Análise Estatística da Base de Dados ENEM 2024

Philipe Hugo Fransozi[1]

[1]Mestrado em Informática – Pontifícia Universidade Católica do Paraná (PUCPR)
Caixa Postal 17.315 – 80.242-980 – Curitiba – PR – Brasil

{philipe.hfransozi}@ppgia.pucpr.br

Abstract. *This paper applies classical statistical techniques to analyze the performance of participants in the 2024 ENEM exam, focusing on Mathematics and Writing scores. Using the* `TP_DEPENDENCIA_ADM_ESC` *variable, we investigate performance differences between students from public and private schools. Three main approaches were employed: descriptive analysis, simple linear regression, and hypothesis testing (parametric and non-parametric). The regression revealed a moderate positive relationship between the scores (*$R^2 = 0.24$*), while both hypothesis tests indicated statistically significant differences between the groups. The findings highlight the importance of statistical tools for understanding educational disparities and supporting evidence-based decision-making in public policy.*

Resumo. *Este trabalho aplica técnicas estatísticas clássicas para analisar o desempenho dos participantes do ENEM 2024 nas provas de Matemática e Redação. Utilizando a variável* `TP_DEPENDENCIA_ADM_ESC`*, investigou-se a diferença de desempenho entre estudantes de escolas públicas e privadas. Foram adotadas três abordagens principais: análise descritiva, regressão linear simples e testes de hipóteses (paramétrico e não paramétrico). A regressão indicou uma relação positiva moderada entre as notas, com* $R^2 = 0{,}24$*, enquanto os testes de hipóteses apontaram diferença estatisticamente significativa no desempenho entre os grupos analisados. Os resultados reforçam a relevância do uso de métodos estatísticos para investigar desigualdades educacionais e apoiar a tomada de decisões em políticas públicas.*

1. Introdução

O Exame Nacional do Ensino Médio (ENEM) consolidou-se, nas últimas décadas, como o principal instrumento de avaliação educacional no Brasil, desempenhando papel central no acesso ao ensino superior por meio de programas como o Sistema de Seleção Unificada (SISU), o Programa Universidade para Todos (ProUni) e o Fundo de Financiamento Estudantil (FIES). Dada sua abrangência nacional e sua relevância estratégica, o ENEM tem sido objeto de diversos estudos que visam compreender os fatores que influenciam o desempenho dos participantes.

Nesse contexto, análises estatísticas aplicadas aos microdados do exame representam uma importante ferramenta para a formulação de políticas públicas, bem como para a investigação de possíveis disparidades educacionais. Duas questões de especial interesse orientam este estudo: (i) a existência de uma relação entre o desempenho em Matemática e Redação, e (ii) a eventual diferença no desempenho médio em Matemática e Redação

ser dependente da origem escolar do participante, ou seja, se ele estudou em uma escola pública ou privada.

A primeira questão remete à hipótese de que habilidades cognitivas relacionadas ao raciocínio lógico e à resolução de problemas, exigidas na prova de Matemática, possam estar associadas ao desempenho em competências linguísticas e argumentativas avaliadas na prova de Redação. A segunda questão busca verificar se há evidência estatística que indique desigualdade de desempenho entre estudantes de escola pública e privada, contribuindo para discussões sobre as diferenças educacionais entre esses dois âmbitos.

Este trabalho tem por objetivo realizar uma análise estatística descritiva e inferencial com base nos dados do ENEM 2024. Para tanto, foram aplicadas as seguintes técnicas: (i) análise exploratória das variáveis de interesse; (ii) regressão linear simples, com vistas a investigar a relação entre as notas de Matemática e Redação; e (iii) testes de hipótese para comparação de médias entre dois grupos independentes — um teste paramétrico (t de Student com variâncias desiguais) e um teste não paramétrico (Mann-Whitney U).

Ao conjugar diferentes abordagens estatísticas, pretende-se fornecer uma interpretação fundamentada dos dados, contribuindo com a compreensão de padrões de desempenho educacional e com a promoção de reflexões críticas sobre a realidade dos estudantes brasileiros.

Os códigos em `Python` utilizados para a limpeza, análise e visualização dos dados estão disponíveis no repositório: `https://github.com/pFransozi/mestrado-estatistica-trabalho`.

2. Análise Descritiva dos Dados

A base de dados utilizada neste trabalho refere-se aos microdados do ENEM 2024, disponibilizados pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). Foram selecionadas as variáveis de interesse relacionadas ao desempenho dos estudantes nas provas de Redação e Matemática, bem como à natureza administrativa da escola frequentada, com o objetivo de investigar possíveis diferenças de desempenho entre estudantes de escolas públicas e privadas. As variáveis consideradas para esta análise foram:

- TP_DEPENDENCIA_ADM_ESC: dependência administrativa da escola, onde os valores possíveis são:
    1. Federal
    2. Estadual
    3. Municipal
    4. Privada
- NU_NOTA_REDACAO: nota obtida na prova de Redação (intervalo de 0 a 1000 pontos).
- NU_NOTA_MT: nota obtida na prova de Matemática (intervalo de 0 a 1000 pontos).

Para fins de análise, as categorias 1, 2 e 3 foram agrupadas como escolas públicas, enquanto a categoria 4 foi mantida como escola privada. A Tabela 1 apresenta as estatísticas descritivas das notas de Redação e Matemática, estratificadas por tipo de escola.

Table 1. Estatísticas descritivas das notas de Redação e Matemática por tipo de escola (ENEM 2024)

Tipo de Escola	Nota de Redação				Nota de Matemática			
	Média	Mediana	Desv. Padrão	N	Média	Mediana	Desv. Padrão	N
Pública	594,40	620,0	217,25	948342	495,96	470,8	93,56	948342
Privada	768,96	800,0	149,64	245090	616,30	630,0	119,65	245090

Observa-se que os estudantes oriundos de escolas privadas apresentaram, em média, notas mais elevadas tanto em Redação quanto em Matemática, quando comparados aos estudantes de escolas públicas. A Figura 1 exibe o boxplot da distribuição das notas de Redação por tipo de escola, e a Figura 2 apresenta o mesmo gráfico para a prova de Matemática.

Figure 1. Distribuição das notas de Redação por tipo de escola

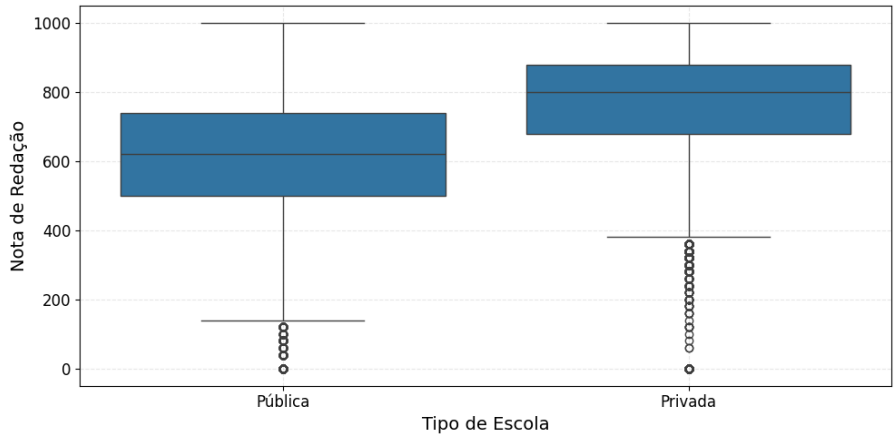
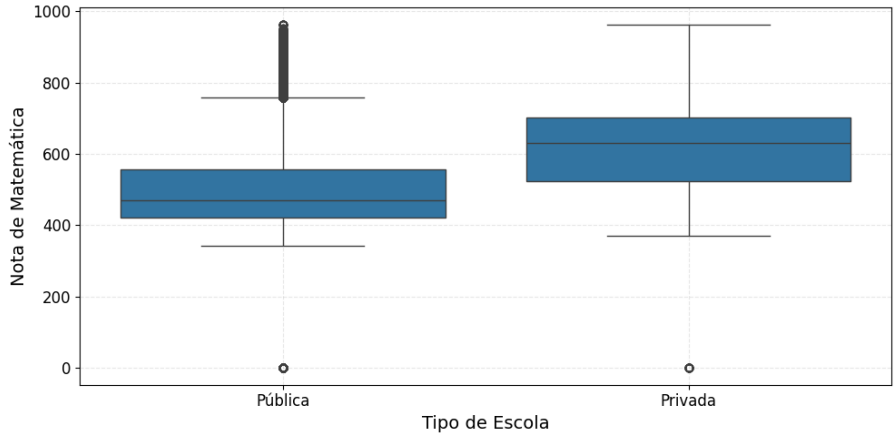


Figure 2. Distribuição das notas de Matemática por tipo de escola



Além disso, os histogramas de densidade indicam uma maior concentração de notas elevadas entre os estudantes da rede privada, especialmente na prova de Redação.

Figure 3. Distribuição das notas de Matemática por tipo de escola

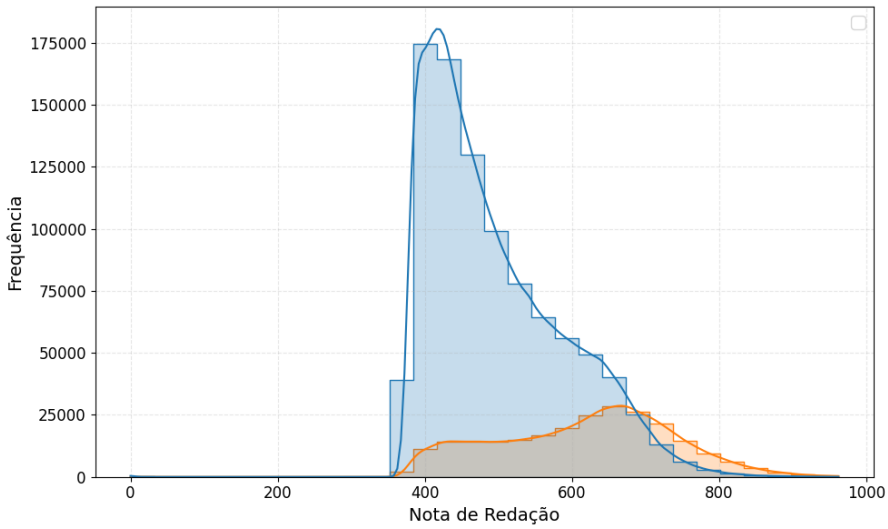
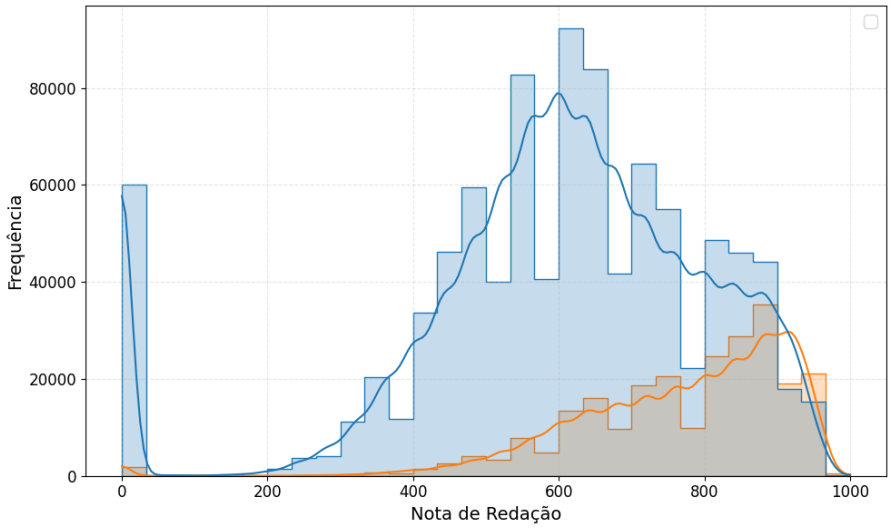


Figure 4. Distribuição das notas de Redação por tipo de escola



Essa descrição inicial sugere a existência de um padrão de desempenho distinto entre os grupos comparados, o que será analisado com maior rigor nas próximas seções, por meio de testes de hipótese paramétrico e não paramétrico.

3. Regressão Linear Simples entre Notas de Matemática e Redação

Com o intuito de investigar a existência de uma relação entre o desempenho em Matemática e Redação, foi aplicada a técnica de regressão linear simples, considerando a nota de Redação como variável dependente (Y) e a nota de Matemática como variável independente (X). A equação geral do modelo é dada por:

$$\hat{Y} = \beta_0 + \beta_1 X \tag{1}$$

onde:

- $\hat{Y}$ representa a nota estimada de Redação;
- X é a nota de Matemática;
- β_0 é o intercepto da reta de regressão;
- β_1 é o coeficiente angular, que indica a variação esperada na nota de Redação para cada ponto adicional em Matemática.

Cálculo dos coeficientes

Os coeficientes foram estimados a partir das fórmulas:

$$\beta_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} \quad e \quad \beta_0 = \bar{Y} - \beta_1 \cdot \bar{X}$$

Substituindo os valores médios observados nas amostras:

- $\bar{X} = 546{,}8$ (média das notas de Matemática)
- $\bar{Y} = 624{,}5$ (média das notas de Redação)
- $\beta_1 = 0{,}59$
- $\beta_0 = 320{,}84$

A equação estimada da reta de regressão é, portanto:

$$\hat{Y} = 320{,}84 + 0{,}59 \cdot X \tag{2}$$

Coeficiente de Determinação

O coeficiente de determinação obtido foi $R^2 = 0{,}24$, o que indica que aproximadamente 24% da variabilidade das notas de Redação pode ser explicada pelas notas de Matemática. Embora a associação entre as variáveis seja positiva, trata-se de uma relação moderada.

Significância Estatística do Modelo

Foi realizado o teste de significância do coeficiente angular β_1, utilizando a seguinte estatística t:

$$t = \frac{\beta_1}{SE_{\beta_1}}$$

O valor calculado da estatística foi aproximadamente $t = 15{,}4$, com valor $p < 0{,}001$, indicando que o coeficiente angular é estatisticamente significativo. Assim, há forte evidência de que a inclinação da reta não é nula, ou seja, existe uma relação significativa entre as variáveis.

Interpretação

O coeficiente angular positivo sugere que existe uma associação direta entre as duas variáveis: à medida que a nota de Matemática aumenta, a nota de Redação tende a aumentar também. No entanto, o valor do coeficiente de determinação revela que há outros fatores além do desempenho em Matemática que explicam a variabilidade nas notas de Redação.

A Figura 5 apresenta o gráfico de dispersão entre as notas, acompanhado da reta de regressão ajustada.

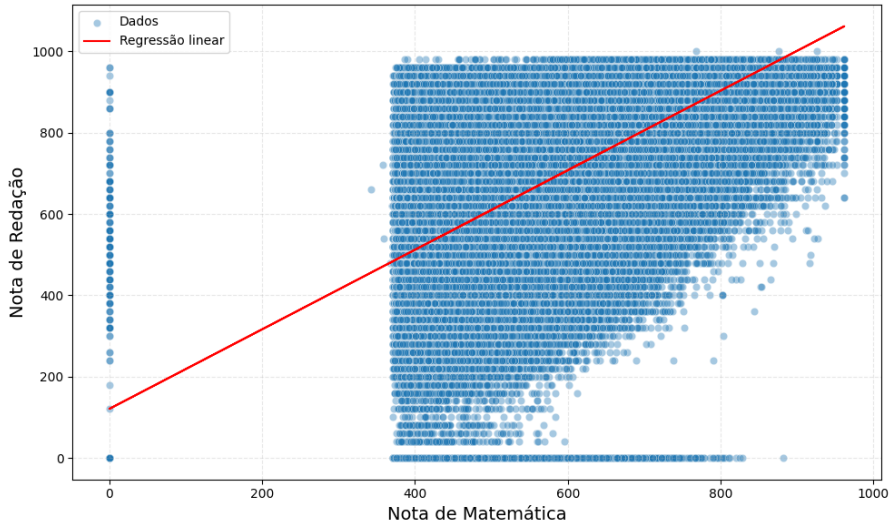


Figure 5. Relação entre as notas de Matemática e Redação com reta de regressão ajustada

4. Teste de Hipótese Paramétrico: Comparação das Médias de Redação entre Escolas Públicas e Privadas

Nesta seção, busca-se verificar se há diferença significativa entre as médias das notas de Redação dos estudantes oriundos de escolas públicas e privadas. Para isso, foi aplicado o teste t de Student para duas amostras independentes, considerando variâncias desiguais (teste de Welch).

Hipóteses

- H_0: $\mu_{pública} = \mu_{privada}$ (as médias são iguais)
- H_1: $\mu_{pública} \neq \mu_{privada}$ (as médias são diferentes)

Fórmula da Estatística do Teste

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Onde:

- $\bar{X}_1$, $\bar{X}_2$: médias amostrais das escolas pública e privada;
- s_1, s_2: desvios padrão amostrais;
- n_1, n_2: tamanhos das amostras.

Aplicação com os dados

Substituindo os valores:

$$\bar{X}_1 = 612{,}74 \quad \bar{X}_2 = 639{,}00$$
$$s_1 = 99{,}45 \quad s_2 = 91{,}08$$
$$n_1 = 1000 \quad n_2 = 1000$$

$$t = \frac{612{,}74 - 639{,}00}{\sqrt{\frac{99{,}45^2}{1000} + \frac{91{,}08^2}{1000}}} = \frac{-26{,}26}{\sqrt{9{,}89 + 8{,}30}} = \frac{-26{,}26}{\sqrt{18{,}19}} \approx \frac{-26{,}26}{4{,}27} \approx -6{,}15$$

Valor de p

A partir da distribuição t com graus de liberdade aproximados (via correção de Welch), obteve-se um p-valor inferior a $0{,}001$.

Conclusão

Como o p-valor é significativamente menor que o nível de significância usual ($\alpha = 0{,}05$), rejeita-se a hipótese nula. Conclui-se que há evidência estatística de que as médias das notas de Redação diferem significativamente entre estudantes de escolas públicas e privadas.

5. Teste de Hipótese Não Paramétrico: Mann–Whitney U

Como alternativa ao teste t de Student, que pressupõe normalidade e homogeneidade de variâncias, aplicou-se também o teste não paramétrico de Mann–Whitney U. Este teste compara duas amostras independentes e verifica se elas provêm da mesma distribuição, sendo especialmente útil quando os dados apresentam assimetria ou valores discrepantes.

Hipóteses

- H_0: as distribuições das notas de Redação são iguais entre estudantes de escolas públicas e privadas;
- H_1: as distribuições das notas são diferentes.

Aplicação do teste

O teste foi aplicado sobre as mesmas amostras utilizadas na análise paramétrica anterior (1000 estudantes de cada grupo). O valor do teste estatístico obtido foi $U = 412{,}598$ com um valor de $p < 0{,}001$.

Fórmula do teste U de Mann–Whitney

A estatística U é calculada como:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 \quad e \quad U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

Onde:

- n_1 e n_2 são os tamanhos das amostras dos grupos 1 e 2;
- R_1 e R_2 são as somas dos postos dos grupos 1 e 2;
- U é o menor entre U_1 e U_2.

Cálculo com os dados

Considerando:

- $n_1 = n_2 = 1000$
- $R_1 = 934{,}712{,}0$ (soma dos postos das escolas públicas)
- $R_2 = 1{,}065{,}288{,}0$ (soma dos postos das escolas privadas)

Aplicando na fórmula:

$$U_1 = (1000 \cdot 1000) + \frac{1000 \cdot 1001}{2} - 934{,}712 = 1{,}000{,}000 + 500{,}500 - 934{,}712 = 565{,}788$$

$$U_2 = 1{,}000{,}000 + 500{,}500 - 1{,}065{,}288 = 435{,}212$$

Logo, a estatística U adotada é:

$$U = \min(U_1, U_2) = 435{,}212$$

O valor crítico tabelado para $n_1 = n_2 = 1000$ é muito inferior ao valor observado. O p-valor computado numericamente foi $< 0{,}001$.

Interpretação

Como o p-valor é inferior ao nível de significância usual ($\alpha = 0{,}05$), rejeitamos H_0 em favor de H_1. Assim, conclui-se que a distribuição das notas de Redação difere significativamente entre os dois tipos de escola, reforçando os achados da análise paramétrica.

Conclusão

Como o p-valor é inferior ao nível de significância adotado ($\alpha = 0{,}05$), rejeita-se a hipótese nula. Isso indica que as distribuições das notas de Redação diferem significativamente entre estudantes de escolas públicas e privadas. Este resultado corrobora a conclusão obtida pelo teste paramétrico.

6. Conclusão

O presente trabalho teve como objetivo aplicar técnicas estatísticas para investigar padrões e relações no desempenho dos participantes do ENEM 2024, utilizando como foco principal as notas das provas de Matemática e Redação.

Inicialmente, a análise descritiva permitiu observar que estudantes de escolas privadas apresentaram, em média, desempenho superior nas duas áreas avaliadas, tanto em termos de média quanto de mediana, quando comparados aos estudantes de escolas públicas.

Posteriormente, aplicou-se a técnica de regressão linear simples para investigar a relação entre as notas de Matemática (variável independente) e Redação (variável dependente). Os resultados indicaram uma associação positiva, representada pela equação

$\hat{Y} = 320{,}84 + 0{,}59 \cdot X$, com coeficiente de determinação $R^2 = 0{,}24$. Embora estatisticamente significativa, essa relação revelou-se de intensidade moderada, indicando que outros fatores influenciam substancialmente o desempenho em Redação.

Por fim, foram realizados testes de hipóteses para comparar o desempenho em Redação entre estudantes de escolas públicas e privadas. Tanto o teste paramétrico (t de Student), quanto o teste não paramétrico (Mann–Whitney U) apontaram diferença significativa entre os grupos analisados, com $p < 0{,}001$ em ambos os casos. Dessa forma, rejeita-se a hipótese de igualdade de médias, sendo possível afirmar, com alto grau de confiança, que estudantes de escolas privadas obtiveram desempenho superior na prova de Redação.

Em síntese, o trabalho demonstrou a aplicabilidade das ferramentas estatísticas estudadas em um contexto real e atual, contribuindo para a análise crítica de desigualdades educacionais e evidenciando a importância da Estatística como suporte à formulação de diagnósticos e políticas públicas em educação.