

**REPROGRAMA**

**ANÁLISE DE DADOS COM PYTHON**

Turma: On-29-Python

RELATÓRIO DE ANÁLISE DE DADOS

TESTE DE HIPÓTESES

CAMPINA GRANDE

2024

# INTRODUÇÃO

A estatística está presente em várias áreas da nossa vida, desde jogar os dados em um jogo de tabuleiro, até o medicamento que nós utilizamos. Assim como os usos diários, a estatística também auxilia a tomada de decisão em diversas áreas, como a pesquisa científica, marketing, vendas, políticas públicas e na área de tecnologia da informação, principalmente na análise de dados

A área de análise de dados está intimamente ligada à estatística, quer seja ela descritiva ou inferencial. A estatística descritiva, como o próprio nome diz, traz a descrição dos dados e como esses dados se comportam, identificando padrões, tendências e outras características relevantes, porém, não consegue apresentar conclusões mais robustas para além do conjunto de dados examinado. A estatística inferencial utiliza métodos estatísticos para testar hipóteses e estimar parâmetros, e a partir da análise desses dados, por meio da confirmação ou rejeição das hipóteses, o analista possa auxiliar os gestores nas tomadas de decisão.

Assim, com base nos assuntos discutidos em aula, traremos a análise de dois bancos de dados, a fim de aplicar os conceitos e códigos aprendidos durante as semanas 10, 11, 12 e 13 do curso de Análise de Dados com Python.

# METODOLOGIA

Os testes estatísticos aplicados foram o teste t, teste Z, teste qui-quadrado (X2) e análise de variância (ANOVA).

* 1. Conjunto de Dados

Foram utilizados dois bancos de dados de acesso público, por meio da base de dados do site Kaggle:

* Banco de dados 1: Students Performance Dataset[[1]](#footnote-1)

Este conjunto de dados contém informações sobre 2.392 alunos do ensino médio, detalhando seus dados demográficos, hábitos de estudo, envolvimento dos pais, atividades extracurriculares e desempenho acadêmico.

As variáveis utilizadas foram “GradeClass: Classification of students' grades based on GPA”, para a aplicação do teste t e “Sports” e “Gender”, para o teste qui-quadrado.

* Banco de dados 2: Análise de Dados ENEM 2019 - [EBAC][[2]](#footnote-2)

Este conjunto de dados contém os registros reais de todas as inscrições do Enem de 2019, com 5.095.270 linhas (cada linha representa a inscrição de uma pessoa distinta) e 18 variáveis relativas a dados socioeconômicos e ao Enem.

Teste de Hipóteses: Para analisar a relação entre as variáveis categóricas "Gênero" e "Preferência por Ação", utilizaremos o teste Qui-Quadrado de Independência. Este teste compara as frequências observadas de cada combinação de categorias com as frequências esperadas sob a hipótese de independência entre as variáveis.

Análise: Utilizaremos a biblioteca scipy.stats do Python para realizar o teste Qui-Quadrado. Os passos da análise incluem:

Importar as bibliotecas necessárias.

Carregar o conjunto de dados.

Criar a tabela de contingência com as frequências observadas.

Realizar o teste Qui-Quadrado.

Interpretar os resultados do teste, incluindo o valor p e a estatística do teste.

3. Resultados

A tabela de contingência abaixo mostra as frequências observadas de cada combinação de gênero e preferência por filmes de ação:

O teste Qui-Quadrado retornou os seguintes resultados:

Estatística Qui-Quadrado: 4.00

Valor p: 0.0455

Graus de Liberdade: 1

4. Discussão

O valor p obtido (0.0455) é menor que o nível de significância usual de 0.05. Isso indica que há evidências suficientes para rejeitar a hipótese nula de independência entre gênero e preferência por filmes de ação. Portanto, concluímos que existe uma relação significativa entre essas duas variáveis.

É importante observar que o teste Qui-Quadrado não indica a direção da relação. Para entender melhor a natureza dessa relação, precisamos analisar a tabela de contingência. Podemos observar que a proporção de homens que preferem filmes de ação é maior do que a proporção de mulheres que preferem esse gênero.

5. Conclusão

A análise realizada com o teste Qui-Quadrado sugere que existe uma relação significativa entre o gênero e a preferência por filmes de ação. Os resultados indicam que a proporção de homens que preferem filmes de ação é significativamente maior do que a proporção de mulheres. Essa informação pode ser útil para direcionar melhor as campanhas de marketing para esse gênero de filmes.

6. Limitações

É importante reconhecer as limitações da análise. O conjunto de dados utilizado é relativamente pequeno, e os resultados podem não ser generalizáveis para toda a população. Além disso, a análise não considera outros fatores que podem influenciar a preferência por filmes de ação, como idade, localização geográfica e outros interesses.

1. https://www.kaggle.com/datasets/rabieelkharoua/students-performance-dataset [↑](#footnote-ref-1)
2. https://www.kaggle.com/code/lauroliveira/analise-de-dados-enem-2019-ebac/input [↑](#footnote-ref-2)