

Relatorio

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

2025-11-11

Analise da area de conhecimento de Humanas: Heitor

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.

Percentual de Presença na Prova de CH

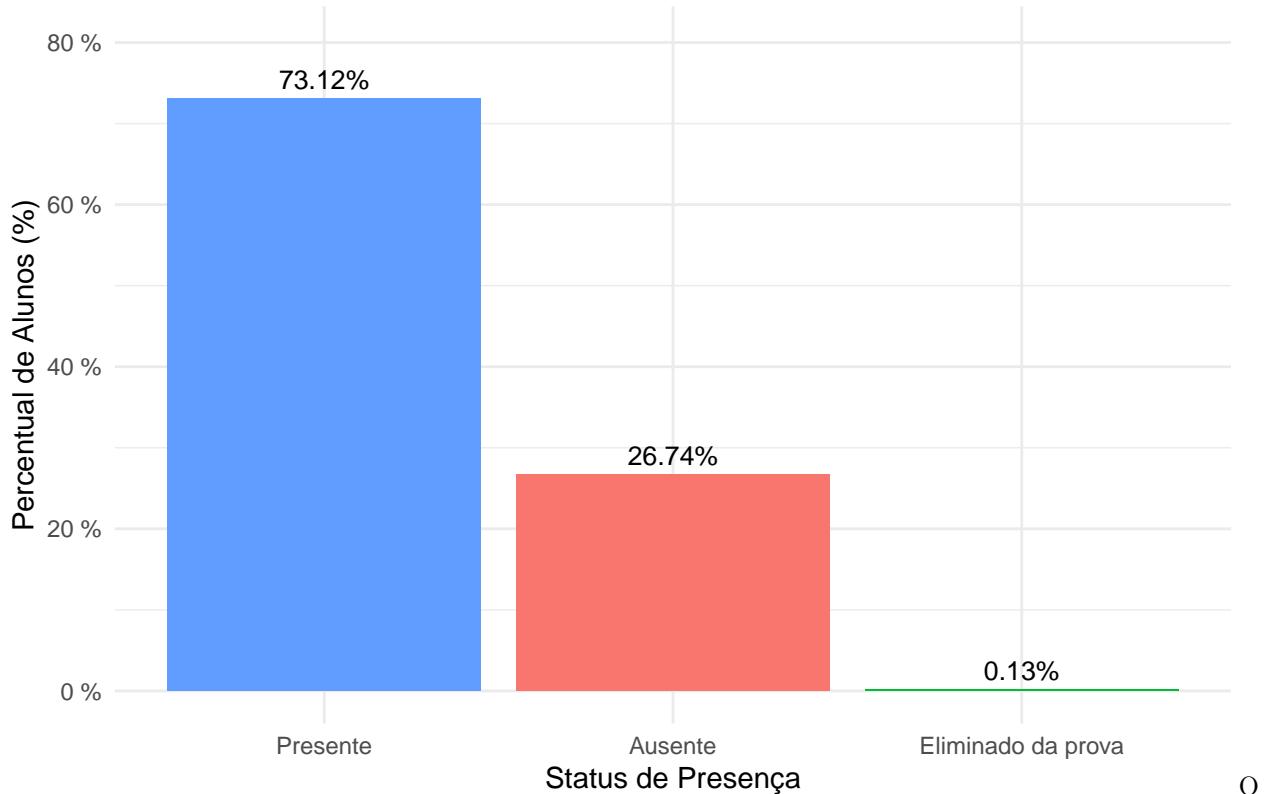


gráfico “Percentual de Presença na Prova de CH”, trata dos dados da variável ‘TP_PRESENCA_CH’. Uma vez que a variável é do tipo qualitativa, a abordagem mais convencional é um gráfico de barras dos percentuais.

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

Regiao	Media	Mediana	Variancia	Desvio Padrao	Minimo	Maximo
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.

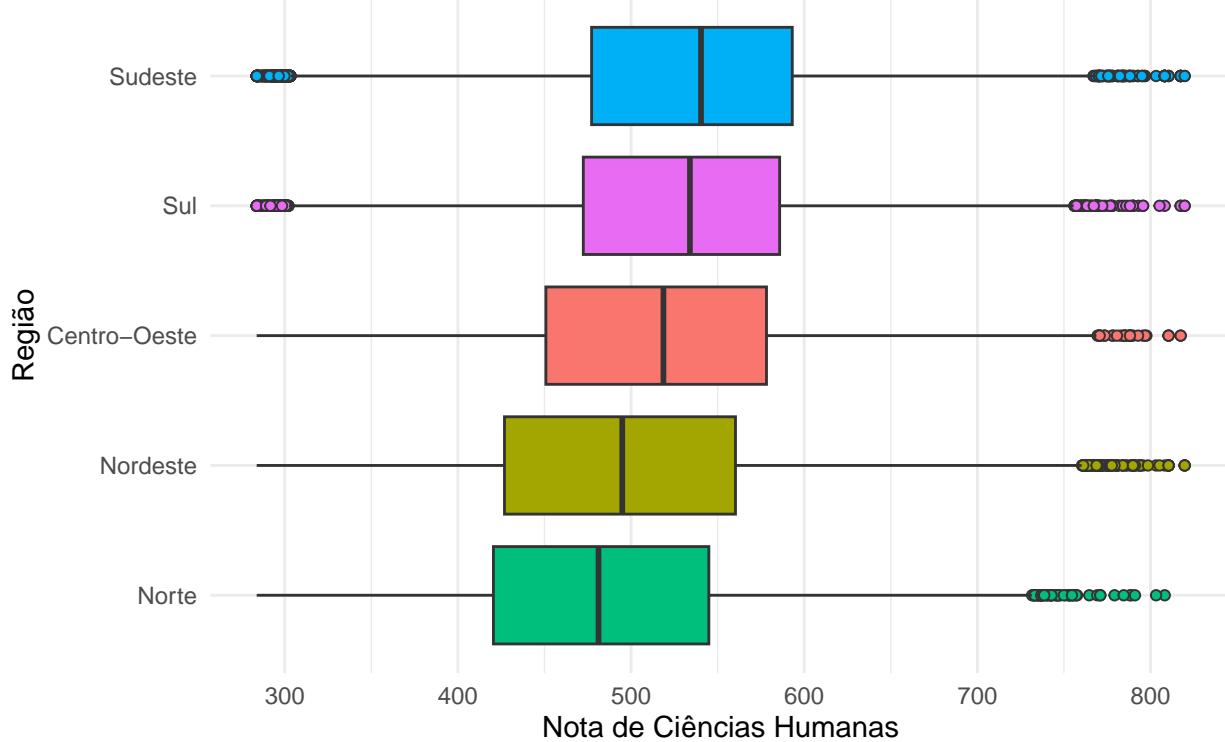
Liderança Clara: O Sudeste lidera em ambos os indicadores de performance, possuindo a maior Média (593,12) e, mais importante, a maior Mediana (601,8).

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

Gráfico: Boxplot das Notas de CH por Região
Ordenado pela Mediana



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.

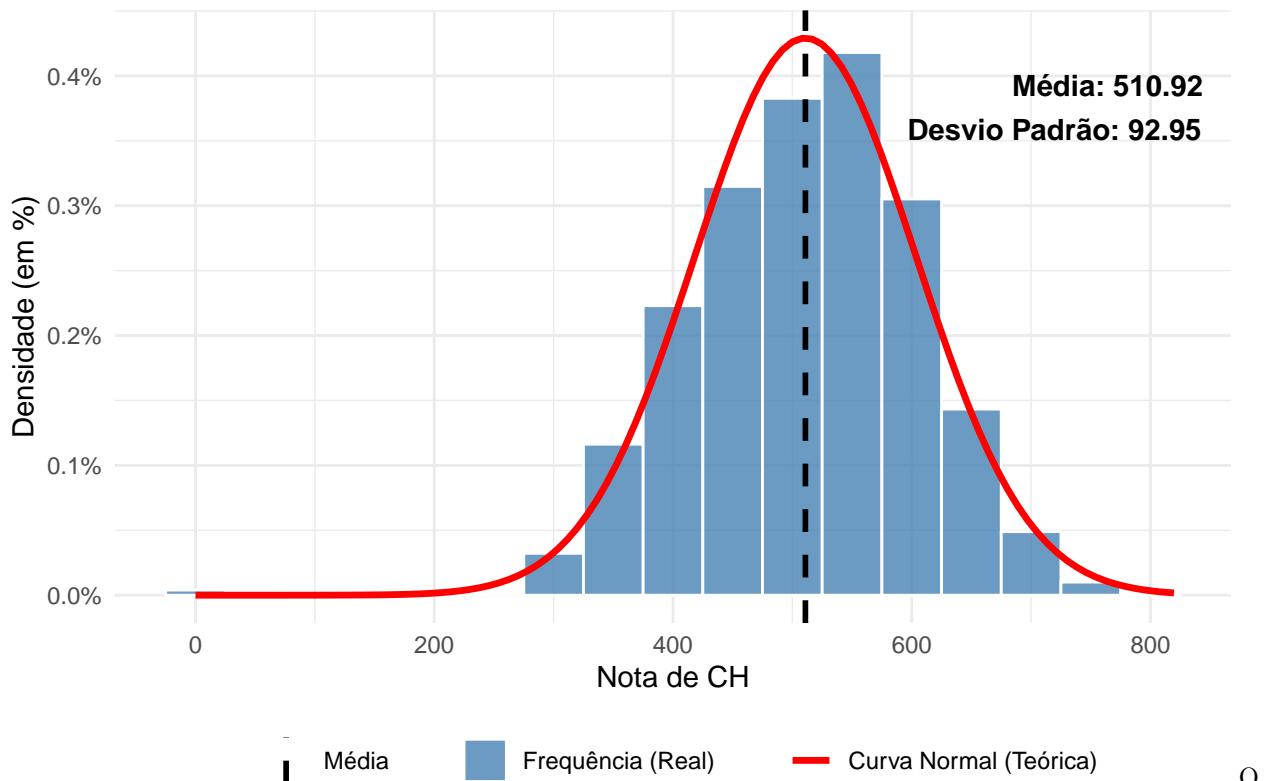
As caixas das regiões Sul e Sudeste são visivelmente mais longas (largas) do que as das outras regiões. Isto significa que a diferença de nota entre o aluno do percentil 25 e o do percentil 75 é maior. Ou seja, embora tenham o melhor desempenho, são regiões internamente mais “desiguais” ou “inconstantes”.

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 desenvolver um histograma com uma Normal sobreposta.

Histograma de Densidade com Média e Desvio Padrão



“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 identificar 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

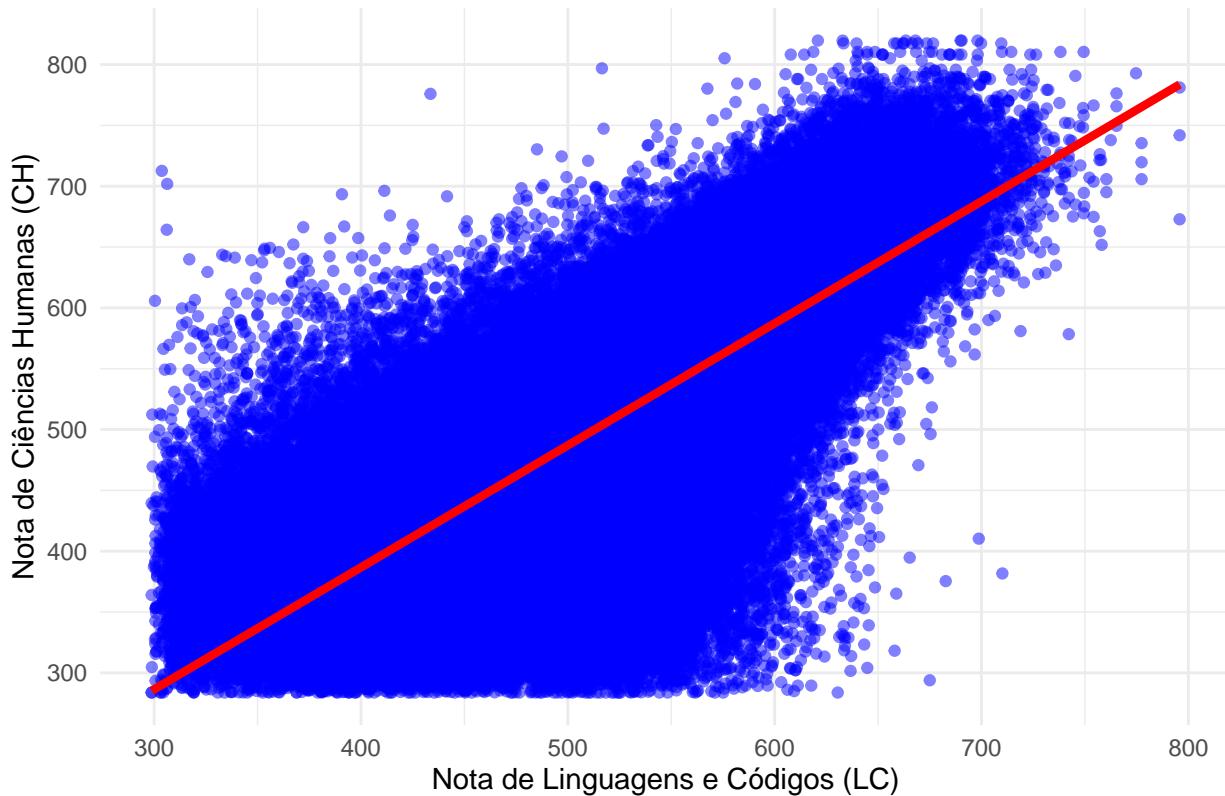
```
## O Coeficiente de Correlação (r) entre LC e CH é: 0.7615
```

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: $\text{Nota_CH} = 93.30 + 0.82 * \text{Nota_LC}$. 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.