o	**Avaliação 02**	
Curso: **Engenharia de Software**		
Disciplina: **Estatística Orientada à Ciência de Dados**	Turma:	Período:
Professor: **Fábio Kravetz**	Data: **27.11.2025**	Nota:
Aluno (a):		

Instruções:

- Favor guardar celulares, computadores e estojos.
- Não se comunique com os demais estudantes nem troque material com eles.
- Use caneta esferográfica azul ou preta, tanto para marcar as questões, as repostas das questões objetivas quanto para escrever as respostas das questões discursivas.
- Responda cada questão discursiva em tópicos ou texto. Qualquer texto ultrapasse o espaço destinado à resposta será desconsiderado.
- Você terá três horas para responder às questões de múltipla escolha e discursivas da prova.
- **As questões que exigem cálculos devem apresentar os mesmos de forma detalhadada e organizada, caso contrário a questão será anulada.**
- **O USO DA CALCULADORA NO CELULAR É EXPRESSAMENTE PROIBIDO.**

Questão 1 - (0,75) Um pesquisador está planejando uma pesquisa de mercado para estimar a proporção de clientes que preferem a nova embalagem de um produto. A população-alvo é muito grande (população considerada infinita) e o estudo deve utilizar amostragem probabilística. Para garantir a confiabilidade dos resultados, ele precisa determinar o tamanho ideal da amostra.

a) Probabilidade Estimada (p):

O pesquisador não possui dados prévios sobre a preferência dos clientes. Qual valor ele deve usar para a Probabilidade Estimada neste cenário, e qual é a justificativa estatística para essa escolha? Justifique sua resposta.

R: O valor que deve ser usado para a Probabilidade Estimada é **0,5 (ou 50%)** caso esta não seja conhecida antes de realizar a amostragem**.**

Esta estimativa é a mais conservadora e é utilizada para maximizar o tamanho da amostra necessário, garantindo assim que a estimativa seja confiável, independentemente da proporção real na população.

b) Valor Crítico (Z) e Nível de Confiança:

Se o pesquisador decidir que a pesquisa deve ter um Nível de Confiança de 95%, determine o Valor Crítico correspondente. Explique o papel do Valor Crítico (Z) nesse cálculo e como ele é obtido através da simetria e das áreas da curva normal padrão. Justifique sua resposta.

R: O Valor Crítico (Z) para um Nível de Confiança de 95% é 1,96.

O Valor Crítico (Z) é fundamental porque ele **define as fronteiras** para decisões estatísticas, estabelecendo os limites do intervalo de confiança. Ele separa a área central da curva (correspondente ao nível de confiança) das áreas extremas (as "caudas").

Obtenção para 95%: Para um Nível de Confiança de 95% (Área Central = 0,95), a área restante (as "caudas") é de 5% (100% - 95%).

Devido à simetria da curva normal, essa área restante de 5% é dividida igualmente entre a cauda da extrema esquerda e a cauda da extrema direita, resultando em 2,5% (ou 0,025) para cada cauda. O valor de 1,96 é encontrado somando a área central (95%) com a área da cauda esquerda (2,5%), resultando na área acumulada à esquerda de de 97,5% (ou 0,9750). Ao utilizar a tabela para procurar o valor 0,9750, obtém-se Z = 1,96.

Questão 2 - (0,75) Um pesquisador está analisando um conjunto de dados sobre o tempo de reação de motoristas após o consumo de cafeína. Ele deseja aplicar um teste estatístico paramétrico para verificar se o tempo médio de reação no grupo com cafeína é menor do que o tempo médio padrão populacional (μ_0).
Analise as afirmações abaixo sobre o processo de teste de hipótese e a Distribuição Normal, marcando (V) para Verdadeiro e (F) para Falso.

a) (**F**) A hipótese que o pesquisador quer provar (tempo médio menor) deve ser representada na Hipótese Nula, pois esta é a afirmação que se tenta refutar;
b) (**F**) Como o pesquisador está buscando evidências de que a média é "menor que" o valor de referência, o teste estatístico adequado para essa formulação de hipóteses será um Teste Unicaudal à Direita;
c) (**F**) Mesmo que os dados do tempo de reação não sigam rigorosamente a distribuição normal, é sempre possível aplicar testes paramétricos, como o t de Student, desde que a amostra seja muito pequena.
d) (**V**) Se o p-valor calculado for 0,001 e o Nível de Significância for 0,05, há evidências estatísticas para rejeitar a Hipótese Nula, pois um p-valor baixo fornece forte evidência contra H_0.
e) (**F**) Se a Estatística de Teste cair na Região de Rejeição, isso significa que o resultado da amostra é "comum" e, portanto, não existem evidências fortes para rejeitar a Hipótese Nula.

a) (F) A hipótese que o pesquisador quer provar (tempo médio menor) deve ser representada na Hipótese Nula, pois esta é a afirmação que se tenta refutar.

   R: A hipótese alternativa representa o que o pesquisador acredita ou quer provar. A hipótese nula é a hipótese do "*status quo*" que o pesquisador tenta refutar, e sempre contém a condição de igualdade.

b) (F) Como o pesquisador está buscando evidências de que a média é "menor que" o valor de referência, o teste estatístico adequado para essa formulação de hipóteses será um Teste Unicaudal à Direita.

   R: Se a hipótese alternativa afirma que a média da população é estritamente menor que o valor de referência, o teste é classificado como Teste Unicaudal à

Esquerda. O Teste Unicaudal à Direita busca evidências de que a média é estritamente maior.

c) (F) Mesmo que os dados do tempo de reação não sigam rigorosamente a distribuição normal, é sempre possível aplicar testes paramétricos, como o t de Student, desde que a amostra seja muito pequena.

R: Testes paramétricos, como o t de Student, assumem que os dados seguem uma distribuição de probabilidade específica, geralmente a distribuição normal. Se a suposição de normalidade é violada, ou se a amostra for muito pequena, são indicados os Testes Não Paramétricos.

d) (V) Se o p-valor calculado for 0,001 e o Nível de Significância for 0,05, há evidências estatísticas para rejeitar a Hipótese Nula, pois um p-valor baixo fornece forte evidência contra H_0.

R: Se o p-valor for menor ou igual ao nível de significância (α), o resultado é estatisticamente significativo e há evidências para rejeitar H_0. Um p-valor baixo (como 0,001) sugere que os dados observados seriam muito improváveis se a H_0 fosse verdadeira, fornecendo evidências fortes para rejeitá-la.

e) (F) Se a Estatística de Teste cair na Região de Rejeição, isso significa que o resultado da amostra é "comum" e, portanto, não existem evidências fortes para rejeitar a Hipótese Nula.

R: Se a Estatística de Teste cair na região de rejeição (ou seja, for mais extrema que o Valor Crítico), rejeita-se H_0. Quanto mais longe de zero for a Estatística de Teste, mais forte é a evidência para rejeitar a Hipótese Nula.

Questão 3 - (0,50). Um instituto de pesquisa precisa determinar o tamanho mínimo da amostra para um estudo em uma universidade com 5.000 alunos. A pesquisa deve ter um nível de confiança de 95% e uma margem de erro de 4%. Utilizando a linguagem R, calcule o tamanho da amostra necessário. Assuma a estimativa mais conservadora para a proporção, já que nenhuma pesquisa prévia foi realizada.

```
N <- 5000  # Tamanho da População
e <- 0.04   # Erro amostral (4%)
p <- 0.5    # Probabilidade (estimativa conservadora)

# Valor Z para 95% de confiança (valor direto)
Z <- 1.96

# 2. Calcular o termo Z^2 * p * (1-p)
# Este termo é a variância máxima multiplicada pelo Z^2
z_p_q <- (Z^2) * p * (1 - p)

# 3. Calcular 'n' usando a fórmula para populações finitas
# A função ceiling() arredonda o resultado para o próximo inteiro
n_final <- ceiling((N * z_p_q) / ((N * e^2) + z_p_q))

# 4. Imprimir o resultado final
print(n_final)
```

Questão 4 - (0,50) Explique de forma detalhada o código, em linguagem R, abaixo. Ressalta-se que é necessário explicar as funções utilizadas e o contexto geral do código. Qual a aplicação deste código?

```
install.packages('plotly')
library(plotly)


set.seed(123)

dados <- rnorm(150, mean = 0, sd = 1)

windows()
hist(dados)

plot_ly(x=dados, type = 'histogram')

windows()
qqnorm(dados)
qqline(dados)

shapiro.test(dados)
```

set.seed(123)

Garante que, toda vez que o código for executado, os números aleatórios gerados na próxima linha serão exatamente os mesmos, tornando o código reprodutível.

dados <- rnorm(150, mean = 0, sd = 1)

Gera 150 números aleatórios que seguem uma distribuição normal.

n = 150: O número de valores a serem gerados.
mean = 0: A média da distribuição.
sd = 1: Desvio padrão da distribuição.

Os 150 valores gerados são armazenados na variável dados.

windows()

Abre uma nova janela gráfica separada.

hist(dados)

Gera um Histograma simples para a variável dados. Um histograma é um gráfico que mostra a frequência (contagem) com que os valores da variável ocorrem em diferentes intervalos.

plot_ly(x=dados, type = 'histogram')

Cria um Histograma Interativo usando o pacote plotly.
x=dados: Define a variável dados como os valores a serem plotados no eixo X.
type = 'histogram': Especifica que o tipo de gráfico é um histograma.

qqnorm(dados)

Cria um Gráfico de Probabilidade Quantil-Quantil (Q-Q Plot). Caso os pontos no gráfico caírem aproximadamente sobre uma linha reta diagonal, isso sugere que os dados são normalmente distribuídos.

qqline(dados)

Adiciona uma linha de referência ao Q-Q Plot criado pela função anterior. Esta linha representa onde os pontos estariam se os dados fossem perfeitamente normais.

shapiro.test(dados)

Este é um teste estatístico formal para determinar se uma amostra de dados veio de uma população normalmente distribuída.

Se p > 0,05, não se rejeita a hipótese nula de que os dados são normalmente distribuídos.

Se p <= 0,05, se rejeita a hipótese nula, concluindo que os dados não são normalmente distribuídos.

Imagine a seguinte situação, se no lugar da linha "dados <- rnorm(150, mean = 0, sd = 1)" tivessemos o seguinte comando "dados <- rgamma(150, shape = 2, scale = 1)", qual seria a diferença?

Neste caso, tem-se a distribuição Gamma e não mais a distribuição normal, o gráfico é assimétrico e a cauda se estende para a direita.

Um valor de shape maior que 1 move o pico da distribuição para longe de zero.

O parâmetro scale determina a dispersão da distribuição. A distribuição não é nem comprimida nem muito esticada horizontalmente.

Questão 5 - (0,50) Um pesquisador está analisando dois métodos de coleta de dados (Método A e Método B) para um experimento. Antes de aplicar um Teste T para comparar as médias, ele precisa verificar se as variâncias dos dois grupos são iguais, que é uma premissa importante para o teste.

```
set.seed(101)

obs_metodo_A <- rnorm(25, mean = 5, sd = 2)
obs_metodo_B <- rnorm(25, mean = 5, sd = 4)


fator_id <- rep(c("A", "B"), each = 25)

dados_analise <- data.frame(Grupo = fator_id,
                            Resultado = c(obs_metodo_A, obs_metodo_B))

windows()
boxplot(obs_metodo_A, obs_metodo_B, names = c("Método A", "Método B"))

bartlett.test(Resultado ~ Grupo, data = dados_analise)
```

Após a execução do script acima, foi exibido no terminal do software RStudio o seguinte resultado:

```
        Bartlett test of homogeneity of variances
data: Resultado by Grupo
Bartlett's K-squared = 8.3528, df = 1, p-value = 0.003851
```

A partir das informações apresentadas responda as questões abaixo:

1. **Qual é a hipótese nula e a hipótese alternativa do bartlett.test realizado?**

   Hipótese Nula: As variâncias são iguais em todos os grupos.
   Hipótese Alternativa: Pelo menos um grupo tem uma variância diferente dos demais.

2. **Com base no p-value apresentado, o que o pesquisador deve concluir sobre as variâncias dos métodos A e B?**

   Como o p-valor é muito menor que 0,05 deve-se rejeitar a H_0. Portanto, como o p-valor (0.003851) é menor que o nível de significância (0.05). Portanto, rejeita-se a hipótese nula de que as variâncias são iguais.

3. **O pesquisador pode prosseguir com um teste T padrão (que assume variâncias iguais)? Por quê?**

   O pesquisador não pode usar o Teste T de Student (o teste padrão que assume variâncias iguais);

Em vez disso, ele deve usar o Teste T de Welch, que é uma adaptação do Teste T desenvolvida especificamente para situações em que as variâncias dos grupos são diferentes.

Questão 6 - (0,50) Uma empresa de agrotecnologia está testando dois novos tipos de fertilizantes (Tipo A e Tipo B) no cultivo de milho. Para avaliar a eficácia, um pesquisador aplicou os fertilizantes em diferentes talhões de teste.
Como a análise completa da fazenda inteira seria inviável, o pesquisador coletou **uma pequena amostra aleatória** da produção de cada tipo de tratamento, medindo a produtividade em sacas por hectare.
Os dados coletados (em sacas/hectare) foram:

- **Fertilizante A:** [20, 22, 19, 21, 23]
- **Fertilizante B:** [23, 25, 22, 24, 25, 23]

Apresente as hipóteses nula e alternativa. Por fim apresente o valor da estatística T.

1. Hipóteses:

   H_0: $\mu_A = \mu_B$ (A produtividade média é a mesma).
   H_1: $\mu_A \neq \mu_B$ (A produtividade média é diferente).

2. Dados:

   Nível de Significância (α): 0,05

Média da amostra do grupo 1 (Fertilizante A):

$$\overline{x_1} = \frac{\sum_{i=1}^{n} x_i}{N}$$

$$\overline{x_1} = \frac{20+22+19+21+23}{5} = \frac{105}{5} = 21,00$$

Média da amostra do grupo 2 (Fertilizante B):

$$\overline{x_2} = \frac{\sum_{i=1}^{n} x_i}{N}$$

$$\overline{x_2} = \frac{23+25+22+24+25+23}{6} = \frac{142}{6} = 23,67$$

Variância amostral do grupo 1:

$$s_1^2 = \frac{\left(\sum_{i=1}^{n} x_i - \mu\right)^2}{n-1}$$

$$(20-21)^2=(-1)^2=1$$
$$(22-21)^2=(1)^2=1$$
$$(19-21)^2=(-2)^2=4$$
$$(21-21)^2=(0)^2=0$$
$$(23-21)^2=(2)^2=4$$

Desta forma

$$s_1^2=\frac{\left(\sum_{i=1}^{n}x_i-\mu\right)^2}{n-1}$$
$$s_1^2=\frac{1+1+4+0+4}{5-1}$$
$$s_1^2=\frac{10}{4}=2,50$$

Variância amostral do grupo 2:

$$s_2^2=\frac{\left(\sum_{i=1}^{n}x_i-\mu\right)^2}{n-1}$$
$$(23-23,67)^2=(-0,67)^2=0,4489$$
$$(25-23,67)^2=(1,33)^2=1,7689$$
$$(22-23,67)^2=(-1,67)^2=2,7889$$
$$(24-23,67)^2=(0,33)^2=0,1089$$
$$(25-23,67)^2=(1,33)^2=1,7689$$
$$(23-23,67)^2=(-0,67)^2=0,4489$$

Desta forma

$$s_2^2=\frac{\left(\sum_{i=1}^{n}x_i-\mu\right)^2}{n-1}$$
$$s_2^2=\frac{0,4489+1,7689+2,7889+0,1089+1,7689+0,4489}{6-1}$$
$$s_2^2=\frac{7,33}{5}=1,47$$

Tamanho da amostra do grupo 1 (n_1): 5
Tamanho da amostra do grupo 2 (n_2): 6

1. Cálculo da Estatística T: **(DUAS AMOSTRAS INDEPENDENTES – VARIÂNCIAS HETEROGÊNEAS)**

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{21,00 - 23,67}{\sqrt{\frac{2,50}{5} + \frac{1,47}{6}}} = \frac{-2,67}{\sqrt{0,50 + 0,245}} = -3,09$$

Questão 7 - (0,50) Analise os gráficos Q-Q Plot (A e B) abaixo e apresente suas **CONCLUSÕES DETALHADAS** sobre a normalidade dos dados que eles representam:

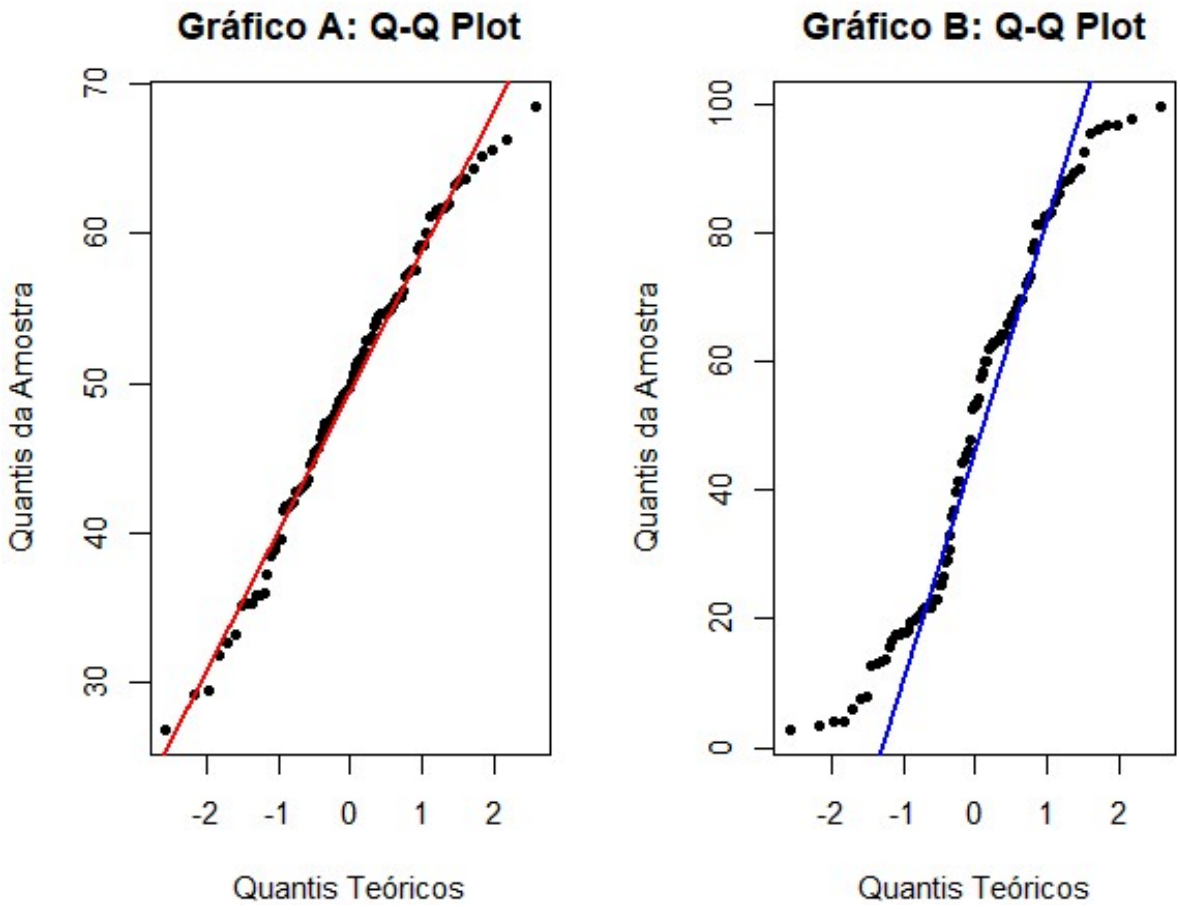

Gráfico A

Os quantis da amostra (pontos pretos) estão alinhados quase de maneira perfeita ao longo da linha de referência vermelha.

Portanto, os dados do **Gráfico A** se ajustam bem a uma distribuição normal. A pequena variação nas pontas é esperada e considerada normal, principalmente com **amostras** de dados.

Gráfico B

Os quantis da amostra (pontos pretos) desviam-se claramente da linha de referência azul, formando um padrão "S" bem definido. Nas extremidades (caudas) os pontos da amostra estão fora da linha. Na cauda esquerda os pontos estão acima da linha. Na cauda direita os pontos estão abaixo da linha.

Portanto, os dados do Gráfico B possui "caudas leves" em comparação com a distribuição normal. Ou seja, os dados da amostra têm menos valores extremos do

que seria esperado para uma distribuiçã normal. Deste modo, o desvio sistemático em forma de “S” é uma evidência clara de que o Gráfico B representa um conjunto de dados que não segue uma distribuição normal.

Formulário:

teste

$$\mu = \frac{\left(\sum_{i=1}^{N} x_i\right)}{N}$$

$$\bar{x} = \frac{\left(\sum_{i=1}^{n} x_i\right)}{n}$$

$$\sigma^2 = \frac{\left(\sum_{i=1}^{N} x_i - \mu\right)^2}{N}$$

$$s^2 = \frac{\left(\sum_{i=1}^{n} x_i - \bar{x}\right)^2}{n-1}$$

$$\sigma = \sqrt{\frac{\left(\sum_{i=1}^{N} x_i - \mu\right)^2}{N}}$$

$$s = \sqrt{\frac{\left(\sum_{i=1}^{n} x_i - \bar{x}\right)^2}{n-1}}$$

$$cv = \left(\frac{DesvioPadrão}{Média}\right)$$

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

$$P(A \cap B) = P(A) \cdot P(B)$$

$$P(E) = \frac{N^{o}\, de\, \mathrm{Re}\, sultados Favoráveis}{N^{o}\, Total De\, \mathrm{Re}\, sultados}$$