Prova de Estatística Inferencial

MBA CDIA

Escolha e resolva 5 das 15 questões abaixo, sendo no máximo 3 questões do mesmo tipo (teóricas ou práticas). Por exemplo: você pode optar por 3 práticas e 2 teóricas ou vice-versa.

Cada questão vale 2 pontos.

As respostas (tanto teóricas quanto práticas) devem ser entregues preferencialmente em arquivo ".ipynb" do Google Colab.

Questões Teóricas:

1. Erros Tipo I e Tipo II
Considere que um órgão público afirma que o tempo médio de espera por um atendimento não ultrapassa 15 minutos. Você, entretanto, decide desafiar este valor, pois trabalha há dois anos no setor e acredita que, devido a modificações recentes e à saída de alguns servidores experientes, o tempo de espera atual é maior. Explique o que seria um erro tipo I e um erro tipo II neste contexto, e dê um exemplo prático de cada situação.

2. Teorema Central do Limite (TCL)
Uma prefeitura deseja estimar a média do tempo de deslocamento diário dos cidadãos para o trabalho. Explique em que condições o Teorema Central do Limite pode ser aplicado para garantir que a média amostral tenha distribuição aproximadamente normal. O que acontece se essas condições não forem satisfeitas?

3. Avaliação de Amostragem
Uma agência reguladora deseja avaliar a satisfação de cidadãos com serviços públicos digitais e decide entrevistar as primeiras 30 pessoas que acessarem o portal online em um determinado dia. Classifique o tipo de amostragem utilizada, explique se é um bom método e sugira uma técnica alternativa, justificando a sua escolha.

4. Análise Crítica do Teste z para Proporções
Um município realizou uma pesquisa sobre aprovação de uma nova política pública com uma amostra de 200 cidadãos e obteve uma proporção de aprovação de 65%. Um teste z para proporções foi realizado, resultando em uma estatística de teste $z = 2,5$ e um p-valor = 0,0124 (nível de significância $\alpha = 0,05$).

Explique brevemente o que significa este p-valor. O que aconteceria com o p-valor e a conclusão do teste se:

- a estatística de teste aumentasse para 3,0?
- o tamanho da amostra aumentasse para 500, mantendo a mesma proporção de aprovação?
- o nível de significância fosse alterado para $\alpha = 0,01$?

5. Distinção entre Teste t Independente e Pareado

Um secretário de educação deseja avaliar o impacto de um programa de capacitação de professores no desempenho dos alunos. Dois pesquisadores propõem desenhos experimentais diferentes:

Pesquisador A: Selecionar aleatoriamente 30 escolas que receberão o programa e 30 escolas controle, comparando as médias de desempenho dos alunos após 6 meses usando teste t para amostras independentes.

Pesquisador B: Selecionar 30 escolas e medir o desempenho dos alunos antes e depois do programa, usando teste t pareado para comparar as médias.

Explique a diferença fundamental entre os dois tipos de teste t em termos de pressupostos e estrutura dos dados.

6. Teste Qui-Quadrado de Independência

Uma pesquisa investigou a relação entre o nível de escolaridade dos cidadãos (Fundamental, Médio, Superior) e sua preferência por canais de atendimento público (Presencial, Digital). Foi obtida a seguinte tabela de contingência:

Escolaridade/Atendimento	Presencial	Digital	**Total**
Fundamental	120	30	150
Médio	80	70	150
Superior	40	110	150
Total	240	210	450

O teste qui-quadrado resultou em $\chi^2 = 89.3$, com $p < 0.001$.

a) Explique o que a hipótese nula e alternativa representam neste contexto de política pública.

b) Interprete o resultado do teste. O que ele nos diz sobre a relação entre escolaridade e preferência de canal?

7. Intervalos de Confiança no Teste de Tukey
Um estudo comparou o tempo de resposta de 3 diferentes canais de atendimento ao cidadão (presencial, telefônico e digital) em uma prefeitura. O teste de Tukey HSD foi aplicado após uma ANOVA significativa, gerando os seguintes intervalos de confiança (95%) para as diferenças entre médias:

- Presencial - Telefônico: [2.3, 5.7] minutos
- Presencial - Digital: [-1.2, 3.1] minutos
- Telefônico - Digital: [-6.8, -2.1] minutos

Identifique quais pares de canais apresentam diferenças estatisticamente significativas e explique o porquê.

8. Seleção de Variáveis em Regressão Linear Múltipla Um modelo de regressão linear múltipla foi desenvolvido para prever o índice de satisfação cidadã (Y) com base em três variáveis preditoras relacionadas a serviços públicos:

- X_1: Tempo de espera (minutos)
- X_2: Número de documentos exigidos
- X_3: Disponibilidade de atendimento digital (0=não, 1=sim)

Os resultados da análise de regressão são apresentados na tabela abaixo:

Variável	Coeficiente (β)	Erro Padrão	Estatística t	p-valor
Intercepto	85.20	5.57	15.30	< 0.001
X_1 (Tempo de espera)	-2.10	0.47	-4.50	< 0.001
X_2 (Documentos exigidos)	-0.80	0.67	-1.20	0.231
X_3 (Atendimento digital)	12.50	3.29	3.80	< 0.001

a) Com base nos testes t individuais ($\alpha = 0,05$), quais variáveis devem ser mantidas no modelo? Justifique.
b) Em que situações não se deve automaticamente remover todas as variáveis não significativas de uma só vez?

Questões Práticas (Python):

9. Determinação do Tamanho Amostral
Um gestor público deseja estimar a proporção de usuários satisfeitos com um novo serviço com margem de erro de no máximo 3% e nível de confiança de 95%. Sabendo que em uma pesquisa

preliminar a satisfação ficou em torno de 70%, determine em Python o tamanho da amostra necessário.

10. Intervalo de Confiança
Uma pesquisa com 50 cidadãos encontrou um tempo médio de espera em uma repartição pública de 20 minutos, com desvio padrão amostral de 5 minutos. Construa em Python um intervalo de confiança de 95% para o tempo médio populacional.

11. Teste t para duas amostras independentes
Em duas regiões administrativas diferentes, foram medidos tempos médios de atendimento ao cidadão:

- Região A: n = 40, média = 12 minutos, desvio padrão = 3 minutos
- Região B: n = 35, média = 14 minutos, desvio padrão = 4 minutos

Utilizando Python, teste se há uma diferença estatisticamente significativa entre os tempos médios ao nível de 5% de significância. Comente brevemente os resultados.

12. Teste t Simples
Um órgão público afirma que o tempo médio para resolver um processo administrativo é de 10 dias. Uma amostra com 25 processos apresentou média de 12 dias e desvio padrão amostral de 3 dias. Utilizando Python, realize um teste t simples para verificar se há evidências suficientes para rejeitar a afirmação do órgão ao nível de significância de 5%. Comente brevemente os resultados.

13. Teste Qui-Quadrado para Avaliação de Política Pública
Uma prefeitura implementou um programa de inclusão digital e deseja verificar se existe associação entre a participação no programa e a empregabilidade dos cidadãos. Os dados coletados de 500 participantes foram:

- **Participantes do programa:** 200 pessoas (120 empregados, 80 desempregados)
- **Não participantes:** 300 pessoas (135 empregados, 165 desempregados)

Utilizando Python:

Realize o teste qui-quadrado de independência ($\alpha = 0,05$) para verificar se há associação entre participação no programa e situação de emprego. Interprete o resultado.

14. ANOVA One-Way e Teste de Tukey Uma secretaria de saúde deseja comparar o tempo médio de espera para consultas em 4 diferentes unidades básicas de saúde (UBS). Os dados coletados (em minutos) foram:

- UBS A: [25, 30, 28, 32, 27, 29, 31, 26, 28, 30]
- UBS B: [35, 38, 40, 37, 39, 36, 41, 38, 37, 39]
- UBS C: [22, 25, 23, 24, 26, 21, 24, 23, 25, 22]
- UBS D: [28, 31, 29, 30, 32, 29, 31, 30, 28, 31]

Utilizando Python:

a) Realize uma ANOVA one-way para testar se existe diferença significativa entre os tempos médios das 4 unidades ($\alpha = 0,05$)

b) Se a ANOVA for significativa, aplique o teste de Tukey HSD para identificar quais pares de UBS diferem significativamente.

15. Regressão Linear Múltipla Um pesquisador coletou dados de 50 municípios para modelar o Índice de Desenvolvimento da Educação Básica (IDEB) com base em três variáveis:

- investimento_per_capita: investimento em educação por aluno (em R$ milhares)
- razao_aluno_professor: número de alunos por professor
- internet_escolas: percentual de escolas com internet banda larga

Código python:

```
# Dados simulados para o exercício
import numpy as np
np.random.seed(42)
n = 50

investimento_per_capita = np.random.uniform(2, 8, n)
razao_aluno_professor = np.random.uniform(15, 30, n)
internet_escolas = np.random.uniform(20, 95, n)

# IDEB simulado com relação linear + ruído
ideb = (3.5 +
    0.4 * investimento_per_capita -
    0.08 * razao_aluno_professor +
```

```
    0.015 * internet_escolas +
    np.random.normal(0, 0.3, n))
```

Utilizando o código Python e os dados simulados fornecidos acima:

a) Ajuste um modelo de regressão linear múltipla utilizando a abordagem de divisão entre base de treino e teste.

b) Apresente e interprete os coeficientes estimados e seus testes de significância

c) Calcule e interprete o R^2.

d) Verifique os pressupostos do modelo através de análise de resíduos

e) Com base nos resultados, que recomendações você faria para melhorar o IDEB nos municípios?
