Pesquisa Quantitativa (2044007)

Prof. Lupércio F. Bessegato

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. Será importante também a apresentação adequada de relatório contendo a análise e suas conclusões. As normas serão aquelas aplicáveis a relatórios técnicos

b. **Relatório**: A análise deverá ser apresentada na forma de relatório técnico, compreendendo o problema proposto, sua modelagem e resolução, bem como os resultados e sua análise. Isto é, o trabalho deverá ganhar um título. Um pequeno resumo vem em seguida, para que o eventual leitor tenha uma ideia geral do conteúdo do trabalho. O corpo do trabalho é o próximo, dividido em três partes clássicas: introdução, desenvolvimento e conclusão. Por fim, deverão ser apresentadas as referências bibliográficas (livros, revistas, relatórios, etc.) que foram consultadas. Indique também o software utilizado.

   Além disso, o relatório 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;
   viii. Apresentação das observações ou sugestões a respeito do presente trabalho.

c. **Avaliação:** O trabalho será avaliado com base nos seguintes quesitos:

Quesito		*Percentual*
Resolução	Uso de estatística	30%
	Análise dos resultados	30%
Apresentação	Apresentação/Relatório	40%

d. **Recomendações:** O foco deverá sempre ser a análise das características dos dados, assim como o tratamento adequado dos valores relevantes do conjunto de dados selecionado, de maneira a extrair informações e a alicerçar conclusões. Espera-se que você crie os gráficos apropriados aos dados e comente sobre qualquer coisa de interesse que veja nos gráficos, em particular se observar algum comportamento não normal que possa fazê-lo sentir-se pouco à vontade para aplicar testes formais de inferência estatística. você é encorajado a olhar sempre os dados e estabelecer conjecturas a serem posteriormente verificadas formalmente como consequência dessa análise exploratória.

   No decorrer da análise do conjunto de dados solicita-se que o aluno responda também (no corpo do relatório) 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.
`Ttrab`	tipo de trabalho, só para os que trabalham: 1 = empregado com carteira; 2 = empregado sem carteira; 3 = profissional liberal; 4 = autônomo; 5 = rural.
`Itrab`	idade que começou a trabalhar, em anos.
`Renda`	renda familiar em faixas de reais: 1 = [0, 150); 2 = [150, 300); 3 = [300, 450); 4 = [450, 900); 5 = [900, 1500); 6 = [1500, ∞).
`Acompu`	acesso a computador: 1 = sim; 2 = não.
`Serief`	série em que parou de estudar: 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`, `Resid` e `Renda`.
14. Teste se o número médio de residentes em casas da população das comunidades de baixa renda do Butantã – SP é inferior a 4. Indique as suposições adicionais necessárias.
15. Verifique estatisticamente se a proporção de trabalhadores com carteiras assinadas, na população das comunidades de baixa renda do Butantã – SP é inferior a 40%. Use nível de significância de 5%.

16. Teste se a média da variável `Itrab` na população das comunidades de baixa renda do Butantã – SP é a mesma nas subpopulações definidas pelo estado civil dos residentes. Repita o procedimento com as subpopulações definidas pelo local de moradia.
17. Que variáveis você introduziria no estudo para compreender melhor a 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.