

## Lista nº 4 – Análise Exploratória de Dados

### Instruções:

- a. **Objetivo:** O objetivo desta atividade é explorar, analisar, descrever e interpretar um conjunto de dados. Como objetivo colateral está o uso de pacote computacional em aplicações de estatística descritiva.
- b. **Resposta:** Sua solução deverá conter:
  - i. Identificação dos tipos de variáveis encontradas no Banco de Dados;
  - ii. Construção de tabelas e gráficos das variáveis de interesse;
  - iii. Cálculo, onde aplicável, das medidas descritivas para cada variável (medidas de tendência central, de posição, de dispersão, etc.);
  - iv. Breve comentário sobre os resultados obtidos;
  - v. Análise das relações mais relevantes entre as variáveis apresentadas;
  - vi. Um resumo das principais conclusões a respeito dos dados apresentados, a partir da interpretação dos resultados obtidos;
  - vii. Indicação do tipo de levantamento adicional que poderia ser efetuado no sentido de melhorar as condições de interpretação dos dados;
- c. **Recomendações:** Você é encorajado a olhar sempre os dados e estabelecer conjecturas a serem posteriormente verificadas formalmente como consequência dessa análise exploratória. Solicita-se que o aluno responda também as questões de 1 a 17 descritas abaixo.

Os itens a seguir referem-se aos dados contidos no arquivo de nome *aeusp.txt*, que contém parte dos dados de uma pesquisa realizada pela Associação dos Educadores da USP (AEUSP), sobre aspectos sócio-econômicos e culturais de comunidades de baixa renda da região do Butantã, São Paulo. O questionário foi respondido por um dos moradores da casa sorteada para participar da pesquisa. Sendo um conjunto de dados reais, poderão aparecer incoerências oriundas de equívocos na digitação ou na coleta de dados. Nestes casos, adote uma alternativa que permita contornar a dificuldade encontrada, justificando-a. Os dados estão organizados da seguinte forma:

Num	número do questionário
Comum	comunidade
Sexo	1 = masculino; 2 = feminino.
Idade	idade em faixas: 1 = [14, 25); 2 = [25, 35); 3 = [35, 45); 4 = [45, ∞).
Ecivil	estado civil: 1 = solteiro; 2 = casado; 3 = divorciado; 4 = viúvo; 5 = outro.
Reproce	região de procedência
Temposp	tempo de residência em São Paulo, em anos.
Resid	número de residentes na casa.
Trab	Trabalho: 1 = sim;

	2 = não;
	3 = aposentado.
	tipo de trabalho, só para os que trabalham:
	1 = empregado com carteira;
Ttrab	2 = empregado sem carteira;
	3 = profissional liberal;
	4 = autônomo;
	5 = rural.
Itrab	idade que começou a trabalhar, em anos.
	renda familiar em faixas de reais:
	1 = [0, 150);
Renda	2 = [150, 300);
	3 = [300, 450);
	4 = [450, 900);
	5 = [900, 1500);
	6 = [1500, ∞).
Acompu	acesso a computador:
	1 = sim; 2 = não.
	série em que parou de estudar:
Serief	Branco = não parou de estudar;
	1 a 8 = séries do ensino fundamental;
	9 a 12 = séries do ensino médio.

1. Explore o conjunto de dados e classifique as variáveis. Verifique se existem variáveis com valores incompatíveis ou inválidos e proponha alternativas para a solução do problema. Observe que existem variáveis com respostas em branco e discuta por isso acontece.
2. Estude a variável Renda em função de Comum. Você diria que os moradores da COHAB e do Jardim d'Abril têm a mesma renda? Justifique sua resposta baseando-se em gráficos ou tabelas de frequência.
3. Verifique se o comportamento da variável Temposp é influenciado pelo tipo de trabalho (variável Ttrab).
4. Estude a variável Itrab em função de Reproce através de *box-plot*. Você diria que há diferença de padrão entre as várias comunidades. Repita o procedimento para analisar a variável Itrab em função de Comum.
5. Compare, usando um QQ-plot o comportamento da variável Temposp nos bairros Jardim Raposo e Jardim d'Abril.
6. Obtenha as estatísticas descritivas básicas para as variáveis Itrab e Renda. Repita para cada uma das comunidades. Existem diferenças entre elas?
7. Utilizando os valores da variável Serief, divida os moradores em três categorias: os que não pararam de estudar, aqueles que pararam até a 8ª série e os demais. Para cada uma das categorias, obtenha as medidas de posição e de variância da variável Itrab.
8. Baseado nas variáveis Sexo e Itrab, você diria que os homens começam a trabalhar mais cedo?
9. Construa uma tabela de dupla entrada (tabela de contingência) com as variáveis Comum e Renda. Você diria que existe associação entre elas? Repita para as variáveis Reproce e Trab.

10. O que pode ser dito da associação entre número de residentes (variável  $Resid$ ) e idade que começou a trabalhar (variável  $Itrab$ ).
11. Para cada região de procedência, construa um histograma para a variável  $Temposp$ . Compare os gráficos. Existe diferença entre eles? Algum palpite sobre os modelos teóricos que poderiam ser adequados?
12. Qual a percentagem de:
  - a. Entrevistados com idade inferior a 35 anos;
  - b. Mulheres dentre os entrevistados com idade inferior a 35anos;
  - c. Moradores do Jardim Raposo que tenham acesso a computador;
  - d. Mulheres, vindas do nordeste e que estejam trabalhando. Dentre essas, qual a percentagem das que possuem carteira assinada?
13. Obtenha as estimativas da média e da variância da população das comunidades de baixa renda do Butantã – SP para as seguintes variáveis:  $Idade$ ,  $Temposp$ ,  $Reside$  e  $Renda$ .
14. Que variáveis você introduziria no estudo para compreender melhor a população em questão? Que críticas, se houver, você faz ao delineamento da pesquisa em questão?

Fonte: MAGALHÃES, M. N.; LIMA, A. C. P. *Noções de Probabilidade e Estatística*. 7ª Ed. São Paulo: Edusp, 2013.