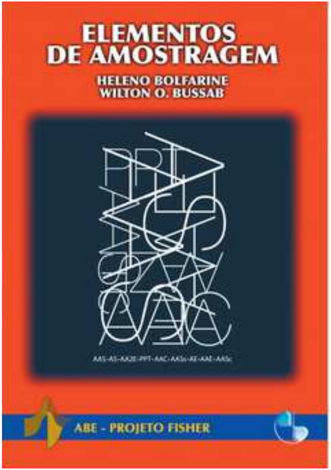
$$= \frac{1.57^2 + 0.03^2 + 0.47^2}{3} = 0.8956$$

$$EQM_{AAS_5}(Y) = 0.8956 + 0.03^2$$

$$= 0.8965$$

Referências

Referências



Bons Estudos!