Universidade Estadual de Santa Cruz - UESC
Departamento de Ciências Exatas e Tecnológicas - DCET
CET083 - Probabilidade e Estatística
Curso de Ciência da Computação
Prof. José Cláudio Faria

Prova Final
Pontuação total: 10
Prazo: 15/12/2020 - 28/11/2020

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_v9.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: `Y1` (medido em *un*), `Y2` (medido em *un*) e `Sexo`. Adicionalmente, assuma que `Y1` e `Y2` não podem assumir valores reais negativos. Os dados são fictícios e tem finalidades exclusivamente didáticas para fins de avaliação prática em análise de dados.

1 Probabilidade (2.0)

Ao final da votação uma urna contém 4 cédulas com votos para o candidato **A** e 3 para **B**. Suponha que estas cédulas sejam removidas da urna uma a uma.

1. (1.0) Liste todos os resultados possíveis?
2. (1.0) Qual a probabilidade do candidato **A** estar sempre na frente na contagem?

2 Análise exploratória de dados (8.0)

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

3 AED: Apresentações tabulares e gráficas (2.0)

3.1 Diagrama de caixa (boxplot) para Y1 e Y2 (1.0)

1. (0.5) Antes e após a eliminação de possíveis outliers[1];
2. (0.5) Após a eliminação de possíveis outliers[2].

3.2 Para Y1 (1.0)

1. (0.5) Uma apresentação tabular contendo apenas as freqüências: absoluta (Fi), relativa (Fr, %) e acumulada (Fac, %), nessa ordem[2];
2. (0.5) Histograma e o polígono de freqüência acumulada dos dados[2].

4 AED: Medidas estatísticas básicas (2.0)

4.1 AED: Medidas determinadas a partir dos vetores (1.0)

Para as variáveis Y1 e Y2 elaborar apresentações tabulares[2] contendo as seguintes estimativas:

1. (0.33) Tendência central: média, mediana e moda;
2. (0.33) Posição: quartis e decis;
3. (0.33) Dispersão: amplitude total, variância, desvio padrão e coeficiente de variação.

[1] Não distinguindo sexo
[2] Para cada sexo: M seguido de F

4.2 AED: Medidas determinadas a partir de apresentações tabulares (1.0)

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. (0.33) Tendência central: média, mediana e moda;
2. (0.33) Posição: quartis e decis;
3. (0.33) Dispersão: amplitude total, variância, desvio padrão e coeficiente de variação.

5 AED: Medidas estatísticas de associação e regressão linear (3.0)

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

5.1 Associação (1.0)

1. (0.33) Estimativas: covariância e correlação linear simples[2];
2. (0.33) Diagramas de dispersão dos dados[2,3];
3. (0.33) 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.

5.2 Regressão linear (2.0)

Considere os dados gerados pela função `gerar_dados_rl` 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. (0.33) Apresente um diagrama de dispersão dos dados[4] com o melhor modelo.
3. (0.33) Qual modelo melhor explica o fenômeno em estudo? Justifique com fundamentação estatística.
4. (0.33) 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.

6 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` 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.
- Cada hora de atrazo na entrega da avaliação implica na perda de 25%. Portanto, após 4 horas não entregue.

[3] Considere Y2 no eixo das ordenadas e Y1 no eixo das abscissas
[4] Considere Y no eixo das ordenadas e X no eixo das abscissas