

Relatorio

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

2025-11-11

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

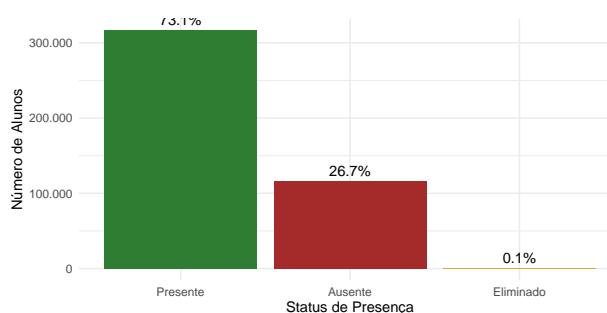


Tabela e gráfico de presença

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.

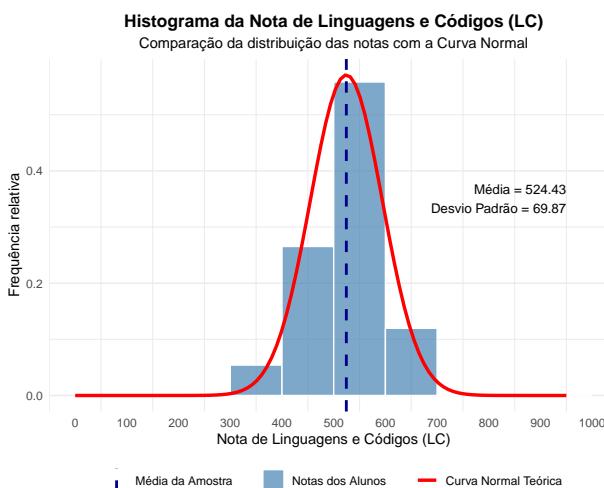


Tabela 1: Tabela Comparativa: LC vs MT (para 298.976 alunos presentes em ambas e com nota > 0)

Prova	Média	Mediana	Desv. Padrão	Min	Máx
LC	526.70	533.3	68.03	298.8	795.8
MT	527.24	499.1	113.91	342.8	961.9

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).

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

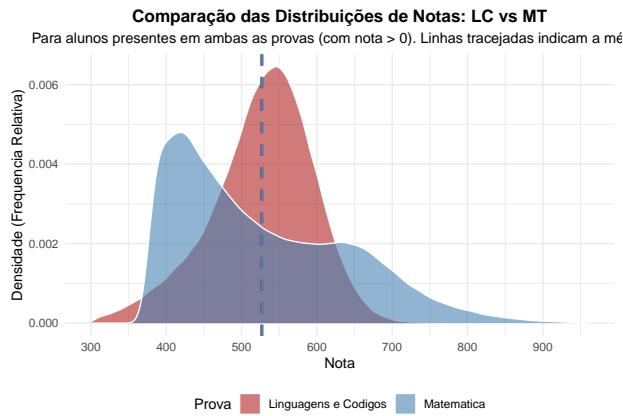
Tabela e gráfico LC vs MT

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.

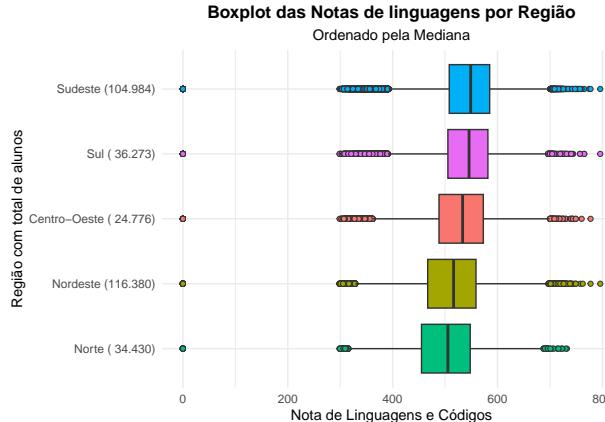
O gráfico “comparação das distribuições de notas:LC vs MT” exibe os gráficos de distribuição de notas de linguagens e matemática sobrepostos, oferecendo uma comparação visual direta e clara da distribuição de notas das duas matérias.

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 achatada 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

Região	Média	Mediana	Var	D. Padrão	Min	Máx
S	542.13	548.0	3935.23	62.73	298.8	795.8
CO	527.52	533.5	4542.66	67.40	298.8	777.4
NE	510.60	516.3	4889.35	69.92	298.8	795.8
N	499.47	505.4	4805.12	69.32	298.8	732.3

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 “caixas” 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.

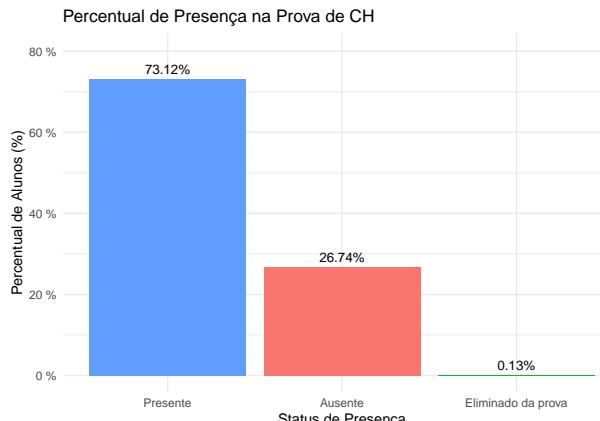
Análise da variável de presença na prova de humanas

Foi feita uma análise da área de conhecimento de Ciências Humanas do Enem, utilizando a variável: TP_PRESENCA_CH. Tal variável, 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 identificar 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 análise 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 aproximadamente 2,73 vezes maior que o número de alunos ausentes. Além disso, percebe-se que a quantidade de alunos eliminados 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

Região	Média	Mediana	Variância	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

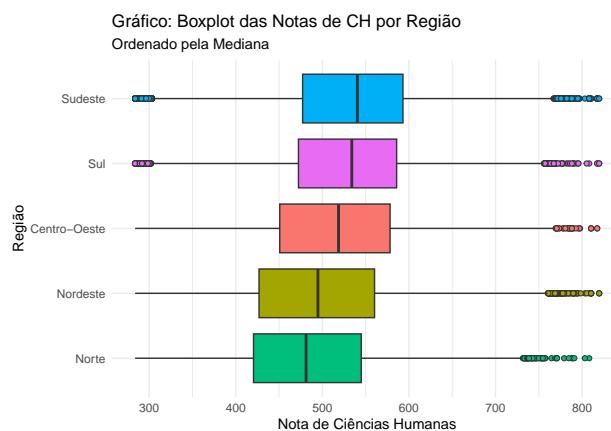
A “Tabela Resumo: Estatísticas das Notas de CH por Região” apresenta a distribuição das medidas: media, mediana, variância, desvio padrão, valor mínimo, valor máximo e a frequência relativa percentual. Essa divisão foi feita por região do Brasil. Com base nisso, podemos identificar claramente o desempenho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



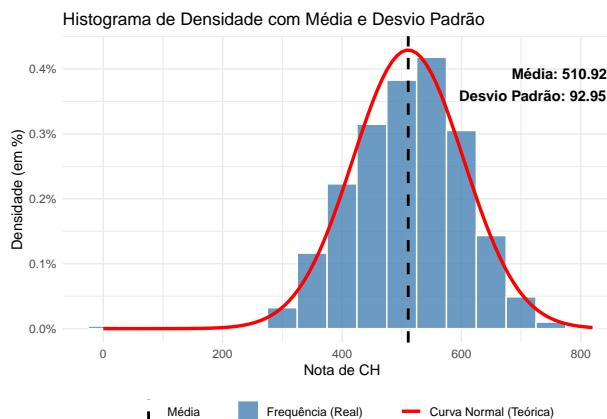
O gráfico “Boxplot das Notas de CH por Região”, compara o desempenho central (a mediana) das notas de Ciências Humanas (CH) entre as cinco grandes regiões do Brasil. Além disso, o gráfico representa os valores “extremos”, os outliers, da variável ‘NU_NOTA_CH’, contribuindo para uma análise de desempenho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

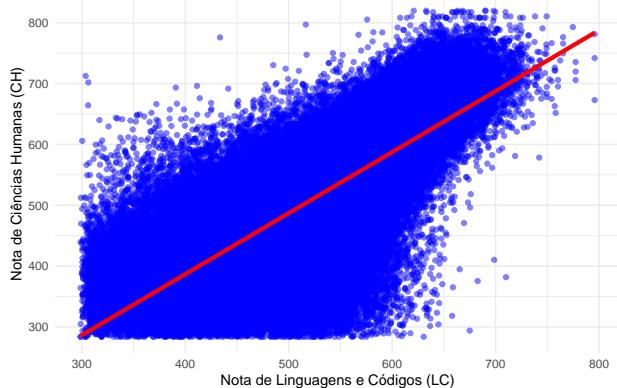
O Coeficiente de Correlação (r) entre LC e CH é: 0.7615

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'

Gráfico de Regressão: Nota de Humanas vs. Nota de Linguagens



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, confirmado a tendência de que notas altas em LC acompanham notas altas em CH. Para o modelo elaborado, temos a seguinte equação: $Nota_{CH} = 93.30 + 0.82 * Nota_{LC}$. Através dessa equação, é possível dizer que para cada 1 ponto que um aluno ganha em linguagens (NU_NOTA_LC), espera-se que sua nota em humanas (NU_NOTA_CH) aumente, em média, 0,82 pontos. Logo, é possível

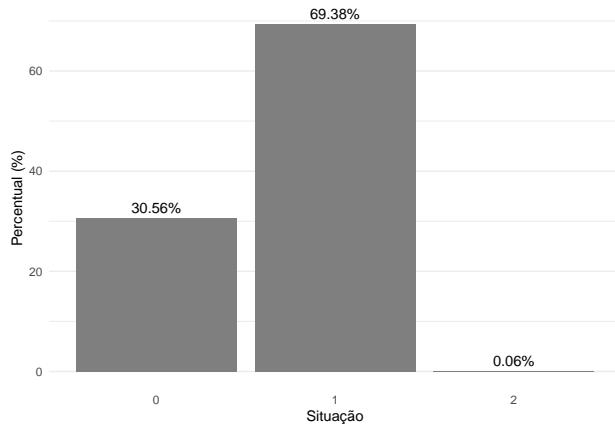
afirmar que a nota de Linguagens em geral consegue prever a nota de Humanas.

Análise das variáveis de Ciências da Natureza

Presença na prova	Percentual	Frequência
0	30.56%	132415
1	69.38%	300640
2	0.06%	239

A partir da Tabela 1, 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

```
## Warning: No shared levels found between `na
## data's fill values.
## No shared levels found between `names(value
## data's fill values.
## No shared levels found between `names(value
## data's fill values.
```



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.

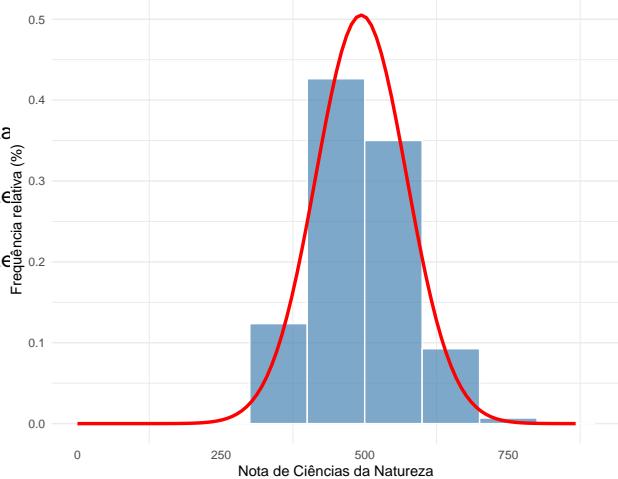
Tabela 2: Frequência Intervalar da Nota de Ciências da Natureza

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 2 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

```
## Warning: Removed 132654 rows containing non-finite outs
## (`stat_bin()`).
```



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.