

Prova final
Análise Exploratória de Dados - AED
Pontuação total: 10
Prazo: 10/07/2024 - 15/07/2024

Nome: _____ Matrícula: _____

Na resolução da prova use as funções geradoras de dados (`gerar_dados`, `gerar_dados_rl` e `gerar_tdf`), todas disponíveis no arquivo [gerar_dados_v10.R](#) na [página da disciplina](#).

A função `gerar_dados` foi programada para gerar uma amostra aleatória estratificada de pessoas do município X. As variáveis aleatórias de interesse são: Y_1 (medido em *un*), Y_2 (medido em *un*) e **Sexo**. Adicionalmente, assuma que Y_1 e Y_2 não podem assumir valores reais negativos. Os dados são fictícios e tem finalidades exclusivamente didáticas para fins de avaliação prática em análise de dados.

Realizar a análise exploratória dos dados com respostas às seguintes questões:

1 AED: Apresentações tabulares e gráficas (2.0)

1.1 Diagrama de caixa (boxplot) para Y_1 e Y_2 (1.0)

1. (0.5) Antes e após a eliminação de possíveis outliers¹;
2. (0.5) Após a eliminação de possíveis outliers².

1.2 Para Y_1 (1.0)

1. (0.5) Uma apresentação tabular contendo apenas as frequências: absoluta (F_i), relativa (Fr , %) e acumulada (F_{ac} , %), nessa ordem²;
2. (0.5) Histograma e o polígono de frequência acumulada dos dados².

2 AED: Medidas estatísticas básicas (3.0)

2.1 AED: Medidas determinadas a partir dos vetores (1.5)

Para as variáveis Y_1 e Y_2 elaborar apresentações tabulares² contendo as seguintes estimativas:

1. (0.5) Tendência central: média, mediana e moda;
2. (0.5) Posição: quartis e decis;
3. (0.5) Dispersão: amplitude total, variância, desvio padrão e coeficiente de variação.

2.2 AED: Medidas determinadas a partir de apresentações tabulares (1.5)

A função `gerar_tdf` foi programada para gerar uma tabela de distribuição de frequências do tipo comum, dessas que se encontra em publicações. Considere que esta tabela descreve um assunto de seu interesse - publicado - e que é necessário determinar as medidas estatísticas básicas com finalidades de entendimento e comparações.

Elabore uma apresentação tabular contendo:

1. (0.5) Tendência central: média, mediana e moda;
2. (0.5) Posição: quartis e decis;
3. (0.5) Dispersão: amplitude total, variância, desvio padrão e coeficiente de variação.

¹Não distinguindo sexo

²Para cada sexo: M seguido de F

3 AED: Medidas estatísticas de associação e regressão linear (4.0)

Considere os dados gerados pela função `gerar_dados` para a questão subsequente:

3.1 Associação (1.5)

1. (0.5) Estimativas: covariância e correlação linear simples²;
2. (0.5) Diagramas de dispersão dos dados^{2,3};
3. (0.5) Um estudo semelhante foi realizado em um outro município, por outras pessoas. Contudo, as unidades de medida usadas foram: $Y1$ ($100 * un$) e $Y2$ ($100 * un$).

Para comparar associações entre as variáveis de ambos os estudos, qual seria a medida estatística recomendada? Justifique.

3.2 Regressão linear (2.5)

Considere os dados gerados pela função `gerar_dados_r1` como uma amostra de um estudo da influência de uma variável fixa X (medido em un) sobre uma variável aleatória Y (medido em $un.dia^{-1}$).

Os dados são fictícios e tem finalidades exclusivamente didáticas para fins de avaliação prática em análise quantitativa de dados.

1. (1.0) Ajuste aos dados dois modelos de regressão linear: polinômios de grau I e II (ambos não forçado para a origem);
2. (0.5) Apresente um diagrama de dispersão dos dados⁴ com o melhor modelo.
3. (0.5) Qual modelo melhor explica o fenômeno em estudo? Justifique com fundamentação estatística.
4. (0.5) Pelos critérios de ajustamento e escolha de modelos vistos em aula, os coeficientes de determinação (r^2) de modelos lineares ajustados (forçados e não forçados para a origem) são comparáveis? Justifique com fundamentação estatística.

4 Contextualização (1.0)

Localize um artigo científico (periódico **Qualis A** ou **B**) em área de seu interesse no qual a análise exploratória de dados (AED - possivelmente com medidas de associação e uso de regressão linear como modelo explicativo) teve papel preponderante. Discuta o artigo com ênfase nos recursos da AED usados e também na adequação das normas básicas das apresentações gráficas e tabulares adotada pelo periódico.

Observações:

- Para possibilitar a correção, anexe esta prova devidamente preenchida com nomes e matrículas efetivamente usadas na primeira página das respostas.
- As normas para apresentações gráficas e tabulares são obrigatórias, serão observadas e corrigidas.
- Sugere-se (mas não é obrigatório) o uso do ambiente R na resolução das questões propostas.
- A prova poderá ser entregue até as 24:00hs do prazo final.

³Considere $Y2$ no eixo das ordenadas e $Y1$ no eixo das abscissas

⁴Considere Y no eixo das ordenadas e X no eixo das abscissas