IEC -Estatística para Ciência de Dados

PUC Minas

Trabalho Final

Inferência Estatística

Valor - 30 pts (6 pts cada questão)

Para cada uma das situações abaixo, utilize o teste de hipótese mais adequado. O objetivo é testar a adequação dos dados aos testes paramétricos e quando as suposições não forem atendidas ou o tipo de dado não for adequado, usar o teste correspondente não paramétrico.

Tente construir as análises e conclusões de forma que você fará no seu trabalho no futuro, utilize gráficos se achar útil e não se esqueça de escrever as hipóteses que estão sendo testadas em cada teste de forma adequada. As bases de dados contém várias outras informações que não serão usadas, mas eu vou mantê-las para que vocês possam explorar e tentar fazer outros testes para colocar no portfólio de vocês.

(A minha dica é: Para o trabalho, foque apenas nas variáveis que vamos utilizar e depois quando tiver mais tempo, aproveite para explorar as outras informações e fazer outros testes).

Para todos os cenários abaixo é necessário escrever as hipóteses testadas, testar as suposições e escrever as conclusões dos resultados obtidos (não se esqueça de fazer o post hoc quando necessário).

1. O TOC, ou Transtorno Obsessivo-Compulsivo, é um distúrbio psiquiátrico caracterizado pela presença de obsessões e compulsões. As obsessões são pensamentos, imagens ou impulsos indesejados e intrusivos que causam ansiedade significativa. Já as compulsões são comportamentos repetitivos que uma pessoa se sente compelida a realizar como uma resposta para aliviar a ansiedade associada às obsessões.

   O Y-BOCS, ou Yale-Brown Obsessive Compulsive Scale (Escala Yale-Brown para Transtorno Obsessivo-Compulsivo), é uma ferramenta de avaliação

desenvolvida para medir a gravidade dos sintomas em pessoas com TOC. Essa escala é dividida em duas partes: uma para avaliar as obsessões e outra para avaliar as compulsões.

A base de dados estruturada "ocd_patient_dataset.csv" é um conjunto de dados de pacientes com TOC que possui dados demográficos e clínicos que contém informações abrangentes sobre 1.500 indivíduos com diagnóstico de transtorno obsessivo-compulsivo (TOC). Observação: o conjunto de dados é inteiramente fictício e não deve ser usado para quaisquer fins clínicos ou de pesquisa reais.
Use um nível de significância de 5% para testar a afirmativa de que existe diferença no Score (Pontuação) Y-BOCS (Obsessões) entre pacientes dos sexo feminino e masculino. Com base neste resultado, podemos concluir que existe diferença?

Informações extras: Pontuação Y-BOCS para Obsessões:

A pontuação da parte de obsessões do Y-BOCS avalia a gravidade das obsessões em termos de tempo gasto, interferência nas atividades diárias, angústia causada pelas obsessões e resistência em controlá-las. A pontuação total para obsessões varia de 0 a 40, sendo uma pontuação mais alta indicativa de sintomas mais graves.

0 a 7: Sintomas leves a moderados.
8 a 15: Sintomas moderados a graves.
16 a 40: Sintomas graves a extremamente graves.

Use as colunas:
(Y-BOCS Score (Obsessions) - Score de obsessão e Gender - sexo).

2. O conjunto de dados "**Sleep_health_and_lifestyle_dataset.csv"** compreende 400 linhas e 13 colunas, cobrindo uma ampla gama de variáveis relacionadas ao sono e hábitos diários. Inclui detalhes como sexo, idade, ocupação, duração do sono, qualidade do sono, nível de atividade física, níveis de estresse, categoria de IMC, pressão arterial, frequência cardíaca, passos diários e presença ou ausência do distúrbio do sono (Insônia).

   O nosso objetivo é avaliar se existe diferença na proporção de pessoas que possuem ou não o distúrbio do sono (Insônia) para as diferenças ocupações na base de dados. Use o nível de significância de 5% e caso encontre diferenças, relate em quais das ocupações vemos essa diferença significativa.

Use as colunas: (Occupation - ocupação e Sleep Disorder - Distúrbio do sono (Insônia ou não).

3. Uma empresa de tecnologia está constantemente analisando o mercado e buscando novas formas de converter os clientes (fechar contratos). Pensando nisso, ela investiu em uma empresa de treinamento de vendedores (VendaPro Academy) que diz que o seu método de vendas **é superior aos métodos convencionais**. Para verificar essa hipótese, a empresa de tecnologia vai avaliar o desempenho dos funcionários.

   Vamos comparar o desempenho de 35 vendedores ao utilizar o método convencional de vendas e o método obtido pelo treinamento com a empresa VendaPro Academy, medindo o **tempo de trabalho em horas necessário para alcançar uma meta de 10 contratos vendidos.** Os dados estão na base de dados "**treinamentoVendedores.csv".** Determine ao nível de 2% de significância, se o método da empresa VendaPro Academy é realmente superior ao método convencional da empresa.

4. Em um mundo onde a música desempenha um papel fundamental nas experiências diárias, a imersão em uma variedade de gêneros musicais tornou-se uma parte intrínseca da vida cotidiana. Com a ascensão de plataformas de streaming e a facilidade de acesso à música, as pessoas desfrutam da liberdade de escolher seus gêneros preferidos e dedicar tempo considerável à apreciação das composições que ressoam com suas emoções.

   Este fenômeno despertou um interesse significativo em compreender o tempo dedicado diariamente às músicas favoritas das pessoas, bem como explorar se existem diferenças notáveis nos hábitos de escuta entre os diversos gêneros musicais.

   No conjunto de dados "Musicas.csv", encontramos informações detalhadas sobre usuários de várias idades, incluindo dados sobre seus hábitos musicais, como o gênero favorito e o tempo diário dedicado à audição musical.

   O objetivo central deste estudo é testar a hipótese de que existe uma diferença significativa no tempo, em horas, que as pessoas dedicam diariamente à audição de música, com base em seus gêneros musicais favoritos.

   Use as colunas: (Fav genre - gênero favorito e Hours per day - horas por dia). Use o nível de significância de 5% para verificar essa hipótese.

Dica: Use o pacote pingouin para facilitar a análise de suposições do teste escolhido.

5. O sucesso de uma equipe esportiva não se limita apenas às habilidades físicas dos atletas, mas também é influenciado por fatores psicológicos, incluindo o estilo de treinamento adotado pelo treinador. **Diferentes treinadores têm abordagens únicas para motivar, orientar e interagir com seus atletas.** Pensando nisso, um estudo foi feito durante um ano para explorar o impacto do comportamento do treinador no desempenho de equipes de atletas, para isso, avaliou-se o desempenho de equipes para diferentes tipos de treinadores que treinaram a equipe durante dois meses cada um.

   A hipótese em teste é que o comportamento do treinador durante o treinamento pode ter um efeito significativo no desempenho das equipes na corrida. Acredita-se que certos tipos de treinadores podem influenciar positivamente os resultados, enquanto outros podem ter diferentes implicações no desempenho dos atletas. **Entenda como impacto positivo, uma diminuição no tempo que as equipes levaram para finalizar a corrida.**

   Use a base de dados “ImpactoTreinador.csv” e conclua se existe essa diferença significativa do comportamento do treinador no tempo de corrida das equipes, use o nível de significância de 5%.

   Com base nos resultados conseguimos identificar uma diferença? Se sim, **entre quais tipos de comportamentos vemos um tempo menor para finalização da corrida?**

Referências:
Exercício 1: Base de dados do kaggle:
https://www.kaggle.com/datasets/ohinhaque/ocd-patient-dataset-demographics-and-clinical-data/

Exercício 2: Base de dados *modificada* do kaggle:
https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset

Exercício 3: Base de dados simulada que eu criei.

Exercício 4: Base de dados *modificada* do kaggle:
https://www.kaggle.com/datasets/catherinerasgaitis/mxmh-survey-results

Exercício 5: Base de dados simulada que eu criei.