UFRRJ

INSTITUTO DE CIÊNCIAS EXATAS

PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

DISSERTAÇÃO

Gestão escolar baseada em dados: análise da frequência estudantil e modelos de otimização

Carolina Pinheiro dos Santos

2025

UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

GESTÃO ESCOLAR BASEADA EM DADOS: ANÁLISE DA FREQUÊNCIA ESTUDANTIL E MODELOS DE OTIMIZAÇÃO

CAROLINA PINHEIRO DOS SANTOS

Sob orientação de
Ronaldo Malheiros Gregório

e co-orientação de
Felipe Leite Coelho da Silva
e
Marcelo Dib Cruz

Dissertação submetida como requisito parcial para obtenção do grau de **Mestre** no Programa de Pós-Graduação em Modelagem Matemática e Computacional, Área de Concentração em Modelagem Matemática e Computacional.

Seropédica, RJ, Brasil
Fevereiro de 2025

Universidade Federal Rural do Rio de Janeiro
Biblioteca Central / Seção de Processamento Técnico

Ficha catalográfica elaborada
com os dados fornecidos pelo(a) autor(a)

S237g

Santos, Carolina Pinheiro dos, 1994-
Gestão escolar baseada em dados: análise da frequência estudantil e modelos de otimização / Carolina Pinheiro dos Santos. - Nova Iguaçu, 2025.
69 f.

Orientador: Ronaldo Malheiros Gregório.
Coorientador: Felipe Leite Coelho da Silva.
Coorientador: Marcelo Dib Cruz.
Dissertação(Mestrado). -- Universidade Federal Rural do Rio de Janeiro, Programa de Pós-graduação em Modelagem Matemática e Computacional, 2025.

1. Gestão Escolar. 2. Frequência Estudantil. 3. ANOVA. 4. Modelo de Alocação de Salas de Aula. 5. Modelo Quadrático. I. Gregório, Ronaldo Malheiros, 1978-, orient. II. Silva, Felipe Leite Coelho da, 1981-, coorient. III. Cruz, Marcelo Dib, 1967-, coorient. IV Universidade Federal Rural do Rio de Janeiro. Programa de Pós-graduação em Modelagem Matemática e Computacional. V. Título.

MINISTÉRIO DA EDUCAÇÃO

UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO

INSTITUTO DE CIÊNCIAS EXATAS

PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

Seropédica-RJ, 27 de fevereiro de 2025.

CAROLINA PINHEIRO DOS SANTOS

Dissertação submetida como requisito parcial para a obtenção de grau de **Mestra**, no Programa de Pós-Graduação em Modelagem Matemática e Computacional PPGMMC, área de Concentração em Modelagem Matemática e Computacional.

DISSERTAÇÃO APROVADA EM 27/02/2025

Ronaldo Malheiros Gregório Drº UFRRJ (Orientador, Presidente da Banca)

Felipe Leite Coelho da Silva Drº UFRRJ (membro interno)

Marcelo Dib Cruz Drº (membro interno)

Erito Marques de Souza Filho Drº (UFRRJ- Externo ao Programa)

Luidi Gelaberti Simonetti Drº (UFRJ- Externo à Instituição)

ATA Nº ata/2025 - ICE (12.28.01.23)
(Nº do Documento: 586)

(Nº do Protocolo: NÃO PROTOCOLADO)

(Assinado digitalmente em 10/03/2025 16:24)
ERITO MARQUES DE SOUZA FILHO
PROFESSOR DO MAGISTERIO SUPERIOR
DeptTL/IM (12.28.01.00.00.90)
Matrícula: ###448#3

(Assinado digitalmente em 10/03/2025 18:42)
FELIPE LEITE COELHO DA SILVA
PROFESSOR DO MAGISTERIO SUPERIOR
DeptM (12.28.01.00.00.00.63)
Matrícula: ###398#2

(Assinado digitalmente em 11/03/2025 10:51)
MARCELO DIB CRUZ
PROFESSOR DO MAGISTERIO SUPERIOR
DCOMP (11.39.97)
Matrícula: ###680#1

(Assinado digitalmente em 10/03/2025 19:19)
RONALDO MALHEIROS GREGORIO
PROFESSOR DO MAGISTERIO SUPERIOR
DeptTL/IM (12.28.01.00.00.90)
Matrícula: ###696#7

(Assinado digitalmente em 10/03/2025 16:57)
LUIDI GELABERT SIMONETTI
ASSINANTE EXTERNO
CPF: ###.###.927-##

*Visualize o documento original em https://sipac.ufrrj.br/documentos/ informando seu número: **586**, ano: **2025**, tipo: **ATA**, data de emissão: **10/03/2025** e o código de verificação: **54c73f3d76***

Para Zuko, Shiro e Arya.

AGRADECIMENTOS

À minha família: minha mãe, Nice Maria, e meu pai, Rildo, pelo apoio à minha jornada acadêmica. Aos meus irmãos, por sempre torcerem por mim. Aos meus tios, pelo carinho e incentivo. E à memória de minha avó Adalgiza, que nos ensinou o valor da educação.

Ao meu orientador, professor Ronaldo, e aos meus coorientadores, professores Felipe e Marcelo, pela paciência, dedicação e pelas valiosas orientações ao longo desta pesquisa. A todo o corpo docente do Programa de Pós-graduação em Modelagem Matemática e Computacional (PPGMMC/UFRRJ), que contribuiu significativamente para minha formação, e aos funcionários do programa, sempre solícitos e prestativos.

A Renan, por constantemente me lembrar de que sou capaz e por emprestar seu ouvido aos meus intermináveis desabafos. A Andreza e Winnie, minhas companheiras no PPGMMC, por tornarem essa jornada mais leve.

À Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ), pelo apoio financeiro ao projeto E-26/210.191/2022, do qual esta pesquisa nasceu. O incentivo à pesquisa é fundamental para o avanço da ciência e a formação de pesquisadores, possibilitando o desenvolvimento de estudos que contribuem tanto para o conhecimento acadêmico quanto para a sociedade.

A todos que contribuíram, direta ou indiretamente, com o desenvolvimento deste trabalho, meus mais sinceros agradecimentos.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior — Brasil (CAPES) — Finance Code 001.

RESUMO

SANTOS, Carolina Pinheiro dos. **Gestão escolar baseada em dados: análise da frequência estudantil e modelos de otimização**. 2025. 69f. Dissertação (Mestrado em Modelagem Matemática e Computacional). Instituto de Ciências Exatas, Universidade Federal Rural do Rio de Janeiro, Seropédica, RJ, 2025.

A frequência estudantil na Educação Básica é um fator crucial para o sucesso escolar, mas os parâmetros que influenciam a decisão dos alunos de frequentar ou não a escola em determinados dias da semana ainda são pouco explorados na literatura. Este trabalho propõe um estudo sobre a influência dos dias da semana e dos quadros de horários de disciplinas sobre a frequência escolar, com o objetivo de otimizar a gestão de recursos materiais, humanos e financeiros. A análise de dados foi realizada a partir de um estudo de caso no Colégio Estadual Engenheiro Arêa Leão, em Nova Iguaçu, Rio de Janeiro, utilizando análise exploratória e análise de variância de 1-fator sobre os dados de frequência de 2022. Os resultados indicaram que a frequência média é maior nas terças-feiras, especialmente em disciplinas como Português e Matemática, e menor nas sextas-feiras, associada a disciplinas de natureza humana. A ponderação dos pesos disciplinares levou à adaptação de um modelo de múltiplas mochilas para a construção do quadro de horários de disciplinas por turmas, enquanto os dados de frequência inspiraram a criação de um modelo de minimização quadrático, baseado na variância dos dados, para gerar uma frequência esperada. Esses modelos podem contribuir para uma gestão escolar mais eficiente, auxiliando na alocação de recursos e no planejamento das atividades escolares.

Palavras-chave: Gestão Escolar, Frequência Estudantil, ANOVA, Modelo de Alocação de Salas de Aula, Modelo Quadrático.

ABSTRACT

SANTOS, Carolina Pinheiro dos. **Data-driven school management: analysis of student attendance and optimization models**. 2025. 69p. Dissertation (Master in Mathematical and Computational Modeling). Instituto de Ciências Exatas, Universidade Federal Rural do Rio de Janeiro, Seropédica, RJ, 2025.

School attendance in Basic Education is a crucial factor for academic success, yet the parameters influencing students' decisions to attend or skip school on specific weekdays remain underexplored in the literature. This study examines the influence of weekdays and the course schedule on school attendance, aiming to optimize the management of material, human, and financial resources. Data analysis was conducted through a case study at Colégio Estadual Engenheiro Arêa Leão, in Nova Iguaçu, Rio de Janeiro, using exploratory analysis and one-way ANOVA on 2022 attendance data. The results indicated that average attendance is highest on tuesdays, particularly in subjects like Portuguese and Mathematics, and lowest on fridays, mainly in humanities-related subjects. The weighting of subject impact led to the adaptation of a multiple knapsack model for constructing class schedules, while attendance data inspired the development of a quadratic minimization model, based on data variance, to generate an expected attendance distribution. These models can contribute to more efficient school management, aiding in resource allocation and the planning of educational activities.

Keywords: School Management, Student Attendance, ANOVA, Classroom Assignment Model, Quadratic Programming.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

1 INTRODUÇÃO

A **frequência estudantil** refere-se à presença dos estudantes nas aulas e atividades de sua instituição de ensino e constitui um fator determinante no desempenho acadêmico, além de ser um indicador essencial para a gestão educacional. As taxas ou médias de frequência são amplamente utilizadas por gestores e instituições governamentais para diferentes propósitos, incluindo aspectos legislativos, pedagógicos, acadêmicos e financeiros.

Estudos sobre fatores capazes de impactar a frequência vêm sendo utilizados para justificar políticas públicas visando o aumento das taxas. Algumas dessas medidas são voltadas à família, como o Programa Bolsa Família (PBF), instituído na Lei nº 10.836, de 9 de janeiro de 2004 (BRASIL, 2004), que trabalha sua melhoria de vida e *status* socioeconômico aliados ao cumprimento de uma taxa mínima de frequência dos dependentes em idade escolar. Outras medidas buscam tornar o ambiente de aprendizagem mais atrativo, por exemplo, através da oferta de merenda, regulamentada pela Lei nº 11.947, de 16 de junho de 2009, que estabelece diretrizes para a alimentação escolar (obrigatória na educação básica pública) e outros aspectos relacionados à educação básica no Brasil (BRASIL, 2009).

O presente trabalho busca relacionar os padrões de frequência dos estudantes das escolas estaduais do Rio de Janeiro a fatores internos à escola, de responsabilidade da gestão escolar, que influenciam sua assiduidade, visando embasar a formulação de estratégias que possam reduzir impactos negativos no processo ensino-aprendizagem. O incentivo à frequência é amplamente justificado pelo seu impacto no **sucesso escolar**, conforme documentado na literatura ao longo dos anos. Portanto, antes de definir formalmente a pesquisa, para contextualizá-la e fundamentá-la, é essencial revisar as discussões acadêmicas sobre frequência escolar, que aqui será sinônimo de frequência estudantil, e analisar as disposições legais e diretrizes da educação brasileira.

1.1 SUCESSO ESCOLAR

Na época de 80, já era bem estabelecido que o tempo dedicado ao estudo tinha uma relação positiva com bons resultados escolares, ou sucesso escolar. Walberg (1988) aponta que existem nove fatores que influenciam a aprendizagem: sucessos em etapas escolares anteriores ou habilidades já adquiridas, desenvolvimento, motivação, tempo dedicado à aprendizagem, qualidade da instrução, ambiente familiar, moral da sala de aula, socialização fora da escola e o tempo dedicado ao lazer. Supondo as demais condições propícias à aprendizagem, o tempo dedicado ao estudo é o diferencial para alcançar bons resultados. Como normalmente os oito demais fatores variam muito e dificilmente estão todos em concordância, Walberg discute a

importância de algo além do tempo dedicado ao aprender: o tempo de produtividade.

A produtividade, de acordo com o autor, é tanto efetiva quanto eficiente para a aprendizagem, pois representa um melhor uso do tempo e traz benefícios à moral. No entanto, as salas de aula tradicionais representam um obstáculo ao tempo de produtividade, visto que não são capazes de atender às necessidades particulares de cada estudante, trabalhando de maneira singular com uma pluralidade estudantil, com diferentes velocidades de aprendizagem e conhecimentos prévios. Somente uma fração do tempo dedicado aos estudos é, de fato, produtiva, tornando o tempo um fator precioso e o processo de aprendizagem mais dependente da maturidade e atividade do aluno. Para Walberg, mais do que comparecer às aulas, é também importante engajar nas atividades escolares. Para a escola, resta o desafio de tornar as aulas, as atividades e o ambiente escolar mais atrativos.

Desde então, diversos estudos têm evidenciado a pertinência dos fatores sintetizados por Walberg. Com enfoque na frequência escolar como fator de sucesso, Silva Junior e Gonçalves (2016) mostram que a presença na creche e pré-escola tem impacto positivo e significativo estatisticamente nas notas de Português e Matemática do ensino fundamental. O estudo de Ansari e Pianta (2019) revela que crianças com um histórico de altas taxas de faltas nos primeiros anos de escolaridade tendem a manter padrões de baixa frequência na adolescência, o que impacta negativamente seus resultados acadêmicos e socio-comportamentais. Cicuto e Torres (2020) observam que o baixo engajamento, resultado de ausência nas aulas ou baixa participação durante as mesmas, comprometem a construção do conhecimento. Além disso, um dos fatores apontados como responsáveis pelo insucesso escolar, é a organização curricular, com sobrecarga de créditos.

Outros estudos abordam o fluxo escolar, que diz respeito à continuidade ou interrupção do processo de aprendizagem, levando em consideração a permanência dos estudantes na escola, bem como as taxas de evasão, reprovação e aprovação. Soares, Alves e Fonseca (2021) destacam a permanência dos alunos na escola como um fator essencial para a aprendizagem significativa. O trabalho desses autores propõe a inclusão da análise das trajetórias educacionais como uma ferramenta crucial no monitoramento da qualidade da educação nos municípios. Já no estudo de Pontili e Kassouf (2007) as autoras afirmam que para melhorar o nível médio de escolaridade do país deve-se elevar a frequência escolar e manter a criança na escola, garantindo o avanço dos níveis educacionais. A pesquisa evidencia que fatores internos à escola são capazes de influenciar o comportamento estudantil, revelando que a escolaridade do professor e a disponibilidade de bibliotecas e laboratórios nas escolas pode afetar de maneira positiva a frequência.

Embora diversos estudos ressaltem a importância da frequência escolar, a literatura sobre o tema ainda é limitada e desatualizada. Grande parte das pesquisas sobre frequência estudantil aborda fatores externos ao ambiente escolar, como questões socioeconômicas, raciais, de gênero e religiosas. Banerji e Mathur (2021) concluem que meninas e alunos mais velhos têm maior probabilidade de se ausentar, destacando também a influência de fatores religiosos e de classe social. Os autores argumentam que as altas taxas de matrícula não se refletem necessariamente na frequência escolar. Eles enfatizam que é necessário adotar medidas para aumentar a frequência e o engajamento dos alunos, pois esses fatores estão diretamente relacionados ao desempenho acadêmico e à qualidade da aprendizagem.

Há pouca ou nenhuma publicação sobre a relação entre a frequência escolar e as disciplinas ofertadas ou o quadro de horários. Silva (2023) investigou as razões que levam os estudantes a abrir mão de determinados tempos de aula. No entanto, a pesquisa revela que 8 entre 10 motivos relatados pelos próprios alunos estão relacionados a questões de saúde, sem fornecer informações sobre as aulas das quais se ausentaram.

1.2 LEIS E PROGRAMAS

A Constituição Federal de 1988 (BRASIL, 2024), conhecida como "Constituição Cidadã", institui, no artigo 208, a educação básica obrigatória e gratuita dos 4 aos 17 anos, assegurada também para os que não tiveram acesso na idade própria. Já no artigo 214, estabelece a criação do Plano Nacional de Educação (PNE), que define diretrizes, objetivos, metas e estratégias para a manutenção e o desenvolvimento do ensino em seus diversos níveis e modalidades, visando, entre outras coisas, a melhoria da qualidade do ensino.

O PNE, aprovado pela Lei nº 13.005/2014 (BRASIL, 2014) e vigente até 2024, estabeleceu como meta a universalização do Ensino Fundamental e a garantia de que pelo menos 95% dos alunos o concluíssem na idade recomendada. No entanto, essas metas foram impactadas pela pandemia nos anos de 2020 e 2021, resultando em uma cobertura do ensino de apenas 95,9%, enquanto a taxa de conclusão na idade adequada ficou em 81,1%. (SENADO FEDERAL, 2023, online).

A meta de conclusão no tempo adequado reflete, implicitamente, a expectativa de que os alunos tenham um desempenho satisfatório ao longo de sua trajetória escolar. De modo geral, as normativas educacionais enfatizam mais as taxas de matrícula do que a frequência escolar, embora a mera inclusão dos alunos no sistema de ensino não seja suficiente para assegurar seu sucesso acadêmico.

O Programa Bolsa Escola (BRASIL, 2001), precursor do PBF, foi uma iniciativa de transferência de recursos destinada à manutenção das crianças na escola, sendo uma das primeiras políticas públicas a obter sucesso em aumentar as taxas de frequência escolar. Isso se deve ao fato de que o Bolsa Escola exigia das crianças cadastradas uma frequência mínima de 85% para que a família recebesse o benefício. Lavinas e Barbosa (2001) destacam os impactos positivos do programa na educação, evidenciando a redução da pobreza a curto prazo e a sua eficácia no combate à repetência escolar, apontado pelas autoras como um dos principais fatores da evasão escolar no Brasil.

Ao ser institucionalizado, o Programa Bolsa Família manteve a exigência de frequência mínima de 85% para seus beneficiários em idade escolar. As análises de Melo e Duarte (2010) demonstram que, no ano de 2005, o programa contribuiu diretamente para o aumento da taxa de frequência escolar no Nordeste, com um incremento de 5,4 a 5,9 pontos percentuais. Porém, Cacciamali, Tatei e Batista (2010) em análise para o ano de 2004, apontam que simples transferências de recursos não eram suficiente para reduzir a incidência do trabalho infatil ou corrigir a dificuldade de acesso à escola para famílias de rendas extremamente baixas. Já Cavalcanti, Costa e Silva (2013) relatam impactos positivos do programa, afirmando que o aumento da frequência escolar é tão ou mais importante do que a transferência monetária em si pelo sucesso escolar ser capaz de aumentar as chances de melhoria de vida das famílias. Para os autores, o PBF era melhor justificado pelas condicionalidades do que pelo repasse de renda.

Desde então, o valor do benefício do PBF vem sendo reajustado, tendo sido alterado pela última vez no ano de 2016. Atualmente, o programa exige apenas 75% de frequência escolar mínima para seus beneficiários de seis a dezoito anos que não tenham concluído a educação básica. (BRASIL, 2023)

Os estudos a respeito do impacto do PBF apontam a importância de outras ferramentas para o controle do sucesso escolar. Pires (2013) reforça que a frequência escolar mínima foi estabelecida com o intuito de romper o ciclo intergeracional de pobreza, através de possíveis benefícios consequentes do sucesso escolar. Mas o autor afirma que a mera matrícula escolar ou a garantia do fluxo escolar não oferecem, por si só, possibilidades de maior mobilidade social se não for considerada a qualidade do ensino e o aproveitamento do aluno.

Dada a influência da frequência escolar sobre o sucesso acadêmico, e reconhecendo que nem todo o tempo gasto em sala de aula é efetivamente produtivo, garantir apenas uma taxa mínima de frequência não deve ser a única preocupação das normas educacionais. Essas normas precisam buscar a maximização da frequência por meio de ferramentas que identifiquem e analisem os fatores que a impactam, além de promover ações que visem corrigi-los. Estabelecer taxas de frequência fixas, por exemplo, desconsidera a distribuição das faltas entre as diferentes disciplinas, que possuem diferentes demandas horárias e distribuem-se ao longo da semana de forma desigual. Ou seja, um aluno que tende a faltar mais em um determinado dia pode comprometer seu desempenho no conjunto de disciplinas oferecidas nesse dia.

1.3 PROBLEMA DE PESQUISA

Este estudo busca descrever e analisar o comportamento da frequência estudantil em resposta a determinadas ações da gestão escolar. Especificamente, investiga-se se a organização do quadro de horários influencia a decisão dos alunos de comparecer ou não à escola em determinados dias da semana e, caso essa influência exista, de que forma ela se manifesta. Através da compreensão desse fenômeno, busca-se embasar estratégias que visem à melhoria das taxas de frequência e à melhoria do aproveitamento acadêmico.

Para isso, será realizada uma análise de dados buscando identificar padrões de frequência associados à distribuição das disciplinas ao longo da semana, investigando a hipótese de que certos arranjos do quadro de horários impactam positivamente a assiduidade, elevando as médias de frequência de um dia específico, enquanto outros podem desestimular a presença dos alunos, favorecendo faltas. Além disso, a metodologia estabelecida neste estudo também poderá servir como base para futuras pesquisas sobre a influência de outros fatores da gestão escolar, como o cardápio da merenda ofertada.

Com base nos dados analisados, também serão propostos modelos de otimização tanto para a construção de quadros de horário que consideram a distribuição de disciplinas pelo seu impacto na frequência, quanto para a geração de um modelo de frequência esperada, que seja capaz de permitir previsões mais realistas das quantidades diárias de alunos e a alocação mais eficaz de recursos. Esses modelos poderão auxiliar gestores escolares na tomada de decisões estratégicas, contribuindo para um planejamento mais dinâmico e alinhado às necessidades dos alunos e da instituição.

1.4 OBJETIVOS

A principal meta deste estudo é analisar a relação entre a organização do quadro de horários e a frequência estudantil, investigando se a distribuição das disciplinas ao longo da semana influencia a presença dos alunos e de que forma essa influência ocorre. A partir dessa análise, pretende-se desenvolver modelos de otimização que auxiliem a gestão escolar na alocação de disciplinas por tempos de aula e na previsão da frequência esperada. Para tal, foram definidos os seguintes objetivos específicos:

a. Analisar as variações entre as médias de frequência entre os dias da semana e os fatores associados a essas diferenças;

b. Verificar os impactos da distribuição de disciplinas nas médias de frequência diárias, determinando quais disciplinas exercem influência positiva ou negativa sobre a mesma;

c. Estabelecer um modelo de otimização para a organização do quadro de horários, utilizando técnicas estatísticas para balancear a alocação das disciplinas, com base em seus impactos sobre a frequência;

d. Estabelecer um modelo para estimar a frequência esperada.

1.5 METODOLOGIA

Este estudo adota uma abordagem quantitativa, fundamentada na análise estatística dos dados de frequência escolar e no desenvolvimento de modelos matemáticos para otimização da alocação de horários e previsão da frequência estudantil. A pesquisa é estruturada em duas partes que se complementam ao longo do trabalho.

A primeira parte concentra-se na análise de dados e utiliza os registros de frequência do Colégio Estadual Engenheiro Arêa Leão, localizado em Nova Iguaçu, Rio de Janeiro, referentes ao ano letivo de 2022. Foram considerados a frequência diária dos alunos por turma, a distribuição das disciplinas no quadro de horários e relatos de professores e gestores. Inicialmente, os dados passaram por uma análise exploratória para identificar padrões e formular hipóteses. Em seguida, foi realizada uma análise de variância de 1-fator para verificar se as médias de frequência variavam significativamente ao longo da semana, permitindo avaliar a influência da organização curricular na presença dos alunos.

Na segunda parte, com base nos padrões identificados na análise estatística, investiga-se suas implicações na modelagem do quadro de horários. Para isso, propõe-se um modelo baseado no problema de múltiplas mochilas, no qual as disciplinas são distribuídas ao longo da semana, considerando seus pesos disciplinares, definidos em função da frequência observada. O objetivo é minimizar as discrepâncias na distribuição da frequência estudantil ao longo da semana. Além disso, sugere-se um modelo de frequência esperada, formulado como um problema de minimização das distâncias quadráticas entre os dados. Esse modelo busca ajustar os dados de frequência, regularizando a variância e gerando projeções de auxílio ao planejamento escolar.

Os modelos estatísticos e matemáticos foram implementados em Python (Python Software Foundation, 2024, online), com o suporte de bibliotecas especializadas, cujos detalhes são apresentados ao longo do trabalho.

1.6 ORGANIZAÇÃO

O Capítulo 2 apresenta as ferramentas de análise unidimensional dos dados utilizadas na investigação sobre a frequência estudantil. Ele está estruturado em duas seções principais: a análise exploratória, que abrange medidas de posição, dispersão e a interpretação de gráficos, e a inferência estatística, que introduz conceitos fundamentais e detalha o modelo de análise de variância, além de discutir ferramentas auxiliares para a adequação dos dados ao modelo.

O Capítulo 3 apresenta a aplicação das ferramentas estatísticas ao estudo de caso. Ele descreve a coleta e tratamento dos dados, a forma que o modelo foi aplicado e discute os resultados da análise exploratória e da análise de variância.

No Capítulo 4, são explorados os aspectos teóricos da otimização, apresentando conceitos fundamentais de otimização linear, otimização inteira binária e otimização quadrática que inspiraram aplicações na gestão escolar.

No Capítulo 5, são discutidas as implicações dos resultados estatísticos nos modelos de otimização para a gestão escolar. São analisadas e sugeridas aplicações no problema do quadro

de horários e no problema da frequência esperada, assim como discutidos os seus impactos na tomada de decisão por parte dos gestores. Também são apresentadas possibilidades para a modelagem do problema da merenda.

Por fim, o Capítulo 6 conclui o trabalho sintetizando os principais resultados, as contribuições do estudo e discutindo brevemente as dificuldades encontradas e limitações dos modelos, assim como sugestões para trabalhos futuros.

2 ANÁLISE UNIDIMENSIONAL

Um estudo de caso é uma estratégia metodológica que se aprofunda em um fenômeno social complexo, buscando desvendar seus processos e mecanismos significativos, baseando-se em questões de pesquisa do tipo “qual”, “como” ou “por que” (Sátyro & D’Albuquerque, 2020). O estudo realizado se dá majoritariamente pela via quantitativa, ou seja, utiliza técnicas estatísticas para validar ou rejeitar hipóteses levantadas em cima de dados numéricos. O estudo é parcialmente guiado pelas experiências e discussões de professores e gestores envolvidos no caso, como será devidamente explicado no Capítulo 3. Aqui serão apresentadas as ferramentas estatísticas utilizadas na pesquisa.

Todo conteúdo de estatística experimental aqui abordado baseia-se fundamentalmente em três bibliografias: *An Introduction to Probability and Statistical Inference* (Roussas, 2003), *Design and Analysis of Experiments* (Montgomery, 2013) e Estatística Básica (Bussab & Morettin, 2011), que contém informações mais detalhadas e complementares sobre o tema, podendo ser utilizadas como material suporte ao leitor.

2.1 ANÁLISE EXPLORATÓRIA

Através da análise exploratória de dados (AED) é possível obter informações que indicam modelos plausíveis a serem utilizados para tentar responder as questões de pesquisa (Bussab & Morettin, 2011). Algumas das ferramentas-chave para a AED são as medidas de posição, de dispersão e as técnicas gráficas aqui apresentadas para conjuntos de dados quantitativos.

2.1.1 Medidas de posição e dispersão

As medidas de posição são três: **moda**, **mediana** e **média**. A moda é o valor mais frequente dentro de um conjunto de dados observáveis e ela não necessariamente é única. A mediana é o valor que ocupa a posição central do conjunto de dados, quando organizados em uma série crescente. Já a média é uma medida central que pode ser definida de diversas formas, porém aqui será considerada a média aritmética, que corresponde a soma dos valores observados dividida pela quantidade de dados do conjunto.

Essas medidas muitas vezes são utilizadas para representar um conjunto de valores de forma simplificada, e a escolha de se utilizar uma ou todas elas para analisar um conjunto de dados depende dos objetivos almejados. Uma das limitações da utilização das medidas de posição é que não trazem informações sobre a variabilidade dos dados dentro do conjunto de observações. (Bussab & Morettin, 2011)

Tentando suprir essa limitação, pode-se trabalhar com as medidas que medem a dispersão dos dados em torno da média, como o **desvio médio** e a **variância**, definidos respectivamente por:

$$\mathrm{dm}(X) = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}, \tag{2.1}$$

$$\mathrm{var}(X) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}, \tag{2.2}$$

onde x_i $(i = 1, \ldots, n)$ são as observações do conjunto e $\bar{x}$ é a média. Além destas, também utiliza-se o **desvio padrão**, definido como a raiz quadrada positiva da variância. Tanto o desvio médio quanto o desvio padrão são indicativos do erro médio obtido ao substituir as observações do conjunto por uma medida resumo, como a média. Porém, as medidas resumo e as medidas de dispersão são muito afetadas por valores extremos e não são capazes de traduzir a simetria da distribuição dos dados. (Bussab & Morettin, 2011)

Para tal, consideram-se os **quantis de ordem** p ou **p-quantis**, indicados por $q(p)$, onde p é uma proporção $(0 < p < 1)$ tal que $100p\%$ das observações são menores do que $q(p)$. Os quantis mais utilizados são:

- $q(0,25) = q_1$: 1º Quartil – abaixo deste valor, encontram-se 25% dos dados;
- $q(0,50) = q_2$: 2º Quartil – equivale à mediana;
- $q(0,75) = q_3$: 3º Quartil – abaixo deste valor, encontram-se 75% dos dados.

Uma outra medida de dispersão é a **distância interquartil**, definida como a diferença entre o terceiro e o primeiro quartis:

$$d_q = q_3 - q_1.$$

Com os dados em ordem crescente, x_1 e x_n são, respectivamente, o menor e o maior valor do conjunto. Os principais valores para se obter uma ideia da distribuição dos dados são, então: x_1, q_1, q_2, q_3 e x_n. A diferença $q_2 - x_1$ é chamada de **dispersão inferior** e $x_n - q_2$ é chamada de **dispersão superior**. Quando as duas dispersões são aproximadamente iguais, tem-se uma distribuição aproximadamente simétrica, como é o caso da **distribuição normal**, discutida mais adiante.

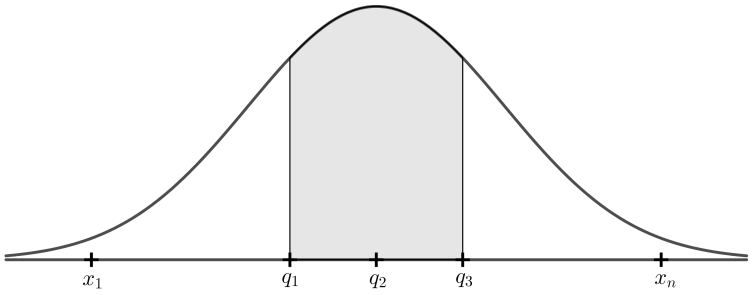

Figura 2.1 – Representação de uma distribuição simétrica.

2.1.2 *Boxplots*

Uma das ferramentas gráficas mais utilizadas para a visualização de resumos de dados é o diagrama chamado de *boxplot*, que dá ideia da posição, dispersão, assimetria e dados discrepantes do conjunto (Bussab & Morettin, 2011). Ele é composto por um retângulo e duas linhas dentro de uma área determinada pelo conjunto de dados. O retângulo identifica três medidas importantes: o primeiro, segundo e terceiro quartis, como pode-se observar na Figura 2.2. Abaixo e acima do retângulo seguem linhas que não devem exceder o limite superior ou o inferior, dados por $LS = q_3 + (1.5)d_q$ e $LI = q_1 - (1.5)d_q$, respectivamente, onde d_q é a medida da dispersão dos dados.

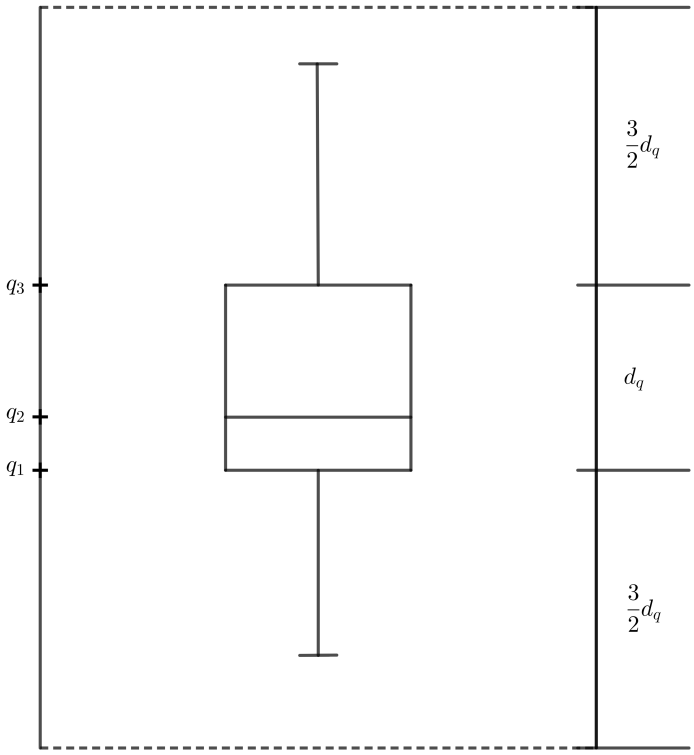

Figura 2.2 – Representação de um *boxplot*.

Observações acima do limite superior ou abaixo do limite inferior são chamados **pontos exteriores** e destoam dos demais dados. Esses pontos podem ser valores atípicos, erros de arredondamento ou erros humanos na construção do conjunto de dados.

2.2 INFERÊNCIA ESTATÍSTICA

A inferência estatística utiliza-se das informações obtidas na fase de análise exploratória para reduzir, analisar e modelar os dados, objetivando fazer previsões sobre as quais decisões podem ser tomadas (Bussab & Morettin, 2011). Neste trabalho, utiliza-se a **análise de variância de 1-fator** (ANOVA), técnica que permite a avaliação ou comparação de médias entre grupos de uma mesma população ou entre populações, sendo **população** o conjunto de todos os elementos ou observações a serem estudados. Essas comparações são feitas por meio de estimativas pontuais, intervalos e testes de hipóteses, uma vez que essas técnicas se aplicam ao modelo considerado.

2.2.1 Introdução à inferência

Uma **amostra** é um subconjunto qualquer de uma população. Quando esse subconjunto $(X_1, \ldots, X_n)$ é escolhido ao acaso de uma população X de tal forma que cada elemento seja uma variável aleatória independente com a mesma distribuição de X, chama-se **amostra aleatória simples**. Uma função dessa amostra é chamada de **estatística**, ou seja, T é uma estatística se $T = f(X_1, \ldots, X_n)$. Uma medida usada para descrever uma característica de uma população é chamada de **parâmetro**. A inferência estatística ocupa-se de fazer afirmações sobre os parâmetros de uma população através de amostras. (Bussab & Morettin, 2011)

2.2.1.1 Modelos para variáveis aleatórias contínuas

A **função densidade de probabilidade** de uma variável aleatória define uma curva cuja área sob a mesma, dado um certo intervalo, representa a probabilidade da variável cair dentro desse mesmo intervalo. Com base nessa função e em **momentos** como o valor esperado (ou esperança, representado por $E[X]$) e a variância ($\mathrm{Var}[X]$), são criados modelos probabilísticos para descrever e representar as variáveis estudadas. Todos esses conceitos podem ser encontrados em Bussab & Morettin (2011). Aqui serão brevemente apresentados os modelos das distribuições citadas em seções posteriores: a distribuição normal, a distribuição χ^2 e a distribuição **F**.

No modelo de **distribuição normal**, também conhecida como distribuição gaussiana, a função densidade da variável aleatória X é dada por:

$$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}. \tag{2.3}$$

Nesse modelo, $E[X] = \mu$ e $\mathrm{Var}[X] = \sigma^2$. Se X possui distribuição normal, denota-se, simbolicamente, $X \sim N(\mu, \sigma^2)$. O gráfico dessa distribuição é simétrico em μ:

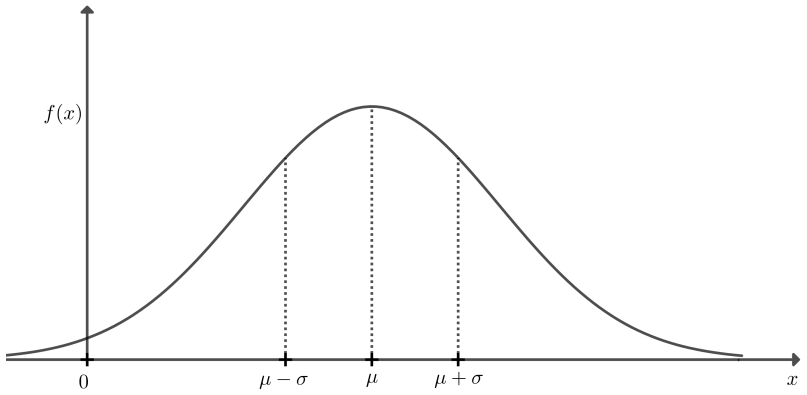

Figura 2.3 – Ilustração de uma curva normal.

O modelo de **distribuição χ^2** com ν graus de liberdade tem função densidade de uma variável aleatória contínua Y dada por:

$$f(y; \nu) = \begin{cases} \dfrac{1}{\Gamma(\nu/2)2^{\nu/2}} y^{\nu/2-1} e^{-y/2}, \text{ se } y > 0, \\ 0, \text{ se } y < 0, \end{cases} \tag{2.4}$$

onde $\Gamma(\alpha)$, $\alpha > 0$, é a **função gama**, e denota-se $Y \sim \chi^2(\upsilon)$. Nesse modelo, $E[Y] = \upsilon$ e $\text{Var}[Y] = 2\upsilon$, e seu gráfico se assemelha a:

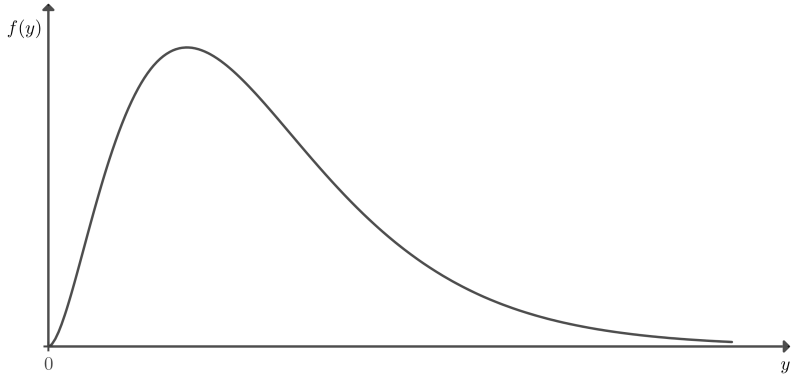

Figura 2.4 – Ilustração de uma curva χ^2, com $\upsilon = 3$.

A **distribuição F de Snedecor** é um caso particular em que uma variável aleatória pode ser definida como o quociente de duas variáveis com distribuição χ^2, de υ_1 e υ_2 graus de liberdade. A variável $W = \frac{U/\upsilon_1}{V/\upsilon_2}$ tem função densidade:

$$g(w;\upsilon_1,\upsilon_2) = \frac{\Gamma((\upsilon_1+\upsilon_2)/2)}{\Gamma(\upsilon_1/2)\Gamma(\upsilon_2/2)}\left(\frac{\upsilon_1}{\upsilon_2}\right)^{\upsilon_1/2}\frac{w^{(\upsilon_1-2)/2}}{(1+\upsilon_1 w/\upsilon_2)^{(\upsilon_1+\upsilon_2)/2}},\ w>0. \tag{2.5}$$

Quando W tem distribuição F com υ_1 e υ_2 graus de liberdade, denota-se $W \sim F(\upsilon_1,\upsilon_2)$ e, nesse caso, $E(W) = \frac{\upsilon_2}{\upsilon_2-2}$ e $\text{Var}[W] = \frac{2\upsilon_2^2(\upsilon_1+\upsilon_2-2)}{\upsilon_1(\upsilon_2-2)^2(\upsilon_2-4)}$. O gráfico dessa distribuição é similar ao da distribuição χ^2, representado na Figura 2.4.

2.2.1.2 Estimadores

Considere uma amostra $(X_1,\ldots,X_n)$ de uma variável aleatória que descreve uma característica de interesse de uma população. Uma estatística T é dita ser estimador de um parâmetro θ se T for utilizado para estimar θ. Nesse caso, $T = g(X_1,\ldots,X_n)$. Estimadores possuem as seguintes propriedades: (Bussab & Morettin, 2011)

(i.) Um estimador é chamado não viciado se $E[T] = \theta$.

(ii.) Um estimador é chamado consistente se $\lim_{n\to\infty} E[T_n] = \theta$ e $\lim_{n\to\infty} \text{Var}[T_n] = 0$, sendo T_n uma estatística baseada numa amostra de tamanho n.

Uma amostra verossímil é aquela que fornece a melhor informação possível sobre um parâmetro de interesse da população. O **princípio da verossimilhança** é que toda a informação disponível de θ fornecida pela amostra X está resumida na função de verossimilhança. Seja $X_1,\ldots,X_n$ uma amostra aleatória simples de uma população X com função densidade de probabilidade $f(x;\theta)$, com vetor paramétrico $\theta \in \Theta$, Θ sendo o espaço paramétrico. A **função de verossimilhança** associa a cada valor de θ a probabilidade da amostra e é dada por:

$$L(\theta;x) = f(x;\theta) = \prod_{i=1}^{n} f(x_i;\theta),\ L:\Theta \longrightarrow \mathbb{R}^+. \tag{2.6}$$

O **estimador de máxima verossimilhança** (EMV) de θ é o valor de $\hat{\theta} \in \Theta$ que maximiza $L(\theta; x)$. Em geral, utiliza-se o logaritmo natural de $L(\theta; x)$ para encontrar o máximo dessa função, denotado por:

$$l(\theta; x) = \log L(\theta; x), \tag{2.7}$$

conhecida por **log-verossimilhança**. Obtém-se o EMV solucionando:

$$\frac{\partial l(\theta; x)}{\partial \theta} = 0, \tag{2.8}$$

e verifica-se ter conseguido um ponto de máximo quando:

$$\frac{\partial^2 l(\theta; x)}{\partial \theta^2} < 0, \text{ para } \hat{\theta} = \theta. \tag{2.9}$$

2.2.1.3 Estimação por intervalos de confiança

Nem sempre é mais interessante obter um valor pontual para um estimador, visto que isso não permite analisar a possível magnitude do erro cometido. É possível criar intervalos de confiança (IC) baseados na distribuição amostral do estimador pontual com determinado coeficiente de confiança γ. (Bussab & Morettin, 2011)

Se T for um estimador de θ, se for conhecida a distribuição amostral de T, sempre será possível achar t_1 e t_2 tais que $P(t_1 < \theta < t_2) = \gamma$, $0 < \gamma < 1$, e o intervalo de confiança para θ com coeficiente de confiança γ é denotado por $IC(\theta; \gamma) =]\, t_1, t_2\, [$.

Para exemplificar, seja $X \sim N(\mu, \sigma^2)$ com variância σ^2 conhecida. Se $X_1, \ldots, X_n$ é uma amostra aleatória simples dessa população, então o intervalo de confiança de nível de confiança $(1 - \alpha)\%$ para a média populacional μ é dado por:

$$\left[\overline{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}; \overline{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right], \tag{2.10}$$

onde $\epsilon = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ é a margem de erro e $\alpha \in (0, 1)$ o nível de significância. O valor de $z_{\alpha/2}$ é tal que $P(Z \geqslant z_{\alpha/2} = \frac{\alpha}{2})$, e $Z = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \sim N(0, 1)$. Esse intervalo de confiança para a média de uma população normal com variância conhecida pode ser usado mesmo quando a população não for normal, desde que $n > 30$.

2.2.1.4 Teste de Hipóteses

Assuma uma variável X associada a uma população e determinada hipótese sobre um parâmetro θ dessa população. A hipótese que desejamos testar chama-se **hipótese nula** (H_0) e é colocada à prova contra sua negativa, a **hipótese alternativa** (H_1). O objetivo de um teste de hipóteses é dizer, usando uma estatística $\hat{\theta}$, se H_0 é aceitável. Caso os testes descartem H_0 como verdadeira, H_1 é considerada aceitável. Qualquer que seja a decisão tomada, erros podem ser cometidos:

- *Erro de tipo I:* rejeitar H_0 quando essa é verdadeira. A probabilidade de cometer esse erro é denotado por α.
- *Erro de tipo II:* não rejeitar H_0 quando essa é falsa. A probabilidade de cometer esse erro é denotado por β.

A decisão em um teste de hipóteses é tomada através da consideração de uma região crítica (RC). Caso o valor observado da estatística pertença a essa região, rejeita-se H_0. Esta região é construída de forma que $P(\hat{\theta} \in \text{RC}|H_0 \text{ é verdadeira})$ seja igual a α, chamado de **nível de significância** do teste. Quanto menor for α, menor a probabilidade de se obter uma amostra com estatística pertencente à região crítica, sendo pouco verossímil a obtenção de uma amostra da população para a qual H_0 seja verdadeira. O passo a passo de um teste de hipóteses é como segue: (Bussab & Morettin, 2011)

- *Passo 1:* Fixe a hipótese a ser testada (H_0) e a hipótese alternativa (H_1).
- *Passo 2:* Determine o melhor estimador para testar a hipótese.
- *Passo 3:* Fixe a probabilidade do erro do tipo I (α), normalmente em 5%, 1% ou 0.1%.
- *Passo 4:* Construa a região crítica sob a hipótese de H_0 ser verdadeira e calcule a estatística do teste com as amostras aleatórias simples.
- *Passo 5:* Se o valor da estatística calculado com os dados da amostra não pertender à RC, não rejeite H_0; caso contrário, rejeite H_0.

Se a hipótese nula for verdadeira, o valor p de um teste de hipótese é a probabilidade de se obter uma estatística amostral com valores tão extremos do que aquela determinada com os dados da amostra. Deve-se rejeitar H_0 ao nível de significância α sempre que $p \leqslant \alpha$. O valor p é a probabilidade de significância do teste. Se T é uma estatística de teste e H_0 é rejeitada para $T > c$, então o p-valor é a probabilidade $P(T > t|H_0)$, onde $t(x)$ é o valor observado de T.

Para exemplificar, seja $X_1, \ldots, X_n$ uma amostra aleatória simples de uma população $N(\mu, \sigma^2)$ com σ^2 conhecido. O teste de hipótese sobre a média μ é:

- *Teste bilateral:* $(H_0 : \mu = \mu_0; H_1 : \mu \neq \mu_0)$

$$\text{RC}(\mu; (1-\alpha)\%) : |z^*| > z_{\alpha/2}; \tag{2.11}$$

- *Teste unilateral à esquerda:* $(H_0 : \mu = \mu_0; H_1 : \mu < \mu_0)$

$$\text{RC}(\mu; (1-\alpha)\%) : z^* < -z_{\alpha/2}; \tag{2.12}$$

- *Teste unilateral à direita:* $(H_0 : \mu = \mu_0; H_1 : \mu > \mu_0)$

$$\text{RC}(\mu; (1-\alpha)\%) : z^* > z_{\alpha/2}, \tag{2.13}$$

onde $z^* = \frac{(\overline{X}-\mu_0)\sqrt{n}}{\sigma}$. Esse teste também pode ser usado quando a população não for normal e para σ^2 desconhecido desde que $n \geqslant 30$.

Mais sobre testes de hipóteses pode ser encontrado em Bussab & Morettin (2011), Montgomery (2013) e Roussas (2003).

2.2.2 Modelo ANOVA

Sejam Y_{ij} variáveis aleatórias de uma população Y com a seguinte estrutura: dados I grupos e J observações dado um certo grupo, o valor da variável y_{ij} encontra-se perto de uma média μ, e a variação de valores ocorre devido a erros e_{ij}. Assume-se que os erros aleatórios e_{ij} são variáveis aleatórias independentes de distribuição normal $N(0,\sigma^2)$ para uma variância desconhecida σ^2. Então,

$$y_{ij} = \mu_i + e_{ij},$$

onde $e_{ij} \sim N(0,\sigma^2)$ são independentes, $i = 1,\ldots,I$, $j = 1,\ldots,J$, e $\mu_i = \mu + \tau_i$, sendo τ_i um parâmetro intrínseco ao i-ésimo grupo.

A esse modelo interessa estimar os parâmetros μ_i e σ^2 e testar a hipótese de não haver diferença entre as médias dos grupos. Para estimar os parâmetros, parte-se da função de verossimilhança do modelo, que é descrita por (Roussas, 2003):

$$L(\boldsymbol{y};\boldsymbol{\mu},\sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{IJ} \exp\left[\frac{-1}{2\sigma^2}\sum_i\sum_j (y_{ij}-\mu_i)^2\right], \tag{2.14}$$

onde $\boldsymbol{y} = (y_1,\ldots,y_j)$ e $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_I)$. Logo,

$$\log L(\boldsymbol{y};\boldsymbol{\mu},\sigma^2) = -\frac{IJ}{2}\log(2\pi) - \frac{IJ}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_i\sum_j (y_{ij}-\mu_i)^2. \tag{2.15}$$

A partir da equação 2.15, vemos que que, para σ^2 fixo, a log-verossimilhança é maximizada em relação a $\mu_1,\ldots,\mu_I$ se $S(\mu_1,\ldots,\mu_I) = \sum_i\sum_j (y_{ij}-\mu_i)^2$ for minimizado também em relação a $\mu_1,\ldots,\mu_I$. Diferenciando:

$$\frac{\partial}{\partial\mu_i} S(\mu_1,\ldots,\mu_I) = -2\sum_j y_{ij} + 2J\mu_i = 0, \tag{2.16}$$

e então,

$$\mu_i = \frac{1}{J}\sum_j y_{ij} \text{ e } \frac{\partial^2}{\partial\mu_i^2} S(\mu_1,\ldots,\mu_I) = 2J. \tag{2.17}$$

Analisando a segunda derivada parcial, a matriz diagonal resultante é definida positiva, visto que, para todo $\lambda_1,\ldots,\lambda_I$, com $\lambda_1^2+\cdots+\lambda_I^2 > 0$,

$$(\lambda_1,\ldots,\lambda_I)\begin{pmatrix} 2J & & 0 \\ & \ddots & \\ 0 & & 2J \end{pmatrix}\begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_I \end{pmatrix} = 2J(\lambda_1^2+\cdots+\lambda_I^2) > 0. \tag{2.18}$$

Segue que os valores encontrados em 2.17 são os EMV's dos μ_i's. Isto é,

$$\hat{\mu}_i = y_{i\cdot}, \text{ onde } y_{i\cdot} = \frac{1}{J}\sum_j y_{ij},\ i = 1,\ldots,I. \tag{2.19}$$

Para estimar σ^2, substitua em 2.15 $S(\mu_1,\ldots,\mu_I)$ por $\hat{S} = \sum_i\sum_j (y_{ij}-y_{i\cdot})^2$. Diferenciando e igualando a zero:

$$\frac{d}{d\sigma^2}\log\hat{L}(\boldsymbol{y};\hat{\boldsymbol{\mu}},\sigma^2) = -\frac{IJ}{2\sigma^2} + \frac{\hat{S}}{2\sigma^4} = 0, \text{ ou } \sigma^2 = \frac{\hat{S}}{IJ}. \tag{2.20}$$

Mais uma vez, pela segunda derivada:

$$\frac{d^2}{d(\sigma^2)^2}\log\hat{L}(\boldsymbol{y};\hat{\boldsymbol{\mu}},\sigma^2) = \frac{IJ}{2(\sigma^2)^2} - \frac{2\hat{S}}{2(\sigma^2)^3}, \tag{2.21}$$

que, quando $\sigma^2 = \frac{\hat{S}}{IJ}$, resulta em $-\frac{(IJ)^3}{2\hat{S}^2} < 0$ e, portanto, o valor de σ^2 em 2.20 é seu EMV. Isto é,

$$\hat{\sigma}^2 = \frac{1}{IJ}SQ_e, \text{ onde } SQ_e = \sum_i\sum_j (y_{ij} - y_{i.})^2, \tag{2.22}$$

onde SQ_e é a soma de quadrados do erro e representa a variabilidade não explicada pelo modelo.

Estimados os parâmetros, resta verificar se as médias dos grupos podem ser consideradas iguais. Em outras palavras, deseja-se testar:

$$\begin{cases} H_0: & \mu_1 = \cdots = \mu_I, \\ H_1: & \text{pelo menos uma das médias } \mu_i \text{ é diferente das demais.} \end{cases} \tag{2.23}$$

O procedimento a seguir é conhecido como **Teste F** e pode ser encontrado em detalhes em Roussas (2003) e em Montgomery (2013). Aplicando a hipótese nula, 2.15 pode ser reescrita como:

$$\log L(\boldsymbol{y};\boldsymbol{\mu},\sigma^2) = -\frac{IJ}{2}\log(2\pi) - \frac{IJ}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_i\sum_j (y_{ij} - \mu)^2, \tag{2.24}$$

e, repetindo o procedimento das derivadas em relação a μ e σ^2, obtém-se novas estimativas $\hat{\mu}$ e $\hat{\sigma^2}_{H_0}$:

$$\hat{\mu} = y_{..}, \text{ onde } y_{..} = \frac{1}{IJ}\sum_i\sum_j y_{ij}, \tag{2.25}$$

$$\hat{\sigma^2}_{H_0} = \frac{1}{IJ}SQ_T, \text{ onde } SQ_T = \sum_i\sum_j (y_{ij} - y_{..})^2, \tag{2.26}$$

onde SQ_T é a soma de quadrados total e representa a variabilidade total dos dados.

O teste baseia-se em uma estatística baseada na razão de verossimilhança $\lambda = \lambda(y)$ para comparar a máxima verossimilhança dos dados sob a hipótese nula com a máxima verossimilhança sem restrições (sob a hipótese alternativa). Note que, sob H_0:

$$\exp\left[-\frac{1}{\hat{\sigma^2}_{H_0}}\sum_i\sum_j (y_{ij} - y_{..})^2\right] = \exp\left(-\frac{IJ}{2SQ_T}\times SQ_T\right) = \exp\left(-\frac{IJ}{2}\right), \tag{2.27}$$

e, sob a hipótese alternativa,

$$\exp\left[-\frac{1}{\hat{\sigma^2}}\sum_i (y_{ij} - y_{..})^2\right] = \exp\left(-\frac{IJ}{2SQ_e}\times SQ_e\right) = \exp\left(-\frac{IJ}{2}\right), \tag{2.28}$$

que conclui a razão de verossimilhança $\lambda = (\hat{\sigma^2}/\hat{\sigma^2}_{H_0})^{IJ/2}$. De acordo com esse resultado, $\lambda < C$ se, e somente se,

$$\left(\frac{\hat{\sigma^2}}{\hat{\sigma^2}_{H_0}}\right)^{IJ/2} < C, \text{ ou } \frac{\hat{\sigma^2}}{\hat{\sigma^2}_{H_0}} < C^{2/IJ}, \text{ ou } \frac{\hat{\sigma^2}_{H_0}}{\hat{\sigma^2}} > \frac{1}{C^{2/IJ}} = C_0. \tag{2.29}$$

Teorema 2.2.1. *$SQ_T = SQ_E + SQ_H$, onde SQ_E e SQ_T são dados por 2.22 e 2.26, respectivamente,*

$$e\ SQ_H = \sum_i \sum_j (y_{i.} - y_{..})^2 = J \sum_i (y_{i.} - y_{..})^2, \tag{2.30}$$

SQ_H sendo a soma de quadrados da hipótese (pois é a estatística utilizada para testar a hipótese nula) e reflete as variações devidas às diferenças entre grupos.

Prova. Roussas (2003, p. 400).

Por esse resultado, a expressão em 2.29 fica:

$$\frac{SQ_T}{SQ_e} > C_0, \text{ ou } \frac{SQ_e + SQ_H}{SQ_e} > C_0, \text{ ou } \frac{SQ_H}{SQ_e} > C_1 = C_0 - 1. \tag{2.31}$$

Ou seja, o teste da razão de verossimilhança rejeita H_0 sempre que $SQ_H/SQ_e > C_1$. Para determinar C_1, é necessário conhecer a distribuição da estatística $\dfrac{SQ_H}{SQ_e}$ sob H_0, com os valores observados substituídos pelas respectivas variáveis aleatórias.

Teorema 2.2.2. *Considere o modelo descrito no 2.2.1 e as expressões de SQ_e, SQ_T e SQ_H, e substitua os valores observados y_{ij}, $y_{i.}$ e $y_{..}$ pelos Y_{ij}, $Y_{i.}$ e $Y_{..}$ das variáveis aleatórias, respectivamente. Então:*

(i.) SQ_e/σ^2 *tem distribuição* $\chi^2_{I(J-1)}$.

(ii.) As estatísticas SQ_e e SQ_H são independentes.

Se a hipótese nula for verdadeira, então:

(iii.) SQ_H/σ^2 *tem distribuição* χ^2_{I-1}.

(iv.) $\frac{SQ_H/(I-J)}{SQ_e/I(J-1)} \equiv F_{I-1,I(J-1)}$.

(v.) SQ_T/σ^2 *tem distribuição* χ^2_{IJ-1}.

Prova. Roussas (2003, p. 401).

É resultado do Teorema 2.2.2 que:

(i.) As EMV's $\hat{\mu}_i = \overline{Y}_i$ são estimadores não viciados de μ_i, para $i = 1, \ldots, I$.

(ii.) A EMV $\hat{\omega}^2 = \frac{SQ_e}{IJ}$ é viciada, mas a estimativa $QM_e = \frac{SQ_e}{I(J-1)}$ é não viciada.

Prova. Roussas (2003, p. 401).

Conclui-se que, para testar 2.23 a nível de significância α, rejeita-se H_0 sempre que:

$$\mathbf{F} = \frac{SQ_H/(I-J)}{SQ_e/I(J-1)} = \frac{QM_H}{QM_e} > F_{I-1,I(J-1);\alpha}, \tag{2.32}$$

onde o valor crítico $F_{m,n;\alpha}$ é determinado de forma que $P(X > F_{m,n;\alpha}) = \alpha$, onde X é uma variável aleatória com distribuição $F_{m,n}$.

Todos os passos realizados para o cálculo do teste são, normalmente, organizados em uma tabela, chamada de tabela da análise de variância de um fator (ou tabela ANOVA), como a Tabela 2.1.

VARIAÇÃO	SOMA DOS QUADRADOS	GRAUS DE LIBERDADE	QUADRADO MÉDIO	F_0
Entre grupos	$SQ_H = J\sum_{i=1}^{I}(Y_{i.} - Y_{..})^2$	$I-1$	$QM_H = \frac{SQ_H}{I-1}$	$\frac{QM_H}{QM_e}$
Erros	$SQ_e = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - Y_{i.})^2$	$I(J-1)$	$QM_e = \frac{SQ_e}{I(J-1)}$	
Total	$SQ_T = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - Y_{..})^2$	$IJ-1$		

Tabela 2.1 – Tabela da análise de variância de um fator.

O teste **F** da ANOVA testa a hipótese da igualdade entre as médias dos grupos contra haver pelo menos uma diferente das demais. O detalhamento dessas informações requer um teste adicional, conhecido como **teste de Tukey**. Além disso, para que a ANOVA corresponda a uma avaliação adequada da diferença entre as médias de tratamento, é necessário que as observações sejam adequadamente descritas pelo modelo, que os erros tenham distribuição normal e sejam independentes, com média zero e variância constante (Montgomery, 2013). Para a verificação dessas suposições, podem ser realizados dois procedimentos auxiliares: o **teste de Levene** e o **teste de Breusch-Pagan**. Caso o modelo não esteja adequado, pode-se realizar uma transformação prévia dos dados. Aqui será apresentada uma forma de transformação, conhecida como **transformação de Box-Cox**.

2.2.2.1 Teste de Levene

Para testar a hipótese de igualdade de variâncias entre os tratamentos, o teste de Levene utiliza os desvios absolutos das observações y_{ij} em cada grupo em relação à mediana do respectivo grupo. A estatística de teste empregada é a mesma estatística **F** da ANOVA, porém aplicada aos desvios absolutos em vez das observações originais. Esse teste é menos sensível à suposição de normalidade, tornando-se uma escolha adequada para dados reais, que raramente seguem uma distribuição normal. Mais sobre o teste de Levene pode ser encontrado em Montgomery (2013).

2.2.2.2 Teste de Breusch-Pagan

O teste de Breusch-Pagan é utilizado para detectar heterocedasticidade do modelo, ou seja, a inconstância da variabilidade dos erros do modelo. Ele realiza um teste de hipótese onde H_0 supõe a homocedasticidade (variância constante dos erros) contra a hipótese alternativa de que a variância dos erros depende das variáveis explicativas. A estatística de teste segue uma distribuição χ^2 e, se o p-valor for pequeno, rejeita-se a homocedasticidade, indicando a presença de heterocedasticidade no modelo. (Breusch & Pagan, 1979)

2.2.2.3 Transformação de Box-Cox

As transformações de dados são geralmente utilizadas para estabilizar variâncias, aproximar a distribuição das variáveis à distribuição normal e melhor ajustar o modelo aos dados (Montgomery, 2013). A transformação de Box-Cox é tal que

$$w_t^{(\lambda)} = \begin{cases} \dfrac{y_t^{\lambda} - 1}{\lambda}, \text{ se } \lambda \neq 0, \\ \log y_t, \text{ se } \lambda = 0, \end{cases} \tag{2.33}$$

onde o parâmetro λ é estimado simultaneamente aos demais parâmetros do modelo, pelo estimador de máxima verossimilhança. O EMV de λ é o mínimo dos quadrados dos erros $SQ_e(\lambda)$, que normalmente é encontrado geometricamente.

Escolhido um valor de λ, um intervalo de confiança deve ser determinado para os pontos em que $SQ_e(\lambda)$ cortam um valor de referência baseado nos graus de liberdade da SQ_e. Se esse intervalo incluir o valor 1, isso sugere que a transformação não é necessária. Caso contrário, utiliza-se $w_t^{(\lambda)}$ para transformar os dados.

2.2.2.4 Teste de Tukey

O teste de Tukey é um teste de hipótese desenhado para comparar os grupos de dados dois a dois. Suponha que, seguido de uma análise de variância que rejeitou a hipótese nula, deseja-se testar todos os pares de média na seguinte comparação:

$$\begin{cases} H_0 : \mu_i = \mu_J \\ H_1 : \mu_i \neq \mu_j, \end{cases} \tag{2.34}$$

para todo $i \neq j$.

No procedimento de Tukey, o nível de significância geral é exatamente α quando o tamanho das amostras são iguais e no máximo α quando os tamanhos são desiguais. Também é possível construir intervalos de confiança para as diferenças dos pares de médias, onde se considera as médias iguais caso o intervalo contenha o zero e diferentes caso não contenha. Para esses intervalos, o nível de confiança simultâneo é $100(1-\alpha)\%$ quando os tamanhos das amostras são iguais e pelo menos $100(1-\alpha)\%$ quando os tamanhos são diferentes. A versão do procesimento de Tukey para amostras de tamanhos diferentes é também conhecida como teste de Tukey-Kramer. Mais sobre o teste de Tukey pode ser encontrado em Montgomery (2013).

3 ESTUDO DE CASO

Como visto, a frequência estudantil está intimamente relacionada a uma série de fatores contextuais e sociais que influenciam diretamente a presença ou ausência dos alunos na escola. Elementos como a condição socioeconômica, questões familiares e até fatores culturais podem impactar os padrões de frequência, tornando a análise dos dados uma tarefa singular.

A abordagem de um estudo de caso possibilita uma análise mais detalhada e contextualizada dos fatores locais que influenciam os padrões de frequência. Supõe-se que, dentro de uma mesma escola, os alunos compartilham condições socioeconômicas parecidas, e estão sujeitos às mesmas condições sazonais, principalmente por residirem, de maneira geral, em uma mesma área próxima à escola. Assim, é possível direcionar a análise para os fatores internos à instituição, como a organização do quadro de horários.

As técnicas estatísticas apresentadas no Capítulo 2 são aqui aplicadas na avaliação da frequência escolar do Colégio Estadual Engenheiro Arêa Leão, que colaborou fornecendo dados de frequência das turmas, sem identificação dos alunos, referentes ao ano de 2022. O objetivo desse estudo foi descrever os efeitos do quadro de horários sobre a frequência escolar, investigando a possível existência de disciplinas que afetassem a assiduidade de forma positiva ou negativa.

3.1 CARACTERIZAÇÃO DA ESCOLA E MUNICÍPIO

O Colégio Estadual Engenheiro Arêa Leão (CEEAL), localizado no município de Nova Iguaçu, Rio de Janeiro, trabalha as modalidades Ensino Regular e Educação para Jovens e Adultos para cerca de 1300 alunos por ano, distribuídos na região metropolitana da Baixada Fluminense – o que o caracteriza como de médio a grande porte. A maioria de seus estudantes concentra-se na cidade de Nova Iguaçu, porém, o colégio recebe estudantes de todo o estado.

De acordo com o Índice de Desenvolvimento da Educação Básica (Ideb, BRASIL, 2021), que avalia fluxo escolar e médias de desempenho nas avaliações da rede pública de ensino em índice que varia de 0 a 10, a cidade de Nova Iguaçu foi avaliada com a nota 4.7 para os anos iniciais do Ensino Fundamental e 4.4 para os anos finais. A meta do Ideb estabelecida para o ano de 2022 (ano dos dados levantados no estudo) era de alcançar a média 6, considerado "um valor correspondente a um sistema educacional de qualidade comparável ao dos países desenvolvidos" pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep, BRASIL, 2021). De acordo com o Indicador de Nível Socioeconômico (Inse, BRASIL, 2021), o colégio enquadra-se no nível socioeconômico 4 – o que significa que os estudantes estão até meio desvio padrão abaixo da média nacional verificada pelo Inse. A descrição

do perfil socioeconômico dos estudantes nesse nível segue:

> *"A maioria dos estudantes respondeu ter em sua casa uma geladeira, dois ou mais celulares com internet, um carro, mesa para estudar, wi-fi, TV por internet, garagem, forno de micro-ondas, máquina de lavar roupa e freezer. Algumas respostas não obtiveram maioria, mas indicam que parte dos estudantes afirmou possuir um ou dois banheiros, uma ou duas televisões, dois ou mais quartos para dormir, aspirador de pó, um computador e escolaridade da mãe (ou responsável) variando entre ensino médio e ensino superior completo e do pai (ou responsável) entre ensino fundamental completo e ensino médio completo."* (BRASIL, 2021)

No ano de 2022, o CEEAL contou com 42 turmas, sendo 17 do Ensino Fundamental (EF) II, 17 do Ensino Médio (EM) e 8 do Núcleo de Educação de Jovens e Adultos (NEJA). O detalhamento das turmas por nível, ano e turno encontra-se abaixo, na Tabela 3.1.

NÍVEL	**ANO/SEMESTRE**	**QUANTIDADE DE TURMAS**	**TURNO**
ENSINO FUNDAMENTAL	6º ano	4	Tarde
	7º ano	4	Tarde
	8º ano	1	Manhã
		3	Tarde
	9º ano	3	Manhã
		2	Tarde
ENSINO MÉDIO	1º ano	5	Manhã
		1	Tarde
	2º ano	5	Manhã
		1	Tarde
	3º ano	4	Manhã
		1	Tarde
NEJA	I	2	Noite
	II	2	Noite
	III	2	Noite
	IV	2	Noite

Tabela 3.1 – Distribuição das turmas do CEEAL em 2022.

É importante destacar dois aspectos contextuais relevantes em relação ao ano de 2022. Primeiro, esse foi o primeiro ano de retomada das aulas presenciais após a pandemia de COVID-19, iniciada em 2020, a qual resultou em déficits de aprendizado constatados pelo Conselho Nacional de Educação (CNE, BRASIL, 2022). Segundo, 2022 marcou a implementação do Novo Ensino Médio, estabelecido pela Lei nº 13.415/2017 (BRASIL, 2017), que alterou a estrutura dessa etapa de ensino, ampliando a carga horária mínima dos estudantes e redefinindo a organização curricular.

No CEEAL, essa transição curricular foi iniciada em 2022, com a aplicação das mudanças apenas ao primeiro ano do Ensino Médio. As principais alterações incluíram a introdução da disciplina Projeto de Vida e a oferta de disciplinas eletivas, que não foram especificadas no presente estudo.

3.2 MATERIAIS E MÉTODOS

3.2.1 Coleta e tratamento dos dados

O CEEAL disponibilizou os diários de classe referentes ao ano de 2022, abrangendo suas 42 turmas. Esses registros continham informações sobre a frequência diária dos estudantes por turma, os quadros de horários e a distribuição dos professores por disciplina.

Para assegurar o anonimato dos alunos, os dados de frequência foram reorganizados de forma a representar exclusivamente a frequência total de cada turma por dia, sem qualquer identificação individual. A análise da frequência teve como objetivo identificar padrões e tendências relacionados à faixa etária dos estudantes, ao nível de ensino e ao turno em que estavam matriculados, de modo que o comportamento individual não foi considerado. As entradas de frequência utilizadas contavam somente com dias letivos em que houve, de fato, aula, desconsiderando feriados, passeios e outras situações particulares.

Além disso, foi possível contar com a colaboração de professores da escola, que contribuiram com informações relevantes sobre o quadro de horários, principalmente em relação à distribuição e organização de professores. Também auxiliaram na leitura dos dados de frequência e compartilharam percepções acerca dos fatores que acreditavam, por suas próprias experiências, influenciar a frequência estudantil.

3.2.2 Aplicação do modelo

O estudo de caso foi orientado pela questão de pesquisa sobre a influência do quadro de horários na decisão dos estudantes de comparecer à escola em determinados dias da semana. Caso essa influência fosse confirmada, interessava saber como se dava essa relação. Para consideração no estudo, a relação frequência e dia da semana corresponde à relação frequência e quadro disciplinar diário, visto que esse é fixo para um determinado dia e invariável ao longo do ano.

Sendo assim, as informações de frequência foram agrupadas em $I = 5$ grupos correspondentes a Segunda, Terça, Quarta, Quinta e Sexta. Interessava estimar as médias de frequência μ_i e testar a hipótese de não haver diferença significativa entre os dias da semana, ou seja, realizar o teste proposto em 2.23. O passo a passo do estudo encontra-se esquematizado abaixo:

Em função da natureza dos dados, foram realizados os testes de Levene e Breusch-Pagan para avaliar a normalidade e homogeneidade das variâncias nas observações de cada turma. Quando necessário, aplicou-se a transformação de Box-Cox antes de proceder com a análise de variância. Para aprofundar a comparação entre as médias, foi utilizado o teste de Tukey.

3.3 RESULTADOS

3.3.1 Análise exploratória

As primeiras observações dos dados foram realizadas por meio dos gráficos de *boxplot* de cada turma, os quais destacaram a variação nas distribuições e chamaram a atenção para quedas e aumentos significativos nas médias de frequência de dias específicos para as turmas. As médias analisadas revelaram um padrão polarizado: das 45 turmas, 25 apresentaram a terça-feira como o dia com a maior média de frequência, enquanto 35 indicaram a sexta-feira como o dia com a menor média. Esses resultados chamam a atenção para as terças e sextas-feiras,

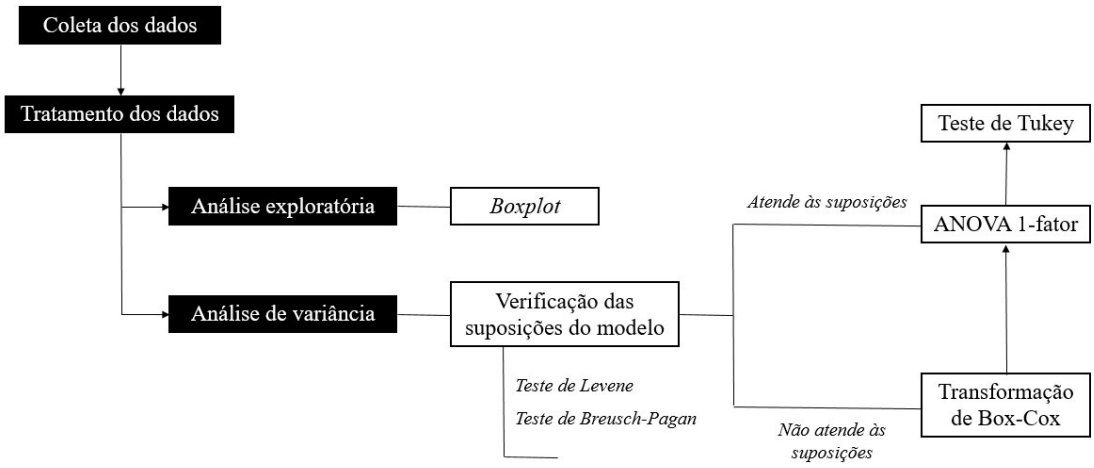

Figura 3.1 – Representação esquemática dos processos do estudo.

sugerindo que esses dias possuem características específicas que influenciam essa disparidade.

As Tabelas 3.2, 3.3 e 3.4 contém as informações de médias de frequência por dia da semana do Ensino Fundamental, Médio e NEJA, respectivamente.

MÉDIAS DE FREQUÊNCIA DO ENSINO FUNDAMENTAL					
TURMA	**SEG**	**TER**	**QUA**	**QUI**	**SEX**
601	26,0892	28,6533	29,9121	29,5873	26,7820
602	27,8488	30,1887	28,9280	28,2670	26,6942
603	26,6296	28,0659	28,2492	27,7913	25,7909
604	17,5639	17,9070	16,9537	17,6413	16,3881
701	32,6009	33,1342	34,2412	33,2571	29,7186
702	33,7341	34,2217	34,5625	33,9475	32,2349
703	29,1584	32,0846	32,4023	32,1169	28,9138
704	27,2566	26,8208	28,7406	28,0777	25,0964
801	30,4015	31,0141	30,1085	31,1732	29,3105
802	32,1217	34,1787	33,7467	35,0052	31,1595
803	33,0909	32,7531	33,0270	34,4107	29,8764
804	34,3464	35,3109	34,9122	33,8712	32,9927
901	33,3610	34,5435	33,5881	34,1743	31,4781
902	27,3165	29,5435	28,5345	28,3463	27,0977
903	27,4694	31,3999	29,7023	28,7586	29,5183
904	31,9327	32,9120	31,8884	32,8381	30,2198
905	28,7130	30,0892	28,7725	29,1191	26,9161

Tabela 3.2 – Médias de frequência por dia da semana das turmas do Ensino Fundamental.

MÉDIAS DE FREQUÊNCIA DO ENSINO MÉDIO					
TURMA	**SEG**	**TER**	**QUA**	**QUI**	**SEX**
1001	35,7823	37,6876	36,9615	36,3270	32,1284
1002	31,3625	31,5126	32,2125	32,0981	25,5029
1003	29,6178	33,1851	32,2074	30,9079	29,3658
1004	30,1462	31,2101	31,2053	31,8374	30,7869
1005	28,0019	29,7811	29,4585	28,4972	26,9400
1006	26,1453	26,9618	25,8968	26,1548	20,1568
2001	27,3843	27,9249	28,6073	27,6867	27,6780
2002	23,8065	24,3333	26,7097	25,6786	22,5000
2003	28,1791	29,8612	29,6145	29,8066	26,2177
2004	27,5552	29,1262	27,0387	26,7228	25,7879
2005	23,2379	23,2477	21,0779	21,7981	21,4822
2006	18,5255	20,0288	19,5256	19,3316	15,7888
3001	28,0758	29,5247	28,8531	28,3496	26,7459
3002	28,9615	30,4225	29,8138	27,1963	27,5424
3003	29,0367	29,7032	29,0222	29,8159	27,2429
3004	24,9908	26,9409	23,7424	25,6123	23,4388
3005	13,5625	14,0312	13,0000	10,9565	10,8621

Tabela 3.3 – Médias de frequência por dia da semana das turmas do Ensino Médio.

MÉDIAS DE FREQUÊNCIA DO NEJA					
TURMA	**SEG**	**TER**	**QUA**	**QUI**	**SEX**
NEJA I 1	18,3607	20,7045	19,2817	17,2359	17,2270
NEJA I 2	7,8111	8,5198	7,3292	6,0263	5,2008
NEJA II 1	23,2473	21,9279	21,7034	21,7001	19,8100
NEJA II 2	17,7160	19,8099	19,4505	18,9685	15,0231
NEJA III 1	19,8090	21,0171	15,4551	16,4228	15,7135
NEJA III 2	12,1344	11,7228	13,0309	10,5639	9,5309
NEJA IV 1	16,1945	17,7998	18,9309	20,7645	13,0854
NEJA IV 2	9,0093	9,5856	9,5148	8,3961	6,7749

Tabela 3.4 – Médias de frequência por dia da semana das turmas do NEJA.

Em diálogo com professores e gestores, observou-se que as sextas-feiras, além de naturalmente atrativas para as faltas devido à proximidade do final de semana, também têm sido preferenciais para a alocação de disciplinas que fogem do normal. Essas disciplinas incluem aquelas que, no início do ano letivo, ainda não tinham professores alocados, resultando em períodos de tempo ociosos, ou as disciplinas eletivas inseridas no quadro do Novo Ensino Médio, que, por não serem avaliadas por notas e serem relativamente novas para os alunos, acabam não despertando o mesmo comprometimento.

Essas observações podem ajudar a explicar o comportamento da frequência observado nos *boxplots* das Figuras 8.1, 8.2 e 8.3 do Apêndice 8. Em particular, as frequências das turmas do primeiro ano do EM, já adequada ao Novo Ensino Médio, que apresentam uma queda visível às sextas-feiras.

A Tabela 3.5 sintetiza as informações sobre os dias da semana em que ocorrem, com maior frequência, as maiores e menores médias de assiduidade em cada ano ou semestre. Além disso, apresenta o primeiro levantamento da relação entre frequência e disciplinas, destacando quais são as mais recorrentes nos dias de maior e menor assiduidade em cada ano.

		MAIOR FREQUÊNCIA		**MENOR FREQUÊNCIA**	
NÍVEL	**ANO/SEMESTRE**	**DIA(S)**	**DISCIPLINA(S)**	**DIA(S)**	**DISCIPLINA(S)**
EF	6º ano	Terça e quarta	Língua Portuguesa	Sexta	*
	7º ano	Quarta	Matemática	Sexta	Ciências
	8º ano	Quinta	*	Sexta	Matemática
	9º ano	Terça	Matemática	Sexta	Arte
EM	1º ano	Terça	Matemática	Sexta	Eletivas
	2º ano	Terça	Matemática	Sexta	*
	3º ano	Terça	Matemática e Língua Portuguesa	Sexta	Tempos vagos
NEJA	I	Terça	Matemática	Sexta	Língua Inglesa
	II	Segunda e terça	Física	Sexta	Biologia
	III	Terça e quarta	Matemática e Língua Portuguesa	Quarta e sexta	*
	IV	Terça e quinta	Química	Sexta	Arte

Tabela 3.5 – Relação de dias mais comuns à maior e menor frequência observada e disciplinas mais comuns nesses dias por nível e ano. (*) Empate entre três ou mais disciplinas.

Mesmo ao agrupar todas as turmas por ano, observa-se que nenhuma delas apresenta mais de dois dias distintos como aqueles de maiores ou menores médias de frequência. Nos casos em que há dois dias destacados, isso indica uma igualdade na quantidade de ocorrências desses dias, tanto como os de maior quanto os de menor média.

Esse padrão revela a predominância de certos dias da semana em relação aos demais, indicando uma regularidade nos dados da frequência, refletindo aspectos comportamentais ou estruturais. Por exemplo, no Ensino Médio, as terças-feiras se destacam como o dia de maior

média de frequência de forma unânime. Por outro lado, no Ensino Fundamental e no Ensino Médio, as sextas-feiras são, em maioria para todos os anos, os dias com as menores taxas de presença.

Começa-se a delinear também uma tendência em relação às disciplinas, especialmente nos dias de maior frequência, onde há um predomínio de Língua Portuguesa e Matemática. Por outro lado, nos dias de menor assiduidade, o cenário se mostra mais diversificado, incluindo disciplinas como Ciências, Biologia, Artes, disciplinas eletivas e, no caso do terceiro ano do Ensino Médio, a ocorrência de horários vagos. A presença de disciplinas eletivas e tempos vagos associados às sextas-feiras com as menores médias de frequência para o EM corrobora o testemunho dos professores.

Considerando os dias com as maiores médias de frequência, 26 das 42 ocorrências estavam associadas à disciplina de Matemática, que se destacou como a mais frequente. A segunda disciplina mais recorrente nesses dias foi Educação Física. Já em relação aos dias com as menores médias de frequência, a disciplina mais comum registrada foi Arte, que apareceu 12 vezes, seguida pelas disciplinas eletivas e pelos períodos de tempo vago, que, juntos, somaram 10 ocorrências.

As relações entre a frequência estudantil e os dias da semana parecem estar bem definidas, enquanto a influência das disciplinas exige uma análise mais detalhada. Para compreender melhor essa relação, é necessário verificar se as variações observadas nas médias de frequência são estatisticamente significativas ou meramente aleatórias. Para isso, aplicou-se a análise de variância, que, como visto, é um método que permite testar se as diferenças entre as médias dos grupos comparados são resultado de fatores sistemáticos ou apenas de variações ocasionais.

3.3.2 ANOVA

Na análise de variância de um fator, utiliza-se o valor **F** e o valor p para interpretar os resultados. Como visto no Capítulo 2, o valor **F** representa a razão entre a variabilidade das médias dos grupos e a variabilidade dentro dos próprios grupos. Quando esse valor é elevado, isso sugere que as diferenças entre as médias dos grupos são consideráveis em comparação à variação interna de cada grupo, o que pode indicar um efeito significativo do fator estudado.

Para determinar se um valor **F** é suficientemente alto para ser considerado significativo, é necessário compará-lo ao valor crítico da distribuição **F** para o nível de significância determinado – nesse caso, 5% – e os graus de liberdade do teste. Aqui, para determinar se essa diferença é estatisticamente significativa, verificou-se o valor p, que representa a probabilidade de obtermos um valor **F** tão extremo quanto o observado caso a hipótese nula (de que não há diferença entre os grupos) seja verdadeira. Se o valor p for menor que o nível de significância estabelecido, rejeita-se a hipótese nula e conclui-se que há uma diferença significativa entre as médias dos grupos analisados (Bussab & Morettin, 2011). Caso contrário, não há evidências suficientes para afirmar que o fator analisado tem impacto significativo sobre os resultados.

As Tabelas 3.6, 3.7 e 3.8 mostram os resultados do teste **F** da ANOVA para cada turma do NEJA, Ensino Fundamental e Ensino Médio, respectivamente. Incluem também os valores p e a conclusão do teste em relação à hipótese nula. A 5% de significância, quando o valor p for menor que 0,05, há a indicação da improbabilidade das diferenças observadas terem ocorrido por acaso. Nesse caso, rejeita-se a hipótese nula que assume igualdade entre as médias dos dias da semana. Assim, ao rejeitar H_0 conclui-se que pelo menos uma das médias semanais difere significativamente das demais, evidenciando uma possível influência do dia da semana sobre a frequência estudantil. O código da implementação do modelo ANOVA, assim como a tabela de dados, pode ser encontrado no Apêndice 8, seção 8.2.

NEJA		
TURMA	**VALOR F**	**VALOR *p***
NEJA I 1	1,329745	0,270000
NEJA I 2	2,648996	0,044799
NEJA II 1	0,819114	0,518256
NEJA II 2	4,491597	0,002887
NEJA III 1	3,244602	0,017802
NEJA III 2	3,305988	0,015800
NEJA IV 1	1,726040	0,157982
NEJA IV 2	2,606151	0,045288

Tabela 3.6 – Resultado da ANOVA para as turmas do NEJA.

ENSINO FUNDAMENTAL		
TURMA	**VALOR F**	**VALOR *p***
601	5,498805	0,000377
602	3,907250	0,004792
603	2,351772	0,056657
604	2,182532	0,073861
701	5,370141	0,000451
702	1,824420	0,127070
703	8,103860	0,000006
704	3,783443	0,005887
801	0,667776	0,615531
802	12,210143	0,000000
803	6,394412	0,000092
804	3,710453	0,006583
901	2,730778	0,031350
902	1,725995	0,147728
903	2,994786	0,020879
904	2,156838	0,076876
905	1,969425	0,102503

Tabela 3.7 – Resultado da ANOVA para as turmas do Ensino Fundamental.

ENSINO MÉDIO		
TURMA	**VALOR F**	**VALOR *p***
1001	6,606183	0,000067
1002	7,745952	0,000012
1003	3,270253	0,013441
1004	0,745177	0,562691
1005	2,076862	0,086851
1006	8,271617	0,000005
2001	0,212735	0,931017
2002	2,747974	0,030485
2003	2,527752	0,043260
2004	1,684126	0,156752
2005	1,432677	0,226074
2006	6,763744	0,000054
3001	1,157652	0,332040
3002	1,788624	0,134461
3003	1,132297	0,343680
3004	2,431787	0,050582
3005	8,051499	0,000007

Tabela 3.8 – Resultado da ANOVA para as turmas do Ensino Médio.

Das 42 turmas analisadas, 23 apresentaram diferenças estatisticamente significativas entre as médias de frequência ao longo da semana. Entre essas, a maior frequência foi mais comumente observada às terças-feiras (13 turmas), seguida pelas quartas-feiras (6 turmas) e quintas-feiras (3 turmas). Já em relação às menores médias de frequência, a sexta-feira se destacou como o dia mais associado à baixa assiduidade, sendo apontada por 20 turmas, enquanto a segunda-feira e a quarta-feira foram apontadas por duas e uma turma, respectivamente.

As diferenças entre as médias foram detalhadas pelo teste de Tukey. O resultado das comparações do teste pode ser encontrado no Apêndice 8, seção 8.5. Assim como na ANOVA, um valor p menor que 0,05 sugere que a diferença entre os grupos comparados é estatisticamente significativa, e fala a favor de que as médias desses grupos são diferentes entre si. Já valores p maiores que 0,05 indicam que não há evidências suficientes para rejeitar a hipótese nula. Assim, é possível determinar exatamente quais dias da semana possuem médias diferentes entre si.

A maioria dos resultados indicou que as diferenças significativas nas médias de frequência ocorreram, principalmente, nas comparações envolvendo a sexta-feira e, em menor grau, a terça-feira. No total, 49 comparações entre a sexta-feira e outro dia da semana apresentaram diferenças estatisticamente significativas, sugerindo que esse dia se destaca em relação aos demais. Já as comparações envolvendo a terça-feira mostraram diferenças significativas em 20 casos. Apenas seis comparações que não envolviam esses dias indicaram diferenças.

TURMA		**DISCIPLINAS**	
602	Língua Portuguesa	História	
804	Língua Portuguesa	Geografia	Ciências
901	História	Matemática	Educação Física
903	História	Matemática	Educação Física
1001	Língua Portuguesa	Matemática	Filosofia
1003	História	Matemática	Biologia
1006	Língua Portuguesa	Matemática	Educação Física
2003	Matemática	Língua Estrangeira	Física
2006	Matemática	Educação Física	Física
3005	Língua Portuguesa	Educação Física	Biologia
NEJA I 2	Língua Portuguesa	Matemática	
NEJA II 2	Física	Projeto de Vida	
NEJA IV 2	Língua Portuguesa		

Tabela 3.9 – Disciplinas alocadas às terças-feiras para as turmas descritas selecionadas na ANOVA.

Ao analisar a distribuição das disciplinas nas terças e sextas-feiras para as turmas em que esses dias se diferenciaram significativamente dos demais, observou-se um padrão semelhante ao identificado na análise exploratória de dados. Das treze turmas que destacaram a terça-feira como o dia de maior frequência, oito tinham aulas de Matemática nesse dia, sete contavam com aulas de Língua Portuguesa e cinco com Educação Física. Ao considerar os diferentes níveis de ensino, percebe-se que o interesse por Educação Física é mais evidente no Ensino Médio, enquanto a presença de Matemática se mantém forte em todos os níveis, reforçando sua possível influência sobre a assiduidade estudantil. A Tabela 3.9 mostra a ocorrência completa

TURMA	DISCIPLINAS		
602	Língua Inglesa	Arte	Ciências
701	Língua Portuguesa	Arte	Ciências
703	Língua Portuguesa	Língua Inglesa	Ciências
704	Geografia	Arte	Ciências
802	Língua Portuguesa	Língua Inglesa	Ciências
803	Língua Portuguesa	Matemática	
804	Língua Inglesa	Matemática	
901	RPM	Arte	Ciências
1001	Projeto de Vida	Eletivas	
1002	Projeto de Vida	Eletivas	
1003	Língua Inglesa	Eletivas	
1006	Projeto de Vida	Eletivas	
2002	Educação Física	Filosofia	Língua Portuguesa
2003	Língua Portuguesa	Matemática	
2006	Matemática	Sociologia	Química
3005	Matemática	Sociologia	Química
NEJA I 2	Língua Inglesa	Geografia	
NEJA II 2	Química	Biologia	Arte
NEJA III 2	Sociologia	Filosofia	
NEJA IV 2	Língua Portuguesa	Arte	Biologia

Tabela 3.10 – Disciplinas alocadas às sextas-feiras para as turmas descritas selecionadas na ANOVA.

de disciplinas para as turmas selecionadas.

Em relação às sextas-feiras, entre as vinte turmas que apresentaram diferenças significativas, sete tinham aulas de Língua Portuguesa, enquanto Língua Inglesa, Arte e Ciências apareceram em seis turmas cada. Observa-se que Língua Portuguesa e Ciências são mais frequentes no Ensino Fundamental, sendo esta última exclusiva desse nível. Já no Ensino Médio, destaca-se a presença de Projeto de Vida e disciplinas eletivas, especialmente para as turmas do primeiro ano. De maneira geral, a distribuição das disciplinas nas sextas-feiras é mais diversificada, sugerindo que outros fatores, além do quadro disciplinar, podem influenciar a frequência dos estudantes nesse dia. A Tabela 3.10 mostra a ocorrência de disciplinas para essas turmas.

Esses resultados foram apresentados ao gestor escolar com o objetivo de oferecer uma alternativa à alocação de disciplinas sem professores ou novas no currículo às sextas-feiras. Como outros fatores parecem afetar negativamente a frequência nesse dia, essa prática pode reforçar a tendência dos estudantes de se ausentarem, prejudicando suas chances de sucesso nas demais disciplinas previstas para esse dia. Além disso, identificar quais disciplinas são mais atrativas para os estudantes pode ajudar a gestão a elaborar um quadro de horários mais equilibrado, distribuindo essas disciplinas de forma estratégica, juntamente com aquelas que geralmente recebem menos atenção dos alunos.

Para auxiliar a gestão a manter essas informações atualizadas, podendo-se refazer as

análises com os dados de frequência mais recentes, foi criado um aplicativo de *desktop*.

3.3.3 Aplicativo para análise de dados

O aplicativo de análise de frequência foi desenvolvido com o objetivo de facilitar a análise de dados educacionais, permitindo que gestores acompanhem as tendências de frequência estudantil de forma interativa e dinâmica. Totalmente desenvolvido em *Python* (Python Software Foundation, 2024, online), foi utilizado o *framework PyQt5* (2016) para a construção da interface gráfica, enquanto as bibliotecas *pandas* (2010), *matplotlib* (2007), *scipy.stats* (2020) e *statsmodels* (2010) são empregadas para o processamento e análise dos dados. A aplicação foi projetada para ser executada em *desktop*, dispensando a necessidade de instalação prévia do *Python* no computador, bastando ao usuário clicar no arquivo executável para iniciar a utilização.

A interface do aplicativo é intuitiva e simplificada. Na parte superior, o usuário pode carregar a tabela de frequência para análise. Caso surjam dúvidas sobre o formato necessário da tabela, há um botão logo abaixo dessa opção que fornece uma explicação detalhada sobre como ela deve ser estruturada. Após o carregamento da tabela, o usuário pode selecionar uma turma e escolher o tipo de análise desejada, com duas opções disponíveis: *Boxplot* ou Análise de Variância. Na Figura 3.2, pode-se observar a interface inicial do aplicativo, com as opções disponíveis antes de carregar qualquer tabela de dados.

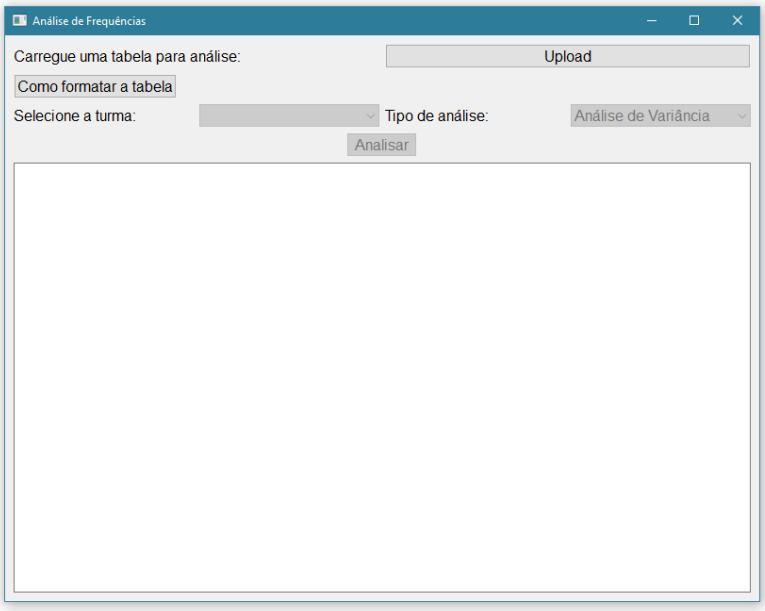

Figura 3.2 – Interface inicial do aplicativo.

O aplicativo é compatível com tabelas nos formatos *Comma-Separated Values* (.csv) e *Excel* (.xlsx). No entanto, essas informações devem ser organizadas em um formato específico, no qual a primeira coluna representa o dia da semana, a segunda coluna indica a data e as colunas subsequentes contêm o somatório das frequências diárias. Se a escola utilizar um sistema de diário eletrônico, a geração dessa tabela se torna um processo simples, já que basta calcular automaticamente o somatório das entradas de dados para cada dia.

O aplicativo é fornecido com um manual do usuário, elaborado para garantir que os

usuários compreendam como usar o sistema de forma eficaz. Este oferece instruções detalhadas sobre como gerar a tabela de dados corretamente e descreve todas as funcionalidades do aplicativo. Além disso, a Figura 3.3 mostra a janela aberta quando o usuário clica em *"Como formatar a tabela"*.

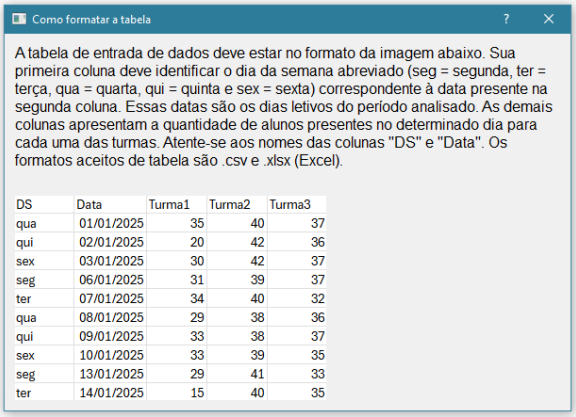

Figura 3.3 – Interface de como formatar a tabela de entrada de dados.

Após o carregamento dos dados, o usuário pode selecionar a turma desejada e o tipo de análise a ser realizado. Os nomes das turmas serão exibidos de acordo com os títulos das colunas nas tabelas carregadas. É necessário clicar em "*Analisar*"para que os resultados sejam gerados.

A Figura 3.4 exemplifica a tela gerada ao escolher a turma 1001-M e selecionar o *boxplot* como tipo de análise. Esse é um gráfico de frequência por dia da semana e permite uma rápida visualização da distribuição dos dados.

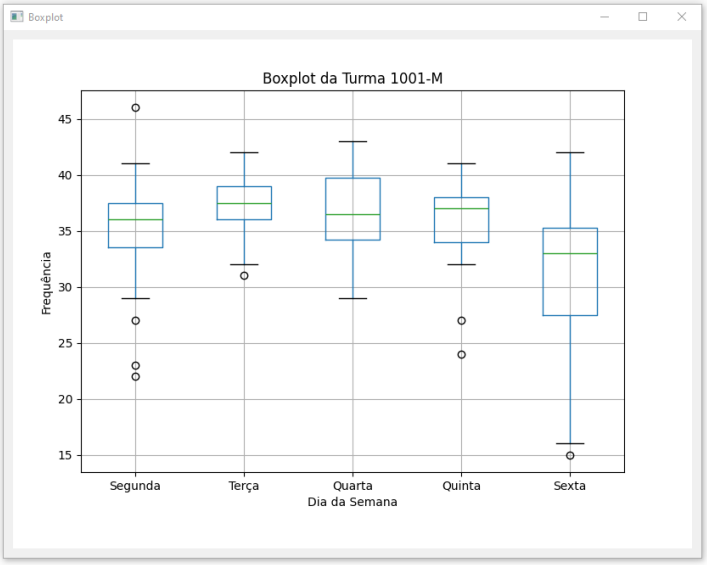

Figura 3.4 – *Boxplot* gerado pelo aplicativo.

A Figura 3.5 apresenta a tela com a análise de variância para a turma 704-T. Nessa tela, são exibidos diretamente na interface inicial os seguintes resultados, em ordem: a variância

da turma, os testes de normalidade e homogeneidade de variância (Shapiro-Wilk e Bartlett, utilizados como alternativas aos testes de Levene e Breusch-Pagan, respectivamente), a tabela de resultados da ANOVA e o resultado do teste de Tukey com as comparações entre os dias da semana e se rejeita ou não a hipótese de igualdade entre as médias.

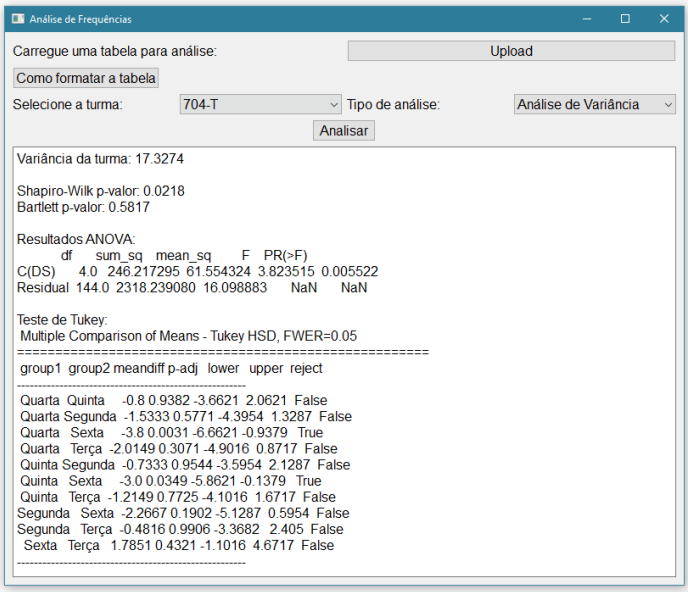

Figura 3.5 – Análise de variância gerada pelo aplicativo.

Futuramente, o aplicativo pode ser modificado para imprimir junto à análise de variância uma interpretação dos dados, afirmando se há ou não diferença significativa entre as médias e quais médias são diferentes das demais. Também é possível integrar algoritmos de previsão de frequência e sugerir estratégias em relação à confecção do quadro de horários.

O código-fonte do aplicativo, seu executável e manual do usuário podem ser encontrados em repositório especificado no Apêndice 8, seção 8.3.

4 OTIMIZAÇÃO

Este capítulo tem como objetivo apresentar uma visão geral para algumas formulações clássicas de problemas de otimização, destacando suas principais características. Para isso, são discutidos conceitos fundamentais e definições que servem de base para o estudo de algoritmos apresentados, por exemplo, nas seguintes bibliografias: *Nonlinear programming: theory and algorithms* (Bazaraa, Sherali & Shetty, 2006), *Otimização combinatória e programação linear: modelos e algoritmos* (Goldbarg & Luna, 2005), *Tópicos em Otimização Inteira* (Macambira *et al.*, 2022), *Métodos Computacionais de Otimização* (Martínez & Santos, 2020), *Introdução à pesquisa operacional* (Hillier & Lieberman, 2013), *Otimização* (Izmailov & Solodov, 2020), *Otimização - volume 2. Métodos Computacionais* (Izmailov & Solodov, 2018) e *Linear and Nonlinear Programming* (Luenberger & Ye, 2016).

De uma maneira geral, considere $K, \Omega \subseteq \mathbb{R}^n$, com $K \subseteq \Omega$ e uma função $f : \Omega \longrightarrow \mathbb{R}$. Um problema de otimização consiste em encontrar um minimizador local para f, em K, isto é, um ponto $x^* \in K$, para o qual existe $\delta > 0$, tal que $f(x^*) \leq f(x)$, para todo $x \in B_\delta(x^*) \cap K$, onde $B_\delta(x^*) = \{x \in \mathbb{R}^n : \|x - x^*\| < \delta\}$, para alguma norma $\|\cdot\|$ em $\mathbb{R}^n$. Quando x^* é tal que $f(x^*) \leq f(x)$, para todo $x \in K$, ele é denominado minimizador global para f, em K.

Em grande parte dos problemas de otimização, um minimizador local se apresenta como algo realmente factível de ser determinado. Já o mesmo não podemos garantir quando há exigência de solucão global. Matematicamente, o problema de otimização descrito é denotado como segue:

$$\begin{aligned} \text{minimizar} \quad & f(x) \\ \text{sujeito a} \quad & x \in K. \end{aligned} \tag{4.1}$$

K é chamado de **conjunto viável** para o problema, enquanto seus elementos são denominados **pontos** ou **soluções viáveis**. f é denominada **função objetivo**. Para problemas cujo o objetivo é encontrar um ponto de máximo (local ou global) para f, é padrão resolver, de maneira equivalente, um problema de minimização, considerando $-f$ como função objetivo.

4.1 OTIMIZAÇÃO LINEAR

4.1.1 Definições

Um problema de otimização linear, ou programação linear (PL), é caracterizado por uma função objetivo linear com restrições que definem o conjunto viável descritas por igualdades ou desigualdades lineares. Em geral, esses problemas se apresentam sob a forma:

$$
\begin{aligned}
\text{minimizar} \quad & c_1 x_1 + \cdots + c_n x_n \\
\text{sujeito a} \quad & a_{11} x_1 + \cdots + a_{1n} x_n = b_1 \\
& \vdots \\
& a_{m1} x_1 + \cdots + a_{mn} x_n = b_m \\
& x_1, \ldots, x_n \geqslant 0,
\end{aligned}
\tag{4.2}
$$

denominada **padrão**, onde $c_j, a_{ij}, b_i \in \mathbb{R}$, $i = 1, \ldots, m$; $j = 1, \ldots, n$, são constantes reais e $x_1, \ldots, x_n$ são as **variáveis de decisão**.

São chamadas de **restrições triviais** ou **canônicas** aquelas que indicam a não-negatividade das variáveis de decisão. Uma solução para um problema de PL consiste em uma n-upla $(x_1, \cdots, x_n)$ de números reais não-negativos que atendem as igualdades lineares presentes no problema e para o qual a função objetivo atinge seu valor ínfimo. Vetorialmente, temos a seguinte representação para um problema de PL

$$
\begin{aligned}
\text{minimizar} \quad & \boldsymbol{c}^T \boldsymbol{x} \\
\text{sujeito a} \quad & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \\
& \boldsymbol{x} \geqslant \boldsymbol{0},
\end{aligned}
\tag{4.3}
$$

onde $\boldsymbol{x}$ e $\boldsymbol{c}$ são vetores coluna n-dimensionais, $\boldsymbol{A}$ é uma matriz $m \times n$ e $\boldsymbol{b}$ um vetor m-dimensional. $\boldsymbol{c}^T$ denota a transposição realizada sobre $\boldsymbol{c}$, transformando-o em um vetor linha. A notação $\boldsymbol{x} \geqslant \boldsymbol{0}$ indica que todas as componentes de $\boldsymbol{x}$ são não-negativas, isto é, remetem as restricões triviais associadas ao problema de PL. Um vetor $\boldsymbol{x}$ que satisfaz as restrições do problema 4.3 é chamado de **solução viável** ou **factível**.

Caso o problema possua restrições de desigualdade, estas podem ser convertidas em igualdades por meio da adição de **variáveis de folga**, no caso de restrições do tipo ($\leq$), ou subtração de **variáveis de excesso**, no caso de restrições do tipo ($\geq$) (Luenberger & Ye, 2016, p. 12). Caso o problema tenha uma ou mais variáveis não-canônicas, denominadas **livres**, estas podem ser reescritas em função de outras variáveis canônicas, estejam elas já presentes na formulação original do problema ou introduzidas artificialmente. (Luenberger & Ye, 2016, p. 13)

Dado o problema 4.3, suponha que, dentre as n colunas de $\boldsymbol{A}$, seja possível selecionar um conjunto de m colunas linearmente independentes, ou seja, suponha que o posto de $\boldsymbol{A}$ seja m. Estas colunas podem ser reorganizadas para corresponderem às primeiras m colunas de $\boldsymbol{A}$. Denote, então, por $\boldsymbol{B}$, a matriz $m \times m$ determinada por essas colunas. $\boldsymbol{B}$ é não-singular e a equação

$$
\boldsymbol{B}\boldsymbol{x}_{\boldsymbol{B}} = \boldsymbol{b} \tag{4.4}
$$

possui uma única solução x_B. Definindo $\boldsymbol{x} = (\boldsymbol{x}_{\boldsymbol{B}}, \boldsymbol{0})$, isto é, atribuindo às primeiras m componentes de $\boldsymbol{x}$ os valores de $\boldsymbol{x}_{\boldsymbol{B}}$ e definindo as demais como zero, obtém-se uma solução para o sistema linear formado pelas restrições não-triviais de 4.3. As componentes de $\boldsymbol{x}$ associados às colunas de $\boldsymbol{B}$ são chamadas de **variáveis básicas** e a solução $\boldsymbol{x}$ assim obtida é chamada de **solução básica** para 4.3, com respeito à **base** $\boldsymbol{B}$, sendo também denominada **viável** caso $\boldsymbol{x}_{\boldsymbol{B}} \geq \boldsymbol{0}$.

Se alguma variável básica se anula em uma solução básica então a mesma é denominada **solução básica degenerada**. Caso contrário, trata-se de uma **solução básica não-degenerada**. Em face de viabilidade, as soluções básicas degeneradas ou não-degeneradas podem ainda ser classificadas como factíveis ou não-factíveis.

Um sistema linear $Ax = b$ sob as condições $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $\boldsymbol{b} \in \mathbb{R}^m$, $m < n$ e posto($\boldsymbol{A}$)= m, está na **forma canônica** se cada variável básica aparece uma única vez em apenas uma igualdade linear, com coeficiente 1. Em outras palavras, sejam $J = \{1,\ldots,n\}$ o conjunto dos índices das colunas de $\boldsymbol{A}$ e $\overline{J} = \{j_1,\ldots,j_m\} \subset J$ os índices das m colunas de $\boldsymbol{B}$ ou, equivalentemente, os índices das colunas linearmente independentes de $\boldsymbol{A}$. Então, o sistema linear canônico associado ao sistema linear $\boldsymbol{Ax = b}$ apresenta-se sob a forma:

$$\begin{array}{lllll} x_{j_1}\ldots & +y_{1j_{m+1}}x_{j_{m+1}} & +\ldots & +y_{1j_n}x_{j_n} & =b_1 \\ \ddots & +\vdots & +\ldots & +\vdots & =\vdots \\ x_{j_m} & +y_{mj_{m+1}}x_{j_{m+1}} & +\ldots & +y_{mj_n}x_{j_n} & =b_m \end{array} \tag{4.5}$$

A matriz ampliada do sistema em formato canônico denomina-se **tableau**. As variáveis cujos índices pertencem a $J - \overline{J}$ são chamadas de **não-básicas**.

Em geral, o conjunto viável para um problema de PL é da forma $V = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{Ax = b}; \boldsymbol{x \geqslant 0}\}$. Afirmar que um PL é viável equivale a dizer que o conjunto V é não-vazio.

Teorema 4.1.1. ***Teorema Fundamental da Programação Linear.*** *Dado um problema da forma 4.3, onde $\boldsymbol{A}$ é uma matriz $m \times n$, com* posto$(A) = m$*:*

a. Se existe uma solução viável, então existe uma solução básica viável.

b. Se existe uma solução viável ótima, então existe uma solução básica viável ótima.

Prova. Luenberger & Ye (2016, p. 21).

O Teorema 4.1.1 reduz a tarefa de resolver um problema de otimização linear à busca por soluções básicas viáveis e fundamenta um método eficiente para tal: o **método Simplex**. Esse é amplamente discutido em Arenales *et al.* (2011), Goldbarg (2005), Luenberger & Ye (2016), e Macambira *et al.* (2022).

4.1.2 Conjuntos convexos

Um conjunto não-vazio $C \subset \mathbb{R}^n$ é dito **convexo** se, para quaisquer $\boldsymbol{x_1}, \boldsymbol{x_2} \in C$, o segmento de reta conectando $\boldsymbol{x_1}$ a $\boldsymbol{x_2}$ está inteiramente contido em C. Isso equivale a dizer que, para $\alpha \in [0,1]$, tem-se que $\alpha\boldsymbol{x_1} + (1-\alpha)\boldsymbol{x_2} \in C$ (Izmailov & Solodov, 2020). Um exemplo de conjunto convexo é o conjunto viável de um PL, em seu formato padrão, ou seja, $V = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{Ax = b}; \boldsymbol{x \geqslant 0}\}$.

Um ponto $\boldsymbol{x}$ em um conjunto convexo C é chamado de **ponto extremo** de C se não houverem pontos distintos $\boldsymbol{x_1}$ e $\boldsymbol{x_2}$ em C tais que $\boldsymbol{x} = \alpha\boldsymbol{x_1} + (1-\alpha)\boldsymbol{x_2}$ para algum α, $0 < \alpha < 1$.

Teorema 4.1.2. *Seja $\boldsymbol{A}$ uma matriz $m \times n$, com* posto$(A) = m$ *e $\boldsymbol{b}$ um vetor m-dimensional. Seja V o poliedro convexo formado por todos os vetores $\boldsymbol{x} \in \mathbb{R}^n$ que satisfazem*

$$\boldsymbol{Ax = b\ ;\ x \geqslant 0}. \tag{4.6}$$

$\boldsymbol{x}$ é um ponto extremo de V se, e somente se, $\boldsymbol{x}$ é uma solução básica viável para V.

Prova. Luenberger & Ye (2016, p. 23).

Alguns resultados importantes decorrem da correspondência entre pontos extremos, também chamados de vértices, e soluções básicas, conforme estabelecido pelo Teorema 4.1.2. Primeiramente, se o conjunto viável V for não vazio, então ele contém pelo menos um ponto extremo. Além disso, se um problema de programação linear possui uma solução ótima finita, então existe pelo menos uma solução ótima que é um ponto extremo de V. Como V possui um número finito de pontos extremos, a busca pela solução ótima pode ser restringida a esses pontos, demonstrando que uma solução pode ser obtida em tempo computacional finito.

Se, além de ser não vazio, V também for limitado, ele forma um poliedro convexo, ou seja, seus pontos podem ser expressos como combinações convexas de um número finito de pontos extremos. Nesse caso, diz-se que o problema de PL é bem formulado.

O Teorema Fundamental da Programação Linear, juntamente com o Teorema 4.1.2, estabelece que as soluções ótimas podem ser encontradas nos vértices (ou soluções básicas) da região viável. Com base nesse princípio, os algoritmos de otimização, como o método Simplex, são projetados para explorar essas propriedades. O Simplex opera garantindo que, a cada iteração, a solução permaneça viável, movendo-se de um vértice a outro dentro da região factível. Ele também assegura melhorias sucessivas na função objetivo, garantindo que o processo converge para uma solução ótima (caso esta exista).

4.2 OTIMIZAÇÃO INTEIRA BINÁRIA

Quando as variáveis de um problema de otimização correspondem à ocorrência ou não de um certo evento, essa dicotomia é modelada por uma variável binária x tal que (Arenales, 2007):

$$x = \begin{cases} 1, & \text{se o evento ocorre,} \\ 0, & \text{se o evento não ocorre.} \end{cases} \tag{4.7}$$

Um problema de Programação Inteira Binária (PIB) pode ser formulado como segue em 4.8 , onde $\boldsymbol{x}$ é o vetor de variáveis binárias, c_i são os coeficientes da função objetivo, $\boldsymbol{A}$ é uma matriz de coeficientes das restrições lineares e $\boldsymbol{b}$ é o vetor de termos independentes das restrições:

$$\begin{aligned} &\text{maximizar } f(\boldsymbol{x}) = c_1 x_1 + c_2 x_2 + \cdots + c_n x_n \\ &\quad \text{sujeito a } \boldsymbol{Ax} \leqslant \boldsymbol{b}, \\ &\qquad x_i \in \{0,1\}, \forall\, i \in \{1,\ldots,n\}. \end{aligned} \tag{4.8}$$

Os problemas de PIB pertencem à classe NP-difícil, pois como cada variável binária pode assumir dois valores, um problema com n variáveis possui até 2^n soluções possíveis, ou seja, cada vez que n é incrementado em uma unidade, o número de soluções dobra (Hillier & Lieberman, 2014). Métodos de enumeração explícita, que exploram todas as possíveis combinações de variáveis, tornam-se computacionalmente inviáveis para valores altos de n, apesar da quantidade de pontos a serem visitados ser finita.

4.2.1 Problema da mochila

Problemas que envolvam a alocação ou não alocação de itens em compartimentos de forma a maximizar uma função de utilidade são conhecidos como **problemas da mochila**.

Alternativamente, pode-se entendê-lo como o desafio de encher uma mochila sem ultrapassar um limite de peso. Apesar disso, sua formulação pode ser aplicada em uma série de diferentes problemas, como investimento de capital, problemas de corte e empacotamento, carregamento de veículos, orçamento, etc. (Goldbarg, 2005)

O problema da mochila 0-1, onde assume-se que exista um objeto de cada tipo dentro de um conjunto de objetos que podem ocupar as diferentes partes da mochila, pode ser formulado como:

$$\begin{aligned} &\text{maximizar} && z = \textstyle\sum_{j=1}^{n} c_j x_j \\ &\text{sujeito a} && \textstyle\sum_{j=1}^{n} w_j x_j \leqslant b \\ & && x_j \in \{0,1\},\ j = 1,\ldots,n, \end{aligned} \tag{4.9}$$

onde x_j representa se o objeto j foi selecionado ou não, b representa a capacidade da mochila, c_j representa o valor de cada item e w_j seu peso.

Em particular, considere n itens que devem ser colocados em m mochilas de capacidades distintas $b = 1,\ldots,m$, e deseja-se selecionar m subconjuntos de itens tal que cada subconjunto ocupe uma capacidade de, no máximo, b_i e o lucro total seja maximizado. Se cada item j tem lucratividade p_j e peso w_j, e x_{ij} for uma variável binária que corresponda à alocação ($x_{ij} = 1$) ou não alocação ($x_{ij} = 0$) desse item, o problema é formulado como (Arenales, 2007):

$$\text{maximizar} \sum_{i=1}^{m}\sum_{j=1}^{n} p_j x_{ij} \tag{4.10}$$

$$\text{sujeito a} \sum_{j=1}^{n} w_j x_{ij} \leqslant b_i,\ i = 1,\ldots,m \tag{4.11}$$

$$\sum_{i=1}^{m} x_{ij} \leqslant 1,\ j = 1,\ldots,n, \tag{4.12}$$

$$\boldsymbol{x} \in B^{mn}. \tag{4.13}$$

A esse tipo de problema chama-se **problema de múltiplas mochilas**. Nele, 4.10 representa a maximização do lucro, as restrições 4.11 garantem que a capacidade b_i da mochila i não seja excedida, as restrições 4.12 indicam que se o item j é escolhido, então ele é colocado em uma única mochila, e a restrição 4.13 refere-se ao tipo das variáveis.

4.2.2 *Branch-and-bound*

O método *Branch and Bound* (B&B) é uma técnica fundamental para a resolução de problemas de otimização combinatória. Sua ideia central é particionar o problema original em subproblemas menores no espaço das soluções (*branching*) e utilizar limites superiores e inferiores para eliminar subproblemas inviáveis ou que não possam levar a soluções melhores (*bounding*). Definindo os problemas auxiliares:

$$(P) = \text{ maximizar } \{cx | Ax = b, x \geqslant 0, x \in \mathbb{Z}^+\} \tag{4.14}$$

$$(\overline{P}) = \text{ maximizar } \{cx | Ax = b, x \geqslant 0, x \in \mathbb{R}^+\} \tag{4.15}$$

e sendo $V^*(P)$ e $V^*(\overline{P})$ os valores das funções objetivo no ótimo de (P) e $\overline{P}$, respectivamente, tem-se que $v^*(P) \leqslant V*(\overline{P})$.

Seja qualquer solução viável $\hat{x}$ de (P) e seja $V(\hat{x})$ o valor da função objetivo no ponto $\hat{x}$. Então $V(\hat{x}) \leqslant V^*(P)$ e, dessa forma, $V*(\overline{P})$ é um limite superior para (P) e qualquer de suas soluções viáveis. Se $\overline{x}$ é solução ótima de $(\overline{P})$ tal que $\overline{x}_j$ é não inteiro, tem-se:

$$x_j \geqslant \lfloor \overline{x}_j \rfloor + 1 \text{ ou } x_j \leqslant \lfloor \overline{x}_j \rfloor \tag{4.16}$$

em toda solução viável de (P). Assim, divide-se (P) em dois novos problemas, (P_1) e (P_2), em que a envoltória convexa C de $(P_1) \cup (P_2)$ esteja estritamente contida na envoltória de (P). A envoltória convexa de um conjunto X é o menor conjunto convexo que contém todos os pontos de X. Essa estratégia de separação cria problemas mais restritos, normalmente de mais fácil solução, e pode ser aplicada inúmeras vezes, criando o que é chamado de **árvore branch**. (Goldbarg, 2005)

Um nó na árvore branch corresponde a um problema P_i. Esses nós podem ser descartados na busca de soluções por testes que utilizam técnicas de **relaxação linear**, resolvendo o problema alternativo PL_i, que permite a determinação das variáveis dentre os números reais. Seja $F(P_i)$ a região factível do problema P_i, $F(PL_i)$ a região factível do problema PL_i, z_i o valor ótimo do problema P_i, $\overline{z}_i$ o valor ótimo do problema PL_i, $\boldsymbol{x^*}$ a melhor solução encontrada até determinado momento e z^* o valor do problema para $\boldsymbol{x^*}$. O nó correspondente a P_i é eliminado se satisfizer um dos seguintes testes (Arenales, 2011):

1. $F(PL_i) = \emptyset$: P_i é eliminado por infactibilidade;
2. $\overline{z}_i \leqslant z^*$: P_i é eliminado por qualidade;
3. Se a solução ótima de PL_i é inteira, P_i é eliminado por otimalidade.

Um algoritmo básico do método B&B encontra-se abaixo, onde o nó 0 da árvore branch corresponde ao problema original de programação inteira P, e nós não eliminados pelos testes e ainda sem subdivisões é chamado de **nó ativo** e armazenado em uma lista L (Arenales, 2011):

(i.) *(Inicialização)* Faça $\overline{z} = \infty$, $z^* = -\infty$, $\boldsymbol{x}^* = \emptyset$ e $L = \{P\}$.

(ii.) *(Seleção de nó)* Selecione o nó ativo i, associado ao problema P_i, da lista de nós ativos. Se não houver nó na lista, vá para (vii.).

(iii.) *(Teste de eliminação 1)* Se a região factível de PL_i for vazia, vá para (i).

(iv.) *(Teste de eliminação 2)* Se o valor $\overline{z}_i$ da solução ótima de PL_i é tal que $\overline{z}_i \leqslant z^*$, vá para (i).

(v.) *(Teste de eliminação 3)* Se a solução ótima $\overline{\boldsymbol{x}}_i$ de PL_i é inteira com valor $\overline{z}_i$, e se $\overline{z}_i > z^*$, atualize $\boldsymbol{x}^*$ e z^*. Elimine nós ativos i da lista L tais que $\overline{z}_i \leqslant z^*$ e volte para (i).

(vi.) *(Ramificação)* Selecione uma variável da solução ótima $\overline{\boldsymbol{x}}_i$ de PL_i com valor não inteiro e divida P_i em dois problemas. Adicione estes à lista L e vá para (i).

(vii.) *(Fim)* Se $z^* = -\infty$, não existe solução factível; caso contrário, a solução $\boldsymbol{x}^*$ é uma solução ótima.

Mais sobre o método de *branch-and-bound* pode também ser encontrado em Macambira *et al.* (2022).

4.2.3 Planos de corte

Os algoritmos de plano de corte buscam construir uma aproximação da envoltória convexa da região factível de um problema de otimização inteira, garantindo que essa aproximação contenha um ponto extremo correspondente a uma solução ótima.

Essa aproximação é obtida por meio da adição de cortes, ou **desigualdades válidas**, que restringem progressivamente a região factível sem remover soluções inteiras viáveis. Uma desigualdade da forma $\phi \boldsymbol{x} \leqslant \phi_0$ é considerada válida para um conjunto $X \subset \mathbb{R}^n$ se todos os pontos de X estiverem em um dos semi-espaços definidos pelo hiperplano $\phi \boldsymbol{x} = \phi_0$. (Arenales, 2011)

O passo a passo conhecido como **procedimento de Chvátal-Gomory** para a construção de uma desigualdade válida para um conjunto $X = \{\boldsymbol{x} : \boldsymbol{Ax} \leqslant \boldsymbol{b}, \boldsymbol{x} \in \mathbb{Z}^n_+\}$, tal que $\boldsymbol{A}$ é uma matriz $m \times n$ com colunas $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ e $\boldsymbol{u} \in \mathbb{R}^m_+$, segue (Arenales, 2011):

(i.) A desigualdade $\sum_{j=1}^n \boldsymbol{u^T a}_j x_j \leqslant \boldsymbol{u^T b}$ é válida para X pois $\boldsymbol{u} \geqslant 0$ e $\sum_{j=1}^n \boldsymbol{a}_j x_j \leqslant \boldsymbol{b}$;

(ii.) A desigualdade $\sum_{j=1}^n \lfloor \boldsymbol{u^T a}_j \rfloor x_j \leqslant \boldsymbol{u^T b}$ é válida para X pois $\boldsymbol{x} \geqslant 0$;

(iii.) A desigualdade $\sum_{j=1}^n \lfloor \boldsymbol{u^T a}_j \rfloor x_j \leqslant \lfloor \boldsymbol{u^T b} \rfloor$ é válida para X, pois $\boldsymbol{x}$ é inteiro e, portanto, $\sum_{j=1}^n \lfloor \boldsymbol{u^T a} \rfloor x_j$ é inteiro.

Seja o problema de PI (P) e sua relaxação linear (PL):

$$(P) \quad z = \max \{\boldsymbol{c^T x} : \boldsymbol{Ax = b}, \boldsymbol{x} \in \mathbb{Z}^n_+\}, \tag{4.17}$$

$$(PL) \quad \overline{z} = \max \{\boldsymbol{c^T x} : \boldsymbol{Ax = b}, \boldsymbol{x} \in \mathbb{R}^n_+\}. \tag{4.18}$$

Considerando a solução ótima do PL, em que x_{B_i}, $i = 1, \ldots, m$, representam as variáveis básicas e NB denota o conjunto das variáveis não-básicas, o PL pode ser reescrito como:

$$\begin{aligned} z = \quad & \max \ \hat{a}_{00} + \textstyle\sum_{j \in NB} \hat{a}_{0j} x_j \\ & x_{B_i} + \textstyle\sum_{j \in NB} \hat{a}_{ij} x_j = \hat{a}_{i0}, \ i = 1, \ldots, m, \\ & \boldsymbol{x} \in \mathbb{R}^n_+, \end{aligned} \tag{4.19}$$

onde $\hat{a}_{0j} \leqslant 0$ para $j \in NB$ e $\hat{a}_{i0} \geqslant 0$, $i = 1, \ldots, m$.

Se a solução ótima não é inteira, então existe uma linha i com y_{i0} não inteiro, e o corte de Chvátal-Gomory para essa linha é:

$$x_{B_i} + \sum_{j \in NB} \lfloor \hat{a}_{ij} \rfloor x_j \leqslant \lfloor \hat{a}_{i0} \rfloor. \tag{4.20}$$

Substituindo x_{B_i} em 4.20:

$$\sum_{j \in NB} (\hat{a}_{ij} - \lfloor \hat{a}_{ij} \rfloor) x_j \geqslant \hat{a}_{i0} - \lfloor \hat{a}_{ij} \rfloor, \tag{4.21}$$

ou $\sum_{j \in NB} f_{ij} x_j \geqslant f_{i0}$, com $f_{ij} = \hat{a}_i j - \lfloor \hat{a}_{ij} \rfloor$, $0 \leqslant f_{ij} < 1$ e $f_{i0} = \hat{a}_{i0} - \lfloor \hat{a}_{i0} \rfloor$, $0 < f_{i0} < 1$.

Como na solução básica ótima $x_j = 0$ para todo $j \in NB$, $\sum_{j \in NB} f_{ij} x_j \geqslant f_{i0}$ faz um corte na solução básica ótima. Introduzindo a variável de sobra $s_i \geqslant 0$, tem-se $s_i = \sum_{j \in NB} f_{ij} x_j - f_{i0}$. De 4.19, seque que:

$$x_{B_i} = f_{i0} - \sum_{j \in NB} f_{ij} x_j + \lfloor \hat{a}_{i0} \rfloor - \sum_{j \in NB} \lfloor \hat{a}_{ij} \rfloor x_j, \tag{4.22}$$

ou $x_{B_i} = -s_i + \lfloor \hat{a}_{i0} \rfloor - \sum_{j \in NB} \lfloor \hat{a}_{ij} \rfloor x_j$ e, portanto, s_i é não-negativa e inteira, dado que os outros termos da equação são inteiros. (Arenales, 2011)

O **algoritmo de Gomory** para problemas de otimização inteira consiste na geração iterativa de uma sequência de problemas de programação linear PL_i, $i = 0, \ldots, n$. Nessa sequência, os conjuntos de soluções factíveis seguem a relação $X = \overline{X}_0 \supseteq \cdots \supseteq \overline{X}_n$, enquanto os valores ótimos da função objetivo obedecem $\overline{z} = \overline{z}_0 \geqslant \cdots \geqslant \overline{z}_n \geqslant z$. Aqui, $\overline{X}_i$ e $\overline{z}_i$ representam, respectivamente, a região factível e a solução ótima do problema PL_i. O algoritmo itera até que a solução obtida seja inteira, refinando a região factível por meio da adição de cortes de Gomory, que eliminam apenas soluções fracionárias sem remover soluções inteiras viáveis.

4.2.4 *Branch-and-cut*

O método ***Branch-and-cut*** combina as estratégias do *branch-and-bound* e dos planos de cortes, visando reduzir o número de nós explorados na árvore de busca do B&B. Em cada nó, são adicionadas desigualdades válidas que restringem a região factível, permitindo obter um limitante superior mais preciso e, assim, acelerar a convergência para a solução ótima. O algoritmo inclui uma etapa de pré-processamento, na qual a formulação original do problema é analisada para identificar variáveis e restrições redundantes, além de refinar os limites das variáveis, melhorando a eficiência da resolução. Os detalhes desse pré processamento podem ser encontrados em Arenales (2011). Um algoritmo de *branch-and-cut* que inclui k cortes em cada nó cuja relaxação linear é factível segue:

(i.) (Inicialização) Faça $\overline{z} = \infty$, $z^* = -\infty$, $\boldsymbol{x}^* = \emptyset$. Pré-processe o problema inicial e coloque-o na lista $L = \{P\}$;

(ii.) (Seleção de nó) Selecione o nó ativo i, associado ao problema P_i, da lista de nós ativos. Se a lista estiver vazia, vá para (vii.);

(iii.) (Teste de eliminação 1) Se a região factível de PL_i estiver vazia, volte a (i.);

(iv.) (Corte) Tente eliminar a solução ótima de PL_i. Se não for possível, faça $k = 0$ e vá para (v.). Caso contrário, adicione k cortes a PL_i de forma a obter a formulação PL_{ik}.

(v.) (Teste de eliminação 2) Se $\overline{z}_{ik}$ de PL_{ik} for tal que $\overline{z}_{ik} \leqslant z^*$, volte a (i.);

(vi.) (Teste de eliminação 3) Se a solução ótima $\overline{\boldsymbol{x}}_{ik}$ de PL_{ik} for inteira com valor $\overline{z}_{ik}$, e se $\overline{z}_{ik} > z^*$, atualize $\boldsymbol{x}^*$ e z^*. Elimine nós ativos i da lista L tais que $\overline{z}_i \leqslant z^*$ e volte a (i.);

(vii.) (Ramificação) Selecione uma variável da solução ótima $\overline{\boldsymbol{x}}_{ik}$ de PL_{ik} com valor não inteiro e divida P_{ik} em dois problemas. Adicione estes problemas à lista L e volte a (i.);

(viii.) (Fim) Se $z^* = -\infty$, não existe solução factível; caso contrário, $\boldsymbol{x}^*$ é uma solução ótima.

4.3 OTIMIZAÇÃO QUADRÁTICA

A Programação Quadrática (PQ) é uma área da otimização que trata da resolução de problemas onde a função objetivo é quadrática, ou seja, é um polinômio em n variáveis com termos de até segunda ordem (Martínez, 2020) e as restrições são lineares.

4.3.1 Problemas irrestritos

Seja $\boldsymbol{G} \in \mathbb{R}^{n\times n}$ uma matriz simétrica, $\boldsymbol{b} \in \mathbb{R}^n$ um vetor e $c \in \mathbb{R}$ uma constante. Um problema de otimização quadrática irrestrito é do tipo:

$$\text{minimizar} \quad q(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{G}\boldsymbol{x} + \boldsymbol{b}^{\boldsymbol{T}}\boldsymbol{x} + c. \tag{4.23}$$

Observe que $\nabla q(\boldsymbol{x}) = \boldsymbol{G}\boldsymbol{x} + \boldsymbol{b}$ e $\nabla^2 q(\boldsymbol{x}) = \boldsymbol{G}$, para todo $\boldsymbol{x} \in \mathbb{R}^n$. O fato do gradiente ser uma função linear e a hessiana ser invariável em relação a $\boldsymbol{x}$ contribui para a estrutura analítica do problema, facilitando sua resolução.

Considere o problema 4.23 e seja $\overline{\boldsymbol{x}} \in \mathbb{R}^n$. Se $q(\overline{\boldsymbol{x}}) \leqslant q(\boldsymbol{x})$ para todo $\boldsymbol{x} \in \mathbb{R}^n$, $\overline{\boldsymbol{x}}$ é chamado de **mínimo global**. Se existir uma ϵ-vizinhança $N_\epsilon(\overline{\boldsymbol{x}})$ tal que $q(\overline{\boldsymbol{x}}) \leqslant q(\boldsymbol{x})$ para todo $\boldsymbol{x} \in N_\epsilon(\overline{\boldsymbol{x}})$, $\overline{\boldsymbol{x}}$ é chamado de **mínimo local**.

Dado um ponto $\overline{\boldsymbol{x}} \in \mathbb{R}^n$, para determinar se esse é um mínimo local ou global de uma função q, é necessário caracterizar uma solução que minimize q. Os meios para tal são dados pela própria diferenciabilidade da função. O Corolário 4.3.1 a seguir, em decorrência do Teorema 4.3.1, estabelece condição necessária de primeira ordem para $\overline{\boldsymbol{x}}$ ser um mínimo local.

Teorema 4.3.1. *Suponha que $q : \mathbb{R}^n \longrightarrow \mathbb{R}$ seja diferenciável em $\overline{\boldsymbol{x}}$. Se houver um vetor $\boldsymbol{d}$ tal que $\nabla q(\overline{\boldsymbol{x}})^t\boldsymbol{d} < 0$, então existe $\delta > 0$ tal que $q(\overline{\boldsymbol{x}} + \lambda\boldsymbol{d}) < q(\overline{\boldsymbol{x}})$ para todo $\delta \in (0, \delta)$, tal que $\boldsymbol{d}$ é uma **direção de descida** de q em $\overline{\boldsymbol{x}}$.*

Prova. Bazaraa, Sherali & Shetty (2006, p. 167).

Corolário 4.3.1. ***Condição necessária de primeira ordem.*** *Suponha que $q : \mathbb{R}^n \longrightarrow \mathbb{R}$ seja diferenciável em $\overline{\boldsymbol{x}}$. Se $\overline{\boldsymbol{x}}$ for um mínimo local, então $\nabla q(\overline{\boldsymbol{x}}) = \boldsymbol{0}$.*

Prova. Bazaraa, Sherali & Shetty (2006, p. 167).

Os pontos que anulam o gradiente de uma função são conhecidos como **pontos estacionários**. Pelo Corolário 4.3.1, os pontos estacionários de 4.23 são as soluções do sistema linear:

$$\boldsymbol{G}(\boldsymbol{x}) + \boldsymbol{b} = \boldsymbol{0}. \tag{4.24}$$

Portanto, sua existência e unicidade estão determinados pelas propriedades desse sistema. Ele admitirá ponto estacionário se, e somente se, $b \in R(\boldsymbol{G})$, onde $R(\boldsymbol{G})$ é o espaço coluna de $\boldsymbol{G}$, e esse ponto será único se, e somente se, $\boldsymbol{G}$ for não singular. (Martínez, 2020)

Condições necessárias para a determinação de minimizadores também podem ser expressas em termos da matriz hessiana $\boldsymbol{H}$, cujos elementos correspondem às derivadas parciais de segunda ordem de q. Essas condições são, portanto, conhecidas como condições de segunda ordem.

Teorema 4.3.2. ***Condição necessária de segunda ordem.*** *Suponha que $q : \mathbb{R}^n \longrightarrow \mathbb{R}$ seja duas vezes diferenciável em $\overline{\boldsymbol{x}}$. Se $\overline{\boldsymbol{x}}$ for um mínimo local, $\nabla q(\overline{\boldsymbol{x}}) = 0$ e $\boldsymbol{H}(\overline{\boldsymbol{x}})$ é semidefinida positiva.*

Prova. Bazaraa, Sherali & Shetty (2006, p. 168).

As condições vistas até então são necessárias, ou seja, devem ser verdadeiras para toda solução ótima local. No entanto, existem pontos que satisfazem essas condições mas não são minimizadores de q. A Condição suficiente para a otimalidade é dada pelo Teorema 4.3.3.

Teorema 4.3.3. ***Condição suficiente de otimalidade.*** *Suponha que $q : \mathbb{R}^n \longrightarrow \mathbb{R}$ seja duas vezes diferenciável em $\overline{\boldsymbol{x}}$. Se $\nabla q(\overline{\boldsymbol{x}}) = 0$ e $\boldsymbol{H}(\overline{\boldsymbol{x}})$ for definida positiva, $\overline{\boldsymbol{x}}$ é um mínimo local.*

Prova. Bazaraa, Sherali & Shetty (2006, p. 168).

Particularmente, se 4.23 admitir um minimizador local, então $\boldsymbol{G \succcurlyeq 0}$. Por sua vez, isso implica na convexidade da função objetivo, e todo minimizador local para 4.23 é também minimizador global. (Martínez, 2020)

4.3.2 Problemas em caixas

Alguns problemas de quadrados mínimos estão sujeitos a restrições de intervalo, também conhecidas como **restrições de caixa**, da forma $l_i \leqslant x_i \leqslant u_i$, para todo $i = 1, \ldots, n$. O conjunto $\Omega \subset \mathbb{R}^n$ formado pelos pontos que satisfazem essas restrições é chamado de **caixa**. Esses casos se diferenciam pois a matriz $\boldsymbol{G}$ será a hessiana da função objetivo em determinado ponto e não se pode supor que seja semidefinida positiva.

A forma geral de um problema quadrático em caixa é:

$$\begin{aligned} &\text{minimizar} \quad && q(\boldsymbol{x}) \\ &\text{sujeito a} \quad && \boldsymbol{x} \in \Omega, \end{aligned} \tag{4.25}$$

onde $\Omega = \{\boldsymbol{x} \in \mathbb{R}^n | l \leqslant \boldsymbol{x} \leqslant u, l < u\}$ e $q(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x^T G x + b^T x} + c$.

Seja $\gamma = \min\{u_i - l_i, i = 1, \ldots, n\}$, $\overline{g}(\boldsymbol{x}) \equiv -\nabla q(\boldsymbol{X}) \equiv -(\boldsymbol{Gx + b})$ e $L > 0$ uma cota superior do maior autovalor de $\boldsymbol{G}$. Para todo $\boldsymbol{x}, \boldsymbol{z} \in \mathbb{R}^n$,

$$q(\boldsymbol{z}) - q(\boldsymbol{x}) - \nabla q(\boldsymbol{x})^T(\boldsymbol{z} - \boldsymbol{x}) = \frac{1}{2}(\boldsymbol{z} - \boldsymbol{x})^T G(z - x) \leqslant \frac{L}{2}||\boldsymbol{z} - \boldsymbol{x}||^2. \tag{4.26}$$

Uma **face aberta** de Ω é um conjunto $F_I \subset \Omega$, onde I é um subconjunto de $\{1, \ldots, 2n\}$ que não contenha simultaneamente i e $n + i$, $i \in \{1, \ldots, n\}$, tal que:

$$F_I = \{\boldsymbol{x} \in \Omega | x_i = l_i, \text{ se } i \in I, x_i = u_i, \text{ se } n + i \in I, \text{ ou } l_i < x_i < u_i\}. \tag{4.27}$$

Faces abertas correspondentes a sub-índices diferentes são disjuntas e Ω é a união de todas as suas faces abertas. Chame de $\overline{F}_I$ o fecho de cada face aberta, $V(F_I)$ a menor variedade afim que contenha F_I, $S(F_I)$ o subespaço paralelo a $V(F_I)$ e $\dim F_I$ a dimensão de $S(F_I)$. Se $|I|$ denotar o número de elementos de I, ou, equivalentemente, o número de restrições ativas nos pontos de F_I, então $\dim F_I = n - |I|$. (Martínez, 2020)

Para cada $\boldsymbol{x} \in \Omega$, seja $\overline{g}P(\boldsymbol{x}) \in \mathbb{R}^n$ o gradiente projetado negativo:

$$\overline{g}P(\boldsymbol{x}) = \begin{cases} 0, & \text{se } x_i = l_i \text{ e } [\nabla q(\boldsymbol{x})]_i > 0, \\ 0, & \text{se } x_i = u_i \text{ e } [\nabla q(\boldsymbol{x})]_i < 0, \\ -[\nabla q(\boldsymbol{x})]_i & \text{nos outros casos.} \end{cases} \tag{4.28}$$

Pela aplicação da condição necessária de primeira ordem, observa-se que, se $\boldsymbol{x}$ é minimizador local ou global de 4.25, então $\overline{g}P(\boldsymbol{x}) = \boldsymbol{0}$. Se $\boldsymbol{G \succcurlyeq 0}$, essa condição passa a ser suficiente para $\boldsymbol{x}$ ser um minimizador global.

Alguns algoritmos para minimizar quadráticas em caixas produzem sequências $\{\boldsymbol{x}_k\}$ de aproximações da solução de 4.25 baseadas na minimização parcial da quadrática nas diferentes faces visitadas. Aqui será discutido uma extensão do **método de Newton** para resolver problemas de programação quadrática com restrições de caixa.

4.3.3 Método de Newton

O método de Newton, em sua forma básica, é uma ferramenta para resolver sistemas de equações não-lineares. A motivação de sua utilização como método de otimização vêm de sua convergência rápida devido ao uso de informações de segunda ordem (Izmailov & Solodov, 2018). Como $\nabla q(\boldsymbol{x}) = \mathbf{0}$ é um sistema não linear, o método de Newton pode ser aplicado. Porém, como não é um método de otimização, não dá preferências a minimizadores sobre maximizadores ou sobre qualquer outro possível ponto estacionário (Martínez, 2020). No entanto, esse método serve de base para outros melhor adaptados à otimização, conhecidos como métodos quase-Newton.

Suponha que queira achar um $\boldsymbol{x} \in \mathbb{R}^n$ tal que $\phi(\boldsymbol{x}) = \mathbf{0}$, onde $\phi : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ é diferenciável. Seja $\boldsymbol{x}_k \in \mathbb{R}^n$ uma aproximação a alguma solução $\overline{\boldsymbol{x}}$. Em torno de $\boldsymbol{x}_k$, pode-se aproximar a equação por sua linearização:

$$\phi(\boldsymbol{x}_k) + \phi'(\boldsymbol{x}_k)(\boldsymbol{x} - \boldsymbol{x}_k) = \mathbf{0}. \tag{4.29}$$

A 4.29 é a equação de iteração do método de Newton. Seu algoritmo parte da escolha de um $\boldsymbol{x}_0 \in \mathbb{R}^n$ e, tomando $k = 0$ (Izmailov & Solodov, 2018):

(i.) Calcular $\boldsymbol{x}_{k+1} \in \mathbb{R}^n$ como solução da equação linear 4.29;

(ii.) Tomar $k := k + 1$ e retornar a (i.).

Se $\phi'(\boldsymbol{x}_k)$ for não-singular, o método de Newton pode ser escrito como:

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - (\phi'(\boldsymbol{x}_k))^{-1}\phi(\boldsymbol{x}_k),\ k = 0, 1, \ldots \tag{4.30}$$

Esse algoritmo gera uma sequência $\{\boldsymbol{x}_k\}$ bem definida que converge a $\overline{\boldsymbol{x}}$ a uma taxa de convergência quadrática (Izmailov & Solodov, 2018).

Pode-se modificar o método de Newton de forma a considerar direções de descida. Observe que, quando as direções $\boldsymbol{d}_k$ são geradas como soluções de um sistema linear $\boldsymbol{B}_k\boldsymbol{d}_k = -\nabla f(\boldsymbol{x}_k)$, $\boldsymbol{d}_k^T\boldsymbol{B}_k\boldsymbol{d}_k = -\boldsymbol{d}_k^T\nabla f(\boldsymbol{x}_k)$. Portanto, direções de descida são geradas se $\boldsymbol{B}_k > \mathbf{0}$, e deve-se então impor que as matrizes que geram direções de busca em métodos de minimização sejam definidas positivas.

Assim, dados $\alpha \in (0, 1)$, $\beta > 0$, $\theta \in (0, 1)$ e $\boldsymbol{x}_k \in \mathbb{R}^n$, tem-se o seguinte algoritmo de Newton com busca linear (Martínez, 2020):

(i.) Se $\nabla f(\boldsymbol{x}) = \mathbf{0}$, parar;

(ii.) Tentar a fatoração de Cholensky: $\nabla^2 f(\boldsymbol{x}_k) = LDL^T$;

(iii.) Se houve sucesso em (ii.), obter $\boldsymbol{d}_k$ resolvendo: $L\boldsymbol{z} = -\nabla f(\boldsymbol{x}_k)$ e $DL^T\boldsymbol{d}_k = \boldsymbol{z}$;

(iv.) Se (ii.) fracassou, definir $\boldsymbol{B}_k = \nabla^2 f(\boldsymbol{x}_k) + \mu I$, $\mu > 0$, de maneira que $\boldsymbol{B}_k > 0$. Obter a fatoração de Cholensky: $\boldsymbol{B}_k = \overline{LDL}^T$ e calcular $\boldsymbol{d}_k$ resolvendo $\overline{L}\boldsymbol{z} = -\nabla f(\boldsymbol{x}_k)$ e $\overline{DL}^T\boldsymbol{x}_k = \boldsymbol{z}$;

(v.) Se $\nabla f(\boldsymbol{x}_k)^T\boldsymbol{d}_k > -\theta||\nabla f(\boldsymbol{x}_k)||\,||\boldsymbol{d}_k||$, fazer $\mu \leftarrow \max\{2\mu, 10\}$ e repetir o passo (iv.);

(vi.) Se $||\boldsymbol{d}_k|| < \beta||\nabla f(\boldsymbol{x}_k)||$, corrigir $\boldsymbol{x}_k \leftarrow \beta\frac{||\nabla f(\boldsymbol{x}_k)||}{||\boldsymbol{d}_k||}$;

(vii.) Obter t por *backtracking* de modo a satisfazer $f(\boldsymbol{x}_k + t\boldsymbol{d}_k) \leqslant f(\boldsymbol{x}_k) + \alpha t\nabla f(\boldsymbol{x}_k)^T\boldsymbol{d}_k$, definir $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + t\boldsymbol{d}_k$ e voltar para (i.).

Esse último algoritmo é adequado para a minimização de problemas irrestritos.

O método quase-Newton mais amplamente utilizado é o **método BFGS**, desenvolvido por Broyden, Fletcher, Goldfarb e Shanno. Ele se destaca por ser um método eficiente para otimização irrestrita de funções diferenciáveis, oferecendo uma alternativa ao método de Newton quando a matriz hessiana não está disponível ou seu cálculo é computacionalmente caro (Martínez, 2020).

No método BFGS, aproxima-se a matriz hessiana com atualizações sucessivas baseadas apenas nas informações do gradiente da função objetivo. Seja $\boldsymbol{s}_k = \boldsymbol{x}_{k+1} - \boldsymbol{x}_k$ o deslocamento entre duas iterações, $\boldsymbol{y}_k = \nabla f(\boldsymbol{x}_{k+1}) - \nabla f(\boldsymbol{x}_k)$ a variação do gradiente e $\boldsymbol{B}_k$ a aproximação da hessiana. A atualização de $\boldsymbol{B}_k$ é dada pela seguinte expressão, conhecida como **fórmula BFGS** (Martínez, 2020):

$$\boldsymbol{B}_{k+1} = \boldsymbol{B}_k + \frac{\boldsymbol{y}_k \boldsymbol{y}_k^{\boldsymbol{T}}}{\boldsymbol{y}_k^T \boldsymbol{s}_k} - \frac{\boldsymbol{B}_k \boldsymbol{s}_k \boldsymbol{s}_k^T \boldsymbol{B}_k}{\boldsymbol{s}_k^T \boldsymbol{B}_k \boldsymbol{s}_k}. \tag{4.31}$$

Teorema 4.3.4. *Na fórmula BFGS, se $\boldsymbol{B}_k$ for simétrica definida positiva e $\boldsymbol{s}_k^T \boldsymbol{y}_k > 0$, então $\boldsymbol{B}_{k+1}$ também é simétrica e definida positiva.*

Prova. Martínez (2020, p. 111).

A fórmula BFGS tem, portanto, a característica de gerar matrizes definidas positivas, o que garante direções de descida que geralmente não exigem correções. Além disso, sabe-se que, quando aplicada à minimização de uma função quadrática com hessiana definida positiva, e considerando que o passo t seja determinado como o minimizador da função ao longo da direção $\boldsymbol{d}_k$, a convergência para o minimizador da quadrática é atingida em, no máximo, n iterações (Martínez, 2020).

4.3.3.1 Método L-BFGS-B

O algoritmo L-BFGS-B é um método quase-Newton de memória limitada, baseado em gradiente, projetado para resolver problemas de otimização com restrições de caixa (Byrd, 1995). O "L"refere-se a *limited memory*, pois o algoritmo não precisa armazenar toda a matriz hessiana, e sim apenas um número limitado de atualizações das direções de descida. Já o "B"refere-se às *box constraints* (restrições de caixa).

A aproximação da matriz hessiana por uma matriz de memória limitada permite reduzir o custo computacional, especialmente em problemas de grande escala. Essa aproximação é usada para construir um modelo quadrático da função objetivo, o que facilita a determinação da direção de busca. Para calcular essa direção, o algoritmo aplica o método do gradiente projetado. Nesse processo, são identificadas as variáveis ativas. Em seguida, o modelo quadrático é minimizado de forma aproximada com relação às variáveis livres, ou seja, aquelas que não estão sujeitas a restrições. Esse procedimento permite que a direção de busca seja eficiente, conduzindo o algoritmo em direção ao mínimo da função objetivo, respeitando as restrições de caixa. (Zhu, 1997)

O método do gradiente projetado é amplamente utilizado na programação não-linear para resolver problemas com restrições. O processo começa com o cálculo do gradiente da função objetivo em um ponto inicial, seguido pela determinação da direção de maior diminuição da função, que é obtida ao se mover na direção oposta ao gradiente. O tamanho do passo é então ajustado por meio de uma busca linear. Após cada atualização da posição, a solução é projetada de volta ao conjunto viável, garantindo que o novo ponto esteja dentro das restrições impostas

pelo problema. Esse processo é repetido iterativamente até que um critério de convergência seja alcançado. (Luenberger & Ye, 2016)

O algoritmo L-BFGS-B simplificado encontra-se descrito a seguir. Defina um ponto inicial $\boldsymbol{x}_0$, uma tolerância ϵ para a convergência e o número de iterações de memória limitada m. Calcule o gradiente inicial $\nabla f(\boldsymbol{x}_0)$ e (Zhu, 1997):

(i.) *(Direção de busca)* A cada iteração k, calcule $\boldsymbol{d}_k$ resolvendo o sistema $\boldsymbol{B}_k\boldsymbol{d}_k = -\nabla f(\boldsymbol{x_k})$. Utilize o método de projeção de gradiente para identificar variáveis que devem ser mantidas em seus limites. Resolva uma minimização de linha em relação às variáveis livres, com base no modelo quadrático definido pelo gradiente e pela direção de busca $\boldsymbol{d}_k$;

(ii.) *(Busca linear)* Para encontrar o passo t_k, realize uma busca linear ao longo de $\boldsymbol{d}_k$ utilizando um método de busca de linha;

(iii.) *(Atualização)* Atualize o ponto de iteração $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + t_k\boldsymbol{d}_k$;

(iv.) *(Limite da memória)* Atualize a matriz $\boldsymbol{B}_k$ com base nos pares de vetores armazenados $(\boldsymbol{s}_k, \boldsymbol{y}_k)$, mantendo o número de vetores guardados em memória limitado a m;

(v.) *(Verificação da convergência)* Verifique se $||\nabla f(\boldsymbol{x}_{k+1})|| \leqslant \epsilon$. Se sim, termine; caso contrário, volte a (i.).

Os métodos de busca de linha, mencionado no algoritmo, buscam determinar um ponto mínimo em uma única variável, tratando o problema temporariamente como unidimensional. Muitos algoritmos de programação não-linear procedem da seguinte forma: dado um ponto $\boldsymbol{x}_k$, encontra-se um vetor direcional $\boldsymbol{d}_k$ e, em seguida, um tamanho de passo adequado t_k, gerando um novo ponto $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + t_k\boldsymbol{d}_k$; e o processo é então repetido. Encontrar o tamanho de passo t_k envolve resolver o subproblema de minimizar $f(\boldsymbol{x}_k + t\boldsymbol{d}_k)$, que é unidimensional na variável t. Mais sobre esses métodos pode ser encontrado em Bazaraa, Sherali & Shetty (2006) e Luenberger & Ye (2016).

5 IMPLICAÇÕES NA GESTÃO ESCOLAR

O estudo de caso apresentado no Capítulo 3 indicou que o dia da semana exerce influência sobre a frequência estudantil, tornando-se um parâmetro relevante à sua tendência. Os dados analisados apontam que determinados dias apresentam maior média de frequência, enquanto outros registram um número significativamente menor de estudantes em sala de aula. Apesar desse padrão poder estar associado a diversos fatores, os resultados mostram que a distribuição das disciplinas ao longo da semana exerce influência particular. Compreender essa variação é essencial para a gestão escolar, permitindo a formulação de estratégias que incentivem a presença e minimizem os impactos negativos que as faltas causam no processo ensino-aprendizagem.

O Capítulo 5 apresenta ferramentas para auxiliar a gestão escolar a lidar com os problemas identificados ao longo do estudo. Primeiramente, discute uma modelagem do quadro de horários, buscando equilibrar a distribuição das disciplinas com base em seus impactos na frequência estudantil. Em seguida, explora um modelo de previsão de frequência, fundamental para o planejamento financeiro e a alocação de recursos da instituição. Finalmente, são abordadas outras análises e modelagens que podem contribuir para uma gestão mais eficiente e embasada em dados.

5.1 PROBLEMA DO QUADRO DE HORÁRIOS

O estudo de Santos (2024) propõe o primeiro modelo de organização do quadro de horários fundamentado na análise estatística da frequência estudantil apresentada no Capítulo 3. Esse modelo busca otimizar a distribuição das disciplinas ao longo da semana, levando em consideração o impacto que cada matéria tem na frequência escolar. Para isso, introduz-se o conceito de **peso da disciplina**, um indicador calculado a partir da média de frequência de cada matéria em relação à soma das médias de todas as disciplinas para um determinado ano escolar.

A determinação do peso disciplinar passa por três etapas:

1ª. *(Média da frequência da disciplina na turma)* Como uma mesma disciplina pode ser lecionada em dias diferentes, a média de frequência da mesma em uma determinada turma é dada por:

$$A_{mt} = \frac{\sum_{i=1}^{I} a_{it}}{I}, \tag{5.1}$$

onde I é o número de dias da semana em que a disciplina é lecionada, $a_i t$ é a média aritmética da frequência do dia i na turma t e A_{mt} é a média da frequência da disciplina m na turma t;

2ª. *(Peso da disciplina na turma)* O peso da disciplina m na turma t é dado então por:

$$P_{mt} = \frac{A_{mt}}{\sum_{j=1}^{M} A_{jt}}, \tag{5.2}$$

onde P_{mt} é o peso da disciplina m na turma t, M é a quantidade de disciplinas da turma, e A_{jt} é a média da frequência da disciplina j na turma t. O denominador corresponde à soma da média de todas as matérias em uma determinada turma;

3ª. *(Peso final da disciplina)* O peso final de uma disciplina é dado por:

$$PF_{my} = \frac{\sum_{t=1}^{T} P_{mt}}{T}, \tag{5.3}$$

onde PF_{my} é o peso final da disciplina m no ano y e T é o número de turmas no respectivo ano letivo.

Observe que, devido à forma como os pesos disciplinares são calculados, a soma dos pesos para o conjunto de disciplinas de um determinado ano escolar deve ser igual a 1. Essa característica simplifica a comparação entre os pesos, permitindo a identificação de disciplinas que apresentam menores índices de frequência, indicando um possível desinteresse ou negligência por parte dos alunos. Da mesma forma, torna-se possível reconhecer quais disciplinas têm maior capacidade de engajamento.

Os pesos calculados para as disciplinas ofertadas no CEEAL para o Ensino Fundamental, Ensino Médio e NEJA encontram-se detalhados nas Tabelas 5.1, 5.2 e 5.3, respectivamente. As matérias cujos pesos foram deixados em branco não são ofertadas ao determinado ano escolar.

DISCIPLINAS	**6º ANO**	**7º ANO**	**8º ANO**	**9º ANO**
Geografia	0,0989	0,0998	0,1001	0,0997
Matemática	0,1015	0,1025	0,0984	0,1023
Português	0,1027	0,099	0,099	0,1009
História	0,1013	0,0993	0,0997	0,1022
Educação Física	0,0988	0,1024	0,1017	0,1022
Inglês	0,1002	0,0986	0,0972	0,0984
Ciências	0,0995	0,098	0,1005	0,1001
Letramento de Matemática	0,1010	0,1023	0,0998	
Letramento de Português	0,0976	0,1037	0,1020	
Artes	0,0985	0,0943	0,1015	0,0960
Resolução de Problemas Matemáticos				0,0982
Produção Textual				0,1000
SOMA	1	1	1	1

Tabela 5.1 – Pesos para as disciplinas do Ensino Fundamental do ano 2022 do CEEAL. Fonte: Santos, 2024.

DISCIPLINAS	1º ANO	2º ANO	3º ANO
Geografia	0,0930	0,0883	0,0933
Matemática	0,0944	0,0852	0,0939
Português	0,0931	0,0858	0,0973
História	0,0949	0,0884	0,0703
Filosofia	0,0930	0,0844	0,0930
Projeto de Vida	0,0844		
Química	0,0915	0,0868	0,0879
Biologia	0,0948	0,0664	0,0939
Educação Física	0,0920	0,0851	0,0933
Inglês	0,0760		
Física	0,0928	0,0883	0,0925
Artes		0,0873	
Língua Estrangeira		0,0689	0,0960
Sociologia		0,0851	0,0887
SOMA	1	1	1

Tabela 5.2 – Pesos para as disciplinas do Ensino Médio do ano 2022 do CEEAL. Fonte: Santos, 2024.

DISCIPLINAS	NEJA I	NEJA II	NEJA III	NEJA IV
Geografia	0,1359		0,1571	
Matemática	0,1647	0,1441	0,1492	0,1630
Português	0,1621	0,1508	0,1523	0,1385
História	0,1524		0,1372	
Filosofia	0,1312		0,1406	
Projeto de Vida		0,1514		
Química		0,1322		0,1529
Biologia		0,1306		0,1342
Educação Física	0,1312		0,1396	
Inglês	0,1224			0,1428
Física		0,1515		0,1532
Artes		0,1394		0,1152
Sociologia			0,1240	
SOMA	1	1	1	1

Tabela 5.3 – Pesos para as disciplinas do NEJA do ano 2022 do CEEAL. Fonte: Santos, 2024.

O modelo de quadro de horários baseou-se em um modelo de múltiplas mochilas, visto no Capítulo 4. Considerando as disciplinas de um determinado ano escolar, cada disciplina m, $m = 1,\ldots,M$, necessita de T tempos de aula que devem ser alocados ao longo de D dias da semana. Cada dia d tem uma capacidade de seis tempos e cada disciplina m tem um peso p_m, como calculado. O objetivo do modelo é que cada dia tenha o máximo de pesos possíveis. (Santos, 2024)

$$\text{maximizar } z \tag{5.4}$$

$$\text{sujeito a } \sum_{m=1}^{M}\sum_{t=1}^{T} p_m x_{dmt} \geqslant z,\ d = 1,2,3,4,5 \tag{5.5}$$

$$\sum_{m=1}^{M}\sum_{t=1}^{T} x_{dmt} \leqslant 6,\ d = 1,2,3,4,5 \tag{5.6}$$

$$\sum_{d=1}^{D}\sum_{t=1}^{T} x_{dmt} = C_m,\ m = 1,\ldots,M \tag{5.7}$$

$$x_{dm1} \leqslant x_{dm2},\ m = 1,\ldots,M,\ d = 1,\ldots,5 \tag{5.8}$$

$$x_{dm3} \leqslant x_{dm_4},\ m = 1,\ldots,M,\ d = 1,\ldots,5 \tag{5.9}$$

$$x_{dm3} \leqslant x_{dm_5},\ m = 1,\ldots,M,\ d = 1,\ldots,5 \tag{5.10}$$

$$\sum_{d=1}^{D} x_{dmt} = 1,\ m = 1,\ldots,M,\ t = 1,\ldots,T \tag{5.11}$$

$$\sum_{t=1}^{T} x_{dmt} \leqslant 4,\ m = 1,\ldots,M,\ d = 1,\ldots,5 \tag{5.12}$$

$$x \in B^{dmt} \tag{5.13}$$

Para esse modelo, foi utilizada uma variável artificial z que representa a função objetivo 5.4, e as variáveis x_{dmt} são tais que:

$$x_{dmt} = \begin{cases} 1, \text{ se o tempo } t \text{ da matéria } m \text{ é alocado no dia } d, \\ 0, \text{ caso contrário.} \end{cases} \tag{5.14}$$

A função objetivo 5.4 e as restrições 5.5 equilibram a quantidade de pesos entre os dias. As restrições 5.6 impedem o modelo de alocar mais disciplinas do que comportam os seis tempos diários de aula. As restrições 5.7 alocam C_m tempos de aula da matéria m na semana, já que as disciplinas possuem diferentes quantidades de tempo de aula obrigatórias. As restrições 5.8 e 5.9 alocam sempre dois tempos de aula seguidos. As restrições 5.10 alocam três tempos seguidos caso a matéria necessite de cinco tempos semanais. A restrições 5.11 garantem as três restrições anteriores. As restrições 5.12 impedem a alocação de mais de quatro tempos de aula de uma mesma disciplina em um mesmo dia. Por fim, a restrição 5.13 refere-se a x pertencer ao espaço binário de cardinalidade dmt. (Santos, 2024)

A implementação do modelo utilizou-se do solver GLPK (GNU Linear Programming Kit) (2024), *software* de código aberto, em Python (Python Software Foundation, 2024, online). A principal metodologia do solver para a resolução de problemas inteiros é o *branch-and-cut*, visto no Capítulo 4 na seção 4.2.4.

Esse modelo distribui as disciplinas ao longo da semana de tal forma que a soma dos pesos disciplinares de cada dia seja sempre o maior possível. Como os pesos foram calculados com base nas médias de frequência, idealmente o quadro de horários gerado ocasionará em

médias de frequência mais equilibradas, amenizando possíveis desvantagens de um dia específico. Esses resultados podem ser observados na Tabela 5.1 abaixo, que contém um exemplo de organização disciplinar sintetizada como resposta do modelo, a soma dos pesos das disciplinas alocadas no dia e a comparação com a soma dos pesos das disciplinas no quadro de horários original.

	SEGUNDA	**TERÇA**	**QUARTA**	**QUINTA**	**SEXTA**
	Matemática	Geografia	Projeto de Vida	Matemática	Português
	Matemática	Geografia	Projeto de Vida	Matemática	Português
	Educação Física	Português	História	Química	Física
	Educação Física	Português	História	Química	Física
		Filosofia	Inglês	Biologia	
		Filosofia	Inglês	Biologia	
$\sum p_m$ **(MODELO)**	18,64	27,91	25,53	28,07	18,59
$\sum p_m$ **(ORIGINAL)**	26,18	28,05	27,79	28,28	8,44

Tabela 5.4 – Sugestão de quadro de horários fornecido pelo modelo com somatório de pesos do modelo e do quadro original. Fonte: Santos, 2024.

Um dos desafios do modelo, conforme apontado por Santos (2024), é a ausência da consideração da disponibilidade dos professores, que, à época do estudo, era o principal critério para a organização do quadro de horários no CEEAL. Nesse formato, o modelo funciona como uma ferramenta de apoio à tomada de decisão, oferecendo sugestões iniciais que podem ser ajustadas pelo gestor conforme as necessidades e restrições da escola.

5.2 PROBLEMA DA FREQUÊNCIA ESPERADA

Uma das maiores dificuldades na gestão escolar é o manejo adequado dos recursos financeiros. A oferta de merenda, por exemplo, frequentemente resulta em desperdício de alimentos, enquanto os recursos poderiam ser melhor alocados para oferecer refeições mais completas, com receitas mais elaboradas e atraentes. Nesse contexto, um modelo de previsão de frequência escolar poderia ser uma ferramenta valiosa, permitindo antecipar padrões de frequência com base em dados históricos e variáveis contextuais. Isso possibilitaria aos gestores adotar estratégias mais eficazes para manutenção das instalações, organização de cardápios, gestão da equipe docente, além de estimar o consumo de água e energia elétrica.

No entanto, a modelagem da frequência escolar enfrenta desafios devido às suas variações ao longo do ano, influenciadas por fatores como sazonalidade, eventos escolares e características específicas das turmas. Uma tentativa inicial de descrever esse comportamento foi o ajuste dos dados a uma curva polinomial, mas mesmo com polinômios de grau elevado, os resultados não foram satisfatórios, pois o modelo não capturava adequadamente os padrões e oscilações. Esse comportamento resulta em uma alta variabilidade dos dados, tornando a previsão pouco confiável.

Para lidar com esse problema, foi proposto um modelo baseado na minimização das distâncias quadráticas ponderadas entre os dados, o que contribui para a regularização da variância. Ele objetiva, ao reduzir oscilações extremas e suavizar a curva de frequência, fornecer uma estimativa mais estável da frequência estudantil.

O modelo foi baseado nas formulações de quantização por mínimos quadrados propostas por Lloyd (1982), que estabelecem a distribuição otimizada de pesos para a função, atribuindo valores diferenciados conforme a região em que as variáveis estão localizadas. A intenção é gerar uma frequência esperada baseada nos dados de frequências reais, minimizando a distância entre os dados ao penalizar valores muito abaixo do **teto de frequência**, o valor mais alto observado dentro de um período considerado. O modelo é da forma:

$$\text{minimizar} \quad \sum_{i=1}^{n}\sum_{j=1}^{n} w_j(y_i - c_j)^2 \tag{5.15}$$

$$\text{sujeito a} \quad y_i \geqslant c_j, \qquad i = 1,\ldots,n,\ j = 1,\ldots,n \tag{5.16}$$

$$y_i \leqslant t, \qquad i = 1,\ldots,n \tag{5.17}$$

onde o índice i representa a quantidade de observações de frequências reais, c_j os valores de frequência diários, y_i as variáveis de decisão que determinam a frequência esperada, w_j os pesos da função, e t o teto de frequência. Naturalmente, os valores de frequência observados são não-negativos e, portanto, está implícito que $y_i \geqslant 0$.

Os pesos são dados por:

$$w_j = \frac{t - c_j}{t}, \tag{5.18}$$

de forma que, quanto mais próximo do teto for o valor observado, menor será o peso e, consequentemente, menor será o valor da função objetivo, gerando valores de y_i com pouca ou nenhuma correção. Por outro lado, quanto mais distantes os valores reais estiverem de t, maior o peso e maior será a punição da função objetivo, que priorizará a correção do y_i associado.

Dessa forma, o modelo gera um vetor de frequência esperada que nunca será inferior à frequência real no período observado, ao mesmo tempo em que proporciona previsões mais adequadas, contribuindo para a redução de desperdícios.

O modelo foi implementado em Python (2024), com suporte das bibliotecas *numpy* (Harris *et al.*, 2020) e *scipy* (2020). O código pode ser encontrado no Apêndice 8, seção 8.4, junto com os dados de frequência utilizados. Para a otimização, foi utilizado o método L-BFGS-B, conforme discutido no Capítulo 4, devido à natureza das restrições do tipo caixa. Esse método apresentou rápida convergência, atingindo os resultados em média após três iterações.

Para a geração das frequências esperadas, os dados de frequência foram segmentados por bimestre e turno, com o objetivo de isolar possíveis efeitos sazonais. A comparação dos valores reais de frequência com os valores gerados pelo modelo encontram-se no Apêndice 8, seção 8.6, em gráficos de linha. É possível perceber o comportamento observado: a frequência esperada corrige valores muito abaixo do teto e tende a coincidir com os valores reais caso estes estejam próximos do teto. Isso estabiliza o comportamento da frequência.

A Tabela 8.7 do Apêndice 8, apresenta uma comparação entre o total de alunos presentes nas observações reais e o total de alunos esperados, considerando diferentes expectativas. Para isso, os dados de frequência foram somados para cada bimestre e turno, e a diferença entre os valores reais e os valores esperados foi calculada. Como referência para a comparação com o desperdício gerado pelo somatório dos dados fornecidos pelo modelo, foram utilizadas algumas medidas de tendência central, como a mediana e a moda, e o teto. As diferenças evidenciam quantos alunos foram previstos a mais ou a menos no período considerado. Por isso, a média não foi analisada, visto que a comparação daria sempre zero.

É fácil notar que em nove dos dez casos apresentados na tabela os valores dados pelo

modelo geraram o menor desperdício ao prever uma quantidade de alunos presentes mais próxima do real.

Uma observação importante é que, apesar da média sugerir que não causa desperdício pelo somatório das frequências esperadas, ela gera um problema logístico em comum com os valores negativos das diferenças na tabela. Esperar menos alunos do que a frequência real é pior para a gestão do que gerar um desperdício. Ela pode culminar em, por exemplo, o refeitório não poder fornecer merenda a todos os estudantes e não ter tempo hábil para corrigir o erro pois os alimentos não foram devidamente preparados, descongelados ou comprados.

Esse é o principal problema da utilização de medidas centrais como frequência esperada – comumente elas se encontram muito abaixo da frequência real. Por outro lado, esperar sempre o máximo de alunos possível gera muito desperdício. O modelo proposto oferece uma alternativa baseada em dados que, por suas próprias restrições, nunca irão prever menos alunos do que as quantidades reais de ocorrências passadas.

No entanto, o modelo apresenta algumas limitações que precisam ser consideradas. Um dos principais desafios é que ele se baseia unicamente em dados históricos, o que pode resultar em previsões imprecisas quando ocorrem variações significativas nos padrões de frequência. Fatores imprevistos, como atividades escolares ou feriados, mudanças no comportamento dos alunos ou até mesmo eventos externos, como condições climáticas adversas, podem alterar significativamente a frequência real, comprometendo a precisão das previsões. Além disso, o modelo não leva em consideração a flexibilidade necessária para adaptar-se a tais mudanças contextuais, o que limita sua eficácia em cenários dinâmicos.

Para melhorar a precisão e a aplicabilidade do modelo, seria fundamental incorporar variáveis externas que influenciem a frequência escolar, como a própria organização do quadro de horários e eventos especiais. Além disso, uma melhoria significativa seria a implementação de ajustes em tempo real, permitindo que as previsões sejam continuamente atualizadas com base em dados mais recentes. Isso poderia ajudar a minimizar a diferença entre os valores previstos e os reais, promovendo uma gestão mais eficiente e evitando desperdício de recursos, como observado anteriormente.

5.3 TRABALHOS FUTUROS

Durante a realização deste estudo, ficou evidente que as metodologias empregadas também poderiam ser aplicadas à questão da merenda escolar, especialmente no que diz respeito à composição do cardápio. Em particular, a análise de variância poderia ser utilizada para investigar se a frequência dos alunos apresenta alguma tendência comportamental influenciada pelo tipo de refeição oferecida. Essa hipótese surgiu a partir de relatos de professores do CEEAL, que observaram o interesse dos alunos em determinadas refeições, enquanto outras pareciam menos atraentes.

A merenda escolar, obrigatória nas escolas públicas, é de responsabilidade da gestão escolar, embora a Secretaria de Estado de Educação do Rio de Janeiro (SEEDUC-RJ, 2024) forneça cardápios sugeridos. No entanto, conforme relatado pelo gestor responsável do CEEAL, nem sempre é possível seguir essas recomendações devido a desafios logísticos, como a dificuldade de descongelar determinados alimentos a tempo para as refeições servidas pela manhã da segunda-feira.

Como não havia um registro sistemático das refeições realmente servidas, a aplicação do modelo de análise de variância tornou-se inviável neste estudo de caso. No entanto, é possível investigar os efeitos da merenda sobre a frequência escolar ao agrupar os dados com base em determinados tipos de refeições. Outra abordagem interessante seria analisar se os alunos

demonstram alguma tendência em relação ao tipo de proteína oferecida.

A identificação de uma tendência na frequência escolar associada a determinados alimentos permitiria a atribuição de pesos a esses itens, de maneira semelhante à atribuição de pesos disciplinares no problema do quadro de horários. Esses pesos representariam a influência de cada alimento na presença dos alunos, permitindo ajustes estratégicos na composição dos cardápios. Com isso, um modelo de geração de cardápios poderia utilizar esses pesos para equilibrar as refeições ao longo da semana, buscando otimizar a frequência sem comprometer os requisitos nutricionais.

O problema de geração de cardápios consiste em planejar quais alimentos serão servidos em cada refeição ao longo de um período determinado (por exemplo, uma semana ou um mês), garantindo uma distribuição equilibrada e variada. Esse planejamento deve levar em conta não apenas a escolha dos alimentos, mas também a quantidade adequada de cada item, respeitando critérios nutricionais, orçamentários e logísticos. Além disso, é necessário evitar a repetição de cardápios em dias consecutivos, garantindo diversidade na alimentação e maior aceitação por parte dos estudantes. Esse tipo de problema pode ser abordado por meio de métodos de otimização, que buscam encontrar combinações ideais de refeições considerando restrições e objetivos específicos.

5.3.1 Modelo de geração de cardápio

Um modelo de geração de cardápio deve levar em conta que alguns alimentos podem ser fracionados, permitindo que suas quantidades sejam representadas por números reais, enquanto outros só podem ser servidos em unidades inteiras. Para formalizar essa distinção, utilizamos os seguintes índices:

- $\overline{i}$: identifica o grupo alimentar ao qual o alimento pertence, $\overline{i} = 1, \ldots, \overline{N}$;
- $\overline{j}_{\overline{i}}$: representa os alimentos do grupo $\overline{i}$ que devem podem ser fracionados;
- $\hat{j}_{\overline{i}}$: representa os alimentos do grupo $\overline{i}$ que devem ser servidos em quantidades inteiras;
- k: indica o dia da semana em que o alimento será servido, $k = 1, \ldots, P$.

Considere que o $\overline{i}$-ésimo grupo alimentar contenha $\overline{M}_{\overline{i}}$ alimentos que podem ser fracionados e $\hat{M}_{\overline{i}}$ alimentos que devem ser servidos inteiros. As variáveis do problema serão identificadas por $y_{\overline{j}_{\overline{i}}k}$ e $y_{\hat{j}_{\overline{i}}k}$ e correspondem às quantidades de porções dos alimentos $\overline{j}_{\overline{i}}$ e $\hat{j}_{\overline{i}}$, respectivamente, do $\overline{i}$-ésimo grupo a ser servido no k-ésimo dia.

Se o objetivo do modelo for minimizar custos, sejam $\overline{c}_{\overline{j}_{\overline{i}}}$ e $\overline{c}_{\hat{j}_{\overline{i}}}$ os custos relacionados às porções de alimentos fracionários e inteiros, respectivamente. A função objetivo toma a forma:

$$\sum_{\overline{i}=1}^{\overline{N}} \sum_{\overline{j}_{\overline{i}}=1}^{\overline{M}_{\overline{i}}} \sum_{k=1}^{P} \overline{c}_{\overline{j}_{\overline{i}}} y_{\overline{j}_{\overline{i}}k} + \sum_{\overline{i}=1}^{\overline{N}} \sum_{\hat{j}_{\overline{i}}=1}^{\hat{M}_{\overline{i}}} \sum_{k=1}^{P} \overline{c}_{\hat{j}_{\overline{i}}} y_{\hat{j}_{\overline{i}}k}. \tag{5.19}$$

Como o cardápio escolar deve assegurar níveis nutricionais adequados, as variáveis devem estar sujeitas a uma série de restrições que garantam esse equilíbrio. Além disso, devem ser estabelecidos limites mínimo e máximo para as porções de cada grupo alimentar, assegurando variedade em cada refeição. Para evitar repetições excessivas, certas categorias de alimentos,

como carnes e saladas, podem ter suas porções limitadas dentro do período P, garantindo a diversificação dos cardápios ao longo dos dias. Esse é modelo é amplamente discutido no trabalho de Brasil (2023).

Uma possível adaptação desse modelo, caso se verifique variação nos pesos dos alimentos em relação à frequência, seria reformular o problema para a maximização dos pesos calculados estatisticamente. Para garantir que os custos sejam adequadamente considerados, pode-se incorporar uma restrição orçamentária, levando em conta o valor da verba V repassada pelo Programa Nacional de Alimentação Escolar (PNAE, 2024):

$$\sum_{\bar{i}=1}^{\bar{N}} \sum_{\bar{j}_{\bar{i}}=1}^{\bar{M}_{\bar{i}}} \sum_{k=1}^{P} \bar{c}_{\bar{j}_{\bar{i}}} y_{\bar{j}_{\bar{i}}k} + \sum_{\hat{i}=1}^{\hat{N}} \sum_{\hat{j}_{\hat{i}}=1}^{\hat{M}_{\hat{i}}} \sum_{k=1}^{P} \bar{c}_{\hat{j}_{\hat{i}}} y_{\hat{j}_{\hat{i}}k} \leqslant V. \tag{5.20}$$

Essa é apenas uma sugestão de como abordar o problema. Se forem considerados pesos de diferentes tipos de refeições ofertadas ao longo de um mesmo turno, como café da manhã, lanche e almoço, e tratar as variáveis como refeições já preparadas, é possível também elaborar o problema como um modelo de múltiplas mochilas que objetiva a maximização do somatório dos pesos ao longo do dia.

6

CONCLUSÕES

Este estudo teve como objetivo explorar a relação entre a frequência estudantil e o quadro de horários, além de propor ações baseadas em dados para otimizar a gestão dos recursos escolares. Para alcançar esses objetivos, a pesquisa foi conduzida em duas etapas principais: primeiro, a análise de dados por meio de modelos estatísticos, e, em seguida, a aplicação dos resultados obtidos em modelos de otimização.

Por meio do estudo de caso e da aplicação do modelo de análise de variância, identificou-se uma forte relação entre a frequência estudantil e o dia da semana. As sextas-feiras se destacaram negativamente, registrando, de maneira geral, as menores médias de frequência. Entre as 42 turmas analisadas, 20 apresentaram diferenças estatisticamente significativas entre as médias de frequência da sexta-feira e de pelo menos outro dia da semana, conforme indicado pelo teste de Tukey. Esse resultado acentuadamente negativo pode estar associado a fatores externos, alheios ao controle da gestão escolar. No entanto, também foram identificados fatores internos que podem ter contribuído para essa queda na frequência, como a alocação predominante de tempos vagos e disciplinas eletivas às sextas-feiras.

Por outro lado, as terças-feiras apresentaram, em sua maioria, as maiores médias de frequência. Em 18 turmas, as médias das terças-feiras se destacaram significativamente em relação a pelo menos um outro dia da semana, conforme indicado pelo teste de Tukey. À primeira vista, esse resultado não parecia estar associado a características específicas do dia, já que as terças-feiras seguiam uma estrutura regular, sem um número excessivo de tempos vagos, disciplinas sem professores ou eletivas – características semelhantes às dos demais dias da semana, com exceção das sextas-feiras.

A análise mais aprofundada da relação entre frequência e dia da semana concentrou-se, por fim, nas disciplinas ministradas em cada dia. Entre as turmas que apresentaram altas médias de frequência às terças-feiras, 8 tinham aulas de Matemática, 7 de Língua Portuguesa e 5 de Educação Física, totalizando 11 disciplinas identificadas. Observou-se uma forte correlação entre a presença dos alunos e essas matérias. Por outro lado, as baixas taxas de frequência às sextas-feiras abrangeram um conjunto mais diversificado de disciplinas. Das 14 matérias lecionadas nesse dia, 7 turmas tinham aulas de Língua Portuguesa, 6 de Língua Inglesa, 6 de Arte e 6 de Ciências.

A identificação da correlação entre frequência estudantil, dia da semana e distribuição das disciplinas é a principal contribuição desta pesquisa para a literatura sobre fatores que influenciam a frequência escolar. Como discutido, assegurar a matrícula, a permanência do aluno na escola ou o cumprimento de uma frequência mínima não é suficiente se o ambiente educacional não for capaz de despertar e manter o interesse dos estudantes. Para além de simplesmente mo-

nitorar a assiduidade, a gestão escolar deve atuar estrategicamente para promover um tempo de aprendizado equilibrado e produtivo entre as disciplinas, evitando que algumas matérias sejam implicitamente percebidas como mais relevantes do que outras.

O *software* desenvolvido para a análise de variância e visualização dos dados de frequência foi concebido como uma ferramenta estratégica para o monitoramento da frequência estudantil. Com ele, os gestores podem identificar padrões e problemas, permitindo intervenções rápidas e eficazes. Sua aplicação se torna ainda mais simples e eficiente se a escola for integrada a um diário eletrônico de frequências, que já sistematiza os registros diários, automatizando a coleta e o processamento das informações.

Os resultados obtidos sobre a influência das disciplinas na frequência escolar motivaram a formulação de um modelo de otimização para a organização do quadro de horários, baseado no problema de múltiplas mochilas. O objetivo do modelo é maximizar o somatório dos pesos diários das disciplinas distribuídas ao longo da semana, levando em conta a relação entre a frequência dos alunos e as matérias lecionadas. O modelo considera um ponderamento das disciplinas, refletindo seu impacto na assiduidade dos estudantes. No entanto, uma limitação importante é que ele desconsidera um fator essencial na organização do quadro de horários no CEEAL: a disponibilidade dos professores, que frequentemente orienta a distribuição das aulas.

Além disso, foi desenvolvido o problema da frequência esperada, um modelo baseado em mínimos quadrados, cujo objetivo é minimizar a variabilidade entre os dados, gerando um vetor de frequência esperada com comportamento mais estável para previsões. Os resultados oferecem uma alternativa mais precisa para a estimativa de presença de alunos, evitando os inconvenientes causados pelas medidas centrais, que comumente se encontram abaixo dos valores reais, e do uso do teto de estudantes matriculados, que pode levar a desperdícios. No entanto, como modelo preditivo, ainda apresenta limitações, pois se baseia apenas nas observações de um único ano. Idealmente, sua formulação deveria considerar uma série histórica de dados para aumentar sua precisão.

Todas as ferramentas e modelos propostos podem e devem ser aprimorados para uma aplicação mais eficaz na gestão escolar. Um ponto fundamental sobre os estudos em frequência é que eles devem ser conduzidos de forma contínua e dinâmica, considerando fatores sazonais e eventos que possam impactar a assiduidade dos alunos. A análise da frequência do ano anterior se torna uma ferramenta valiosa quando comparada aos resultados acadêmicos e utilizada no acompanhamento das turmas à medida que avançam para novos anos e níveis de ensino. Esse monitoramento permite a identificação de padrões e a implementação de estratégias mais assertivas para a melhoria do desempenho e da permanência estudantil.

Trabalhos futuros podem explorar as possíveis implicações dos estudos de frequência na elaboração de um modelo de geração de cardápios para a merenda escolar. Para isso, seria necessário definir critérios para agrupar as frequências de acordo com os alimentos servidos e aplicar a análise de variância para identificar se há diferenças estatisticamente significativas no comportamento dos alunos em função das refeições oferecidas. Caso essa relação seja comprovada, os pesos atribuídos às refeições poderiam ser incorporados ao modelo de geração de cardápios, permitindo um planejamento mais estratégico da alimentação escolar.

Em resumo, este estudo destaca que a frequência escolar também é influenciada por fatores institucionais que podem ser otimizados. Ao adotar estratégias baseadas em análise de dados, os gestores educacionais podem criar ambientes mais propícios ao aprendizado, promovendo melhores resultados para os estudantes.

7

REFERÊNCIAS BIBLIOGRÁFICAS

1 ARENALES, M. *et al*. *Pesquisa Operacional*. 1 ed. Rio de Janeiro: Elsevier, 2011. 524 p.

2 ANSARI, A.; PIANTA, R. C. School absenteeism in the first decade of education and outcomes in adolescence. *Journal of School Psychology*, v. 76, p. 48-61, 2019. Disponível em: <https://doi.org/10.1016/j.jsp.2019.07.010>. Acesso em 18 nov. 2024.

3 BANERJI, M.; MATHUR, K. Understanding school attendance: The missing link in "Schooling for All". *International Journal of Educational Development*, v. 87, 2021. Disponível em: <https://doi.org/10.1016/j.ijedudev.2021.102481>. Acesso em 18 nov. 2024.

4 BAZARAA, M. S.; SHERALI, H. D.; SHETTY, C. M. *Nonlinear programming: theory and algorithms*. 3rd ed. USA: John Wiley & Sons, Inc, 2006. 880 p.

5 BRASIL, A. M. M. *Modelos de otimização combinatória aplicados à gestão de demandas operacionais em centros universitários*. Dissertação: Programa de Pós-graduação em Modelagem Matemática e Computacional, Universidade Federal Rural do Rio de Janeiro. Rio de Janeiro. 2023. 57 p.

6 BRASIL. Conselho Nacional de Educação. *Nota de esclarecimento: retorno das atividades escolares e acadêmicas presenciais em 2022*. Brasília, DF, 27 jan. 2022. Disponível em: <http://portal.mec.gov.br/index.php?option=com_docman&view=download&alias=232651-nota-de-esclarecimento-covid-19-2022&category_slug=dezembro-2021-pdf&Itemid=30192>. Acesso em 23 jan. 2025.

7 BRASIL. Constituição (1988). *Constituição da República Federativa do Brasil*. Brasília, DF, 1988.

8 BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. *Nota técnica: índice de desenvolvimento da educação básica-Ideb*. Brasília, 2021.

9 BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). *Resumo Técnico: Censo Escolar da Educação Básica 2021*. Brasília, 2021.

10 BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). *Saeb 2021: Indicador de Nível Socioeconômico do Saeb 2021: nota técnica.* Brasília, DF: Inep, 2023.

11 BRASIL. Lei nº 13.415, de 16 de fevereiro de 2017. Altera as Leis nº 9.394, de 20 de dezembro de 1996, que estabelece as diretrizes e bases da educação nacional, e nº 11.494, de 20 de junho de 2007, que regulamenta o Fundo de Manutenção e Desenvolvimento da Educação Básica e de Valorização dos Profissionais da Educação (FUNDEB), e dá outras providências. *Diário Oficial da União:* seção 1, Brasília, DF, 17 fev. 2017. Disponível em: <https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2017/lei/l13415.htm>. Acesso em 23 jan. 2025.

12 BRASIL. Lei nº 10.836, de 9 de janeiro de 2004. Cria o Programa Bolsa Família, altera a Lei nº 10.689, de 13 de junho de 2003, e dá outras providências. *Diário Oficial da União:* seção 1, Brasília, DF, 12 jan. 2004. Disponível em: <https://www2.camara.leg.br/legin/fed/lei/2004/lei-10836-9-janeiro-2004-490604-publicacaooriginal-1-pl.html>. Acesso em 12 mar. 2025.

13 BRASIL. Lei nº 14.601, de 19 de junho de 2023. Altera a Lei nº 10.836, de 9 de janeiro de 2004, que institui o Programa Bolsa Família, para dispor sobre condicionalidades na área de educação e medidas de acompanhamento das famílias beneficiárias. *Diário Oficial da União:* seção 1, Brasília, DF, 20 jun. 2023. Disponível em: <http://www.planalto.gov.br/ccivil_03/_Ato2023-2026/2023/Lei/L14601.htm>. Acesso em 23 jan. 2025.

14 BRASIL. Lei nº 11.947, de 16 de junho de 2009. Dispõe sobre o atendimento da alimentação escolar e do Programa Dinheiro Direto na Escola aos alunos da educação básica. *Diário Oficial da União:* seção 1, Brasília, DF, 17 jun. 2009. Disponível em: <https://www.planalto.gov.br/ccivil_03/_ato2007-2010/2009/lei/l11947.htm>. Acesso em 25 jan. 2025.

15 BRASIL. Lei nº 13.005, de 25 de junho de 2014. Aprova o Plano Nacional de Educação – PNE e dá outras providências. *Diário Oficial da União:* seção 1, p. 1, 26 jun. 2014. Disponível em: <https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2014/lei/l13005.htm>. Acesso em 25 jan. 2025.

16 BRASIL. Lei nº 10.219, de 11 de abril de 2001. Cria o Programa Nacional de Renda Mínima vinculada à Educação – "Bolsa Escola", e dá outras providências. *Diário Oficial da União:* seção 1, Brasília, DF, 12 abr. 2001. Disponível em: <https://www.planalto.gov.br/ccivil_03/leis/leis_2001/l10219.htm>. Acesso em 25 jan. 2025.

17 BRASIL. *Programa Nacional de Alimentação Escolar (PNAE)*. Disponível em: <https://www.gov.br/fnde/pt-br/acesso-a-informacao/acoes-e-programas/programas/pnae>. Acesso em 15 jan. 2025.

18 Breusch, T.S.; Pagan, A.R. A Simple Test for Heteroscedasticity and Random Coefficient Variation. *Econometrica*, v. 47, n. 05, p. 1287-1294, 1979. Disponível em: <https://doi.org/10.2307/1911963>. Acesso em 18 nov. 2024.

19 BUSSAB, W. O.; MORETTIN, P. A. *Estatística Básica*. 7 ed. São Paulo: Saraiva, 2011. 540 p.

20 BYRD, R. H.; *et al*. A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM Journal on Scientific Computing*, v. 16, n. 5, p. 1190-1208, 1995. Disponível em: <https://doi.org/10.1137/0916069>. Acesso em 25 jan. 2025.

21 CACCIAMALI, M.C.; TATEI, F.; BATISTA, N.F. Impactos do programa bolsa família federal sobre o trabalho infantil e a frequência escolar. *Revista de Economia Contemporânea*, v. 14, n. 2, p. 269-301, 2010. Disponível em: <https://doi.org/10.1590/S1415-98482010000200003>. Acesso em 22 nov. 2024.

22 CAVALCANTI, D.M.; COSTA, E.M.; SILVA, J.L.M. Programa bolsa família e o Nordeste: impactos na renda e na educação, nos anos de 2004 e 2006. *Revista de Economia Contemporânea*, v. 17, n. 1, p. 99-128, 2013. Disponível em: <https://doi.org/10.1590/S1415-98482013000100004>. Acesso em 22 nov. 2024.

23 CICUTO, C. A. T.; TORRES, B. B. Influência da frequência e participação no desempenho em um ambiente de aprendizagem centrado no aluno. *Química Nova*, v. 43, n. 2, p. 239-248, 2020. Disponível em: <http://dx.doi.org/10.21577/0100-4042.20170464>. Acesso em 18 nov. 2024.

24 GALHARDO, E. *et al*. Desempenho acadêmico e frequência dos estudantes ingressantes pelo Programa de Inclusão da UNESP. *Avaliação:* Revista da Avaliação da Educação Superior (Campinas), Sorocaba, v. 25, n. 03, p. 701-723, 2020. Disponível em: <http://dx.doi.org/10.1590/S1414-40772020000300010>. Acesso em 18 nov. 2024.

25 GARCIA, R. A.; RIOS-NETO, E. L. G.; MIRANDA-RIBEIRO, A. Efeitos rendimento escolar, infraestrutura e prática docente na qualidade do ensino médio no Brasil. *Revista brasileira de Estudos de População*, v. 38, p. 1-32, 2021. Disponível em: <https://doi.org/10.20947/S0102-3098a0152>. Acesso em 18 nov. 2024.

26 GNU PROJECT. *GNU Linear Programming Kit (GLPK), version 5.0.* Disponível em: <https://www.gnu.org/software/glpk/>. Acesso em 5 fev. 2025.

27 GOLDBARG, M. C.; LUNA, H. P. L. *Otimização combinatória e programação linear: modelos e algoritmos*. 2 ed. Rio de Janeiro: Elsevier, 2005. 519 p.

28 HARRIS, C. R. *et al*. Array programming with NumPy. *Nature*, v. 585, n. 7825, p. 357–362, 2020. Disponível em: https://doi.org/10.1038/s41586-020-2649-2. Acesso em 5 fev. 2025.

29 HILLIER, F. S.; LIEBERMAN, G. J. *Introdução à pesquisa operacional*. 9 ed. Porto Alegre: AMGH, 2013. 1005 p.

30 HUNTER, J. D. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, v. 9(3), p. 90-95, 2007. Disponível em: <https://doi.org/10.1109/MCSE.2007.55>. Acesso em 30 dez. 2024.

31 IZMAILOV, A.; SOLODOV, M. *Otimização*. 4 ed, v 1. Rio de Janeiro: IMPA, 2020. 256 p.

32 IZMAILOV, A.; SOLODOV, M. *Otimização - volume 2. Métodos Computacionais*. 3 ed. Rio de Janeiro: IMPA, 2018. 494 p.

33 LAVINAS, L.; BARBOSA, M.L.O. Combater a pobreza estimulando a freqüência escolar: o estudo de caso do Programa Bolsa-Escola do Recife. *Dados*, v. 43, n. 3, 2001. Disponível em: <https://doi.org/10.1590/S0011-52582000000300002>. Acesso em 18 nov. 2024.

34 LLOYD, S. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, v. 28, n. 2, p. 129-137, 1982. Disponível em: https://hal.science/hal-04614938v1. Acesso em 22 nov. 2024.

35 LUENBERGER, D. G.; YE, Y. *Linear and Nonlinear Programming*. 4th ed. Springer, 2016. 555 p.

36 MCKINNEY, W. *pandas: A fast, powerful, flexible, and easy-to-use open source data analysis and manipulation library.* 2010. Disponível em: <https://pandas.pydata.org>. Acesso em 30 dez. 2024.

37 MACAMBIRA, A. F. U.; *et al*. *Tópicos em Otimização Inteira*. 1 ed. Rio de Janeiro: UFRJ, 2022. 258 p.

38 MARTÍNEZ, J. M.; SANTOS, S. A. *Métodos Computacionais de Otimização*. Versão revisada. São Paulo: IMECC, UNICAMP, 2020. 228 p.

39 MELO, R.M.S.; DUARTE, G.B. Impacto do Programa Bolsa Família sobre a Frequência Escolar: o caso da agricultura familiar no Nordeste do Brasil. *Revista de Economia e Sociologia Rural*, v. 48, n. 3, p. 635-656, 2010. Disponível em: <https://doi.org/10.1590/S0103-20032010000300007>. Acesso em 22 nov. 2024.

40 MÍGUEZ, D. P. Factores asociados al desempeño entre estudiantes de bajo estatus sociocultural en Brasil, Chile y Argentina. *Revista Brasileira de Educação*, v. 28, 2023. Disponível em: <https://doi.org/10.1590/S1413-24782023280020>. Acesso em 18 nov. 2024.

41 MONTGOMERY, D. C. *Design and Analysis of Experiments*. 8th ed. USA: John Wiley & Sons, Inc, 2013. 724 p.

42 PIRES, A. Afinal, para que servem as condicionalidades em educação do Programa Bolsa Família? *Ensaio: Avaliação e Políticas Públicas em Educação*, v. 21, n. 80, p. 513-532, 2013. Disponível em: <https://doi.org/10.1590/S0104-40362013000300007>. Acesso em 22 nov. 2024.

43 PONTILI, R. M.; KASSOUF, A. L. Fatores que afetam a frequência e o atraso escolar nos meios urbano e rural, de São Paulo e Pernambuco. *Revista de Economia e Sociologia Rural*, v. 45, n. 1, p. 27-47, 2007. Disponível em: <https://doi.org/10.1590/S0103-20032007000100002>. Acesso em 22 nov. 2024.

44 PYTHON SOFTWARE FOUNDATION. *Python Programming Language*. Disponível em: <https://www.python.org/>. Acesso em 30 dez. 2024.

45 RIO DE JANEIRO. Secretaria de Estado de Educação. *Alimentação escolar*. Disponível em: <https://www.seeduc.rj.gov.br/cidad%C3%A3o/alimenta%C3%A7%C3%A3o-escolar>. Acesso em: 18 nov. 2024.

46 RIVERBANK COMPUTING. (2016). *PyQt5: Python bindings for Qt5*. 2016. Disponível em: <https://www.riverbankcomputing.com/software/pyqt/intro>. Acesso em 30 dez. 2024.

47 ROUSSAS, G. *An Introduction to Probability and Statistical Inference*. USA: Elsevier Science, 2003. 523 p.

48 SÁTYRO, N. G. D.; D'ALBUQUERQUE, R. W. O que é um Estudo de Caso e quais as suas potencialidades? *Sociedade e Cultura*, Goiânia, v. 23, 2020. DOI: 10.5216/sec.v23i.55631. Disponível em: <https://revistas.ufg.br/fcs/article/view/55631>. Acesso em: 18 nov. 2024.

49 SILVA, B. S. Principais motivos de saída antecipada dos alunos durante o período de aula. *Educação e Pesquisa*, v. 49, 2023. Disponível em: <https://doi.org/10.1590/S1678-4634202349249413por>. Acesso em 22 nov. 2024.

50 SILVA JUNIOR, W. S.; GONÇALVES, F. O. Evidências da relação entre a frequência no ensino infantil e o desempenho dos alunos do ensino fundamental público no Brasil. *Revista brasileira de Estudos de População*, Rio de Janeiro, v. 33, n. 2, p. 283-301, 2016. Disponível em: <https://doi.org/10.20947/S0102-30982016a0015>. Acesso em 18 nov. 2024.

51 SANTOS, R. V. *Otimização do quadro de horário de disciplinas em uma escola da rede estadual de ensino do estado do Rio de Janeiro: uma proposta baseada em análise estatística da frequência estudantil*. Monografia: Licenciatura em Matemática, Instituto Multidisciplinar, Universidade Federal Rural do Rio de Janeiro. Rio de Janeiro. 2024. 54 p.

52 SEABOLD, S.; PERKTOLD, J. Statsmodels: Econometric and statistical modeling with Python. In *9th Python in Science Conference*, p. 92-96, 2010. Disponível em: <https://www.statsmodels.org>. Acesso em 30 dez. 2024.

53 SENADO FEDERAL. *As 20 metas do PNE e a avaliação do Inep*. Brasília, DF, 17 fev. 2023. Disponível em: <https://www12.senado.leg.br/noticias/materias/2023/02/17/as-20-metas-do-pne-e-a-avaliacao-do-inep>. Acesso em 18 nov. 2024.

54 SOARES, J. F.; ALVES, M. T. G.; FONSECA, J. A. Trajetórias educacionais como evidência da qualidade da educação básica brasileira. *Revista brasileira de Estudos de População*, v. 38, p. 1-21, 2021. Disponível em: <http://dx.doi.org/10.20947/S0102-3098a0167>. Acesso em: 18 nov. 2024.

55 VIRTANEN, P.; *et al*. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, v. 17(3), p. 261–272, 2020. Disponível em: <https://doi.org/10.1038/s41592-019-0686-2>. Acesso em 30 dez. 2024.

56 WALBERG, H. J. Synthesis of research on time and learning. *Educational Leadership*, v. 45, p. 76-85, 1988. Disponível em: <https://www.researchgate.net/publication/234750966_Synthesis_of_Research_on_Time_and_Learning>. Acesso em 18 nov. 2024.

57 ZHU, C.; *et al*. Algorithm 778: L-BFGS-B: Fortran Subroutines for Large-Scale Bound Constrained Optimization. *ACM Transactions on Mathematical Software*, v. 23, n. 4, p. 550-560, 1997. Disponível em: <https://dl.acm.org/doi/pdf/10.1145/279232.279236>. Acesso em 25 jan. 2025.

8 APÊNDICES

8.1 *Boxplots*

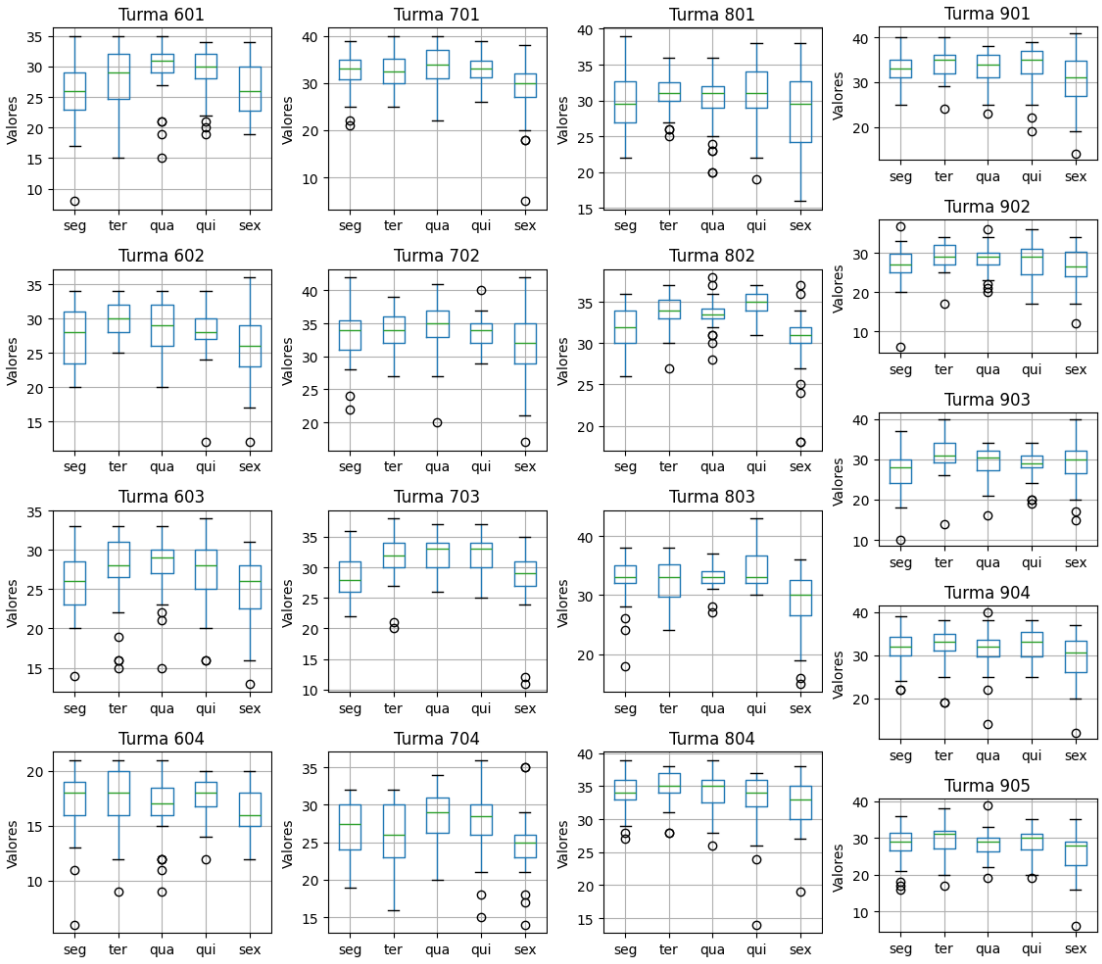

Figura 8.1 – *Boxplots* de frequências das turmas do Ensino Fundamental.

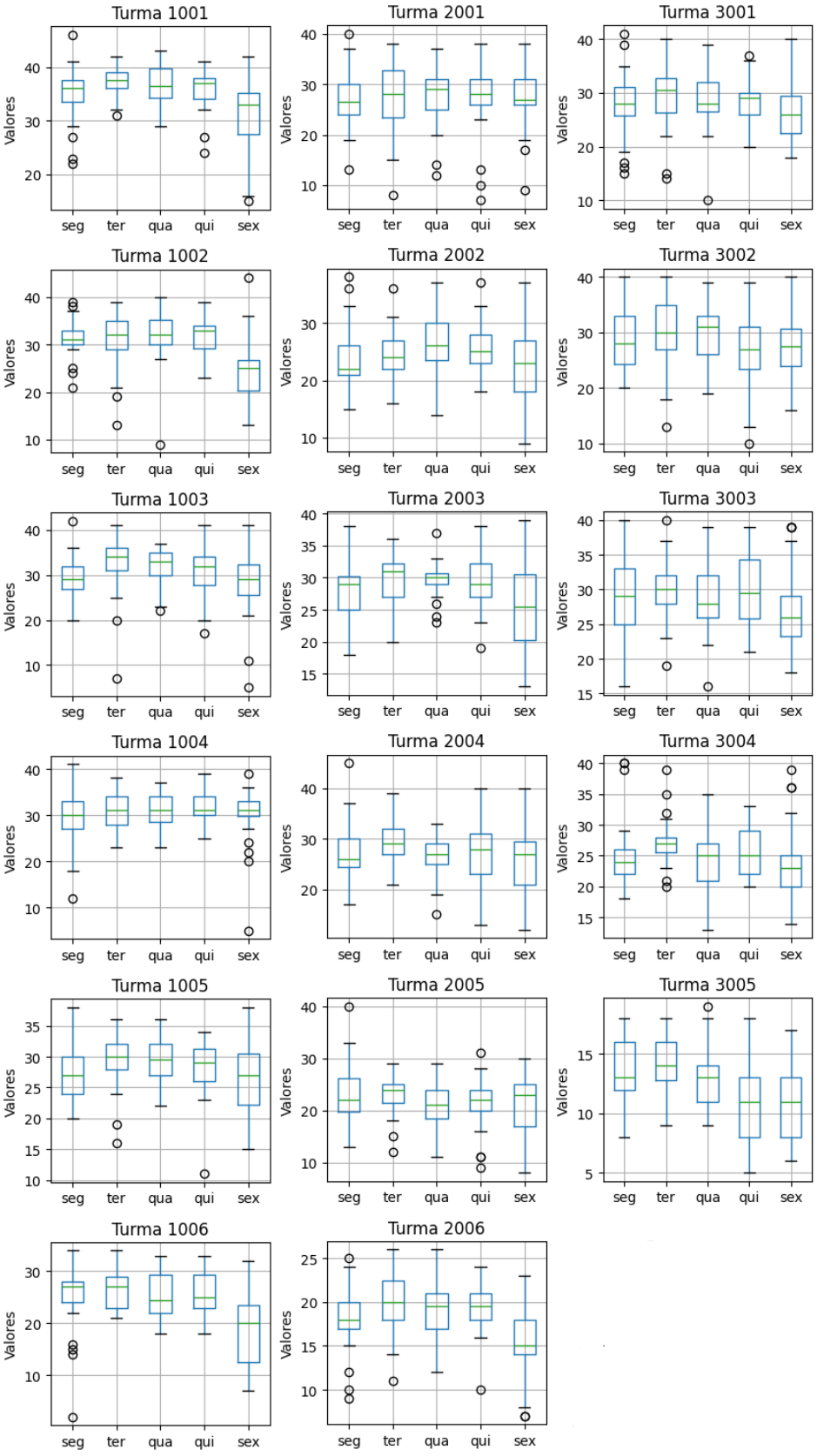

Figura 8.2 – *Boxplots* de frequências das turmas do Ensino Médio.

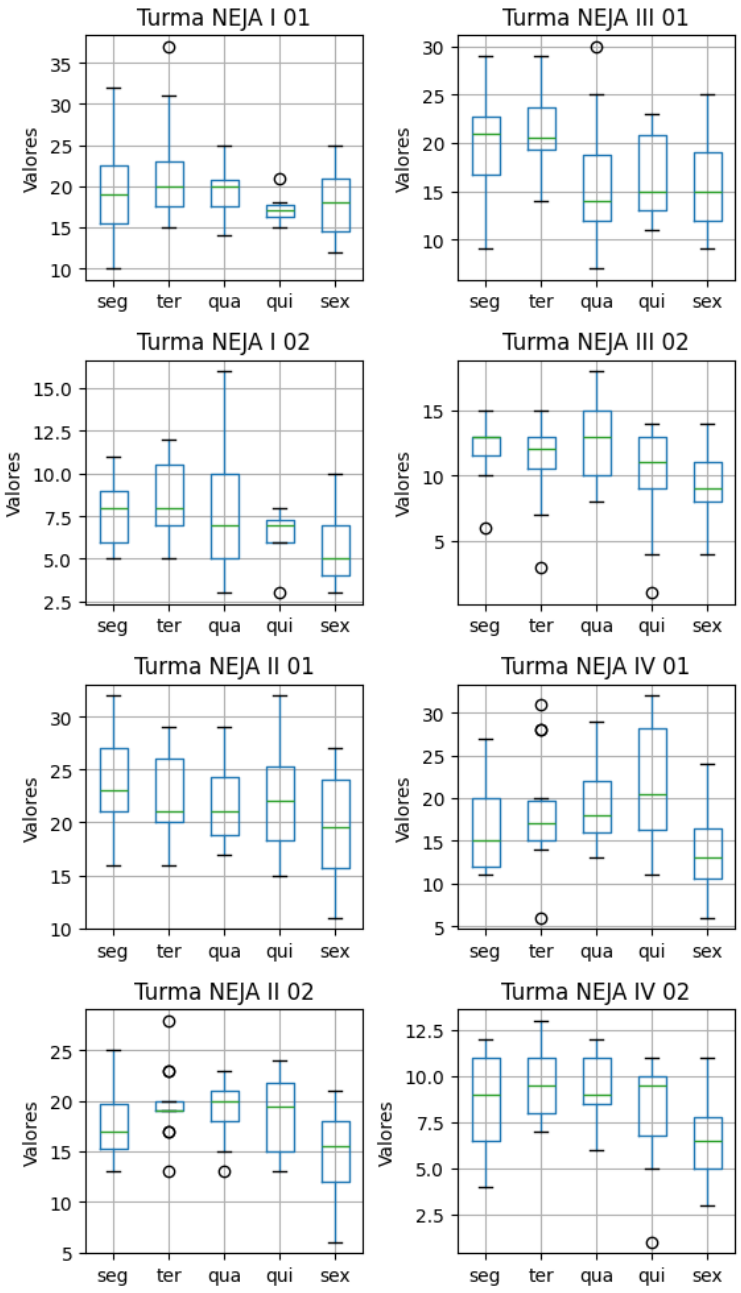

Figura 8.3 – *Boxplots* de frequências das turmas do NEJA.

8.2 Dados do modelo ANOVA

- Código da análise de frequência;
- Tabelas de dados.

8.3 Repositório do *software*

- Código do aplicativo desenvolvido;
- Executável;
- Manual do usuário.

8.4 Dados do modelo de minimização quadrática

- Código da implementação do modelo;
- Dados de frequência por bimestre e turno.

8.5 Tabela: Valor-p (a 5% de significância) dos testes de Tukey para cada uma das turmas selecionadas após o teste F da ANOVA.

	TESTE DE TUKEY (VALOR p)									
TURMA	**SEG-TER**	**SEG-QUA**	**SEG-QUI**	**SEG-SEX**	**TER-QUA**	**TER-QUI**	**TER-SEX**	**QUA-QUI**	**QUA-SEX**	**QUI-SEX**
601	0,1104	0,0077	0,0103	0,9262	0,8260	0,8788	0,4562	0,9999	0,0629	0,0809
602	0,0890	0,7832	0,9910	0,6815	0,6427	0,2533	0,0015	0,9622	0,1041	0,4028
701	0,9861	0,6683	0,9525	0,0161	0,9195	0,9995	0,0030	0,9731	0,0002	0,0017
703	0,0314	0,0104	0,0195	0,9788	0,9773	0,9990	0,0040	0,9976	0,0013	0,0024
704	0,9906	0,5771	0,9544	0,1902	0,3071	0,7725	0,4321	0,9382	0,0031	0,0349
802	0,0453	0,1818	0,0012	0,3284	0,9884	0,7283	0,0000	0,4571	0,0007	0,0000
803	0,9983	0,9988	0,5973	0,0044	0,9794	0,4106	0,0105	0,7713	0,0021	0,0000
804	0,7804	0,9827	0,8506	0,3454	0,9714	0,1876	0,0191	0,5107	0,1025	0,9231
901	0,7734	0,9990	0,9883	0,1686	0,9043	0,9648	0,0062	0,9994	0,1025	0,0544
903	0,0136	0,3261	0,6944	0,4905	0,7152	0,2862	0,4419	0,9667	0,9964	0,9980
1001	0,2273	0,6912	0,9565	0,0034	0,9393	0,6562	0,0000	0,9770	0,0000	0,0004
1002	0,9999	0,9933	0,9890	0,0001	0,9788	0,9702	0,0001	1,0000	0,0000	0,0000
1003	0,1761	0,3404	0,9174	0,9567	0,9974	0,6114	0,0427	0,8211	0,1026	0,5570
1006	0,8395	1,0000	0,9981	0,0000	0,8960	0,9596	0,0000	0,9996	0,0000	0,0000
2002	0,9950	0,2142	0,6691	0,8770	0,3962	0,8664	0,6589	0,9477	0,0219	0,1671
2003	0,6267	0,7030	0,6813	0,2844	1,0000	1,0000	0,0105	1,0000	0,0163	0,0161
2006	0,4678	0,7704	0,8711	0,0151	0,9887	0,9665	0,0001	0,9998	0,0004	0,0010
3005	0,9618	0,9317	0,0071	0,0021	0,5918	0,0008	0,0002	0,0673	0,0302	1,0000
NEJA I 2	0,9324	1,0000	0,7649	0,1565	0,9496	0,4509	0,0330	0,8198	0,2641	0,9841
NEJA II 2	0,4973	0,6962	0,8886	0,2292	0,9989	0,9613	0,0033	0,9949	0,0110	0,0282
NEJA III 1	0,9758	0,2002	0,4985	0,2455	0,0549	0,2138	0,0735	0,9952	1,0000	0,9978
NEJA III 2	0,9920	0,9641	0,5438	0,1926	0,7699	0,7671	0,3452	0,1592	0,0276	0,9702
NEJA IV 2	0,9569	0,9737	0,9518	0,1855	1,0000	0,6405	0,0452	0,7166	0,0720	0,5708

8.6 Gráficos das frequências reais e esperadas por bimestre e turno

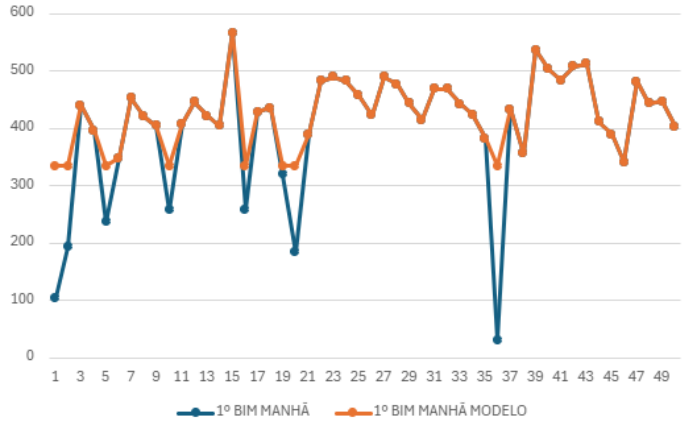

Figura 8.4 – Gráfico de linhas das frequências reais e das frequências esperadas geradas pelo modelo do 1º bimestre, turno da manhã.

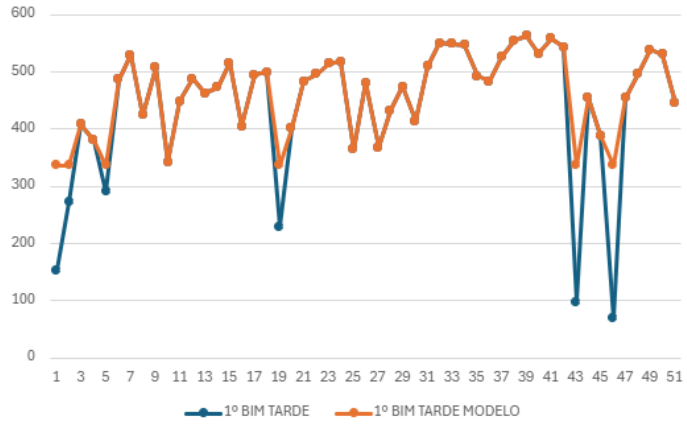

Figura 8.5 – Gráfico de linhas das frequências reais e das frequências esperadas geradas pelo modelo do 1º bimestre, turno da tarde.

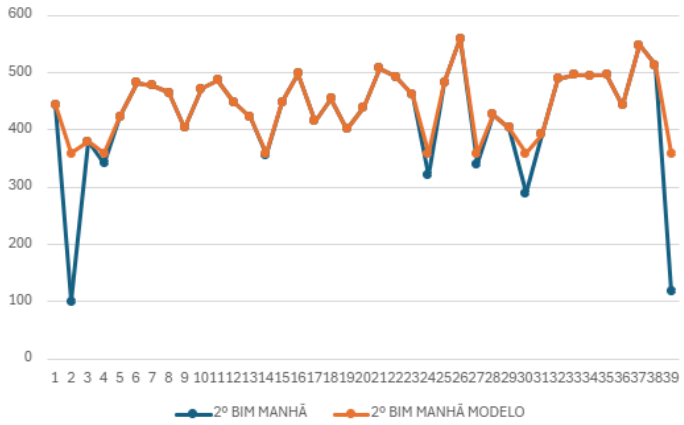

Figura 8.6 – Gráfico de linhas das frequências reais e das frequências esperadas geradas pelo modelo do 2º bimestre, turno da manhã.

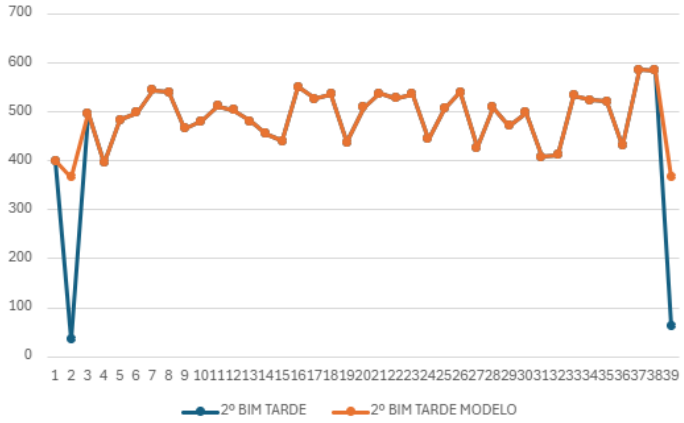

Figura 8.7 – Gráfico de linhas das frequências reais e das frequências esperadas geradas pelo modelo do 2º bimestre, turno da tarde.

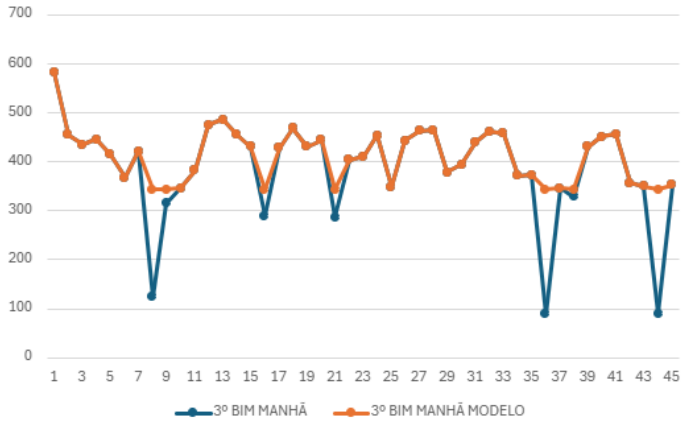

Figura 8.8 – Gráfico de linhas das frequências reais e das frequências esperadas geradas pelo modelo do 3º bimestre, turno da manhã.

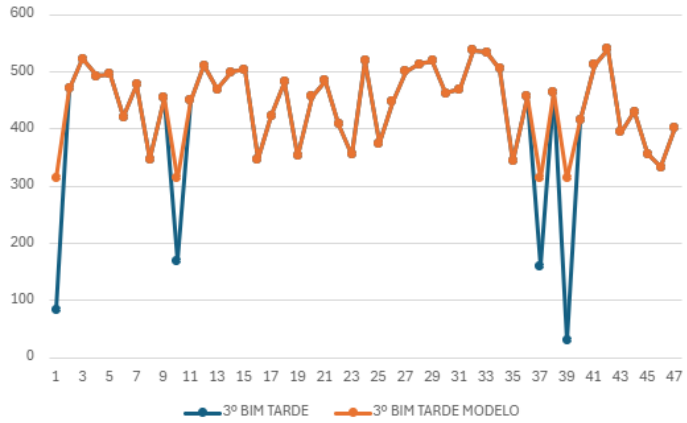

Figura 8.9 – Gráfico de linhas das frequências reais e das frequências esperadas geradas pelo modelo do 3º bimestre, turno da tarde.

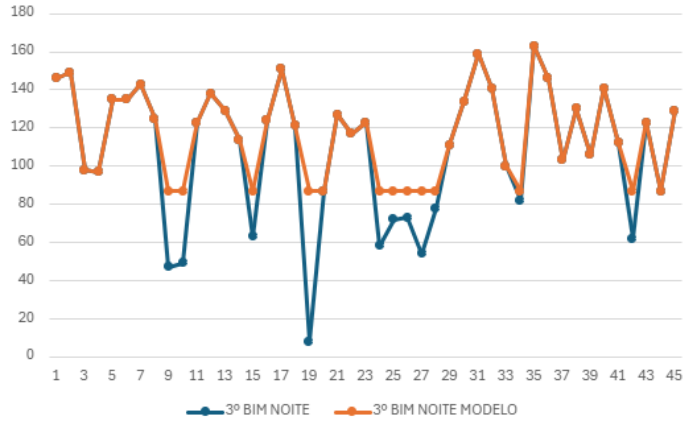

Figura 8.10 – Gráfico de linhas das frequências reais e das frequências esperadas geradas pelo modelo do 3º bimestre, turno da noite.

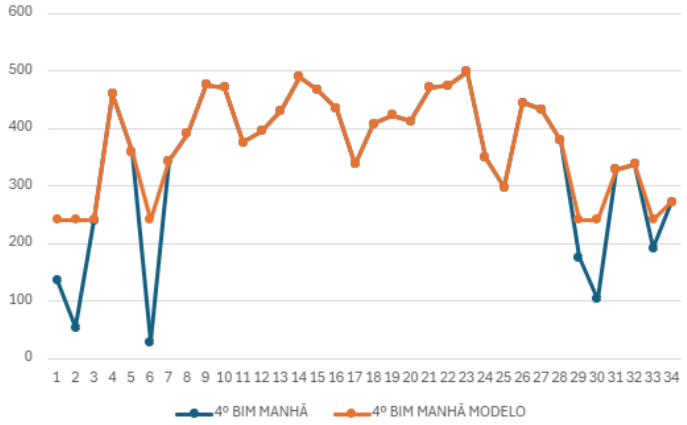

Figura 8.11 – Gráfico de linhas das frequências reais e das frequências esperadas geradas pelo modelo do 4º bimestre, turno da manhã.

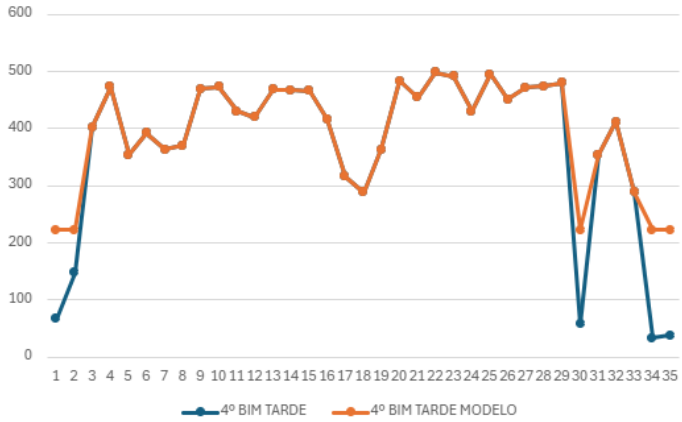

Figura 8.12 – Gráfico de linhas das frequências reais e das frequências esperadas geradas pelo modelo do 4º bimestre, turno da tarde.

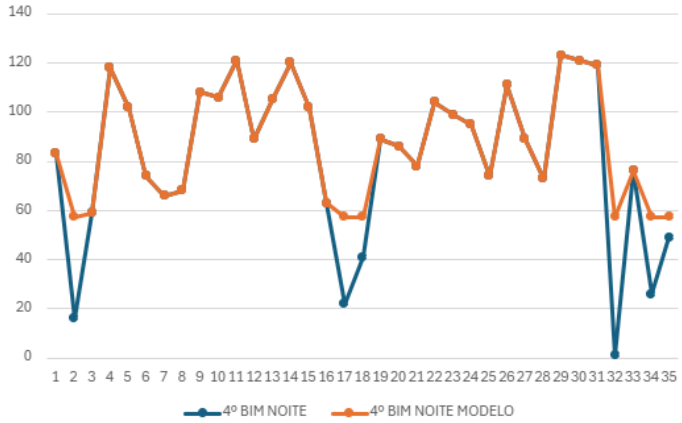

Figura 8.13 – Gráfico de linhas das frequências reais e das frequências esperadas geradas pelo modelo do 4º bimestre, turno da noite.

8.7 Tabela: Comparações entre valores reais e diferentes expectativas de frequência.

COMPARAÇÕES ENTRE VALORES REAIS E DIFERENTES EXPECTATIVAS								
	Total de alunos ($\sum c_j$)	**Total de alunos esperados** ($\sum y_i$)	**Diferença** ($\sum y_i - \sum c_j$)	**Total de alunos esperados pela mediana** (nm_e)	**Diferença** (nm_e - $\sum c_j$)	**Total de alunos esperados pela moda** (nm_o)	**Diferença** (nm_o - $\sum c_j$)	**Total de alunos esperados pelo teto** (nt)	**Diferença** ($nt - \sum c_j$)
1ºB MAN	20139,00	21229,95	1090,95	21300,00	1161,00	20200,00	61,00	28300,00	8161,00
1ºB TAR	22647,00	23558,35	911,35	24633,00	1986,00	24633,00	1986,00	28713,00	6066,00
2ºB MAN	16632,00	17280,18	648,18	17511,00	879,00	15756,00	-876,00	21840,00	5208,00
2ºB TAR	18335,00	18970,11	635,11	19422,00	1087,00	19422,00	1087,00	22815,00	4480,00
3ºB MAN	17588,00	18464,84	876,84	18945,00	1357,00	20475,00	2887,00	26190,00	8602,00
3ºB TAR	19898,00	20712,64	814,64	21479,00	1581,00	16309,00	-3589,00	25380,00	5482,00
3ºB NOI	4913,00	5223,94	310,94	5445,00	532,00	5535,00	622,00	7335,00	2422,00
4ºB MAN	11891,00	12651,11	760,11	13090,00	1199,00	16014,00	4123,00	16966,00	5075,00
4ºB TAR	13057,00	13824,33	767,33	14700,00	1643,00	16555,00	3498,00	17430,00	4373,00
4ºB NOI	2876,00	3065,46	189,46	3204,00	328,00	3204,00	328,00	4428,00	1552,00