

**Como a composição de gênero da coorte de
ensino médio influencia a escolha de
graduação dos estudantes?**

Uma abordagem de ciência de dados

Dayanne Cristina Pereira Gomes

RELATÓRIO APRESENTADO AO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DA UNIVERSIDADE DE SÃO PAULO
PARA EXAME DE QUALIFICAÇÃO DE
MESTRADO EM CIÊNCIAS

Programa: Ciência da Computação
Orientador: Prof. Dr. Fabio Kon

São Paulo
Dezembro de 2023

**Como a composição de gênero da coorte de
ensino médio influencia a escolha de
graduação dos estudantes?**

Uma abordagem de ciência de dados

Dayanne Cristina Pereira Gomes

Esta é a versão original do texto de
qualificação elaborado pela candidata
Dayanne Cristina Pereira Gomes, tal
como submetido à Comissão Julgadora.

*O conteúdo deste trabalho é publicado sob a licença CC BY 4.0
(Creative Commons Attribution 4.0 International License)*

Do. Or do not. There is no try.

— Mestre Yoda

[illegible]

Resumo

Dayanne Cristina Pereira Gomes. **Como a composição de gênero da coorte de ensino médio influencia a escolha de graduação dos estudantes?: Uma abordagem de ciência de dados.** Exame de Qualificação (Mestrado). Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2023.

[illegible]

Palavras-chave: Palavra-chave1. Palavra-chave2. Palavra-chave3.

Abstract

Dayanne Cristina Pereira Gomes. **How does the gender composition of a high school cohort influences students' major choice?: A data science approach.** Qualifying Exam (Master's). Institute of Mathematics and Statistics, University of São Paulo, São Paulo, 2023.

[illegible]

Keywords: Keyword1. Keyword2. Keyword3.

Lista de Abreviaturas

SISU	Sistema de Seleção Unificada (<i>Continuous Fourier Transform</i>)
IBGE	Instituto Brasileiro de Geografia e Estatística (<i>Discrete Fourier Transform</i>)
UNICAMP	Universidade Estadual de Campinas (<i>Electron-Ion Interaction Potentials</i>)
STFT	Transformada de Fourier de tempo reduzido (<i>Short-Time Fourier Transform</i>)
ABNT	Associação Brasileira de Normas Técnicas
URL	Localizador Uniforme de Recursos (<i>Uniform Resource Locator</i>)
IME	Instituto de Matemática e Estatística
USP	Universidade de São Paulo

Lista de Símbolos

ω	Frequência angular
ψ	Função de análise <i>wavelet</i>
Ψ	Transformada de Fourier de ψ

Lista de Figuras

2.1	Tipos de estudos observacionais	3
5.1	Distribuição das notas do ENEM por gênero	27
5.2	Distribuição das notas do SISU por situação de aprovação	28
5.3	Distribuição das unidades federativas das instituições de ensino	29
5.4	Distribuição das unidades federativas de residência dos alunos por situação de aprovação	29

Lista de Tabelas

3.1	Estatísticas descritivas das variáveis dependentes por gênero. A coluna Diferença reporta o coeficiente do teste-t. P-valor <0,01. Desvio padrão em parênteses. Adaptado de SILVA BORGES (2021)	11
4.1	Tabulação cruzada das variáveis TP_ST_CONCLUSAO e Q046	17
4.2	Tabulação da variável TP_ANO_CONCLUIU	18
4.3	Tabulação da variável TP_ENSINO	18
4.4	Tabulação dos filtros aplicados no conjunto de dados do ENEM	19
4.5	Tabulação de inscrições e inscritos no SISU, divididos por listagem	19
4.6	Tabulação de quantidade de cursos optados no SISU	19
4.7	Tabulação da quantidade de inscritos por caso de inscrição em listagem no SISU	20
4.8	Tabulação da quantidade de inscritos pelo total de inscrições no SISU	20

5.1	Estatísticas descritivas dos participantes do ENEM por gênero	24
5.2	Estatísticas descritivas de cursos e participantes do SISU por situação de aprovação	24
5.3	Observações de notas dos participantes do ENEM por gênero	25
5.4	Observações de notas dos participantes do SISU por situação de aprovação	25
5.5	Estatísticas descritivas das notas e idade do ENEM por gênero	25
5.6	Estatísticas descritivas das notas e idade do SISU por situação de aprovação	26
5.7	Distribuição da classificação dos cursos do SISU por gênero	26
5.8	Top 10 universidades com mais inscrições	26

Lista de Programas

Sumário

1	Introdução	1
1.0.1	Motivação e Objetivos	2
2	Fundamental teórico	3
2.1	Estudo de coorte	3
2.2	Estratégias empíricas: Variações idiossincráticas na composição de gênero	5
2.2.1	Efeitos de gênero em escolhas de graduação	6
3	Trabalhos relacionados	9
4	Metodologia	13
4.1	Bases de dados	13
4.1.1	ENEM	13
4.1.2	SISU	14
4.1.3	Censo Escolar	15
4.2	Pré-processamento	15
4.3	Análise exploratória de dados	17
5	Resultados preliminares	23
5.1	Análise descritiva	23
5.1.1	Perfil socioeconômico	23
6	Plano de trabalho	31

Apêndices

Anexos

Referências 33

Índice Remissivo 37

Capítulo 1

Introdução

O momento da escolha de carreira profissional é um acontecimento de grande importância na vida de um indivíduo. Esse processo é determinante na definição de características individuais e coletivas de pessoas que se veem na responsabilidade de tomar uma decisão impactante (AKOSAH-TWUMASI *et al.*, 2018) cujas consequências podem influenciar uma vida inteira. As implicações de optar por um determinado caminho, tanto no âmbito acadêmico como profissional, vão além da preferência por uma área do conhecimento. Elas podem influenciar de forma significativa na qualidade de vida futura, já que sua ocupação tem poder de influenciar personalidade, nível de renda, status social e grupos sociais nos quais os sujeitos se caracterizarão (SHAHID KAZI e AKHLAQ, 2017).

As mudanças ocasionadas pelas transformações sociais fizeram com que se percebesse a necessidade de entender como a juventude é afetada no mundo contemporâneo (UNESCO, 2006). A responsabilidade de tomar decisões significativas não é alheia a esses jovens. Segundo GATI e SAKA (2001), adolescentes estão envolvidos diretamente no processo de escolhas complexas. Preocupações com questões educacionais, como estudos e carreira, mostram um nível de entendimento dos riscos e implicações envolvidos, além das dificuldades que podem surgir.

Esses temas podem ser observados de forma muito expressiva em estudantes do ensino médio. Para aqueles que estão vivendo o término da adolescência e início da vida adulta, a transição envolve não só o amadurecimento etário, mas também a passagem da educação secundarista para a superior. Para uma grande maioria desses estudantes, há a aspiração de adentrar em algum tipo de curso de graduação (VENEZIA e JAEGER, 2013). A escolha de frequentar a universidade pode começar cedo durante a jornada educacional, e nesse caminho até adentrar o curso desejado existem diversas etapas (CABRERA e NASA, 2000). Em todo o mundo, exames educacionais competitivos são aplicados para medir o desempenho de alunos secundaristas. Os resultados desses exames podem ser utilizados para ranquear os melhores candidatos e convocá-los para admissão em uma instituição. Alguns exemplos são ACT e SAT, nos Estados Unidos, Baccalauréat, na França e GCSE, no Reino Unido.

No Brasil, tais exames são conhecidos como vestibular. Os participantes costumam realizá-lo no último ano do ensino médio, quando estão encerrando seus estudos, ou até que consigam a classificação para a vaga pretendida. As instituições têm autonomia para

adotar uma prova específica, como o vestibular tradicional, cuja aplicação é realizada pela própria instituição. Outra opção é a adoção do **ENEM** como forma de entrada. O Exame Nacional do Ensino Médio é uma prova anual com abrangência em todo território brasileiro. A nota desse exame, além de avaliar o desempenho de estudantes secundaristas, pode ser utilizada para adentrar algum curso superior através de programas governamentais, como SISU, ProUni e FIES.

Dadas diversas formas de ingresso, ainda é parte da escolha do estudante qual curso de graduação ele deseja seguir. Múltiplos fatores devem ser considerados nesse processo. Para **BORCHERT (2001)**, existem três grandes áreas que afetam a escolha de carreira: oportunidade, personalidade e ambiente. Segundo **ABBAGNANO (2012)**, ambiente pode ser definido como um complexo conjunto de relações entre mundo natural e ser vivo, que influem na vida e no comportamento do mesmo. O gênero tem um grande papel nas questões relacionadas ao ambiente no qual os jovens se encontram. No cenário brasileiro, as questões de gênero são relevantes e consideradas no processo de levantamento de indicadores sociais. Um exemplo é o estudo **Estatísticas de gênero: indicadores sociais das mulheres no Brasil**, do Instituto Brasileiro de Geografia e Estatística (IBGE). Nele, são levantados diferentes aspectos da vida da população, incluindo desigualdades no mercado de trabalho e na educação.

1.0.1 Motivação e Objetivos

Neste contexto, queremos descrever o perfil dos estudantes, em especial de mulheres, e entender como um fator específico, o gênero, impacta na decisão de carreira. Para o processo de análise das escolhas de um curso de graduação, em particular cursos com histórica baixa representatividade feminina, bem como a composição de turmas do ensino médio, levantamos as seguintes questões de pesquisa:

- **Q1** Há diferença no desempenho de homens e mulheres nas disciplinas do ENEM?
- **Q2** Homens e mulheres que realizam o processo seletivo do SISU fazem escolhas distintas? Em quais cursos estão mais concentrados? Há uma diferença na proporção daqueles que selecionam cursos STEM?
- **Q3** Quais são as características das escolas de ensino médio nas quais os estudantes se formaram?
- **Q4** A composição de gênero da coorte do ensino médio influencia a escolha de graduação das estudantes? Mulheres em escolas com mais colegas do sexo feminino são mais ou menos propensas a selecionarem cursos STEM? Há um efeito sobre a probabilidade das mulheres escolherem o curso de ciência da computação?

Para alcançar os objetivos dessa pesquisa, realizaremos um estudo de coortes levando em consideração os efeitos da composição de gênero. Serão aplicadas técnicas de ciência de dados para alcançar os objetivos deste trabalho. A ciência de dados é uma área multidisciplinar que envolve técnicas estatísticas, matemáticas e computacionais para resolver problemas de um determinado domínio. Os dados utilizados nesse processo passam uma série de etapas, como tratamento, análise e visualização, a fim de dar suporte à tomada de decisões estratégicas. Trabalharemos com dados educacionais a nível nacional de múltiplas fontes, incluindo ENEM, SISU e Censo Escolar.

Capítulo 2

Fundamental teórico

2.1 Estudo de coorte

A palavra coorte tem origem no latim *cohors*. Esse termo era utilizado para nomear uma unidade militar do Império Romano, que compunha uma legião. Pode ser considerada equivalente ao conceito moderno de batalhão. No contexto científico, uma coorte pode ser definida como um grupo de pessoas que possuem uma característica ou experiência em comum. Coortes podem ser estabelecidas com vários propósitos. Um deles é a análise de grupos dentro de um determinado domínio, como em estudos econométricos, epidemiológicos e demográficos. Dentro das metodologias utilizadas para explorar uma questão específica estão os estudos observacionais. O que diferencia os estudos de caráter observacional de outros estudos é a realização de intervenções. Nesses estudos, os pesquisadores não interferem nos fenômenos estudados, apenas os observam, fazendo com que a variável considerada não esteja sob controle (SONG e CHUNG, 2010).

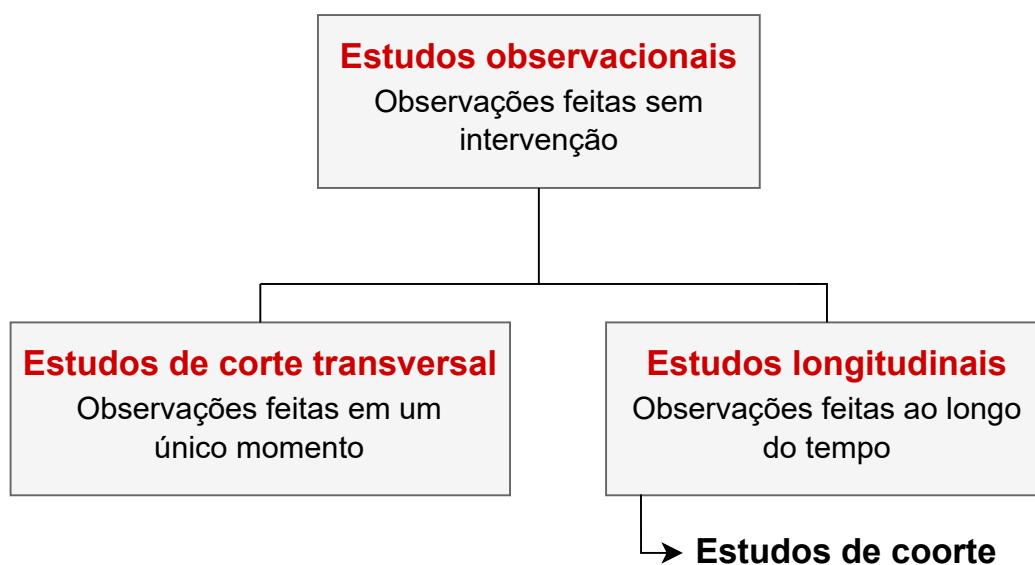


Figura 2.1: Tipos de estudos observacionais

Estudos observacionais podem ser divididos de acordo com o período de coleta de dados. Quando as observações são feitas em um momento específico, coletando dados em um curto intervalo de tempo, chamamos de estudo de corte transversal. Eles são úteis para observar o estado e as condições atuais dos participantes, sem que haja um acompanhamento dos mesmos (ZANGIROLAMI-RAIMUNDO *et al.*, 2018). Alguns exemplos de estudos que seguem essa estrutura no Brasil são o **Censo Demográfico**, realizado a cada 10 anos pelo IBGE, e o **Censo Escolar**, realizado anualmente pelo INEP.

Já quando os indivíduos são acompanhados por longos períodos de tempo, os estudos são classificados como longitudinais. São realizadas coletas de dados contínuas ou repetidas em intervalos regulares, como dias, meses e anos. Como os dados levam em consideração um grupo pré-definido, é possível ajustar os métodos estatísticos para mudanças observadas ao longo do tempo para o grupo como um todo ou para sujeitos específicos (CARUANA *et al.*, 2015). Temos como exemplos o **índice Bovespa**, disponibilizado pela B3 e **taxa Selic**, disponibilizada pelo Banco Central.

Os estudos de coorte são um tipo de estudo longitudinal que leva em consideração uma segmentação específica da população, que é a coorte. Um exemplo de estudo de coorte é a observação do desenvolvimento de doenças em uma população. A amostra populacional pode ser dividida em duas coortes: a coorte 1, de expostos à doença e a coorte 2, de não expostos. Assim, é possível comparar a ocorrência da doença entre os grupos. Estudos de coorte podem ser conduzidos de duas maneiras principais:

- Nos estudos prospectivos, os indivíduos são acompanhados do início do estudo, no presente, durante um período de tempo, para analisar os dados recolhidos no futuro;
- Nos estudos retrospectivos, os dados são coletados e mensurados no passado, acompanhados por um determinado período de tempo, para a realização de uma análise no presente.

Na área da educação, estudos de coorte podem ser utilizados para analisar como questões específicas impactam sujeitos inseridos no sistema educacional, como professores, estudantes e outros envolvidos. BJÖRKENSTAM *et al.* (2011) investigou a associação da performance escolar com taxas de suicídio em estágios posteriores da vida, utilizando uma coorte de nascidos entre 1972 e 1981 na Suécia. Outro exemplo foi o trabalho de ENSMINGER e SLUSARCICK (1992), que examinou os caminhos de desenvolvimento de uma coorte de estudantes negros de uma escola de Chicago entre 1966 e 1977. No sentido dessa pesquisa, pode-se delimitar uma coorte como um grupo de estudantes que estão inseridos no mesmo ambiente escolar. Eles interagem entre si, podem desenvolver relacionamentos e compartilhar experiências. A realização de estudos de coorte educacionais, em particular os retrospectivos, como é o atual, é facilitada pela disponibilização de dados públicos de diferentes abrangências. No Brasil, há a iniciativa do **Portal Brasileiro de Dados Abertos**, com dados a nível federal e local de múltiplas áreas.

2.2 Estratégias empíricas: Variações idiossincráticas na composição de gênero

Variações idiossincráticas podem ser definidas como particularidades e diferenças individuais únicas que podem influenciar resultados ou comportamentos. Esses fatores comuns ou gerais que afetam um grupo de indivíduos também influenciam os sistemas nos quais estes estão inseridos (MEISTER, 1991). Essas variações podem ser diversas, abrangendo uma ampla gama de características, experiências e situações pessoais. Alguns exemplos de variações são:

- **Contexto pessoal:** Dinâmica familiar, status socioeconômico, crenças, *background* cultural, valores familiares;
- **Carreira:** Aspirações e objetivos profissionais, experiência prévia de trabalho, habilidades técnicas;
- **Educação:** Qualidade escolar, atividades extracurriculares, exposição prematura a tópicos avançados.

Por serem altamente individualizadas e frequentemente envolverem múltiplos fatores, é um desafio conduzir uma pesquisa que isole e entenda o impacto de fatores específicos, como o efeito de pares. O efeito de pares (do inglês *peer effect*) se refere à influência e impactos que indivíduos dentro do círculo próximo (familiar, social ou educacional) têm nas suas decisões e comportamentos. Um trabalho muito relevante no entendimento dos efeitos sociais do grupo que uma pessoa se insere é o de MANSKI (1993). Ele aponta a existência de efeitos de pares correlatos, que são comportamentos similares em pessoas do mesmo grupo por conta da semelhança de características individuais ou ambientes institucionais.

Isso é especialmente importante no contexto educacional, já que estudantes do ensino médio compartilham o ambiente escolar durante sua formação. Inseridos em classes diversificadas, eles devem conviver diariamente com outros adolescentes por uma parte relevante de suas vidas. Os colegas de escola podem ser uma importante força social não só na performance acadêmica, mas também nas aspirações profissionais e decisões de seguir uma área específica (TANG *et al.*, 2008).

A literatura observa que os pares têm grande relevância em comportamentos, escolhas e resultados educacionais (SACERDOTE, 2014; ZIMMERMAN, 2003). Para entender melhor esse efeito, é importante considerar como estão estruturados os grupos de referências de estudantes, já que eles são altamente vulneráveis à influência uns dos outros pela exposição contínua e proximidade. Os efeitos dos pares podem se apresentar tanto de maneira positiva, quanto negativa, que se refletem em pontuações de prova, motivação e hábitos de estudo, por exemplo. Uma maneira de fazer isso é através da observação da composição das turmas, isto é, como diferentes distribuições dos pares nas classes de aula afetam os alunos.

Um importante precursor dessas observações é o estudo de HOXBY (2000). Ela identificou e mensurou a existência de efeitos dos pares em coortes escolares que diferem na composição de fatores específicos, como o gênero. Seus resultados sugerem que um grupo

de pares com maioria feminina eleva as pontuações de meninos e meninas em matemática e leitura.

Nesse sentido, múltiplos trabalhos exploram o problema com uma metodologia similar, denominada de variações idiossincráticas na composição de gênero (SCHØNE *et al.*, 2019; LAVY e SCHLOSSER, 2011; SCHNEEWEIS e ZWEIMÜLLER, 2012; BRENØE e ZÖLITZ, 2020; SILVA BORGES, 2021). Eles utilizam estratégias empíricas para levantar hipóteses que ajudem a estabelecer um modelo econométrico para investigar os efeitos de gênero entre os pares.

Os modelos econométricos, apesar de semelhantes, diferem-se por conta de serem estimados utilizando conjuntos de dados educacionais de diferentes países (Noruega, Israel, Áustria, Dinamarca e Brasil), que se configuram em sistemas de ensino distintos, bem como se apoiam em outros recursos, como registros demográficos. Além disso, são consideradas diferentes etapas da educação básica, como anos iniciais e finais do fundamental e médio. Uma lista não exaustiva de variáveis analisadas inclui parcela de estudantes do gênero feminino observadas na coorte escolar, características individuais e escolares e efeitos fixos de escola e coorte.

Como parte da nossa metodologia se propõe a analisar o efeito da composição de gênero nas escolhas de graduação, utilizaremos como referência o trabalho de SILVA BORGES, 2021, que se volta para coortes de escolas de ensino médio brasileiras, um contexto similar ao nosso. Para esse propósito, ela emprega dados do Censo Escolar e do vestibular da Universidade Estadual de Campinas (UNICAMP). As definições teóricas apresentadas, assim como a equação definida, foram extraídas do capítulo *Gender peer effects on major choice* (SILVA BORGES, 2021).

2.2.1 Efeitos de gênero em escolhas de graduação

A hipótese da estratégia empírica é de que o efeito de pares tem um papel significativo na influência das escolhas do curso de graduação. Pais e alunos podem escolher suas turmas potenciais, que pode ser um problema por conta do viés da auto-seleção (*self-selection bias*). Apesar das escolhas pessoais das famílias, elas não conseguem prever corretamente as variações coorte a coorte na composição de gênero dos estudantes, que seguem processos aleatórios. Assim, explorando a variação idiossincrática de coortes na proporção de alunas, temos a estimação do seguinte modelo econométrico, onde i indexa estudante, e escolas, c coorte e t ano do vestibular:

$$y_{iect} = \gamma_0 + \beta_1 \text{Feminino}_i + \beta_2 \text{Prop_FemEM}_{iec} + \beta_3 \text{Feminino}_i \times \text{Prop_FemEM}_{iec} + \alpha_e + \alpha_t + X_i \omega + Z_{ec} \delta + \gamma_{et} + \epsilon_{iect} \quad (2.1)$$

A equação representa um modelo de regressão para estimar o impacto de alguns fatores na variável dependente y_{iect} . Ela é composta por:

- Feminino_i , indicador de gênero;
- Prop_FemEM_{iec} , proporção de colegas do gênero feminino na escola e no ano de

conclusão do ensino médio c ;

- α_e , efeitos fixos da escola;
- α_t , efeitos fixos do ano do vestibular;
- X_i , vetor de características individuais do aluno (notas do ENEM, idade, indicador de trabalho durante o ensino médio);
- Z_{ec} , variáveis da coorte escolar: tamanho da turma, proporção de alunos que frequentam aulas diurnas, idade média dos colegas;
- γ_{et} , tendência linear de tempo específica da escola;
- β_1 , coeficiente de diferenças de gênero nas variáveis de resultados;
- β_2 , coeficiente de efeitos de pares de gênero aplicáveis tanto a homens quanto a mulheres;
- β_3 , coeficiente do impacto diferencial de mulheres terem uma proporção maior de colegas do sexo feminino.

As variáveis de resultado (y_{iect}) podem ser agrupadas de acordo com o objetivo principal. Para analisar se maiores proporções de colegas do sexo feminino durante o ensino médio influenciam a escolha de graduação das alunas, são consideradas duas variáveis: um indicador de área intensiva em matemática e física, outro de escolha de curso STEM. Já para analisar o efeito das colegas do sexo feminino na composição de gênero de um curso, são mensurados se o curso tem maioria de mulheres (*female-dominated*), homens (*male-dominated*) ou é balanceado entre os gêneros (*gender-balanced*) e a média de aplicantes do gênero feminino. Por fim, para analisar o impactos das colegas na probabilidade de escolher cursos competitivos, foram consideradas as médias de candidatos por vaga e de notas de corte.

De forma similar, utilizaremos a metodologia de [SILVA BORGES \(2021\)](#) para analisar os efeitos da composição de gênero das coortes de ensino médio nas opções de carreira dos estudantes, dessa vez numa perspectiva mais abrangente, que é proporcionada pelo conjunto de dados do SISU. Abordaremos com mais detalhes no [Capítulo 4](#) como isso será feito.

Capítulo 3

Trabalhos relacionados

Um estudo de grande relevância para o entendimento de como os pares afetam os colegas na sala de aula é o de [HOXBY \(2000\)](#). Nessa pesquisa, a economista Caroline Hoxby explora como a composição de uma turma escolar pode influenciar a experiência de aprendizado dos estudantes e suas conquistas acadêmicas. São utilizados dados de alunos do 4º ao 7º ano do ensino fundamental de escolas públicas do Texas, Estados Unidos na década de 1990. A autora aborda duas fontes de variação idiossincrática, sendo elas as mudanças na composição de gênero e raça de uma turma escolar em anos adjacentes. Duas estratégias empíricas discutem essas variações: na primeira, são avaliados os efeitos de ter um grupo com maioria feminina e diferente composição racial; na segunda, são avaliados os efeitos das conquistas dos pares em grupos masculinos e femininos.

De forma similar, outros trabalhos abordam os efeitos da composição de gênero através de variações idiossincráticas. [SCHØNE *et al.* \(2019\)](#) visaram entender como a composição de gênero afeta escolhas educacionais. Para isso, foram utilizados dados de alunos dos anos finais do ensino fundamental da Noruega, entre os anos de 2003 e 2008. Nesse cenário, são considerados os efeitos da parcela feminina de pares nas escolhas de áreas e disciplinas do ensino médio, mais especificamente naqueles orientados a STEM. Algumas das hipóteses levantadas investigam se há efeitos na performance escolar, na perpetuação de estereótipos de gênero e na competitividade.

[LAVY e SCHLOSSER \(2011\)](#) apresentam a extensão dos efeitos da composição de gênero na função de produção educacional. Nesse trabalho, são investigados as conquistas educacionais de meninos e meninas em diferentes estágios do sistema educacional. Para isso, são utilizados dados de estudantes de Israel do ensino fundamental, entre os anos de 2002 a 2005, e médio, de 1993 a 2000. São utilizados como resultados as notas de disciplinas e performance em exames de entrada. Alguns dos mecanismos destacados são a disrupção e violência na sala de aula, interação entre estudantes, relacionamento aluno-professor e senso de fadiga de professores na sua profissão.

Outro aspecto considerado pela literatura é a diferença no impacto de escolas segregadas por sexo e coeducacionais, onde meninos e meninas são ensinados juntos, como no trabalho de [SCHNEEWEIS e ZWEIMÜLLER \(2012\)](#). Schneeweis e Zweimüller identificam o impacto causal da composição de gênero na escolha de um campo acadêmico. Os dados cobrem

os anos de 1988 e 2006 da cidade de Linz, Áustria, do ensino fundamental, com foco no último ano. No contexto austríaco, os estudantes podem escolher uma área dentro de uma escola vocacional, que prepara para uma vaga de trabalho, ou seguir estudos superiores na universidade. Além de levar em consideração o interesse vocacional, também é estimado o impacto de estudar em turmas com mais colegas do gênero feminino, que pode levar a escolhas de áreas mais técnicas, a depender da composição.

No trabalho de [BRENØE e ZÖLITZ \(2020\)](#) é observado o efeito dos pares a longo prazo. A utilização de dados de registros de estudantes que ingressam na linha de matemática no ensino médio entre os anos de 1980 e 1994 possibilitou o acompanhamento de toda essa população pelo período de 20 anos. Assim, a investigação dos efeitos da composição de gênero na participação na área STEM na Dinamarca pode acompanhar não só as escolhas educacionais, mas as consequências diretas e retardadas ao longo do tempo em turmas de ensino médio. Leva-se em consideração se a exposição a mais pares femininos está correlacionada com a disparidade de gênero. São observadas probabilidade da entrada e finalização de um curso STEM, bem como os ganhos salariais e a fertilidade de homens e mulheres em diferentes etapas da vida.

Ainda sobre estudos que observam variações idiossincráticas na composição de gênero, temos um recente trabalho que considera o contexto nacional. O trabalho de [SILVA BORGES \(2021\)](#) utiliza essa metodologia para avaliar se a composição de gênero de coortes do ensino médio influencia a escolha de curso de graduação de estudantes, em especial as mulheres. Borges levantou algumas questões específicas, como meninas estudantes expostas a maiores proporções de colegas de seu gênero são mais prováveis de escolher áreas de estudo focadas em matemática. Também são consideradas a seleção em cursos balanceados ou com maioria em relação ao gênero e a competitividade, ilustrada através das taxas de admissão dos cursos. Essa análise foi realizada com dados do vestibular entre os anos 2000 a 2008 de uma universidade pública, a UNICAMP, que foram relacionados aos dados do Censo Escolar. Sendo assim, foi possível traçar as escolhas e os perfis dos estudantes inseridos em turmas de ensino médio. O modelo econométrico de Borges identificou que mulheres são menos prováveis a se aplicarem a áreas focadas em matemática e cursos STEM. Outra observação foi que mulheres estão mais concentradas em cursos de maioria feminina, além de escolherem cursos com alto número de concorrentes por vaga, mas com nota de corte inferior à dos homens. Alguns dos seus resultados podem ser vistos na [Tabela 3.1](#).

Outros estudos também utilizaram dados educacionais brasileiros para analisar diferentes fatores. [MACHADO e SZERMAN \(2021\)](#) utilizou dados entre os anos de 2010 e 2017 do SISU e do Censo Escolar para investigar os impactos de sistemas de admissão centralizados na composição de estudantes. Machado observou características dos estudantes como gênero, idade, etnia e migração, além de características das escolas para mensurar os efeitos do SISU na atração de candidatos de diversos perfis. [MELLO \(2022\)](#) analisa como reformas educacionais que expandiram o acesso à educação superior impactaram na admissão de estudantes de baixa renda. Essas políticas incluem a expansão da centralização de aplicações com o SISU e mais oferta de cotas de ações afirmativas. São utilizados dados do Censo da Educação Superior dos anos de 2010 a 2015 e do ENEM dos anos de 2009 a 2014. [OTERO et al. \(2021\)](#) também conduziu um estudo sobre as consequências de ações afirmativas no contexto de admissão em universidades brasileiras. Ele explorou questões como escolhas de área, frequência e persistência na universidade e rendimentos projetados.

Variáveis dependentes	Todos	Mulheres	Homens	Diferença
Matemática ou física são disciplinas prioritárias	0,42 (0,49)	0,26 (0,44)	0,60 (0,49)	-0,33
Curso STEM	0,47 (0,50)	0,34 (0,47)	0,60 (0,49)	-0,27
Área com maioria masculina	0,26 (0,44)	0,08 (0,26)	0,45 (0,50)	-0,38
Área com maioria feminina	0,16 (0,37)	0,26 (0,44)	0,06 (0,23)	0,21
Área balanceada entre os gêneros	0,58 (0,49)	0,66 (0,47)	0,49 (0,50)	0,17
Média de participação de candidatas na carreira	0,50 (0,24)	0,62 (0,18)	0,38 (0,24)	0,24
Média de candidatos por vaga	33,45 (25,94)	36,49 (27,70)	30,25 (23,53)	6,24
Média de nota de corte	525,66 (72,80)	522,14 (78,60)	529,38 (65,93)	-7,24
Observações	139896	71742	68154	139896

Tabela 3.1: Estatísticas descritivas das variáveis dependentes por gênero. A coluna Diferença reporta o coeficiente do teste-*t*. *P*-valor <0,01. Desvio padrão em parênteses. Adaptado de [SILVA BORGES \(2021\)](#)

Para tal, são utilizados dados do ENEM de 2009 a 2015, SISU de 2016, Censo da Educação Superior entre 2009 e 2019 e Relação Anual de Informações Sociais de 2017.

Assim, como foi possível observar, existem diversos trabalhos na literatura que abordam como múltiplos fatores influenciam jovens inseridos no contexto educacional, sendo um deles o efeito da composição de gênero. Dentre aqueles que considerassem o contexto brasileiro, porém, foram encontrados poucos estudos que focassem nesse fator específico. Entre os trabalhos apresentados, o trabalho de [SILVA BORGES \(2021\)](#) é o que mais se assemelha ao que estamos fazendo, porém seus dados se restringem a uma universidade específica. A nossa grande motivação é basear-se na metodologia de variações idiossincráticas para expandir a análise ao nível nacional, utilizando dados do ENEM, SISU e Censo Escolar para observar como estudantes realizam suas escolhas de curso superior e como elas são influenciadas por seus pares no ensino médio.

Capítulo 4

Metodologia

Neste capítulo, definimos a metodologia empregada para responder as questões levantada na pesquisa. As bases de dados utilizadas estão descritas na [Seção 4.1](#). Já as etapas do processo de ciência de dados, que explora os dados obtidos, estão descritas nas seções subsequentes. O pré-processamento está descrito na [Seção 4.2](#) e a análise exploratória de dados está descrita na [Seção 4.3](#).

4.1 Bases de dados

Nesta seção, descrevemos as bases de dados utilizadas nesta etapa da pesquisa. Para realizar a análise das escolhas de graduação, utilizamos dados educacionais dos estudantes. Uma das fontes é o Exame Nacional do Ensino Médio (ENEM). Através dele, conseguimos obter informações sobre alunos concludentes e que já concluíram o ensino médio, bem como de suas escolas. A outra fonte utilizada é o Sistema de Seleção Unificada (SISU). Com ela, obtemos informações relativas à inscrição dos alunos em cursos de nível superior, além de detalhes das instituições e cursos ofertados. Uma versão resumida dos dicionários de dados, que explicam as variáveis das bases, está disponível no Apêndice A.

4.1.1 ENEM

O Exame Nacional do Ensino Médio (ENEM) é um exame realizado pelo Ministério da Educação cujo objetivo é avaliar o desempenho escolar no final da educação básica. Desde 2009, ele passou a ser utilizado como mecanismo de ingresso à educação superior, cujas notas podem ser aproveitadas no Sistema de Seleção Unificada (SISU) e Programa Universidade para Todos (ProUni). O exame também possibilita o pleito de certificação do ensino médio. Os participantes realizam provas em quatro áreas de conhecimento: linguagens, ciências humanas, ciências da natureza e matemática. Além disso, eles devem desenvolver um texto dissertativo-argumentativo dada uma situação-problema, conhecido como redação (INEP, 2023).

Os dados do Exame Nacional do Ensino Médio são disponibilizados publicamente através do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). Os dados abertos do INEP incluem os microdados do ENEM, que reúnem um conjunto

de informações relativas ao exame. Os microdados são o menor nível de desagregação de dados recolhidos, sendo disponibilizados dados dos anos de 1998 a 2022. Nesta etapa preliminar da pesquisa, utilizamos os dados do ano de 2016. Além dos microdados relativos às edições anuais, também são disponibilizados outros arquivos relevantes, como dicionário de dados, documentos técnicos, provas, gabaritos e programa para leitura da base.

Com o passar dos anos, os microdados foram se diferenciando à medida que eram incluídas ou retiradas determinadas variáveis, mas pode-se observar uma estrutura comum entre as edições. Em 2016, os dados estão divididos nas categorias:

- Dados do participante;
- Dados da escola;
- Dados dos pedidos de atendimento especializado;
- Dados dos pedidos de atendimento específico;
- Dados dos pedidos de recursos especializados e específicos para realização das provas;
- Dados dos pedidos de certificação do ensino médio;
- Dados do local de aplicação da prova;
- Dados da prova objetiva;
- Dados da redação;
- Dados do questionário socioeconômico.

Como parte da política adotada pela Lei Geral de Proteção de Dados Pessoais (LGPD), os dados passam por um tratamento antes de serem publicados. Isso significa que dados cadastrais e sensíveis, como nome, endereço, RG, etc, não são disponibilizados ou passam por uma máscara para anonimizá-lo, como é o caso do número de inscrição. Os arquivos de microdados são disponibilizados no formato .csv (valores separados por vírgulas). Os dados de 2016 constituem-se por apenas um arquivo, com uma tabela. Cada linha da tabela representa a inscrição de um candidato de forma individual, bem como as colunas são as variáveis definidas anteriormente, que caracterizam o participante.

4.1.2 SISU

O Sistema de Seleção Unificada (SISU) é um sistema eletrônico do Ministério da Educação, no qual instituições públicas de ensino superior de todo o Brasil oferecem vagas para estudantes participantes do Exame Nacional do Ensino Médio. Durante o período da oferta de vagas, os alunos são ranqueados de acordo com as notas no exame e, aqueles com melhor classificação, são selecionados. Em cada processo seletivo do SISU, que tem duas aberturas anuais, o candidato pode escolher até duas opções de curso. É possível verificar informações sobre as vagas oferecidas, como cursos, instituições e localizações, turnos e modalidade de concorrência (MEC, 2023).

No ato da inscrição, o sistema recupera as notas da edição mais recente anterior do ENEM. Por exemplo, o SISU 2023 leva em consideração a edição do ENEM 2022. Apenas

aqueles que obtiveram nota superior a zero na redação e não têm o status de treineiro no ENEM podem se inscrever. O processo é totalmente digital e gratuito, sendo o estudante o responsável por acompanhar o status da sua inscrição durante o mesmo. Quando não há a aprovação em uma das duas opções selecionadas, conhecido por chamada regular, ainda é possível a disputa por vaga através da lista de espera.

Os dados do Sistema de Seleção Unificada foram obtidos através do Portal de Dados Abertos do MEC. O portal é uma plataforma que disponibiliza dados e informações públicas do Ministério da Educação, que podem ser usadas no desenvolvimento de aplicativos e ações. Além do SISU, é possível observar conjuntos de dados de outros programas como FIES, ProUni e PRONATEC. São disponibilizados dados relacionados às inscrições realizadas nos processos seletivos dos anos de 2017 a 2022. Nesta etapa preliminar da pesquisa, utilizamos os dados do ano de 2017.

São fornecidas informações detalhadas sobre o participante, como dados pessoais e desempenho nas provas do ENEM, a vaga para qual ele se inscreve, além da classificação e aprovação. Diferente do ENEM, não há especificação de categoria dos dados no dicionário fornecido. Também há a aplicação de máscara para anonimizar dados sensíveis, como CPF e número de inscrição. Os arquivos também são disponibilizados no formato .csv. Desta vez, são constituídos por múltiplas tabelas. Cada tabela representa uma etapa do processo de convocação dos candidatos, sendo divididas entre chamada regular e lista de espera. Ocorre duas chamadas regulares ao longo do ano, uma em cada semestre. Já a quantidade de listas de espera varia, conforme o preenchimento de vagas nas etapas anteriores. Nesta etapa da pesquisa, utilizamos apenas os dados das chamadas regulares. Cada linha da tabela representa uma inscrição de um candidato, sendo possível que um candidato tenha múltiplas inscrições por conta das duas aberturas do processo ao longo do ano e por poder se inscrever em mais de um curso.

4.1.3 Censo Escolar

O Censo Escolar é um levantamento de informações da educação básica brasileira em escolas e instituições de ensino por todo o país. Essa ferramenta demográfica realiza coletas anuais em colaboração entre o Inep e as secretarias estaduais e municipais de educação, contando com a participação de todas as escolas públicas (federais, estaduais e municipais) e privadas da rede de ensino. O Censo abrange diferentes etapas e modalidades de ensino da educação básica e profissional. Ele permite a obtenção de dados individualizados, em diversos aspectos, de estudantes, professores, turmas e escolas. A pesquisa é realizada em duas etapas: a primeira coleta informações sobre os estabelecimentos de ensino, gestores, turmas, alunos e profissionais escolares em sala de aula; a segunda, informações sobre o movimento e o rendimento escolar dos alunos. Os dados do Censo Escolar, de forma semelhante aos anteriores, são disponibilizados ao público pelo INEP no formato .csv.

4.2 Pré-processamento

Na etapa da metodologia de pré-processamento, os dados são tratados a fim de adaptá-los às necessidades do projeto. Com isso, otimizamos as etapas posteriores através de obtenção de um conjunto de dados que seja mais relevante para a pesquisa, facilitando o

processo de análise. As técnicas aplicadas podem ser agrupadas em redução, integração, limpeza e transformação de dados. Utilizamos como referência os trabalhos de JAFARI (2022) e GARCÍA *et al.* (2016), que provêm definições e exemplos práticos de como realizar esses processos.

É importante frisar que o processo de ciência de dados não é rígido e estático, mas sim um processo que se flexibiliza à medida que novas necessidades vão surgindo durante o projeto. Assim, as técnicas aplicadas no pré-processamento também são utilizadas além da etapa inicial. Para isso, utilizamos a linguagem Python, com as bibliotecas pandas e NumPy para análise e manipulação de dados.

Os datasets originais possuem um grande volume de dados, tanto pela quantidade de participantes inscritos, quanto pela quantidade de informações armazenadas sobre eles, expressas pelas colunas (ou *features*). Uma consequência disso é uma maior utilização de recursos computacionais para processá-los, seja nos processos de leitura e escrita de arquivo quanto no armazenamento. Isso é particularmente importante pela inclusão de múltiplas fontes de dados. Assim, visamos diminuir a quantidade de dados pouco relevantes para a pesquisa, gerando dados menos volumosos e mais representativos.

Uma técnica aplicada na redução foi a seleção de *features*. Na seleção de *features*, visando diminuir a dimensionalidade do conjunto de dados, é gerado um subconjunto das *features* originais através da identificação e remoção daquelas pouco relevantes ou redundantes. Isso foi realizado no *dataset* do ENEM, que possui uma grande quantidade de colunas, e algumas informações, como as referentes à aplicação da prova, não são importantes na análise a ser realizada. Outra técnica aplicada foi a tipagem explícita de dados. Uma particularidade da biblioteca utilizada, pandas, é que a inferência de tipos nem sempre é a mais eficiente, o que pode levar a uma utilização de memória maior que o esperada, além de dificultar operações específicas, como manipulação de *strings* e realização de cálculos matemáticos. Para contornar esse problema, fizemos uma análise dos valores, a fim de mapeá-los para tipos de dados mais precisos, que pudessem melhorar os resultados da análise.

Outra característica dos *datasets* é que eles são imperfeitos. Apesar da presença de documentação auxiliar que define como os dados se comportam, na prática, os dados têm um estado diferente, incompletos e com "sujeiras". Como a qualidade dos dados interfere nos resultados obtidos, foi necessário observar características dos dados para definir uma abordagem adequada. Um dos problemas notados foi a ausência de valores em alguns campos. Por exemplo, foram encontrados registros de inscrições do SISU que não possuíam o curso de graduação do candidato. Optamos por não descartar esses registros que tivessem valores faltando, já que poderia causar uma perda de acurácia e um viés de auto-seleção, bem como a desconsideração de algumas *features* mais importantes que outras.

Visando solucionar essa questão, criamos novas classes de valores para representar uma instância de informação ausente e, quando possível, utilizamos funções de probabilidade para inserir valores inferidos. Também tornamos os dados consistentes entre si, já que para a combinação das bases de dados, é necessário que os valores sejam do mesmo tipo e estejam uniformes em cada uma delas. Ao realizarmos uma análise de anomalias para detectar valores problemáticos, observamos uma situação particular na qualidade dos dados do SISU. Por conter mais de um *dataset* da chamada regular, era necessário agrupá-los

para obter o cenário do ano como um todo, mas um deles estava parcialmente corrompido. A utilização inadvertida de dados não-estruturados pode levar a interpretações incorretas e falsas conclusões no processo de análise. Assim, realizamos um filtro dessas anomalias, relacionadas aos nomes das instituições de ensino, seus campi e cursos de graduação, que foram identificadas e tratadas para refletir o comportamento esperado.

4.3 Análise exploratória de dados

A partir da obtenção dos conjuntos de dados pré-processados, pudemos realizar uma análise exploratória de dados. Nesta etapa, iremos utilizar técnicas de manipulação de dados e ferramentas estatísticas para investigar os dados. Nela, conseguimos ter um entendimento melhor do cenário geral a ser explorado, como os dados estão distribuídos, quais são as variáveis, como elas se relacionam entre si e como podemos utilizá-las para responder as questões de pesquisa.

Por definição, o *dataset* do ENEM possui os dados do exame de estudantes de todo o país, que totalizam 8627367 registros. Cada registro corresponde a inscrição de um único estudante. No ato de realização do exame, esses participantes estão em diferentes situações em relação ao ensino médio, podendo já ter concluído, estar cursando ou não ter concluído e não estar cursando. Essa informação está expressa em duas variáveis: TP_ST_CONCLUSAO, Situação de conclusão do Ensino Médio, e Q046, que é a resposta do questionário socioeconômico à pergunta "Você já concluiu ou está concluindo o Ensino Médio?", visualizadas na [Tabela 4.1](#). Em ambas, mais da metade dos estudantes já concluiu o ensino médio, seguidos daqueles que estão cursando e irão concluir em 2016.

Situação	TP_ST_CONCLUSAO		Q046	
	Quantidade	Percentual	Quantidade	Percentual
Já concluí o Ensino Médio	4928251	57.12%	4947935	57.35%
Estou cursando e concluirei o Ensino Médio em 2016	1882278	21.82%	1872570	21.70%
Estou cursando e concluirei o Ensino Médio após 2016	1344085	15.58%	1331073	15.43%
Não concluí e não estou cursando o Ensino Médio	472753	5.48%	475785	5.51%

Tabela 4.1: Tabulação cruzada das variáveis TP_ST_CONCLUSAO e Q046

Outros dados relevantes são o ano e o tipo de ensino em que o participante concluiu o ensino médio. A primeira informação está expressa na variável TP_ANO_CONCLUIU, que explicita dos anos anteriores ao exame até 2007, agrupando o restante em anterior a 2007 ou não informado, que pode ser visualizado na [Tabela 4.2](#). No caso de anos não informados, isso acontece tanto por haver valores ausentes da fonte, quanto pelos alunos concluintes não terem um ano de conclusão especificado. A segunda está expressa em TP_ENSINO, que descreve o tipo de ensino da instituição na qual o aluno concluiu o ensino médio, que pode ser visualizada na [Tabela 4.3](#); nem todos os alunos possuem essa informação. Pode-se observar que um valor significativo dos estudantes não têm o ano de conclusão informado (42,88%), além de que aqueles que concluíram em ensino regular constituírem a maioria

(19,15%). Uma variável importante para essa análise da composição das turmas de ensino médio é CO_ESCOLA. Ela representa um código que identifica de maneira unificada a instituição junto ao Ministério da Educação. É através dela que poderemos associar a qual escola esse estudante pertence dentro da base do Censo Escolar. De todos os estudantes, apenas 21.81% possuem esse código.

Ano	Quantidade	Percentual
2015	966842	11.21%
2014	699987	8.11%
2013	527310	6.11%
2012	416454	4.83%
2011	317364	3.68%
2010	294214	3.41%
2009	244461	2.83%
2008	199619	2.31%
2007	178091	2.06%
Anterior a 2007	1083909	12.56%
Não informado	3699116	42.88%

Tabela 4.2: *Tabulação da variável TP_ANO_CONCLUIU*

Tipo de ensino	Quantidade	Percentual
Ensino Regular	1652485	19.15%
Educação Especial - Modalidade Substitutiva	10295	0.12%
Educação de Jovens e Adultos	218532	2.53%

Tabela 4.3: *Tabulação da variável TP_ENSINO*

Como nem todos os registros presentes no conjunto são relevantes para a pesquisa, estabelecemos alguns critérios para a aplicação de filtros.

- Registros que possuam informação de gênero, já que sem ela não podemos analisar o efeito da composição de gênero;
- Registros de estudantes que já concluíram o ensino médio ou que irão concluir em 2016, excluindo aqueles que concluirão após 2016 ou não estão cursando e não concluíram;
- Registros de estudantes que tenham concluído o ensino médio após 2007, já que não há especificação do ano anterior a 2007;
- Registros de estudantes que realizaram ensino médio no Brasil, excluindo aqueles que estudaram no exterior por não haver informação dessas escolas;
- Registros de estudantes que não estão realizando o exame como treineiros, já que treineiros não estão aptos a utilizar a nota do ENEM para ingresso em universidade;
- Registros que possuam informação da escola de ensino médio, já que sem ela não é possível identificar a escola em outra base;

- Registros que possuam informação de todas as notas.

A [Tabela 4.4](#) descreve as observações durante a aplicação dos filtros. Cada filtro foi aplicado sequencialmente, sendo que os valores mostrados correspondem ao total do conjunto original subtraindo os total de registros excluídos até o momento. Após a aplicação de todos os filtros, restaram-se 1388044 registros, que correspondem a 16,09% do conjunto original. Esse subconjunto de dados filtrados do original é o que passa a ser utilizado nas análises subsequentes.

Filtro	Quantidade	Percentual
Conjunto original	8627367	100.00%
Com informação de gênero	8627367	100.00%
Já concluíram ou concluirão EM em 2016	6810529	78.94%
Concluíram após 2007	5726620	66.38%
Concluíram EM no Brasil	5725635	66.37%
Concluíram EM no ensino regular	5496816	63.71%
Não são treineiros	5496816	63.71%
Com informação presente da escola de EM	1652471	19.15%
Com informação presente de notas	1388044	16.09%

Tabela 4.4: *Tabulação dos filtros aplicados no conjunto de dados do ENEM*

Já no *dataset* do SISU, os dados estão estruturados de forma diferente, com uma listagem (ou abertura) por semestre na realização do processo seletivo. Cada registro corresponde a uma inscrição de um estudante em um curso, totalizando 6665892 registros. Na [Tabela 4.5](#), podemos observar a quantidade de inscrições e participantes inscritos, divididos por listagem. Em ambos os casos, a primeira listagem concentra a maioria dos registros.

Listagem	Quantidade de inscrições	Quantidade de inscritos
1	4868545	2494173
2	1797347	935538
Total	6665892	3429711

Tabela 4.5: *Tabulação de inscrições e inscritos no SISU, divididos por listagem*

Quando segmentado pela opção de cursos nos quais o candidato se inscreveu, que pode ser de apenas 1 ou 2 cursos, pode-se perceber que 94,3% dos participantes opta por se inscrever nas duas opções disponíveis, como mostra a [Tabela 4.6](#).

Opção de curso	Quantidade de inscritos
Apenas 1 curso	193530
2 cursos	3236181
Total	3429711

Tabela 4.6: *Tabulação de quantidade de cursos optados no SISU*

É possível que um candidato realize inscrição em somente uma das aberturas ou em ambas, observado na [Tabela 4.7](#). Mais da metade (64,4%) optou por se inscrever apenas na primeira abertura.

Caso de inscrição	Quantidade de inscritos
Apenas na listagem 1	1689691
Apenas na listagem 2	131056
Ambas as listagens	804482
Total	2625229

Tabela 4.7: *Tabulação da quantidade de inscritos por caso de inscrição em listagem no SISU*

Também é possível que o candidato tenha até 4 inscrições em cursos ao longo do ano, com o máximo de 2 por semestre. A [Tabela 4.8](#) mostra a quantidade de inscritos pelo total de inscrições. A maior parte dos estudantes realizou 2 inscrições (65,08%), com o segundo maior total sendo de 4 inscrições (27,90%).

Total de inscrições	Quantidade de inscritos	Porcentagem
1	116802	4.45%
2	1708599	65.08%
3	67420	2.57%
4	732408	27.90%

Tabela 4.8: *Tabulação da quantidade de inscritos pelo total de inscrições no SISU*

De forma semelhante ao caso do ENEM, foram realizados alguns filtros no *dataset* do SISU, sendo eles a exclusão de registros que não possuam informação de gênero e de registros que não possuam a informação de todas as notas. Não houve retirada de nenhum registro após a aplicação dos filtros, já que essas informações estavam disponíveis em todos eles. Também foi aplicada uma técnica de transformação de dados, de modo a agrupar todas as inscrições de um candidato em um único registro. Assim, houve uma redução de 60,62% no *dataset*, passando de 6665892 a 2625230 registros. Agora, o registro deixa de ser a inscrição de um candidato em um único curso para ser a inscrição no SISU como um todo.

Para identificar cada participante unicamente nas bases de dados, é necessário atribuir um campo como chave identificadora. Boas chaves candidatas para esse propósito seriam o CPF ou número de inscrição. Diferente de outros estudos da literatura, estamos lidando com dados disponibilizados publicamente. Isso significa que informações sensíveis, tais como essas chaves, não foram fornecidas. Assim, foi necessário desenvolver uma nova chave a partir dos campos existentes. Para isso, realizamos uma investigação nas notas das provas, que foram agrupadas para observar se é possível utilizá-las como identificador único.

No *dataset* do ENEM, as notas agrupadas foram suficientes para identificar cada um dos alunos. Isso porque não há valores duplicados para essas notas, ou seja, cada aluno possui notas únicas. No SISU, isso não foi possível de ser observado, já que 22 notas estão repetidas. Essas notas pertencem a 75 inscritos. As notas repetidas se enquadram em dois

casos: 4 notas objetivas zeradas, com exceção da redação (ex.: 0 - 0 - 0 - 0 - 520) ou todas as notas maiores que zero (ex.: 570,4 - 619,6 - 433,5 - 546,1 - 520). Não há caso de candidatos com todas as notas iguais a zero, pois o SISU impede a inscrição quando a nota da redação é zerada. Então, utilizamos outros campos para compor o identificador. O identificador final constitui-se das notas agrupadas, idade, unidade federativa de residência e gênero. Como esses campos estão disponíveis em ambas as bases, é possível identificar o mesmo candidato nas bases com a mesma chave identificadora.

Os cursos de graduação foram agrupados de acordo com a área do conhecimento, utilizando a classificação documentada pelo International Standard Classification of Education¹, realizada pela UNESCO. Nessa classificação, 25 áreas do conhecimento, e seus respectivos cursos, estão organizadas em 9 grupos:

- Educação;
- Humanidades e artes;
- Ciências sociais, negócios e direito;
- Ciência;
- Engenharia, manufatura e construção;
- Agricultura;
- Saúde e bem-estar;
- Serviços;
- Não informado.

Através do identificador criado, foi realizada a combinação das bases de dados. O resultado foi uma base única que associa todas as informações dos candidatos presentes nos *datasets* do ENEM e SISU. Nessa base, são encontrados 712526 registros, que representam participantes encontrados tanto no SISU, quanto no ENEM. 27.14% dos participantes do ENEM foram encontrados no SISU, ao passo que 51.33% dos participantes do SISU foram encontrados no ENEM. Vale notar que, para se inscrever no SISU, é obrigatório a realização do ENEM, mas nesse estudo, estamos utilizando apenas uma parte dos dados do exame, por isso a correspondência de participantes do SISU no ENEM não é de 100%.

O conjunto de dados obtido será utilizado na combinação com uma terceira fonte, a do Censo Escolar. Através do Censo, poderemos caracterizar as escolas de ensino médio nas quais os estudantes se formaram. Os dados dessa fonte passarão por pré-processamento e análise exploratória, semelhante ao já realizado. Também com esses dados, realizaremos a aplicação da metodologia de variações idiossincráticas na composição de gênero através das coortes escolares. Isso será realizado em etapas futuras da pesquisa, que serão descritas no **Capítulo 6**. O próximo capítulo apresentará os resultados da análise descritiva dos dados levantados até então. Essa análise está segmentada por gênero e por aprovação nos cursos.

¹ <https://uis.unesco.org/sites/default/files/documents/international-standard-classification-of-education-iscd-2011-en.pdf>

Capítulo 5

Resultados preliminares

5.1 Análise descritiva

5.1.1 Perfil socioeconômico

Variável	Gênero	Feminino	Masculino
	Descrição		
Dependência administrativa da escola	Estadual	77,33%	73,80%
	Federal	2,21%	2,99%
	Municipal	0,92%	0,88%
	Privada	19,53%	22,32%
Escolaridade mãe	Com EM completo	50,26%	55,69%
	Sem EM completo	49,74%	44,31%
Escolaridade pai	Com EM completo	39,01%	44,72%
	Sem EM completo	60,99%	55,28%
Idade	15 anos ou menos	0,16%	0,11%
	16 a 20 anos	97,15%	96,97%
	21 a 25 anos	1,78%	2,38%
	26 a 30 anos	0,39%	0,28%
	31 a 35 anos	0,23%	0,11%
	36 a 40 anos	0,14%	0,06%
	41 a 45 anos	0,08%	0,04%
	46 a 50 anos	0,04%	0,02%
	50 anos ou mais	0,03%	0,02%
Raça	Amarela	2,45%	1,86%
	Branca	40,53%	41,57%
	Indígena	0,54%	0,61%
	Não declarado	1,44%	1,83%
	Parda	44,30%	42,40%
	Preta	10,73%	11,72%
	Sem informação	0,00%	0,00%

Tabela 5.1: Estatísticas descritivas dos participantes do ENEM por gênero

Variável	Situação	Aprovado	Não aprovado
	Descrição		
Grau	Bacharelado	62,82%	66,39%
	Licenciatura	26,08%	21,62%
	Tecnológico	7,48%	9,52%
	Área Básica de Ingresso (ABI)	3,63%	2,46%
Gênero	Feminino	46,49%	57,94%
	Masculino	53,51%	42,06%
Modalidade	Ampla concorrência	45,78%	45,76%
	Ações afirmativas	54,22%	54,24%
Turno	Integral	44,27%	45,34%
	Matutino	12,34%	12,28%
	Noturno	35,45%	34,67%
	Vespertino	7,94%	7,70%

Tabela 5.2: Estatísticas descritivas de cursos e participantes do SISU por situação de aprovação

Intervalo		(0, 200]	(200, 400]	(400, 600]	(600, 800]	(800, 1000]
Prova	Gênero					
Ciências	Feminino	0,00%	4,06%	78,43%	17,51%	0,01%
Humanas	Masculino	0,00%	3,90%	71,81%	24,28%	0,02%
Ciências da	Feminino	0,00%	12,75%	80,99%	6,25%	0,00%
Natureza	Masculino	0,00%	9,14%	80,03%	10,80%	0,03%
Linguagens	Feminino	0,00%	4,18%	83,41%	12,42%	0,00%
	Masculino	0,00%	4,51%	81,05%	14,44%	0,00%
Matemática	Feminino	0,00%	22,11%	66,32%	11,02%	0,55%
	Masculino	0,00%	13,67%	63,26%	21,16%	1,90%
Redação	Feminino	0,24%	11,00%	57,36%	26,08%	5,32%
	Masculino	0,43%	15,48%	58,48%	21,72%	3,88%

Tabela 5.3: Observações de notas dos participantes do ENEM por gênero

Intervalo		(0, 200]	(200, 400]	(400, 600]	(600, 800]	(800, 1000]
Prova	Situação					
Ciências	Aprovado	0,00%	0,02%	20,60%	79,19%	0,18%
Humanas	Não aprovado	0,06%	1,95%	69,81%	28,16%	0,02%
Ciências da	Aprovado	0,00%	0,30%	53,09%	46,42%	0,19%
Natureza	Não aprovado	0,06%	7,80%	80,93%	11,19%	0,02%
Linguagens	Aprovado	0,00%	0,15%	42,26%	57,59%	0,00%
	Não aprovado	0,01%	2,44%	79,82%	17,73%	0,00%
Matemática	Aprovado	0,00%	1,47%	32,22%	57,44%	8,87%
	Não aprovado	0,01%	14,04%	64,61%	19,88%	1,46%
Redação	Aprovado	0,24%	0,22%	15,26%	53,90%	30,62%
	Não aprovado	0,12%	8,33%	55,70%	29,53%	6,32%

Tabela 5.4: Observações de notas dos participantes do SISU por situação de aprovação

Gênero		Feminino			Masculino		
Estatística	Média	Desvio padrão	p-valor	Média	Desvio padrão	p-valor	
Variável							
Idade	17,69	1,96	0,0	17,77	1,64	0,0	
Nota Ciências Humanas	532,12	72,75	0,0	544,01	76,72	0,0	
Nota Ciências da Natureza	474,44	71,18	0,0	495,02	78,94	0,0	
Nota Linguagens	522,25	68,01	0,0	524,28	71,02	0,0	
Nota Matemática	476,82	96,86	0,0	518,36	115,32	0,0	
Nota Redação	559,52	150,15	0,0	531,96	155,58	0,0	

Tabela 5.5: Estatísticas descritivas das notas e idade do ENEM por gênero

Situação		Aprovado		Não aprovado		
Estatística	Média	Desvio padrão	p-valor	Média	Desvio padrão	p-valor
Variável						
Idade	20.73	5.77	0,0	22.38	7.11	0,0
Nota Ciências Humanas	638.67	50.47	0,0	557.41	72.75	0,0
Nota Ciências da Natureza	593.83	69.85	0,0	497.64	78.74	0,0
Nota Linguagens	605.11	48.45	0,0	538.94	65.55	0,0
Nota Matemática	646.52	114.98	0,0	513.66	110.82	0,0
Nota Redação	746.67	120.40	0,0	585.07	133.80	0,0

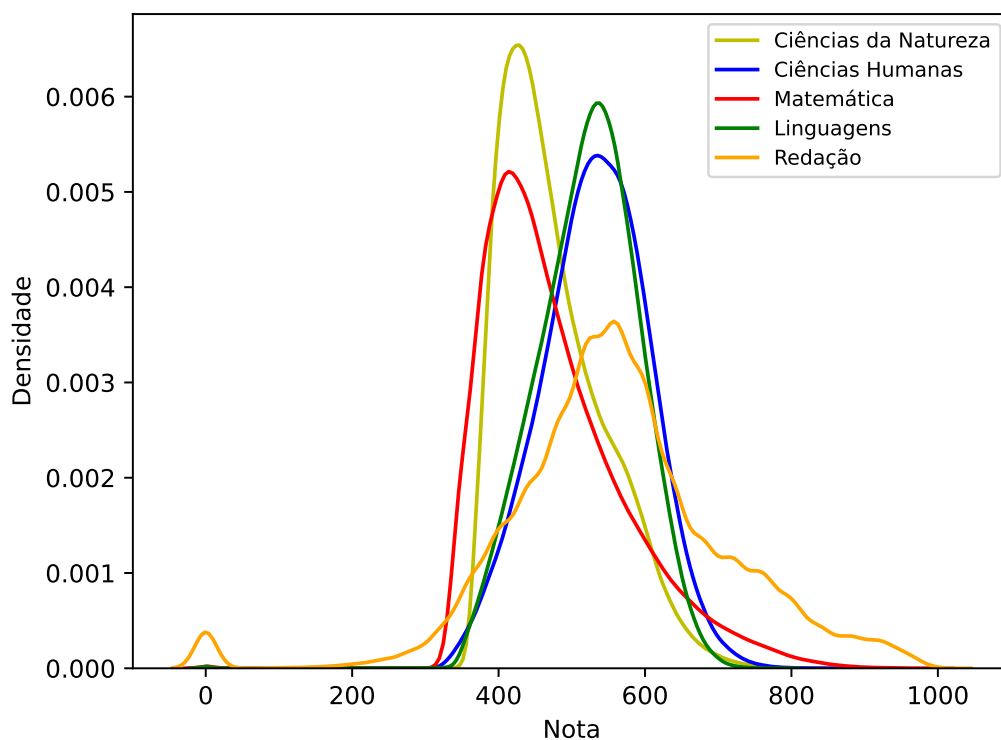
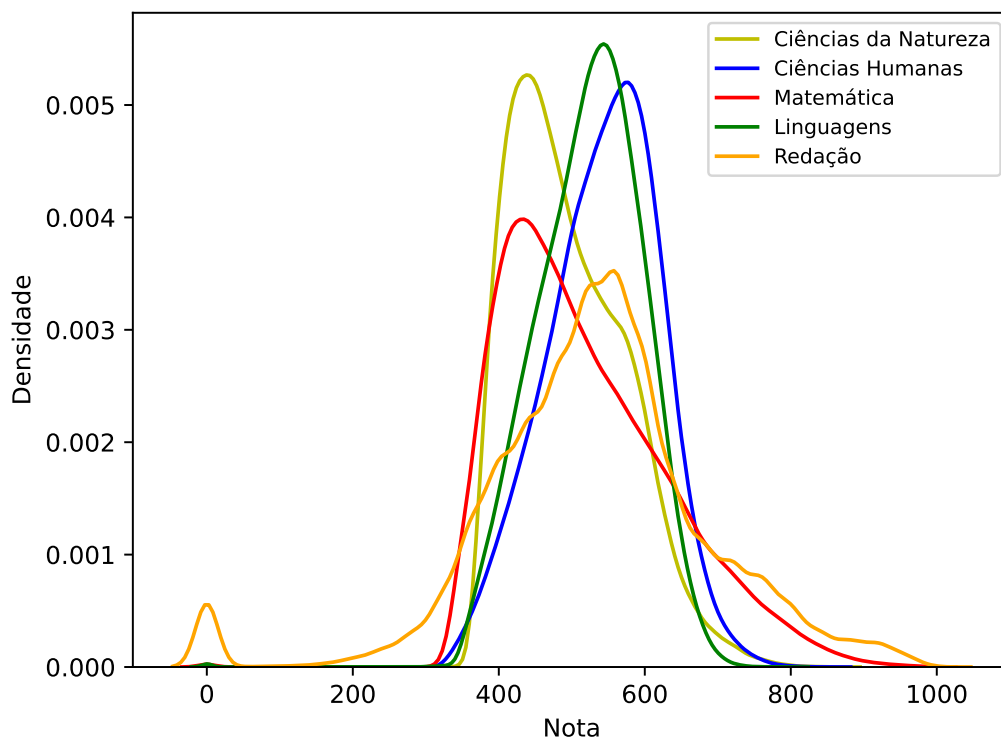
Tabela 5.6: Estatísticas descritivas das notas e idade do SISU por situação de aprovação

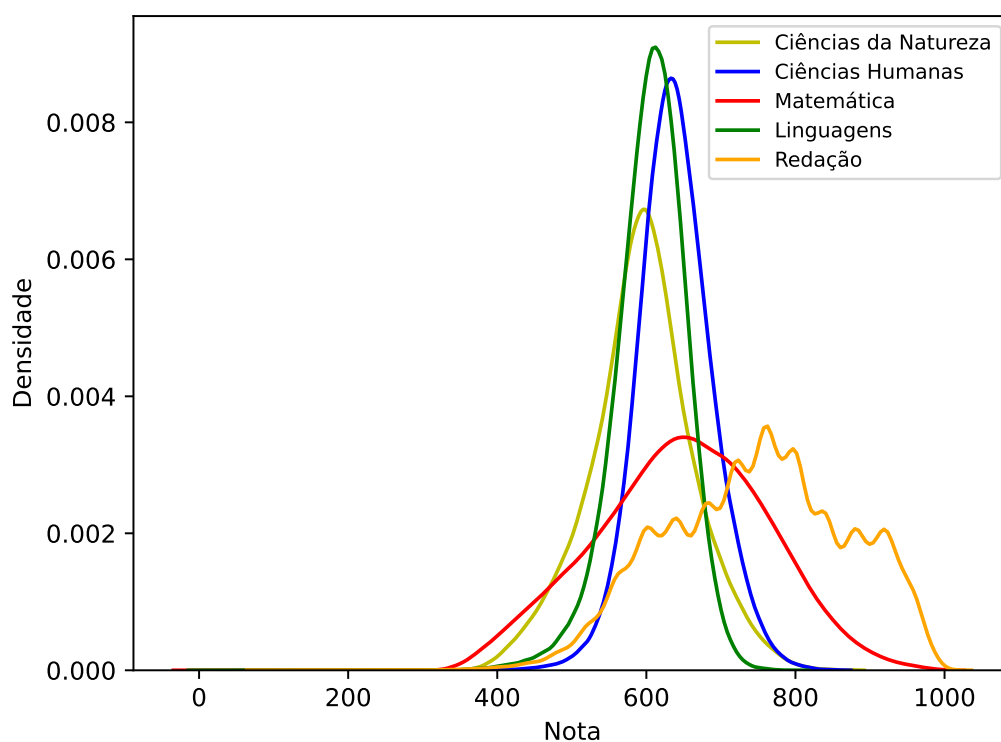
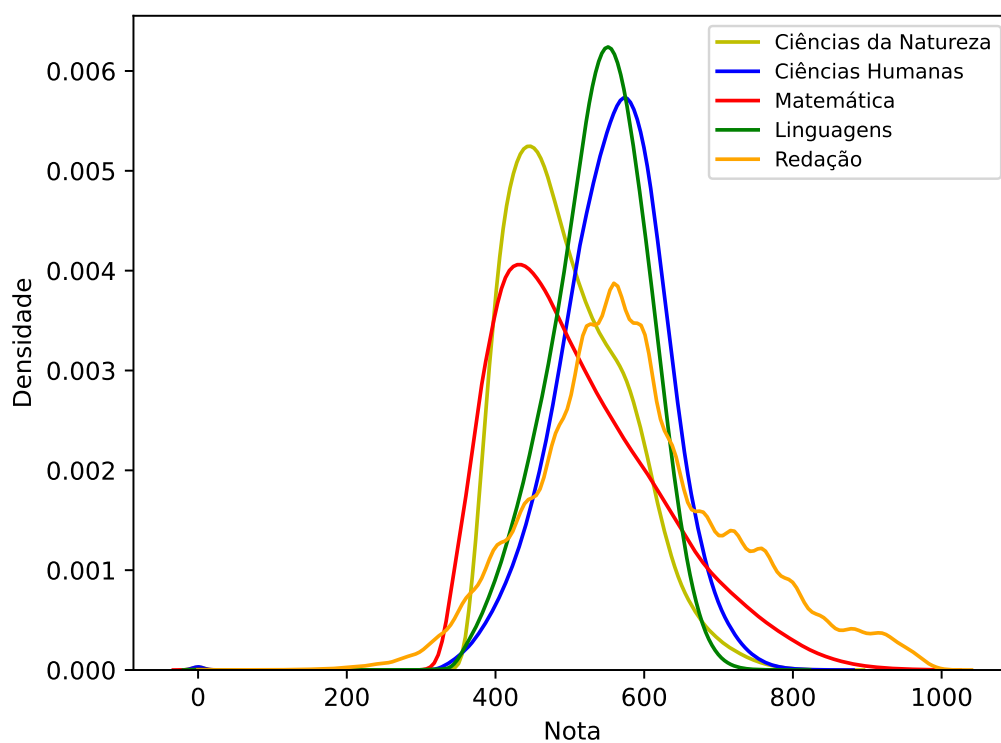
Classificação	Feminino	Masculino
Educação	6,29%	1,81%
Humanidades e artes	8,94%	7,69%
Ciências sociais, negócios e direito	22,24%	20,85%
Ciência	13,01%	21,85%
Engenharia, manufatura e construção	9,76%	19,96%
Agricultura	5,19%	4,90%
Saúde e bem-estar	22,95%	10,42%
Serviços	6,53%	7,76%
Não informado	5,10%	7,76%

Tabela 5.7: Distribuição da classificação dos cursos do SISU por gênero

Instituição de Ensino	Porcentagem
Universidade Federal do Maranhão	3,79%
Universidade Federal do Rio de Janeiro	3,43%
Universidade Federal Fluminense	3,13%
Universidade Federal da Bahia	3,01%
Instituto Federal de Educação, Ciência e Tecnologia do São Paulo	2,63%
Universidade Federal do Piauí	2,60%
Universidade Federal da Paraíba	2,60%
Universidade Federal de Minas Gerais	2,58%
Universidade Tecnológica Federal do Paraná	2,29%
Universidade Federal de Pernambuco	2,17%

Tabela 5.8: Top 10 universidades com mais inscrições

(a) *Feminino*(b) *Masculino***Figura 5.1:** Distribuição das notas do ENEM por gênero

(a) *Aprovado*(b) *Não aprovado***Figura 5.2:** Distribuição das notas do SISU por situação de aprovação

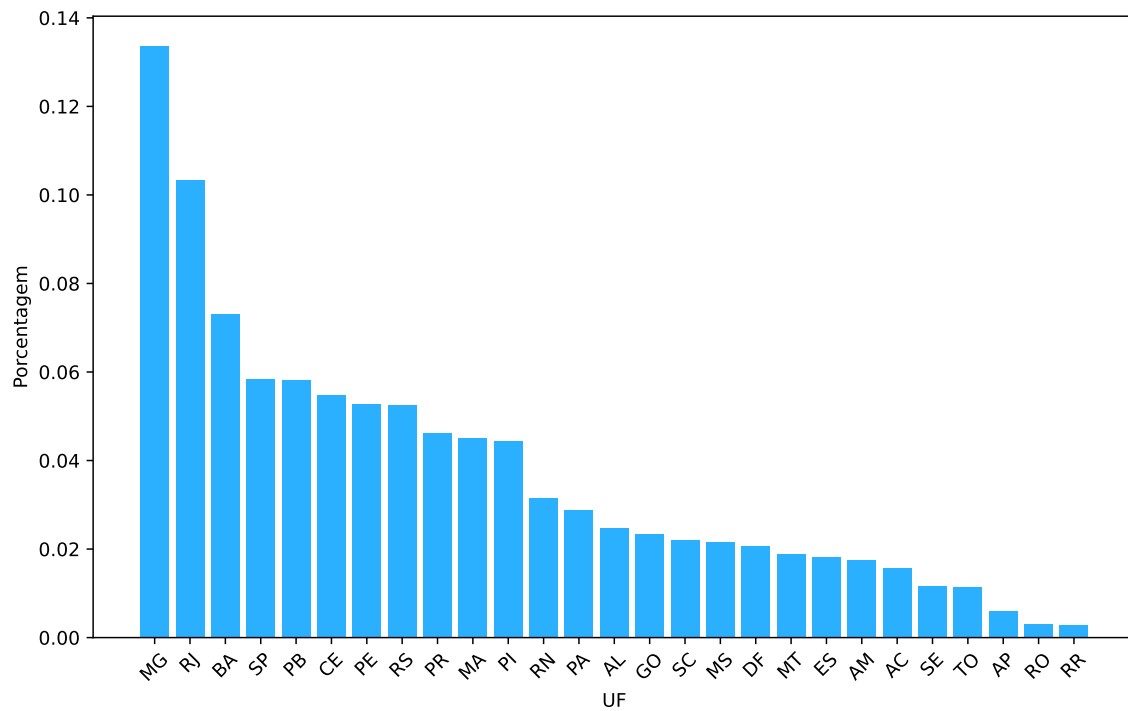


Figura 5.3: Distribuição das unidades federativas das instituições de ensino

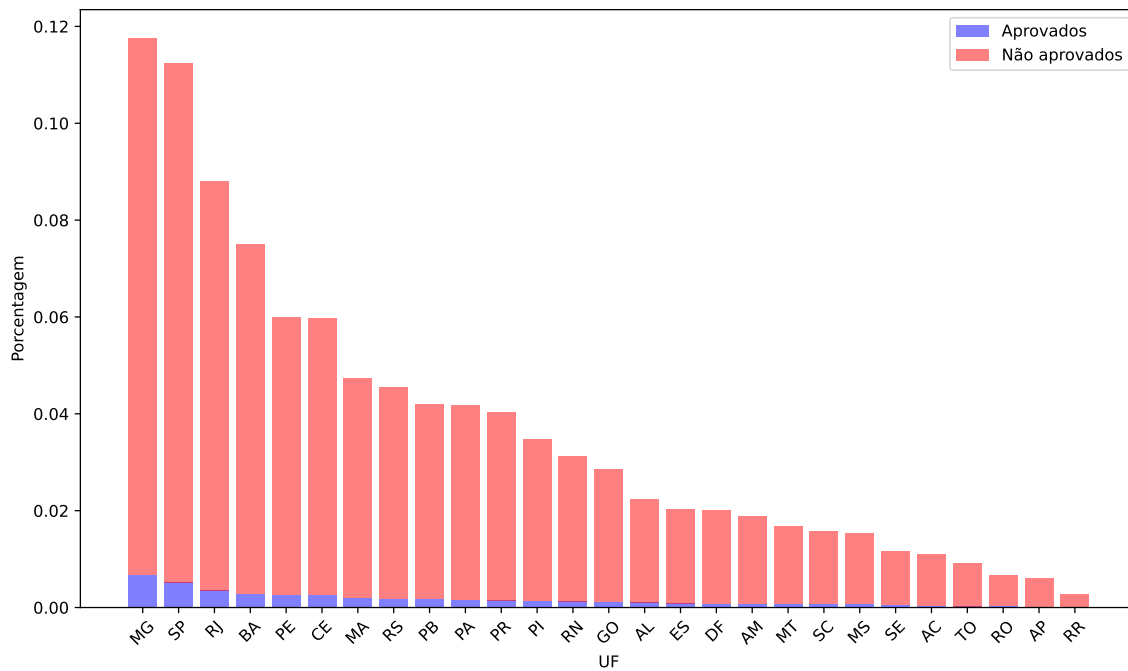


Figura 5.4: Distribuição das unidades federativas de residência dos alunos por situação de aprovação

Capítulo 6

Plano de trabalho

Oii

Referências

- [ABBAGNANO 2012] Nicola ABBAGNANO. *Dicionário de filosofia*. Mar. de 2012 (citado na pg. 2).
- [MEC 2023] Portal Único de ACESSO AO ENSINO SUPERIOR. SISU. 2023. URL: <https://accessunico.mec.gov.br/sisu> (acesso em 13/10/2023) (citado na pg. 14).
- [AKOSAH-TWUMASI *et al.* 2018] Peter AKOSAH-TWUMASI, Theophilus I. EMETO, Daniel LINDSAY, Komla TSEY e Bunmi S. MALAU-ADULI. “A systematic review of factors that influence youths career choices—the role of culture”. Em: *Frontiers in Education* 3 (jul. de 2018). DOI: [10.3389/feduc.2018.00058](https://doi.org/10.3389/feduc.2018.00058). URL: <https://doi.org/10.3389/feduc.2018.00058> (citado na pg. 1).
- [BJÖRKENSTAM *et al.* 2011] Charlotte BJÖRKENSTAM *et al.* “School grades, parental education and suicide—a national register-based cohort study”. Em: *Journal of Epidemiology & Community Health* 65.11 (2011), pgs. 993–998. ISSN: 0143-005X. DOI: [10.1136/jech.2010.117226](https://doi.org/10.1136/jech.2010.117226). eprint: <https://jech.bmj.com/content/65/11/993.full.pdf>. URL: <https://jech.bmj.com/content/65/11/993> (citado na pg. 4).
- [BORCHERT 2001] Michael BORCHERT. “Career choice factors of high school students”. Em: *Career choice factors* (nov. de 2001). URL: <https://minds.wisconsin.edu/bitstream/handle/1793/40311/2002borchertm.pdf?sequence=1> (citado na pg. 2).
- [BRENØE e ZÖLITZ 2020] Anne Ardila BRENØE e Ulf ZÖLITZ. “Exposure to more female peers widens the gender gap in STEM participation”. Em: *Journal of Labor Economics* 38.4 (out. de 2020), pgs. 1009–1054. DOI: [10.1086/706646](https://doi.org/10.1086/706646). URL: <https://doi.org/10.1086/706646> (citado nas pgs. 6, 10).
- [CABRERA e NASA 2000] Alberto F. CABRERA e Steven M. La NASA. “Understanding the college-choice process”. Em: *New Directions for Institutional Research* 2000.107 (2000), pgs. 5–22. DOI: [10.1002/ir.10701](https://doi.org/10.1002/ir.10701). URL: <https://doi.org/10.1002/ir.10701> (citado na pg. 1).
- [CARUANA *et al.* 2015] Edward Joseph CARUANA, Marius ROMAN, Jules HERNÁNDEZ-SÁNCHEZ e Piergiorgio SOLLI. “Longitudinal studies”. Em: *Journal of Thoracic Disease* 7.11 (2015). ISSN: 2077-6624. URL: <https://jtd.amegroups.org/article/view/5822> (citado na pg. 4).

- [ENSMINGER e SLUSARCICK 1992] Margaret E. ENSMINGER e Anita L. SLUSARCICK. “Paths to high school graduation or dropout: a longitudinal study of a first-grade cohort”. Em: *Sociology of Education* 65.2 (abr. de 1992), pg. 95. DOI: [10.2307/2112677](https://doi.org/10.2307/2112677). URL: <https://doi.org/10.2307/2112677> (citado na pg. 4).
- [INEP 2023] Instituto Nacional de ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA. *Exame Nacional do Ensino Médio (Enem)*. 2023. URL: <https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem> (acesso em 13/10/2023) (citado na pg. 13).
- [GARCÍA *et al.* 2016] Salvador GARCÍA, Sergio RAMÍREZ-GALLEGO, Julián LUENGO, José Manuel BENÍTEZ e Francisco HERRERA. “Big data preprocessing: methods and prospects”. Em: *Big Data Analytics* 1.1 (nov. de 2016). DOI: [10.1186/s41044-016-0014-0](https://doi.org/10.1186/s41044-016-0014-0). URL: <https://doi.org/10.1186/s41044-016-0014-0> (citado na pg. 16).
- [GATI e SAKA 2001] Itamar GATI e Noa SAKA. “High school students’ career-related decision-making difficulties”. Em: *Journal of Counseling & Development* 79.3 (jul. de 2001), pgs. 331–340. DOI: [10.1002/j.1556-6676.2001.tb01978.x](https://doi.org/10.1002/j.1556-6676.2001.tb01978.x). URL: <https://doi.org/10.1002/j.1556-6676.2001.tb01978.x> (citado na pg. 1).
- [HOXBY 2000] Caroline HOXBY. *Peer Effects in the Classroom: Learning from Gender and Race Variation*. Rel. técn. Ago. de 2000. DOI: [10.3386/w7867](https://doi.org/10.3386/w7867). URL: <https://doi.org/10.3386/w7867> (citado nas pgs. 5, 9).
- [JAFARI 2022] Roy JAFARI. *Hands-On Data Preprocessing in Python*. en. Birmingham, England: Packt Publishing, mai. de 2022 (citado na pg. 16).
- [LAVY e SCHLOSSER 2011] Victor LAVY e Analía SCHLOSSER. “Mechanisms and impacts of gender peer effects at school”. Em: *American Economic Journal: Applied Economics* 3.2 (abr. de 2011), pgs. 1–33. DOI: [10.1257/app.3.2.1](https://doi.org/10.1257/app.3.2.1). URL: <https://doi.org/10.1257/app.3.2.1> (citado nas pgs. 6, 9).
- [MACHADO e SZERMAN 2021] Cecilia MACHADO e Christiane SZERMAN. “Centralized college admissions and student composition”. Em: *Economics of Education Review* 85 (2021), pg. 102184. ISSN: 0272-7757. DOI: <https://doi.org/10.1016/j.econedurev.2021.102184>. URL: <https://www.sciencedirect.com/science/article/pii/S027277572100100X> (citado na pg. 10).
- [MANSKI 1993] Charles F. MANSKI. “Identification of endogenous social effects: the reflection problem”. Em: *The Review of Economic Studies* 60.3 (1993), pgs. 531–542. ISSN: 00346527, 1467937X. URL: <http://www.jstor.org/stable/2298123> (acesso em 24/10/2023) (citado na pg. 5).
- [MEISTER 1991] “Chapter 6 - idiosyncratic variables”. Em: *Psychology of System Design*. Ed. por David MEISTER. Vol. 17. Advances in Human Factors/Ergonomics. Elsevier, 1991, pgs. 245–265. DOI: <https://doi.org/10.1016/B978-0-444-88378-0.50011-4>. URL: <https://www.sciencedirect.com/science/article/pii/B9780444883780500114> (citado na pg. 5).

- [MELLO 2022] Ursula MELLO. “Centralized admissions, affirmative action, and access of low-income students to higher education”. Em: *American Economic Journal: Economic Policy* 14.3 (ago. de 2022), pgs. 166–197. DOI: [10.1257/pol.20190639](https://doi.org/10.1257/pol.20190639). URL: <https://doi.org/10.1257/pol.20190639> (citado na pg. 10).
- [OTERO *et al.* 2021] Sebastián OTERO, Nano BARAHONA e Cauê DOBBIN. “Affirmative action in centralized college admission systems: evidence from brazil”. Em: (nov. de 2021). URL: <https://siepr.stanford.edu/publications/working-paper/affirmative-action-centralized-college-admission-systems-evidence-brazil> (citado na pg. 10).
- [SACERDOTE 2014] Bruce SACERDOTE. “Experimental and quasi-experimental analysis of peer effects: two steps forward?” Em: *Annual Review of Economics* 6.1 (ago. de 2014), pgs. 253–272. DOI: [10.1146/annurev-economics-071813-104217](https://doi.org/10.1146/annurev-economics-071813-104217). URL: <https://doi.org/10.1146/annurev-economics-071813-104217> (citado na pg. 5).
- [SCHNEEWEIS e ZWEIMÜLLER 2012] Nicole SCHNEEWEIS e Martina ZWEIMÜLLER. “Girls, girls, girls: gender composition and female school choice”. Em: *Economics of Education Review* 31.4 (ago. de 2012), pgs. 482–500. DOI: [10.1016/j.econedurev.2011.11.002](https://doi.org/10.1016/j.econedurev.2011.11.002). URL: <https://doi.org/10.1016/j.econedurev.2011.11.002> (citado nas pgs. 6, 9).
- [SCHØNE *et al.* 2019] Pål SCHØNE, Kristine von SIMSON e Marte STRØM. “Peer gender and educational choices”. Em: *Empirical Economics* 59.4 (abr. de 2019), pgs. 1763–1797. DOI: [10.1007/s00181-019-01697-2](https://doi.org/10.1007/s00181-019-01697-2). URL: <https://doi.org/10.1007/s00181-019-01697-2> (citado nas pgs. 6, 9).
- [SHAHID KAZI e AKHLAQ 2017] Asma SHAHID KAZI e Abeeda AKHLAQ. “Factors affecting students’ career choice”. Em: *Journal of Research and Reflections in Education* 11 (dez. de 2017), pgs. 187–196 (citado na pg. 1).
- [SILVA BORGES 2021] Bruna Pugialli da SILVA BORGES. “Gender in higher education”. Tese de dout. 2021. DOI: [10.11606/t.12.2021.tde-27052021-215611](https://doi.org/10.11606/t.12.2021.tde-27052021-215611). URL: <https://doi.org/10.11606/t.12.2021.tde-27052021-215611> (citado nas pgs. viii, 6, 7, 10, 11).
- [SONG e CHUNG 2010] Jae W. SONG e Kevin C. CHUNG. “Observational studies: cohort and case-control studies”. Em: *Plastic and Reconstructive Surgery* 126.6 (dez. de 2010), pgs. 2234–2242. DOI: [10.1097/prs.0b013e3181f44abc](https://doi.org/10.1097/prs.0b013e3181f44abc). URL: <https://doi.org/10.1097%2Fprs.0b013e3181f44abc> (citado na pg. 3).
- [TANG *et al.* 2008] Mei TANG, Wei PAN e Mark D. NEWMAYER. “Factors influencing high school students’ career aspirations”. Em: *Professional School Counseling* 11.5 (jun. de 2008), pg. 2156759X0801100. DOI: [10.1177/2156759x0801100502](https://doi.org/10.1177/2156759x0801100502). URL: <https://doi.org/10.1177/2156759x0801100502> (citado na pg. 5).
- [UNESCO 2006] UNESCO. *Juventude e contemporaneidade: possibilidades e limites*. pt. 2006. URL: <https://unesdoc.unesco.org/ark:/48223/pf0000154569> (citado na pg. 1).

- [VENEZIA e JAEGER 2013] Andrea VENEZIA e Laura JAEGER. “Transitions from high school to college”. Em: *The Future of Children* 23.1 (2013), pgs. 117–136. ISSN: 10548289, 15501558. URL: <http://www.jstor.org/stable/23409491> (acesso em 20/10/2023) (citado na pg. 1).
- [ZANGIROLAMI-RAIMUNDO *et al.* 2018] Juliana ZANGIROLAMI-RAIMUNDO, Jorge De Oliveira ECHEIMBERG e Claudio LEONE. “Research methodology topics: cross-sectional studies”. Em: *Journal of Human Growth and Development* 28.3 (nov. de 2018), pgs. 356–360. DOI: [10.7322/jhgd.152198](https://doi.org/10.7322/jhgd.152198). URL: <https://doi.org/10.7322%2Fjhgd.152198> (citado na pg. 4).
- [ZIMMERMAN 2003] David J. ZIMMERMAN. “Peer effects in academic outcomes: evidence from a natural experiment”. Em: *Review of Economics and Statistics* 85.1 (fev. de 2003), pgs. 9–23. DOI: [10.1162/003465303762687677](https://doi.org/10.1162/003465303762687677). URL: <https://doi.org/10.1162/003465303762687677> (citado na pg. 5).

Índice Remissivo

C

Captions, *veja* Legendas

Código-fonte, *veja* Floats

E

Equações, *veja* Modo Matemático

F

Figuras, *veja* Floats

Floats

Algoritmo, *veja* Floats, Ordem

Fórmulas, *veja* Modo Matemático

I

Inglês, *veja* Língua estrangeira

P

Palavras estrangeiras, *veja* Língua es-

trangeira

R

Rodapé, notas, *veja* Notas de rodapé

S

Subcaptions, *veja* Subfiguras

Sublegendas, *veja* Subfiguras

T

Tabelas, *veja* Floats

V

Versão corrigida, *veja* Tese/Dissertação,
versões

Versão original, *veja* Tese/Dissertação,
versões