

Relatorio

Fabio Firanzi, Heitor Dias, Julia Fideles, Matheus Soares, Tiago Braga

2025-11-11

O trabalho elaborado pelos alunos Fabio Firanzi, Heitor Dias, Julia Fideles, Matheus Soares e Tiago Braga tem como objetivo a análise estatística dos resultados do ENEM 2024, por meio de uma amostra aleatória que abrange 10% dos resultados. As variáveis utilizadas para a análise são: TP_PRESENCA_CN, TP_PRESENCA_CH, TP_PRESENCA_LC, TP_PRESENCA_MT (presença nas provas de Ciências da Natureza, Ciências Humanas, Linguagens e Códigos e Matemática, respectivamente), NU_NOTA_CN, NU_NOTA_CH, NU_NOTA_LC, NU_NOTA_MT (Notas nas quatro áreas do conhecimento presentes na prova), TP_LINGUA (língua estrangeira escolhida pelo participante), TP_STATUS_REDACAO (Indicação do motivo para a redação ser zerada, além de auxiliar para filtragem de redações não zeradas para os cálculos referentes as notas), NU_NOTA_COMP1, NU_NOTA_COMP2, NU_NOTA_COMP3, NU_NOTA_COMP4, NU_NOTA_COMP5 (notas das cinco competências) e NU_NOTA_REDACAO (notas das redações).

Os resultados obtidos indicam a relação de participantes frequentes, ausentes e eliminados; a análise das notas de cada área do conhecimento e da redação; relação entre as escolhas da língua estrangeira; apresentação dos motivos que levaram uma redação a ser zerada; distribuição das notas das competências.

Presença na prova	Percentual	Frequência
Ausente	26.74%	115880
Presente	73.12%	316843
Eliminado	0.13%	571

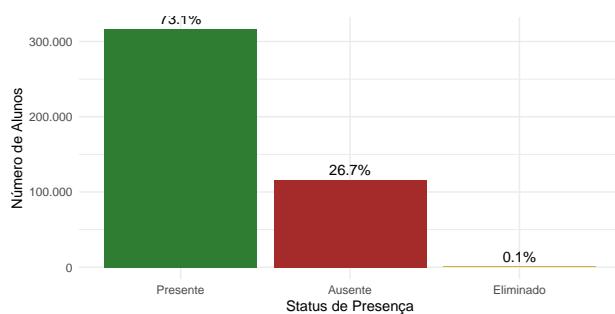
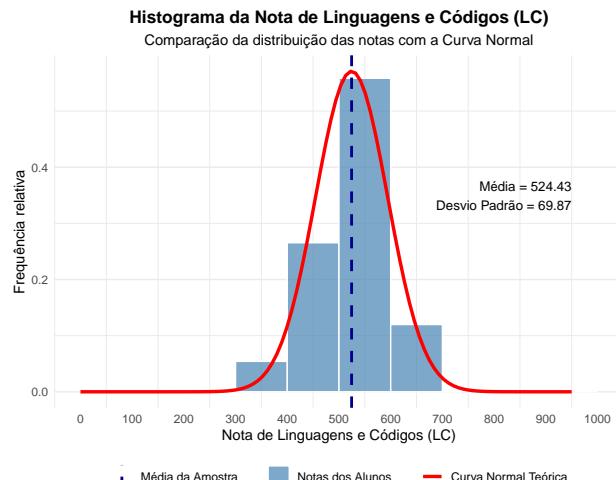


Tabela e gráfico de presença

A tabela “Frequência de presença na prova de LC” e o Gráfico “Presença na prova de LC” exibem a porcentagem de alunos presentes, ausentes e eliminados na prova de linguagens e códigos. Com base nessa tabela, pode-se identificar que a grande maioria, 73,1%, dos alunos estava presente na avaliação, 26,7% ausente e uma minoria de 0,1% foi eliminada antes, durante ou após a prova. Embora a maioria tenha completado a avaliação, o alto índice de ausência sugere a necessidade de estudos futuros para investigar os fatores que contribuem para essa abstenção.



Histograma da nota de LC

Para a criação desse histograma foi utilizado a nota dos alunos presentes e não eliminados na avaliação de Linguagens e códigos, com esse dado sendo analizado em sua frequência, média, desvio padrão e distribuição.

O “Histograma da Nota de LC” ilustra a distribuição de frequência das notas de Linguagens e Códigos para todos os alunos presentes. Ele foi construído sobrepondo uma curva normal teórica (linha vermelha), calculada a partir da média e do desvio padrão da amostra, sobre o histograma das notas reais (barras azuis).

Tabela 1: Tabela Comparativa: LC vs MT (para 298.976 alunos presentes em ambas e com nota > 0)

Prova	Média	Mediana	D. Padrão	Min	MáxNE	Região	Média	Mediana	Var	D. Padrão	Min	Máx
LC	526.70	533.3	68.03	298.8	795.8	S	542.13	548.0	3935.23	62.73	298.8	795.8
MT	527.24	499.1	113.91	342.8	961.9	CO	527.52	533.5	4542.66	67.40	298.8	777.4

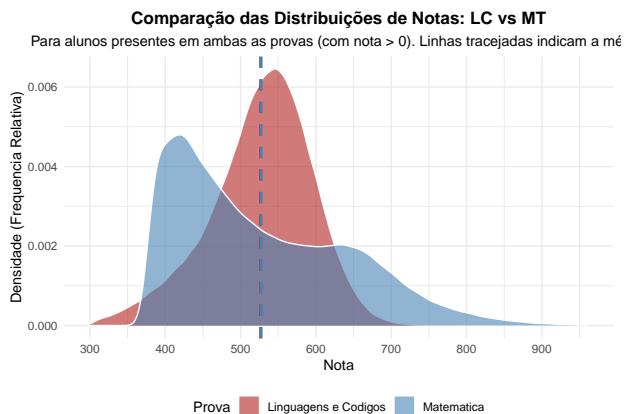
A maior concentração de notas é de 500-600 pontos, contendo aproximadamente 55% dos alunos, o desvio padrão das notas é de 69,87.

Tabela e gráfico LC vs MT

A tabela “Tabela comparativa: LC vs MT” ilustra as, médias, medianas, desvios-padrão, notas mínimas e notas máximas das provas de linguagens e matemática, considerando apenas os alunos que estiveram presentes em ambas delas.

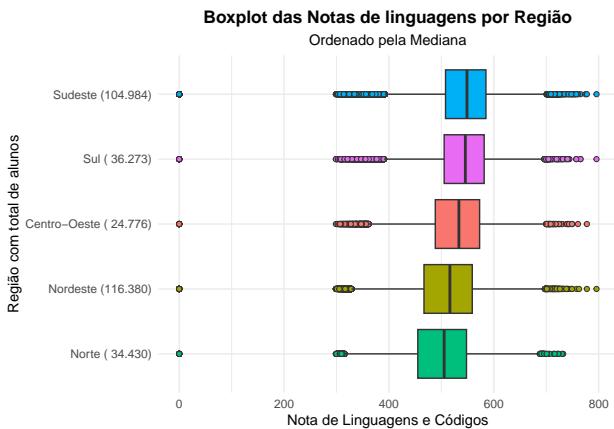
O gráfico “comparação das distribuições de notas:LC vs MT” exibe os gráficos de distribuição de notas de linguagens e matemática sobrepostos, oferecendo uma comparação visual direta e clara da distribuição de notas das duas matérias.

Embora as médias de linguagens e matemática tenham sido muito próximas, 526,7 e 527,2 respectivamente, a distribuição de notas da prova de matemática é totalmente diferente da de linguagens (que se aproxima de uma normal perfeita) tendo suas notas muito mais distribuídas no gráfico, evidenciando um desvio padrão muito maior.



O Gráfico de Densidade oferece uma comparação visual direta entre as distribuições das notas de Linguagens (vermelho) e Matemática (azul). Este gráfico foi gerado utilizando apenas os alunos que compareceram e obtiveram nota maior que zero em ambas as provas. Percebe-se que a curva de Matemática é mais acha-

tada e espalhada (refletindo o maior desvio padrão), indicando que há uma variabilidade muito maior nas notas. Em contraste, as notas de Linguagens são mais “pontudas” e concentradas em torno de sua média.



O Boxplot complementa a tabela estatística anterior, visualizando a distribuição das notas de Linguagens e Códigos por região. Ele foi criado agrupando os alunos por região e ordenando o eixo Y pela mediana das notas, da maior para a menor. Este gráfico permite uma visualização clara não apenas da mediana (a linha central na caixa), mas também da dispersão do “meio” dos alunos (o tamanho da caixa, ou Intervalo Interquartil) e dos outliers (pontos). Observa-se que as regiões Sudeste e Sul apresentam as medianas mais elevadas, enquanto as regiões Nordeste e Norte mostram um desempenho mediano inferior e uma dispersão de notas (tamanho da caixa) ligeiramente maior, indicando maior variação no desempenho dos alunos dessas regiões.

Tabela e boxplot de notas LC por região

A tabela “Estatísticas das notas de LC por Região” exibe as médias, medianas, variâncias, desvios-padrão, notas mínimas, notas máximas e número aproximado de alunos de cada região do Brasil.

O “Boxplot das notas de linguagens por Região”, ilustra a performance dos alunos de cada uma das regiões do Brasil exibindo a distribuição das notas. As “ca-

xas” representam os 50% centrais dos alunos, com os pontos sendo os “outliers”, ou seja, os fora da média, tanto para baixo quanto para cima e a linha dentro da caixa representa a mediana. Com base no boxplot, a região Sudeste obteve o melhor resultado, com sua média de 542,74 sendo a mais alta seguida de perto pela região Sul e sua média de 540,36, o boxplot também deixa evidente a desigualdade do país, com as regiões Norte e nordeste ficando significativamente atrás das demais.

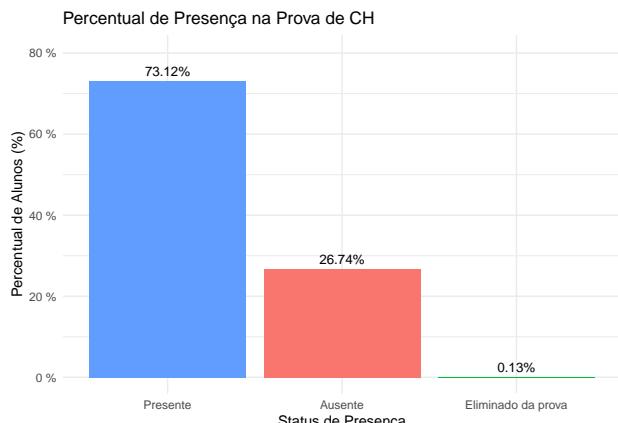
Analise da variavel de presenca na prova de humanas

Foi feita uma analise da area de conhecimento de Ciências Humanas do Enem, utilizando a varivel: TP_PRESENCA_CH. Tal variavel, é classificada como qualitativa possuindo 3 possíveis valores:

- 0: Ausente
- 1: Presente
- 2: Eliminado da Prova

Gráfico de Frequências das presenças no dia da prova

Busca indentificar se o aluno estava presente, ausente ou se foi eliminado da prova de ciências humanas, e, com isso, expressar a porcentagem e os valores absolutos da variável TP_PRESENCA_CH.



O gráfico “Percentual de Presença na Prova de CH”, trata dos dados da variável ‘TP_PRESENCA_CH’. Uma vez que a variável é do tipo qualitativa, a abordagem mais convencional é um gráfico de barras dos percentuais.

Com base na analise do gráfico “Percentual de Presença na Prova CH” foi possível determinar que no Exame Nacional do Ensino Médio (ENEM), edição de 2024, o número de alunos presentes foi de aproxi-

madamente 2,73 vezes maior que o número de alunos ausentes. Além disso, percebe-se que a quantidade de alunos elimanados na prova de Ciências Humanas foi extremamente pequena - 0,1% - comparado com os percentuais da coluna “Presença” e da coluna “Ausente”.

Analise da variavel Notas da prova de Ciências Humanas

Tabela 2: Tabela Resumo: Estatísticas das Notas de CH por Região

Regiao	Media	Mediana	Variancia	Desvio Padrão	Mínimo	Máximo
Sudeste	533.58	540.4	7482.78	86.50	283.8	819.7
Sul	527.92	534.0	7095.26	84.23	283.8	819.7
Centro-Oeste	514.69	518.7	8192.05	90.51	283.8	817.4
Nordeste	495.07	494.9	8216.95	90.65	283.8	819.7
Norte	484.05	481.2	7511.50	86.67	283.8	808.2

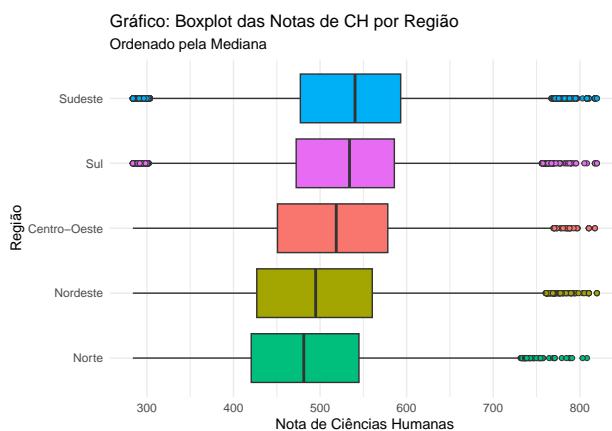
A “Tabela Resumo: Estatísticas das Notas de CH por Região” apresenta a distribuição das medidas: media, mediana, variância, desvio padrão, valor mínimo, valor máximo e a frequência relativa percentual. Essa divisão foi feita por região do Brasil. Com base nisso, podemos identificar claramente o desempenho superior da região Sudeste.

Liderança Clara: O Sudeste lidera em ambos os indicadores de performance, possuindo a maior Média (593,12) e, mais importante, a maior Mediana (601,8).

O “Grupo de Ponta”: Embora o Sudeste seja o primeiro, ele faz parte de um “grupo de alta performance” juntamente com as regiões Sul (Mediana 597,5) e Centro-Oeste (Mediana 591,3). Estas três regiões estão claramente destacadas das regiões Nordeste (Mediana 562,9) e Norte (Mediana 555,0).

A Armadilha da Média: Em todas as regiões, a Média é “puxada” para baixo por notas mais fracas (assimetria à esquerda). Por isso, a Mediana é a métrica mais justa para a comparação, e nela o Sudeste também vence.

Gráfico Boxplot da variável notas



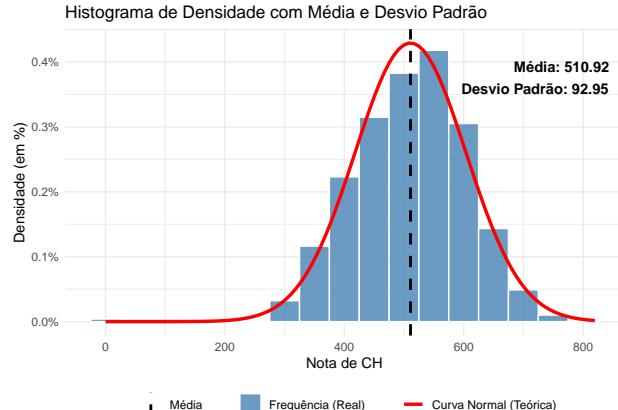
O gráfico “Boxplot das Notas de CH por Região”, compara o desempenho central (a mediana) das notas de Ciências Humanas (CH) entre as cinco grandes regiões do Brasil. Além disso, o gráfico representa os valores “extremos”, os outliers, da variável ‘NU_NOTA_CH’, contribuindo para uma análise de desempenho na prova do Enem.

As caixas das regiões Sul e Sudeste são visivelmente mais longas (largas) do que as das outras regiões. Isto significa que a diferença de nota entre o aluno do percentil 25 e o do percentil 75 é maior. Ou seja, embora tenham o melhor desempenho, são regiões internamente mais “desiguais” ou “inconstantes”.

O Boxplot confirma que a região Sudeste tem o melhor desempenho geral em Ciências Humanas, não apenas na mediana, mas no “corpo” principal dos seus alunos (o miolo de 50%). No entanto, esta alta performance vem acompanhada de uma maior desigualdade interna (maior dispersão) representada pelo comprimento maior e um alto número de Outliers, tanto Outliers superiores (notas > 750) quanto os Outliers inferiores (notas < 300), um padrão também visto na região Sul.

Histograma de Densidade com Média e Desvio Padrão

Para os dados quantitativos contínuos da variável NU_NOTA_CH, que representa as notas dos alunos na prova de ciências humanas, criamos classes (faixas de valores) para a desenvolver um histograma com uma Normal sobreposta.



O “Histograma de Densidade com Média e Desvio Padrão” apresenta a densidade das notas de Ciências Humanas (CH), indicando a distribuição dos valores observados. As barras em azul representam a frequência relativa das notas, enquanto a linha vermelha mostra a curva normal teórica ajustada a partir dos dados. Além disso, a linha pontilhada vertical identifica a média das notas (510,92 pontos). Nesse cenário, essa variável possui O desvio-padrão igual a 92,95 pontos, informado no canto superior direito do gráfico. Todos esses fatores auxiliam para a execução de uma análise a cerca da distribuição das notas da prova de Ciências Humanas.

A faixa de nota com maior frequência (a Classe Modal) é [500, 550], contendo 21,05% dos alunos.

As notas de Ciências Humanas apresentam uma distribuição aproximadamente normal, bem representada pela curva teórica sobreposta ao histograma. A média foi de 510,92 pontos, indicando o desempenho central dos estudantes, enquanto o desvio-padrão de 92,95 pontos mostra dispersão moderada ao redor da média. A forma da distribuição confirma que os dados seguem um padrão típico e estável, adequado para análises baseadas em normalidade.

1.3 Correlação e Regressão Linear Simples

Para essa análise, buscou-se indentificar o poder de correlação de duas variáveis, notas de Ciências Humanas e notas de Linguagens e Códigos.

Nesta análise, vamos investigar a relação entre duas variáveis quantitativas: NU_NOTA_LC (Linguagens e Códigos) e NU_NOTA_CH (Ciências Humanas).

- Variável Independente (X): NU_NOTA_LC
- Variável Dependente (Y): NU_NOTA_CH

O objetivo é responder: “A nota de Linguagens pode prever a nota de Humanas?”

Para isso, foi feita um processo de filtragem, matendo somente os alunos com notas válidas (sem N) e os alunos com notas maiores que zero em ambas as variáveis. Com isso, o número total de observações válidas para a regressão: foi de 316201.

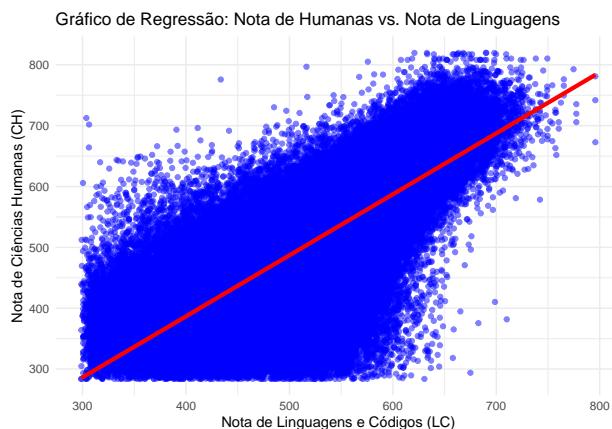
Coeficiente de Correlação de Pearson

`## O Coeficiente de Correlação (r) entre LC e CH é: 0.7613`

Isso significa que, em geral, alunos que tiram notas mais altas em Linguagens também tiram notas mais altas em Humanas. Isso demonstra que as duas variáveis possuem uma forte correlação.

Regressão Linear

`## `geom_smooth()` using formula = 'y ~ x'`



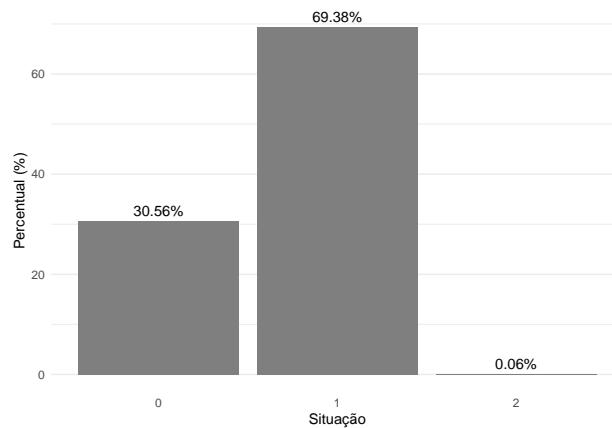
O “Gráfico de Regressão: Nota de Humanas vs. Nota de Linguagens” confirma a correlação positiva. Nesse cenário, os pontos estão razoavelmente agrupados ao redor da linha vermelha, que sobe da esquerda para a direita, confirmando a tendência de que notas altas em LC acompanham notas altas em CH. Para o modelo elaborado, temos a seguinte equação: $\text{Nota_CH} = 93.30 + 0.82 * \text{Nota_LC}$. Através dessa equação, é possível dizer que para cada 1 ponto que um aluno ganha em na prova de linguagens (NU_NOTA_LC), espera-se que sua nota em humanas (NU_NOTA_CH) aumente, em média, 0.82 pontos. Logo, é possível afirmar que a nota de Linguagens em geral consegue prever a nota de Humanas.

Análise das variáveis de Ciências da Natureza

Presença na prova	Percentual	Frequência
Ausente	30.56%	132415
Presente	69.38%	300640
Eliminado	0.06%	239

A partir da Tabela 2, observa-se a distribuição dos participantes em relação à presença na prova de Ciências da Natureza, destacando-se a proporção de estudantes que compareceram, se ausentaram ou foram eliminados. Essa informação é importante para entender o engajamento dos inscritos na aplicação da prova.

Gráfico 1: Presença na Prova de Ciências da Natureza



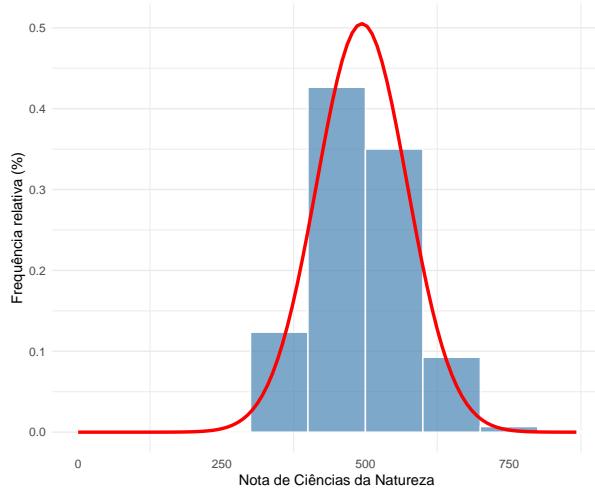
O Gráfico 1 apresenta visualmente a comparação entre estudantes presentes, ausentes e eliminados na prova de Ciências da Natureza, facilitando a interpretação da participação na avaliação.

Tabela 3: Frequência Intervalar da Nota de Ciências da Natureza

Intervalo de nota	Percentual
0 – 100	0.02%
100 – 200	0%
200 – 300	0%
300 – 400	12.32%
400 – 500	42.64%
500 – 600	35.04%
600 – 700	9.28%
700 – 800	0.69%
800 – 900	0.01%
900 – 1000	0%

A Tabela 3 apresenta a distribuição das notas de Ciências da Natureza em classes de 100 pontos, permitindo identificar em quais faixas se concentram os maiores e os menores desempenhos dos estudantes.

Gráfico 2: Histograma das Notas de Ciências da Natureza com Curva Normal

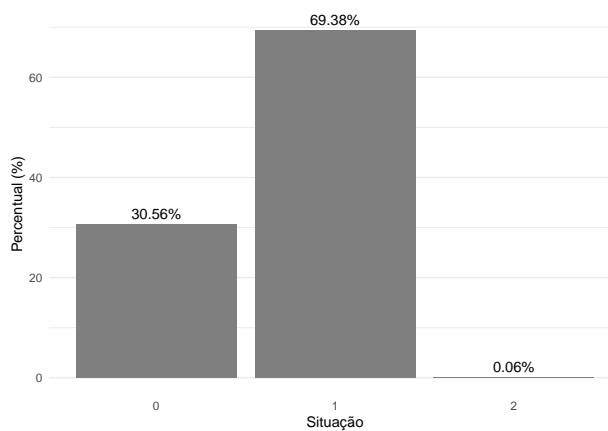


O Gráfico 2 compara a distribuição relativa das notas de Ciências da Natureza com uma curva normal ajustada pela média e pelo desvio padrão da amostra, permitindo avaliar se o desempenho se aproxima de um padrão aproximadamente normal ou se há assimetrias importantes.

Tabela 4: Frequência - Presença na Prova de Matemática

Presença na prova	Percentual	Frequência
Ausente	30.56%	132415
Presente	69.38%	300640
Eliminado	0.06%	239

Gráfico 3: Presença na Prova



1. Análise da Presença na Prova de Matemática

A Tabela 4 mostra que 30,56% dos inscritos não compareceram, enquanto 69,38% estiveram presentes e

apenas 0,06% foram eliminados. Essa discrepância revela dois pontos centrais:

1.1 Comparação entre presença e ausência

A quantidade de presentes é mais do que o dobro da de ausentes, indicando que, embora a taxa de ausência seja significativa, a maior parte dos estudantes permanece engajada e comparece à prova.

1.2 Ausências ainda são um desafio estrutural

Com mais de 132 mil ausentes, observa-se uma dificuldade que pode estar associada a:

- Logística de deslocamento
- Realização da prova em dia único
- Desmotivação ou insegurança com o exame

O gráfico 3 reforça visualmente essa diferença, tornando evidente o contraste entre o volume de presentes e ausentes.

Tabela 5: Frequência Intervalar da Nota de Matemática

Intervalo de nota	Percentual
0 – 100	0.03%
100 – 200	0%
200 – 300	0%
300 – 400	10.1%
400 – 500	40.17%
500 – 600	22.48%
600 – 700	18.46%
700 – 800	7.06%
800 – 900	1.52%
900 – 1000	0.18%

3. Distribuição das Notas de Matemática

A Tabela 5 divide as notas em intervalos de 100 pontos, permitindo observar o comportamento geral do desempenho dos estudantes.

3.1 Concentração forte em torno do meio da distribuição

Os intervalos com maiores proporções são:

- 400–500: 40,17%
- 500–600: 22,48%

Somando esses dois grupos, mais de 62% dos participantes estão entre 400 e 600 pontos, indicando que:

- A maior parte dos estudantes teve desempenho mediano, nem muito baixo, nem muito alto.
- A dificuldade da prova parece adequadamente calibrada para centralizar alunos em torno da média.

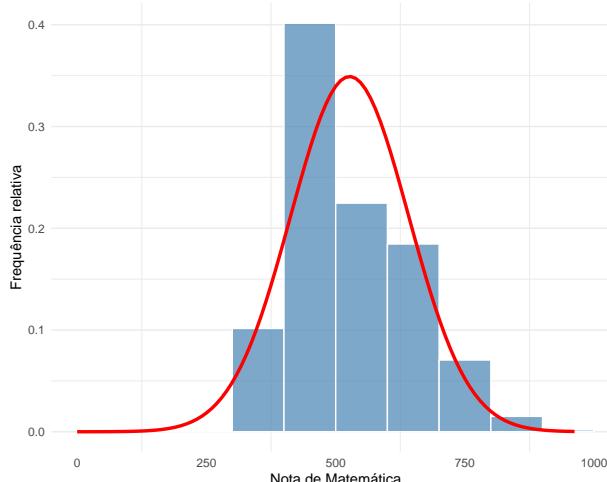
3.2 Baixa incidência de notas extremas

Notas muito baixas (<300) e muito altas (>800) representam percentual muito pequeno:

- <100 pontos: 0,03%
- 800–900: 1,52%
- 900–1000: 0,18%

Esses valores sugerem que a prova discrimina bem alunos com domínio intermediário, mas raramente produz desempenhos extremos.

Gráfico 4: Histograma das Notas de Matemática com Curva Normal



4. Ajuste à Curva Normal da Distribuição

O histograma com a curva normal ajustada mostra que:

4.1 A distribuição é aproximadamente normal

A forma do gráfico indica:

- Picos de densidade próximos ao centro
- Decaimento simétrico nos extremos

- Ausência de longas caudas que caracterizariam assimetria acentuada

Isso sugere que:

- O exame conseguiu distribuir os alunos segundo um padrão próximo ao esperado estatisticamente.
- Há equilíbrio entre questões fáceis, intermediárias e difíceis.

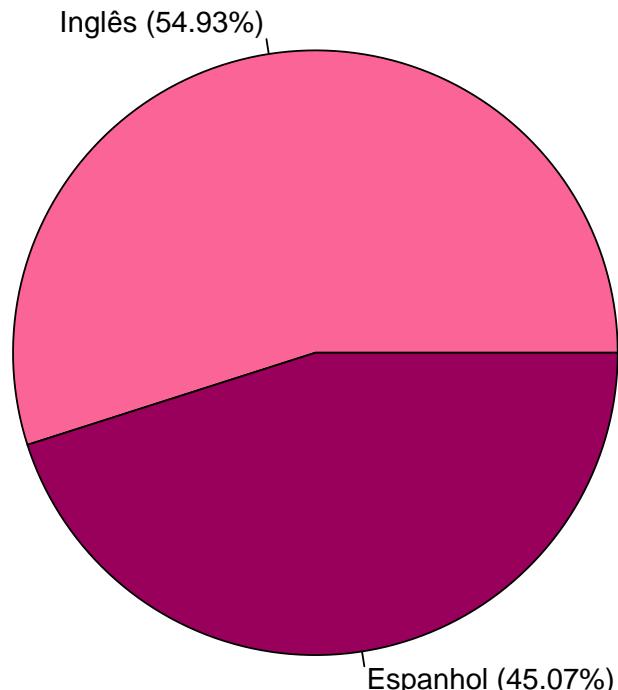
4.2 Diferença entre curva teórica e real

Pequenas diferenças entre a curva real e a curva normal ajustada indicam que:

- Alguns níveis de nota são mais frequentes que o previsto pela normal (natural em testes educacionais).

Língua Estrangeira	Percentual	Frequência
Inglês	54.93%	237997
Espanhol	45.07%	195297

Observa-se, a partir da tabela, que aproximadamente 55% dos participantes escolheu a opção “inglês” para a prova de língua estrangeira e 45% dos participantes escolheu “espanhol”, o que evidencia que a maior parte dos estudantes optou pela língua inglesa.



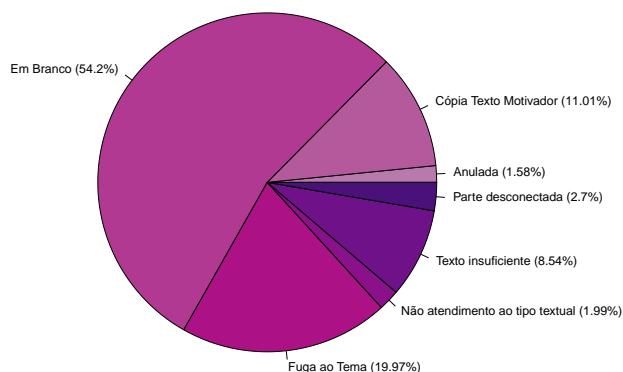
O gráfico apresenta, de forma imagética, o contraste entre a quantidade do todo que optou pela língua inglesa e a quantidade que optou pela língua espanhola.

Tabela 6: Frequência - Motivo do Zero na Redação

Motivo	Percentual
Anulada	1.58%
Cópia Texto Motivador	11.01%
Em Branco	54.2%
Fuga ao Tema	19.97%
Não atendimento ao tipo textual	1.99%
Texto insuficiente	8.54%
Parte desconectada	2.7%

Observando a tabela, percebe-se que mais da metade (54,2%) das redações zeradas teve como motivo a redação em branco, sendo o principal motivo para isso. Os demais motivos, em ordem após esse, são fuga ao tema (19,97%), cópia do texto motivador (11,01%), texto insuficiente (8,54%), parte desconectada (2,7%), não atendimento ao tipo textual (1,99%) e anulação (1,58%). Com isso, é possível constatar que, a maior parte dos participantes que teve a redação zerada não escreveu ou escreveu insuficientemente (54,2% em branco e 8,54% texto insuficiente), não entendeu o tema ou não foi capaz de relacionar o tema a conhecimentos prévios e copiou os textos motivadores.

Gráfico 4: Motivos para Redação Zerada



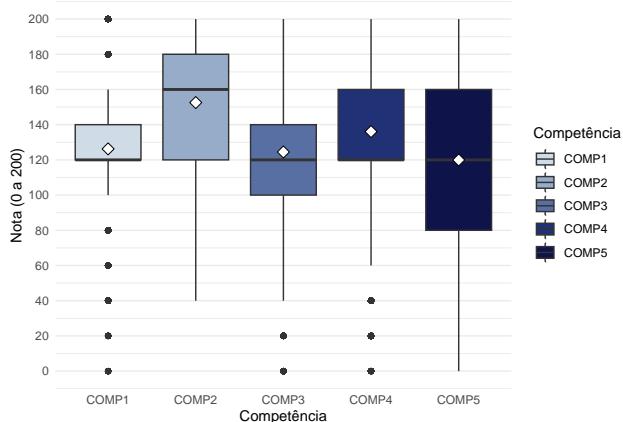
O gráfico, apresenta de forma imagética, a relação entre os motivos das redações zeradas e a frequência relativa deles. É possível observar que mais da metade das redações foi zerada por um motivo (Em Branco) e a outra metade foi por motivos diversos, o que evidencia que um motivo se sobressai aos demais.

Tabela 7: Frequência - Notas nas Competências

Nota	C1	C2	C3	C4	C5
0	0.01%	0%	0.02%	0.02%	6.67%
20	0.01%	0%	0.03%	0.04%	2.83%
40	0.21%	0.8%	1.81%	0.35%	4.49%
60	0.96%	0.54%	2.74%	1.49%	4.57%
80	7.78%	2.03%	10.38%	7.34%	10.98%
100	13.92%	3.55%	13.08%	10.85%	10.31%
120	37.04%	27.06%	34.67%	32.03%	15.22%
140	17.12%	11.24%	13.59%	12.97%	9.66%
160	21.53%	18.89%	14.87%	14.99%	11.82%
180	1.35%	17.39%	6.52%	10.55%	9.27%
200	0.07%	18.51%	2.28%	9.37%	14.19%

A tabela apresenta a relação entre as notas e a frequência relativa da aparição da mesma em cada uma das 5 competências da redação, em redações que não foram zeradas. Ao comparar as frequências das notas em todas as competências, as principais constatações são: “A competência com maior taxa de 0 é a competência 5 (6,67%)”, “A competência 1 possui a maior parte das notas concentradas entre 100 e 160 (com porcentagens mais significativas, entre 13,92% e 37,04%)”, “A competência 1 é a única cuja taxa de notas 200 é significativamente menor”, “A competência 5 possui maior distribuição de notas que as demais (todas as porcentagens estão acima de 2%)” e “A competência 2 é a competência com maior taxa de notas entre 120 e 200 (93,09% das notas está nessa faixa)”.

Gráfico 5: Distribuição das Notas por Competência



O boxplot representa, graficamente, as informações presentes na tabela. Os pontos indicam valores que estão fora da concentração de resultados, mas que tiveram aparição (outliers). As principais constatações são: A competência 1 é a que possui menor dispersão, ou seja, a maioria dos resultados está concentrada em uma faixa pequena (100 a 160) e possui outliers em todas as notas fora dessa faixa. A competência 2 possui concentração de notas mais altas que as demais,

o que é evidenciado por abranger notas mais altas e não possui outliers, ou seja, todos os resultados estão na faixa estipulada. A competência 5 possui maior dispersão, ou seja, abrange todas as notas possíveis e está mais distribuída.

Tabela 9: Frequência - Notas da Redação

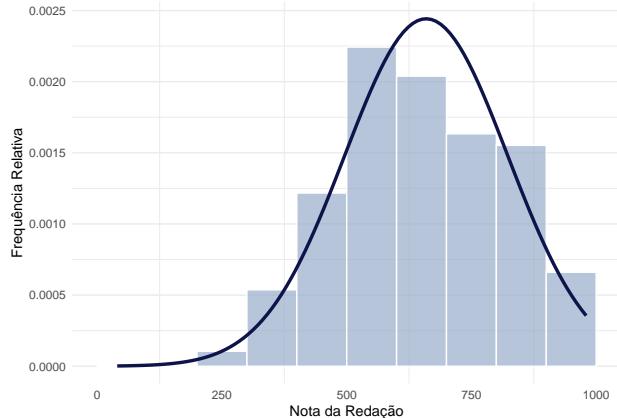
Intervalo de Notas	Percentual
$x \leq 100$	0.02%
$100 < x \leq 200$	0.1%
$200 < x \leq 300$	1.05%
$300 < x \leq 400$	5.37%
$400 < x \leq 500$	12.17%
$500 < x \leq 600$	22.43%
$600 < x \leq 700$	20.39%
$700 < x \leq 800$	16.34%
$800 < x \leq 900$	15.52%
$900 < x$	6.61%

Média das notas: 659.33

Desvio Padrão: 163.34

A tabela indica a frequência relativa, em porcentagem, das faixas de notas da redação. Aproximadamente metade dos resultados está centrado entre 500 e 700 ($22,43\% + 20,39\%$), enquanto os demais intervalos têm frequência cada vez menor conforme se afasta do centro, tendendo para os intervalos extremos.

Gráfico 6: Distribuição das Notas da Redação



O gráfico relaciona o comportamento da frequência dos intervalos das notas da redação (presentes na tabela 4), indicado pelas barras, com o comportamento da normal, linha, para a verificação da semelhança entre ambos. É possível observar que a maior frequência de aparição dos valores converge para o centro, enquanto diminui ao se aproximar das extremidades, o que se assemelha ao comportamento teórico da normal.

Com base na análise estatística, é possível constatar

que a presença no primeiro dia de prova (provas de humanas e linguagens) é maior que no segundo dia (natureza e matemática), o que mostra que, embora não seja uma taxa muito grande, ainda há desistência após o primeiro dia de prova. A média das quatro áreas do conhecimento varia entre 400 e 600, o que mostra que a distribuição das notas tende ao centro e tem menor distribuição nas extremidades, convergindo para o padrão da normal, fator que torna a prova justa por nivelar os níveis das questões. Além disso, as notas da redação é centrada entre 500 e 700, o que também mantém a tendência central de notas observada.