UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
TRABALHO DE CONCLUSÃO DE CURSO

ANÁLISE DE CORRESPONDÊNCIA APLICADA À PESQUISA DE DESIGUALDADE NA EDUCAÇÃO BRASILEIRA

TRABALHO DE CONCLUSÃO DE CURSO

VICTOR ARDUIN

PORTO ALEGRE, RS
2022

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
TRABALHO DE CONCLUSÃO DE CURSO

ANÁLISE DE CORRESPONDÊNCIA APLICADA À PESQUISA DE DESIGUALDADE NA EDUCAÇÃO BRASILEIRA

VICTOR ARDUIN

Trabalho de Conclusão de curso apresentado como requisito parcial para obtenção do título de Bacharel em Estatística.

Orientador:
Prof. MARCIO VALK

PORTO ALEGRE, RS
2022

Arduin, Victor.
ANÁLISE DE CORRESPONDÊNCIA APLICADA À PESQUISA DE DESIGUALDADE NA EDUCAÇÃO BRASILEIRA / VICTOR ARDUIN. -- 2022.
60 f.

Orientador: MARCIO VALK

Trabalho (Conclusão de curso) - Universidade Federal do Rio Grande do Sul, Instituto de Matemática e Estatística, Porto Alegre, BR-RS, 2022.

Ações Afirmativas, Educação, Estatística Multivariada, Análise de Correspondência, Enem I. , orient. II. Título.

Agradecimentos

Aprender não é um caminho fácil. Muitas vezes envolve irmos ao limite do nosso possível e assim estendermos a fronteira do que conhecemos. Entretanto, somente assim somos capazes de transformar o que entendemos. Mais importante, precisa-se ter consciência que de nada serve sermos capazes de moldar o mundo se não o fazemos para o bem, contribuindo para o desenvolvimento da nossa sociedade e as pessoas que dela participam. Agora, formado como estatístico, espero ser capaz de influenciar nosso país para um caminho de mais igualdade e prosperidade.

Escolher estudar estatística foi uma das decisões mais importantes da minha vida. Desde o começo tive apoio da minha família. Meus pais e irmão acompanharam o sacrífico realizado, o suor escorrido e as lagrimas derramadas, até lograr alcançar a conclusão do curso, por isso lhes entrego a vitória. Amigos deram suporte nos momentos felizes e nos difíceis, sem eles eu não entenderia a importância do curso, por isso dedico-lhes minha alegria. Aos professores por sua paciência e paixão por seu trabalho, destino-lhes o meu futuro.

Resumo

Há uma profunda desigualdade social no Brasil. Claramente o modelo desenvolvimento proposto até aqui fez pouco para equilibrar as oportunidades entre a camada social mais abastada e a camada social mais vulnverável. Analisando as diferentes etnias do país, observa-se que a população não branca enfrenta mais dificuldades em sua trajetória escolar e, posteriormente, em sua vida adulta. Políticas afirmativas têm sido implementadas para corrigir essas distorções, mas ainda estão longe de mitigar os efeitos negativos da situação atual. Além do desenvolvimento de ações para atenuar as desigualdades no país, mostra-se fundamental avaliar e implementar medidas efetivas no combate da desigualdade. Nesse sentido, uso da estatística é importante para o processamento e a análise de dados. Técnicas de Estatística Multivariada estão sendo amplamente utilizadas para realizar análise exploratórias dos dados, ajudando pesquisadores na formulação de hipóteses das causas de problemas enfretandos por nossa sociedade. A trajetória escolar dos estudantes brasileiros na Educação Básica é divergente a depender da sua etnia. Portanto, identificar quais fatores econômicos e culturais se diferenciam entre alunos pode auxiliar na formulação e implementação de ações públicas. Análise de Correspondência é um método estatístico para dados categóricos que, através da redução de dimensionalidade, apresenta graficamente as relações entre variáveis e categórias de uma matriz de dados. Usando dados públicos do Exame Nacional do Ensino Médio, esse trabalho de conclusão de curso estudou a relação entre as etnias e variáveis socioeconômicas de estudantes brasileiros.

Palavras-chave: Ações Afirmativas, Educação, Estatística Multivariada, Análise de Correspondência, Enem

Abstract

There is deep social inequality in Brazil. Clearly, the proposed model has been little developed so far to balance the opportunities between a more affluent and a more vulnerable social strata. Analyzing the different ethnicities in the country, it is observed that the non-white population faces more difficulties in their school trajectory and, later, in their adult life. Affirmative policies have been avoided to correct these distortions, but policies are being corrected based on the current situation. In addition to developing actions to mitigate inequalities in the country, it is essential to evaluate and implement effective measures to combat inequality. In this sense, the use of statistics is important for data processing and analysis. Multivariate Statistical Techniques are being widely used to carry out problem analysis in researching hypotheses about the causes of our society. The school trajectory of Brazilian students in Basic Education is different depending on their ethnicity. Therefore, what are the differentiating and cultural factors among the students that can help in the formulation and implementation. Correspondence Analysis is a statistical method for categoricals that, through dimensionality data reduction, graphically presents the variables between variables and categoricals of a data matrix. Public data from the National High School Exam, this course conclusion work studied the relationship between ethnicities and socioeconomic variables of Brazilian students.

Key-words: Affirmative Actions, Education, Multivariate Statistics, Correspondence Analysis, Enem

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Desigualdade social é um dos maiores desafios do país. As consequências da falta de oportunidades iguais para a população são latentes nos índices sociais de educação, segurança e saúde. Se por um lado reconhece-se esse como um problema comum de todos, ainda não foi possível implantar uma ampla política pública para combater os prejuízos causados à população mais vulnerável.

Educação é um dos intrumentos de política pública mais importantes em um governo. Através dela é possível mudar o destino de milhares de pessoas. Oportunidades são geradas através da qualificação de estudantes, principalmente para aqueles em situação social mais precária. Um dos maiores desafios que permite a manutenção da desigualdae social no Brasil são as taxas de analfabetismo, baixo acúmulo da escolarização, desigual desempenho na conclusão da educação básica, ou seja, a falta de priorização da educação.

Esse último fator é o tema central deste trabalho de conclusão de curso. A escolarização é um período importante na vida de uma pessoa. Um ambiente educacional que permita as crianças desenvolverem plenamente suas capacidades é essencial para sua formação profissional e pessoal. Conforme palavras de (CASTRO, 2009):

> A educação, tendo como uma de suas formas de atuação mais importantes a escolarização, é um fator capaz de desenvolver nos indivíduos suas potencialidades ao permitir o "pleno desenvolvimento da pessoa, seu preparo para o exercício da cidadania e sua qualificação para o trabalho", como previsto na Constituição de 1988.

Dados do (GOES et al., 2020) mostram como o retrato da desigualdade é distribuído de maneira distinta para cada raça. Negros, pardos e indíginas são pior remunerados do que brancos. Identificar os fatores que perpetuam essa diferença é fundamental, e grande parte, reside na falta de um sistema de educação melhor. Compreender o fenômeno da desigualdade educacional entre estudantes brasileiros mostra-se um passo importante para construção de políticas sociais efetivas contra a desigualdade do país. Neste sentido, a estatística é um intrumento científico poderoso na produção de análises de dados que evidenciem os pontos de geração de desigualdades.

Estudar o que há por detrás dos dados é um passo importante para desenvolver políticas públicas melhores e mais assertivas. Empresas e governos cada vez mais investem mais em centros de processamento e análise de dados, visto que a recente digitalização tem permido a geração e armazenamento de informação. Por tanto, técnicas estatísticas fornecem uma metodologia fundamental para produção científica na sociedade (BENAKOUCHE, 1999).

1.1 Problema

Há uma grande desigualdade entre raças no Brasil. Seja no mercado de trabalho ou na educação, básica e superior, a população negra tem muita desvantagem em relação à população branca. Estudos como (BARRETO, 2015) apontam que a desigualdade entre raças crescem a partir do Ensino Médio. Portanto, esse trabalho parte da problemática de identificar quais fatores econômicos e culturais diferenciam estudantes ao fim da sua vida escolar.

1.1.1 Objetivo Geral

Estatística é uma ferramenta poderosa para estudar dados. Através dos seus métodos, um pesquisador pode explorar dados para formular hipóteses e evidenciar relações de causa. Usando Estatística Multivariada, esse trabalho aplicará Análise de Correspondência para explorar quais variáveis socioeconômicas estão associadas a cada raça dos alunos no fim da vida escolar.

1.1.2 Objetivos Específicos

De maneira complementar, serão desenvolvidas análises para apoiar a pesquisa. Como Análise de Correspondência possui caráter exploratório, também haverão testes confirmatórios para suportar as hipóteses formuladas no objetivo geral. Os objetivos específicos são:

1. Anova;
2. Teste de Tukey;
3. Estatística Qui-quadrado;

Esse trabalho está dividido em quatro partes mais introdução e conclusão. A primeira parte, revisão bibliográfica, discute tópicos de estatística multivariada, focando em técnicas de análise exploratória para observação de dados. Apresentou-se conceitos de medidas de similaridade, sua importância para aproximação de variáveis e observações em mapas perceptuais e gráficos de correspondência. A segunda parte faz uma retrospectiva da desigualdade da educação brasileira, concentrando-se em dados socioeconômicos para evidenciar a divisão entre as diversas raças do Brasil. A terceira parte apresenta a metodologia da monografia. Como é executada a Análise de Correspondência, suas características e vantagens. São apresentadas também testes confirmátorios tais como: Estatística Qui-Quadrado, Análise de Variância e Teste de Tukey. Por último, utilizando dados do Enem 2021, aplicou-se os métodos estatísticos anteriores para formular hipóteses dos porquês das desigualdades entre diferentes raças de alunos usando variáveis culturais e econômicas.

Capítulo 2

Revisão Bibliográfica

Esse capítulo é dedicado a apresentar uma revisão dos tópicos mais relevantes para a discussão sobre Análise de Correspondência. As técnicas de Estatística Multivariada foram bastante popularizadas ao longo dos últimos anos e encontra-se presente nas mais diversas áreas do conhecimento humano.

2.1 Estatística Multivariada

As técnicas multivariadas são extensões da análise univariada, estas analisam distribuições de uma única variável, ou da análise bivariada, cujo objetivo é analisar duas variáveis simultaneamente. Entretanto, muitas vezes, mostra-se necessário analisar de maneira conjunta diversas variáveis, sendo a principal característica da Análise Multivariada medir, explicar e prever o grau de relação entre várias variáveis. Conforme palavras (HAIR et al., 2009), uma verdadeira análise multivariada consiste:

> [...] para ser considerada verdadeiramente multivariada, todas as variáveis devem ser aleatórias e inter-relacionadas de tal maneira que seus diferentes efeitos não podem ser significativamente interpretados em separado.

Por causa da sua habilidade para gerar hipóteses relacionadas a um conjunto de dados, são técnicas exploratórias poderosas, mas que também podem ser usadas para fins confirmatórios.

Ainda, dentro do campo da Estatística Multivariada, as técnicas de agrupamento são bastante populares. Esses métodos classificatórios são utilizados quando se deseja explorar similaridades entre objetos através de diversas variáveis. A ideia consiste em agrupar os objetos homogêneos em um espaço n-dimensional através dos seus coeficientes de similaridades. Segundo palavras dos autores (HAIR et al., 2009):

> Análise de agrupamentos é um grupo de técnicas multivariadas cuja finalidade principal é agregar objetos com base nas características que eles possuem. Ela tem sido chamada de análise Q, construção de tipologia, análise de classificação e taxonomia numérica.

Um dos principais desafios é organizar os dados de uma maneira que permita sua análise, agregando essas informações em uma tabela de dados. Essas técnicas são realizadas quando se deseja explorar as similaridades entre indivíduos ou variáveis, classificando-as em grupos conforme suas características em comum. Esse processo acaba por tratar a heterogeneidade dos dados, pois dentro de um conjunto de dados existem p variáveis mais associadas aos n objetos. Existem diversos métodos para processar a disposição gráfica dos vetores multivariados, sendo um dos mais comuns o método de mínimos quadrados, em que a soma dos quadrados das distâncias dos pontos até os eixos é minimizada (BENZÉCRI, 1977).

Existem muitos métodos de agrupamento. Contudo todos convergem para um objetivo comum que é dividir um conjunto de objetos em dois ou mais grupos com base em uma similaridade. Em outras palavras, formar grupos homogêneos. Logo, torna-se necessário conhecer as características do conjunto de dados amostrais sob análise. Quando se observar similaridades, um método de agrupamento poderá ser utilizado (VICINI, 2005).

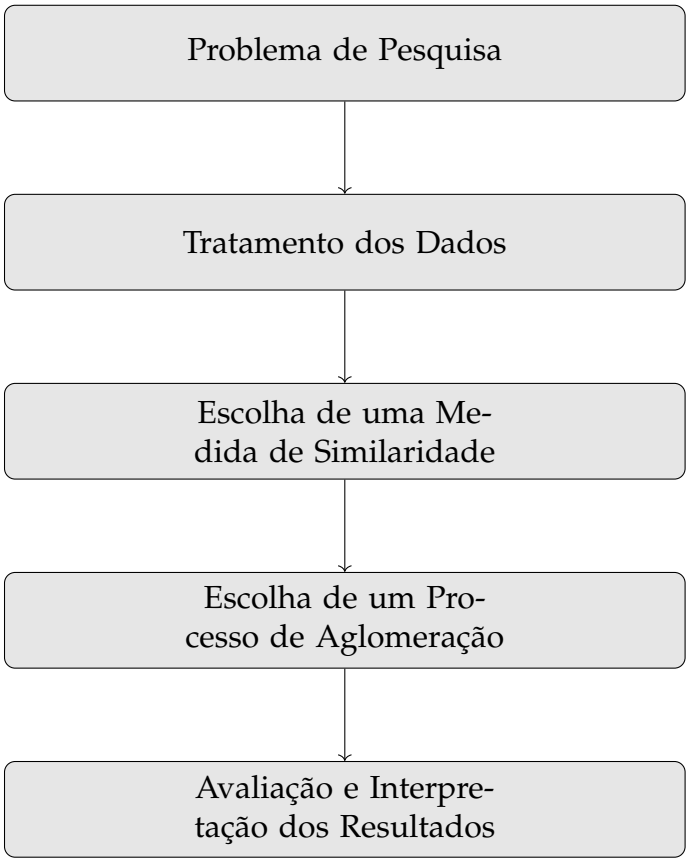
2.2 Medidas de Similaridade

As premissas mais importantes a respeito da análise de agrupamento é a medida de similaridade, ou dissimilaridade, entre os indivíduos e a justificativa teórica para

estruturar dados em grupos. Afinal, todo o processo é guiado através da teoria e da lógica da análise de agrupamento (HAIR et al., 2009).

Importante salientar que o conceito de variável estatística é uma questão central para a análise de agrupamentos, pois ela é responsável pelo agrupamento do conjunto de variáveis que apresentam atributos similares usados para comparar objetos de um estudo. Portanto, o foco não é obter uma estatística, ou seja, estimar uma medida, mas ter uma base comum de comparação entre os objetos. Em busca de estabelecer um método de divisão de dois ou mais grupos de um conjunto de dados, Análise de Grupos deve ser empregada de tal maneira a seguir os seguintes estágios:

FIGURA 2.1. Etapas de uma Análise de Agrupamento



Fonte: Hair et al., Análise Multivariada de Dados (Bookman), pag. 437.

O primeiro estágio do fluxograma se refere ao principal objetivo da análise de agrupamentos. O pesquisador pode querer realizar uma descrição taxonômica, simplificação de dados, evidenciação de relações através do agrupamento. Já o segundo

estágio organiza o planejamento da pesquisa e tratamento dos dados. Se existem observações atípicas ou dados que devem ser padronizados, finalizando com a escolha da metodologia mais adequada de aproximação. Estágio 3 define como deverão ser agrupados os dados, qual medida de similaridade será aplicado. Em seguida, o estágio 4, define qual o processo de aglomeração da pesquisa. Por último, o estágio 5 análise resultados e conclui interpretações em relação ao problema da pesquisa (VICINI, 2005).

As medidas de correlação, amplamente utilizadas por conta do seu apelo intuitivo, não são as técnicas multivariadas comumente empregadas em análise de agrupamentos. As medidas de similaridade mais aplicadas são as que mensuram a distância entre objetos, essas acabam por avaliar o nível de proximidade das observações com as variáveis através da variável estatística de agrupamento. Pode-se pensar, também, sobre as medidas de distância, como valores de dissimilaridade, conforme (HAIR et al., 2009):

> Essas medidas de distância representam similaridade como proximidade de observações umas com as outras ao longo de variáveis na variável estatística de agrupamento. As medidas de distância são, na verdade, uma medida de dissimilaridade, com valores maiores denotando menor similaridade. A distância é convertida em uma medida de similaridade pelo uso de uma relação inversa.

Nota-se, pois, que uma análise de agrupamentos parte do desafio de escolher uma medida de similaridade entre os objetos do estudos, essa métrica é o denominador comum de comparação. Por exemplo, ao realizar um estudo entre similaridades de objetos através de um plano bidimensional, precisa-se de uma regra que permita separar esses objetos. De maneira geral, existem diversos métodos que podem ser encontrados na literatura para agrupar dados, precisando o pesquisador adotar aquela técnica mais adequada ao seu problema de pesquisa. Diferentes abordagens geram diferentes resultados. Logo, a escolha da medida de proximidade mais adequada é de grande importância para análise de agrupamento (BLASHFIELD; ALDENDERFER, 1978).

2.2.1 Distância Euclidiana

Uma das medidas de distância mais comumente utilizadas, muitas vezes chamada de distância em linha reta, a Distância Euclidiana é aplicada no contexto de análise de agrupamento. Sua distância é representada pelo comprimento da hipotenusa de um triângulo retângulo[1]

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}, \tag{2.1}$$

em que x_i e y_i, são, respectivamente, a i-ésima coordenada (variável) dos pontos do $\mathbb{R}^n$ (objetos, vetores) $x = (x_1, x_2, \ldots, x_n)$ e $y = (y_1, y_2, \ldots, y_n)$.

Embora com padrões bastante intuitivos para similaridades, essa dissimilaridade tem a tendência de fazer com que as características que tenham os maiores valores dominem o resultado (LACHI; ROCHA, 2005).

Dois pontos do plano cartesiano possuem uma distância *d*, onde cada par de pontos x e y assumem um número real positivo. As seguintes propriedades são verificadas:

1. se $0 \leq d(x,y)$ e $d(x,y) = 0$, se e somente se, $x = y$;
2. $d(x,y) = d(y,x)$ (Simetria);
3. $d(x,z) \leq d(x,y) + d(z,y)$, para z um ponto qualquer do plano (Desigualdade Triangular).

2.2.2 Distância City-Block (de Manhattan)

Conhecida como a distância "*city block distance*", a medida de Manhattan calcula, diferentemente da Distância Euclidiana, a soma das diferenças absolutas das variáveis:

[1] A distância euclidiana é um conceito matemático que representa a menor distância existente entre dois pontos no $\mathbb{R}^n$ na Geometria Euclidiana. Esta geometria foi construída pelo matemático grego Euclides (ÁVILA, 2001).

$$d_{Cb}(x,y) = \sum_{i=1}^{n} |(x_i - y_i)|, \tag{2.2}$$

Esse é um método de agrupamento mais simples do que a Distância Euclidiana, e muitas vezes esse procedimento pode conduzir a agrupamentos inválidos, principalmente se as variáveis forem altamente correlacionadas (LACHI; ROCHA, 2005).

2.2.3 Distância de Chebychev

Também conhecida como distância do máximo, outro método disponível na literatura é a Distância de Chebyshev, que representa a diferença máxima entre duas variáveis:

$$d_C(x,y) = \max_{i \in [0,\ldots,n]} |(x_i - y_i)|, \tag{2.3}$$

A distância é calculada pela maior diferença entre as coordenadas. Sua aplicação é bastante suscetível a diferenças em escalas ao longo das variáveis, uma vez que valores com alta dispersão (desvios-padrões) têm maior impacto sobre o valor da similaridade final (VICINI, 2005).

2.2.4 Distância de Mahalanobis (D^2)

A similaridade entre dados com características correlacionadas, em que se considera o grau de dependência entre as variáveis com repetição de dados, gera um desafio a mais a ser levado em conta. A medida mais utilizada para o caso em que x e y são dois vetores com a mesma distribuição e com matriz de covariância S é a Distância Mahalanobis (D^2), dada por

$$D^2(x,y) = \sqrt{(x-y)^T S^{-1} (x-y)}, \tag{2.4}$$

Se a matriz de covariância S é a identidade, a distância de Mahalanobis é equivalente a distância euclidiana. Se a matriz de covariância S é diagonal, então a medida de

similaridade resultante é chamada distância euclidiana normalizada:

$$d_n(x, y) = \sqrt{\sum_{i=1}^{p} \frac{(x_i - y_i)^2}{\sigma_i^2}},$$

onde σ_i é o desvio-padrão de x_i na amostra.

A distância de Mahalanobis leva em conta a variabilidade de cada unidade amostral, bastante sugerida para dados provenientes de delineamento experimentais e variáveis correlacionadas (VICINI, 2005).

2.3 Escalonamento Multidimensional

Técnicas de mapeamento perceptual, ou Escalonamento Multidimensional (MDS), são procedimentos que permitem um pesquisador avaliar um conjunto de dados através de um espaço multidimensional. Utilizando medidas de dissimilaridade, esse método busca apresentar como as distâncias entre pontos em espaço dimensional pequeno de $\mathbb{R}^2$, ou $\mathbb{R}^3$, proporcionando uma ideia de aproximação, oferecendo ao pesquisador uma interpretação geométrica dos dados. A visualização gráfica facilita a exploração dos dados muitas vezes difícil quando estudada através de uma matriz numérica associada. Objetivo fundamental da MDS, portanto, é estimar a posição relativa de cada objeto em um mapa perceptual (BORG; GROENEN, 2005).

Em um estudo perceptual, como demarcar se dois objetos A e B possuem maior proximidade do que A e C? Qual a interpretação deve ser inferida a respeito dos eixos do espaço multidimensional? Desafios às perguntas como estas são resolvidas pelos métodos MDS. Uma vez que as dimensões perceptuais são definidas, as comparações relativas entre objetos podem ser realizadas. Estes quando caracterizados podem ter medidas objetivas, propriedades quantificáveis, ou medidas subjetivas, onde são avaliados através das percepções dos indivíduos (HAIR et al., 2009).

Diferentemente de outras técnicas do campo da estatística multivariada, o escalonamento multidimensional não utiliza uma variável estatística para a inferência dos dados. As variáveis que formam a métrica de decisão são resultado do método

de similaridade empregado para os objetos da pesquisa. Conforme palavras do autor (HAIR et al., 2009):

> Em uma analogia simples, isso é como fornecer a variável dependente (similaridade entre objetos) e descobrir quais devem ser as variáveis independentes (dimensões perceptuais). O MDS tem a vantagem de reduzir a influência do pesquisador, uma vez que não requer a especificação das variáveis a serem usadas na comparação de objetos, como se faz em análise de agrupamentos.

Os autores (BORG; GROENEN, 2005) argumentam que existem quatro propósitos que tornam a MDS prático e útil para um estudo. Primeiro é seu caráter de análise exploratória que representa as proximidades dos dados através dos seus pontos em um plano, tornando fácil a inspeção das correlações existentes. Segundo é como a técnica permite testar se e quando um certo critério de similaridade está correspondendo às expectativas empíricas de uma pesquisa. Terceira é sua abordagem analítica que permite concluir quantas dimensões fundamentam as dissimilaridades dos dados. Por fim, quarto propósito, como em um modelo psicológico, esse método explica os julgamentos de dissimilaridades em termos de uma regra que reproduz uma função de distância entre objetos de um experimento que está sob análise.

2.4 Análise de Correspondência

Análise de Correspondência (AC) é uma técnica de interdependência que se popularizou para redução dimensional e mapeamento perceptual. Segundo palavras dos autores (HAIR et al., 2009) "Também é conhecida como escalonamento ou escore ótimo, média recíproca ou análise de homogeneidade". Ela se destaca em relação às técnicas MDS por ser um método composicional, visto que o mapa perceptual é baseado na associação entre objetos. Possui, também, aplicação mais direta na retratação de correspondência de categorias das variáveis. O maior benefício da AC é a maneira como representa linhas e colunas em um espaço conjunto.

A história da AC remonta a mais de 50 anos atrás, sendo referido por uma variedade de nomes como escala dual, método recíproco de médias e discriminante de categorias. A ideia consistia em estabelecer um método de transformar linhas e colunas como pontos em um gráfico bidimensional ou tridimensional. Conforme citada por (DOEY; KURTA, 2011):

> Rows with comparable patterns of counts will have points that are close together on the biplot and columns with comparable patterns of counts will also have points that are close together on the biplot. The row and column points are shown on the same graphical display allowing for easier visualization of the associations among variables.

Em seu caso mais simples, AC examina as relações entre categorias de dados nominais em uma tabela de contingência. A forma mais comum da tabela de contingência é a tabulação cruzada de duas variáveis categóricas não-métricas. Um exemplo para dados de tabulação cruzada no livro (HAIR et al., 2009):

TABELA 2.1. Dados de tabulação cruzada detalhando vendas de produtos por categoria etária

	Vendas			
Categoria Etária	A	B	C	TOTAL
Jovens Adultos (18-35 anos)	20	20	20	60
Meia-idade (36-55 anos)	40	10	40	90
Indivíduos idosos (56 anos ou mais)	20	10	40	70
TOTAL	80	40	100	220

Fonte: (HAIR et al., 2009)

Conforme exemplo, as vendas se diferenciam bastante entre os produtos e entre os grupos etários. Uma AC tenta, através de um mapa, estabelecer um padrão de comportamento entre a idade dos indivíduos e as vendas dos produtos. A identificação desses padrões distintos é alcançada através da frequência observada de cada célula, onde contagens são realizadas para verificar os totais de linhas e colunas da tabela. Como estas geralmente são desiguais, caso mais comum, comparamos o valor observado contra o valor esperado que reflita os valores esperados dos totais específicos de linha e coluna daquela célula.

Diferente de outras técnicas estatísticas que usam teste de hipótese, AC é uma técnica exploratória que explora variáveis categóricas sem uma hipótese específica formada. Assim como a Análise Fatorial, esse método tenta explicar a variabilidade de um modelo através da decomposição da sua variância em uma representação geométrica (DOEY; KURTA, 2011)

AC é um método de obtenção de coordenadas que reproduz categorias das variáveis linha e coluna da tabela de tal forma que a predição de correlação seja representada graficamente. Além de descrever graficamente os dados dispostos em tabelas de contingência, esse método decompõe a estatística qui-quadrada do teste de independência. Algumas das suas vantagens são permitir verificar relações que não seriam identificadas caso essa abordagem fosse realizada aos pares, metodologia usada nos método de escalonamento multidimensional (GREENACRE; BLASIUS, 1994).

AC é uma técnica bastante valorizada pela vasta utilização em diferentes áreas. Alguns exemplos são suas aplicações nas áreas da ecologia, marketing, sociologia, medicina e muitas outras, pois é comum a coleta de dados de natureza categórica. Como dito anteriormente, por seu caráter exploratório, permite um pesquisador levantar questionamentos acerca dos dados. Ademais, o levantamento de dados categóricos muitas vezes é mais fácil e rápido para essas áreas (DOEY; KURTA, 2011).

Importante compreender o processo para padronizar os valores de frequência da tabela de contingência e formar a base para associação entre as variáveis no método de AC. O valor do Qui-Quadrado, um dos principais conceitos estatísticos, serve para mensurar os valores observados contra valores esperados. Conforme exemplo da tabela de contingência anterior, cada célula contém os valores para um combinação específica de linha e coluna. A obtenção do teste Qui-Quadrado segue os seguintes passos (HAIR et al., 2009):

1. calcular o valor esperado para uma célula como se não houvesse qualquer associação nos dados, esse valor é obtido através das probabilidades marginais conjuntas entre linhas e colunas e o total geral:

$$\text{Contagem Esperada da Célula} = \frac{\text{Total Coluna da Célula} \times \text{Total Linha da Célula}}{\text{Total Geral}} \tag{2.5}$$

2. realizar a diferença entre o valor esperado e o valor real, em que o valor absoluto denota a magnitude da diferença e o sinal direção de associação (positivo para associação maior que o esperado ou negativo para associação menor que o esperado):

$$\text{Diferença} = \text{Frequência Esperada} \text{Frequência Real} \tag{2.6}$$

3. depois, padroniza-se as diferenças ao longo das células para formar comparações das variáveis, formando, portanto, o valor Qui-Quadrado através da divisão da diferença dos valores da células ao quadrado pelas respectivas frequência esperadas:

$$\text{Valor Qui-Quadrado} \chi^2 = \frac{\text{Diferença}^2}{\text{Frequência Esperada da Célula}} \tag{2.7}$$

4. por fim, converte-se o valor do Qui-Quadrado para uma medida de similaridade em que a estatística obtida significará o grau de associação entre as variáveis.

A AC é um método híbrido em relação ao Escalonamento Multidimensional ao usar dados não-métricos cruzados para reproduzir mapas de percepção através do posicionamento de todas as variáveis em único mapa. AC tem sido bastante popularizada quando comparada a outros métodos estatísticos por permitir acomodar tanto dados não métricos quanto relações não-lineares (HAIR et al., 2009).

Seu uso é importante para Ciências Sociais. Muitos trabalhos têm alcançado resultados importantes através do uso da Análise de Correspondência. Por exemplo, Nascimento et al. (2017) investigaram associação de variáveis econômicas e culturais para traçar o perfil dos alunos e seu desempenho de um exame nacional de grande porte como o Enem. Evidenciou-se uma grave desigualdade social no sistema educacional brasileiro, alinhado com outros trabalhos produzidos nas diversas áreas, concluindo a consistência dos resultados encontrados.

Capítulo 3

Desigualdade da Educação Básica Brasileira

Um dos problemas mais persistentes do Brasil é a desigualdade social. Essa discrepância entre ricos e pobres é amplamente tratada no debate nacional. Seja no campo acadêmico, nas discussões políticas ou sociais, o tema da desigualdade é de grande relevância para a população. Um dos componentes centrais da estruturação das desigualdades sociais no país é a cor da pele. As dimensões do tecido social brasileiro expõe em praticamente todos os lugares as diferenças raciais nele existentes. Dados do autor (GUIMARÃES, 2006) mostram que pretos e pardos são piores remunerados e possuem menos acesso à educação do que brancos, o que reforça a vulnerabilidade social deste grupo geração após geração. Esse ciclo vicioso gera um problema social ao país. Além disso, embora muita da discussão da desigualdade seja pautada entre pretos e brancos, é importante ressaltar que dados apontam que indígenas também se encontram em desvantagem social, às vezes até maior que os outros estratos. Portanto, o país precisa avançar em políticas públicas que possam resolver as questões econômicas e sociais.

3.1 Desigualdade Social: um aspecto persistente no Brasil

A Educação Básica possui grande importância na vida de qualquer aluno. Na escola, os estudantes desenvolvem suas capacidades cognitivas que serão usadas nas mais diversas áreas profissionais das suas vidas adulta. As escolas também cumprem

protagonismo na construção de valores sociais básicos indispensáveis para o aprimoramento de uma sociedade que são transmitidos de geração em geração. Não obstante, um sistema escolar desigual, onde parcela da sociedade recebe educação de alta qualidade e outra parcela recebe educação de baixa qualidade, impossibilita o desenvolvimento social do país. No contexto brasileiro há um processo ainda mais complexo e que merece atenção - a desigualdade racial. Poucos países têm em seu processo de formação a diversidade que o Brasil possui, existindo em sua história imigrantes das mais diversas regiões do mundo. Todavia, infelizmente, as oportunidades não foram, e ainda não são, distribuídas de forma igualitária, resultando em altos índices de desigualdades, principalmente para a população negra e indígena. A educação pode ser um instrumento transformador desse cenário negativo, contribuindo para a igualdade da sociedade brasileira. Logo, identificar quais são as características que compõem essas desigualdades são essenciais para a promoção de políticas públicas eficientes, sejam estas realizadas por agentes públicos ou por agentes privados.

Pesquisa realizada por (NERI, 2019) mostra o índice de Gini 2 alcançou 0,6003 em 2019 no país. O autor comenta que a desigualdade de renda entre pobres e ricos têm crescido desde 2014, reforçando o cenário de vulnerabilidade social da sociedade brasileira. A principal razão apontada pelo autor para o aumento da desigualdade no período foi o crescimento do desemprego. Ainda, trabalhadores com mais anos de estudo ganham vantagem na procura por emprego, reforçando uma das características mais associadas às disparidades sociais do país que são os anos de escolaridade. As políticas públicas são, por conseguinte, essenciais para reduzir essas distâncias, em que ações educacionais voltadas para a educação básica podem potencializar o crescimento inclusivo da população mais pobre aos empregos de melhor qualidade, possibilitando-os a maiores salários e bem-estar.

Entre brancos e negros observa-se desigualdades bastante evidentes. O rendimento médio mensal das pessoas ocupadas brancas foi de R$ 2.796,00, enquanto que o das pessoas pretas e pardas foi de R$ 1.608,00, aproximadamente 42% menor, em 2018. Da mesma forma, os dados apontam diferenças para a taxa de ocupação entre brancos e negros, sendo de forma mais desfavorável para o último grupo. Por exemplo, em 2018 aproximadamente 55% da força de trabalho no país era composta por negros, não obstante foram o grupo mais desocupado no país 64,2% e mais subutilizados

66,1%; à medida que a população branca de desocupados foi de 34,6% e subutilizados de 32,7%, ou seja, bastante menor. Também, ao analisar os dados dos 10% mais ricos do país constatou-se que 72,7% são brancos, enquanto que dos 10% mais pobres há sobreposição de pretos ou pardos em 75,2% (IBGE, 2019).

A persistência da desigualdade entre grupos raciais no Brasil é tema do debate nacional. Diversos estudos e trabalhos científicos buscam explanar quais são os obstáculos e mecanismos que dificultam a mobilidade social no país. A tradição sociológica sempre deu ênfase aos fatores socioeconômicos para explicar o porquê da falta de mobilidade. Um dos argumentos é a falta de condições iguais entre os diferentes indivíduos. Se por um lado quando o país realizou a abolição da escravatura 3 removeu a barreira formal que separava a sociedade brasileira, por outro faltaram ações que reduzissem os recursos sociais distribuídos de forma dissonante entre brancos, negros e índiginas, fatores como: educação, terras e outros. Logo, os ex-escravos, após a Lei Áurea, precisaram percorrer distâncias maiores que os brancos para superarem a desigualdade social imposta a esses. Enquanto perdurar as diferenças de oportunidades o Brasil não será capaz de reduzir sua desigualdade social (THEODORO et al., 2008).

De fato um relatório produzido por (ELBERS et al., 2004) constatou que os recursos sociais como educação, terra e capital são distribuídos de forma mais assimétrica no Brasil do que em outros países. Dentre esses recursos, a educação tem papel primordial na mobilidade social, em que políticas públicas bem aplicadas na base educacional do país podem ajudar na redução das desigualdades observadas na esfera social. Conforme o estudo, a educação possui forte associação com níveis de renda, sendo os salários positivamente influenciados por cada ano de escolaridade mais observados entre os indivíduos. Em outras palavras, o acréscimo marginal de anos de escolaridade aumenta, em média, a remuneração dos assalariados do país. A teoria do capital humano estabelece existir relação direta entre investimentos individuais em educação e retornos no mercado de trabalho, onde a oferta de educação de qualidade às crianças ajudam-nas a se tornarem indivíduos com capacidades cognitivas mais altas. Portanto, as desigualdades do país podem ser consequência da falta promoção de oportunidades iguais desses recursos, em especial à educação infantil que atende crianças em seus anos iniciais de escolaridade, etapa essa crucial para prepará-las para as próximas fases da educação.

3.2 Retrato da Educação Básica no Brasil

O acesso à educação no Brasil melhorou bastante após a redemocratização. A frequência de alunos pretos ou pardos na infância alcançou 53% em 2018, bastante próximo da frequência de crianças brancas de 55,8%. Outro dado importante é a frequência escolar líquida 5 que apresentou melhora para todos os grupos, principalmente para jovens pretos e pardos que tiveram significativa melhora na situação de atraso escolar. Sem dúvidas ações afirmativas estão contribuindo para a expansão e aprimoramento do ensino no Brasil, possibilitando avanços expressivos - principalmente na Educa- ção Infantil. Nota-se que a frequência escolar líquida praticamente não tem diferença entre crianças brancas e negras até 10 anos, idade que corresponde aos anos iniciais do ensino fundamental, sendo suas proporções, respectivamente, de 96,5% e de 95,8% Embora vários indicadores mostrem melhora na trajetória da educação básica, ainda há notável diferença em relação a alunos brancos e negros adultos. Por exemplo, a proporção de negros entre 18 a 24 anos com estudo inferior a 11 anos e que não frequentavam a escola foi de 28,8% em 2018, enquanto que brancos sob a mesma condição foi de 17,4% (IBGE, 2019).

Políticas públicas voltadas para o combate da desigualdade racial no Brasil são relativamente recentes. (JACCOUD et al., 2009) argumenta que somente nos anos 2000 que iniciativas passaram a ser desenvolvidas com mais intensidade para combater diferenças raciais históricas. Programas como cursinhos pré-vestibulares para alunos negros e cotas no Ensino Superior contribuíram para dar oportunidade de pretos, pardos e indigínas acessarem a universidade. Se por um lado medidas como as apresentadas possuem resultados positivos no que tange o aumento da educação formal para negros e indígenas, por outro evidencia a disparidade dos alunos da rede pública do Ensino Básico. Outra característica recente, porém muito importante, é o desenvolvimento de ações para a valorização da cultura e da história negra 6 , reforçando o caráter multirracial da identidade brasileira. Contudo, um problema reforçado pela autora é o reduzido número de especialistas em história e cultura africana atuantes no Brasil. Muito desse fator é explicado em decorrência de que poucos cursos da área escolar incluem em seus currículos disciplinas relacionadas à história da África.

Fica bastante evidente pelos argumentos expostos que a educação básica é cen-

tral para os formuladores de políticas públicas no que tange a redução das desigualdades e das oportunidades no Brasil. Segundo Pieri (2018), ainda que os gastos com alunos tenham aumentado nos últimos anos, problemas continuam a persistir como o grande número de evasão escolar e pouco avanço na aprendizagem. Muitos alunos não concluem o ensino médio na idade certa e outros que concluem não adquirem as habilidades esperadas. O desenvolvimento das crianças durante a infância e a adolescência é de suma importância, merecendo relevância para políticas educacionais no país. Nessa fase das suas vidas, entre o nascimento e os 21 anos de idade, o cérebro está em formação, assim como suas habilidades cognitivas.

Logo, quando essas características citadas não são desenvolvidas, um passivo educacional é carregado durante sua trajetória escolar que dificilmente é revertido, tornando-os indivíduos com menos chances de concorrer no mercado laboral. O aprimoramento da gestão escolar é necessário na busca da melhora da educação no país. Através da estipulação de metas e medição de resultados, políticas públicas podem ser melhor desenvolvidas com os estudantes a fim que consigam desenvolver as habilidades necessárias para a sua formação. Importante ressaltar também o desafio de promover essas políticas no Brasil dado o seu grande número de estudantes e suas marcas de desigualdades. No ano de 2016, por exemplo, foram matriculados aproximadamente 49 milhões de alunos nas escolas brasileiras, número esse maior do que a população de muitos países. Portanto, com os recursos limitados e o grande contingente de discentes, a eficiência dos recursos investido na educação brasileira é fator essencial para o sucesso de boas ações nos centros escolares brasileiros (PIERI, 2018)

Capítulo 4

Metodologia

Estudar como estão associadas variáveis socioeconômicas com as raças dos estudantes da educação básica não é simples, sendo necessário implementar técnicas estatísticas para tal tarefa. Em sua forma mais simples, a estatística univariada estuda variáveis de maneira isolada, enquanto que em uma forma mais complexa existe a estatística multivariada que estuda variáveis de maneira conjunta. Os métodos multivariados são realizados de acordo com os objetivos da pesquisa, gerando suposições para o estudo. Em contrapartida, as técnicas confirmatórias, como o Teste de Hipótese, são utilizadas para dar confiabilidade a um resultado. No que concerne a este trabalho, busca-se juntar técnicas exploratórias e confirmatórias para construção de uma resposta para um problema profundo na sociedade brasileira: a desigualdade étnica na Educação Básica brasileira.

Conforme trabalho desenvolvido por Nascimento et al. (2017) utilizando microdados do Enem de 2014, os autores, através da Estatística Multivariada, investigaram o desempenho dos alunos no exame nacional de acordo com seu perfil socioeconômico. As variáveis presentes em seu estudo foram: dependência administrativa da escola, raças autodeclarada do aluno, desempenho obtido no exame e uma variável de índice de capital econômico e capital cultural. A conclusão dos autores foi a presença de uma grande desigualdade social no sistema educacional brasileiro, onde o acesso ao Ensino Superior está mais próximo para raça branca e alunos oriundos de escolas federais e particulares. O uso da AC permitiu obter dados consistentes utilizando várias variáveis simultaneamente.

Esse trabalho reproduz de maneira similar Análise de Correspondência com os

microdados do Enem de 2021. Separou-se duas dimensões de variáveis para verificar suas relações com as raças autodeclaradas dos alunos, seguindo as suposições de construção de dimensões de Silva e Hasenbalg (2000):

1. A primeira dimensão refere-se aos recursos econômicos, usualmente mensurada através da renda ou riqueza familiar. Os recursos físicos facilitam o aprendizado das crianças como lugar fixo para estudar, materiais didáticos, etc [1].

2. Segunda dimensão foca nos recursos educacionais, chamado de capital cultural estuda o nível de educação entre os membros adultos da família. Acredita-se que há grande influência dos pais na educação dos filhos, principalmente no papel da educação materna. Esse "clima educacional"da família supõe que pais mais educados percebem melhor os benefícios futuros da educação de seus filhos.

Ademais, há a terceira dimensão que trata da estrutura familiar que não será abordada nesta monografia. A forma como se constitui a família é de grande influência na trajetória de aprendizagem de uma criança. Diferente do capital econômico e cultural que são mensurados mais facilmente, compreender as relações que residem no seio familiar é difícil e exige estabelecer o contexto no qual o capital econômico e cultural dos pais são convertidos em desempenho escolar das crianças.

4.1 Dados

Os dados foram obtidos através do Ministério da Educação (MEC). Lá estão os microdados para consultas dos dados do Exame Nacional do Ensino Médio (Enem). Instituído em 1998, esse exame busca avaliar o desempenho escolar dos estudantes ao término da educação básica. O contexto da sua criação foi desenvolver um ajustamento conceitual e pedagógico para recente inclusão do ensino médio na educação básica. Os objetivos do Enem foram expressos como Corti (2013):

[1] Acredita-se que quanto maior a renda de uma família, maior será a demanda por educação da mesma.

> 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.

O Enem foi escolhido como fonte de dados para essa pesquisa por ser uma das principais ferramentas para avaliar as habilidades e competências de concluintes do Ensino Médio, diferente dos vestibulares locais que buscam selecionar candidatos para o ensino superior. Desde sua criação o exame nacional traz reflexões para compreensão e melhora da qualidade da educação no Brasil. Utilizando as informações preenchidas pelos estudantes no questionário socioeconômico e os resultados obtidos na prova, os dados coletados permitem vislumbrar qual é a dimensão da desigualdade da educação no país.

A prova do Enem possui 180 questões, divididas em 45 itens para cada área do conhecimento: Linguagens e Códigos, Ciências Humanas, Ciências da Natureza e Matemática. Cada pergunta conta com cinco alternativas, sendo apenas uma verdadeira. Além da prova objetiva, há a redação que propõe ao estudante dissertar sobre uma frase tema.

4.2 Análise Exploratória

A melhor maneira para resumir um montante de informações com análise estatística é utilizar Análise Exploratória de Dados (AED). Ela permite visualizar informações além de um modelo estatístico ou um processo de teste de hipóteses, contribuindo

para maior compreensão dos padrões dos dados. A EAD pode ocorrer de diferentes maneiras, sendo algumas delas: tabelas, gráficos, medidas numéricas. A partir das evidências apresentadas pelos dados através da AED, um pesquisador pode realizar suposições e propor modelos para estudar o problema em análise. Técnicas de análise exploratórias são aplicadas para estudar teoricamente dados amorfos, ou seja, dados que não estão vinculados a uma teoria explícita que determine o padrão esperado. Seu objetivo maior é auxiliar um pesquisador para entender a estrutura dos dados (BORG; GROENEN, 2005).

Explorar dados envolve aplicar técnicas estatísticas para observar padrões que podem estar ocultos em um conjunto de dados. Uma das técnicas exploratórias mais famosas é o Diagrama de Caixas. Através de um gráfico, essa técnica permite comparar visualmente através das medianas, quartis aproximados e os pontos baixos e altos de diferentes grupos. Estudar a dispersão e simetria das observações em estudo é importante e facilita a construção de suposições da pesquisa. Apelo visual para interpretar dados é fundamental para exploração de dados, pois muitas vezes os dados organizados apenas em tabelas torna difícil conjecturar sobre os dados. Logo, utilização do *Box Plot*, como também é denominado o Diagrama de Caixas, melhora o raciocínio sobre informações quantitativas (WILLIAMSON et al., 1989).

Nesse sentido, o Diagrama de Caixa é uma técnica simples e poderosa para comparar variabilidades e medianas dos objetos que compõem esses agrupamentos. Diagrama de Caixa consiste em uma caixa que representa a variabilidade numérica por meio de quartis. Suas medidas descritivas permitem observar: primeiro quartil (25%), segundo quartil (50%), terceiro quartil (50%), limite inferior e limite superior; sendo os valores além dos limites considerados outliers. Há vantagens interessantes nessa técnica em relação às demais como sua natureza não paramétrica, isto é, apresenta a variação dos dados sem fazer suposição da distribuição estatística subjacente. Outra vantagem é o uso da mediana em detrimento da média, visto que a última sofre forte influência dos valores discrepantes do conjunto de dados (WILLIAMSON et al., 1989).

4.2.1 Análise de Correspondência

Análise de Correspondência (AC) é uma técnica multivariada utilizada para dados categóricos. O método utiliza frequências das tabelas de contingência para construir um gráfico das associações dos dados em um plano dimensional de caráter qualitativo capaz de explicar grande parte da variabilidade dos dados. Utilizando os valores residuais entre linhas e colunas de uma matriz de dados, pode-se observar o grau de proximidade das variáveis do estudo. Um aspecto importante que diferencia esse método das outras análises estatísticas é a sua característica multivariada, que permite vislumbrar relações de maior complexidade que dificilmente são percebidas em simples comparações de tabelas. Embora a grande variedade de exemplos da AC seja aplicada para casos simples, isto é, quando se explora a relação entre dois fatores, o método permite a extensão para mais variáveis quando se busca envolver mais fatores em sua investigação. Logo, o uso desse método gráfico é de grande valor para ciências sociais aplicadas, permitindo investigar correlações dos dados e identificar os desafios do tema abordado (GREENACRE; BLASIUS, 1994).

AC é uma extensão de outro método de agrupamento chamado de Análise de Componentes Principais. Ela reúne variáveis qualitativas em uma tabela de contingência que conta o número de ocorrências, permitindo analisar o grau de interação entre as mesmas. Os principais conceitos aplicados para essa análise são os perfis de linha ou coluna e a distância qui-quadrada. Conforme palavras de Infantosi et al. (2014):

> Em breves palavras, a AC é um método de análise gráfica de tabelas de contingência, e seus conceitos principais foram descritos em 1940 por Fisher, que os exemplificou com uma análise de associação entre cor dos olhos e tipos de cabelo de habitantes da cidade escocesa de Caithness.

Supondo uma tabela de dimensões $I \times J$, em que a soma de todos os campos seja n_{++}, define-se cada elemento do perfil de linha i em relação às categorias dispostas nas colunas j como:

$$r_{ij} = \frac{n_{ij}}{n_{i+}}, \tag{4.1}$$

onde n_{ij} é o valor do capo i, j e n_{i+} é a soma total da i-ésima linha, logo:

$$n_{i+} = \sum_{j=1}^{J} n_{ij}, \tag{4.2}$$

Por definição, entende-se que $1 \leq i \leq I$ e $1 \leq j \leq J$, $I, J \in \boldsymbol{N}$ para este trabalho. AC possui uma interpretação geométrica importante. O perfil da i-ésima linha é considerado um vetor no espaço J-dimensional, cujas coordenadas são dadas por cada elemento r_{ij}, também compreendido como o vetor $\boldsymbol{r_i} = [r_{i1}, r_{i2}, r_{i3}]$. Analogamente se pode obter os perfis das colunas:

$$c_{ij} = \frac{n_{ij}}{n_{+j}} \tag{4.3}$$

em que

$$n_{+j} = \sum_{i=1}^{I} n_{ij}, \tag{4.4}$$

e

$$\boldsymbol{c_j} = [c_{1j}, c_{2j}, c_{3j}], \tag{4.5}$$

Construindo uma matriz $\boldsymbol{A}(i,:) = [\boldsymbol{r_i}]$ contendo os perfis de linhas e outra mariz $\boldsymbol{B}(:,j) = [\boldsymbol{c_j}]$ contendo os perfis de colunas. Os vetores $\boldsymbol{c} = [n_{+1}, n_{+2}, n_{+J}]^t$ e $\boldsymbol{r} = [n_{1+}, n_{2+}, n_{I+}]^t$ representam, respectivamente, os vetores de totais de coluna e de linha. Ainda, dentro da nomenclatura da AC, os perfis médios são representados pelo centróides de linha e coluna conforme expressos abaixo:

$$r_0 = [r_{01}, r_{02}, \cdots, r_{0J}] = [r_{0j}] = \left[\frac{n_{+1}}{n_{++}}, \frac{n_{+2}}{n_{++}}, \cdots, \frac{n_{+j}}{n_{++}}\right] = \frac{c}{n_{++}} \tag{4.6}$$

e

$$c_0 = [c_{01}, c_{02}, \cdots, c_{0I}]^t = [c_{0i}]^t = \left[\frac{n_{1+}}{n_{++}}, \frac{n_{2+}}{n_{++}}, \cdots, \frac{n_{j+}}{n_{++}}\right] = \frac{r}{n_{++}}, \tag{4.7}$$

A importância de uma categoria em relação às demais fica evidenciada através dos elementos de um centróide, visto que são comparadas com o total da tabela[2]. O

[2]Esses cálculos dos elementos em relação ao total da tabela são chamados de massa

centróide de uma linha é a média ponderada entre os perfis de linha e suas respectivas massas:

$$\sum_{i=1}^{I} c_{0i}\boldsymbol{r_i} = \sum_{i=1}^{I} \frac{n_{i+}}{n_{n++}} \cdot \frac{n_{ij}}{n_{i+}} = \frac{1}{n_{++}} \sum_{i=1}^{I} n_{ij} = \frac{n_{+j}}{n_{++}} = r_{0j}, \tag{4.8}$$

Analogicamente, para os perfis de coluna:

$$\sum_{j=1}^{J} r_{0j}\boldsymbol{c_j} = \sum_{j=1}^{J} \frac{n_{+j}}{n_{n++}} \cdot \frac{n_{ij}}{n_{+j}} = \frac{1}{n_{++}} \sum_{j=1}^{J} n_{ij} = \frac{n_{i+}}{n_{++}} = c_{0i}, \tag{4.9}$$

Uma maneira de verificar a dependência entre os perfis das linhas i , $i^{'}$ é aplicar a distância qui-quadrado:

$$D_{qui}^{i,i^{'}} = \sqrt{\sum_{j=1}^{J} \frac{1}{r_{0j}} \cdot (r_{ij} - r_{i^{'}j})^2}, \tag{4.10}$$

Substituindo a equação (4.1) em (4.10), a distância qui-quadrado entre dois perfis de linha é apresentado como:

$$D_{qui}^{i,i^{'}} = \sqrt{\sum_{j=1}^{J} \frac{1}{r_{0j}} \cdot (\frac{n_{ij}}{n_{i+}} - \frac{n_{i^{'}j}}{n_{i+}})^2} = \sqrt{\sum_{j=1}^{J} (\frac{n_{ij}}{\sqrt{r_{0j}} \cdot n_{++} \cdot c_{0i^{'}}} - \frac{n_{i^{'}j}}{\sqrt{r_{0j}} \cdot n_{++} \cdot c_{0i^{'}}})}, \tag{4.11}$$

Portanto, conseguimos identificar que a distância qui-quadrado entre perfis de linha é a distância Euclidiana com as seguintes coordenadas:

$$\frac{n_{ij}}{n_{++}} = p_{ij}; \frac{p_{ij}}{\sqrt{r_{0j}} \cdot c_{0i}} = s_{ij}^{l}, \tag{4.12}$$

tem-se

$$D_{euc}^{i,i^{'}} = \sqrt{\sum_{j=1}^{J} (s_{ij}^{l} - s_{i^{'}j}^{l})^2}, \tag{4.13}$$

Como $D_r = diag(r_0)$ e $D_c = diag(c_0)$ são as matrizes diagonais dos centróides de linha (JxJ) e coluna (IxI) obtém-se:

$$\boldsymbol{S^I = D_c^{-1} \times P \times D_r^{-0,5}}, \tag{4.14}$$

A nova matriz obtida S^I é o resultado da padronização da matriz de perfis de linha (IxJ), e P a matriz com elementos P_{ij}. Analogamente para os perfis de coluna:

$$D_{qui}^{j,j'} = \sqrt{\sum_{i=1}^{I} \frac{1}{c_{0i}} \times (c_{ij} - c_{ij'})^2}, \tag{4.15}$$

Logo,

$$\boldsymbol{S^c = D_c^{-0,5} \times P \times D_r^{-1}}, \tag{4.16}$$

Usando as definições de perfil de linha e coluna, consegue-se observar o princípio distributivo, no qual dois perfis similares podem ser unidos em um perfil com massa igual à soma das massas individuais relativas aos perfis considerados. Como suas posições no espaço são iguais, ou aproximadas, a nova linha de contingência é dada por $2[n_{i1}, n_{i2}, \ldots, n_{iJ}]$ e massa $\frac{2n_i}{n_{++}}$, como a soma total de cada coluna n_{+j} não se modifica, resulta um mesmo centróide de linha r_{0j}. Logo, as distâncias dos perfis aos centróides de linha não mudam (ABDI; BÉRA, 2014).

Usando os perfis de linha em conjunto com as categorias nas colunas, pode-se realizar uma interpretação geométrica, pois a nova matriz obtida dos calculos acima gera uma figura geométrica regular em $(J-1)$ dimensões para os perfis de linha.

Geralmente, quando da disposição gráfica dos vetores multivariados, é comum utilizar uma nova origem para as variáveis da análise por meio do Método dos Mínimos Quadrados (MMQ). De acordo com Infantosi et al. (2014), o método computacional mais usado para obtenção da minimização da distância dos pontos até o eixo:

> O método computacional mais utilizado para tal minimização é o algoritmo de Decomposição por Valores Singulares (DVS), em que a matriz de perfis é fatorada em três matrizes, uma das quais diagonal com os valores singulares

Esse método é de fundamental importância para análise matricial. Um grande volume de dados pode ser armazenado, processado e analisado. Logo, possibilita evidenciar o "conteúdo dominante"existente nos dados, facilitando a informação a ser apresentada para o pesquisador (ANDRADE; SANTOS, 2020).

4.3 Testes Confirmatórios

4.3.1 Teste de Hipóteses

Em experimentos que utilizam estatística é comum, se não imprescindível, a realização de algum teste de hipóteses para verificar as hipóteses de uma pesquisa. Esses testes podem ser divididos em paramétricos e não paramétricos conforme a necessidade de cada pesquisa, onde o primeiro exige maior rigor teórico, visto que é baseado nas distribuições dos nossos dados, e o segundo com propriedades mais simples e flexíveis.

A validade de um resultado de um método empregado para estudar uma hipótese de pesquisa é observado através da execução de algum teste de hipótese. Em geral são constituídos duas hipóteses: a Hipótese Nula (H_0) e a Hipótese Alternativa (H_1). Enquanto a primeira reflete a expectativa de não relação entre variáveis que estão sendo estudadas, a segunda corresponde à uma situação de relação entre as variáveis, sendo essa última geralmente o que um pesquisador está tentando estabelecer (FIRMINO, 2015).

4.3.1.1 Estatística Qui-Quadrado

Uma das distribuições mais utilizadas na estatística inferencial é a estatística Qui-Quadrado. A aplicação do Qui-Quadrado permite avaliar de maneira quantitativamente um valor de dispersão para duas variáveis categóricas. Por ser um método não-paramétrico, não depende de parâmetros populacionais como média e variância, sendo possível estender seu uso para variáveis qualitativas. Logo, pode-se observar se há evidências de associação entre as variáveis de uma tabela de contingência, 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.

Esse teste é adequado para execução quando os elementos da amostra são divididos em duas ou mais categorias. A ideia fundamental é verificar se existem diferenças significativas entre as respostas observadas e esperadas de cada grupo. Em outras palavras, busca estabelecer se as respostas entre as diferentes categorias estão equitativamente distribuídas (FIRMINO, 2015).

Inicialmente define-se a hipótese nula e alternativa da pesquisa. A primeira supõe que as distribuições das categorias estudadas são homogêneas. Já a hipótese alternativa evidencia que existem diferenças entre as categorias. O teste usa as frequências observadas contra as frequências esperadas de cada categoria, em que as últimas provêm de uma distribuição hipotética dos dados observados supondo que a hipótese nula é verdadeira (LARSON et al., 2009).

Definimos frequência esperada como:

$$E_i = np_i, \tag{4.17}$$

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

Por ser um método adequado para comparar proporções, isto é, estudar possíveis divergências entre frequências observadas e esperadas para diferentes grupos, sua aplicação é importante para AC. Um ajuste para distribuição qui-quadrado com $k-1$ graus de liberdade, onde J é o número de categorias, é expressado por:

$$\chi^2 = \sum_{i,j=1}^{I,J} \frac{(n_{ij} - E_{ij})^2}{E_{ij}}, \tag{4.18}$$

em que n_{ij} e E_{ij} são, respectivamente, os valores observados e esperados de cada célula da tabela de contingência. Supondo a independência entre as variáveis, o valor esperado será o produto da probabilidade de ocorrência de cada uma delas,

$$E_{ij} = \frac{n_{i+} \cdot n_{+j}}{n_{++}}, \tag{4.19}$$

Aplicando a razão a estatística qui-quadrado pelo total da tabela de contingência, chega-se:

$$\rho^2 = \frac{\chi^2}{n_{++}}, \tag{4.20}$$

O resultado é o coeficiente de contingência de Pearson. Esse valor representa as diferenças entre os valores observados (n_{ij}) e os valores esperados (E_{ij}), quanto maior o valor, maior é a dispersão entre os dados.

4.3.1.2 Distribuição Qui-Quadrado

Pertencente ao grupo das distribuições contínuas, a distribuição qui-quadrada é especificada pelos graus de liberdade e parâmetro de não centralidade. Uma característica importante dessa distribuição é sua assimetria positiva, que diminui quanto mais graus de liberdade existirem, aproximando cada vez mais para uma densidade Normal[3]. Sejam $Z_1, Z_2, \ldots, Z_n$ variáveis aleatórias independentes com distribuição normal, cada uma com média 0 e variância 1. A soma dessas variáveis ao quadrado terá uma distribuição qui-quadrada, ou seja

$$X = Z_1^2 + Z_2^2 + \cdots + Z_n^2 \sim \chi_q^2 \tag{4.21}$$

cuja função densidade com q graus de liberdade é dada por:

$$f(x;q) = \begin{cases} 0 & x \leq 0; \\ \frac{1}{\Gamma\left(\frac{q}{2}\right)} \left(\frac{1}{2}\right)^{q/2} x^{q/2-1} e^{-\frac{x}{2}} & x > 0, \end{cases}$$

em que Γ representa a função gamma.

Como todos os valores de qui-quadrado são tomados ao quadrado, a soma de quadrados resultará em um número positivo real. O primeiro e segundo momento, a média e a variância, estão relacionados aos graus de liberdade. Conforme:

$$\mu = E(X) = q, \tag{4.22}$$

e

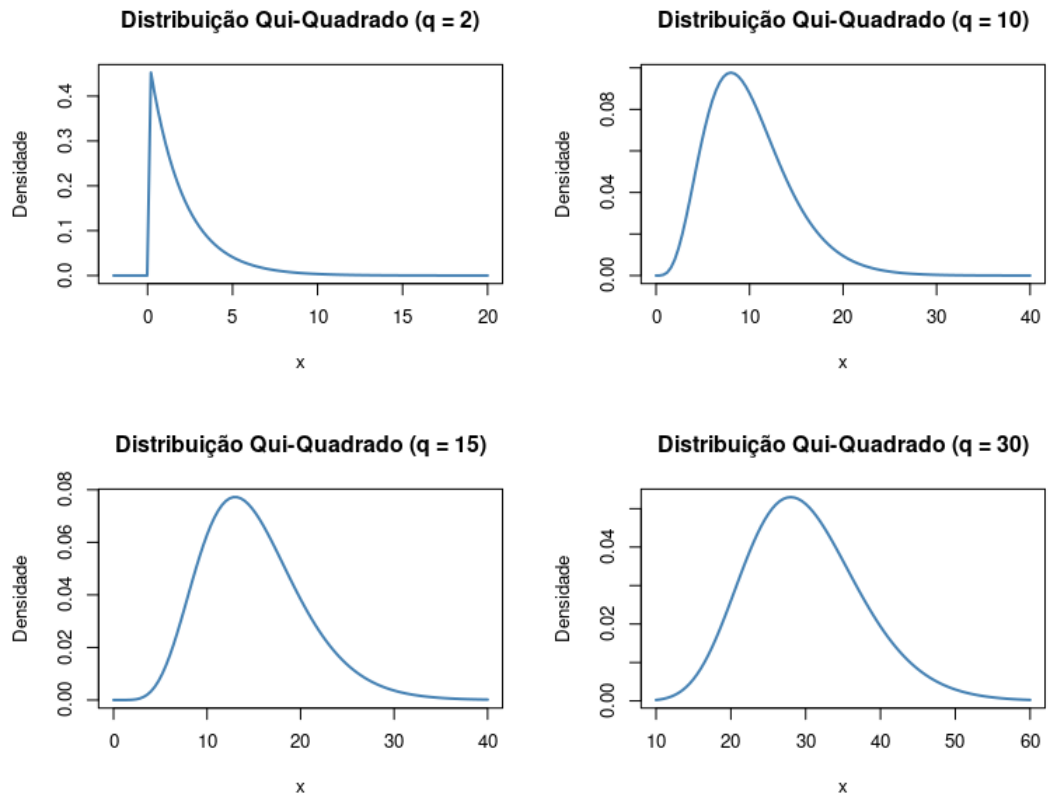
$$\sigma^2 = Var(X) = 2q, \tag{4.23}$$

em que a média é igual ao número dos graus de liberdade e a variância duas vezes esse valor.

[3]Densidade Normal é uma função contínua e simétrica

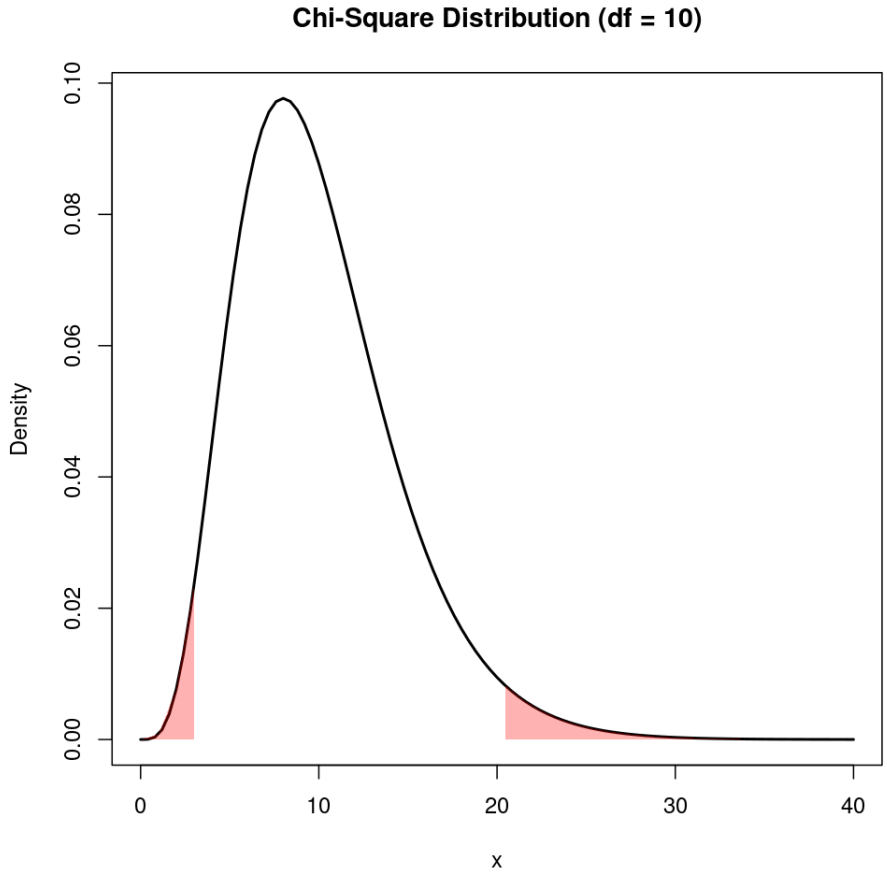
O maior uso da distribuição qui-quadrado reside no teste de hipótese, visto sua relação com a distribuição Normal Padronizada. Seus valores começam em zero e vão ao infinito, jamais assumindo valores negativos, uma vez que toma-se o quadrado dos valores de uma Normal Padronizada. A Figura 4.1 é um exemplo de diferentes distribuições Qui-Quadrados geradas a partir de diferentes graus de liberdade:

FIGURA 4.1. Densidades qui-quadrado



O valor obtido pela estatística Qui-Quadrado (como em (4.18)) que revela a discrepância entre valores observados e esperados para um conjunto de dados é aplicado à distribuição para buscar evidências contra ou a favor da hipótese nula. Caso a estatística seja um valor pequeno, ou seja, os valores observados estão alinhados com os valores esperados, então a estatística falhará em rejeitar a hipótese nula. Por outro lado, caso os valores sejam discrepantes, então a estatística cairá na região de rejeição. A figura 4.2 mostram as regiões de rejeição para uma distribuição qui-quadrado com 10 graus de liberdade:

FIGURA 4.2. Região de Rejeição qui-quadrado



4.3.2 Anova

Muitas vezes pesquisadores precisam determinar se existem diferenças entre os resultados de diversos grupos. Embora visualmente dados agregados em diferentes grupos pareçam ser diferentes, um teste confirmatória é importante. Um método bastante utilziado é Análise de Variância (ANOVA). O ponto de partida são duas hipóteses conhecidas como hipótese nula H_0 e hipótese alternativa H_1. A primeira estabelece que não existe diferença entre os grupos, ou médias, da análise, enquanto a segunda determina se há alguma diferença. Portanto, por serem complementares, pode-se aceitar ou rejeitar a H_0 a favor da H_1 (ST et al., 1989). ANOVA é uma técnica estatística poderosa que é geralmente aplicada para mostrar grau de variabilidade entre dois ou mais grupos através da variância entre grupos e a variância dentro dos grupo. O calculo do coeficiente F da Análise de Variância:

$$F = \frac{MST}{MSE}, \tag{4.24}$$

onde,

1. MST - Média da Soma ao Quadrado do Tratamento;

2. MSE - Média da Soma de Quadrados do Erro.

Algumas suposições são consideradas ao usar ANOVA. Primeiro, assume-se que os dados da amostra tenham sido selecionados de maneira independentes. Segundo, a variância dos dados entre diferentes grupos deve ser similar. Por último, cada amostra é proveniente de uma Distribuição Normal (LARSON, 2008).

O resultado em (4.24) é chamado de Razão F. Se a hipótese nula é verdadeira, então não há diferença signifcativa entre os grupos e o resultado F deverá ser próximo de 1. Por outro lado, quando a hipótese alternativa é verdadeira, isto é, há diferença significativa em ao menos um grupo, então o valor F será baixo o suficiente ($p < 0,05$), evidenciando que as médias dos grupos variam o suficiente para ser significamente diferentes (CONNELLY, 2021).

Embora o teste F traga evidências se existem grupos significamente diferentes, ele não indica qual ou quais são divergentes dos demais. Ainda é necessário determinar quais pares de grupos são signifcamente diferentes através de um método *post hoc*. Existem diferentes teste que podem realizar essa análise, utilizando esse trabalho o método a seguir.

4.3.3 Teste de Tukey

Quando um pesquisador precisa determinar se três ou mais grupos de amostras independentes apresentam diferenças significativas, usa-se algum método de comparação múltipla, em que um dos mais populares é Teste de Tukey[4] (SMITH, 1971).

[4]Também é conhecido como Teste de Tukey HSD (Teste de Tukey da Diferença Honestamente Significativa)

O Teste de Tukey compara todos os possíveis pares de médias. Ele baseia-se na diferença mínima significativa (D.M.S) através dos percentis do grupo. Fundamentalmente o resultado apresentará:

1. Diferença Mínima Signifcativa, módulo da diferença média entre os pares de grupos.

2. Intervalo de Confiança, nível de confiabilidade do resultado.

3. p-valor, quando um par de médias apresenta diferença significativa ($p < 0,05$)

O tamanho de uma população em estudo sempre possui importância para uma análise estatística, uma vez que o tamanho da população cresce a variância diminui. Portanto, recomenda-se amostras grandes para desempenhar comparações de médias entre grupos para minimiar o erro do tipo 1 no Teste Tukey para Comparações Múltiplas. (NANDA et al., 2021).

4.4 Aplicação de Análise de Correspondência em Software Computacional

Um dos *softwares* estatísticos mais utilizados na atualidade é o R. Baseado sistema *open source* da linguagem S, essa linguagem computacional permite uma variedade de análises através das suas funções nativas e bibliotecas disponíveis para importação (MICHEAUX et al., 2013).

Há vastas opções de pacotes que podem ser utilizados para realizar uma AC no R. Um dos mais usados é o *FactoMineR package*. Esse possui funções que permitem análises exploratórias de agrupamento, visualização e descrição de dataframes[5]. Para realizar análise de correspondência será utilizada a função *CA()* (KASSAMBARA, 2017).

[5]DataFrame é uma estrutura de dados bidimensional com os dados alinhados de forma tabular em linhas e colunas, mutável em tamanho, similar à uma matriz, mas potencialmente heterogênea entre suas colunas.

4.4.1 Formato dos Dados

Os dados disponível no pacote *FactoMineR* são resultados de uma pesquisa sobre crianças do primário que sofreram intoxicação alimentar. A tabela de dados possui 55 observações e 15 variáveis. Cada observação é um aluno de uma escola. Autor seleciona parte dos dados e reproduz em uma plotagem as primeiras linhas, conforme segue:

```
poison.active <- poison[1:55, 5:15]
head(poison.active[, 1:6], 3)
```

FIGURA 4.3. Exemplo Tabela de Dados

```
##   Age Time   Sick Sex   Nausea Vomiting Abdominals
## 1   9   22 Sick_y   F Nausea_y  Vomit_n     Abdo_y
## 2   5    0 Sick_n   F Nausea_n  Vomit_n     Abdo_n
## 3   6   16 Sick_y   F Nausea_n  Vomit_y     Abdo_y
```

Ainda, há uma função nativa do *software R* que permite visualizar a frequências das categorias das variáveis:

4.4.2 MCA

A função *MCA()* do pacote *FactoMiner* é utilizada através dos seguintes argumentos:

1. X: tabela de dados contendo n linhas (observações) e p colunas (variáveis categóricas).

2. ncp: número de dimensões mantidas no resultados final.

3. graph: valor lógico. Se verdadeiro, um gráfico é produzido.

Abaixo o código R é implementado:

```
res.mca <- MCA(poison.active, graph = FALSE)
```

O resultado serão várias informações distribuídas em diferentes objetos tipo listas e tipo matrizes resultantes da função.

4.4.3 Visualização e Interpretação

Em ordem para facilitar a interpretação e visualização da análise de correspondência múltipla, existem funções que ajudam extrair a informação gerada. Várias dessas funções etão no *factorxtra package*, sendo apresentadas a seguir:

```
res.mca <- MCA(poison.active, graph = FALSE)
```

4.4.3.1 Função get_eigenvalue

Parte da variação contida nos dados é retida em diferentes dimensões. Utilizando a função *get_eigenvalue* e *fviz_screeplot* alcança-se qual porcentagem da variabilidae é explicada pelo gráfico:

```
library("factoextra")
eig.val <- get_eigenvalue(res.mca)

fviz_screeplot(res.mca, addlabels = TRUE, ylim = c(0, 45))
```

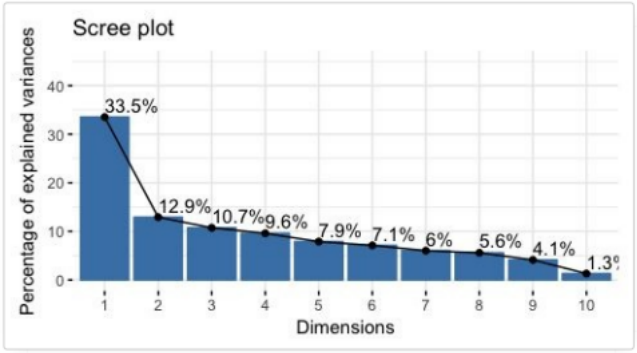


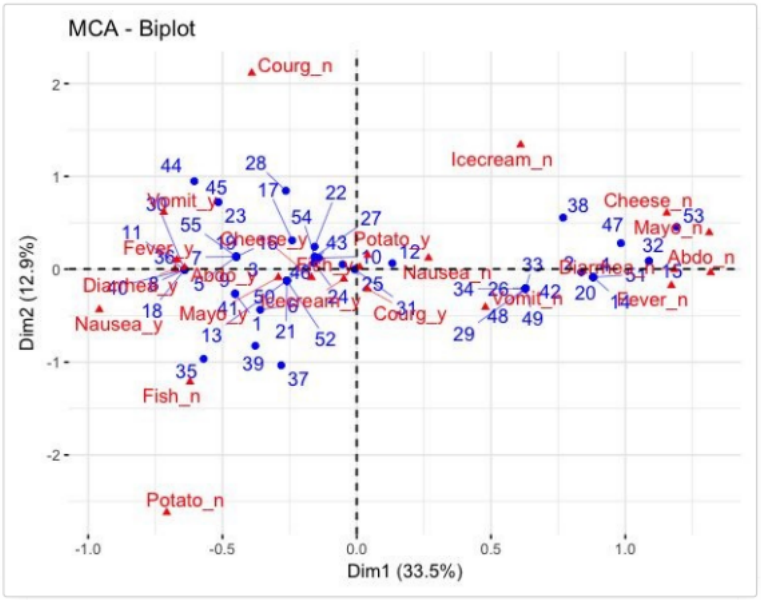
FIGURA 4.4. Percentual de Inércia Explicada

4.4.3.2 Função fviz_mca_biplot

Ao final, pode-se plotar um gráfico bidimensional para apresentar o mapa de correspondência entre as observações e variáveis categóricas do banco de dados. A função *fviz_mca_biplot* reproduz o seguinte resultado:

```
fviz_mca_biplot(res.mca,
repel = TRUE, # Avoid text overlapping (slow if many point)
ggtheme = theme_minimal())
```

FIGURA 4.5. Biplot



Fonte: Kassambara 2017

De acordo com o gráfico acima duas dimensões conseguem explicar 56% da variabilidade dos dados. A cor azul identificam os alunos, enquanto a cor vermelha os alimentos consumidos e as doenças associadas a cada estudante. Percebe-se que o número 38 possui associação com queijo e sorvete. Já o aluno 11 está associado com naúsea e febre. Portanto, a Análise de Correspondência serve como uma ferramenta exploratória para buscar eventuais associações que existam nos dados.

Capítulo 5

Resultados e Discussão

5.1 Base de Dados

O Enem é uma prova de admissão à educação superior. Realizada pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), autarquia federal vinculada ao Ministério da Educação (MEC), foi criada inicialmente em 1998 avaliar a qualidade do ensino médio no Brasil.

Como objetivo deste trabalho é analisar a desigualdade da educação básica brasileira, aplicou-se filtros para limitar respostas de estudantes com o perfil de análise deste trabalho. A primeira resposta filtrada foi a situação de conclusão do ensino médio. Somente respondentes que estavam cursando o Ensino Médio em 2021 e que concluiriam a escola naquele ano ou depois. Ainda, foram considerados alunos menores de 20 anos para limitar estudantes que permaneceram de forma contínua na Educação Básica. Por último, foram consideradas somente respostas com etinia identificada e que compareceram em todas as provas. Ao final o banco de dados teve a seguinte distribuição:

Dentre as cinco raças estão distribuídos 585.707,00 estudantes. Do total de respostas, aproximadamente 62% são do genêro feminino e 48% masculino. Os alunos originalmente pertencentes às diferentes escolas tiveram, respectivamente, a seguinte frequência: estadual (171.952), privada (96.292), federal (19.898) e municipal (2.289).

Algumas modicações foram implementadas para facilitar a AC. Ao invés de usar a escala de salários disponível no banco de dados uso-se a escala do IBGE (Instituto

Distribuição dos Aplicantes do Exame		
Observações por Raça	**Abs**	**%**
Indíginas	2.405	0,41%
Pretos	54.356	9,28%
Pardos	228.593	39,03%
Amarelos	11.713	9,28%
Brancos	288.640	49,28%
TOTAL	**585.707**	**100%**

TABELA 5.1. Dados Enem 2021, após filtragem de dados.

Brasileiro de Geográfia e Estatística) dividada em cinco estratos: A, B, C, D e E.

5.2 Análise Exploratória

A análise exploratória é um importante passo para compreensão do problema em que se está investigando. Quando se trabalha com informações acerca de grupos, utilizar gráficos mostra-se essencial na busca de evidenciar diferenças existentes.

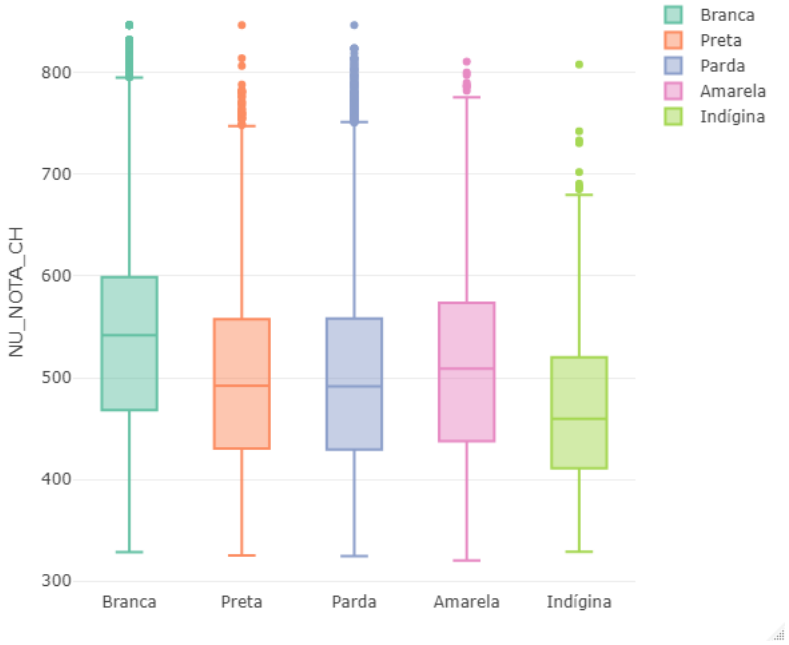
O *Box Plot* foi utilizado para estudar como estão distribuídas as notas obtidas no exame entre as diferentes raças dos estudantes. Se existe diferença na aprendizagem e formação de conhecimento dos estudantes do Ensino Básico brasileiro, as medianas devem divergir entre os agrupamentos de estudantes. Ainda, pode ser que as divergências sejam maiores em alguma área do conhecimento da prova.

5.2.1 Prova de Ciências Humanas

A prova de Ciências Humanas avalia o conhecimento dos alunos em problemas históricos e da atualidade, sem exigir conhecimento de datas ou fatos passados, e sim aspectos sociais, culturais e econômicos. As temáticas abordadas são: Filosofia, Geografia, História e Sociologia. Realizando a análise exploratória nas notas obtidas no Enem, verifica-se que a raça branca possui mediana superior às demais. O menor desempenho é obtido pela raça indígina.

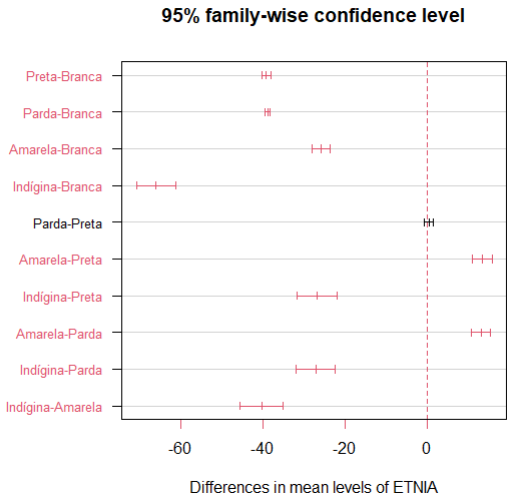
Aplicando análise de variância para confirmar se há diferença significativa entre as raças, obtém-se resultado de p-valor $< 0,001$. Portanto, conclui-se que existe ao

FIGURA 5.1. Nota da Prova de Ciências Humanas



menos uma raça com médias destoantes das demais. Aplicando Teste de Tukey para comparações múltiplas para comprar pares de médias:

FIGURA 5.2. Nota da Prova de Ciências Humanas

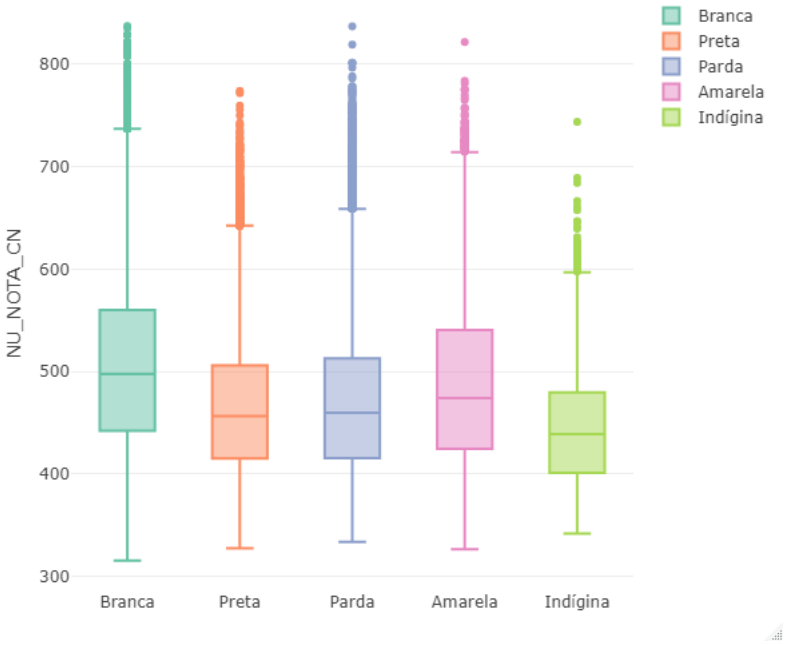


Única raça que não apresentou diferença significativa entre média foram parda e preta. Chama atenção que a maior desigualdade está entre raças Indígenas e Brancas, uma diferença de aproximadamente 60 pontos.

5.2.2 Prova de Ciências da Natureza

Nesta parte do exame o aplicante é testado nas áreas de Biologia, Física e Química. Ao transcurso das questões são abordados problemas de cálculos e fórmulas, mas não sendo esses o foco da avaliação. Objetivo maior é determinar a relação com outros temas da atualidade presente nestes conteúdos como, por exemplo, mudança climática. Diagrama de Caixas apresenta, novamente, mediana mais alta para raça branca, seguida pela raça amarela e, mais abaixo, parda e preta. Raça indígina aparece menor em relação às demais. Diagrama de Caixa:

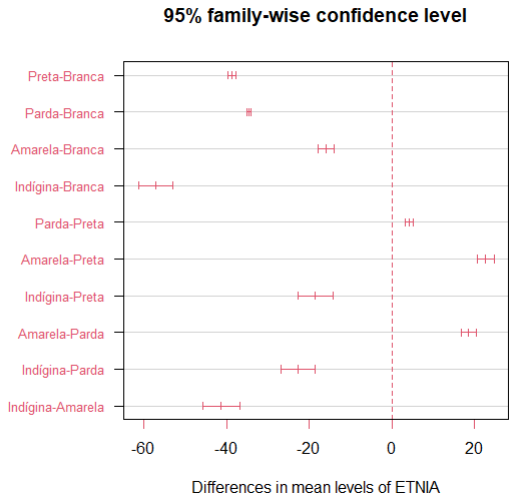
FIGURA 5.3. Nota da prova de Ciências da Natureza



A Análise de Variância foi testada para confirmar se há diferença significativa entre as raças, alcançando valor de p-valor $< 0,001$. Conclui-se, portanto, que existe ao menos uma raça com médias divergentes. Aplicando Teste de Tukey para Comparações Múltiplas:

Todas raças apresentaram médias significativamente diferentes. A maior diferença permaneceu entre estudantes brancos e indígenas. Já a menor diferença de média entre grupos de pretos e pardos.

FIGURA 5.4. Nota da prova de Ciências da Natureza



5.2.3 Prova de Linguagens e Códigos

Áreas de Linguagens, Códigos e suas Tecnologias traz conteúdos de Língua Portuguesa, Artes e Língua Estrangeira. Objetivo é verificar a compreensão dos aplicantes em interpretação de textos, desenhos e entendimento em idioma de inglês ou espanhol. Resultados obtidos:

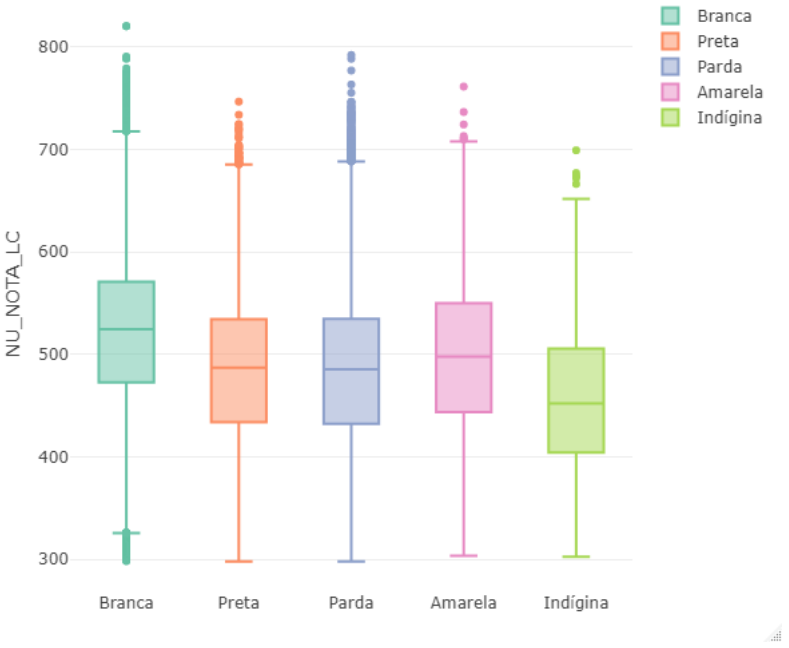
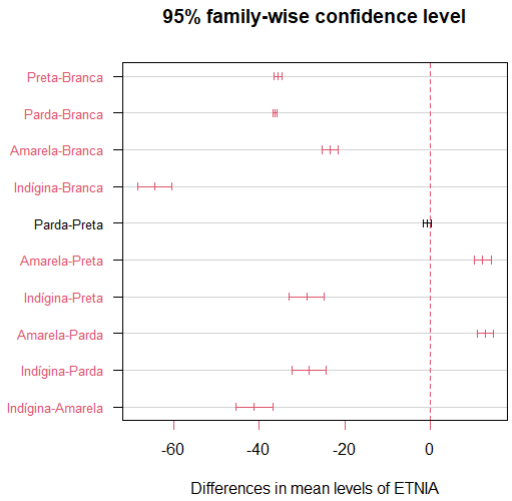


FIGURA 5.5. Nota da Prova de Linguagens e Códigos

Embora visualmente se perceba possíveis diferenças entre os grupos, novamente usa-se teste confirmatório Anova que apresenta p-valor $< 0,001$. Aplicando Tukey HSD para comprar pares de médias:

FIGURA 5.6. Nota da prova de Ciências da Natureza



Raças de pretos e pardos não obtiveram médias significativamente diferentes. Os grupos brancos e indígenas foram as maiores diferenças, seguidos pelos pares de média de amarelos e indígenas.

5.2.4 Prova de Matemática

Os alunos são testados em disciplinas de geometria, probabilidade, estatística, equações e outras. De maneira geral, são aplicados problemas de ordem mais lógica, sem exigir demasiada decoração de cálculos e fórmulas.

Ao testar a diferença significativa entre as raças, novamente tem-se $p - \text{valor} < 0,001$. Aplicando Tukey HSD para comprar pares de médias:

As raças apresentaram todas diferenças significativas entre médias. A maior diferença aconteceu entre grupo de brancos e indígenas, ainda maior do que nas demais provas. A menor diferença foi entre aplicantes do exame das raças de pretos e pardos.

FIGURA 5.7. Nota da Prova de Matemática

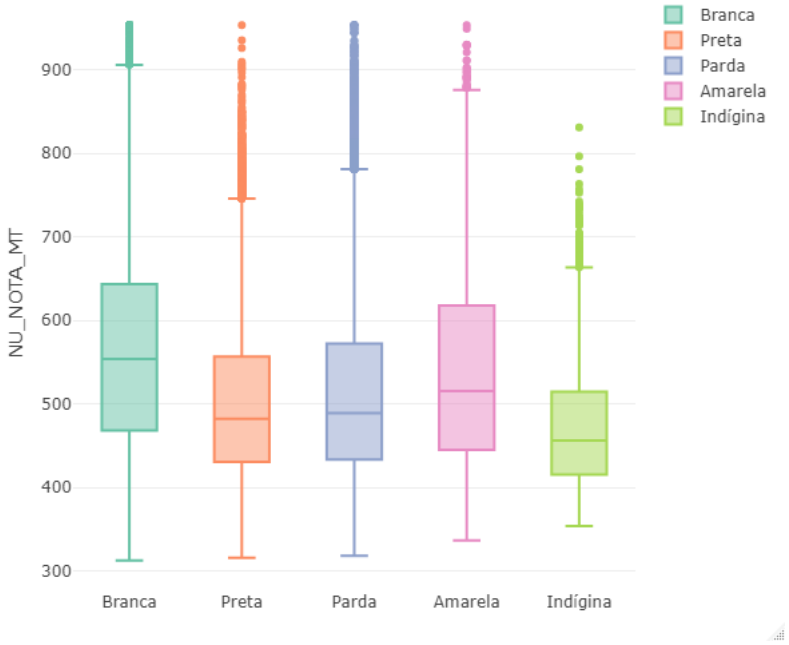
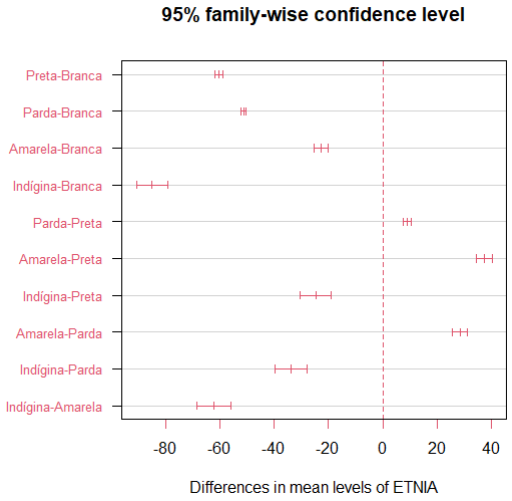


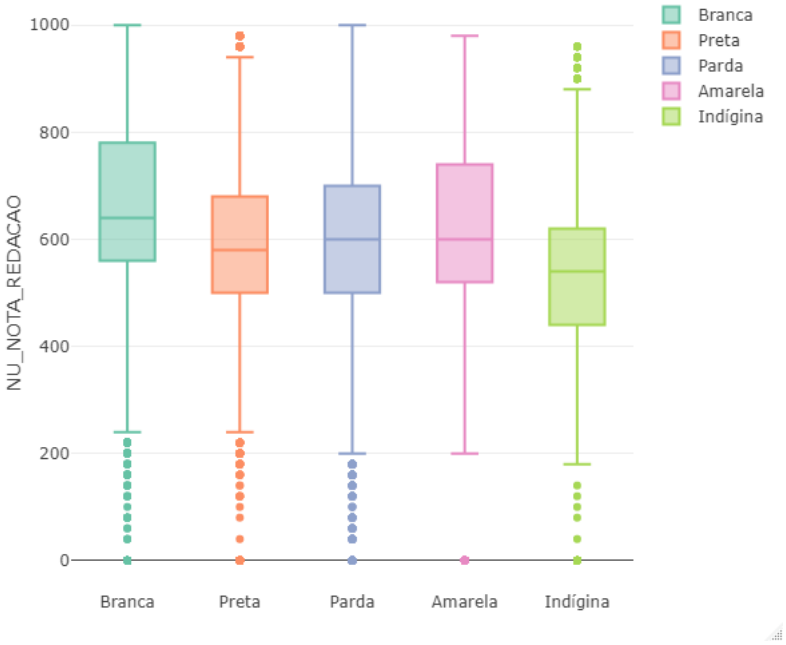
FIGURA 5.8. Nota da prova de Ciências da Natureza



5.2.5 Prova de Redação

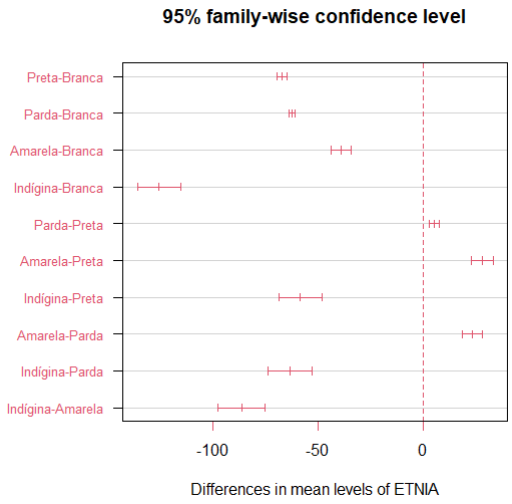
A redação do Enem é composta por uma frase-tema, em que o aluno precisa dissertar de maneira argumentativa sobre um problema atual da sociedade brasileira. Fuga do tema ou textos que não sigam a estrutura exigida podem zerar a prova de redação.

FIGURA 5.9. Nota da prova de redação



Diferentemente das demais provas, a prova de redação apresenta muitos outliers abaixo do primeiro quartil. Há indícios que muitos alunos possuem dificuldade em seguir as normas de avaliação da redação, sendo um sinal importante a ser avaliado pelos formuladores de política pública do Brasil. Resultado da Anova foi de diferenças significativas para as raças de p-valor $< 0,001$. Aplicando Tukey HSD para comparar pares de médias:

FIGURA 5.10. Nota da Prova de Redação



Embora a maior diferença tenha ocorrido de novo para o par de raças de brancos e indígenas, chama atenção que a diferença foi a maior observada em toda análise exploratória. A menor diferença ocorreu entre o par de grupos de pretos e pardos.

5.3 Análise de Correspondência

Mostra-se importante compreender a relação dos grupos de cor com as variáveis de cada dimensão apresentadas por Silva e Hasenbalg (2000). Utilizando a Análise de Correspondência, espera-se que grupos homogêneos apresentam resultados semelhantes. Por outro lado, quando os grupos apresentam dinâmicas distintas, a análise de correspondência apresentará os agrupamentos afastados uns dos outros no gráfico. Utilizando os valores quiquadrado foi calculada a Distância Euclidiana para medir o quão longe esses grupos estão uns dos outros em um plano geométrico.

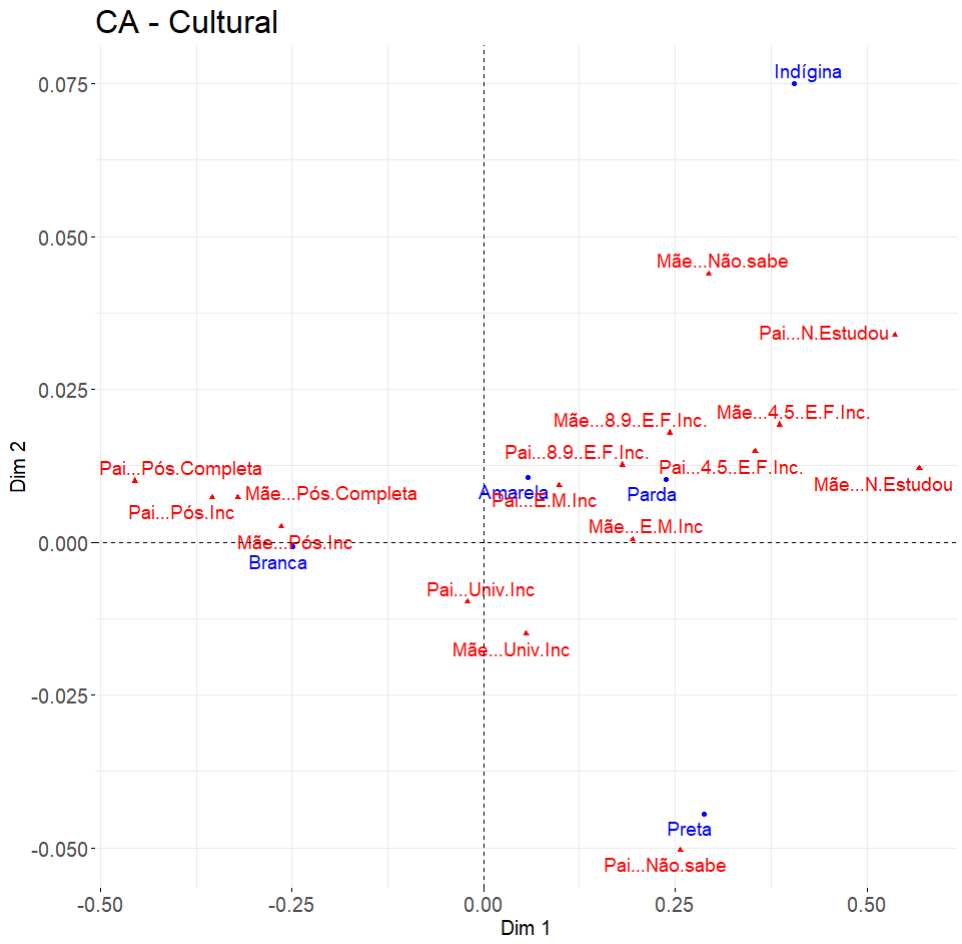
5.3.1 Capital Cultural

Conforme visto no capítulo anterior, compreender os recursos educacionais dos pais, ou tutores, dos alunos é importante para entender o fenômeno da desigualdade da Educação Básica. Os dados socioeconômicos do Enem possuem dados sobre grau de escolaridade dos pais ou responsável pelos alunos. Todas as variáveis selecionadas foram testadas no teste do qui-quadrado, obtendo estatística p-valor $< 0,001$.

Antes de produzir a AC, verificou-se a porcentagem de variância explicada por um gráfico bidimensional, alcançando praticamente 100% de variabilidade que pode ser esclarescida. AC para variáveis culturais na figura 5.11:

A AC resultou que a raça branca está relativamente mais associada com pai e mãe com ensino superior completo ou incompleto. Por outro lado, as demais raças estiveram mais associadas com com Educação Básica incompleta. O grupo indígina está na situação mais vulnerável, mais associada com mãe e pai que não sabem escrever ou que não estudaram. Ademais, como as raças preta e indígina estão bastante afastadas da origem, pode-se supor que existe uma alta divergência desses grupos em relação à estudantes brancos, amarelos e pardos. As raças mais similares foram amarela e parda.

FIGURA 5.11. AC - Cultural



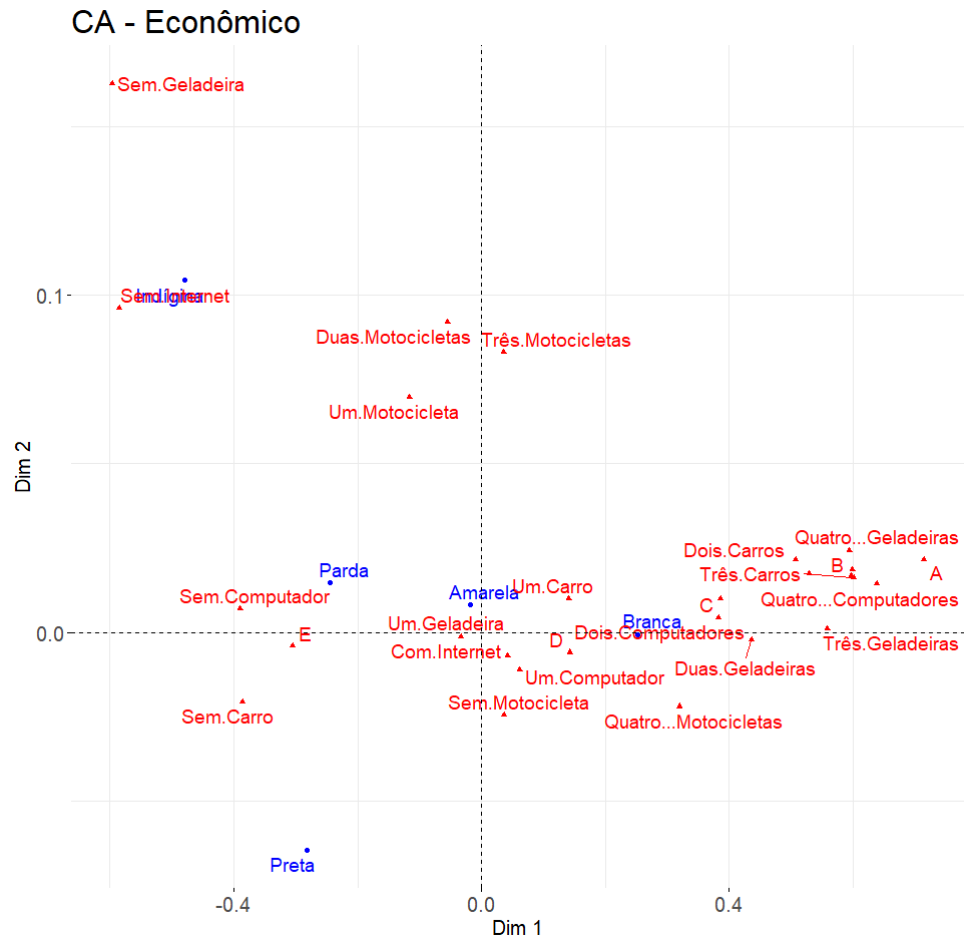
5.3.2 Capital Econômico

Não menos importante, os recursos econômicos também são fundamentais para a aprendizagem dos estudantes. Famílias com mais bens materiais possuem mais facilidade em adquirir educação, conseguem proporcionar melhores ambientes para estudo e outras facilidades que impactam no desenvolvimento escolar; realidade bastante diferente de alunos mais pobres. Novamente, testou-se qui-quadrado para as variáveis selecionadas, resultado significativas com p-valor $< 0,001$.

A porcentagem da variância explicada pelo mapa de correspondência foi de aproximadamente 100% para um gráfico bidiomensional. AC para variáveis econômicas na figura 5.12:

Considerando o aspacto material, raças amarela e branca estão próximas, indi-

FIGURA 5.12. AC - Econômico



cando similaridade. Por outro lado, os grupos de alunos pretos e indíginas estão bastante afastados da origem evidenciando sua desvantagem em possuir bens materiais como: geladeira, carro, computador, motocicleta.

Capítulo 6

Conclusões

De uma maneira geral, ao observar análise exploratória do desempenho nas diferentes áreas do Enem, há um padrão que se repete. Alunos de etnia branca apresentam medianas superiores, seguidos pelas etnias amarela, parda, preta e, por último, discentes da etnia indígina. Logo, entende-se que exista uma estrutura de desigualdade presente nas diversas etnias. Análises estatística através da Anova e Teste de Tukey dão suporte à compreensão que o desenvolvimento escolar ocorre de maneira distinta entre os alunos.

Usando Análise de Correspondência para estudar a relação das variáveis socioeconômicas dos aplicantes do exame, pode-se observar algumas características associadas para cada etnia. Capital Cultural mostrou que brancos possuem famílias com melhor grau de instrução, bastante superior aos demais. Posicionado de maneira oposta, as famílias da etnia indígina estão muito pouco associados à níveis de instrução de ensino superior ou educação básica completa. Etnias de pardos e amarelos estão mais próximos e localizados perto da origem, trazendo evidências de menor diferença estatística do que as demais em relação a esse conjunto de variáveis.

O padrão se repete para análise do Capital Econômico. Etnia branca está mais associada bens materais como computadores, carros, geladeiras, bastante diferente da etnia preta que está mais associado a não ter carros ou motocicletas. Causa preocupação que indíginas estejam altamente associados a falta de bens materiais, sem acesso à internet e geladeira.

Poranto, fica-se subentendido que parte da desigualdade no grau de instrução

dos alunos está associada a carência de instrução dos pais e insuficiência econômica. Há uma visível diferença no nível de escolaridade dos pais dos alunos brancos em relação aos não brancos, reforçando um ciclo vicioso que mantém o caráter desigual da sociedade brasileira. Fatores como esse promovem vantagens do acesso ao Ensino Superior e ao mercado de trabalho para alunos com famílias de capital cultural e econômico mais alto. Políticas públicas tendem a reduzir essas discrepâncias. Embora muitas estejam sendo realizadas hoje, mostra-se importante acompanhar os resultados e medir sua eficiência para lograr o sucesso desejado mais rápido. Oferecer qualificação para os pais de crianças com contexto familiar de vulnerabilidade social pode ser uma ação afirmativa eficiente e inovadora, contribuindo não apenas para uma melhora da trajetória escolar dos estudantes, mas reduzindo o déficit educacional da população adulta.

Dados do IBGE de 2019 mostram que a democratização ao acesso escolar parecer ser uma barreira em superação, principalmente nos anos iniciais, visto que há pouca diferença na frequência entre alunos negros e brancos. Contudo, ao longo da trajetória escolar, esse ritmo não se mantém. Do ponto de vista do Capital Cultural, mostra-se importante o estado brasileiro desenvolver políticas para melhorar o nível de instrução de adultos com filhos. Possivelmente reforçará um ciclo virtuoso de desenvolvimento social, que impactará seus filhos em fase escolar. Os dados revelam uma situação que merece atenção. O grupo indígina encontram-se bastante longe dos demais, há fortes indícios que essa etnia está em uma situação de vulnerabilidade ainda maior do que etnia preta ou parda. Por representar um número absoluto inferior, iniciativas que visem reduzir sua desvantagem escolar e laboral podem ter resultados rápidos.

Referências Bibliográficas

ABDI, H.; BÉRA, M. *Correspondence Analysis.* 2014.

ANDRADE, G. C. de; SANTOS, S. A. Decomposiçao em valores singulares e tecnicas de compressao de dados. *Proceeding Series of the Brazilian Society of Computational and Applied Mathematics*, v. 7, n. 1, 2020.

ÁVILA, G. Euclides, geometria e fundamentos. *Revista do professor de matemática*, v. 45, 2001.

BARRETO, P. C. d. S. Gênero, raça, desigualdades e políticas de ação afirmativa no ensino superior. *Revista Brasileira de Ciência Política*, SciELO Brasil, p. 39–64, 2015.

BENAKOUCHE, T. Tecnologia é sociedade: contra a noção de impacto tecnológico. *Cadernos de pesquisa*, v. 17, p. 1–28, 1999.

BENZÉCRI, J.-P. El análisis de correspondencias. *Cahiers de l'analyse des données*, v. 2, n. 2, p. 125–142, 1977.

BLASHFIELD, R. K.; ALDENDERFER, M. S. The literature on cluster analysis. *Multivariate behavioral research*, Taylor & Francis, v. 13, n. 3, p. 271–295, 1978.

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

CASTRO, J. A. d. Evolução e desigualdade na educação brasileira. *Educação & Sociedade*, SciELO Brasil, v. 30, p. 673–697, 2009.

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

CORTI, A. P. As diversas faces do enem: análise do perfil dos participantes (1999-2007). *Estudos em Avaliação Educacional*, v. 24, n. 55, p. 198–221, 2013.

DOEY, L.; KURTA, J. Correspondence analysis applied to psychological research. *Tutorials in quantitative methods for psychology*, v. 7, n. 1, p. 5–14, 2011.

ELBERS, C.; LANJOUW, P. F.; MISTIAEN, J. A.; ÖZLER, B.; SIMLER, K. On the unequal inequality of poor communities. *The World Bank Economic Review*, Oxford University Press, v. 18, n. 3, p. 401–421, 2004.

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

GOES, E. F.; RAMOS, D. d. O.; FERREIRA, A. J. F. Desigualdades raciais em saúde e a pandemia da covid-19. *Trabalho, Educação e Saúde*, SciELO Brasil, v. 18, 2020.

GREENACRE, M. J.; BLASIUS, J. *Correspondence analysis in the social sciences: Recent developments and applications*. [S.l.: s.n.], 1994.

GUIMARÃES, R. de O. Desigualdade salarial entre negros e brancos no brasil: discriminação ou exclusão? *Econômica*, v. 8, n. 2, 2006.

HAIR, J. F.; BLACK, W. C.; BABIN, B. J.; ANDERSON, R. E.; TATHAM, R. L. *Análise multivariada de dados*. [S.l.]: Bookman editora, 2009.

INFANTOSI, A. F. C.; COSTA, J. C. d. G. D.; ALMEIDA, R. M. V. R. d. Análise de correspondência: bases teóricas na interpretação de dados categóricos em ciências da saúde. *Cadernos de Saúde Pública*, SciELO Brasil, v. 30, p. 473–486, 2014.

JACCOUD, L. d. B. O.; SILVA, F. A. B. d.; DELGADO, G. C.; CASTRO, J. A. d.; JR, J. C. P. C.; THEODORO, M. L.; BEGHIN, N. Questão social e políticas sociais no brasil contemporâneo. Instituto de Pesquisa Econômica Aplicada (Ipea), 2009.

KASSAMBARA, A. *Practical guide to principal component methods in R: PCA, M (CA), FAMD, MFA, HCPC, factoextra*. [S.l.]: Sthda, 2017. v. 2.

LACHI, R. L.; ROCHA, H. V. da. Aspectos básicos de clustering: conceitos e técnicas. *Núcleo de Informática Aplicada à Educação (Nied), UNICAMP-Instituto de Computação–Universidade Estadual de Campinas*, 2005.

LARSON, M. G. Analysis of variance. *Circulation*, Am Heart Assoc, v. 117, n. 1, p. 115–121, 2008.

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

MICHEAUX, P. L. de; DROUILHET, R.; LIQUET, B. *The R software*. [S.l.]: Springer, 2013.

NANDA, A.; MOHAPATRA, B. B.; MAHAPATRA, A.; MAHAPATRA, A. A. P. K.; MAHAPATRA, A. Multiple comparison test by tukey's honestly significant difference (hsd): Do the confident level control type i error. *IJAMS*, v. 6, p. 59–65, 2021.

NASCIMENTO, M. M.; CAVALCANTI, C.; OSTERMANN, F. Análise de correspondência aplicada à pesquisa em ensino de ciências. *Enseñanza de las ciencias: revista de investigación y experiencias didácticas*, n. Extra, p. 1319–1324, 2017.

NERI, M. *A escalada da desigualdade: Qual foi o Impacto da crise sobre distribuição de renda e pobreza?* [S.l.], 2019.

PIERI, R. *Retratos da educação no Brasil*. [S.l.]: Insper São Paulo, 2018.

SILVA, N. d. V.; HASENBALG, C. Tendências da desigualdade educacional no brasil. *Dados*, SciELO Brasil, v. 43, p. 423–445, 2000.

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

ST, L.; WOLD, S. et al. Analysis of variance (anova). *Chemometrics and intelligent laboratory systems*, Elsevier, v. 6, n. 4, p. 259–272, 1989.

THEODORO, M.; JACCOUD, L.; OSÓRIO, R.; SOARES, S. As políticas públicas e a desigualdade racial no brasil: 120 anos após a abolição. *Brasília: Ipea*, p. 69–99, 2008.

VICINI, L. Análise multivariada: da teoria à prática. Universidade Federal de Santa Maria, 2005.

WILLIAMSON, D. F.; PARKER, R. A.; KENDRICK, J. S. The box plot: a simple visual method to interpret data. *Annals of internal medicine*, American College of Physicians, v. 110, n. 11, p. 916–921, 1989.