



IFUSP

Instituto de Física da USP

Relatório iniciação científica

OS LIMITES E POSSIBILIDADES DE TESTES PADRONIZADAS EM LARGA ESCALA
PARA FINS DE PROCESSOS DE SELEÇÃO PARA O ENSINO SUPERIOR

Código do projeto: 2023-430

Período: 26/09/2023 a 25/09/2024

Orientador: Prof. Dr. Ewout ter Haar

Aluno: Luis Felipe Betinassi Martins de Souza

São Paulo

2024

SUMÁRIO

Introdução	3
Objetivo	4
Metodologia	5
Resultados	6
Bibliografia	7

I. Introdução

Neste trabalho foi feita uma análise do Exame Nacional do Ensino Médio (ENEM) utilizando os microdados do ENEM. Com os microdados foi possível analisar desde a parte socioeconômica dos candidatos do ENEM até os desvios padrões das médias do ENEM.

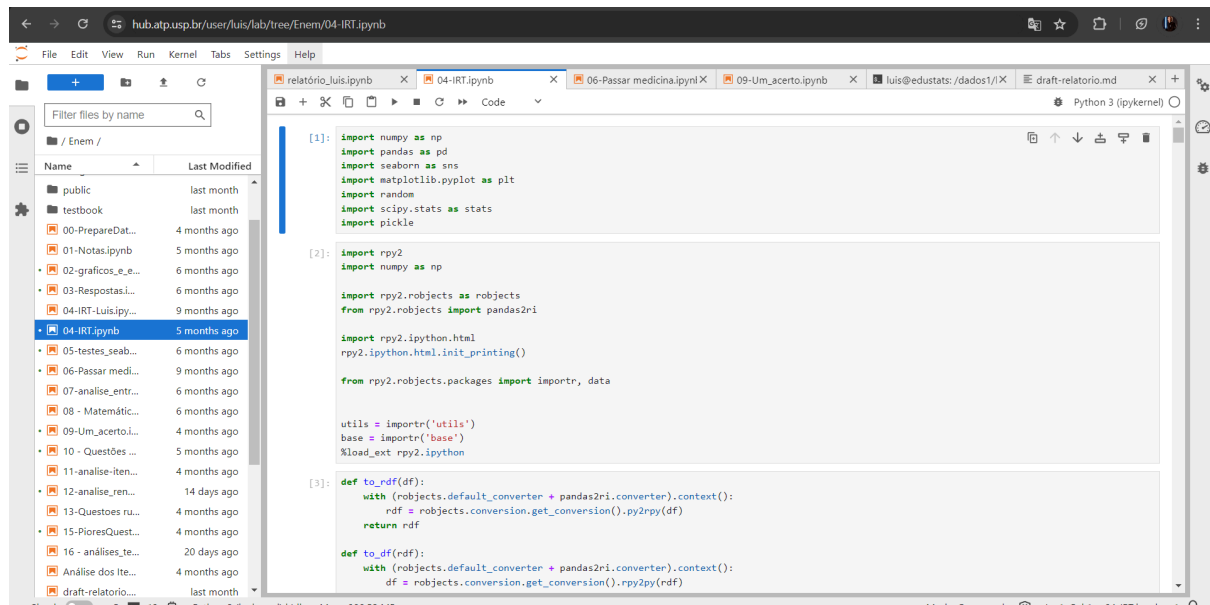
A Teoria de Resposta ao Item é um modelo estatístico da nota do ENEM, portanto ela acaba tendo incertezas, hoje em dia, quando um candidato realiza a prova do ENEM esta incerteza não é calculada pelo INEP. Neste trabalho calculamos a incerteza e mostramos como afetaria a nota dos alunos.

II. Objetivo

Neste projeto buscamos entender o funcionamento da Teoria de Resposta ao Item (TRI) do ENEM. Aliado a isso, fizemos também análises socioeconômicas dos candidatos. Focamos bastante em mostrar que, ao calcular as notas de candidatos no ENEM, há também uma incerteza nesta nota que, atualmente, não é levada em conta na interpretação e uso pela sociedade das notas.

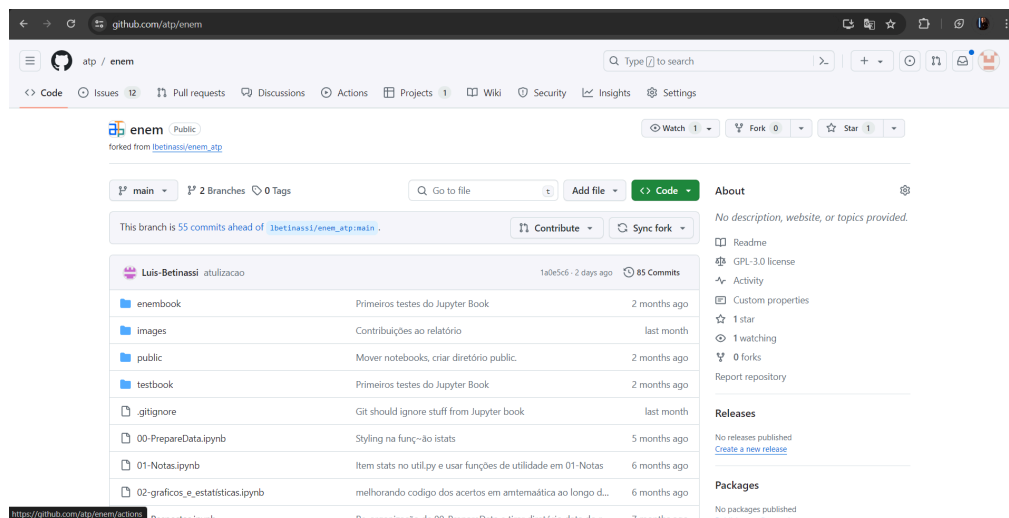
III. Metodologia

Para fazer as análises foram usados os microdados do ENEM. Com eles fizemos análises usando o ambiente virtual do Jupyter notebook, dentro do servidor da Universidade de São Paulo. No ambiente virtual do jupyter usamos a linguagem do python, principalmente as bibliotecas de análise de dados como pandas, matplotlib dentro outras.



Jupyter notebook sendo rodado dentro da máquina virtual.

Além disso, foi usado o princípio de ciência aberta, havendo transparência e divulgação dos resultados em todo processo, ou seja, qualquer um poderia verificar os trabalho que estávamos fazendo. Usamos a plataforma github, uma plataforma de hospedagem de repositórios de código, e que permite um melhor trabalho em grupo, além de permitir que qualquer um do público consiga ver nosso trabalho.



Página do github, disponível em: <https://github.com/atp/enem>

Para nosso trabalho utilizamos os seguintes filtros para selecionar os candidatos do ENEM. Selecionamos apenas uma amostra aleatória de 1% dos candidatos do ENEM, esta escolha foi feita, pois os programas demorariam muito mais tempo para rodar se utilizássemos a amostra inteira e mesmo assim ficamos um valor alto de amostras, aproximadamente 50 mil alunos, dependendo do ano. Ademais, selecionamos apenas os alunos que foram fazer todas as provas, não zeraram nenhuma delas e não eram “treineiros”.

```
def skip(i,fraction=0.01):
    if i == 0:
        return False # precisamos manter o row=0 para ter o header
    else:
        return random.random() > fraction

[skip(i) for i in range(10)]
```

Filtro para selecionar uma amostra aleatória de 1% dos dados.

```
def filterdf(df):
    df = df[df["TP_PRESENCA_CN"] == 1] # presente nas 4 provas
    df = df[df["TP_PRESENCA_CH"] == 1]
    df = df[df["TP_PRESENCA_LC"] == 1]
    df = df[df["TP_PRESENCA_MT"] == 1]
    if "IN_TREINEIRO" in df.columns:
        df = df[(df["IN_TREINEIRO"] == 0) | (df["IN_TREINEIRO"].isna())] # não é treineiro ou não existe
    df = df[df["TP_ST_CONCLUSAO"].isin([1, 2])] # afirma que concluiu ou vai concluir EM em 2019
    # não queremos as notas == 0
    df = df.query("NU_NOTA_CH != 0 and NU_NOTA_CN != 0 and NU_NOTA_LC != 0 and NU_NOTA_MT != 0")
    df.dropna(subset = ['TX_RESPOSTAS_CN', 'TX_RESPOSTAS_CH', 'TX_RESPOSTAS_LC', 'TX_RESPOSTAS_MT']) # achamos linhas com NaN para estas colunas!
    return df

def load_data(ano,frac):
    'carrega dados com determinado ano e fração'
    print(f"processando {ano} com fração = {frac*100:.0f}%",ano, frac)
    def skip(i,fraction=frac):
        if i == 0:
            return False # precisamos manter o row=0 para ter o header
        else:
            return random.random() > fraction

    if ano == 2016:
        caminho = f'../inep/enem/{ano}/DADOS/microdados_enem_{ano}.csv'
    else:
        caminho = f'../inep/enem/{ano}/DADOS/MICRODADOS_ENEM_{ano}.csv'
    #print(caminho)
    df = pd.read_csv(caminho,skiprows = skip, encoding='latin1',sep=";")
    return df

def filter_data(df):
    print("filter data...")
    return filterdf(df)

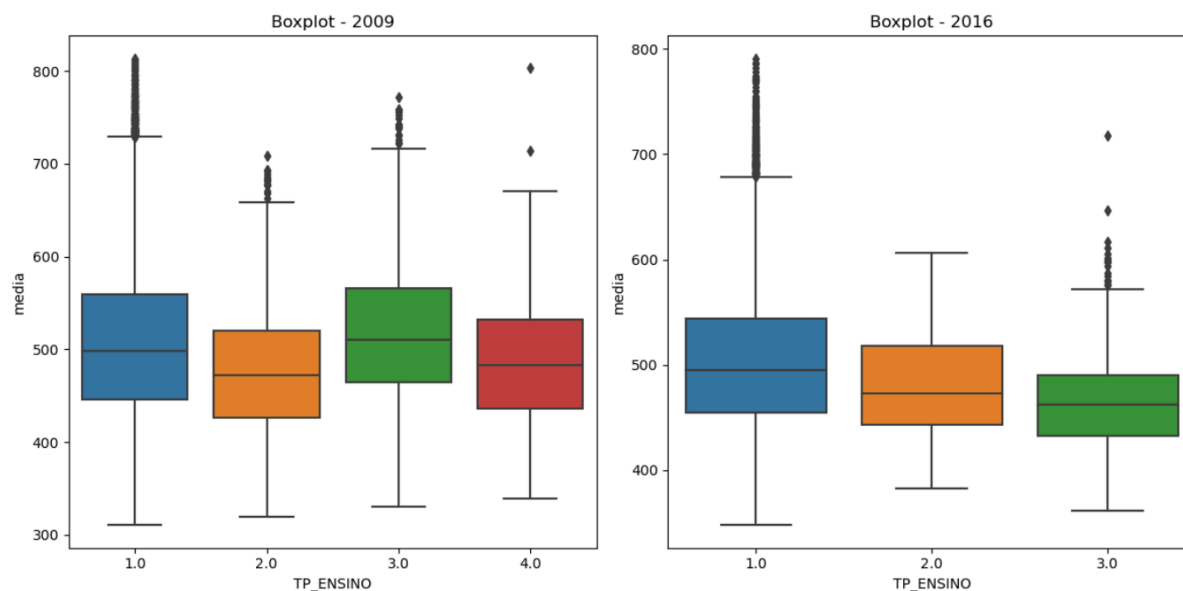
def nome_do_arquivo(ano,frac):
    frac = frac*100
    fn = f'data/enem_{frac:.0f}_{ano}.csv'
    return fn

df = load_data(2014,0.01)
print(len(df))
df = filter_data(df)
print(len(df))
```

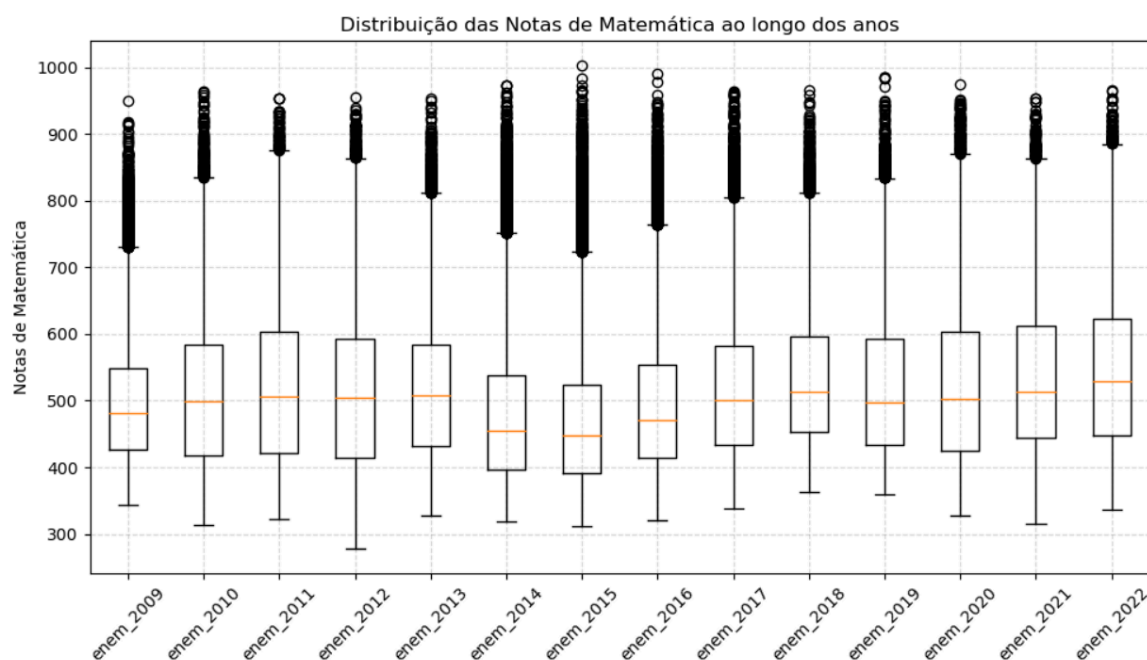
Filtro para selecionar apenas alunos que fizeram todas as provas, não eram treineiros e não zeraram nenhuma delas.

IV. Resultados

Após o pré-processamento dos microdados foi possível fazer várias análises dos candidatos do ENEM. Veja alguns exemplos abaixo.



Comparação da média por tipo de ensino em 2009 e 2016

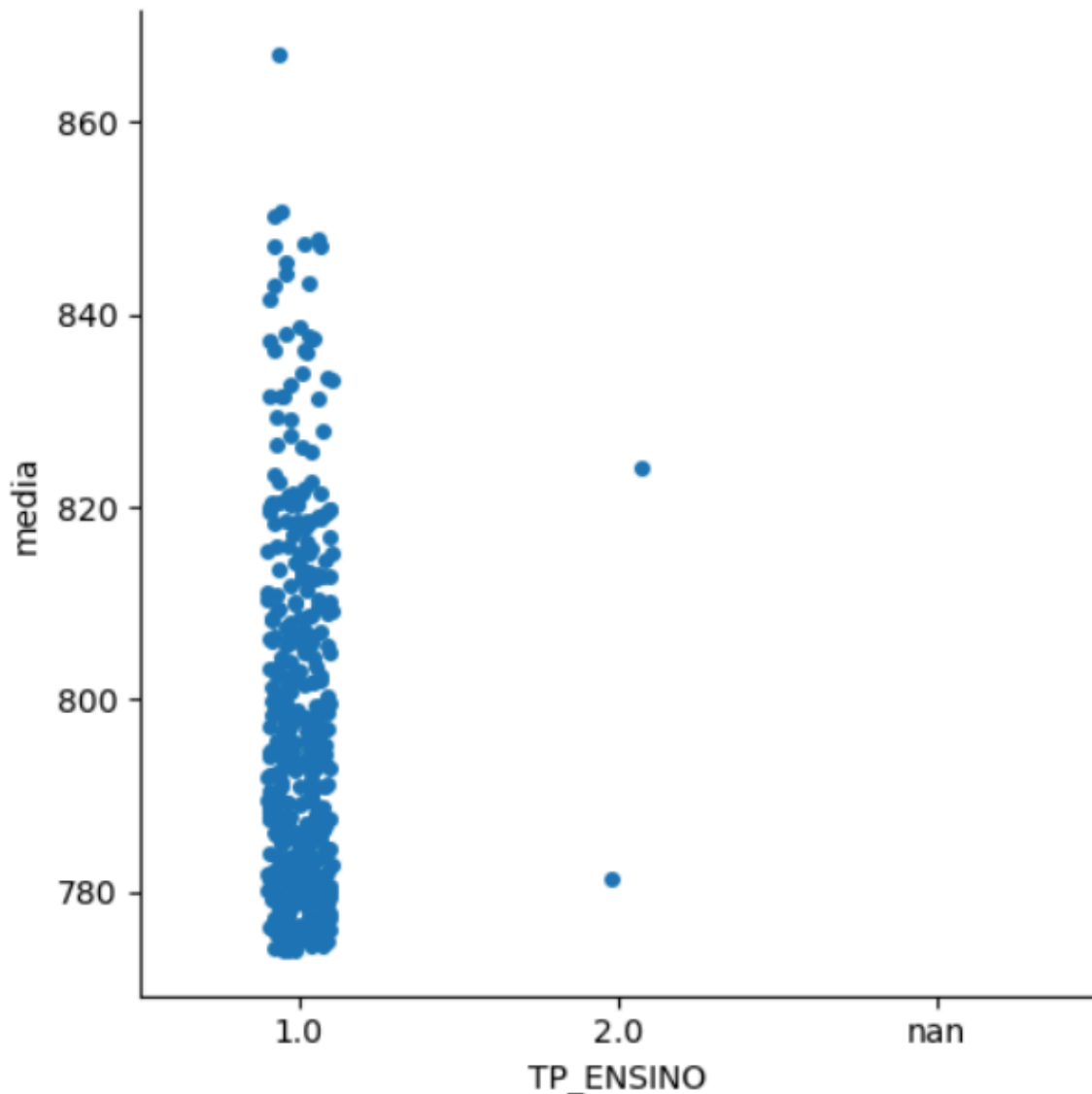


Nota de matemática ao longo dos anos

Para demonstrar um determinado aspecto das desigualdades de acesso ao ensino superior no Brasil, foi feita uma análise dos alunos de escola pública que

teriam nota o suficiente para passar no curso de medicina, o intuito desta análise é reforçar a importância da política de cotas.

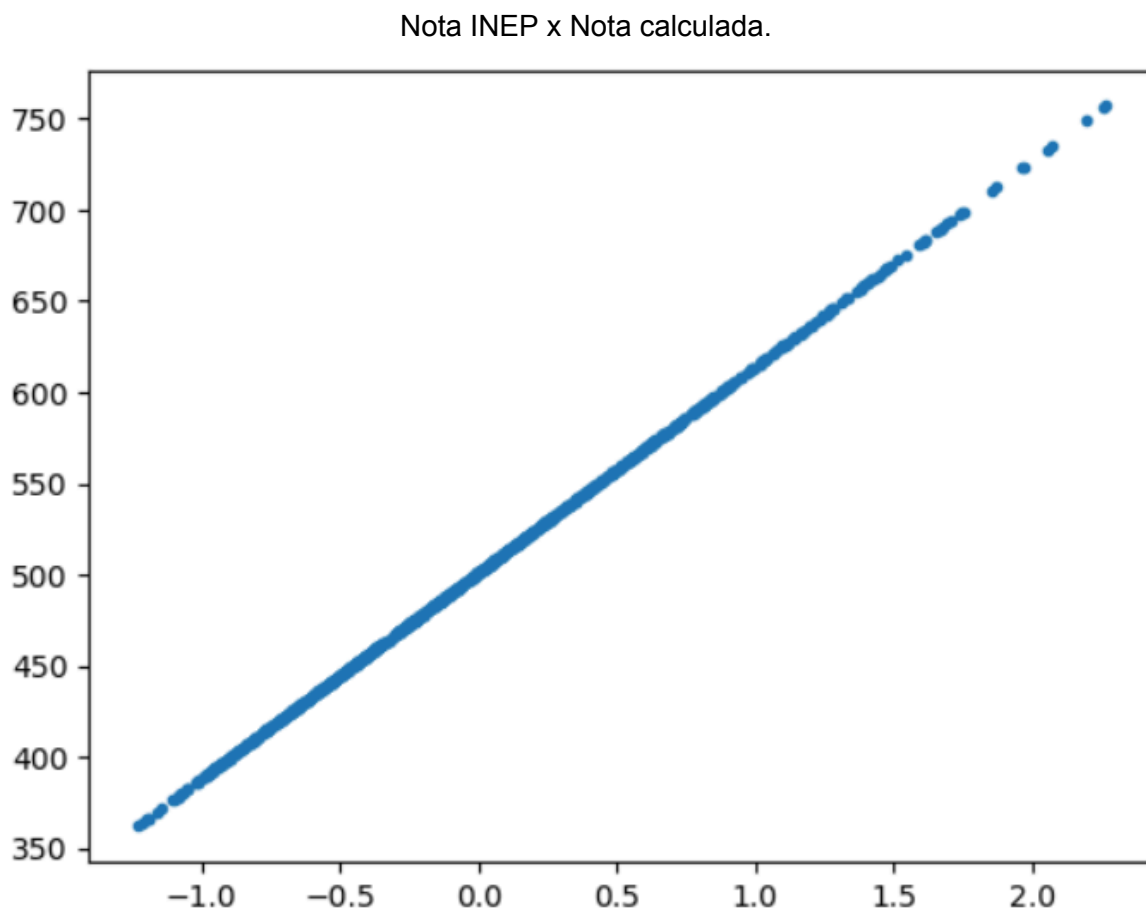
Nesta análise encontrei uma das menores notas de corte em uma faculdade pública no curso de medicina pelo ENEM em 2022, no caso foi na UFERSA. Selecionei alunos que tiraram apenas notas superiores a 772,83 com os pesos 3 para redação, 7 para ciências da natureza, 4 para ciências humanas, 4 para linguagens e códigos e 2 para matemática. De nossa amostra de 1% dos candidatos que prestaram o ENEM obtive 440 alunos da rede particular e apenas 2 alunos da rede pública, uma discrepância muito alta.



Média dos alunos que passariam em medicina no ano de 2022 na UFERSA por tipo de ensino. 1.0 = rede particular, 2.0 = rede pública.

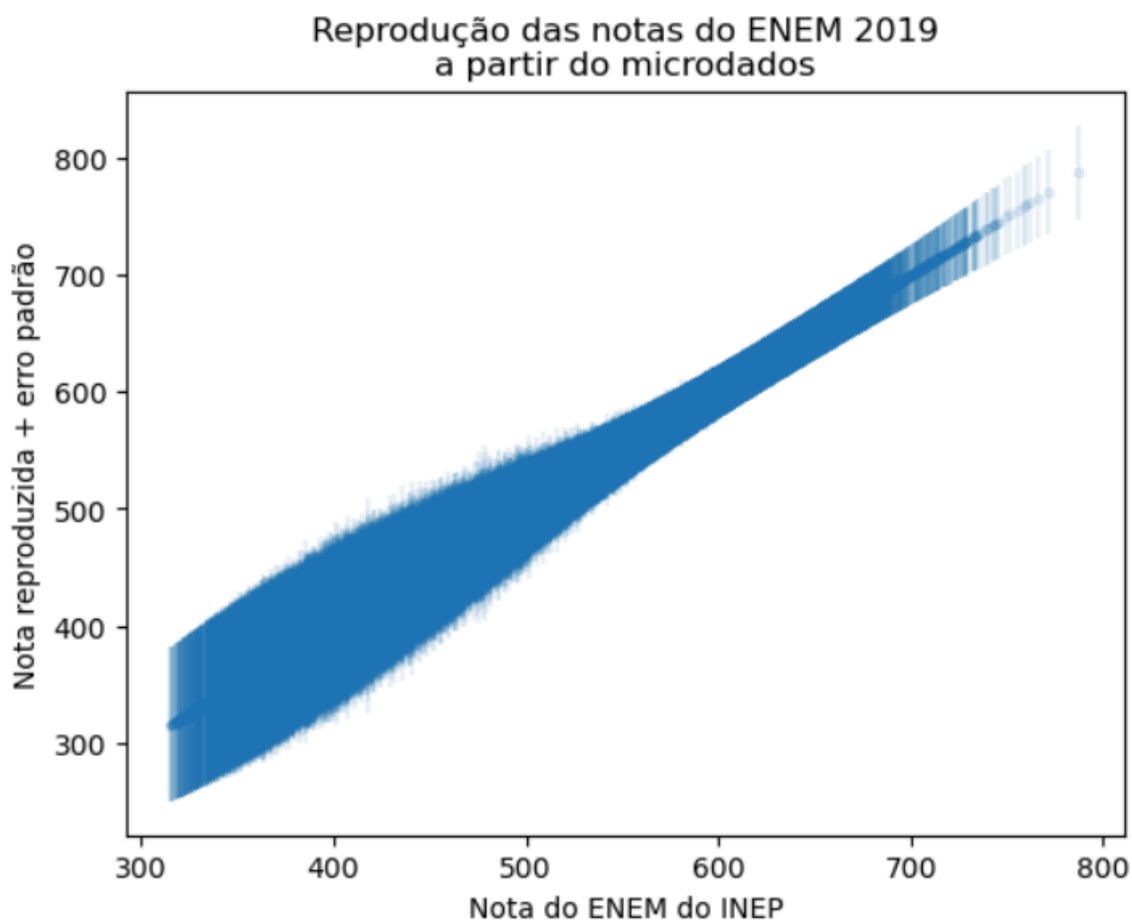
Um dos objetivos do projeto é fazer uma análise das incertezas nas notas do ENEM. Para ter confiança na nossa metodologia, reproduzimos as notas do ENEM

calculado pelo INEP, a partir dos padrões de resposta nos microdados. Até o ano de 2022, não era possível reproduzir as contas do INEP, porque para calcular uma nota usando o TRI é preciso conhecer os parâmetros de cada item (termo adequado para “questão”) e estes parâmetros não eram divulgados pelo INEP anterior ao ano de 2022. Usamos estes parâmetros para calcular as notas do ENEM e ter certeza de que nossa metodologia consegue reproduzir exatamente as notas calculados pelo INEP.



Como podemos observar a nota calculada e a nota obtida foram iguais.

Agora, usando a metodologia de cálculo das notas ENEM foi possível obter o erro padrão nestas notas . Este erro padrão é maior nas notas mais perto da média e menor para as maiores notas, veja na figura abaixo.



Este gráfico mostra que um aluno na faixa dos 400 pontos pode ter uma média de 400 ± 80 . Ou seja, falar que a nota deste aluno é 400 sem apontar a incerteza, oculta uma parte importante do TRI.

Por último, iniciamos uma forma de divulgação de nosso trabalho. Para tanto usamos uma extensão do próprio github, o github pages, que cria uma página relacionada aos notebooks de um repositório e o Voilà,

A questão ENEM

Q Search

Ctrl + K

Apresentação

Testes padronizados X

Teórica Clássica e TRI

Terio de resposta ao Item (TRI)

O ENEM 2009-2022

Desempenho

Características das questões

Análise dos Itens do ENEM 2009-2022

Incerteza

Incerteza da nota ENEM dado pela análise TRI

Contrafactuais

Consequências

Relatórios

RELATÓRIO PARCIAL DE ATIVIDADES DO ALUNO DE

Apresentação

Em 2022, um jovem participou do ENEM e a nota média das quatro provas era 750, permitindo acesso ao curso ENGENHARIA DE CONTROLE E AUTOMAÇÃO da UFRJ com nota de corte 749,45 na modalidade ampla concorrência.[^refsisu] Um outro candidato nesse mesmo ano tirou 720. Como devemos interpretar estes resultados? A principal pergunta

XXX

Hint

O ENEM é o Exame Nacional do Ensino Médio. Desde 2009 o ENEM é uma prova de seleção, com quatro partes objetivos com 45 questões e uma dissertação.

[^refsisu] O SISU é o sistema e processo organizado pelo MEC que organiza as inscrições e distribuição das vagas para cursos de ensino superior no Brasil. Veja [aqui](#) para a relação das notas de corte e distribuição de vagas.

Resumo do projeto

XXX

Principais resultados

Contents

Resumo do projeto

Principais resultados

Página do github pages, disponível em: <https://atp.github.io/enem/>

voila.atp.usp.br/voila/render/enem/Análise%20dos%20Itens%20do%20ENEM.ipynb?

inglês

português

Google Translate

Análise dos item

Aqui analisamos e apresentamos os parâmetros dos itens do ENEM, como foram divulgados pelo INEP (a partir de 2022) (veja ... ref)

Ano: 2022

Area: CN

Sort by: total.r

☒ return_styled

Item	N	mean	sd	total.r	url	pos	area	gab	habilidade	aban	motivo	a	b	c	cor	prova	ano	ling	adaptado	digital	
52	64203	10000	0.17	0.37	0.023	link	133	CN	A	5	1	bisblg < -0.15	nan	nan	nan	AMARELA	1086	2022	nan	0.000000	nan
54	75845	10000	0.26	0.44	0.054	link	95	CN	D	1	1	BLG não rodou	nan	nan	nan	AMARELA	1086	2022	nan	0.000000	nan
74	89554	10000	0.18	0.39	0.064	link	126	CN	D	7	1	BLG não rodou	nan	nan	nan	AMARELA	1086	2022	nan	0.000000	nan
24	141605	10000	0.36	0.48	0.078	link	127	CN	D	19	1	Bilog não converge	nan	nan	nan	AMARELA	1086	2022	nan	0.000000	nan
32	141731	10000	0.13	0.33	0.094	link	107	CN	E	29	0	nan	1.525120	3.192040	0.112830	AMARELA	1086	2022	nan	0.000000	nan
86	97915	10000	0.20	0.40	0.095	link	93	CN	C	24	0	nan	2.398960	2.765880	0.189700	AMARELA	1086	2022	nan	0.000000	nan
8	111694	10000	0.24	0.42	0.095	link	125	CN	A	27	0	nan	2.078840	2.234080	0.171910	AMARELA	1086	2022	nan	0.000000	nan
44	37769	10000	0.28	0.45	0.133	link	135	CN	B	14	0	nan	1.784290	2.500620	0.270710	AMARELA	1086	2022	nan	0.000000	nan
68	85860	10000	0.17	0.38	0.145	link	109	CN	C	24	0	nan	5.316210	2.042880	0.170880	AMARELA	1086	2022	nan	0.000000	nan
64	83901	10000	0.30	0.46	0.156	link	97	CN	B	8	0	nan	0.405830	2.516760	0.009180	AMARELA	1086	2022	nan	0.000000	nan
50	44969	10000	0.27	0.44	0.170	link	110	CN	B	25	0	nan	3.071660	1.891270	0.252720	AMARELA	1086	2022	nan	0.000000	nan
56	78377	10000	0.29	0.46	0.173	link	111	CN	B	6	0	nan	0.602390	2.401370	0.120170	AMARELA	1086	2022	nan	0.000000	nan
6	111668	10000	0.23	0.42	0.178	link	124	CN	B	23	0	nan	1.989640	1.877630	0.194150	AMARELA	1086	2022	nan	0.000000	nan

Página Voilá, disponível em:

<https://voila.atp.usp.br/voila/render/enem/An%C3%A1lise%20dos%20Itens%20do%20ENEM.ipynb?>

V. Conclusão

Para muito o ENEM é uma incógnita, há uma falta de divulgação por parte do INEP de como ele funciona. Torna-se importante uma forte divulgação de como funciona o ENEM, hoje a maior forma de entrada na universidade do país. Esta divulgação deve ser clara, para que o brasileiro possa saber os critérios com que seu futuro será decidido. Ademais, devemos mostrar o ENEM como um modelo estatístico e não como um “algoritmo anti chute” como é mostrado hoje.

Sabemos que hoje não se leva em conta o erro padrão da nota ENEM nos processos decisórios em que é usado. É, portanto, necessário investigar como esta informação a mais poderia ser usado pela sociedade. O erro padrão que o TRI fornece é um limite inferior à incerteza total na nota ENEM. Pretendemos na segunda metade do projeto investigar outras fontes de incerteza. Críticas feitas ao uso no TRI geralmente se limitam a afirmar que não é claro que ele serve para estimar um “quantidade” de conhecimento, porém e que seria impossível medir o conhecimento numa escala unidimensional. Porém, mostraremos com este projeto que esta crítica, embora válido, deveria ser complementado com outra: que mesmo admitindo as premissas do TRI, é preciso interpretar e usar as notas geradas por esta metodologia levando em conta as inevitáveis incertezas nestas notas. Devemos fazer um trabalho de divulgação científica, explicando que há um erro padrão na nota calculada pela TRI, que este erro pode ser grande e que esta informação pode ter consequências para processos decisórios que usam as notas do ENEM.

VI. Bibliografia

[1] - <https://www.gov.br/inep/pt-br/aceso-a-informacao/dados-abertos/microdados>

[2] - <https://www.youtube.com/watch?v=kWZAvuugizw>

[3] -