

# Relatorio

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

2025-11-11

## Analise da area de conhecimento de Humanas: Heitor

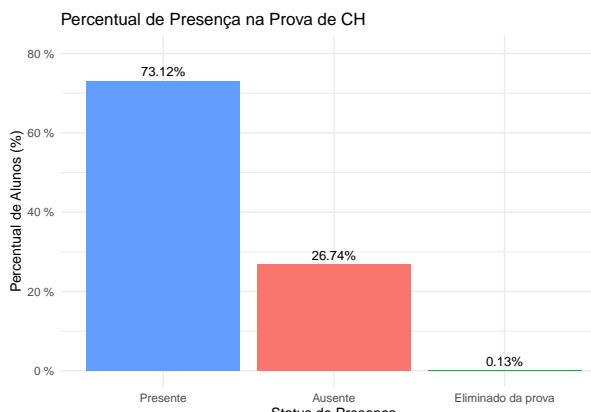
### Analise da variavel de presen a na prova de humanas

Foi feita uma analise da area de conhecimento de Ci ncias Humanas do Enem, utilizando a varivel: TP\_PRESENCA\_CH. Tal variavel,  classificada como qualitativa possuindo 3 poss veis valores:

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

### Gr fico de Frequ ncias das presen as no dia da prova

Busca indentificar se o aluno estava presente, ausente ou se foi eliminado da prova de ci ncias humanas, e, com isso, expressar a porcentagem e os valores absolutos da vari vel TP\_PRESENCA\_CH.



O gr fico “Percentual de Presen a na Prova de CH”, trata dos dados da vari vel ‘TP\_PRESENCA\_CH’. Uma vez que a vari vel  do tipo qualitativa, a abordagem mais convencional  um gr fico de barras dos percentuais.

Com base na analise do gr fico “Percentual de Presen a na Prova CH” foi poss vel determinar que no Exame Nacional do Ensino M dico (ENEM), edi o

de 2024, o n mero de alunos presentes foi de aproximadamente 2,73 vezes maior que o n mero de alunos ausentes. Além disso, percebe-se que a quantidade de alunos eliminados na prova de Ci ncias Humanas foi extremamente pequena - 0,1% - comparado com os percentuais da coluna “Presen a” e da coluna “Ausente”.

### Analise da variavel Notas da prova de Ci ncias Humanas

Tabela 1: Tabela Resumo: Estat sticas das Notas de CH por Regi o

Regi�o	Media	Mediana	Variânci�a	Desvio Padr�o	M�nimo	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

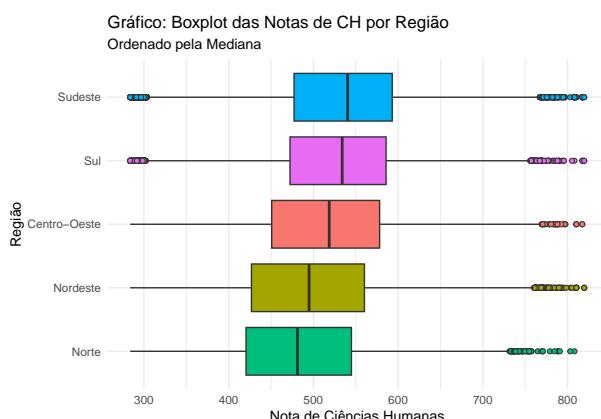
A “Tabela Resumo: Estat sticas das Notas de CH por Regi o” apresenta a distribui o das medidas: media, mediana, variânci a, desvio padr o, valor m nimo, valor m ximo e a frequ cia relativa percentual. Essa divis o foi feita por regi o do Brasil. Com base nisso, podemos identificar claramente o desempenho superior da regi o Sudeste.

Lideran a Clara: O Sudeste lidera em ambos os indicadores de performance, possuindo a maior M dia (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 regi es Sul (Mediana 597,5) e Centro-Oeste (Mediana 591,3). Estas tr s regi es est o claramente destacadas das regi es Nordeste (Mediana 562,9) e Norte (Mediana 555,0).

A Armadilha da M dia: Em todas as regi es, a M dia  “puxada” para baixo por notas mais fracas (assimetria  esquerda). Por isso, a Mediana  a m trica mais justa para a compara o, e nela o Sudeste tamb m vence.

## Gráfico Boxplot da variável notas



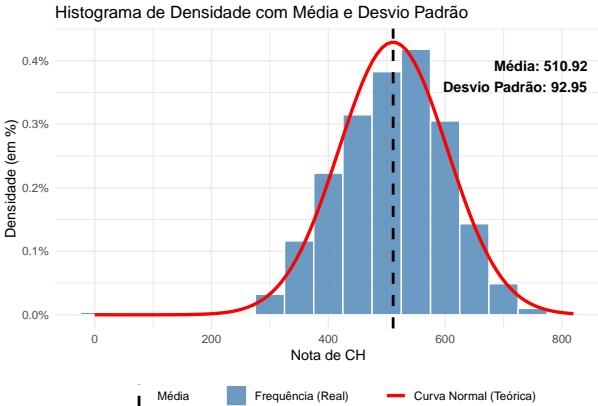
O gráfico "Boxplot das Notas de CH por Região", compara o desempenho central (a mediana) das notas de Ciências Humanas (CH) entre as cinco grandes regiões do Brasil. Além disso, o gráfico representa os valores “extremos”, os outliers, da variável ‘NU\_NOTA\_CH’, contribuindo para uma análise de desempenho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

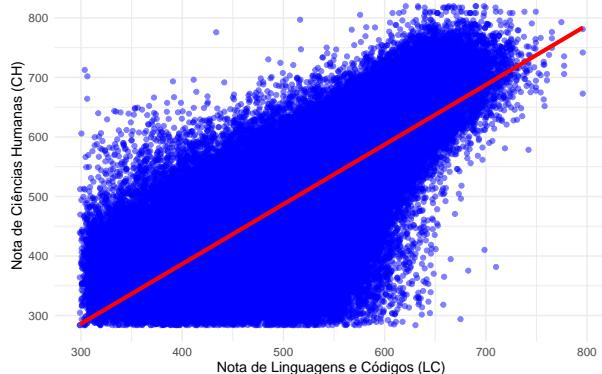
```
## O Coeficiente de Correlação (r) entre LC e CH é: 0.7615
```

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, confirmando a tendência de que notas altas em LC acompanham notas altas em CH. Para o modelo elaborado, temos a seguinte equação:  $\text{Nota}_{\text{CH}} = 93.30 + 0.82 * \text{Nota}_{\text{LC}}$ . 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.