REPROGRAMA
ANÁLISE DE DADOS COM PYTHON

Turma: On-29-Python

Aluna Fátima Daniela Lucio de Amorim

RELATÓRIO DE ANÁLISE DE DADOS
TESTE DE HIPÓTESES

CAMPINA GRANDE
2024

1 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.

2 METODOLOGIA

Os testes estatísticos aplicados foram o teste t, teste Z, teste qui-quadrado (X^2) e análise de variância (ANOVA).

2.1 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]

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]

[1] https://www.kaggle.com/datasets/rabieelkharoua/students-performance-dataset

[2] https://www.kaggle.com/code/lauroliveira/analise-de-dados-enem-2019-ebac/input

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.

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

3.1 Banco de dados 1: Students Performance Dataset

- ESTATÍSTICA T

Tabela 1. Base de dados resumida apresentando a idade, gênero (1=masculino/2=feminino), educação dos pais e GPA

	Age	Gender	Ethnicity	ParentalEducation	GPA
0	17	1	0	2	2.929196
1	18	0	0	1	3.042915
2	15	0	2	3	0.112602
3	17	1	0	3	2.054218
4	17	1	0	2	1.288061
...	...	...	...	...	...
2387	18	1	0	3	3.455509
2388	17	0	0	1	3.279150
2389	16	1	0	2	1.142333
2390	16	1	1	0	1.803297
2391	16	1	0	2	2.140014

2392 rows × 5 columns

As hipóteses determinadas para a análise fora:

H_0: A média do GPA dos alunos é significativamente diferente da média calculada.

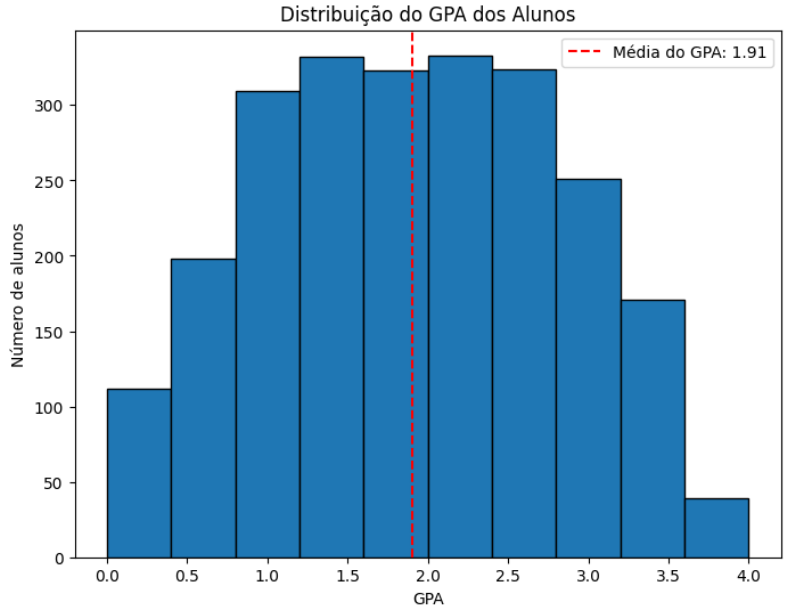
H_1: A média do GPA não é diferente da média calculada.

Significância de 95%

```
# Dados do teste t
t_statistic = 0.0012877131103315757
p_value = 0.9989726612974702
n = 2392  # Tamanho da amostra
```

Para a análise da hipótese nula foi utilizado o teste t. De acordo com o resultado da análise, como o valor de p foi maior que 0,05 como, rejeitamos a hipótese nula, isto é, não há evidências suficientes para concluir que a média do GPA dos alunos é diferente da média calculada.

Gráfico 1. Gráfico da distribuição normal da média calculada do GPA



- ESTATÍTICA X^2

  As hipóteses determinadas para a análise fora:

H0 = não existe associação entre gênero e preferência para esportes.

H1 = existe associação entre gênero e preferência para esportes.

Significância de 95%

Para a análise da hipótese nula foi utilizado o teste Qui-quadrado (X^2). Para o teste X2, primeiro faz uma tabela de contingência (tabela 2x2).

Tabela 2. Tabela de contingência com a relação entre gênero e realização de esportes.

Sports	0	1
Gender		
0	810	360
1	856	366

De acordo com a estatística X^2 mostrada abaixo, pode-se concluir que, com valor de p maior que 0,05, não rejeitamos a hipótese nula, mostrando que não há associação entre gênero e preferência para esportes.

```
    print("Estatística Qui-Quadrado:", chi2_stat)
    print("Valor p:", p_value)
```

```
Estatística Qui-Quadrado: 0.15261664988758886
Valor p: 0.6960472310668808
```

3.2 Banco de dados 2: Análise de Dados ENEM 2019 - [EBAC]

- ESTATÍSTICA Z

Teste Z

H0: a média da idade das mulheres que fizeram o Enem em 2019 é de 20 anos.

H1: a média de idade das mulheres é maior que 20 anos.

Significância de 95%

Tabela 3. Base de dados resumida apresentando a distribuição das idades das mulheres que realizaram o Enem, no ano de 2019.

	idade	sexo
1	16	F
2	18	F
6	30	F
7	26	F
9	17	F
...	...	...
5105419	22	F
5105421	30	F
5105422	28	F
5105424	18	F
5105426	25	F

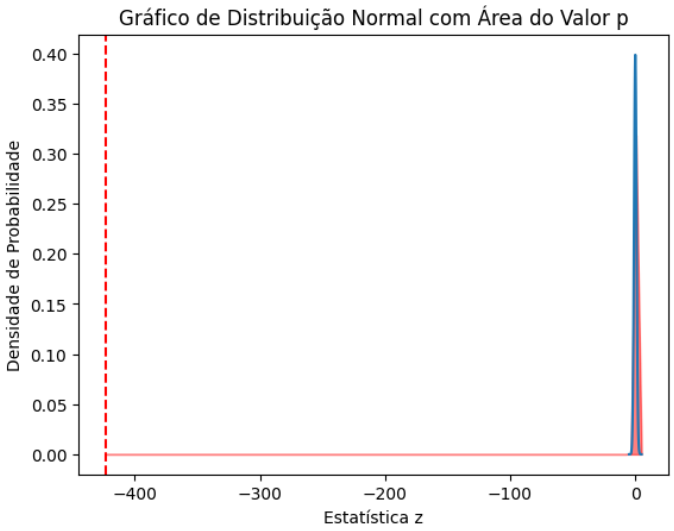
3037657 rows × 2 columns

Para a análise da hipótese nula foi utilizado o teste Z, mostrando que, com valor de p maior que 0,05, não rejeitamos a hipótese nula. Não há evidências suficientes para concluir que a proporção de mulheres com idade maior que 20 anos é maior que 50%.

```
# Executar o teste z
stat, p_value = proportions_ztest(count=x, nobs=n, value=p0, alternative='larger')
# Nível de significância
nivel_significancia = 0.05
print(f"Estatística z: {stat}")
print(f"Valor p: {p_value}")
# Decisão
if p_value < nivel_significancia:
    print("Rejeitamos a hipótese nula. A proporção de mulheres com idade maior que 20 anos é maior que 50%.")
else:
    print("Não rejeitamos a hipótese nula. Não há evidências suficientes para concluir que a proporção de mulheres com idade maior que 20 anos é maior que 50%.
```

[3] Python

```
Estatística z: -423.5003356840059
Valor p: 1.0
Não rejeitamos a hipótese nula. Não há evidências suficientes para concluir que a proporção de mulheres com idade maior que 20 anos é maior que 50%.
```



4 CONCLUSÃO

Podemos concluir que os testes estatísticos foram necessários para apresentar, estatisticamente, a rejeição ou comprovação das principais hipóteses levantadas. Eles são uma metodologia estatística que nos auxilia a tomar decisões sobre uma ou mais populações baseadas nas informações obtidas da amostra.

Ao analisarmos qualquer fato da vida real, sempre realizamos suposições, de forma a tentar estabelecer o que pode ser verdadeiro ou falso. De fato, através de interpretações e da lógica, podemos supor e inferir sensos comuns. Os testes de hipótese, quando corretamente

aplicados, auxiliam o analista na apresentação de possíveis situações e vão basear, de forma segura, o processo de tomada de decisão.

5 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 os resultados.