UNIVERSIDADE FEDERAL DE SERGIPE

CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA

DEPARTAMENTO DE ESTATÍSTICA E CIÊNCIAS ATUARIAIS

Marcelo Tarso de Andrade da Silva

DESIGUALDADES EDUCACIONAIS NO ENEM 2023: UMA ANÁLISE MULTIVARIADA DA ASSOCIAÇÃO ENTRE SEXO, COR/RAÇA, TIPO DE ESCOLA E DESEMPENHO EM SERGIPE.

São Cristóvão – SE

2025

Marcelo Tarso de Andrade da Silva

Desigualdades Educacionais no ENEM 2023: Uma Análise Multivariada da Associação Entre Sexo, Cor/Raça, Tipo de Escola e Desempenho em Sergipe.

Trabalho de Conclusão de Curso apresentado ao Departamento de Estatística e Ciências Atuariais da Universidade Federal de Sergipe, como parte dos requisitos para obtenção do grau de Bacharel em Estatística.

Orientadora: Profa. Dra. Eucymara Franca Nunes Santos

São Cristóvão – SE

2025

Marcelo Tarso de Andrade da Silva

Desigualdades Educacionais no ENEM 2023: Uma Análise Multivariada da Associação Entre Sexo, Cor/Raça, Tipo de Escola e Desempenho em Sergipe.

Trabalho de Conclusão de Curso apresentado ao Departamento de Estatística e Ciências Atuariais da Universidade Federal de Sergipe, como um dos pré-requisitos para obtenção do grau de Bacharel em Estatística.

Profa. Dra. Eucymara Franca Nunes Santos

Orientadora

Prof. Mr. Daniel Francisco Neyra Castaneda

1° Examinador

Prof. Dr. Esdras Adriano Barbosa dos Santos

2° Examinador

RESUMO

Este Trabalho de Conclusão de Curso investiga as desigualdades educacionais entre os participantes do ENEM 2023 em Sergipe, considerando sexo, cor/raça, tipo de escola e desempenho. Foram utilizados os microdados do INEP, compreendendo mais de 65 mil registros. Contudo, um desafio metodológico central foi a alta incidência de dados faltantes em variáveis essenciais, como o tipo de escola, exigindo a aplicação de imputação de dados (algoritmo *MissForest*) para viabilizar a análise. Ciente de que os valores imputados são estimativas, a interpretação dos resultados requer cautela. A metodologia subsequente incluiu análise descritiva, ANOVA, Teste de Tukey HSD, Teste Qui-Quadrado e Análise de Correspondência Múltipla (ACM). Os resultados revelaram disparidades significativas de desempenho entre os grupos raciais, com estudantes brancos, em geral, apresentando médias mais elevadas em todas as áreas avaliadas. Diferenças de desempenho também foram observadas entre os sexos, com homens se sobressaindo em Matemática e Ciências da Natureza, enquanto mulheres obtiveram melhores notas em Redação. Quanto ao tipo de escola, os estudantes de instituições privadas superaram os da rede pública em todas as áreas do conhecimento. A análise multivariada evidenciou que estudantes brancos e do sexo masculino estão mais associados ao ensino privado, enquanto estudantes pretos, pardos, indígenas e do sexo feminino concentram-se no ensino público. Tais achados, analisados à luz das limitações da base de dados, destacam a persistência de desigualdades estruturais no sistema educacional sergipano e indicam a urgência de políticas públicas que promovam maior equidade.

Palavras-chave: Desigualdade educacional; ENEM; Análise de Correspondência Múltipla; Imputação de dados; Equidade.

ABSTRACT

This undergraduate thesis investigates educational inequalities among the participants of the 2023 National High School Exam (ENEM) in the state of Sergipe, considering the variables of gender, race/color, type of school, and test performance. The microdata from INEP, comprising over 65,000 records, were used. However, a central methodological challenge was the high incidence of missing data in essential variables, such as school type, which required the application of data imputation (using the *MissForest* algorithm) to enable the analysis. Aware that imputed values are estimates, the interpretation of the results requires caution. The subsequent methodology included descriptive analysis, ANOVA, Tukey's HSD Test, Chi-Square Test, and Multiple Correspondence Analysis (MCA). The results revealed significant performance disparities among racial groups, with white students generally showing higher average scores in all evaluated areas. Performance differences were also observed between genders, with males outperforming in Mathematics and Natural Sciences, while females achieved better scores in the Essay. Regarding the type of school, students from private institutions outperformed those from the public system in all knowledge areas. The multivariate analysis showed that white and male students are more associated with private education, whereas black, brown, indigenous, and female students are concentrated in public education. These findings, analyzed in light of the database limitations, highlight the persistence of structural inequalities in the educational system of Sergipe and indicate the urgency for public policies that promote greater equity.

Keywords: Educational inequality; ENEM; Data imputation; Multiple Correspondence Analysis; Equity.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

A educação desempenha um papel crucial na construção de uma sociedade mais justa e igualitária, sendo um dos principais instrumentos para o desenvolvimento social e econômico. No Brasil, o Exame Nacional do Ensino Médio (ENEM) é uma ferramenta avaliativa e significativa, utilizada tanto para mensurar o desempenho dos estudantes do ensino médio quanto para fornecer o acesso ao ensino superior por meio de programas como o Sistema de Seleção Unificada (SISU), o Programa Universidade para Todos (PROUNI) e o Fundo de Financiamento Estudantil (FIES). Devido à sua abrangência e relevância, o ENEM permite uma análise detalhada do perfil dos participantes, possibilitando a identificação de desigualdades educacionais relacionadas a fatores socioeconômicos e demográficos.

Instituído em 1998, o ENEM tem como objetivo avaliar o desempenho escolar dos estudantes ao término da educação básica. Sua criação ocorreu no contexto da recente inclusão do ensino médio como etapa final da educação básica, buscando estabelecer uma abordagem conceitual e pedagógica alinhada a essa mudança. Segundo Corti (2013), os objetivos do Enem estão expressos como:

> O Enem será realizado anualmente, com o objetivo fundamental de avaliar o desempenho do aluno ao término da escolaridade básica, para aferir o desenvolvimento de competências fundamentais ao exercício pleno da cidadania. Pretende, ainda, alcançar os seguintes objetivos específicos: a) oferecer uma referência para que cada cidadão possa proceder a sua autoavaliação com vistas às suas escolhas futuras, tanto em relação ao mercado de trabalho quanto em relação à continuidade de estudos; b) estruturar uma avaliação da educação básica que sirva como modalidade alternativa ou complementar aos processos de seleção nos diferentes setores do mundo do trabalho; c) estruturar uma avaliação da educação básica que sirva como modalidade alternativa ou complementar aos exames de acesso aos cursos profissionalizantes pós-médios e ao ensino superior.

Fatores como gênero, cor ou raça e tipo de escola frequentada pelos estudantes influenciam significativamente o acesso e o desempenho educacional. Estudos indicam que as desigualdades educacionais no Brasil estão profundamente ligadas a questões estruturais da sociedade, como a distribuição desigual de renda, o racismo e a discriminação de gênero. Por exemplo, dados apresentados em audiência pública mostram que meninas e mulheres negras enfrentam maiores dificuldades no acesso à educação, desde a educação básica até o ensino superior. Compreender a associação entre essas variáveis no contexto do ENEM é essencial para revelar as dinâmicas sociais que perpetuam as desigualdades no sistema educacional.

O estado de Sergipe, localizado na região Nordeste, apresenta características sociais e econômicas que refletem as disparidades regionais do Brasil, tornando-se um cenário relevante para a investigação das desigualdades educacionais. Sergipe é o menor estado brasileiro em extensão territorial, com uma área de aproximadamente 21.910 km², e possui uma população estimada em cerca de 2,3 milhões de habitantes, conforme dados do Instituto Brasileiro de Geografia e Estatística (IBGE). Segundo o Observatório de Sergipe, em 2022, aproximadamente 140 mil sergipanos saíram da extrema pobreza, reduzindo a proporção de pessoas nessa condição para 8,9%.

Apesar dessa melhoria, as desigualdades socioeconômicas ainda são evidentes, especialmente entre as áreas urbanas e rurais, impactando diretamente o acesso e a qualidade da educação.

A segregação educacional, entendida como a distribuição desigual de estudantes entre escolas públicas e privadas, tem implicações diretas na qualidade da educação recebida, nas oportunidades de acesso ao ensino superior e na inserção no mercado de trabalho. Essa desigualdade é agravada quando se considera o recorte racial no Brasil: estudantes autodeclarados pretos e pardos estão mais frequentemente inseridos nas escolas públicas, que, em geral, apresentam piores condições de infraestrutura e menor oferta de recursos pedagógicos. Segundo Melo e Canegal (2024), escolas com maioria de alunos negros possuem estruturas significativamente inferiores em comparação àquelas frequentadas majoritariamente por alunos brancos, refletindo camadas múltiplas de desigualdade dentro do próprio sistema público. Além disso, Guimarães e Pinto (2024) destacam que esses estudantes enfrentam uma dupla desvantagem — de origem socioeconômica e de discriminação racial — que compromete seu desempenho escolar e limita suas oportunidades futuras. Essas diferenças tendem a acentuar as desigualdades sociais, reproduzindo padrões históricos de exclusão.

Diante desse contexto, a presente pesquisa propõe investigar a associação entre o sexo, cor/raça e tipo de escola dos participantes do ENEM de 2023 no estado de Sergipe. A análise busca não apenas identificar padrões de desigualdade, mas também contribuir para a construção de políticas públicas que visem à promoção da equidade educacional. O estudo se insere em um campo de investigação que busca compreender como as interseccionalidades de sexo, cor/raça e condição socioeconômica afetam o acesso e a permanência na educação, especialmente em uma sociedade marcada por profundas desigualdades sociais.

A abordagem adotada para a realização deste estudo fundamenta-se na utilização de métodos estatísticos que permitem verificar a associação entre as variáveis em análise. A Análise de

Correspondência Múltipla será empregada como ferramenta central para identificar relações entre as categorias de sexo, cor e tipo de escola. Testes confirmatórios, como a Estatística Qui-Quadrado, a Análise de Variância (ANOVA) e o Teste de Tukey, serão utilizados para validar as associações encontradas, proporcionando maior robustez aos resultados obtidos.

Espera-se que os resultados deste estudo contribuam para o entendimento das desigualdades educacionais no estado de Sergipe, evidenciando as relações entre sexo, raça e tipo de escola e oferecendo subsídios para o desenvolvimento de políticas públicas que promovam a inclusão e a equidade no sistema educacional brasileiro.

1.1 Problema

A educação é um dos pilares fundamentais para o desenvolvimento social e econômico de uma sociedade. No Brasil, o Exame Nacional do Ensino Médio (ENEM) é considerado uma das principais portas de acesso ao ensino superior, servindo como instrumento para avaliar o desempenho escolar dos estudantes ao final da educação básica. Apesar de sua importância, o acesso e o desempenho no exame podem estar relacionados a diferentes fatores socioeconômicos e demográficos. Nesse contexto, questões como sexo, cor/raça e o tipo de escola frequentada pelos participantes emergem como variáveis relevantes para compreender as desigualdades no sistema educacional brasileiro.

O estado de Sergipe, localizado na região Nordeste do Brasil, apresenta características sociais e econômicas que podem influenciar o acesso à educação e a participação no ENEM. Compreender como essas variáveis se associam pode contribuir para o desenvolvimento de políticas públicas mais inclusivas e direcionadas para a redução das desigualdades educacionais. Dessa forma, a presente pesquisa propõe investigar a associação entre sexo, cor/raça e o tipo de escola (pública ou privada) dos participantes do ENEM no ano de 2023 no estado de Sergipe, buscando identificar possíveis padrões e disparidades entre esses grupos.

1.2 Objetivo Geral

Investigar as desigualdades educacionais entre os participantes do ENEM 2023 no estado de Sergipe, com foco na associação entre sexo, cor/raça, tipo de escola e o desempenho nas diferentes áreas do conhecimento, a fim de identificar padrões sociodemográficos e suas relações com a qualidade da educação.

1.3 Objetivos Específicos

Analisar a distribuição dos participantes do ENEM 2023 em Sergipe segundo sexo, cor/raça e tipo de escola (pública ou privada);

Verificar se há associação estatisticamente significativa entre as variáveis sexo, cor/raça e tipo de escola por meio da Análise de Correspondência Múltipla (ACM) e testes Qui-Quadrado;

Avaliar as diferenças de desempenho entre os grupos de sexo, cor/raça e tipo de escola nas áreas de Ciências Humanas, Linguagens, Ciências da Natureza, Matemática e Redação, utilizando ANOVA e Teste de Tukey HSD;

Identificar quais grupos sociodemográficos apresentam maiores desigualdades no desempenho educacional;

Fornecer subsídios para a formulação de políticas públicas voltadas à equidade educacional no estado de Sergipe.

1.4 Organização do Trabalho

Este trabalho está dividido em cinco partes. A primeira corresponde à introdução. A segunda apresenta a metodologia da pesquisa, descrevendo como a Análise de Correspondência Múltipla é executada, suas características e vantagens. Além disso, são abordados testes confirmatórios, como a Estatística Qui-Quadrado, a Análise de Variância (ANOVA) e o Teste de Tukey, utilizados para verificar associações estatísticas entre as variáveis estudadas. A terceira seção discute as limitações da pesquisa. A quarta seção é dedicada à apresentação e análise dos resultados. Por fim, a quinta seção reúne as considerações finais, destacando as principais conclusões e recomendações.

2. METODOLOGIA

A análise da relação entre sexo, raça e o tipo de escola dos participantes do ENEM de 2023 em Sergipe exige a aplicação de técnicas estatísticas apropriadas. Inicialmente, a estatística descritiva será utilizada para caracterizar as variáveis de forma isolada. Em seguida, técnicas multivariadas serão adotadas para investigar possíveis associações entre as variáveis. A abordagem combina métodos exploratórios, para identificar padrões, e confirmatórios, por meio de testes de hipótese, com o objetivo de verificar a significância estatística das associações encontradas. Essa estratégia busca fornecer uma análise abrangente e rigorosa da questão proposta.

2.1 Software R

O R é um ambiente computacional e uma linguagem de programação voltada para análise estatística, visualização de dados e computação gráfica. De código aberto e gratuito, o R oferece uma ampla variedade de pacotes e funções que possibilitam desde análises estatísticas básicas até modelagens complexas, com forte ênfase em reprodutibilidade e flexibilidade. É amplamente utilizado nas áreas acadêmica e científica, além de ser adotado em setores como economia, saúde, engenharia e ciência de dados (R CORE TEAM, 2025).

2.2 Dados

Os dados utilizados neste trabalho foram extraídos da base de Microdados do ENEM 2023, disponibilizada pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) por meio de seu portal oficial. Essa base possui um total de 3.933.955 registros e contém informações detalhadas, anonimizadas e de acesso público, abrangendo dados dos participantes que realizaram o exame em todo o território nacional. Após o download do arquivo compactado (formato .zip), foi realizada a leitura e extração dos dados utilizando o software R.

Na sequência, foram selecionadas e padronizadas as variáveis de interesse para a pesquisa, considerando os objetivos propostos. A base original do ENEM é composta por setenta e seis variáveis, abrangendo desde informações básicas de identificação até dados socioeconômicos e notas por área do conhecimento. Dentre essas, destacam-se variáveis relacionadas ao perfil dos participantes, ao desempenho nas provas e ao contexto educacional.

Para a presente análise, foram utilizadas especificamente as variáveis: TP_SEXO (sexo do participante), TP_COR_RACA (cor ou raça autodeclarada), TP_ESCOLA (tipo de escola em

que o participante concluiu o ensino médio), TP_ENSINO (tipo de ensino), NU_NOTA_CH (nota de Ciências Humanas), NU_NOTA_CN (nota de Ciências da Natureza), NU_NOTA_LC (nota de Linguagens e Códigos), NU_NOTA_MT (nota de Matemática) e NU_NOTA_REDACAO (nota da redação). Além disso, utilizou-se a variável SG_UF_PROVA para filtrar os dados referentes ao estado de Sergipe. Após esse processo de tratamento e estruturação, os dados foram organizados em um novo conjunto de dados contendo 65.540 observações.

O ENEM foi selecionado como fonte de dados para esta pesquisa por se tratar de uma das principais ferramentas de avaliação das habilidades e competências dos concluintes do ensino médio no Brasil. Além disso, busca avaliar o desempenho escolar em nível nacional, além de fornecer subsídios para a elaboração de políticas públicas voltadas para a melhoria da qualidade da educação.

A estrutura do exame é composta por 180 questões objetivas, distribuídas igualmente entre quatro áreas do conhecimento: Linguagens e Códigos, Ciências Humanas, Ciências da Natureza e Matemática. Cada área conta com 45 questões de múltipla escolha, com cinco alternativas, sendo apenas uma correta. Além disso, o exame inclui uma redação dissertativa-argumentativa, na qual os participantes devem desenvolver um texto sobre um tema de relevância social.

2.3 Análise Exploratória

A Análise Exploratória de Dados (AED) é uma abordagem fundamental para resumir, descrever e interpretar conjuntos de dados por meio de técnicas estatísticas descritivas. Proposta inicialmente por Tukey (1977), essa metodologia visa explorar os dados de forma intuitiva e visual, permitindo identificar padrões, tendências, discrepâncias e possíveis relações entre variáveis antes da aplicação de modelos estatísticos mais formais. Conforme destacam Cruz et al. (2014) e Mingoti (2005), a AED fornece subsídios importantes para a formulação de hipóteses, seleção de modelos e compreensão mais aprofundada do comportamento das variáveis analisadas.

A AED pode ser realizada por diferentes ferramentas, como tabelas, gráficos e medidas descritivas, que auxiliam na identificação de estruturas e inconsistências nos dados. A partir das evidências observadas, o pesquisador pode formular suposições, propor modelos e direcionar etapas posteriores da pesquisa.

As técnicas exploratórias são especialmente úteis para examinar dados amorfos (aqueles que não estão vinculados a teorias explícitas que determinem padrões esperados). Seu principal objetivo é fornecer uma visão preliminar da estrutura dos dados, facilitando a construção de hipóteses para análises mais aprofundadas (BORG; GROENEN, 2005).

Explorar dados consiste na aplicação de técnicas estatísticas que permitem identificar padrões ocultos em um banco de dados. Entre as técnicas mais utilizadas na análise exploratória está o Diagrama de Caixas, também conhecido como *Box Plot*. Esse método facilita a comparação visual entre grupos, destacando a mediana, os quartis e os valores extremos, além de evidenciar a dispersão e a simetria das observações.

A interpretação visual dos dados é uma etapa fundamental da análise exploratória, pois a apresentação apenas em tabelas pode dificultar a compreensão dos padrões existentes. O *Box Plot* se destaca por fornecer uma visão clara e objetiva das distribuições, auxiliando na formulação de suposições e na compreensão do comportamento das variáveis em estudo (WILLIAMSON et al., 1989).

O Diagrama de Caixas se destaca como uma técnica simples e eficiente para comparar a variabilidade e a mediana entre grupos de dados. Ele representa graficamente a distribuição de um conjunto de valores por meio dos quartis, permitindo visualizar o primeiro quartil (25%), o segundo quartil ou mediana (50%) e o terceiro quartil (75%), além dos limites inferior e superior. Valores que ultrapassam esses limites são classificados como outliers.

Uma característica importante dessa técnica é sua natureza não paramétrica, ou seja, ela não exige pressupostos sobre a distribuição estatística dos dados (TRIOLA, 2017; BUSSAB; MORETTIN, 2017). Além disso, o uso da mediana como medida de tendência central torna o método mais robusto, já que a mediana não sofre influência de valores discrepantes, diferentemente da média, que pode distorcer a interpretação em presença de outliers (TRIOLA, 2017).

2.4 MissForest

O *missForest* é um algoritmo de imputação de dados faltantes que utiliza florestas aleatórias para estimar valores ausentes em conjuntos de dados. Desenvolvido por Daniel J. Stekhoven e Peter Bühlmann (2012), o método é particularmente eficaz em conjuntos de dados com variáveis mistas, ou seja, que contêm tanto variáveis numéricas quanto categóricas.

2.4.1 Funcionamento do MissForest

O algoritmo opera de forma iterativa, seguindo os seguintes passos:

1. Inicialização: Os valores faltantes são inicialmente preenchidos com estimativas simples, como a média para variáveis numéricas e a moda para variáveis categóricas.

2. Treinamento das florestas aleatórias: Para cada variável com dados ausentes, uma floresta aleatória é treinada utilizando as outras variáveis como preditoras. As observações completas servem como conjunto de treinamento.

3. Imputação: As florestas aleatórias treinadas são então utilizadas para prever os valores faltantes da variável em questão.

4. Iteração: Os passos 2 e 3 são repetidos para cada variável com dados ausentes. O processo iterativo continua até que a diferença entre as imputações sucessivas seja mínima ou um número máximo de iterações seja atingido.

Uma das principais vantagens do missForest é sua capacidade de capturar relações não lineares e interações complexas entre variáveis, graças ao uso de florestas aleatórias. Além disso, o algoritmo não requer a especificação de um modelo paramétrico, tornando-o flexível para diferentes tipos de dados.

2.4.2 Desempenho e Aplicações

A escolha deste algoritmo fundamenta-se em sua robustez, amplamente documentada em estudos comparativos que atestam sua superioridade frente a outras técnicas. Em seu artigo seminal, por exemplo, Stekhoven e Bühlmann (2012) avaliaram o MissForest contra métodos populares como a imputação por k-vizinhos mais próximos (kNN) e a imputação múltipla por equações encadeadas (MICE). Os resultados demonstraram que o MissForest produziu o menor erro de imputação, uma vantagem particularmente evidente em conjuntos de dados com interações complexas e relações não-lineares, características esperadas em uma base como a do ENEM.

Essa eficácia foi corroborada em outros domínios; uma pesquisa de Waljee et al. (2013) com dados clínicos complexos, por exemplo, não apenas confirmou a maior precisão do MissForest, mas também revelou que o método foi capaz de gerar estimativas de coeficientes de regressão mais próximas dos valores reais, um fator crucial para a validade de análises inferenciais. O

trabalho de Tang e Ishwaran (2017), realizou uma análise compreensiva dos algoritmos de imputação de dados faltantes baseados em *random forests*. Em seu estudo, os autores não apenas confirmaram o alto desempenho empírico da abordagem iterativa utilizada pelo MissForest, mas também o posicionaram como um método de referência (benchmark) fundamental na área, contra o qual novas propostas são frequentemente comparadas. A pesquisa deles destaca que, apesar de ser computacionalmente mais intensivo, o MissForest se sobressai pela sua capacidade de convergir para estimativas estáveis e precisas, capturando a estrutura de dependência complexa dos dados de forma eficaz. Portanto, a adoção do MissForest representa a tentativa de aplicar a técnica mais acurada disponível para o tratamento dos dados faltantes.

2.4.3 Implementação no R

No ambiente R, o missForest está disponível através do pacote de mesmo nome, facilitando sua aplicação por pesquisadores e profissionais. A função principal, *missForest()*, permite a imputação direta de conjuntos de dados com valores faltantes, oferecendo parâmetros para ajustar o número de árvores na floresta e o número máximo de iterações.

2.5 Teste de Hipótese

A aplicação de testes de hipóteses é uma etapa fundamental em experimentos que envolvem análises estatísticas, sendo essencial para verificar a validade das suposições estabelecidas na pesquisa. Esses testes podem ser classificados em paramétricos e não paramétricos, dependendo das características dos dados e das premissas do estudo. Os testes paramétricos exigem maior rigor teórico, uma vez que pressupõem que os dados seguem uma distribuição estatística específica, como a distribuição normal. Já os testes não paramétricos, por serem mais flexíveis, não necessitam dessas suposições, o que os torna adequados para dados que não atendem aos critérios exigidos pelos testes paramétricos (BUSSAB; MORETTIN, 2017).

A avaliação da hipótese de pesquisa é realizada por meio da formulação de duas proposições: a Hipótese Nula (H_0) e a Hipótese Alternativa (H_1). A hipótese nula expressa a ausência de associação entre as variáveis analisadas, funcionando como uma espécie de pressuposto inicial que se busca rejeitar. Por outro lado, a hipótese alternativa sugere a existência de relação entre as variáveis, sendo essa, geralmente, a expectativa do pesquisador ao conduzir o estudo. A decisão sobre qual hipótese aceitar é baseada na análise estatística, que determina a probabilidade de os resultados observados ocorrerem ao acaso (FIRMINO, 2015).

2.6 Estatística Qui-Quadrado

O Teste Qui-Quadrado é amplamente utilizado na estatística inferencial para avaliar a associação entre variáveis categóricas. Por ser um método não paramétrico, não depende de parâmetros populacionais como média e variância, o que o torna adequado para a análise de variáveis qualitativas. Sua aplicação permite verificar, de forma quantitativa, se há evidências de associação entre as variáveis em uma tabela de contingência.

De acordo com Infantosi et al. (2014), a Análise de Correspondência é um método que decompõe a estatística qui-quadrado do teste de independência, permitindo descrever graficamente os dados dispostos em tabelas de contingência. Essa abordagem facilita a identificação de relações que não seriam detectadas em análises bivariadas tradicionais. Em resumo, o Teste Qui-Quadrado é uma ferramenta essencial para a análise de associações entre variáveis categóricas, especialmente quando combinado com técnicas como a Análise de Correspondência Múltipla, que enriquecem a interpretação dos dados.

A utilização do Qui-Quadrado possibilita investigar se as frequências observadas diferem significativamente das frequências esperadas sob a hipótese de independência entre as variáveis. Nesse sentido, o teste fornece subsídios para avaliar se existe dependência estatística entre as categorias analisadas. Conforme Infantosi et al. (2014), “a estatística de teste mais comum para inferir sobre a hipótese de independência (ou homogeneidade) de duas variáveis categóricas, dispostas em uma tabela de contingência é a Qui-Quadrado”.

O Teste Qui-Quadrado é indicado para situações em que os elementos da amostra estão agrupados em duas ou mais categorias, permitindo avaliar se a distribuição das frequências observadas difere significativamente das frequências esperadas. A lógica central do teste consiste em verificar se as respostas entre as diferentes categorias apresentam uma distribuição homogênea ou se há evidências de associação entre as variáveis analisadas. Em essência, o método busca identificar se as variações observadas decorrem do acaso ou se indicam um padrão significativo (FIRMINO, 2015).

O primeiro passo na aplicação do teste Qui-Quadrado é a definição das hipóteses da pesquisa. A hipótese nula (H_0) assume que as distribuições das categorias analisadas são homogêneas, ou seja, não há associação entre as variáveis. Por outro lado, a hipótese alternativa (H_1) sugere que existem diferenças significativas entre as categorias, indicando uma possível relação entre as variáveis estudadas.

O procedimento do teste consiste em comparar as frequências observadas com as frequências esperadas, que são calculadas com base em uma distribuição hipotética sob a suposição de que a hipótese nula é verdadeira. A estatística do teste é determinada a partir dessas diferenças, fornecendo uma medida quantitativa para avaliar se as variações nas frequências são suficientemente grandes para rejeitar a hipótese nula (LARSON et al., 2009).

Definimos a frequência esperada como:

$$E_i = np_i, \tag{2.1}$$

em que n é o tamanho do conjunto de dados e p_i a probabilidade da i-ésima categoria.

A estatística do teste segue uma distribuição Qui-Quadrado com $k - 1$ graus de liberdade, onde J representa o número total de categorias analisadas. A expressão matemática que descreve essa estatística é dada por:

$$\chi^2 = \sum_{i,j=1}^{I,J} \frac{(n_{ij} - E_{ij})^2}{E_{ij}} \tag{2.2}$$

em que n_{ij} e E_{ij} são os valores observados e esperados. Correspondem, respectivamente, às frequências empíricas e às frequências teóricas de cada célula da tabela de contingência. Sob a suposição de independência entre as variáveis, o cálculo do valor esperado para cada célula é obtido pelo produto da probabilidade marginal de ocorrência das categorias de cada variável, dividido pelo total de elementos da amostra. A expressão para determinar o valor esperado é dada por:

$$E_{ij} = \frac{n_{i+} \cdot n_{+j}}{n_{++}} \tag{2.3}$$

Ao aplicar a razão entre a estatística Qui-Quadrado e o total de observações da tabela de contingência, obtém-se uma medida padronizada que quantifica a dispersão relativa entre as frequências observadas e esperadas. Essa relação é expressa pela seguinte fórmula:

$$\rho^2 = \frac{\chi^2}{n_{++}} \tag{2.4}$$

O valor obtido a partir dessa razão é conhecido como Coeficiente de Contingência de Pearson. Essa medida expressa o grau de associação entre as variáveis categóricas analisadas, com base nas diferenças entre os valores observados (n_{ij}) e os valores esperados (E_{ij}). Quanto maior o

coeficiente, maior é a dispersão entre as frequências observadas e esperadas, indicando uma associação mais intensa entre as variáveis.

O coeficiente é calculado pela seguinte fórmula:

$$c = \sqrt{\frac{\chi^2}{\chi^2 + T}} \tag{2.5}$$

Onde:

- C é o coeficiente de contingência;
- $\chi 2$ é a estatística Qui-Quadrado calculada;
- T representa o total de observações.

O coeficiente de contingência de Pearson varia entre 0 e 1, sendo que valores próximos de 0 indicam baixa associação, enquanto valores próximos de 1 sugerem uma associação mais forte entre as variáveis. Essa medida é amplamente utilizada por sua simplicidade e capacidade de fornecer uma interpretação intuitiva sobre a força da relação entre variáveis qualitativas.

2.7 Distribuição Qui-Quadrado

A distribuição Qui-Quadrado pertence ao grupo das distribuições contínuas e é definida pelo número de graus de liberdade e, em alguns casos, pelo parâmetro de não centralidade. Uma propriedade marcante dessa distribuição é sua assimetria positiva, especialmente quando há poucos graus de liberdade. Conforme o número de graus de liberdade aumenta, a distribuição torna-se progressivamente mais simétrica, aproximando-se da distribuição Normal, devido ao Teorema Central do Limite (TRIOLA, 2017).

Essa característica torna a distribuição Qui-Quadrado adequada para testar hipóteses relacionadas à variância, tabelas de contingência e ajustes de modelos estatísticos, desempenhando papel central na análise inferencial de dados categóricos. Sejam Z_1, Z_2,..., Z_n variáveis aleatórias independentes, cada uma seguindo uma distribuição normal padrão com média 0 e variância 1. A distribuição Qui-Quadrado surge a partir da soma dos quadrados dessas variáveis aleatórias, sendo expressa por:

$$\chi^2 = \sum_{i=1}^{n} z_i^2 \tag{2.6}$$

Essa soma resulta em uma variável aleatória que segue uma distribuição Qui-Quadrado com n graus de liberdade, onde n representa o número de variáveis somadas. Cada termo ao quadrado contribui para a dispersão total, e os graus de liberdade refletem a quantidade de componentes independentes envolvidos na soma.

Essa propriedade fundamenta o uso da distribuição Qui-Quadrado em testes estatísticos, como o teste de independência em tabelas de contingência e a análise de variâncias, tornando-se uma ferramenta essencial na inferência estatística (JOHNSON; WICHERN, 2007).
A função densidade com q graus de liberdade é dada por:

$$f(x;q) = \frac{1}{2^{q/2}\,\Gamma(q/2)}\, x^{(q/2-1)}\, e^{-x/2}, \quad \text{para } x > 0 \tag{2.7}$$

onde q é o número de graus de liberdade; Γ é a função gama, que generaliza o fatorial: $\Gamma(n) = (n-1)!$ para inteiros positivos; e e é a base do logaritmo natural.

Como todos os valores da distribuição Qui-Quadrado resultam da soma de quadrados de variáveis normais padrão, seus valores são sempre positivos e reais. Os dois primeiros momentos, a média e a variância, estão diretamente relacionados aos graus de liberdade da distribuição.

A média da distribuição Qui-Quadrado é dada por:

$$\mu = E(X) = q \tag{2.8}$$

e

$$\sigma^2 = Var(X) = 2q \tag{2.9}$$

Essas expressões indicam que, à medida que os graus de liberdade aumentam, a média e a variância também aumentam, tornando a distribuição mais simétrica e menos assimétrica à direita.

A principal aplicação da distribuição Qui-Quadrado está nos testes de hipóteses, devido à sua relação direta com a distribuição Normal Padronizada. Essa distribuição é especialmente utilizada para avaliar a independência entre variáveis categóricas e a qualidade do ajuste em modelos estatísticos.

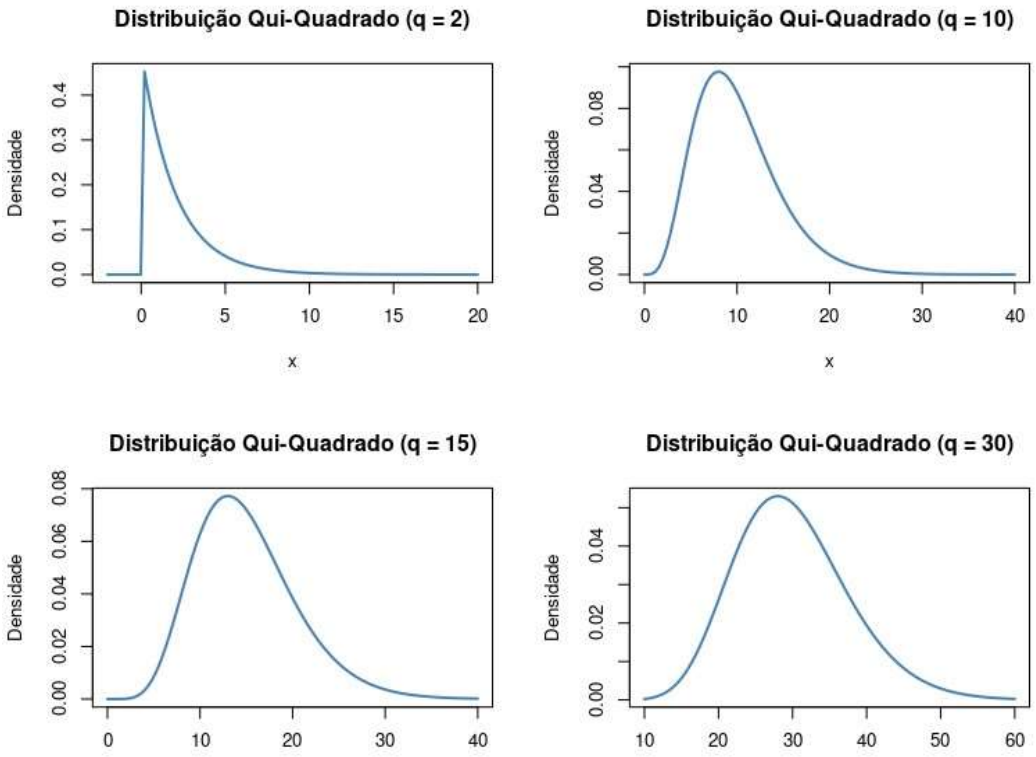

Figura 1: Densidades da distribuição Qui-Quadrado.

Os valores da distribuição Qui-Quadrado variam de zero ao infinito, sem assumir valores negativos, já que derivam da soma de quadrados de variáveis aleatórias normais padrão. Quanto maior o número de graus de liberdade (q), mais simétrica a distribuição se torna, aproximando-se da distribuição normal. A Figura 1 ilustra exemplos de diferentes distribuições Qui-Quadrado geradas com distintos graus de liberdade, evidenciando como a forma da função densidade se modifica à medida que $\boldsymbol{q}$ aumenta. Para valores baixos de graus de liberdade, a curva apresenta assimetria positiva, enquanto para valores mais altos, a distribuição se torna mais simétrica e concentrada em torno da média (BUSSAB; MORETTIN, 2017).

O valor calculado pela estatística Qui-Quadrado, conforme apresentado na equação (2.2), expressa o grau de discrepância entre as frequências observadas e esperadas em um conjunto de dados. Esse valor é utilizado para testar hipóteses, fornecendo evidências que apoiam ou rejeitam a hipótese nula (H_0)**.**

Quando o valor obtido pela estatística Qui-Quadrado é baixo, significa que as frequências observadas estão próximas das esperadas, indicando pouca discrepância entre os grupos analisados. Nesse cenário, não há motivos para rejeitar a hipótese nula.

Por outro lado, se o valor da estatística Qui-Quadrado for alto, revela-se uma diferença significativa entre os valores observados e esperados, sugerindo evidências contra a hipótese nula, o que pode levar à sua rejeição.

A Figura 2 ilustra a distribuição Qui-Quadrado com 10 graus de liberdade, destacando as regiões de rejeição situadas na extremidade direita da curva, onde valores maiores que o valor crítico indicam a rejeição da hipótese nula a um nível de significância pré-estabelecido (MONTGOMERY; RUNGER, 2010).

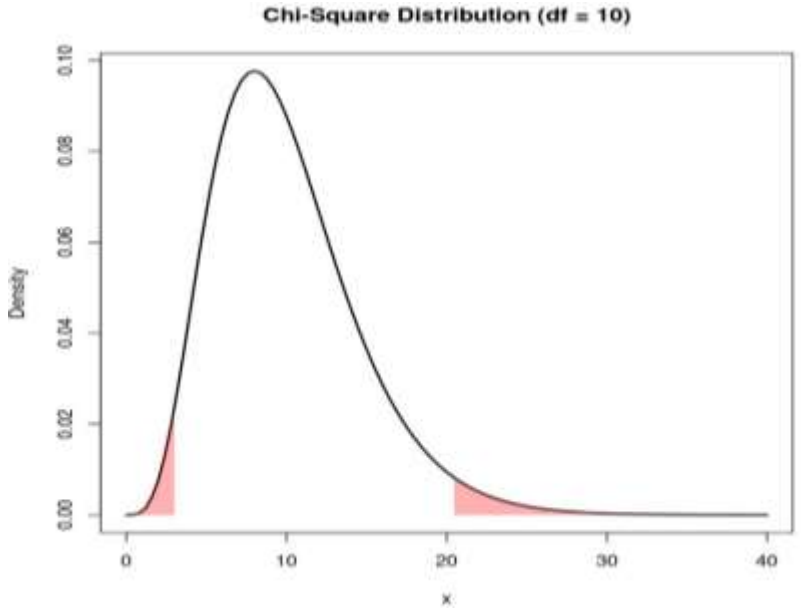

Figura 2: Região de rejeição Qui-Quadrado.

2.8 ANOVA

A Análise de Variância (ANOVA) é uma técnica estatística amplamente utilizada para comparar as médias de dois ou mais grupos e verificar se existem diferenças significativas entre eles. Esse método é especialmente importante quando a análise envolve mais de duas categorias, pois permite avaliar simultaneamente as diferenças entre múltiplos grupos, evitando o aumento da probabilidade de erro associada a comparações múltiplas.

A ANOVA parte da formulação de duas hipóteses:

- Hipótese Nula (H_0): não há diferença significativa entre as médias dos grupos;
- Hipótese Alternativa (H_1): pelo menos uma média de grupo é diferente das demais.

Segundo Bussab e Morettin (2011), o teste ANOVA baseia-se na decomposição da variabilidade total em duas componentes: a variabilidade entre os grupos, que mede as

diferenças entre as médias, e a variabilidade dentro dos grupos, que avalia a dispersão dos dados em torno das médias de cada grupo.

O coeficiente F da ANOVA é calculado como a razão entre a variabilidade entre os grupos e a variabilidade dentro dos grupos, expressa pela seguinte fórmula:

$$F = \frac{MST}{MSE} \tag{2.10}$$

onde, MST é a Média da Soma ao Quadrado do Tratamento; e MSE é a Média da Soma de Quadrados do Erro.

A aplicação da ANOVA requer algumas suposições fundamentais para garantir a validade dos resultados. São elas:

1. Independência das observações: os dados das amostras devem ter sido coletados de maneira independente, sem influência mútua entre as observações;
2. Homogeneidade das variâncias (homocedasticidade): as variâncias populacionais dos grupos devem ser aproximadamente iguais;
3. Normalidade dos dados: cada amostra deve ser proveniente de uma distribuição aproximadamente Normal (LARSON, 2008).

O resultado da estatística apresentada em (2.10) é conhecido como Razão F. Esse valor quantifica a relação entre a variabilidade entre os grupos e a variabilidade dentro dos grupos. Se a hipótese nula for verdadeira, ou seja, se não houver diferença significativa entre as médias dos grupos, espera-se que a Razão F seja próxima de 1.

Por outro lado, se existir alguma diferença significativa entre as médias, a Razão F tenderá a assumir valores elevados. A significância estatística dessa diferença é avaliada através do p-valor. Valores de $p < 0,05$ indicam evidências para rejeitar a hipótese nula, sugerindo que ao menos um grupo difere significativamente dos demais (CONNELLY, 2021).

Apesar de a ANOVA identificar a existência de diferenças entre os grupos, o teste não revela quais grupos específicos diferem entre si. Para isso, é necessário realizar um procedimento post hoc, que permite comparar individualmente as médias dos grupos e identificar quais pares apresentam diferenças estatisticamente significativas. Existem diversas técnicas post hoc disponíveis, e este trabalho utilizará o Teste de Tukey.

2.9 Teste de Tukey

Quando um pesquisador busca identificar quais grupos diferem significativamente após a realização da ANOVA, é necessário aplicar um teste de comparação múltipla. Entre os métodos mais utilizados destaca-se o Teste de Tukey (SMITH, 1971). Esse método é apropriado para comparar todas as combinações possíveis de pares de médias, controlando a probabilidade de erro tipo I ao longo das múltiplas comparações.

O Teste de Tukey avalia a diferença entre as médias dos grupos através da estatística *q* de Tukey, expressa por:

$$q = \frac{\bar{x_i} - \overline{x_J}}{\sqrt{\frac{MSE}{n}}} \tag{2.11}$$

Onde: $\bar{x_i}$ e χ representam as médias dos grupos i e j; MSE é o erro quadrático médio (*Mean Square Error*) obtido pela ANOVA; e *n* é o número de observações por grupo.

A estatística *q* é comparada com o valor crítico da distribuição de Tukey para um nível de significância pré-definido, geralmente $\alpha = 0{,}05$. Quando q observado ultrapassa o valor crítico, conclui-se que há uma diferença significativa entre as médias dos grupos.

O Teste de Tukey é amplamente recomendado por sua capacidade de controlar o erro tipo I, garantindo que a probabilidade de cometer erros em múltiplas comparações não exceda o nível de significância estabelecido (MONTGOMERY, 2017).

2.10 Análise de Correspondência Múltipla

A técnica de Análise de Correspondência (AC) começou a ganhar notoriedade a partir de 1933, com a publicação de diversos estudos voltados à análise multivariada de dados categóricos. No entanto, foi apenas na década de 1960 que Jean-Paul Benzécri propôs uma nova abordagem para o tratamento de tabelas de contingência, especialmente no campo da linguística, ampliando significativamente o alcance e a aplicabilidade da técnica (MINGOTI, 2005).

A partir da década de 1970, a Análise de Correspondência passou a se popularizar, sendo adotada em diversas áreas do conhecimento, como as Ciências Sociais, Biológicas e da Saúde.

Essa difusão deve-se, em grande parte, à facilidade de aplicação da técnica e à possibilidade de interpretação gráfica dos resultados, o que torna a análise mais acessível e visualmente compreensível (CZERMAINSKI, 2004).

A Análise de Correspondência é uma técnica estatística multivariada exploratória voltada ao estudo de associações entre variáveis categóricas, a partir da transformação de uma matriz de dados não negativos (geralmente uma tabela de contingência) em um gráfico bidimensional. Nessa representação, tanto as linhas quanto as colunas da matriz são convertidas em pontos em um espaço de menor dimensionalidade, permitindo observar associações entre categorias das variáveis envolvidas (GREENACRE; HASTIE, apud PAMPLONA, 1998).

A representação gráfica resultante da AC facilita a visualização das relações entre as variáveis, indicando não apenas a associação entre categorias, mas também sua intensidade. Cada ponto no gráfico representa uma categoria, e as distâncias entre esses pontos refletem o grau de proximidade ou associação entre elas.

A Análise de Correspondência Múltipla (ACM), por sua vez, pode ser entendida como uma extensão natural da AC. Enquanto a AC é aplicada a tabelas bidimensionais (duas variáveis), a ACM se destina ao estudo de associações entre três ou mais variáveis categóricas. Sua base metodológica é semelhante à da AC, mas a estrutura dos dados é mais complexa, sendo operacionalizada por meio da Matriz Indicadora e da Matriz de Burt (NASCIMENTO, 2011).

De acordo com Pamplona, na ACM as tabelas “são compostas de linhas por indivíduos e colunas por características observadas, onde cada linha contém todos os códigos correspondentes às modalidades atribuídas a um indivíduo ou elemento para cada uma das características observadas”. A Matriz Indicadora (Z) é formada por variáveis codificadas como dummies, assumindo valores 1 ou 0 conforme a presença ou ausência da categoria.

A Matriz de Burt, por sua vez, é obtida a partir do produto entre a transposta da Matriz Indicadora e ela própria. Essa matriz é composta por todas as tabulações cruzadas possíveis entre as variáveis categóricas, sendo uma matriz simétrica, cuja diagonal principal é formada pelas frequências marginais de cada variável (PAMPLONAWWW, 1998).

Portanto, a ACM constitui uma importante ferramenta na análise exploratória de dados categóricos multidimensionais, oferecendo uma forma gráfica de sintetizar e interpretar relações entre múltiplas variáveis.

2.10.1 Matriz Indicadora

Uma Matriz Indicadora representa um banco de dados com N observações e Q variáveis categóricas (Q > 2). Cada variável Q_q (q = 1, 2, ..., Q) possui J_q categorias distintas. Assim, o número total de categorias envolvidas em uma Análise de Correspondência Múltipla (ACM) corresponde à soma das categorias de todas as variáveis, ou seja:

$$J = \sum_{q=1}^{Q} J_q \tag{2.12}$$

Onde:

- Q_q representa a q-ésima variável categórica, sendo q = 1, 2, ..., Q_q, com Q indicando o número total de variáveis envolvidas na análise;
- J_q corresponde ao número de categorias distintas da variável Q_q;
- N_{qj} representa o número de indivíduos que escolheram a categoria j da variável Q_q, com q = 1, 2, ..., Qq e j = 1, 2, ..., J_q;
- J é o número total de categorias presentes no conjunto de variáveis, obtido pela soma das categorias de todas as variáveis.

Na Análise de Correspondência Múltipla, a visualização gráfica dos dados é representada por duas nuvens de pontos no mapa de correspondência: uma composta pelos indivíduos (linhas da matriz) e outra pelas categorias das variáveis (colunas). A interpretação dessas nuvens baseia-se na proximidade ou distanciamento entre os pontos, o que pode indicar associação (ou ausência dela) entre as categorias analisadas.

O principal objetivo da ACM é reduzir o espaço original de alta dimensionalidade para um subespaço ótimo, facilitando a visualização e análise das relações entre categorias. A dimensionalidade máxima da nuvem de pontos, ou seja, o número de coordenadas geradas pela ACM, é dada por:

$$L < J - Q \tag{2.13}$$

onde:

- L representa o número de dimensões ou eixos fatoriais gerados na análise;
- J é o número total de categorias;

- Q é o número de variáveis categóricas.

Por exemplo, considerando um conjunto de dados com 10 categorias distribuídas em 3 variáveis, a dimensionalidade será:

$$L < 10 – 3 = 7 \; L < 10 - 3 = 7 \; L < 10 – 3 = 7 \tag{2.14}$$

Ou seja, no máximo sete dimensões podem ser obtidas com a aplicação da ACM nesse caso.

Por fim, destaca-se que a ACM se baseia nos mesmos fundamentos algorítmicos da Análise de Correspondência Simples (AC), sendo possível aplicá-la a tabelas de contingência de múltiplas entradas, como a Matriz Indicadora ou a Matriz de Burt. Essas estruturas viabilizam a análise de relações entre várias variáveis categóricas simultaneamente.

2.10.2 Matriz de Burt

Uma alternativa para a análise de correspondência em tabelas multidimensionais é o uso da matriz de Burt. Segundo Greenacre (2008), a matriz de Burt é representada por

$$B = Z^T Z \tag{2.15}$$

sendo uma matriz simétrica e quadrada, organizada por categorias. Essa matriz é composta por tabelas de contingência entre pares de variáveis, incluindo, na diagonal principal, a tabulação cruzada de cada variável com ela mesma.

Por ser simétrica, a análise de correspondência aplicada às linhas da matriz de Burt gera os mesmos resultados obtidos na análise das colunas. Cada linha e cada coluna da matriz correspondem a uma categoria de uma variável categórica. As interseções entre linhas e colunas representam as frequências observadas simultaneamente para os pares de variáveis consideradas.

A estrutura da matriz de Burt apresenta, no triângulo inferior, as tabulações cruzadas de cada variável com as demais, enquanto o triângulo superior contém suas contrapartes transpostas. Já na diagonal principal estão dispostas as frequências marginais de cada categoria, ou seja, o total de observações para cada uma delas.

Segundo Greenacre (2008, p. 190), destacam-se as seguintes propriedades da análise de correspondência baseada na matriz de Burt, especialmente em comparação com a matriz indicadora:

- A análise de correspondência realizada sobre a matriz de Burt produz coordenadas idênticas às obtidas a partir da matriz indicadora;
- As coordenadas principais resultam da multiplicação das coordenadas padrão pela raiz quadrada das respectivas inércias principais;
- As porcentagens de inércia principal derivadas da matriz de Burt correspondem ao quadrado das inércias obtidas com a matriz indicadora.

2.10.3 Procedimento para realização da ACM no R

Inicialmente, é necessário garantir que os dados estejam organizados em um objeto do tipo *data.frame*, composto exclusivamente por variáveis categóricas. Cada linha do conjunto de dados representa um indivíduo ou unidade de observação, enquanto cada coluna representa uma variável qualitativa. É essencial que essas variáveis estejam devidamente codificadas como fatores no R, o que pode ser feito manualmente ou com o auxílio de funções como *factor*().

O procedimento analítico começa com a instalação e o carregamento dos pacotes *FactoMineR* e *factoextra*. O primeiro é responsável pela realização da análise propriamente dita, enquanto o segundo facilita a visualização gráfica dos resultados. Após o preparo dos dados, a função MCA() é utilizada para executar a análise. Essa função recebe como argumento principal o *data frame* contendo as variáveis categóricas e retorna um objeto com os resultados da decomposição fatorial. É possível desabilitar a geração automática de gráficos ao configurar o argumento *graph = FALSE*.

Internamente, a função MCA() do pacote *FactoMineR* utiliza como base de cálculo a matriz indicadora completa, também conhecida como matriz disjuntiva completa, e não a matriz de Burt. A matriz indicadora é uma matriz binária onde cada linha representa um indivíduo e cada coluna representa uma categoria de uma das variáveis. Os elementos dessa matriz assumem o valor 1 quando o indivíduo pertence à respectiva categoria e 0 caso contrário. Essa matriz é a base para o cálculo das distâncias do tipo qui-quadrado entre perfis e, posteriormente, para a decomposição espectral que gera as coordenadas fatoriais das categorias e indivíduos.

Embora a matriz de Burt também possa ser utilizada para a realização da ACM, ela não é o padrão adotado pela função MCA() do FactoMineR. A matriz de Burt é construída a partir do produto matricial da transposta da matriz indicadora por ela mesma (equação 2.15), resultando em uma matriz simétrica que contém, nas diagonais, as frequências marginais de cada categoria e, fora da diagonal, as tabulações cruzadas entre todas as variáveis. A literatura especializada, como Greenacre (2008), demonstra que a ACM baseada na matriz de Burt produz as mesmas coordenadas principais da análise realizada com a matriz indicadora; no entanto, as inércias extraídas da matriz de Burt correspondem ao quadrado das inércias da matriz indicadora.

Após a execução da ACM com MCA(), torna-se fundamental interpretar as dimensões fatoriais extraídas. Cada dimensão representa um eixo de variação que sintetiza, de forma reduzida, a estrutura de associação entre as categorias das variáveis analisadas. Para visualizar essas dimensões, pode-se utilizar a função *fviz_screeplot*(), que gera um gráfico de barras com a porcentagem da variância explicada por cada eixo. Dimensões com maiores valores de inércia são consideradas mais relevantes para a interpretação dos dados.

A visualização gráfica dos resultados da ACM é uma das etapas mais informativas da análise. O gráfico biplot, gerado por meio da função *fviz_mca_biplot*(), permite observar simultaneamente a distribuição dos indivíduos e das categorias no espaço fatorial. Nesse gráfico, categorias posicionadas próximas indicam maior similaridade ou associação. Além disso, é possível identificar agrupamentos de categorias que compartilham padrões semelhantes de resposta. Essa visualização auxilia a interpretar as principais associações e oposições entre os grupos estudados.

3. LIMITAÇÕES DA PESQUISA

Um dos principais desafios metodológicos deste trabalho está relacionado à qualidade da base de dados utilizada. Os microdados do ENEM 2023, disponibilizados pelo INEP, representam a fonte mais abrangente de informações sobre os participantes do exame, mas apresentam falhas significativas em variáveis fundamentais para a compreensão das desigualdades educacionais no Brasil.

Entre essas variáveis, destacam-se TP_COR_RACA (autodeclaração de cor/raça), TP_ENSINO (tipo de instituição que o participante concluiu ou concluirá o Ensino Médio) e TP_ESCOLA (tipo de escola do ensino médio). Essas variáveis, centrais para a investigação proposta neste trabalho, apresentaram proporções distintas de dados faltantes, sendo a situação mais grave aquela verificada na variável TP_ESCOLA, com mais de 70% de registros ausentes.

Diante desse quadro, a estratégia adotada foi a imputação dos dados faltantes exclusivamente para as variáveis TP_COR_RACA, TP_ENSINO e TP_ESCOLA, com o objetivo de viabilizar a aplicação da Análise de Correspondência Múltipla (MCA). Para isso, utilizou-se o algoritmo MissForest, que combina Random Forests e imputação iterativa. Embora esse método seja reconhecido por sua robustez (Stekhoven; Bühlmann, 2012), é fundamental destacar que os valores imputados são estimativas probabilísticas e não observações reais.

A decisão de imputar uma variável com mais de 70% de dados ausentes, como TP_ESCOLA, representa uma limitação metodológica severa. A literatura sobre o tema é cautelosa ao definir um percentual máximo de dados faltantes que possam ser imputados sem causar prejuízos à análise. Embora "regras de bolso" frequentemente sugiram que uma ausência de 5% a 10% seja manejável, não há um consenso sobre um limite universalmente seguro. Autores como Rubin (1987) e Schafer (1999) argumentam que o impacto depende mais do mecanismo que gera a ausência dos dados e da qualidade do método de imputação do que de um percentual fixo. A imputação em larga escala, como a realizada neste estudo, eleva consideravelmente o risco de introduzir vieses, subestimar a variância real dos dados — gerando uma falsa sensação de precisão estatística — e distorcer as correlações entre as variáveis. Portanto, o risco de que as estruturas identificadas pela MCA reflitam mais os padrões do modelo de imputação do que a realidade subjacente é considerável. Assim, todas as conclusões derivadas da MCA devem ser interpretadas com cautela, pois refletem tanto a realidade empírica quanto as premissas do modelo de imputação.

Esse problema não é exclusivo do presente estudo. Pesquisas anteriores já identificaram a baixa completude dos microdados do ENEM como uma limitação recorrente. No que se refere à qualidade das informações, diversos estudos apontaram inconsistências e lacunas nos microdados do ENEM (BRIEGA, 2017; MELLO NETO et al., 2014; SANTOS, 2019; SILVA; MELLETI, 2014). Por exemplo, Silva e Meletti (2014), ao analisar dados limitados a um município, destacaram a diminuição do escopo inicial da pesquisa devido às dificuldades encontradas no uso dos microdados, como a descontinuidade das informações entre edições das provas, a ausência de um identificador único para os estudantes e inconsistências nas variáveis registradas. Soares e Alves destacam que os dados disponíveis não captam integralmente as desigualdades raciais, restringindo a compreensão das barreiras estruturais presentes no sistema educacional brasileiro.

Em outro estudo, Mello Neto et al. (2014) relataram dificuldades nas análises de renda familiar, cujas categorias estavam "[...] definidas por faixa de renda arbitrária, variando de ano para ano" (p. 116), prejudicando comparações consistentes ao longo do tempo. Além disso, Junqueira, Martins e Lacerda (2017) observaram que o processo de coleta de informações no momento da inscrição do exame impacta diretamente a qualidade dos microdados, especialmente no registro de tipo de deficiência ou de recursos de acessibilidade solicitados pelos participantes.

Essa fragilidade ultrapassa o âmbito metodológico e assume contornos institucionais. Variáveis como cor/raça e tipo de escola são fundamentais para o monitoramento das políticas de ação afirmativa e de equidade no ensino superior brasileiro. A ausência de tais informações em larga escala compromete a capacidade de avaliar o impacto das cotas raciais e sociais, criando uma "zona cinzenta" que enfraquece a transparência e dificulta a formulação de políticas públicas baseadas em evidências.

Portanto, o presente estudo, assim como outros que utilizam o ENEM como fonte de dados, assumiu o ônus de aplicar técnicas de imputação para possibilitar a análise pretendida. Recomenda-se que, em edições futuras do exame, o INEP aperfeiçoe seus mecanismos de coleta e integração de dados, de modo a reduzir a incidência de não respostas e garantir maior fidedignidade às informações disponibilizadas. Além disso, futuros trabalhos poderiam validar os achados aqui apresentados a partir de outras bases (como o Censo Escolar) ou em edições subsequentes do ENEM, caso a qualidade da coleta de dados seja aprimorada.

4. RESULTADOS E DISCUSSÃO

4.1 Análise Exploratória

O Exame Nacional do Ensino Médio é uma avaliação utilizada como forma de admissão ao ensino superior. Foi criado em 1998 com o objetivo inicial de avaliar a qualidade do ensino médio no Brasil. O exame é aplicado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), autarquia federal vinculada ao Ministério da Educação (MEC).

Para esta pesquisa, é importante destacar que estão sendo considerados todos os participantes inscritos no exame no estado de Sergipe. De acordo com a base de dados, aproximadamente 62% dos inscritos se identificaram com o sexo feminino e 38% com o sexo masculino (figura 3).

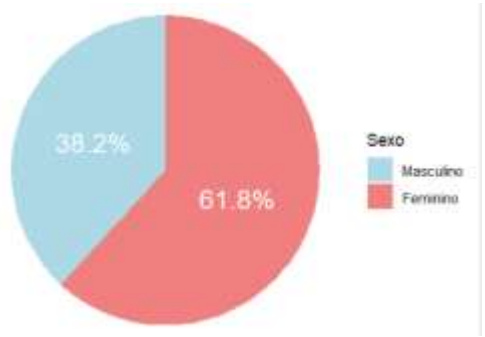

Figura 3: Distribuição dos Participantes do ENEM 2023 em Sergipe por Sexo.

A variável cor ou raça, com base nos dados analisados nesta pesquisa, revela importantes aspectos sociodemográficos dos participantes do ENEM no estado de Sergipe em 2023. Os resultados da tabela 1 indicam que a maioria dos inscritos se autodeclarou parda, o que está em consonância com a composição étnico-racial predominante na região Nordeste, segundo dados do Instituto Brasileiro de Geografia e Estatística (IBGE, 2022). Aproximadamente 25% dos participantes declararam-se brancos, ao passo que cerca de 15% se identificaram como pretos. Além disso, 1,7% dos indivíduos optaram por não declarar sua cor ou raça. Verificou-se ainda, a partir da análise dos dados, que a proporção de autodeclarados indígenas foi extremamente baixa, enquanto o número de participantes que se identificaram como amarelos foi quatro vezes superior ao dos que se declararam indígenas. Esses números reforçam a importância de considerar marcadores étnico-raciais na análise, uma vez que o pertencimento racial pode estar

associado a diferentes condições de acesso e permanência no sistema educacional brasileiro (CARVALHO, 2005; LOPES, 2013).

Tabela 1: Distribuição dos participantes do ENEM 2023 em Sergipe por raça ou cor.

Cor ou raça	Frequência relativa	Valor absoluto
Parda	55,1%	36.132
Branca	22,4%	16.666
Preta	15,2%	9.970
Amarela	2,0%	1.314
Não declarado	1,7%	1.107
Indígena	0,5%	351

4.2 Análise sob a Perspectiva da Cor/Raça

4.2.1 Prova de Ciências Humanas e Suas Tecnologias

A prova de Ciências Humanas e suas Tecnologias tem como objetivo avaliar as competências dos estudantes em relação à compreensão dos fenômenos históricos, geográficos, sociológicos, filosóficos e antropológicos que estruturam a sociedade.

A matriz de referência do ENEM para essa área é composta por habilidades que envolvem a análise crítica de processos sociais, a interpretação de diferentes linguagens e a aplicação de conceitos das Ciências Humanas à realidade cotidiana. Dessa forma, a prova busca aferir a capacidade do candidato de articular conhecimentos interdisciplinares, refletir sobre contextos históricos e sociais e exercer o pensamento crítico, contribuindo para a formação de cidadãos conscientes e participativos.

O desempenho nessa área, portanto, reflete não apenas a memorização de conteúdo, mas também a habilidade de interpretação e argumentação fundamentada, sendo influenciado por múltiplos fatores, inclusive socioeconômicos e educacionais.

A Figura 4 apresenta a distribuição das notas em Ciências Humanas no ENEM de 2023 para o estado de Sergipe, de acordo com a autodeclaração de cor ou raça dos participantes. Observa-se que a mediana das notas varia entre os grupos, sendo mais elevada entre os participantes autodeclarados brancos. Em seguida, aparecem os grupos pardo e preto, com medianas próximas. Os participantes autodeclarados indígenas e amarelos apresentaram as menores medianas de desempenho nessa área do conhecimento.

Além disso, nota-se que a dispersão das notas (representada pela altura das caixas e a presença de outliers) também varia entre os grupos. O grupo branco, por exemplo, apresenta uma distribuição mais alargada para as notas superiores, o que pode indicar uma maior concentração de participantes com desempenho elevado. Já os grupos indígena e amarelo demonstram menor amplitude interquartílica, sugerindo menor variabilidade no desempenho.

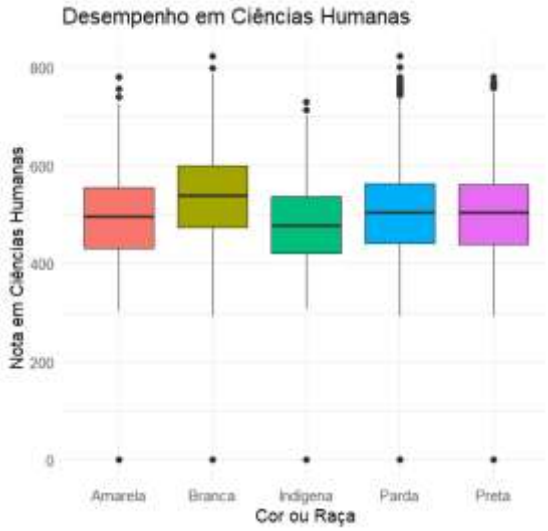

Figura 4: Distribuição das notas em Ciências Humanas segundo a cor ou raça dos participantes de Sergipe.

Esses resultados indicam a persistência de desigualdades no desempenho educacional entre os diferentes grupos étnico-raciais, reforçando a importância de políticas públicas que promovam equidade no acesso e na qualidade da educação (MOURA, 2019; SILVA, BARBOSA3, 2021).

Os resultados da ANOVA indicam que há uma diferença estatisticamente significativa entre os grupos de cor/raça em relação às notas de Ciências Humanas ($F(4, 48938) = 337.8$, $p < 2e\text{-}16$). Como o p-valor é extremamente pequeno (< 0,001), rejeitamos a hipótese nula de que todas as médias são iguais.

O gráfico da Figura 5 exibe as diferenças entre as médias das notas de Ciências Humanas para diferentes pares de grupos de Cor/Raça, apresentando os respectivos intervalos de confiança. Como há cinco categorias de cor/raça analisadas, o número total de pares possíveis para comparação é dado pela combinação das 5 categorias tomadas 2 a 2, ou seja:

$$C(5,2) = \frac{5!}{2!(5-2)!} = \frac{5 \times 4}{2} = 10 \text{ pares} \tag{3.1}$$

O eixo Y do gráfico representa as comparações entre pares de grupos com base na variável cor/raça (por exemplo: "Branca-Amarela"), enquanto o eixo X indica a diferença nas médias das notas de Ciências Humanas entre os respectivos grupos comparados. Cada linha horizontal corresponde a um intervalo de confiança para a diferença entre as médias. A linha tracejada vertical posicionada em zero representa a hipótese nula, ou seja, a ausência de diferença estatisticamente significativa entre os grupos.

Quando o intervalo de confiança cruza essa linha, entende-se que não há evidência estatística suficiente para afirmar uma diferença entre os grupos comparados ($p \geq 0{,}05$). Por outro lado, quando o intervalo não intercepta a linha zero, considera-se que a diferença entre as médias é estatisticamente significativa ($p < 0{,}05$). As cores e os pontos auxiliam na visualização dos resultados: pontos vermelhos indicam comparações com diferença estatisticamente significativa, enquanto pontos azuis correspondem a comparações sem significância estatística. Dessa forma, o gráfico fornece uma síntese visual das desigualdades observadas no desempenho em Ciências Humanas entre os diferentes grupos étnico-raciais.

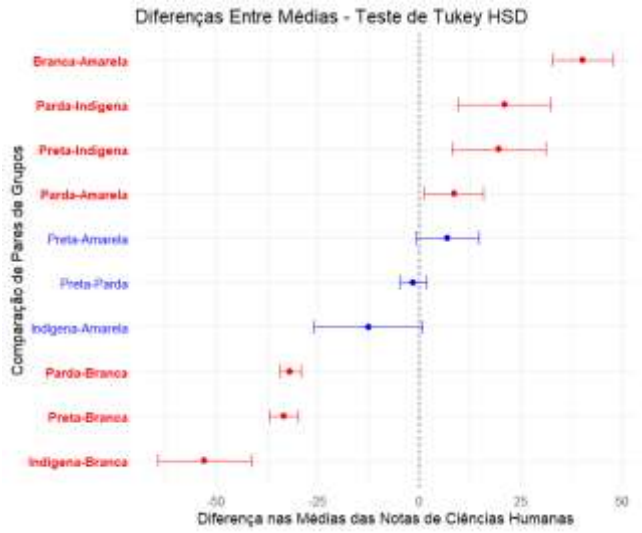

Figura 5: Diferenças nas médias em Ciências Humanas segundo a cor ou raça dos participantes de Sergipe.

A análise das diferenças entre as médias das notas em Ciências Humanas, utilizando o teste de comparações múltiplas de Tukey, permite identificar quais grupos étnico-raciais apresentaram desempenhos significativamente distintos no ENEM.

As comparações estatisticamente significativas, destacadas em vermelho no gráfico, indicam que há diferenças reais entre os grupos comparados. Exemplos disso incluem as comparações entre os grupos Indígena e Branca, Parda e Branca, e Branca e Amarela. Nesses casos, a diferença nas médias das notas é considerada estatisticamente confiável, com base no nível de significância adotado ($p < 0,05$).

Por outro lado, algumas comparações não apresentaram significância estatística, como Indígena e Amarela, Preta e Parda, e Preta e Amarela, sendo representadas na cor azul. Nessas situações, os intervalos de confiança incluem o valor zero, o que indica que os grupos comparados possuem desempenho semelhante em Ciências Humanas.

A direção das diferenças também é importante para a interpretação dos resultados. Quando a diferença entre médias é negativa, significa que o primeiro grupo listado teve, em média, desempenho inferior ao segundo grupo. Por exemplo, na comparação entre Indígena e Branca, o grupo indígena obteve nota média significativamente menor. Já diferenças positivas indicam que o primeiro grupo teve média superior ao segundo.

De forma geral, os resultados apontam para a existência de desigualdades no desempenho entre alguns grupos raciais, ao passo que em outras comparações os desempenhos são equivalentes. O teste de Tukey permite identificar de forma precisa quais dessas diferenças são estatisticamente significativas, contribuindo para uma análise mais robusta sobre a relação entre cor ou raça e o rendimento escolar.

4.2.2 Prova de Linguagens, Códigos e Suas Tecnologias

A prova de Linguagens, Códigos e suas Tecnologias tem como objetivo avaliar competências relacionadas à leitura, interpretação e uso da linguagem em diferentes contextos socioculturais. Composta por 45 questões, essa prova não se limita ao domínio gramatical, mas explora majoritariamente a habilidade dos participantes em compreender textos verbais e não verbais, reconhecer variações linguísticas, analisar efeitos de sentido e utilizar a linguagem como instrumento de interação social.

A estrutura da prova reflete uma abordagem interdisciplinar, contemplando conteúdos de Língua Portuguesa, Literatura, Língua Estrangeira (Inglês ou Espanhol), Artes, Educação Física e Tecnologias da Informação. Tais conteúdos demandam dos participantes uma leitura crítica da realidade e a capacidade de relacionar textos com diferentes linguagens, como imagens, charges, propagandas e letras de músicas.

Na análise estatística dos resultados, os desempenhos observados na prova de Linguagens fornecem importantes indícios sobre o nível de letramento e a competência comunicativa dos estudantes, aspectos que, por sua vez, podem estar associados a fatores socioeconômicos, ao tipo de escola frequentada, bem como às características sociodemográficas dos candidatos, como gênero e cor/raça.

Dessa forma, a interpretação dos dados obtidos nesta área do conhecimento permite observar possíveis desigualdades educacionais, refletidas na distribuição das notas médias entre os diferentes grupos analisados. O desempenho nessa prova pode indicar tanto o acesso a práticas sociais de leitura quanto as condições objetivas de ensino-aprendizagem nos contextos escolares dos participantes do ENEM.

A Figura 6 apresenta um *boxplot* com a distribuição das notas na Prova de Linguagens, Códigos e suas Tecnologias, categorizadas segundo a variável cor ou raça dos participantes. É possível observar diferenças significativas entre os grupos analisados, evidenciando desigualdades no desempenho educacional.

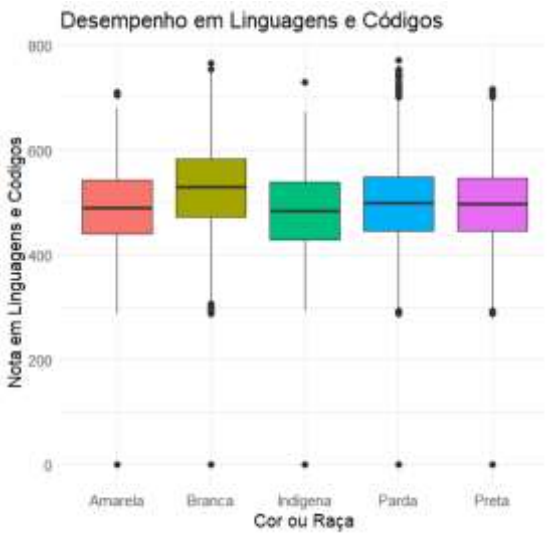

Figura 6: Distribuição das notas em Linguagens, Códigos e Suas Tecnologias segundo a cor ou raça dos participantes de Sergipe.

Os participantes que se autodeclararam brancos apresentaram, em média, os maiores desempenhos, com uma mediana mais elevada em comparação aos demais grupos. Em contraste, candidatos pretos, pardos e indígenas exibiram medianas inferiores e maior concentração de notas em faixas mais baixas, o que indica desempenho acadêmico relativamente menor nesse componente curricular. O grupo amarelo apresentou uma mediana intermediária, mas com ampla dispersão, o que sugere heterogeneidade no desempenho desse segmento.

Além disso, nota-se a presença de outliers em todos os grupos, representando participantes com notas significativamente acima ou abaixo da distribuição central. Esses casos extremos, embora isolados, não alteram a tendência geral observada no gráfico.

Os resultados indicam a persistência de desigualdades raciais na educação, especialmente no que se refere às competências linguísticas avaliadas pelo ENEM. Tais desigualdades podem estar associadas a fatores estruturais, como diferenças no acesso à educação básica de qualidade, recursos pedagógicos disponíveis e capital cultural familiar, conforme discutido por autores como BOURDIEU (1989) e SILVA (2009).

A Figura 7 apresenta os resultados do Teste de Tukey HSD, utilizado para verificar se há diferenças estatisticamente significativas entre as médias das notas de Linguagens entre os diferentes grupos de cor.

Observa-se que os participantes brancos obtiveram médias significativamente superior em relação aos grupos indígena, preto, pardo e amarelo. Em contrapartida, os grupos indígena, preto e pardo apresentaram desempenhos inferiores em relação ao grupo branco, com diferenças negativas estatisticamente significativas, especialmente nas comparações Indígena-Branca, Preta-Branca e Parda-Branca.

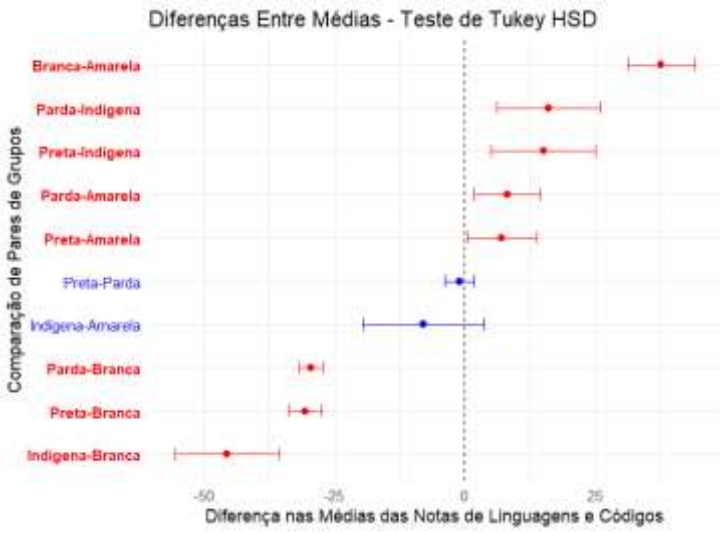

Figura 7: Diferenças nas médias em Linguagens, Códigos e Suas Tecnologias segundo a cor ou raça dos participantes de Sergipe.

Algumas comparações, como Preta–Parda e Indígena–Amarela, não apresentaram diferença estatisticamente significativa, pois os intervalos de confiança incluem o valor zero.
Esses resultados confirmam a existência de desigualdades raciais no desempenho em Linguagens, reforçando as evidências de que fatores históricos e estruturais impactam de forma diferenciada o acesso à educação e ao capital cultural entre os diferentes grupos sociais.

4.2.3 Prova de Ciências da Natureza e Suas Tecnologias

A Prova de Ciências da Natureza e suas Tecnologias é composta por 45 questões de múltipla escolha e abrange conteúdos de Química, Física e Biologia. Seu principal objetivo é avaliar a capacidade dos participantes de compreender fenômenos naturais, interpretar experimentos, analisar dados científicos e aplicar conceitos teóricos a situações do cotidiano.

A abordagem do exame valoriza o raciocínio lógico, a leitura crítica de gráficos e tabelas, e a capacidade de relacionar conhecimentos científicos a temas sociais, ambientais e tecnológicos. Ao invés de testar memorização, a prova exige habilidades de interpretação, análise e resolução de problemas, muitas vezes a partir de textos e contextos interdisciplinares.

Por essa razão, o desempenho nessa área pode refletir não apenas o domínio dos conteúdos escolares, mas também o acesso a uma formação científica de qualidade, frequentemente associada às condições socioeconômicas e ao tipo de escola frequentada.

A Figura 8 apresenta a distribuição das notas na Prova de Ciências da Natureza e suas Tecnologias, segundo a cor ou raça dos participantes. Observa-se que os estudantes brancos obtiveram, em média, os maiores desempenhos, com mediana superior às dos demais grupos. Já os participantes pretos, pardos e indígenas apresentaram medianas mais baixas, indicando menor desempenho médio nessa área do conhecimento. O grupo amarelo também teve mediana inferior à do grupo branco, embora com certa variabilidade nos dados.

A presença de outliers em todos os grupos evidencia que há participantes com desempenhos significativamente distintos dentro de cada categoria racial.

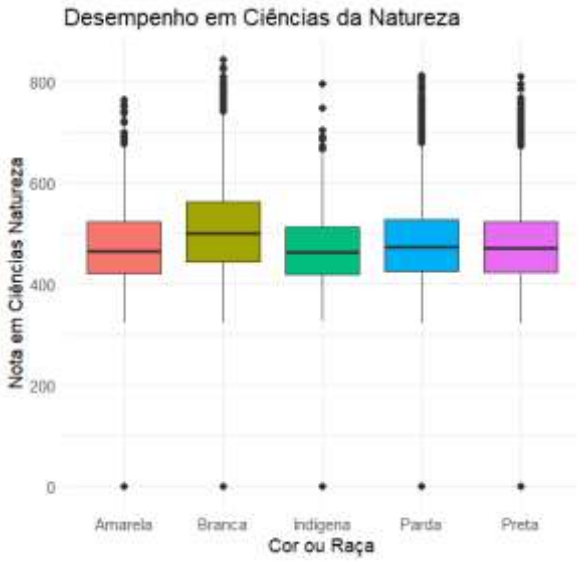

Figura 8: Distribuição das notas em Ciências da Natureza e Suas Tecnologias segundo a cor ou raça dos participantes.

A Figura 9 apresenta os resultados do Teste de Tukey HSD, aplicado para comparar as médias das notas em Ciências da Natureza entre os diferentes grupos de cor ou raça.

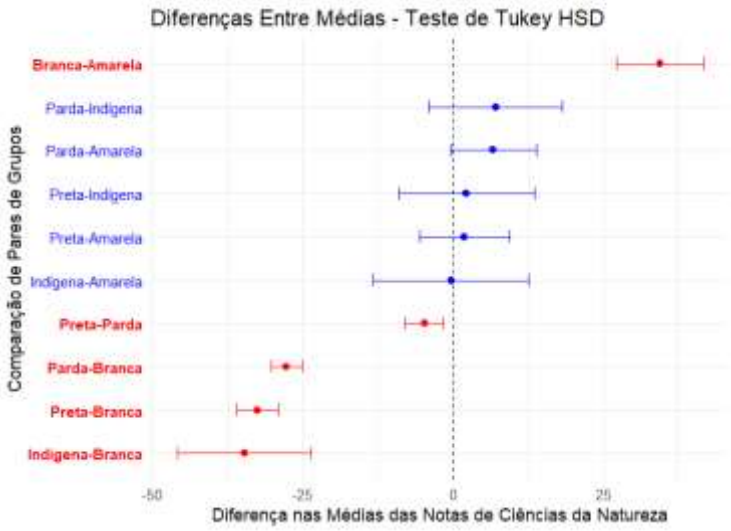

Figura 9: Diferenças nas médias em Ciências da natureza e Suas Tecnologias segundo a cor ou raça dos participantes de Sergipe.

Os resultados mostram que o grupo branco obteve média significativamente superior em relação aos grupos preto, pardo e indígena, como indicado pelas comparações em vermelho. Em especial, a diferença entre os grupos Indígena-Branca e Preta-Branca foram mais acentuadas, com as maiores discrepâncias negativas.

Por outro lado, diversas comparações entre grupos como Parda-Indígena, Preta-Indígena e Preta-Amarela (em azul) não apresentaram diferenças estatisticamente significativas, sugerindo desempenho semelhante entre esses segmentos.

Tais resultados evidenciam desigualdades no desempenho por cor/raça, especialmente entre estudantes brancos e os demais grupos, refletindo possíveis disparidades no acesso ao ensino de ciências e aos recursos educacionais de apoio à aprendizagem científica.

4.2.4 Prova de Matemática e Suas Tecnologias

A Prova de Matemática e suas Tecnologias é composta por 45 questões objetivas e tem como objetivo avaliar a capacidade dos estudantes de resolver problemas, interpretar dados, compreender gráficos, tabelas e aplicar raciocínio lógico em contextos variados do cotidiano.

A abordagem da prova valoriza mais a compreensão conceitual e a aplicação prática da matemática do que a simples memorização de fórmulas. Os itens exploram temas como

aritmética, álgebra, geometria, estatística, funções e análise de informações numéricas em contextos sociais, econômicos e científicos.

O desempenho nessa prova está frequentemente relacionado à qualidade do ensino de matemática recebido durante a educação básica, e tende a refletir desigualdades educacionais vinculadas ao tipo de escola, à infraestrutura pedagógica e ao contexto socioeconômico dos estudantes.

A figura 10 apresenta a distribuição das notas em Matemática, segundo a cor ou raça dos participantes. Observa-se que os estudantes brancos obtiveram as maiores medianas, seguidos pelos grupos pardos e pretos, enquanto os grupos amarelo e indígena apresentaram medianas inferiores. Essa distribuição reforça um padrão de desigualdade no desempenho conforme o grupo racial.

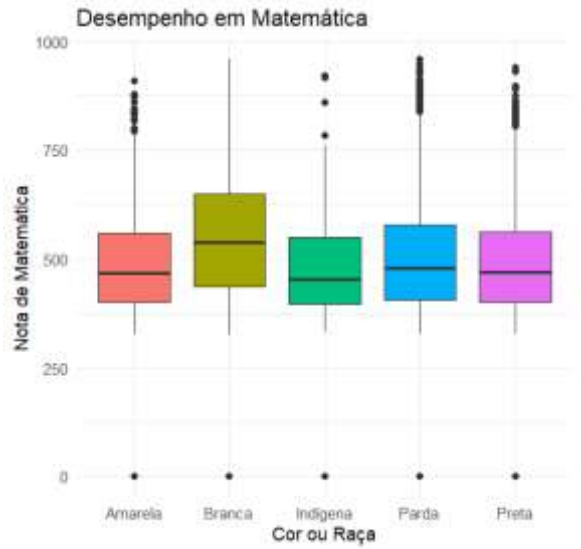

Figura 10: Distribuição das notas em Matemática e Suas Tecnologias segundo a cor ou raça dos participantes em Sergipe.

Na figura 11, o Teste de Tukey revela quais dessas diferenças são estatisticamente significativas. As comparações em vermelho indicam diferenças positivas significativas, especialmente nas comparações entre o grupo branco e os demais. Por exemplo, o grupo branco obteve média significativamente superior às dos grupos preto, pardo, indígena e amarelo.

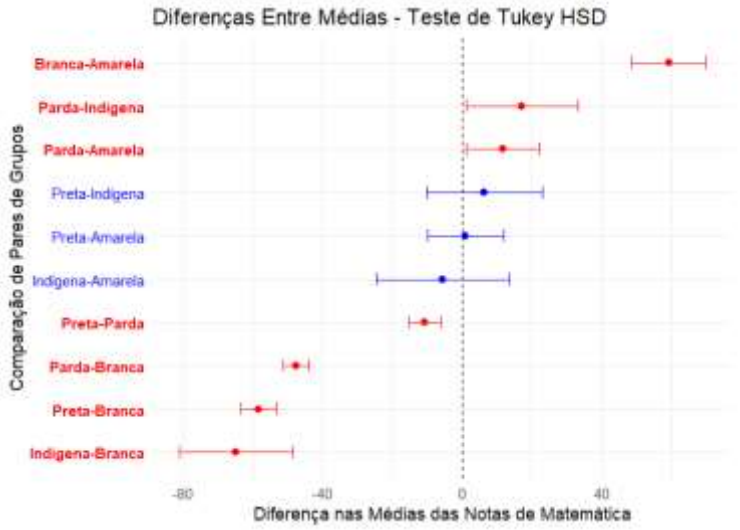

Figura 11: Diferenças nas médias em Matemática e Suas Tecnologias segundo a cor ou raça dos participantes em Sergipe.

As comparações em azul, como entre preta e indígena ou preta e amarela, não apresentaram diferenças significativas, sugerindo desempenho semelhante entre esses grupos. No entanto, as diferenças mais acentuadas foram entre Indígena-Branca e Preta-Branca, indicando uma lacuna preocupante no desempenho matemático entre esses segmentos populacionais.

4.2.5 Prova de Redação

A Prova de Redação tem como objetivo avaliar a competência dos participantes em produzir um texto dissertativo-argumentativo a partir de uma situação-problema proposta. A redação deve apresentar uma tese clara, argumentos bem estruturados e uma proposta de intervenção coerente e respeitosa aos direitos humanos. A correção considera cinco competências, cada uma valendo 200 pontos, totalizando 1.000 pontos. Erros graves, como fuga ao tema ou cópia integral dos textos motivadores levam à nota zero.

Na análise dos resultados por cor ou raça (figura 12), observa-se que os estudantes brancos apresentam as maiores medianas de nota em Redação, seguidos pelos grupos amarelo, pardo e indígena. O grupo preto também apresenta desempenho inferior ao grupo branco, ainda que semelhante aos demais grupos não brancos.

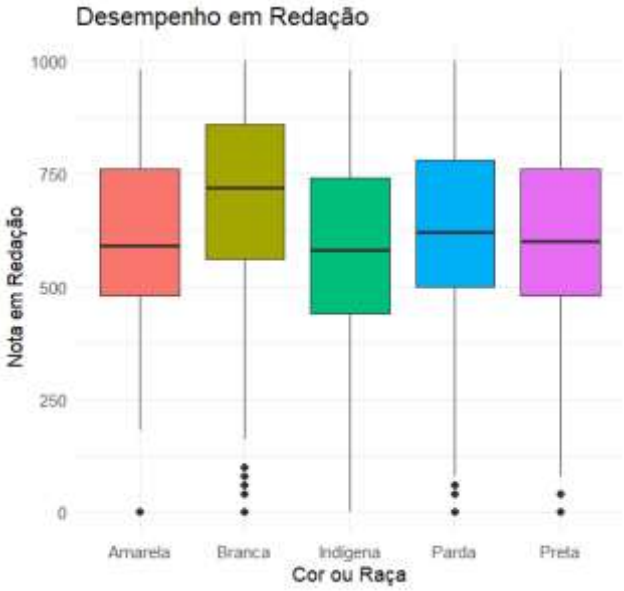

Figura 12: Distribuição das notas em Redação segundo a cor ou raça dos participantes em Sergipe.

O Teste de Tukey (figura 13) reforça essas desigualdades. As comparações significativas em vermelho mostram que os estudantes brancos têm desempenho estatisticamente superior aos grupos preto, pardo e indígena. A maior diferença média é observada entre os grupos indígena e branco, com uma diferença negativa acima de 100 pontos. Já as comparações em azul, como entre preta e indígena ou preta e amarela, indicam ausência de diferença estatística significativa, sugerindo desempenho similar entre esses grupos.

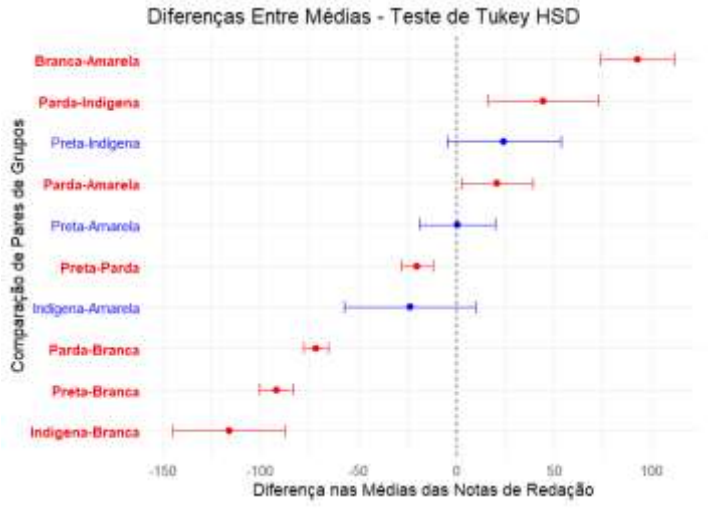

Figura 13: Diferenças nas médias em Redação segundo a cor ou raça dos participantes em Sergipe.

Esses dados evidenciam disparidades raciais no desempenho em Redação, que podem estar associadas a desigualdades no acesso à leitura, escrita formal e orientação escolar, elementos fundamentais para o bom desempenho nessa prova.

4.3 Análise do Desempenho por Sexo nas Diferentes Áreas do Conhecimento

A análise do desempenho dos participantes foi igualmente realizada a partir da variável sexo, visando à comparação dos resultados obtidos por indivíduos do sexo masculino e feminino nas distintas áreas do conhecimento contempladas pelo exame.

No que se refere à área de Ciências Humanas (figura 14), verificou-se que as medianas das notas foram bastante próximas entre os sexos feminino e masculino, situando-se na faixa de 500 a 520 pontos. A análise da dispersão revelou que o grupo feminino apresentou uma concentração maior dos dados em torno da mediana, com uma amplitude interquartil ligeiramente inferior ao do grupo masculino. Apesar disso, ambos os grupos apresentaram ampla variabilidade total e a presença de outliers em ambas as extremidades da distribuição.

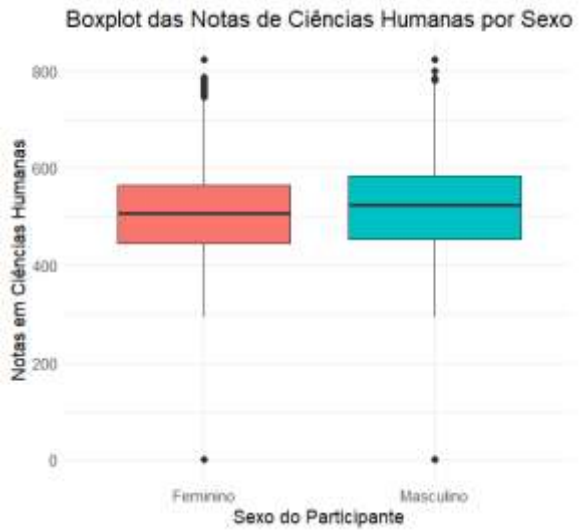

Figura 14: Boxplot das notas de Ciências Humanas de acordo com o sexo do participante.

Em relação à Matemática (figura 15), observou-se uma diferença mais pronunciada entre os sexos. O grupo masculino obteve uma mediana consideravelmente superior, situada entre 520 e 530 pontos, enquanto o grupo feminino apresentou uma mediana abaixo de 500 pontos. A dispersão central dos dados (amplitude interquartil) foi ligeiramente maior entre os homens,

sugerindo maior heterogeneidade no desempenho desse grupo. Além disso, foram identificados outliers em ambos os sexos, abrangendo desde notas muito baixas até pontuações elevadas.

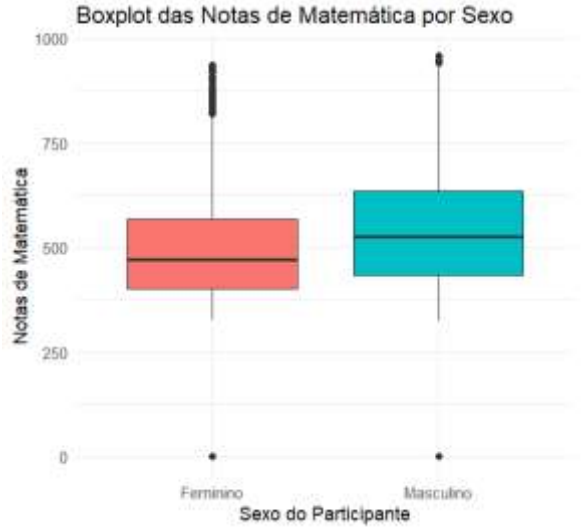

Figura 15: Boxplot das notas de Matemática de acordo com o sexo dos participantes.

Na área de Ciências da Natureza (figura 16), as medianas apresentaram-se bastante similares entre os dois sexos, ambas situadas entre 480 e 500 pontos, com uma pequena vantagem para o grupo masculino. A dispersão dos dados centrais foi um pouco maior entre os homens, e, assim como nas demais áreas, observou-se a presença de valores discrepantes nos dois grupos.

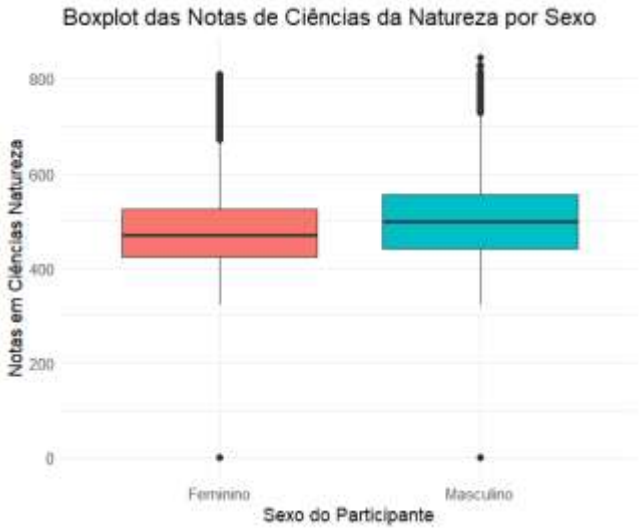

Figura 16: Boxplot das notas de Ciências da Natureza de acordo com o sexo dos participantes.

No campo de Linguagens e Códigos (figura 17), os resultados apontaram para uma notável semelhança no desempenho entre os sexos. As medianas das notas foram praticamente idênticas, variando entre 500 e 510 pontos, e a amplitude interquartil também foi bastante semelhante. Esses dados sugerem uma distribuição de desempenho bastante equilibrada entre homens e mulheres nessa área, com a presença de outliers em ambas as caudas da distribuição.

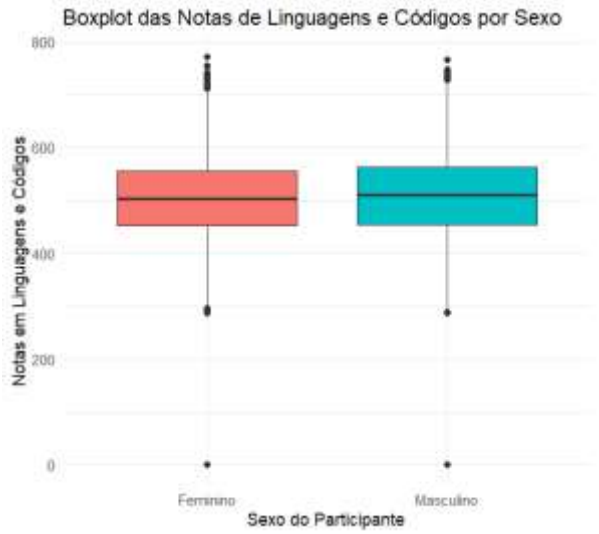

Figura 17: Boxplot das notas em Linguagens e Códigos de acordo com o sexo dos participantes.

Por fim, a análise das notas de Redação (figura 18) evidenciou a maior disparidade de desempenho entre os sexos. As participantes do sexo feminino obtiveram uma mediana significativamente superior, variando entre 670 e 680 pontos, enquanto a mediana do grupo masculino ficou entre 600 e 610 pontos.

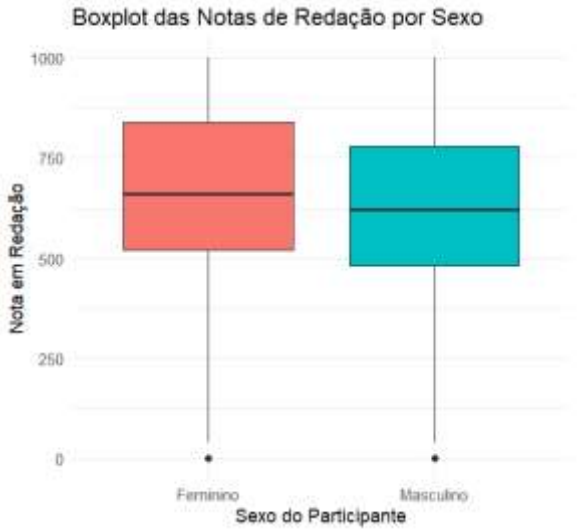

Figura 18: Boxplot das notas de Redação segundo o sexo dos participantes.

A dispersão das notas centrais foi ligeiramente menor entre as mulheres, o que indica maior consistência no desempenho desse grupo. Outliers também foram identificados em ambos os grupos, reforçando a existência de variações extremas nas pontuações.

De forma geral, os dados analisados indicam que, embora em algumas áreas o desempenho entre os sexos tenha sido bastante semelhante, em outras — como Matemática e Redação — foram observadas diferenças mais expressivas, o que pode refletir desigualdades no acesso ao conhecimento, nas metodologias de ensino ou em fatores socioculturais que influenciam o desempenho escolar.

4.4 Análise do Desempenho nas Provas Sob a Perspectiva da Rede de Ensino

Ao analisar o desempenho dos estudantes por tipo de escola, observa-se uma tendência clara de superioridade das instituições privadas em todas as áreas avaliadas.

Em Ciências da Natureza, os alunos da rede privada apresentaram uma mediana mais alta e uma amplitude interquartílica mais concentrada (figura 19), indicando maior consistência nos resultados. Em contraste, os estudantes da rede pública demonstraram maior variabilidade e presença significativa de notas muito baixas.

Figura 19: Desempenho dos participantes de Sergipe na prova de Ciências da Natureza segundo o tipo de escola.

Em Ciências Humanas, a situação se repete: alunos de escolas privadas obtiveram mediana superior e distribuição mais favorável, enquanto a rede pública apresentou assimetria negativa (figura 20).

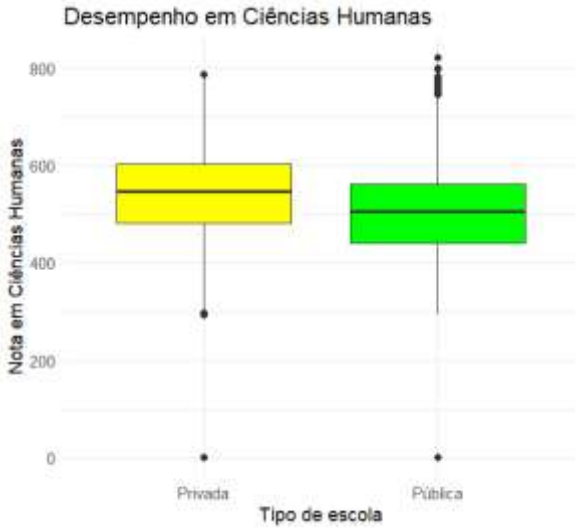

Figura 20: Desempenho dos participantes de Sergipe na prova de Ciências Humanas segundo o tipo de escola.

A Redação apresentou uma das maiores disparidades entre os grupos, com a mediana das escolas privadas se aproximando dos 800 pontos, contra aproximadamente 600 nas públicas (figura 21).

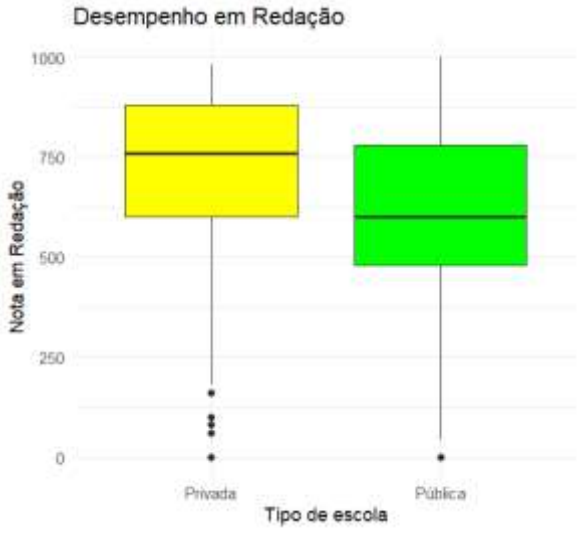

Figura 21: Desempenho dos participantes de Sergipe na prova de Redação segundo o tipo de escola.

Em Matemática, a desigualdade também se mostrou expressiva: a rede privada obteve uma mediana visivelmente superior e menos alunos com notas baixas, ao passo que os estudantes da rede pública se concentraram principalmente abaixo dos 500 pontos, evidenciando lacunas de aprendizagem.

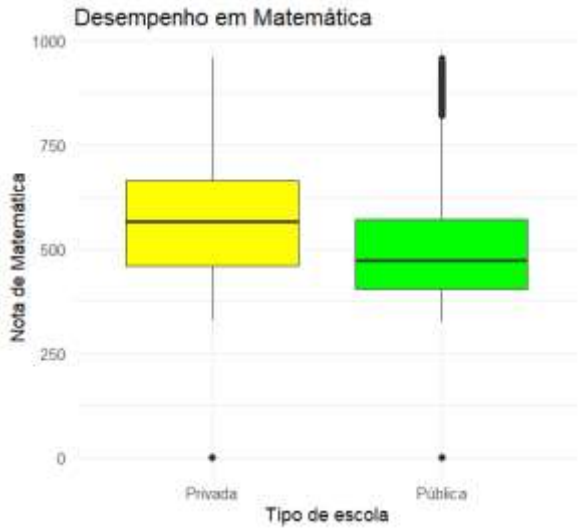

Figura 22: Desempenho dos participantes de Sergipe na prova de Matemática segundo o tipo de escola.

Por fim, em Linguagens e Códigos, ainda que a diferença entre os grupos seja um pouco menos acentuada, os alunos das escolas privadas continuaram com desempenho médio mais alto, e as escolas públicas apresentaram maior dispersão interna, indicando heterogeneidade no nível de aprendizagem.

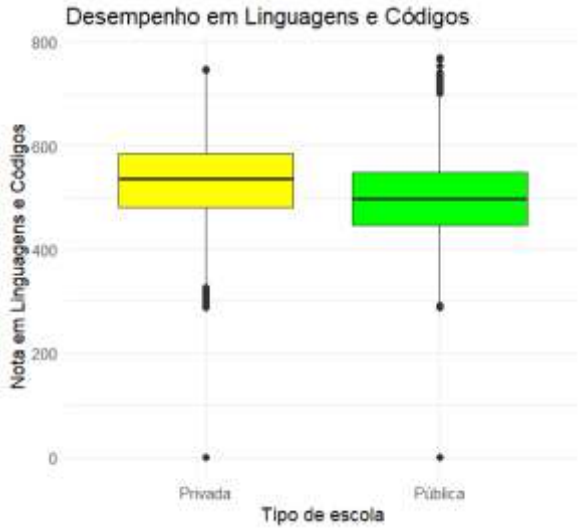

Figura 23: Desempenho dos participantes de Sergipe na prova de Linguagens e Códigos segundo o tipo de escola.

4.5 Resultados dos Testes Qui-Quadrado

Foram realizados testes de associação do tipo qui-quadrado para verificar relações entre variáveis sociodemográficas dos participantes do ENEM em Sergipe. O nível de significância adotado foi de 5% ($\alpha = 0,05$). Os resultados indicam associações estatisticamente significativas entre as variáveis analisadas, conforme apresentado na Tabela 2.

Tabela 2: Resultados dos testes qui-quadrado das variáveis sociodemográficas.

Variável 1	Variável 2	p-valor
TP_ESCOLA	TP_SEXO	0,0000902
TP_ESCOLA	TP_COR_RACA	0,0000000
TP_SEXO	TP_COR_RACA	$1,54 \times 10^{-44}$

Os resultados obtidos por meio do teste qui-quadrado evidenciaram relações estatisticamente significativas entre as variáveis analisadas, considerando o nível de significância de 5%. A associação entre o tipo de escola frequentada e o sexo dos participantes (TP_ESCOLA x TP_SEXO) apresentou um p-valor de $9,02 \times 10^{-5}$, o que permite rejeitar a hipótese nula de independência entre essas variáveis. Esse resultado indica que a distribuição dos estudantes entre as categorias de escola (pública ou privada) varia de acordo com o sexo.

Em relação à associação entre tipo de escola e cor/raça (TP_ESCOLA x TP_COR_RACA), observou-se um p-valor praticamente nulo, denotando uma associação extremamente significativa. Tal evidência sugere a existência de desigualdades raciais no acesso aos diferentes tipos de escola, possivelmente relacionadas a fatores estruturais do sistema educacional brasileiro. De acordo com relatório do Banco Interamericano de Desenvolvimento (BID, 2023), mesmo quando se controla o nível socioeconômico, alunos brancos tendem a estar mais representados em escolas com melhor infraestrutura e desempenho, enquanto alunos pretos e pardos concentram-se na rede pública, com menores condições de ensino, evidenciando um padrão sistemático de desigualdade racial. Complementarmente, dados do relatório *Mapa Preto da Educação* (2024) mostram que essa desigualdade não se dá apenas por raça, mas também por gênero, afetando de forma ainda mais significativa meninas negras.

Por fim, a associação entre sexo e cor/raça (TP_SEXO x TP_COR_RACA) resultou em um p-valor de $1,54 \times 10^{-44}$, também indicando uma forte dependência entre essas variáveis. Esse resultado estatístico corrobora a análise qualitativa apresentada por esse relatório, sugerindo que a distribuição dos sexos não ocorre de forma independente em relação aos grupos raciais, o que pode refletir padrões demográficos específicos ou condicionantes de natureza sociocultural.

4.6 Imputação dos Dados usando o MissForest

A Tabela 3 apresenta o número de registros antes e depois da imputação de dados nas variáveis cor ou raça, tipo de escola e tipo de ensino, cujos valores ausentes foram estimados utilizando o algoritmo missForest. Os maiores acréscimos ocorreram nas categorias "Ensino Regular", "Escola Pública" e "Privada", evidenciando uma elevada proporção de dados faltantes nessas variáveis. O erro de imputação estimado pelo algoritmo é PFC = 0,1995746, com um máximo de 25 iterações e 400 árvores por floresta.

Tabela 3: Quantidade de registros pré e pós imputação por missForest.

Variável	Categoria	Antes da imputação	Total preenchido via imputação	Depois da imputação
tipo de ensino	Ensino Regular	17.521	45.357	62.878
tipo de ensino	Educação Especial - Modalidade Substitutiva	93	2.569	2.662
tipo de escola	Pública	15.266	38.146	53.412
tipo de escola	Privada	3.105	9.023	12.128
cor ou raça	Amarela	1.314	335	1.649
cor ou raça	Indígena	351	292	643
cor ou raça	Branca	16.666	165	16.831
cor ou raça	Parda	36.132	160	36.292
cor ou raça	Preta	9.970	155	10.125

De acordo com Stekhoven e Bühlmann (2012), o erro de imputação para variáveis categóricas no método missForest é medido pela Proportion of Falsely Classified (PFC), que corresponde à proporção de valores imputados incorretamente em relação ao total de valores faltantes da variável. Embora os autores não definam um limiar específico para a confiabilidade da imputação, a literatura posterior e a documentação do pacote missForest indicam que valores de PFC próximos de 0 são considerados indicativos de uma imputação confiável e com bom desempenho, enquanto valores próximos de 1 são considerados ruins (STEKHOVEN & BÜHLMANN, 2012; STEKHOVEN, 2021; PLOSZAJ et al., 2023).

4.7 Análise de Correspondência Múltipla

A Análise de Correspondência Múltipla (MCA) é uma técnica estatística utilizada para examinar as relações entre variáveis categóricas. Por meio de representações gráficas, essa técnica permite visualizar como diferentes categorias dessas variáveis se associam entre si, facilitando a identificação de padrões nos dados.

4.7.1 Análise da Associação entre Sexo, Raça e o Tipo de Escola

A figura 24 corresponde a um *scree plot* gerado a partir da MCA, o qual ilustra a variância explicada por cada uma das dimensões extraídas. No eixo horizontal, têm-se as dimensões que representam eixos ortogonais responsáveis por captar variações nos dados categóricos analisados. Já o eixo vertical indica o percentual da variância explicada por cada uma dessas dimensões, ou seja, a proporção da inércia total atribuída a cada componente.

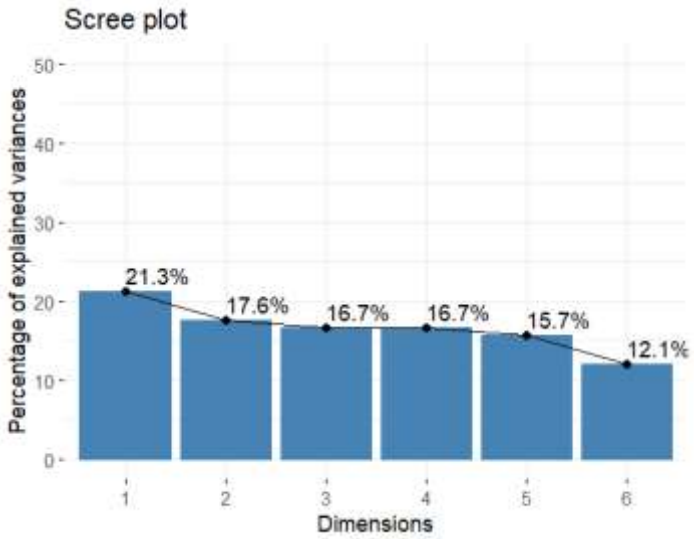

Figura 24: Scree plot.

Os valores observados revelam que a Dimensão 1 explica 21,3% da variância total, sendo a que mais contribui para a separação das categorias. A Dimensão 2 é também relevante, com 17,6% da variância explicada, embora seu peso seja ligeiramente inferior ao da primeira. As Dimensões 3 e 4 apresentam valores idênticos, ambas explicando 16,7% da variância, o que sugere que capturam proporções semelhantes de variabilidade. A Dimensão 5 contribui com 15,7%, enquanto a Dimensão 6 explica 12,1%, sendo esta a de menor importância entre as seis consideradas.

A análise conjunta das variâncias acumuladas permite avaliar a qualidade da representação dos dados em um espaço de menor dimensionalidade. No *scree plot* não se observa um ponto de inflexão nítido (o chamado “cotovelo”), o que indica que a variância está relativamente bem distribuída entre as dimensões, sem uma predominância evidente de apenas uma ou duas. Nesse contexto, embora três dimensões acumulem 55,6% da variância total, a projeção bidimensional (Dimensões 1 e 2), que explica 38,9% da inércia, foi considerada suficiente para os fins

analíticos desta pesquisa. De forma semelhante, no estudo de Nascimento, Massi, Moris e Agostini (2025), intitulado “Análise estatística e pluriescalar das desigualdades educacionais: aspirações científicas e desempenho de estudantes no ENEM”, a projeção bidimensional explica cerca de 40% da inércia total, evidenciando que a variância está distribuída entre múltiplas dimensões, sem que uma ou duas se sobressaiam significativamente.

A decisão por utilizar apenas duas dimensões baseou-se tanto em critérios estatísticos quanto interpretativos. Segundo Hair Jr. et al. (2009), cada dimensão adicionada à solução aumenta a variância explicada da análise, porém em proporção decrescente, ou seja, a primeira dimensão explica a maior parte da inércia, a segunda, a segunda maior parte, e assim por diante. No entanto, acrescentar mais dimensões também torna o processo interpretativo mais complexo, já que mapas perceptuais com mais de três dimensões se tornam progressivamente difíceis de analisar.

Nesse sentido, optou-se por adotar a representação gráfica bidimensional com base no princípio da parcimônia, que recomenda o uso do modelo mais simples possível, desde que adequado à complexidade dos dados analisados. Essa escolha permite preservar, de forma clara e interpretável, as relações mais relevantes entre as categorias, ao mesmo tempo em que mantém a simplicidade gráfica da solução.

Conforme a figura 25, na Dimensão 1, destacam-se as categorias “Privada” e “Branca”, que apresentaram as maiores contribuições percentuais (acima de 35%). Isso indica que essa dimensão está fortemente associada ao tipo de escola e à cor/raça dos participantes.

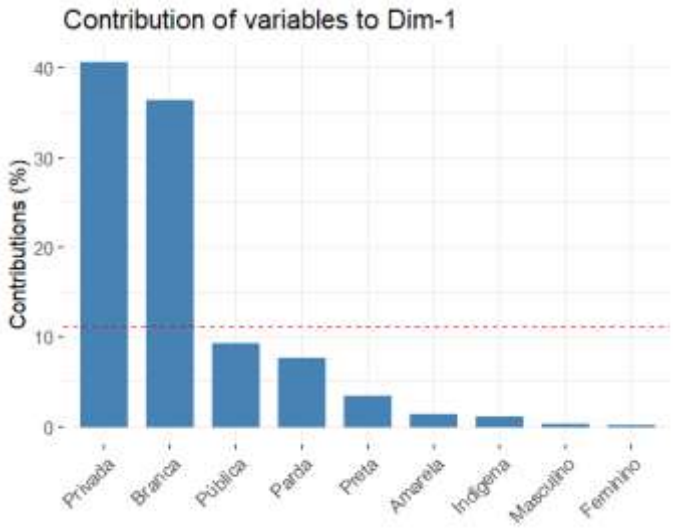

Figura 25: Contribuições das categorias para a Dimensão 1.

Já na Dimensão 2 (figura 26), as maiores contribuições foram observadas para as categorias “Masculino”, “Amarela” e “Preta”, todas com valores superiores a 20%, sugerindo que essa dimensão está mais relacionada ao sexo e a grupos raciais menos representados.

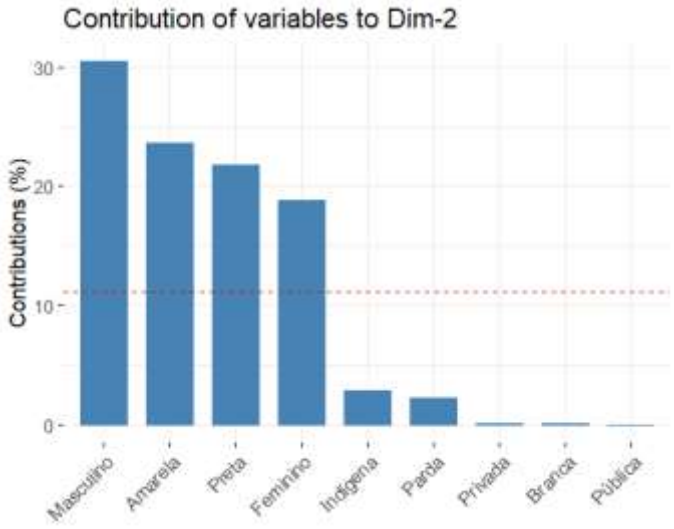

Figura 26: Contribuições das categorias para a Dimensão 2.

A soma das contribuições nas duas primeiras dimensões reforça a relevância das categorias “Privada”, “Branca” e “Masculino” na explicação da variabilidade total. Essas variáveis possuem maior influência na configuração do espaço dimensional da análise, o que revela uma possível associação entre tipo de escola, raça/cor e gênero dos participantes do ENEM em Sergipe (figura 27).

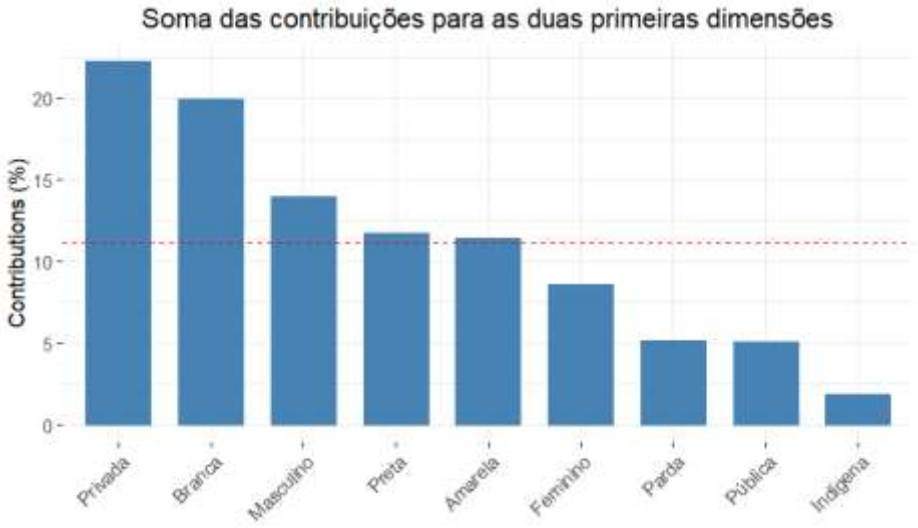

Figura 27: Soma das contribuições das duas primeiras dimensões.

Analisando o mapa perceptual bidimensional, a Dimensão 1 foi fortemente influenciada pelas categorias relacionadas ao tipo de escola e à variável raça. Conforme evidenciado nos gráficos de contribuição das variáveis, as categorias "Privada" e "Branca" apresentaram as maiores contribuições para essa dimensão, ao passo que as categorias "Pública", "Parda", "Preta" e "Indígena" situaram-se no extremo oposto do eixo. Esses resultados indicam a existência de uma associação entre alunos autodeclarados brancos e escolas privadas, enquanto estudantes pretos, pardos e indígenas se associam predominantemente às escolas públicas.

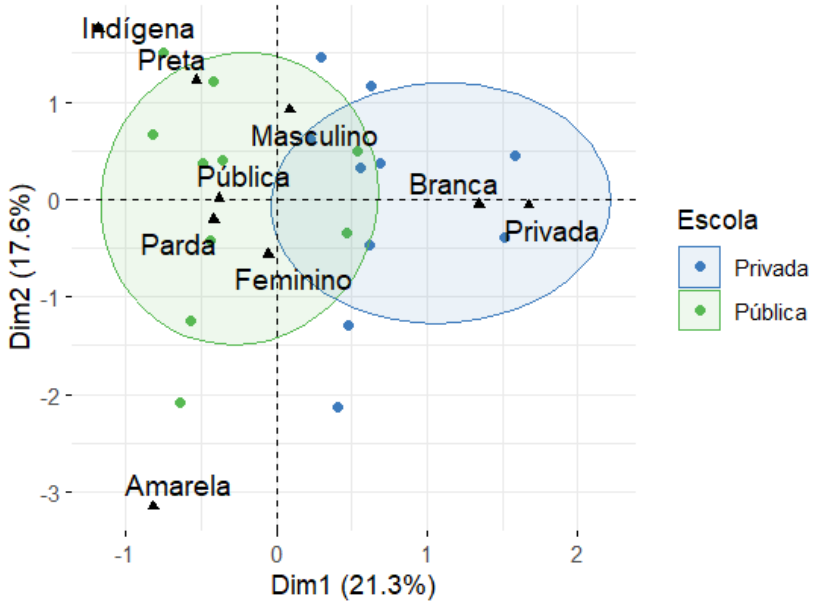

Figura 28: Mapa perceptual bidimensional.

A Dimensão 2 foi influenciada principalmente pelas categorias relacionadas ao sexo, bem como pelas categorias raciais "Preta" e "Amarela". No plano fatorial, observa-se uma clara oposição entre os sexos: a categoria "Masculino" está posicionada na parte superior do eixo vertical, enquanto "Feminino" aparece na parte inferior, indicando que o sexo é um fator relevante na separação dos grupos.

Adicionalmente, a categoria "Amarela" apresenta uma localização distinta em relação às demais, o que pode indicar um padrão de distribuição específico que merece análise aprofundada em estudos futuros.

Com base na interpretação conjunta dos gráficos de contribuição e do plano fatorial, destacam-se as seguintes associações: a variável tipo de escola mostra uma relação direta com a cor/raça

dos estudantes. Alunos brancos estão mais fortemente associados à rede privada de ensino, enquanto estudantes pardos, pretos e indígenas estão mais presentes na rede pública. As categorias do sexo apresentam um padrão de distribuição distinto, com os homens mais próximos das categorias "Branca" e "Privada", enquanto as mulheres se associam mais às categorias "Parda" e "Pública". A distribuição da categoria "Amarela" sugere um perfil diferenciado, o que pode refletir fatores socioculturais e econômicos específicos, sendo necessário aprofundar a investigação para compreender tais nuances.

Os resultados obtidos por meio da ACM apontam para a existência de associações significativas entre cor/raça, gênero e tipo de escola frequentada. As desigualdades observadas evidenciam um cenário em que fatores étnico-raciais e de gênero ainda influenciam fortemente o acesso à educação no estado de Sergipe, especialmente no que se refere à distinção entre ensino público e privado. Esses dados fornecem subsídios relevantes para discussões sobre políticas públicas de equidade educacional.

4.7.2 Análise da Associação entre Sexo, Raça e Tipo de Ensino

Na figura 29, a categoria "Modalidade Substitutiva" é a que mais contribui para a Dimensão 1, com aproximadamente 40% de influência, destacando-se significativamente das demais categorias. Em segundo plano, observam-se contribuições relevantes das categorias "Amarela" e "Branca", ambas relacionadas à variável de cor/raça. Por outro lado, categorias como "Masculino", "Preta", "Feminino" e "Parda" apresentam contribuições menores, situando-se abaixo da linha de referência do gráfico. As categorias "Ensino Regular" e "Indígena" praticamente não contribuem para essa dimensão. Assim, conclui-se que a Dimensão 1 está fortemente associada à variável "Modalidade Substitutiva", seguida pelas categorias de raça/cor "Amarela" e "Branca". Tal configuração sugere que essa dimensão pode estar refletindo um eixo relacionado à modalidade de ensino e à raça.

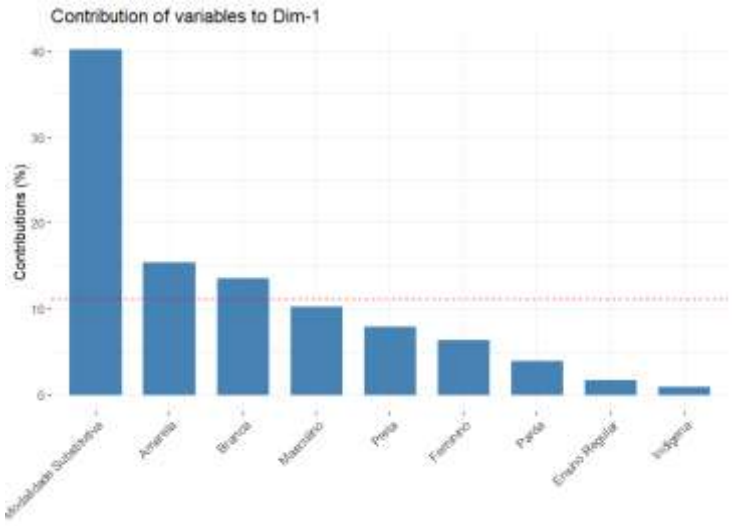

Figura 29: Contribuições das categorias para a Dimensão 1.

No que diz respeito ao gráfico 30, as categorias com maior peso são "Masculino", "Branca" e "Parda", todas com contribuições acima da linha de referência. Em contraste, a categoria "Modalidade Substitutiva", que teve papel central na Dimensão 1, possui uma contribuição muito reduzida nesta segunda dimensão. A variável "Ensino Regular", mais uma vez, exibe um impacto pouco expressivo. Dessa forma, pode-se inferir que a Dimensão 2 está mais fortemente associada à distinção de gênero e de cor/raça, especialmente com destaque para os perfis "Masculino", "Branca" e "Parda". Essa dimensão, portanto, pode estar refletindo um eixo de diferenciação sociodemográfica.

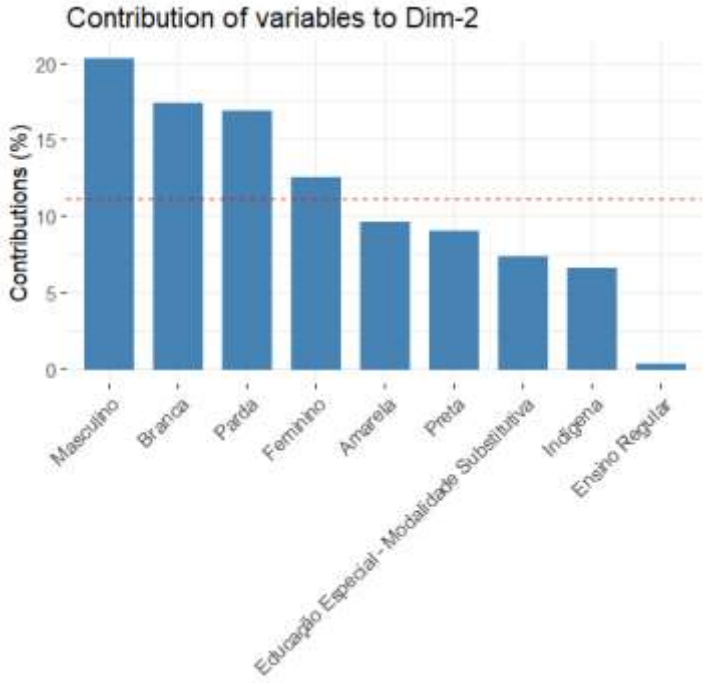

Figura 30: Contribuições das categorias para a Dimensão 2.

Ao se analisar o gráfico 31, que representa a soma das contribuições das Dimensões 1 e 2, verifica-se que a variável "Modalidade Substitutiva" continua sendo a mais influente no modelo, com cerca de 24% de contribuição total. Em seguida, destacam-se as categorias "Branca" e "Masculino", ambas com impacto relevante. Já as categorias "Amarela", "Parda", "Feminino" e "Preta" apresentam contribuições intermediárias. Por fim, "Indígena" e "Ensino Regular" permanecem como as variáveis de menor influência no conjunto das duas dimensões.

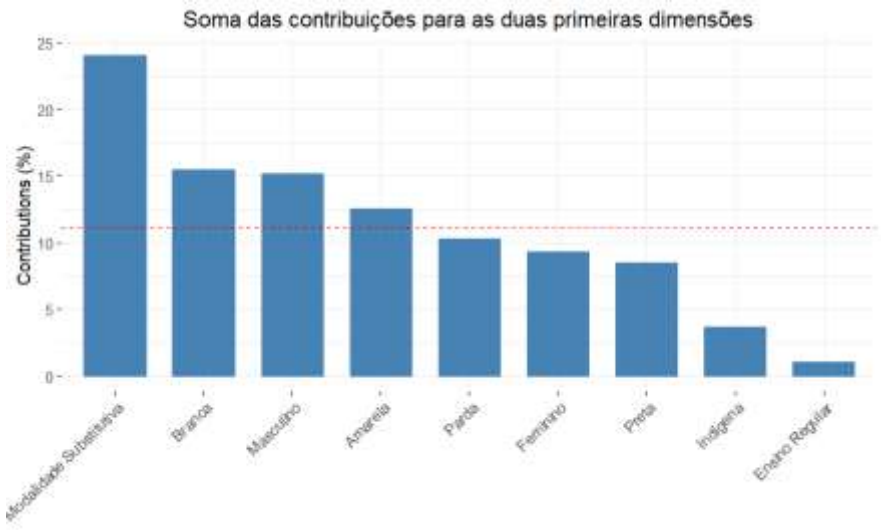

Figura 31: Soma das contribuições das duas primeiras dimensões.

Do ponto de vista interpretativo, esses resultados sugerem uma associação substancial entre o tipo de modalidade de ensino e o perfil racial dos participantes. Além disso, as dimensões analisadas capturam também diferenças de gênero associadas às categorias raciais na Dimensão 2.

O gráfico perceptual apresentado na figura 32 exibe a representação bidimensional gerada pela Análise de Correspondência Múltipla, com o objetivo de visualizar a associação entre as categorias das variáveis qualitativas cor ou raça**,** sexo **e** tipo de ensino. As duas primeiras dimensões explicam, respectivamente, 18,2% e 17,5% da inércia total, somando 35,7% da variabilidade dos dados. Embora esse percentual não represente a totalidade da variância, a opção por manter apenas duas dimensões é justificada com base no princípio da parcimônia, o qual recomenda a utilização do modelo mais simples que seja estatisticamente adequado e interpretativamente informativo (HAIR JR. et al., 2009).

No eixo horizontal, observa-se uma clara separação entre os dois tipos de ensino. Na dimensão 1 concentram-se as categorias associadas ao Ensino Regular como "Masculino", "Branca",

"Preta", "Parda" e "Feminino", enquanto a Dimensão 2 diferencia, em menor grau, categorias como Indígena e Amarela, que se apresentam mais afastadas da massa central de pontos.

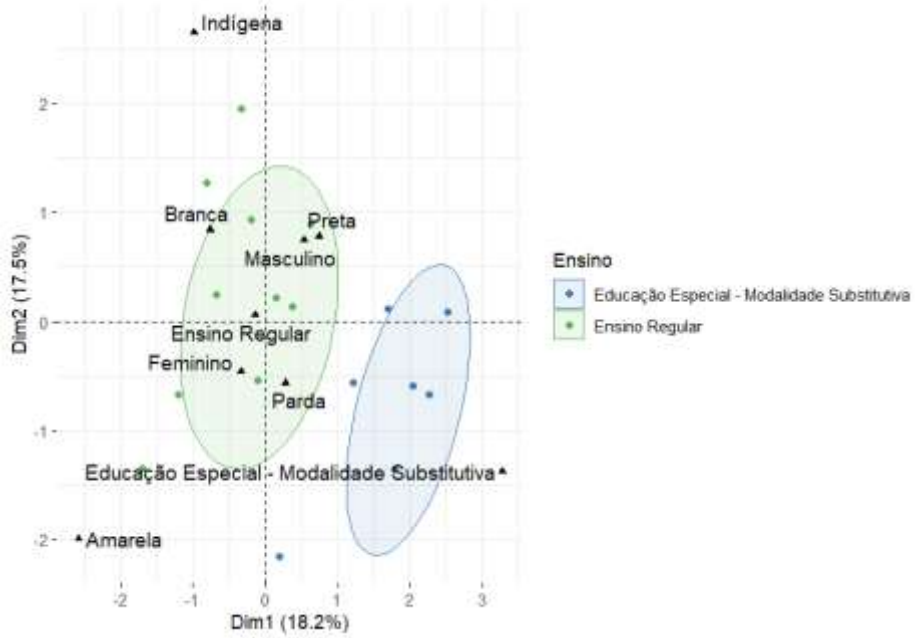

Figura 32: Mapa perceptual bidimensional.

A análise da Dimensão 2 (eixo vertical) evidencia variações ligadas a características sociodemográficas, especialmente raça/cor e sexo. A categoria "Indígena" encontra-se no quadrante superior, distante das demais, o que indica que sua distribuição nos dados é singular em relação às demais categorias. Em contrapartida, "Amarela" localiza-se na extremidade inferior do gráfico, denotando um comportamento distinto, mais relacionado à Educação Especial – Modalidade Substitutiva.

As elipses, que agrupam as observações por tipo de ensino, reforçam a segmentação identificada: o grupo do Ensino Regular se distribui majoritariamente no lado esquerdo e superior do gráfico, enquanto o grupo da Educação Especial – Modalidade Substitutiva está concentrado no lado direito e inferior. Essa disposição espacial das categorias revela uma associação entre as variáveis analisadas e os dois tipos de ensino.

A disposição espacial das categorias indica que o Ensino Regular está associado a uma maior diversidade racial e de gênero, ao passo que a Educação Especial – Modalidade Substitutiva se apresenta como um grupo mais coeso, com menor sobreposição com as demais categorias — o que pode refletir tanto especificidades do público atendido por essa modalidade quanto possíveis desigualdades de acesso.

5. CONCLUSÃO

A presente pesquisa teve como objetivo principal investigar as desigualdades educacionais entre os participantes do ENEM de 2023 no estado de Sergipe, considerando as variáveis sexo, cor/raça, tipo de escola e desempenho nas provas. A partir da análise de uma amostra composta por mais de 65 mil registros, foi possível identificar padrões consistentes de desigualdade que refletem processos históricos e estruturais da sociedade brasileira.

Os resultados revelaram disparidades significativas entre os grupos raciais: estudantes autodeclarados brancos apresentaram, em média, os melhores desempenhos em todas as áreas do conhecimento, especialmente em Linguagens, Matemática, Ciências da Natureza e Redação. Por outro lado, estudantes pretos, pardos e indígenas registraram médias mais baixas, com diferenças estatisticamente significativas, conforme indicaram os testes de ANOVA e Tukey HSD. Esses dados reforçam evidências de que o pertencimento racial ainda exerce forte influência sobre o desempenho escolar.

No que diz respeito ao sexo, identificou-se que os homens obtiveram desempenho superior em Matemática e Ciências da Natureza, enquanto as mulheres se destacaram na Redação, apresentando maior mediana e menor dispersão das notas. As demais áreas do conhecimento apresentaram diferenças menos acentuadas entre os sexos.

As análises por tipo de escola revelaram que estudantes oriundos da rede privada superaram amplamente os da rede pública em todas as áreas avaliadas. Essa diferença sugere que a qualidade da educação, frequentemente vinculada à infraestrutura escolar e à oferta de recursos pedagógicos, é um fator determinante para o desempenho no ENEM.

A Análise de Correspondência Múltipla e os testes Qui-Quadrado permitiram observar associações significativas entre cor/raça, sexo e tipo de escola. Verificou-se que estudantes brancos e do sexo masculino concentram-se majoritariamente nas escolas privadas, enquanto estudantes pretos, pardos, indígenas e do sexo feminino predominam nas instituições públicas. Essa configuração evidencia uma sobreposição de vulnerabilidades sociais, raciais e educacionais que aprofunda as desigualdades de acesso e sucesso escolar.

Dessa forma, conclui-se que o sistema educacional sergipano ainda reflete profundas desigualdades estruturais, exigindo a implementação de políticas públicas voltadas à equidade racial, de gênero e institucional. Tais políticas devem promover não apenas o acesso à escola, mas também a permanência, a aprendizagem de qualidade e a valorização das diversidades sociais.

Cabe destacar que as conclusões aqui apresentadas foram obtidas a partir de dados submetidos a um processo de imputação pelo método MissForest, especificamente aplicado às variáveis cor/raça, tipo de escola e tipo de ensino. Essa abordagem permitiu reduzir perdas de informação e viabilizar as análises estatísticas, mas também impõe limites à interpretação dos resultados, uma vez que se baseia em estimativas inferidas.

Dessa forma, a presente pesquisa avança o debate acadêmico e institucional acerca da desigualdade educacional, ao mesmo tempo em que evidencia a urgência de políticas intersetoriais coordenadas, visando a mitigar as profundas repercussões da desigualdade no sistema de ensino do Brasil.

REFERÊNCIAS

Borg, I.; Groenen, P. J. *Modern multidimensional scaling: Theory and applications*. [S.l.]: Springer Science & Business Media, 2005.

BRIEGA, Diléia Aparecida Martins. *O Enem como via de acesso do surdo ao ensino superior brasileiro*. 2017. 121 f. Tese (Doutorado em Educação Especial). São Carlos: Universidade Federal de São Carlos, 2017. Disponível em: <https://repositorio.ufscar.br/handle/ufscar/8831>.

Briega, Diléia Aparecida Martins. *O Enem como via de acesso do surdo ao ensino superior brasileiro*. 2017. 121 f. Tese (Doutorado em Educação Especial). São Carlos: Universidade Federal de São Carlos, 2017. Disponível em: <https://repositorio.ufscar.br/items/2e933993-20bf-4e62-a421-ec0c86a6a462>.

Bussab, W. O.; Morettin, P. A. *Estatística básica*. 9. ed. São Paulo: Saraiva, 2017.

Connelly, L. M. *Introduction to analysis of variance (anova)*. Medsurg Nursing, Anthony J. Jannetti, Inc., v. 30, n. 3, p. 218–158, 2021.

Cruz, C. D.; Carneiro, P. C. S.; Regazzi, A. J. *Modelos biométricos aplicados ao melhoramento genético: volume 2*. 3. ed Viçosa, MG: UFV, 2014.

Czermainski, A.B. *Análise de correspondência*. Piracicaba, 2004.

Dixneuf, P., Errico, F., & Glaus, M. (2021). *A computational study on imputation methods for missing environmental data*. arXiv preprint arXiv:2108.09500.

Elacqua, Gregory; Dias, Isabella; Nascimento, Danielle; Pérez-Nuñez, Graciela; Rodrigues, Mateus. *O círculo vicioso da desigualdade racial na educação do Brasil*. Washington, DC: Banco Interamericano de Desenvolvimento – Divisão de Educação, nov. 2024. Nota técnica IDB-TN-3046. Disponível em: <publications.iadb.org+2publications.iadb.org+2publications.iadb.org+2>. Acesso em: 14 jun. 2025.

Firmino, M. J. d. A. C. d. S. *Testes de hipóteses: Uma abordagem não paramétrica*. Tese (Doutorado), 2015.

Freitas, E. A.; Silva, M. R.; Carvalho, T. *Two 'Brazils': socioeconomic status and education performance in Brazil*. International Journal of Educational Research, v. 123, p. 102287, 2024.

Geledés – Instituto da Mulher Negra. *Desigualdades no acesso à educação afetam principalmente meninas e mulheres negras*. Geledés, 02 out. 2023. Disponível em: <https://www.geledes.org.br/desigualdades-no-acesso-a-educacao-afetam-principalmente-meninas-e-mulheres-negras/>. Acesso em: 6 jun. 2025.

Greenacre, M. *Correspondence Analysis in Practice*. 2. ed. Boca Raton: Chapman & Hall/CRC, 2008.

Guimarães, Ana Paula; Pinto, Maria. *Identificando a discriminação racial pelo diferencial de desempenho dos estudantes do Ensino Médio*. Revista de Economia Política, São Paulo, v. 44, n. 2, 2024. Disponível em <https://www.scielo.br/j/rep/a/BnFsqgB94FpX7MqzSZzKWMD/>. Acesso em: 27 Abr. 2025.

Hair JR., Joseph F. et al. *Análise Multivariada de Dados*. 6. ed. Porto Alegre: Bookman, 2009.

IBGE. *Síntese de Indicadores Sociais 2023*. Rio de Janeiro: IBGE, 06 dez. 2023. Disponível via Observatório de Sergipe.

Iniciativa Mapa Preto da Educação. *Relatório Nacional – Mapa Preto da Educação*. São Paulo: Instituto Mude, 2024. Disponível em: <https://mapapretodaeducacao.org/Relatorio_Brasil.pdf>. Acesso em: 14 jun. 2025.

Johnson, R. A.; Wichern, D. W. *Applied Multivariate Statistical Analysis*. 6th ed. Upper Saddle River, NJ: Pearson Prentice Hall, 2007.

Junqueira, Rogério Diniz; Martins, Diléia Aparecida; Lacerda, Cristina Broglia Feitosa. *Política de acessibilidade e Exame Nacional do Ensino Médio (Enem)*. Educação & Sociedade, Campinas, v. 38, n. 139, p. 453-471, 1 abr. 2017. Disponível em: <http://dx.doi.org/10.1590/es0101-733020171151513>

Larson, R.; Farber, E.; Farber, E. *Elementary statistics: Picturing the world*. [S.l.]: Pearson Prentice Hall, 2009.

Melo, Nayara; Canegal, Carolina. *Escolas com maioria de alunos negros têm estrutura pior, aponta pesquisa*. Folha de S.Paulo, São Paulo, 15 abr. 2024. Disponível em: <https://www1.folha.uol.com.br/educacao/2024/04/escolas-com-maioria-de-alunos-negros-tem-infraestrutura-pior-aponta-pesquisa.shtml>. Acesso em: 26 Abr. 2025.

Mello Neto, Ruy de Deus et al. *O Impacto do Enem nas políticas de democratização do acesso ao Ensino Superior brasileiro*. Comunicações, Piracicaba, v. 21, n. 3, p. 109-123, jul./dez. 2014. <(PDF) O Impacto do Enem Nas Políticas de Democratização do Acesso ao Ensino Superior Brasileiro >

Mingoti, Sílvia Aparecida. *Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada*. Belo Horizonte: UFMG, 2005.

Nascimento, Matheus Monteiro et al. *Análise estatística e pluriescalar das desigualdades educacionais: aspirações científicas e desempenho de estudantes no ENEM*. Sociologias, Porto Alegre, v. 27, e130399, 2025. Disponível em: <https://doi.org/10.1590/1807-0337/e130399>. Acesso em: set. 2025.

Observatório de Sergipe / Governo de Sergipe. *IBGE: cerca de 140 mil sergipanos saíram da extrema pobreza em 2022*, 07 dez. 2023. Disponível em: <https://observatorio.se.gov.br/ibge-cerca-de-140-mil-sergipanos-sairam-da-extrema-pobreza-em-2022/ório> Acesso em: 07 mar. 2025.

Pamplona, A. S. *Análise de Correspondência para dados com estrutura de grupo*.1998. 180f. Dissertação (Mestrado) - Instituto de Matemática, Estatística e Computação Cientificada, Universidade Estadual de Campinas, São Paulo, 1998. Disponível em: <https://repositorio.unicamp.br/acervo/detalhe/125777>. Acesso em: 05 mar. 2025.

Płoszaj, A., Kowal, M., & Nowak, M. (2023). *Evaluation of imputation methods for missing categorical data: practical guidelines*. PLOS Computational Biology, 19(4), e1010154. Disponível em: <https://doi.org/10.1371/journal.pcbi.1010154>. Acesso em: 27 fev. 2025.

R Core Team**.** *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing, 2025. Disponível em: <https://cran.r-project.org/doc/manuals/r-release/fullrefman.pdf>. Acesso em: 16 jun. 2025.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons.

Sangari, S., & Ray, H. E. (2021). *Evaluation of imputation techniques with varying percentage of missing data*. arXiv preprint arXiv:2109.04227.

Santos, Gustavo de Quadros. *Os efeitos das desigualdades regionais nos resultados do Enem: uma análise a partir dos microdados de 2018*. 2019. 70 f. Trabalho de Conclusão de Curso (Graduação de Ciências Econômicas). Porto Alegre: Universidade Federal do Rio Grande do Sul, 2019. Disponível em: <https://lume.ufrgs.br/handle/10183/205590>.

Schafer, J. L. (1999). *Multiple imputation: a primer*. Statistical Methods in Medical Research, 8(1), 3-15.

Silva, Mariana Cesar Verçosa; Meletti, Silvia Márcia Ferreira. *Estudantes com necessidades educacionais especiais nas avaliações em larga escala: Prova Brasil e Enem*. Revista Brasileira de Educação Especial, Marília, v. 20, n. 1, p. 53-68, mar. 2014. <https://doi.org/10.1590/S1413-65382014000100005>

Smith, R. A. *The effect of unequal group size on tukey's hsd procedure*. Psychometrika, Springer, v. 36, n. 1, p. 31–34, 1971.

Stekhoven, D. J., & Bühlmann, P. (2012). *MissForest—non-parametric missing value imputation for mixed-type data*. Bioinformatics, 28(1), 112-118.

Stekhoven, D. J., & Bühlmann, P. (2012). *MissForest—non-parametric missing value imputation for mixed-type data*. Bioinformatics, 28(1), 112–118. Disponível em: <https://doi.org/10.1093/bioinformatics/btr597>. Acesso em: 25 fev. 2025.

Stekhoven, D. J. (2021). *missForest: Nonparametric Missing Value Imputation using Random Forest [R package vignette]*. Disponível em: <https://cran.r-project.org/web/packages/missForest/vignettes/missForest_1.5.pdf>. Acesso em: 25 fev. 2025.

Tang, Fei; Ishwaran, Hemant. Random forest missing data algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, v. 10, n. 6, p. 363–377, 2017. DOI: <https://doi.org/10.1002/sam.11348>

Triola, M. F. *Introdução à estatística*. 12. ed. Rio de Janeiro: LTC, 2017.

Waljee, Akbar K. et al. *Comparison of imputation methods for missing laboratory data in medicine*. *BMJ Open*, v. 3, n. 8, e002847, 2013. DOI: <https://doi.org/10.1136/bmjopen-2013-002847>

Williamson, D. F.; Parker, R. A.; Kendrick, J. S. *The box plot: a simple visual method to interpret data*. Annals of.

Zeidan, Rodrigo; Almeida, Silvio Luiz de; Bó, Inácio; Lewis Jr., Neil. *Racial and income-based affirmative action in higher education admissions: lessons from the Brazilian experience*. arXiv, [s. l.], 27 abr. 2023. Disponível em: <https://arxiv.org/abs/2304.13936>. Acesso em: 16 jun. 2025.