Técnicas de Amostragem - Aula 01

Definições e Notações Básicas I: População, Amostra e Planejamento Amostral

Kaique Matias de Andrade Roberto

Ciências Atuariais

HECSA - Escola de Negócios

FIAM-FAAM-FMU

Conteúdo

Conceitos que aprendemos em Aulas anteriores

Conceitos que aprendemos em Aulas anteriores

- recapitulamos os tipos de variáveis;
- aprendemos como classificar variáveis;
- recapitulamos alguns tópicos de Estatística Descritiva;
- relembramos o conceito de Variável Aleatória;
- listamos as principais distribuições discretas e contínuas;
- tivemos um primeiro contato com os conceitos de Amostragem.

População

Definição 2.1

A **População ou Universo** é o conjunto $\mathcal{U}$ de todas as unidades elementares de interesse. É indicado por

$$\mathcal{U} = \{1, 2, ..., N\}$$

onde N é o tamanho fixo (e na maioria dos casos, desconhecido) da população.

População

Definição 2.2

Elemento Populacional é a nomenclatura usada para denotar qualquer elemento $i \in \mathcal{U}$. É também conhecido por **Unidade Elementar**.

População

Definição 2.3

Característica(s) de Interesse é a nomenclatura que será usada para denotar a variável ou vetor de informações associado a cada elemento da população. Será representada por

$$Y_i,\ i \in \mathcal{U},$$

ou no caso multivariado,

$$\boldsymbol{Y}_i = (Y_{i1}, ..., Y_{ip}),\ i \in \mathcal{U}.$$

População

Definição 2.4

Parâmetro Populacional é a nomenclatura utilizada para denotar o vetor correspondente a todos os valores de uma variável de interesse que denota-se por

$$\boldsymbol{D} = (Y_1, ..., Y_N),$$

no caso de uma única característica de interesse, e pela matriz

$$\boldsymbol{D} = (\boldsymbol{Y}_1, ..., \boldsymbol{Y}_N)$$

no caso em que para cada unidade da população tem-se associado um vetor $\boldsymbol{Y}_i$ de características de interesse.

Definição 2.5

Função Paramétrica Populacional é uma característica numérica qualquer da população, ou seja, uma expressão numérica que condensa funcionalmente os Y_i's (ou $\boldsymbol{Y}_i$'s), $i \in \mathcal{U}$. Tal função será denotada por

$$\theta(\boldsymbol{D}).$$

População

Esta função pode ser, por exemplo, o total, as médias, ou ainda o quociente de dois totais. É comum utilizar-se a expressão parâmetro populacional de interesse, ou simplesmente parâmetro populacional.

Exemplo 2.6

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-01-exemplo".

População

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

População

Portanto, para os dados descritos na planilha "aula-01-exemplo", os seguintes parâmetros populacionais podem ser definidos:

População

- para a variável idade,

$$\boldsymbol{D} = (20, 30, 40) = \boldsymbol{Y};$$

População

- para o vetor $\begin{pmatrix} F_i \\ T_i \end{pmatrix}$ (renda e número de trabalhadores),

$$\begin{pmatrix} 12 & 30 & 18 \\ 1 & 3 & 2 \end{pmatrix}.$$

População

Com relação à funções paramétricas populacionais, tem-se:

População

- idade média,

$$\theta(\boldsymbol{Y}) = \theta(\boldsymbol{D}) = \frac{20 + 30 + 40}{3} = 30;$$

População

- média das variáveis renda e número de trabalhadores,

$$\theta(\boldsymbol{D}) = \begin{pmatrix} \dfrac{12+30+18}{3} \\ \dfrac{1+3+2}{3} \end{pmatrix} = \begin{pmatrix} 20 \\ 2 \end{pmatrix};$$

População

- renda média por trabalhador,

$$\theta(\boldsymbol{D}) = \frac{12 + 30 + 18}{1 + 3 + 2} = 10.$$

População

Para uma variável de interesse, os parâmetros populacionais mais usados são:

População

- total populacional,

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

População

- média populacional,

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

População

- variância populacional, representada por

$$\theta(\boldsymbol{D}) = \theta(\boldsymbol{Y}) = \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \mu)^2,$$

ou às vezes,

$$\theta(\boldsymbol{D}) = \theta(\boldsymbol{Y}) = s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \mu)^2.$$

População

Para vetores bidimensionais, isto é, duas variáveis de interesse, representadas por (X, Y) são bastante usuais os seguintes parâmetros:

População

- covariância populacional,

$$\theta(\boldsymbol{D}) = \sigma_{XY} = \text{Cov}[X, Y] = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu_X)(Y_i - \mu_Y);$$

ou às vezes

$$\theta(\boldsymbol{D}) = s_{XY} = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \mu_X)(Y_i - \mu_Y);$$

onde μ_X e μ_Y denotam as médias populacionais correspondentes às variáveis de interesse X e Y, respectivamente;

População

- correlação populacional,

$$\theta(\boldsymbol{D}) = \rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y};$$

População

- razão populacional,

$$\theta(\boldsymbol{D}) = \frac{\tau_Y}{\tau_X} = \frac{\mu_Y}{\mu_X} = R,$$

População

- razão média populacional,

$$\theta(\boldsymbol{D}) = \overline{R} = \frac{1}{N}\sum_{i=1}^{N}\frac{Y_i}{X_i}.$$

População

Lembre-se: estas características populacionais **raramente são conhecidas** e na prática, usamos procedimentos amostrais para **estimá-las**.

Amostra

Amostra

Consideremos uma população fixa

$$\mathcal{U} = \{1, 2, ..., N\}.$$

Definição 3.1

Uma sequência qualquer de n unidades de $\mathcal{U}$, é denominada **amostra ordenada** de $\mathcal{U}$, isto é,

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

A entrada k_i é chamada de i-**ésimo componente** de $\boldsymbol{s}$.

Amostra

Exemplo 3.2

Seja $\mathcal{U} = \{1, 2, 3\}$ (vide planilha "aula-01-exemplo"). Os vetores

$$\boldsymbol{s}_1 = (1, 2)$$
$$\boldsymbol{s}_2 = (2, 1)$$
$$\boldsymbol{s}_3 = (1, 1, 3)$$
$$\boldsymbol{s}_4 = (3)$$
$$\boldsymbol{s}_5 = (2, 2, 1, 3, 2)$$

são exemplos de amostras ordenadas de $\mathcal{U}$.

Amostra

Definição 3.3

Seja $f_i(\boldsymbol{s})$ a variável que indica o número de vezes (freqüência) que a i-ésima unidade populacional aparece na amostra $\boldsymbol{s}$. Seja $\delta_i(\boldsymbol{s})$ a variável binária que indica a presença ou não da i-ésima unidade na amostra $\boldsymbol{s}$, isto é,

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

Amostra

Exemplo 3.4

Seja $\mathcal{U} = \{1, 2, 3\}$ (vide planilha "aula-01-exemplo"). Considere as amostras

$$
\begin{aligned}
\boldsymbol{s}_1 &= (1, 2) \\
\boldsymbol{s}_2 &= (2, 1) \\
\boldsymbol{s}_3 &= (1, 1, 3) \\
\boldsymbol{s}_4 &= (3) \\
\boldsymbol{s}_5 &= (2, 2, 1, 3, 2)
\end{aligned}
$$

Vamos calcular $f_i(s_j)$ e $\delta_i(s_j)$.

Definição 3.5

Chama-se **tamanho** $n(\boldsymbol{s})$ da amostra s a soma das frequências das unidades populacionais na amostra, isto é,

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

Amostra

Definição 3.6

Chama-se **tamanho efetivo** $\nu(\boldsymbol{s})$ da amostra s o número de unidades populacionais na amostra, isto é,

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

Exemplo 3.7

Seja $\mathcal{U} = \{1, 2, 3\}$ (vide planilha "aula-01-exemplo"). Considere as amostras

$$\begin{aligned}
\boldsymbol{s}_1 &= (1, 2) \\
\boldsymbol{s}_2 &= (2, 1) \\
\boldsymbol{s}_3 &= (1, 1, 3) \\
\boldsymbol{s}_4 &= (3) \\
\boldsymbol{s}_5 &= (2, 2, 1, 3, 2)
\end{aligned}$$

Vamos calcular $n(s_j)$ e $\nu(s_j)$.

Definição 3.8

Seja $\mathcal{S}(\mathcal{U})$, ou simplesmente $\mathcal{S}$, o conjunto de todas as amostras (sequências ordenadas) de $\mathcal{U}$, de qualquer tamanho. E seja $\mathcal{S}_n(\mathcal{U})$, a subclasse de todas as amostras de tamanho n. Muitas vezes, $\mathcal{S}(\mathcal{U})$ é denominado **espaço amostral**.

Exemplo 3.9

Seja $\mathcal{U} = \{1, 2, 3\}$ (vide planilha "aula-01-exemplo"). Vamos descrever $\mathcal{S}(\mathcal{U})$ e $\mathcal{S}_2(\mathcal{U})$.

Algumas vezes é interessante trabalhar com amostras não ordenadas (por exemplo, considerar as amostras $(1, 2)$ e $(2, 1)$ como sendo as mesmas). No caso de amostras não ordenadas sem reposição, uma amostra é simplesmente um subconjunto de elementos de $\mathcal{U}$.

Amostra

O número de amostras ordenadas de tamanho n, com reposição, é N^n, enquanto que, sem reposição, é dado pelo coeficiente binomial $\binom{N}{n}$.

Planejamento Amostral

Planejamento Amostral

Conforme mencionado anteriormente, um dos nossos objetivos é apresentar procedimentos amostrais probabilísticos, ou seja, aqueles que permitem associar a cada amostra uma probabilidade conhecida de ser sorteada.

Planejamento Amostral

O modo como essas probabilidades são associadas é que irá definir um planejamento amostral. Isto nos leva à seguinte definição:

Planejamento Amostral

Definição 4.1

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 4.2

Considere $\mathcal{U} = \{1, 2, 3\}$ (vide planilha "aula-01-exemplo").
Considere os seguintes exemplos de planejamentos amostrais:

Plano A

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

Plano B

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

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

Plano C

$$
\begin{aligned}
P(2) &= 1/3 \\
P(12) &= P(32) = 1/9 \\
P(112) &= P(132) = P(332) = P(312) = 1/27 \\
P(111) &= P(113) = P(131) = P(311) = 1/27 \\
P(133) &= P(313) = P(331) = P(333) = 1/27 \\
P(\boldsymbol{s}) &= 0, \text{ para as demais } \boldsymbol{s} \in \mathcal{S}.
\end{aligned}
$$

Plano D

$$
\begin{aligned}
P(12) &= 1/10 \\
P(21) &= 1/6 \\
P(13) &= 1/15 \\
P(31) &= 1/12 \\
P(23) &= 1/3 \\
P(32) &= 1/4 \\
P(\boldsymbol{s}) &= 0, \text{ para as demais } \boldsymbol{s} \in \mathcal{S}.
\end{aligned}
$$

Plano E

$$P(12) = P(32) = 1/2$$
$$P(\boldsymbol{s}) = 0, \text{ para as demais } \boldsymbol{s} \in \mathcal{S}.$$

Planejamento Amostral

Do exposto anteriormente, constata-se que é possível criar infinitos planejamentos amostrais.

Planejamento Amostral

Entretanto, descrever probabilidades associadas a cada amostra passa a ser uma tarefa bastante árdua, principalmente para populações grandes.

Planejamento Amostral

Seria muito mais fácil se existissem descrições que permitissem associar, ou calcular, as probabilidades correspondentes a cada amostra de $\mathcal{S}$.

Por exemplo, o Plano C poderia ser descrito mais facilmente da seguinte maneira:

Planejamento Amostral

Sorteie uma unidade após a outra, repondo a unidade sorteada antes de sortear a seguinte, até o surgimento da unidade 2 ($i = 2$) ou até que 3 unidades tenham sido sorteadas.

Podem ser usados vários tipos de descritores para representar as probabilidades associadas a cada amostra.

Planejamento Amostral

Um deles muito utilizado na abordagem clássica da amostragem é a descrição do planejamento através de regras para o sorteio da amostra.

Exemplo 4.3

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** (e será detalhado algumas Aulas adiante).

Planejamento Amostral

Observa-se que para a maioria dos planejamentos, atribui-se probabilidade nula para muitas amostras de $\mathcal{S}$.

Por isso é comum ao apresentar um plano amostral A, restringir $\mathcal{S}$ a alguma subclasse $\mathcal{S}_A$, contendo apenas as amostras $\boldsymbol{s}$, tais que $P(\boldsymbol{s}) > 0$. Isso facilita bastante a apresentação dos resultados.

Planejamento Amostral

É evidente que quanto mais complexas as regras que descrevem os planos amostrais, mais difíceis serão os procedimentos para a determinação das probabilidades associadas ao espaço amostral $\mathcal{S}$.

Planejamento Amostral

Neste curso serão abordados os planos amostrais mais simples e mais usados, e que servem de base para planos amostrais mais complexos.

Outro conjunto de planos muito úteis e simples, são aqueles de tamanho fixo, ou seja, possuem probabilidades diferentes de zero apenas para a subclasse $\mathcal{S}_n$.

Planejamento Amostral

Os tipos de planejamentos amostrais mais utilizados são:

Amostragem Aleatória Simples (AAS)

Seleciona-se 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.

Amostragem Estratificada (AE)

A população é dividida em estratos (por exemplo, pelo sexo, renda, bairro, etc.) e a AAS é utilizada na seleção de uma amostra de cada estrato.

Amostragem por Conglomerados (AC)

A população é dividida em subpopulações (conglomerados) distintas (quarteirões, residências, famílias, bairros, etc). Alguns dos conglomerados são selecionados segundo a AAS e todos os indivíduos nos conglomerados selecionados são observados.

Planejamento Amostral

Em geral a AC é menos eficiente que a AAS ou AE, mas por outro lado, é bem mais econômica. Tal procedimento amostral é adequado quando é possível dividir a população em um grande número de pequenas subpopulações.

Amostragem em Dois Estágios (A2E)

Neste caso, a população é dividida em subpopulações como na AE ou na AC. Num primeiro estágio, algumas subpopulações são selecionadas usando a AAS. Num segundo estágio, uma amostra de unidades é selecionada de cada subpopulação selecionada no primeiro estágio.

A AE e a AC podem ser consideradas, para certas finalidades como casos particulares da A2E.

Planejamento Amostral

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. A seleção do primeiro indivíduo pode ser feita de acordo com a AAS. Os demais indivíduos que farão parte da amostra são então selecionados sistematicamente.

Também serão estudados os estimadores razão e regressão para o total e a média populacionais, que exploram uma possível relação linear entre a variável de interesse y e alguma variável auxiliar x, usualmente conhecida como variável independente na teoria de regressão linear.

Comentários Finais

Comentários Finais

Em resumo, na aula de hoje nós lidamos com uma parte das principais definições, nomenclaturas e terminologias da Teoria de Amostragem. A saber:

- população;
- amostra;
- planejamento amostral.

Comentários Finais

Nas próximas aulas nós vamos focar na outra parte das nomenclaturas, no caso:

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

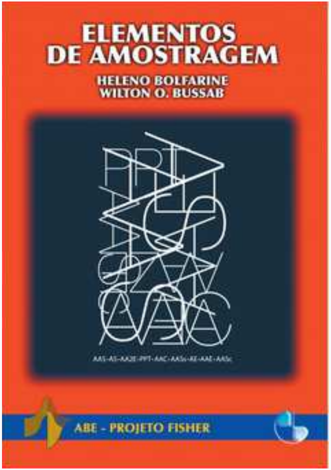
Comentários Finais

ATIVIDADE PARA ENTREGAR (E COMPOR A NOTA N1)

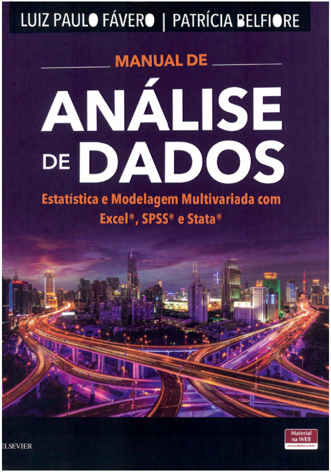
Resolva em grupos de até 4 integrantes os Exercícios 1.1, 1.5, 1.7 e 1.10.

Referências

Referências



Referências



Bons Estudos!