SME 0810 - Métodos Não Paramétricos

Trabalho Prático 1/1

Testes para uma Amostra

Aluno(a) 1: Ada Maris Pereira Mário_____ Nº USP: 12725432
Aluno(a) 2: Emanuel Victor da Silva Favorato_____ Nº USP: 12558151

RELATÓRIO

Passos:
- Descrição da(s) variável(is) para o estudo;
- Definir os objetivos e as hipóteses formuladas.
- Resultados da(s) análise(s) dos dados no *Software* R ou Python;
- Comentários e conclusões.

1 Introdução e Objetivos

Neste projeto, serão utilizadas duas bases de dados relativas às performances de estudantes em provas. O primeiro [Kaggle, 2021] diz respeito a uma pesquisa entre alunos de um curso de matemática no ensino secundário, constituído por 33 colunas e 395 entradas. As variáveis a serem utilizadas neste estudo são:

- `age`: Idade do estudante. Valores discretos variando no intervalo de 15 a 22;
- `G1`: Nota adquirida no exame 1. Valores discretos variando de 0 a 20;
- `G2`: Nota adquirida no exame 2. Valores discretos variando de 0 a 20;
- `G3`: Nota adquirida no exame 3. Valores discretos variando de 0 a 20.

É informado, também, que uma soma de notas ≥ 35 indicaria aprovação no curso inteiro.

Já o segundo conjunto de dados [Kaggle, 2022] apresenta uma temática similar, tratando-se de uma visão geral dos vários fatores que afetam o desempenho dos alunos nos exames. Contém 20 colunas e 6607 linhas. As variáveis a serem analisadas são:

- `Physical_Activity`: Número médio de horas de atividade física por semana;
- `Sleep_Hours`: Número médio de horas de sono por noite.

2 Metodologia

Para a realização de testes não-paramétricos, será utilizado $\alpha = 5\%$. Para a formulação das hipóteses, tomaram-se como base a construção de histogramas de cada variável aqui analisada, de modo a supor previamente as distribuições que estarão descritas nas hipóteses a serem testadas.

Para os testes binomial e do quantil, serão utilizados os dados da base [Kaggle, 2021]. Para os demais, será utilizada a base [Kaggle, 2022]

- **Teste Binomial:** Ao menos metade dos alunos que realizaram os três testes de matemática foram aprovados, ou seja:

$$H_0 : p \geq 0,5 \quad \text{x} \quad H_1 : p < 0,5$$

- **Teste do Quantil:** pelo menos 75% (q_3) dos aluno que realizaram os testes possuem até 18 anos, ou seja:
$$H_0 : P(X \leq 18) \geq 0,75 \quad \text{x} \quad H_1 : P(X \leq 18) < 0,75$$

- **Teste Qui-Quadrado:** a variável `Physical_Activity` é proveniente de uma binomial simétrica ($p = 0.5$) e com $n = 6$, isto é:
$$H_0 : X \sim Bin(6; 0,5) \quad \text{x} \quad X \not\sim Bin(6; 0,5)$$
E, a variável `Physical_Activity` é proveniente de uma distribuição normal com média $\mu = 3$ e variância $\sigma^2 = 1,06$, isto é:
$$H_0 : X \sim N(3; 1,06) \quad \text{x} \quad X \not\sim N(3; 1,06)$$

- **Teste de Kolmogorov-Smirnov:** a variável `Physical_Activity` é proveniente de uma binomial simétrica ($p = 0.5$) e com $n = 6$, isto é:
$$H_0 : X \sim Bin(6; 0,5) \quad \text{x} \quad X \not\sim Bin(6; 0,5)$$
E, a variável `Physical_Activity` é proveniente de uma distribuição normal com média $\mu = 3$ e variância $\sigma^2 = 1,06$, isto é:
$$H_0 : X \sim N(3, 1,06) \quad \text{x} \quad X \not\sim N(3, 1,06)$$

- **Teste Lilliefors:** `Sleep_Hours` segue uma distribuição normal, com média e variância desconhecidas, isto é:
$$H_0 : X \sim N(\mu, \sigma^2) \quad \text{x} \quad X \not\sim N(\mu, \sigma^2)$$

- **Teste Shapiro-Wilk:** `Sleep_Hours` segue uma distribuição normal, com média e variância desconhecidas, isto é:
$$H_0 : X \sim N(\mu, \sigma^2) \quad \text{x} \quad X \not\sim N(\mu, \sigma^2)$$

Para o tratamento, exploração e visualização dos dados, utilizaram-se métodos das bibliotecas `Pandas` e `Matplotlib` da linguagem de programação `Python`. Para os testes, utilizaram-se métodos da biblioteca `statsmodels`, da mesma linguagem. A linguagem de programação `R` com o método `prop.test` foi utilizada para melhor precisão e cálculo da correção nos testes binomial e do quantil, uma vez que $n > 20$ e utilizou-se aproximação para grandes amostras.

3 Resultados

Inicialmente, os dados foram pré-processados verificando a existência de valores nulos, inválidos ou duplicados; os quais, considerando as variáveis aqui utilizadas, não estavam presentes. Para os testes binomial e do quantil as variáveis `G1`, `G2`, `G3` foram utilizadas de modo a criar a variável `pass`, calculada a partir da verificação se soma das três notas dos testes cumpre ou não a nota mínima para passar no curso, logo, a variável criada assume valores binários.

Em seguida, realizaram-se algumas visualizações dos dados observados das variáveis a serem testadas, como forma complementar de obter um diagnóstico prévio. Nas Figuras 1 a 4 estão os histogramas das variáveis testadas.

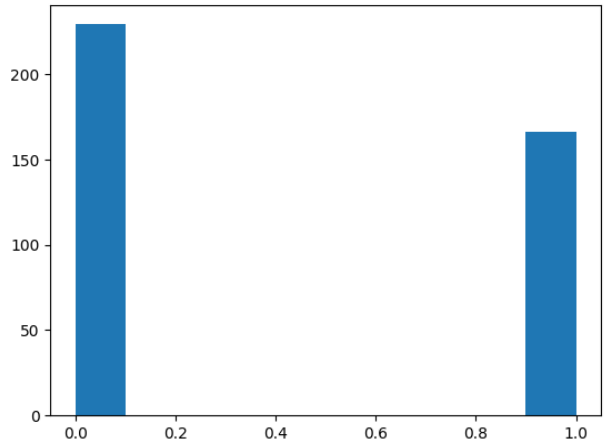

Figura 1: Histograma da variável `pass`

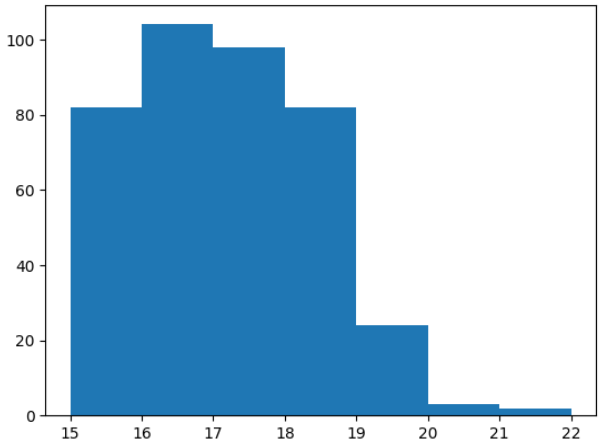

Figura 2: Histograma da variável `age`

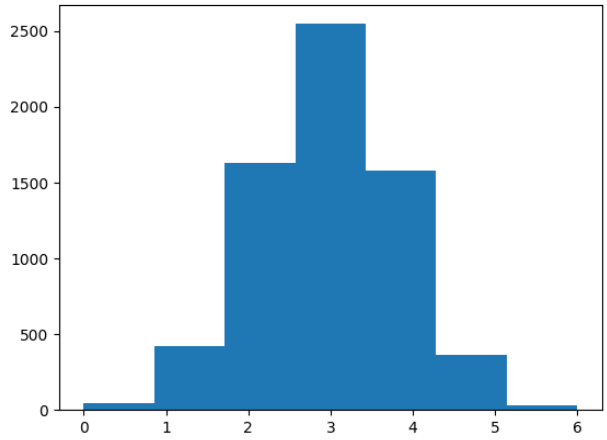

Figura 3:
Histograma da variável `Physical_Activity`

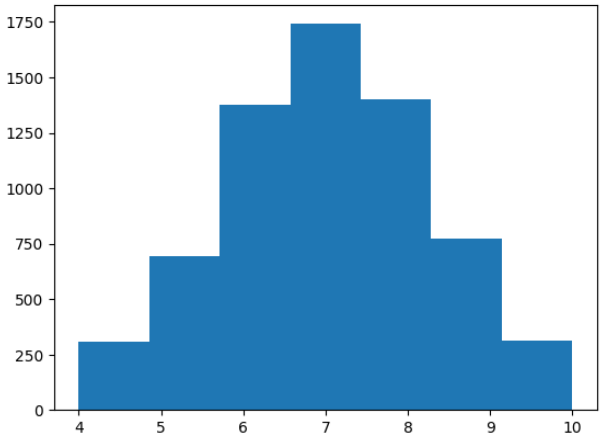

Figura 4:
Histograma da variável `Sleep_Hours`

Observa-se que as hipóteses formuladas são factíveis frente às distribuições de frequências das variáveis estudadas.

Realizados os testes, encontraram-se os seguintes resultados:

- **Teste Binomial:** A estatística de teste calculada foi de $t = 166$. Observou-se $p - valor = 0,001$, rejeitando-se H_0.

  Conclusão: Ao nível de 5% de significância, a análise dos dados coletados fornece evidências de que a proporção de alunos aprovados nos exames de matemática é inferior a 0,5; ou seja, que menos da metade dos estudantes obtiveram êxito nos testes.

- **Teste do Quantil:** A estatística de teste calculada foi de $t = 366$. Observou-se $p - valor = 1$, não se rejeitando H_0.

  Conclusão: Ao nível de 5% de significância, a análise dos dados coletados fornece evidências de que ao menos 75% dos estudantes que realizaram os exames de matemática possuem no máximo 18 anos.

- **Teste Qui-Quadrado:** Para o teste considerando a distribuição binomial, observou-se $t = 368,408$, encontrando-se $p - valor \approx 0$. Logo, rejeita-se H_0.

  Conclusão: Ao nível de significância de 5%, a análise dos dados coletados fornece evidências de que o tempo dedicado a atividades físicas não pode ser explicado por uma distribuição binomial com $n = 6$ e $p = 0,5$.

Já para o teste considerando a distribuição normal, observou-se $t = 8,594$, encontrando-se $p - valor = 0,198$. Logo, não se rejeita H_0.

Conclusão: Ao nível de significância de 5%, a análise dos dados coletados fornece evidências de que o tempo dedicado a atividades físicas provém de uma população com distribuição normal de média $\mu = 3$ e variância $\sigma^2 = 1,06$.

- **Teste de Kolmogorov-Smirnov:** Para os dados, encontrou-se $t = 0,339$. O nível descritivo calculado foi de $p - valor \approx 0$, rejeitando H_0.

  Conclusão: Ao nível de significância de 5%, a análise dos dados coletados fornece evidências de que o tempo dedicado a atividades físicas não pode ser explicado por uma distribuição binomial com $n = 6$ e $p = 0,5$.

  Já para o teste considerando a distribuição normal, observou-se $t = 0,202$, encontrando-se $p - valor \approx 0$. Logo, rejeita-se H_0.

  Conclusão: Ao nível de significância de 5%, a análise dos dados coletados fornece evidências de que o tempo dedicado a atividades físicas não provém de uma população com distribuição normal de média $\mu = 3$ e variância $\sigma^2 = 1,06$.

  É importante notar, neste caso, a discordância entre o teste qui-quadrado e o Kolmogorov-Smirnov, considerando que a variável apresenta muitos empates (como se observa na Figura 3 o histograma com "caudas" leves).

- **Teste Lilliefors:** Para os dados, encontrou-se uma estatística de teste $t = 0,132$. O nível descritivo calculado foi de $p - valor = 0,001$, rejeitando H_0.

  Conclusão: Ao nível de significância de 5%, os dados coletados fornecem evidências de que a quantidade de horas de sono não é proveniente de uma população com distribuição normal.

- **Teste Shapiro-Wilk:** Com os dados, encontrou-se uma estatística de teste $t = 0,953$. O nível descritivo calculado foi de $p - valor \approx 0$, rejeitando H_0.

  Conclusão: Ao nível de significância de 5%, a análise dos dados coletados nos fornece evidências de que a quantidade de horas de sono não é advinda de uma população com distribuição normal.

As informações estão sumarizadas na Tabela 1:

Teste	*Estatística de Teste*	*p-valor*	*Decisão*
Teste Binomial	$t = 166$	0,001	Rejeitar H_0
Teste do Quantil	$t = 366$	1	Não Rejeitar H_0
Teste Qui-Quadrado	$t = 368,408$ (binomial)	≈ 0	Rejeitar H_0
	$t = 8,594$ (normal)	0,198	Não Rejeitar H_0
Teste de Kolmogorov-Smirnov	$t = 0,339$ (binomial)	≈ 0	Rejeitar H_0
	$t = 0,202$ (normal)	≈ 0	Rejeitar H_0
Teste Lilliefors	$t = 0,132$	0,001	Rejeitar H_0
Teste Shapiro-Wilk	$t = 0,953$	≈ 0	Rejeitar H_0

Tabela 1: Sumário dos testes não-paramétricos realizados

4 Considerações Finais

Neste trabalho, aplicamos uma série de testes não-paramétricos para avaliar hipóteses relacionadas ao desempenho acadêmico dos alunos e seus hábitos de saúde, como atividade física e horas de sono.

Os resultados mostraram que, ao nível de significância de 5%, a proporção de alunos aprovados nos testes de matemática é inferior a 50%, o que levanta preocupações sobre o desempenho geral dos estudantes. Além disso, identificamos que pelo menos 75% dos alunos que realizaram os exames possuem até 18 anos, o que está de acordo com a hipótese levantada inicialmente.

No que se refere aos hábitos de atividade física e sono, os resultados indicaram que os testes qui-quadrado e Kolmogorov-Smirnov concordam que o tempo dedicado às atividades físicas não segue uma distribuição binomial; enquanto que, supondo uma distribuição normal, há discordâncias, dado que o qui-quadrado aceita a hipótese e o Kolmogorov-Smirnov rejeita, sugerindo que no caso de variáveis discretas em dados com muitos "empates", como na base utilizada, o teste de Kolmogorov-Smirnov se mostra menos poderoso e robusto em comparação com o teste qui-quadrado.

Da mesma forma, verificou-se que a quantidade de horas de sono também não provém de uma população com distribuição normal, o que pode refletir variações nos padrões de sono entre os estudantes.

Referências

Kaggle. Student performance data, 2021. Disponível em: `https://www.kaggle.com/datasets/devansodariya/student-performance-data/data`. Acesso em: 21 de setembro de 2024.

Kaggle. Student performance factors dataset, 2022. Disponível em: `https://www.kaggle.com/datasets/lainguyn123/student-performance-factors`. Acesso em: 21 de setembro de 2024.