UNIVERSIDADE DO ESTADO DE SANTA CATARINA - UDESC
CENTRO DE CIÊNCIAS TECNOLÓGICAS - CCT

Gabriela da Silva Inácio

ANÁLISES ESTATÍSTICAS REALIZADAS NO PERÍODO DAS FÉRIAS

Joinville
2024

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 OBJETIVO

Realizar análises estatísticas em banco de dados com o intuito de aperfeiçoa os estudos na linguagem de programação R e no segmento estatístico.

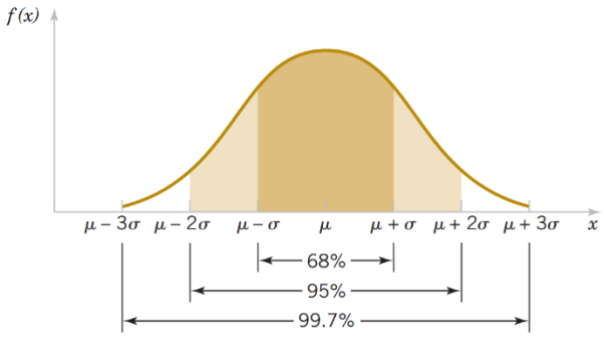
2 FUNDAMENTAÇÃO TEÓRICA

2.1 CORRELAÇÃO ENTRE VARIÁVEIS: CORRELAÇÃO DE PEARSON

A análise de correlação de Pearson é uma ferramenta utilizada para medir a força e a direção da relação linear entre duas variáveis numéricas. Ela está no intervalo de -1 a 1. Dessa maneira, valores próximos de 1 indicam uma correlação linear positiva forte, enquanto valores próximos de -1 indicam uma correlação linear negativa forte. Quanto mais próximo de 0, menor é a força da dessa correlação (BARBETTA; REIS; BORNIA, 2010). No entanto, para que seja realizado o teste de Pearson, considera-se alguns pressupostos, portanto é necessário que haja:

- Distribuição Normal das Variáveis: É como os dados se espalham ao redor de um valor médio. A distribuição normal é representada por uma curva em forma de sino. O meio da curva (o topo do sino) representa a média, que é o valor mais comum. O quanto os dados se afastam da média é controlado pelo desvio padrão. Se o desvio padrão for grande, os dados se espalham mais. Muitos dados (cerca de 68%) estão perto da média, mais dados (cerca de 95%) estão dentro de dois desvios padrão, e quase todos os dados (99.7%) estão dentro de três desvios padrão (STOROPOLI; VILS, 2021).

Figura 1 - Ilustração geral de uma distribuição Normal.



Fonte: Zibetti (2024)

- Homocedasticidade: significa que dos dados variam de forma consistente em diferentes níveis da variável independente. Se a dispersão dos resultados ao longo da linha de ajuste for mais ou menos a mesma, você

tem homocedasticidade. Ela pode ser analisada através de gráficos de resíduos. Gráficos de resíduos acontecem quando cria-se um modelo estatístico, como uma regressão linear. Dessa forma, os resíduos são as diferenças entre os valores previstos pelo modelo e os valores reais. Os gráficos de resíduos mostram visualmente essas diferenças (STOROPOLI; VILS, 2021).

- Relação Linear: A correlação de Pearson mede apenas relações lineares entre variáveis (STOROPOLI; VILS, 2021).
- Ausência de Outliers (STOROPOLI; VILS, 2021).

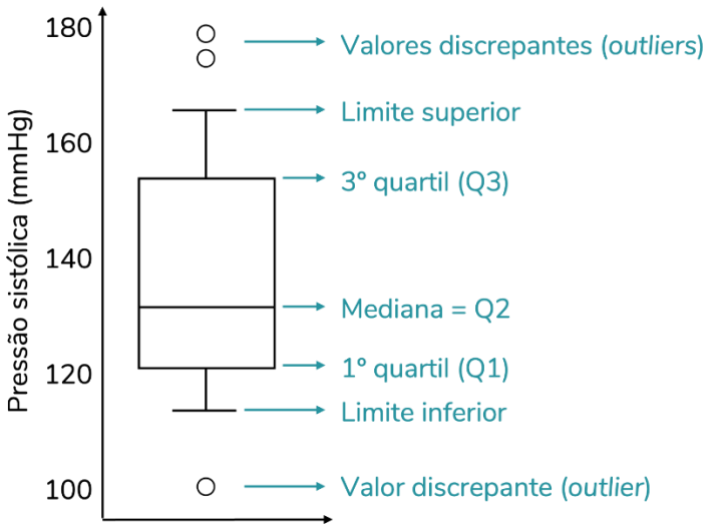
2.2 GRÁFICOS *BOXPLOT*

O boxplot, segundo Peres (2022) é um gráfico utilizado para representar a distribuição de uma variável numérica, já que ele fornece uma representação visual dessa distribuição, incluindo mediana, quartis e possíveis *outliers.* Seus elementos (Figura 4) são:

- *Outliers:* valores discrepantes se comparados com o conjunto de dados; São valores representados fora dos limites superior e inferior (PERES, 2022).
- Limite superior e inferior (PERES, 2022).
- Mediana: é a medida que separa os dados ordenados em 50% superiores e 50% inferiores, sendo, dessa forma, o ponto médio dos dados. No gráfico *boxplot* a mediana coincide com o 2º quartil PERES, 2022).
- 1º e 3º quartis: É semelhante a lógica da mediana, no entanto, ao invés de dividir os dados na metade, é divido por quatro (cada parte/ quartil contendo 25% dos dados) O primeiro quartil (Q1) é o valor abaixo do qual 25% dos dados estão localizados. O terceiro quartil (Q3) é o valor abaixo do qual 75% dos dados estão localizados. O segundo quartil é a mediana, o valor que divide a distribuição ao meio. (PERES, 2022)
- A amplitude interquartil (IQR) é a diferença entre Q3 e Q1 e é representada pelo próprio box no boxplot (a IQR é proporcional ao seu comprimento). (PERES, 2022)

- A "whisker" (linha que se estende para fora do box) geralmente se estende até 1,5 vezes o IQR a partir de Q1 e Q3, e valores fora dessa faixa são considerados outliers. (PERES, 2022).

Figura 2 - Elementos do gráfico *boxplot.*



Fonte: Peres (2022).

Nesse contexto, o gráfico *boxplot* será útil para identificar padrões, comparar distribuições e destacar diferenças, já que poderia ser feito gráficos separados para os dois grupos de canteiros de obras (com água de chuva e sem água de chuva), no qual o eixo y poderia corresponder ao consumo de água. Dessa maneira, poderia ser analisado para ver qual teria maior consumo de água, já que uma mediana mais alta indica que a maioria dos dados está concentrada em valores mais altos, o que sugere um maior consumo médio de água.

2.3 TESTE T

O teste t é empregado para comparar a média de uma variável entre dois grupos. Esse processo envolve o cálculo da diferença entre as médias desses grupos, seguido pela divisão desse resultado pelo desvio padrão da diferença. Essa abordagem permite uma avaliação da diferença padronizada das médias, tornando mais claro quando as diferenças são estatisticamente significativas ou simplesmente casuais. Além disso, o teste utiliza uma distribuição estatística

conhecida como distribuição-t, capaz de fornecer resultados robustos mesmo em amostras pequenas (n < 10) (LIMA JUNIOR, 2023).

No entanto, antes de avaliar a significância estatística de uma diferença entre médias, é crucial verificar se as premissas do teste foram atendidas. É importante notar que existem duas categorias principais de testes estatísticos: testes paramétricos, que partem de pressupostos restritivos sobre a distribuição dos parâmetros e exigem testes adicionais; e testes não paramétricos, que não dependem desses pressupostos, embora geralmente sejam menos poderosos (LIMA JUNIOR, 2023).

O teste t baseia-se nas premissas de normalidade e homocedasticidade. Além disso, presume que os dois grupos de dados, cujas médias estão sendo comparadas, sejam independentes ou dependentes entre si, causando um efeito em uma variável comum. Essa abordagem é essencial para garantir resultados confiáveis e interpretações precisas das diferenças entre as médias de interesse (LIMA JUNIOR, 2023).

3 METODOLOGIA

O presente trabalho visa um estudo estatístico geral para aprimoramento em linguagem R. Diante disso, a metodologia irá dividir-se nas seguintes partes: 1) Análise descritiva de variáveis; 2) Gráfico de dispersão entre variáveis numéricas; 3) Correlação de Pearson entre variáveis quantitativas; 4) Análise para variáveis qualitativas.

3.1 ANÁLISE DESCRITIVA DE VARIÁVEIS

A análise descritiva das variáveis foi conduzida utilizando um banco de dados composto por seis variáveis, adquirido do vídeo "Estatísticas Descritivas no R – Tabelas" (YouTube). Este conjunto de dados proporcionou a obtenção de medidas de tendência central e dispersão, focando na variável numérica associada aos salários dos indivíduos (PERES, 2020).

Faz-se a análise através de funções que pertencem ao pacote *dplyr,* a partir de um agrupamento das variáveis categóricas Gênero e Grau de Instrução, através de uma análise descritiva utilizando a linguagem de programação R.

3.2 GRÁFICO DE DISPERSÃO ENTRE VARIÁVEIS NUMÉRICAS

O banco de dados empregado para gerar o gráfico foi extraído do vídeo "Criando Gráficos no R com o ggplot2 (Parte 1)" (YouTube) e compreende informações relacionadas a filmes, incluindo dados sobre seus lucros e investimentos (PERES, 2021).

A elaboração do gráfico de dispersão foi efetuada por meio de funções incorporadas aos pacotes *dplyr* e *ggplot2*, utilizando a linguagem de programação R. As variáveis numéricas Lucro Mundial e Lucro Local, expressas em dólares, foram utilizadas para a representação visual das relações entre esses dois parâmetros nos filmes analisados.

3.3 CORRELAÇÃO DE PEARSON ENTRE VARIÁVEIS QUANTITATIVAS

O *data frame* empregado para investigar as correlações é proveniente da biblioteca do R, denominado mtcars. Esse conjunto de dados abrange 11 variáveis que descrevem informações detalhadas sobre veículos, sendo o foco

principal da análise observar as relações de correlação entre as milhas por galão e o peso dos veículos.

Para realizar essa investigação, foram necessários os pacotes *corrplot*, *dplyr* e *ggplot2*. A metodologia inicia-se com uma análise de distribuição dessas variáveis por meio do teste de Shapiro-Wilk. Em seguida, é conduzida uma análise da presença de variáveis por meio de um gráfico *boxplot*. Posteriormente, é executado um modelo de Regressão Linear, e a homocedasticidade das variáveis é analisada por meio de gráficos de resíduos derivados do modelo de Regressão. Por fim, é realizada a análise da correlação de Pearson entre as variáveis em questão. Além disso, uma matriz de correlação visual é apresentada para enfatizar as relações entre essas variáveis (ANALYSIS, 2024).

3.4 ANÁLISE PARA VARIÁVEIS QUALITATIVAS

O presente estudo baseia-se em um banco de dados composto por informações de 32 alunos, abrangendo tanto a rede pública quanto a privada. O objetivo é realizar uma análise estatística para investigar se a posição ocupada pelo aluno na sala de aula, especificamente na frente ou nos fundos, exerce algum efeito significativo em suas notas nas disciplinas de Biologia, Física e História. Para realizar essa investigação, uma série de procedimentos foi adotada, utilizando as ferramentas estatísticas disponíveis nos pacotes RVAideMemoire, dplyr, car, ggrastr, gridExtra e ggplot2.

Inicialmente, é realizada uma análise de distribuição das variáveis relevantes por meio do teste de Shapiro-Wilk. Esse procedimento busca verificar a normalidade das distribuições das notas para os grupos de alunos na frente e nos fundos da sala.

Posteriormente, é conduzida uma análise da homogeneidade das variâncias por meio do teste de Levene. Esse teste avalia se as variâncias das notas são estatisticamente iguais entre os grupos, fornecendo *insights* cruciais para a aplicação do teste t.

O teste t de duas amostras é então empregado para avaliar se há diferenças significativas nas médias das notas entre os alunos que ocupam a frente e os que estão nos fundos. Este teste é fundamental para determinar se a

posição na sala influencia de maneira significativa o desempenho dos alunos nas disciplinas selecionadas.

Por fim, para proporcionar uma visualização mais clara das distribuições das notas entre os grupos, é gerado um gráfico de *boxplot*. Esse gráfico destaca estatísticas descritivas como mediana, quartis e possíveis valores atípicos, oferecendo uma perspectiva visual sobre a dispersão dos dados.

4 RESULTADOS E DISCUSSÕES

4.1 ANÁLISE DESCRITIVA DE VARIÁVEIS

Primeiramente, em relação ao grau de instrução, há progressão nos valores médios à medida que se avança do ensino fundamental para o ensino médio e, finalmente, para o nível superior, sobretudo para o gênero masculino. Isso sugere que, em média, o grau de instrução aumenta à medida que os indivíduos avançam em seus estudos. Além disso, nota-se que os homens tendem a ter médias mais altas do que as mulheres, exceto no ensino fundamental.

Ao analisar as medidas centrais, observa-se que, em média, os salários mais altas estão associados ao Gênero Masculino com nível Superior de Instrução, apresentando a média mais elevada (3,28). Em contraste, os participantes do Gênero Masculino com Ensino Fundamental registram a menor média (2,28).

Tabela 1 - Análise descritiva dos salários classificados por gênero e grau de instrução.

Gênero	Grau de Instrução	Mínimo	1° Quartil	Mediana	Média	3° Quartil	Desvio Padrão
F	Ensino Fundamental	2,05	2,45	2,85	2,85	3,25	1,13
F	Ensino Médio	1,00	1,78	1.85	2,35	2.88	1,14
F	Superior	2,20	2,74	3,15	3,22	3,82	0,76
M	Ensino Fundamental	1,45	1,79	2,12	2,28	2,61	0,84
M	Ensino Médio	1,60	2,17	2,75	2,67	3,20	1,03
M	Superior	2,10	2,85	3,15	3,58	4,25	1,21

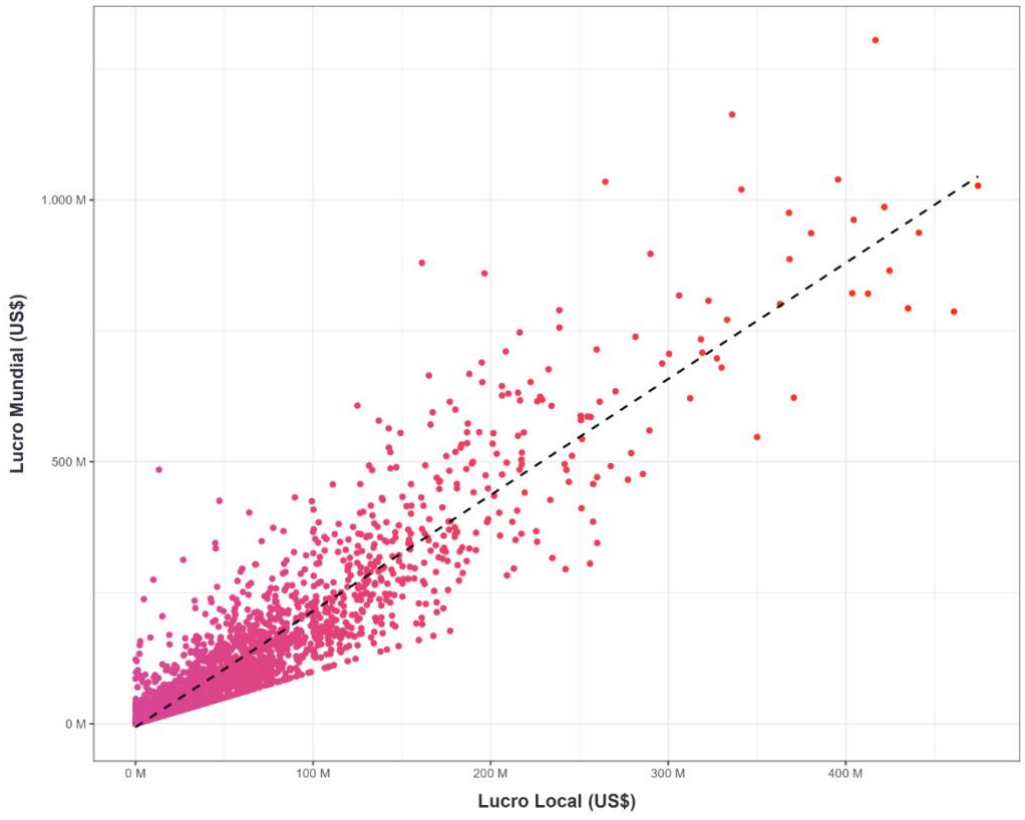
Fonte: Da autora (2024).

Quanto à variabilidade, nota-se que, em geral, as pontuações para o Gênero Feminino apresentam menor dispersão em comparação com o Gênero Masculino em relação aos salários. Isso pode indicar uma maior consistência nos salários das mulheres.

4.2 GRÁFICO DE DISPERSÃO ENTRE VARIÁVEIS NUMÉRICAS

A análise do gráfico de dispersão revela uma relação positiva linear entre as variáveis de lucro local e lucro mundial para os filmes apresentados (Figura 3). Essa tendência sugere que, à medida que o lucro local aumenta, o lucro mundial também tende a crescer proporcionalmente.

Figura 3 - Gráfico de Dispersão: Lucro Mundial X Lucro Local



Fonte: Da autora (2024).

A disposição dos pontos no gráfico forma uma linha de inclinação positiva, indicando uma associação positiva entre as duas variáveis. Isso sugere que filmes que têm um desempenho financeiro sólido em seus mercados locais também tendem a alcançar resultados positivos em escala global.

4.3 CORRELAÇÃO DE PEARSON ENTRE VARIÁVEIS QUANTITATIVAS

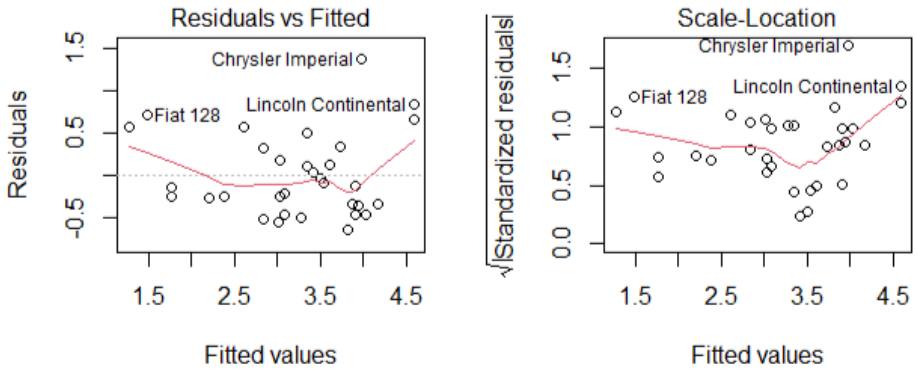
De início, realizou-se um teste de normalidade *Shapiro-Wilk* para avaliar a distribuição da variável peso (*wt).* Dessa forma, obteve-se os resultados da estatística de teste (*W*): 0.94326 e valor p (*p*): 0.09265. De acordo com os resultados, não foi encontrada evidência estatística significativa para rejeitar a hipótese nula de normalidade da variável w*t* (p-valor = 0.09265). Assim, podemos assumir que a variável *wt* segue aproximadamente uma distribuição normal.

Além disso, realizou-se o mesmo teste para avaliar a distribuição da variável milhas por galão (mpg*),* adquirindo os resultados da estatística de teste (*W*): 0.94756 e valor p (*p*): 0.09265. Dessa maneira, tem-se que a variável mpg também segue uma distribuição normal.

Posteriormente fez-se a checagem da presença de outliers através de gráficos *boxplot*. A variável *mpg* não apresenta valores discrepantes, enquanto a variável *wt* apresenta apenas dois. Ademais, verificou-se uma relação de linearidade negativa entre as duas variáveis.

De acordo com o gráfico de resíduos (Figura 4), os dados mostram que a dispersão dos erros permanece constante em diferentes valores da variável independente, indicando homocedasticidade.

Figura 4 - Gráfico de resíduos.

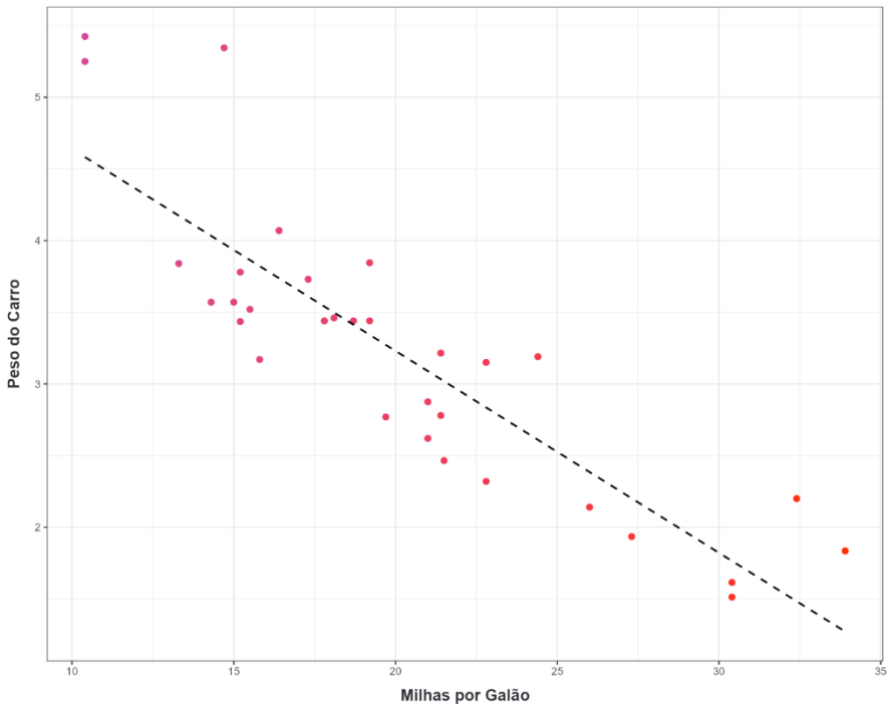


Fonte: Da autora (2024).

Por fim, a análise de correlação de Pearson foi realizada para avaliar a relação entre as variáveis *wt* (peso do carro) e *mpg* (milhas por galão) no

conjunto de dados. Essa relação é expressa graficamente e representada pela Figura 5.

Figura 5 - Gráfico de dispersão das variáveis mpg e wt.

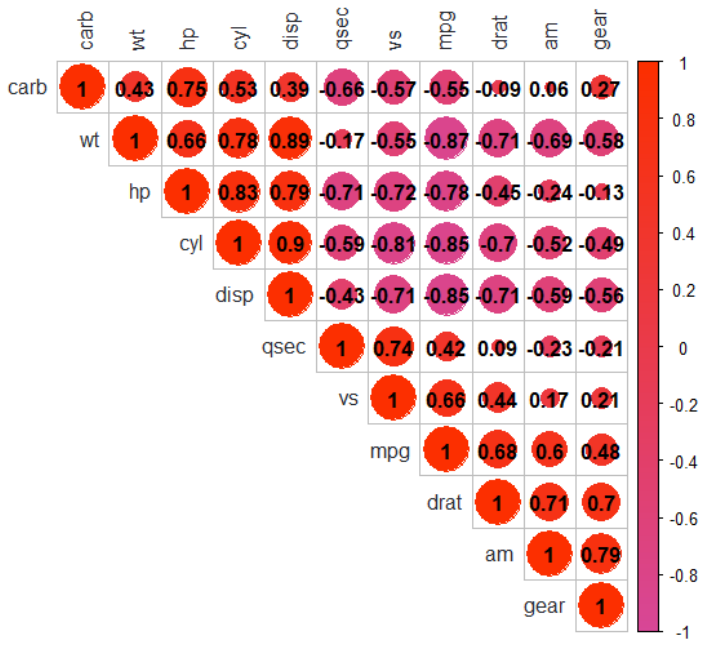


Fonte: Da autora (2024).

Os resultados revelaram uma correlação negativa forte entre as duas variáveis (correlação de Pearson = -0.87, p-value = 1.294^{-10}), indicando que o peso do carro está inversamente relacionado ao consumo de combustível. O intervalo de confiança de 95% para a correlação foi de -0.93 a -0.74.

Além disso, para uma análise mais visual, foi possível estabelecer uma matriz de correlação de Pearson entre todas as variáveis do banco de dados, observada na Figura 6.

Figura 6 - Representação visual de uma matriz de correlação de Pearson.



Fonte: Da autora (2024).

Através dela, constata-se que, dentre todas as variáveis, a relação mais forte da variável milhas por galão é com o peso do veículo, resultado o qual reforça que o peso do carro é o fator que mais interfere no consumo de combustível.

4.4 ANÁLISE PARA VARIÁVEIS QUALITATIVAS

A análise dos dados foi iniciada pela verificação da normalidade das distribuições, utilizando o teste de Shapiro-Wilk. Os resultados (Tabela 2) indicaram que, para todas as disciplinas e grupos (Frente e Fundos), os valores de p foram superiores a 0,05, indicando que não há evidências suficientes para rejeitar a hipótese nula de normalidade. Portanto, pode-se inferir que as notas dos alunos apresentam uma distribuição aproximadamente normal para ambos os grupos.

Tabela 2 - Resultados obtidos no teste de Shapiro-Wilk.

	Biologia	Física	História
Frente	W = 0.9852 p = 0.99312	W = 0.9327, p = 0.2992	W = 0.8936, p = 0.07594
Fundos	W = 0.9003, p = 0.06865	W = 0.9301, p = 0.2186	W = 0.9168, p = 0.13056

Fonte: Da autora (2024).

Em seguida, realizou-se o teste de homogeneidade de variância de Levene, que avalia se as variâncias das notas são iguais entre os grupos. Os resultados (Tabela 3) indicaram significância estatística para as disciplinas de Física e História, sugerindo que as variâncias das notas são diferentes entre os grupos. No entanto, para Biologia, não houve evidências suficientes para rejeitar a hipótese nula de homogeneidade (não há diferenças significativas entre as variâncias das populações representadas por cada grupo).

Tabela 3 - Resultados obtidos no teste de Levene.

Biologia	Física	História
F = 1.0359, p = 0.3169	F = 13.658, p = 0.0008749	F = 14.292, p = 0.0006954

Fonte: Da autora (2024).

Por fim, foi realizado o teste t de duas amostras para investigar se há diferenças significativas nas médias das notas entre os grupos. Os resultados indicaram que, em todas as situações, os alunos que se sentam na frente apresentam notas significativamente maiores, exceto nas matérias de História, onde não foi encontrada diferença significativa.

Tabela 4 - Resultados obtidos no teste t.

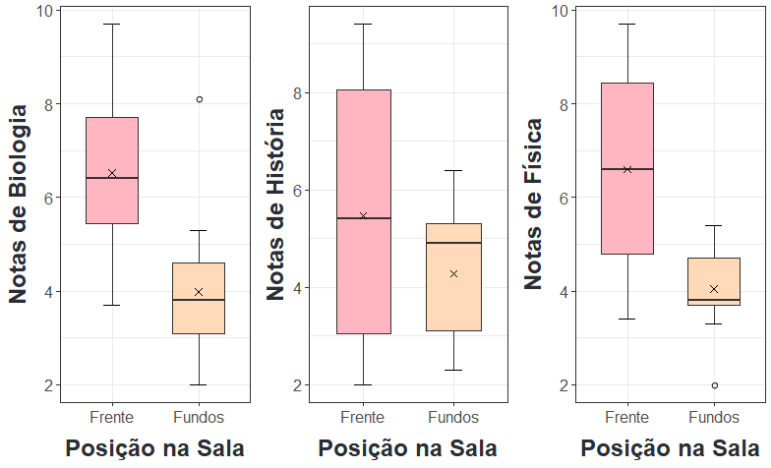
Biologia	Física	História
t = 4.6027, p = 7.136e-05	t = 4.4352, p = 0.0003324	t = 1.5737, p = 0.1313

Fonte: Da autora (2024).

Em suma, constatou-se que, segundo os resultados, a posição na sala de aula influencia as notas dos alunos em Biologia e Física, mas não em História. Alunos que se sentam na frente tendem a ter notas mais altas nessas disciplinas.

Por fim, para proporcionar uma visualização mais clara das distribuições das notas entre os grupos, construiu-se gráficos boxplot (Figura 7) para cada disciplina (Biologia, Física e História).

Figura 7 - Gráfico em grupo boxplot das variáveis notas e posição na sala.



Fonte: Da autora (2024).

Os gráficos destacam a mediana, quartis e possíveis valores atípicos, oferecendo uma perspectiva visual sobre a dispersão dos dados. A média é representada por um “x” sobre a caixa. Observa-se que, contrariamente às outras disciplinas, o boxplot para História não mostra uma diferença significativa nas notas entre os grupos. Ambos os grupos (Frente e Fundos) exibem uma sobreposição substancial nas distribuições, indicando que a posição na sala não parece ter um impacto claro nas notas dessa disciplina, confirmando as conclusões obtidas a partir dos testes estatísticos.

REFERÊNCIAS

ANALYSIS, Statistical Tools For High-Throughput Data. **Visualize correlation matrix using correlogram**. Disponível em: http://www.sthda.com/english/wiki/visualize-correlation-matrix-using-correlogram. Acesso em: 12 jan. 2024.

BARBETTA, Pedro Alberto; REIS, Marcelo Menezes; BORNIA, Antonio Cezar. **Estatística para cursos de engenharia e informática**. 3. ed. São Paulo: Atlas, 2010. 412 p.

LIMA JUNIOR, Paulo. **MÉTODOS QUANTITATIVOS DA PESQUISA EM EDUCAÇÃO**: uma introdução baseada em linguagem r. Brasília: Editora Universidade de Brasília, 2023. 374 p. Disponível em: http://www.realp.unb.br/jspui/bitstream/10482/46607/3/LIVRO_MetodosQuantitativosPesquisa.pdf. Acesso em: 14 jan. 2024.

PERES, Fernanda. **Criando gráficos no R com o ggplot2 (Parte 1)**. 24 ago. 2021. Facebook: FernandaPeres. Disponível em: https://www.youtube.com/watch?v=DYsPRa3vpf0. Acesso em: 10 jan. 2024.

PERES, Fernanda. **Como interpretar (e construir) um gráfico boxplot?** 2022. Disponível em: https://fernandafperes.com.br/blog/interpretacao-boxplot/. Acesso em: 10 jan. 2014.

PERES, Fernanda. **Estatísticas descritivas no R**: tabelas. 12 fev. 2020. YouTube: FernandaPeres. Disponível em: https://www.youtube.com/watch?v=jZvQ4N0nuDY. Acesso em: 08 jan. 2024.

STOROPOLI, Jose; VILS, Leonardo. **Relação entre Variáveis – Correlações**: como que descrevemos a força de associação entre duas variáveis.. Como que descrevemos a força de associação entre duas variáveis.. 2021. Disponível em: https://storopoli.io/Estatistica/5-Correlacoes.html. Acesso em: 11 jan. 2014.

ZIBETTI, André. **Distribuição Normal (Gaussiana)**. Disponível em: https://www.inf.ufsc.br/~andre.zibetti/probabilidade/normal.html. Acesso em: 10 jan. 2024.