UNIOESTE - Mestrado em Ciência da Computação

Metodologia Científica e Técnicas de Experimentação para Ciência da Computação - MCTECC

Análise exploratória de dados - Análise Univariada

1. Em um questionário aplicado aos estudantes que frequentavam a biblioteca do campus, foi perguntado como classificariam o serviço prestado. As respostas foram:

   "ótimo- "bom- "bom- "péssimo- "bom- "bom- "ótimo- "ótimo- "bom- "ótimo- "bom- "ótimo- "bom- "bom- "ótimo- "bom- "péssimo- "bom- "péssimo- "bom" "péssimo- "bom- "bom- "bom- "bom- "ótimo- "bom- "péssimo- "ótimo- "ótimo- "bom- "péssimo"

   (a) Classifique as respostas.

   (b) Construa uma tabela e um gráfico para representar e resumir esta informação. Comente.

2. Considere o conjunto de dados relacionados ao tempo de carga (segundos) de um aplicativo.

4,7	4,9	5,1	5,4	5,7	6	6,3	6,8	7,3	8,4
4,8	4,9	5,1	5,4	5,7	6	6,3	6,8	7,3	8,9
4,8	5	5,2	5,5	5,7	6,2	6,4	6,9	8,2	9,1
4,9	5	5,3	5,6	5,7	6,2	6,5	7	8,2	9,9
4,9	5	5,4	5,6	5,9	6,2	6,7	7,1	8,3	14,1

   (a) Faça uma tabela construindo as classes conforme os critérios apresentados.

   (b) Calcule as frequências: absoluta, relativa, percentual, acumulada, acumulada relativa, acumulada percentual.

   (c) Faça um histograma das classes.

   (d) Faça um diagrama de ramo-e-folhas e interprete os resultados.

   (e) Construa um box-plot e o interprete.

3. Considere as notas de uma turma de alunos.

   6.5, 7.2, 8.0, 5.5, 9.2, 7.8, 6.0, 7.5, 8.5, 6.8, 7.2, 8.9, 9.0, 5.0, 7.0.

   (a) Construa um histograma para visualizar como as notas estão distribuídas.

   (b) Obtenha medidas de tendência central (media, moda, mediana, quartis)

   (c) Obtenha medidas de dispersão (variância, desvio-padrão e coeficiente de variação)

   (d) Calcule o coeficiente de assimetria e curtose.

   (e) Construa um Box-Plot e o interprete.

4. Considere o experimento com dois sistemas operacionais apresentado em aula. Para cada sistema, realize uma análise exploratória dos dados coletados.

   (a) Calcule a média, mediana, quartis, variância, desvio-padrão e coeficiente de variação.

   (b) Medidas de Assimetria e Curtose.

   (c) Faça o gráfico Box-Plot.

Tabela 1: Dados Sistemas Operacionais - tempo (seg)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Micorsofts	14,1	12,8	14,8	13,9	12,1	12,5	13,9	12,2	12	12,4	13	13,1	12,9	14,2
Lunix	12,1	11	13,4	14	12	12,3	13,5	12,9	12,4	11,6	12,8	12,2	11,8	13

5. (Livro Forsyth). Considere o conjunto de dados que mostra o número de barris de petróleo produzidos por ano para os anos de 1880 a 1984, disponibilizados em http://lib.stat.cmu.edu/DASL/Datafiles/Oilproduction.html. A medida de tendência central média é um resumo útil deste conjunto de dados? Por quê?

6. (Livro Forsyth). Considere o conjunto de dados que fornece o conteúdo de sódio e o conteúdo de calorias de três tipos de cachorro-quente disponibilizados em http://lib.stat.cmu.edu/DASL/Datafiles/Hotdogs.html.
   Os tipos são Carne bovina, Aves e Carne (um rótulo bastante vago). Use histogramas condicionais de classe para comparar esses três tipos de cachorro-quente com relação ao conteúdo de sódio e calorias. Faça um box-plot e interprete-o.

7. (Livro Forsyth). Considere o conjunto de dados registrando algumas propriedades de portadores de cartão de crédito taiwaneses disponibilizados em http://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients.
   Este conjunto de dados foi coletado por I-Cheng Yeh e é hospedado pelo UC Irvine Machine Learning Repository. Há uma variável indicando se um portador está inadimplente ou não, e uma variedade de outras variáveis.

   (a) Use histogramas condicionais para investigar se as pessoas que estão inadimplentes têm mais dívidas (use a variável X1 para dívida) do que aquelas que não estão inadimplentes.

   (b) Use box-plot (diagramas de caixa) para investigar se gênero, educação ou estado civil têm algum efeito no valor da dívida (novamente, use X1 para dívida).