Uma Análise Exploratória do dataset do Projeto HS&B

1. Introdução

O presente relatório apresenta uma análise exploratória do dataset **hsb2f.csv**, composto por dados coletados por meio de uma pesquisa de base com alunos do último e segundo ano do ensino médio nos Estados Unidos. O High School and Beyond Project (em português: Projeto Ensino Médio e Além) foi um **estudo longitudinal** dos estudantes do ensino médio e também após o témino de sua formação realizado pelo National Center for Education Statistics (United States Department of Education, 2006). O dataframe utilizado nesta Análise Exploratória de Dados (EAD), denominado **hsb2f.csv** (OPENINTRO, s/d) é uma amostra contendo 200 observações do estudo original, aleatoriamente selecionadas, de características desconhecidas, dos alunos do último ano do ensino médio.

2. Carregando as bibiotecas necessárias

In [50]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from matplotlib.backends.backend_pdf import PdfPages
```

3. Importação dos dados

In [51]:
```
file_path = '~/Library/CloudStorage/OneDrive-Pessoal/Trabalhos 2024/Grupo-2/data/hsb2f.csv' # Observação para uso no meu ambiente
data = pd.read_csv(file_path, delimiter=';')
```

4. Criação das funções utilizadas

In [52]:
```
def autopct_format(values):
    def my_format(pct):
        total = sum(values)
        val = int(round(pct * total / 100.0))
        return f'{pct:.1f}% ({val})'
    return my_format

def plot_categorical(data, column, title):
    plt.figure(figsize=(8, 6))
    values = data[column].value_counts()
    values.plot(kind='pie', autopct=autopct_format(values), startangle=90, colors=sns.color_palette("pastel"))
    plt.title(title)
    plt.ylabel('')
    plt.show()
    plt.close()

def plot_numerical(data, column, title):
    plt.figure(figsize=(8, 6))
    sns.histplot(data[column], kde=True, color='skyblue', bins=10)
    plt.title(title)
    plt.xlabel(f'Notas ({column.capitalize()})')
    plt.ylabel('Número de Alunos')
    plt.show()
    plt.close()
```

5. Análise Descritiva

A análise descritiva é uma técnica de análise de dados que visa resumir, organizar e compreender dados históricos para identificar padrões e relacionamentos. É um dos quatro tipos principais de análise de dados, juntamente com a análise diagnóstica, preditiva e prescritiva (Métricas Boss, 2023). A análise descritiva é essencial para explorar e compreender os dados antes de prosseguir para análises mais avançadas (SIRIUS, 2022). Apesar de ser uma ferramenta simples, realizada no início do trabalho com os dados, a análise descritiva pode ter diferentes tipos, e essa classificação depende da quantidade de elementos que serão interpretados. Os três tipos de classificação são:

- Univariada: análise de dados trabalha com apenas uma variável de forma isolada, sem se relacionar com as outras do dataset sendo analisado. Apresenta apenas uma característica;
- Bivariada: análise feita utilizando-se de duas variáveis. O objetivo é investigar a forma que uma variável se comporta em contato com outra, e medir a relação que existe entre as duas;
- Multivariada: análise realizada simultâneamente entre diversos elementos, relacionando-os entre si permitindo obter inderências mais elaboradas. Em uma *análise univariada*, portanto, analisa-se cada uma das variáveis do *dataset* individualmente. A partir destes resultados pode-se montar um resumo geral dos dados. Na etapa inicial dessa análise serão utilizadas análise descritivas univariadas das dez (10) variáveis do o dataset **hsb2f.csv**. A visualização de dados univariada é feita por meio de gráficos, tais como histogramas, boxplots e gráficos de barras, que ajudam a representar a distribuição dos dados e a destacar características importantes. Para escolher a técnica de visualização univariada, é importante considerar o tipo de dados, ou seja, classificar a variável quanto a seu tipo: se qualitativa (nominal ou ordinal) ou quantitativa (discreta ou contínua).

5.1. Análise Descritiva das variáveis nominais

Para visualizar uma análise univariada de variáveis nominais, de acordo com Mayer (s/d), é possível utilizar gráficos de barras ou de setores, e tabelas de frequências. As variáveis qualitativas ou categóricas podem ser

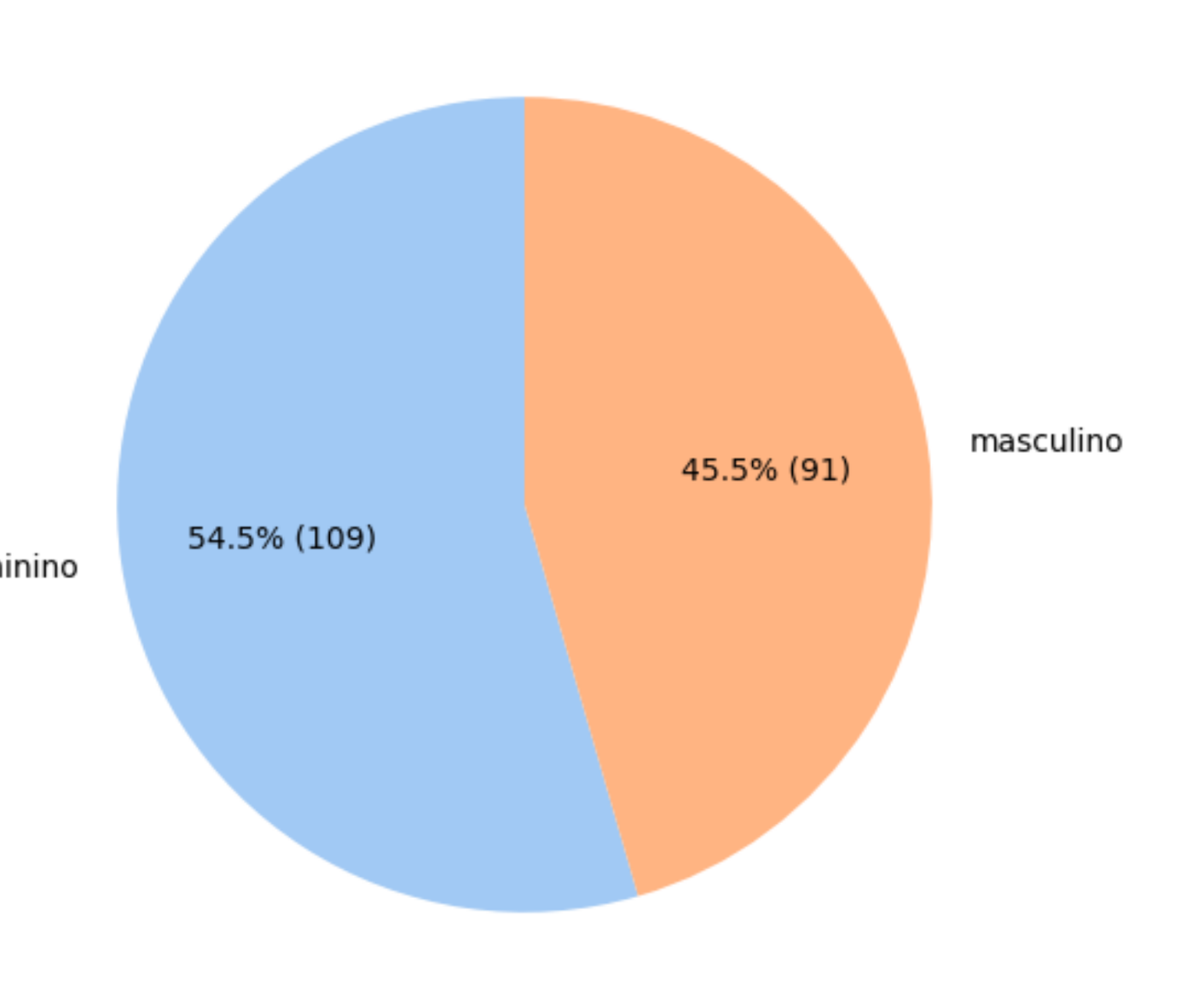
- Nominais: quando as categorias não possuem uma ordem natural, como por exemplo, nome, raça e sexo.
- Ordinais: quando as categorias podem ser ordenadas. Alguns exemplos seriam: classe social (baixa, média, alta) e grau de instrução (básico, médio, graduação, pós-graduação).

5.1.1. Análise Descritiva da variável gênero

O gráfico da **Figura 1** mostra a distribuição de gênero na amostra de dados utilizada.

In [53]:
```
plot_categorical(data, 'genero', 'Figura 1. Distribuição de Gênero')
```

Figura 1. Distribuição de Gênero

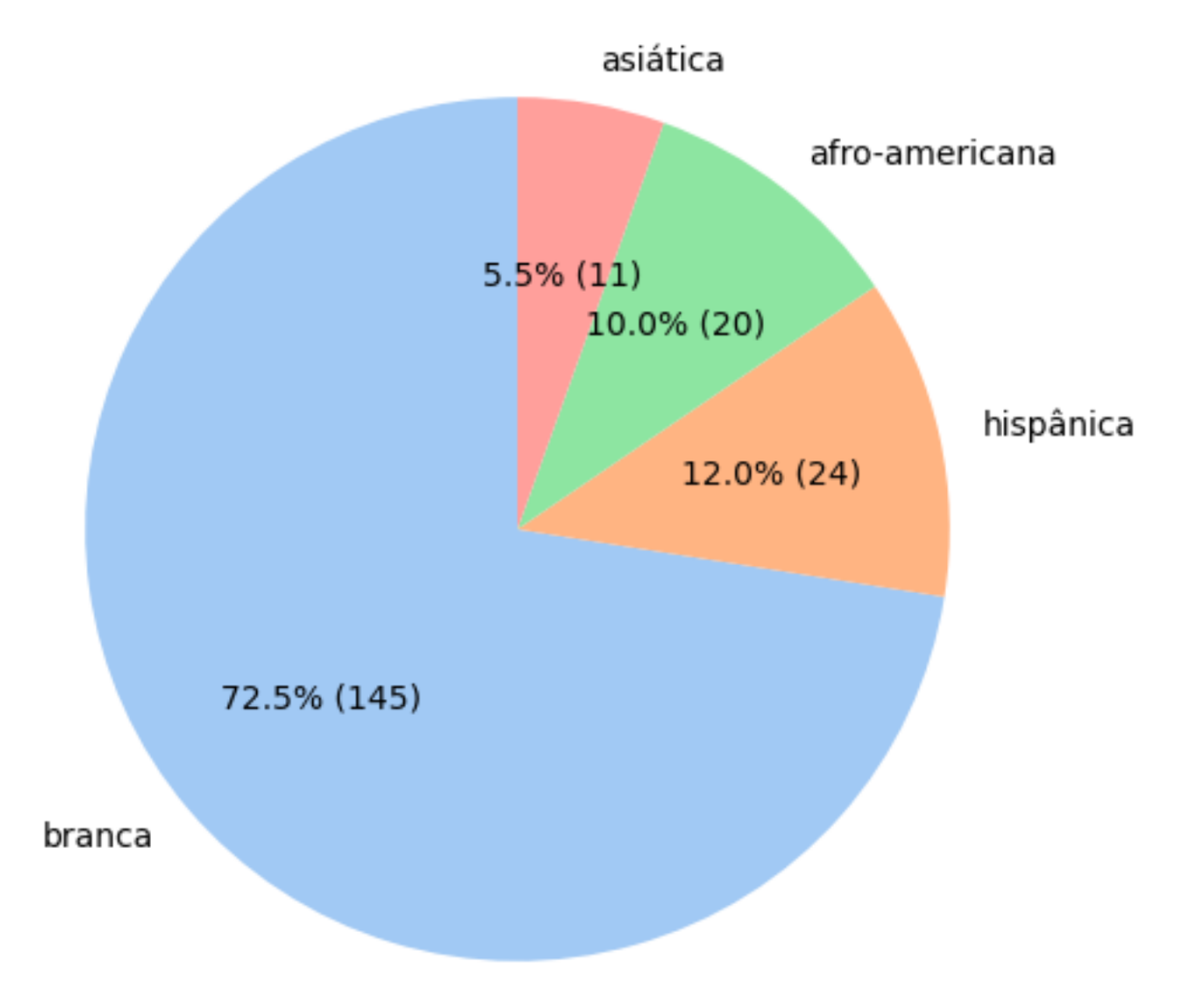


5.1.2. Análise Descritiva da variável raça

O gráfico da **Figura 2** mostra a distribuição das raças presentes na amostra de dados utilizada.

In [54]:
```
plot_categorical(data, 'raca', 'Figura 2. Distribuição de Raça')
```

Figura 2. Distribuição de Raça

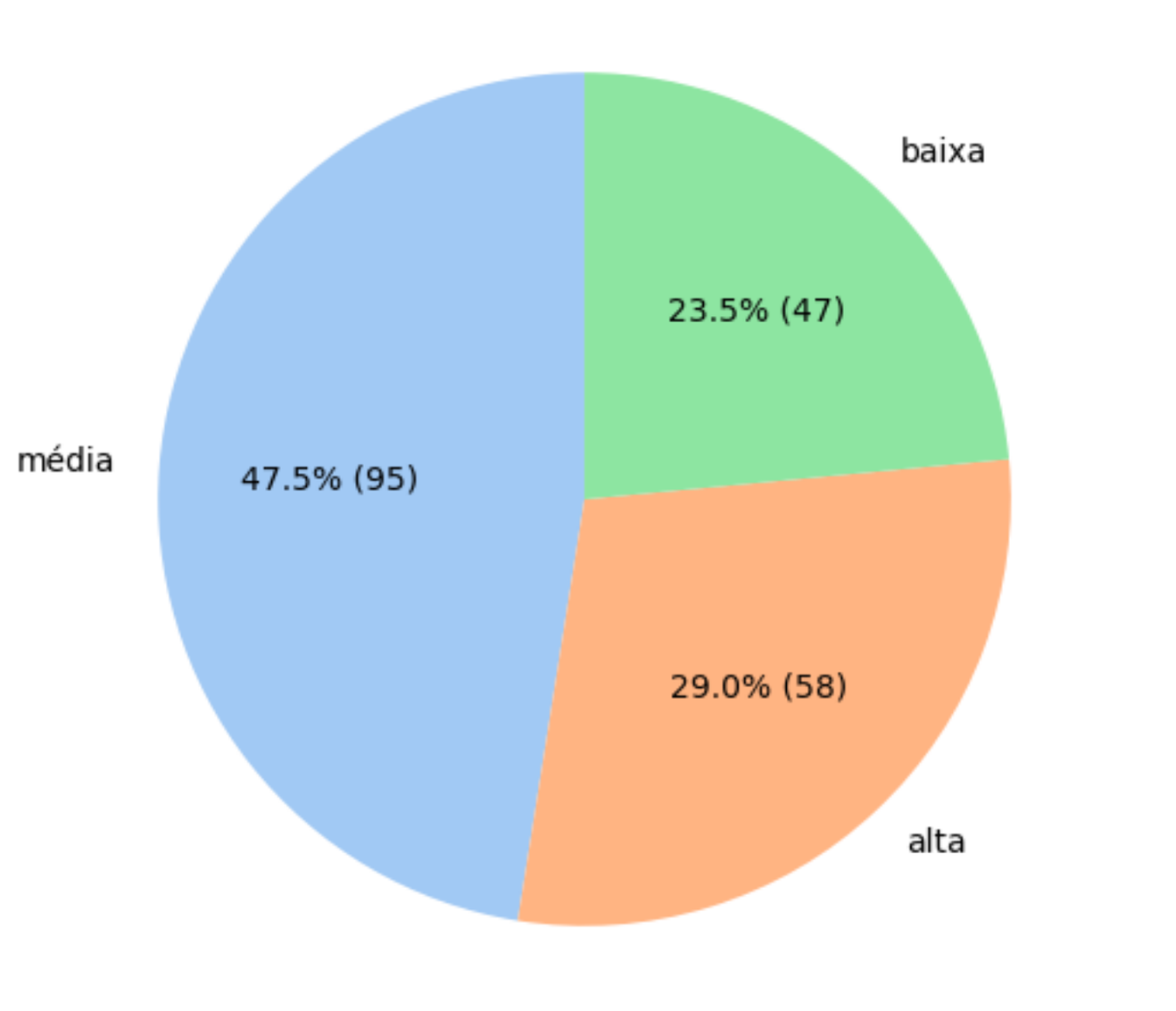


5.1.3. Análise Descritiva da variável raça

O gráfico na **Figura 3** mostra a distribuição das classes sociais na amostra.

In [55]:
```
plot_categorical(data, 'clasocial', 'Figura 3. Distribuição de Classe Social')
```

Figura 3. Distribuição de Classe Social

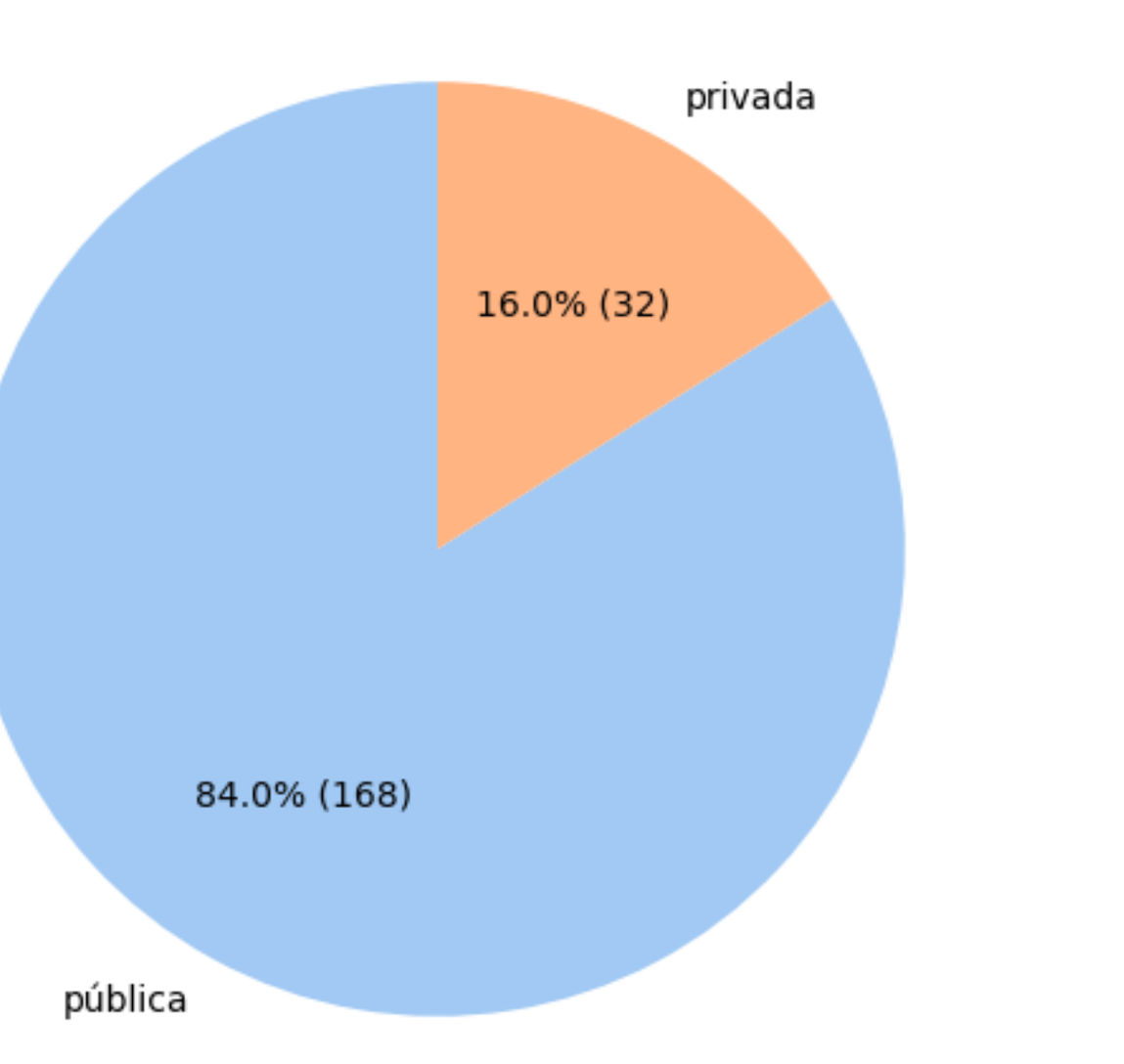


5.1.4. Análise Descritiva da variável tipo de escola

A **Figura 4** representa a distribuição dos tipos de escola da amostra.

In [56]:
```
plot_categorical(data, 'tipescola', 'Figura 4. Distribuição do Tipo de Escola')
```

Figura 4. Distribuição do Tipo de Escola

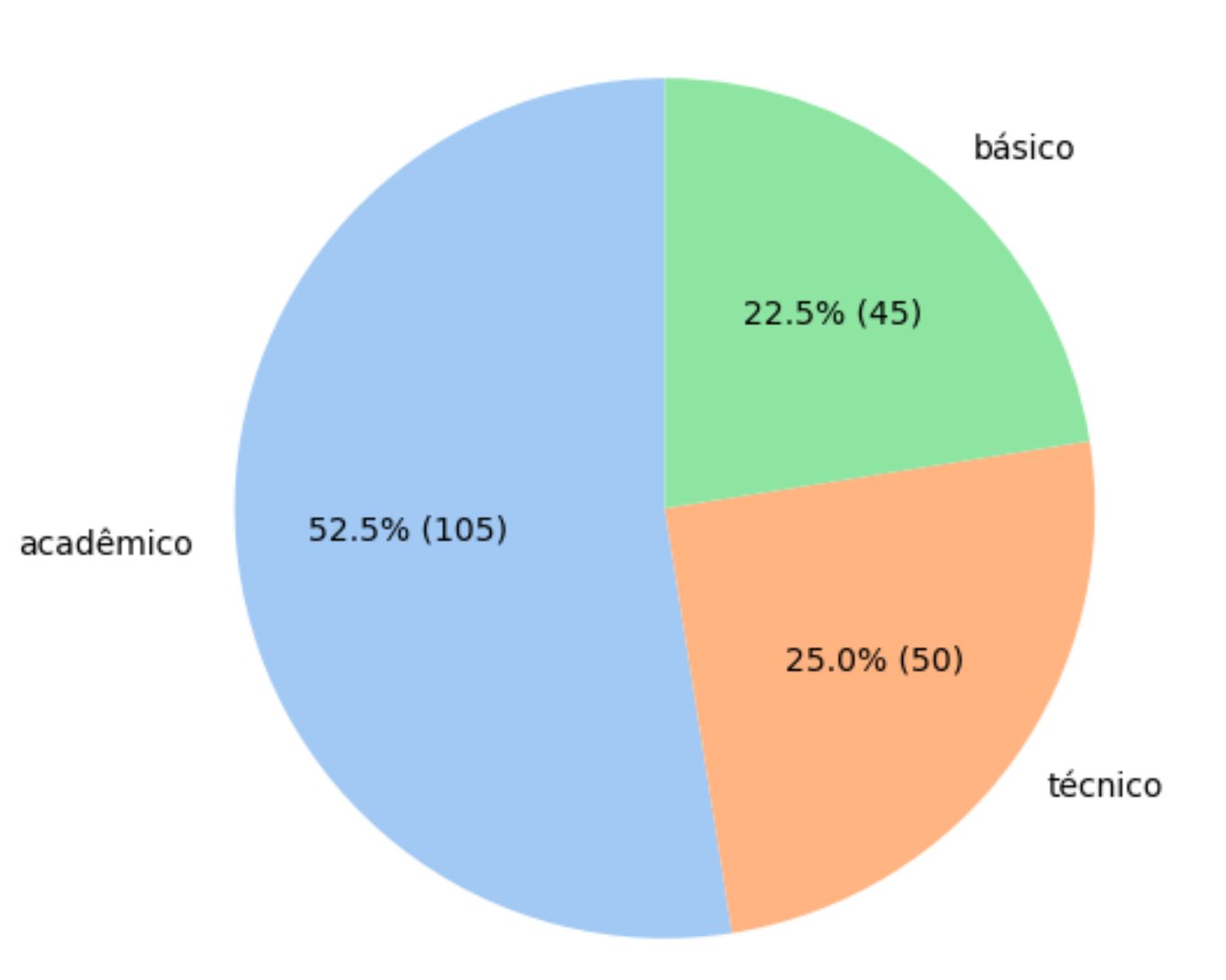


5.1.5. Análise Descritiva da variável tipo de currículo

A **Figura 5** representa a distribuição dos tipos de currículo das escolas na amostra.

In [57]:
```
plot_categorical(data, 'programa', 'Figura 5. Distribuição dos Currículos')
```

Figura 5. Distribuição dos Currículos



5.2. Análise Descritiva das variáveis numéricas

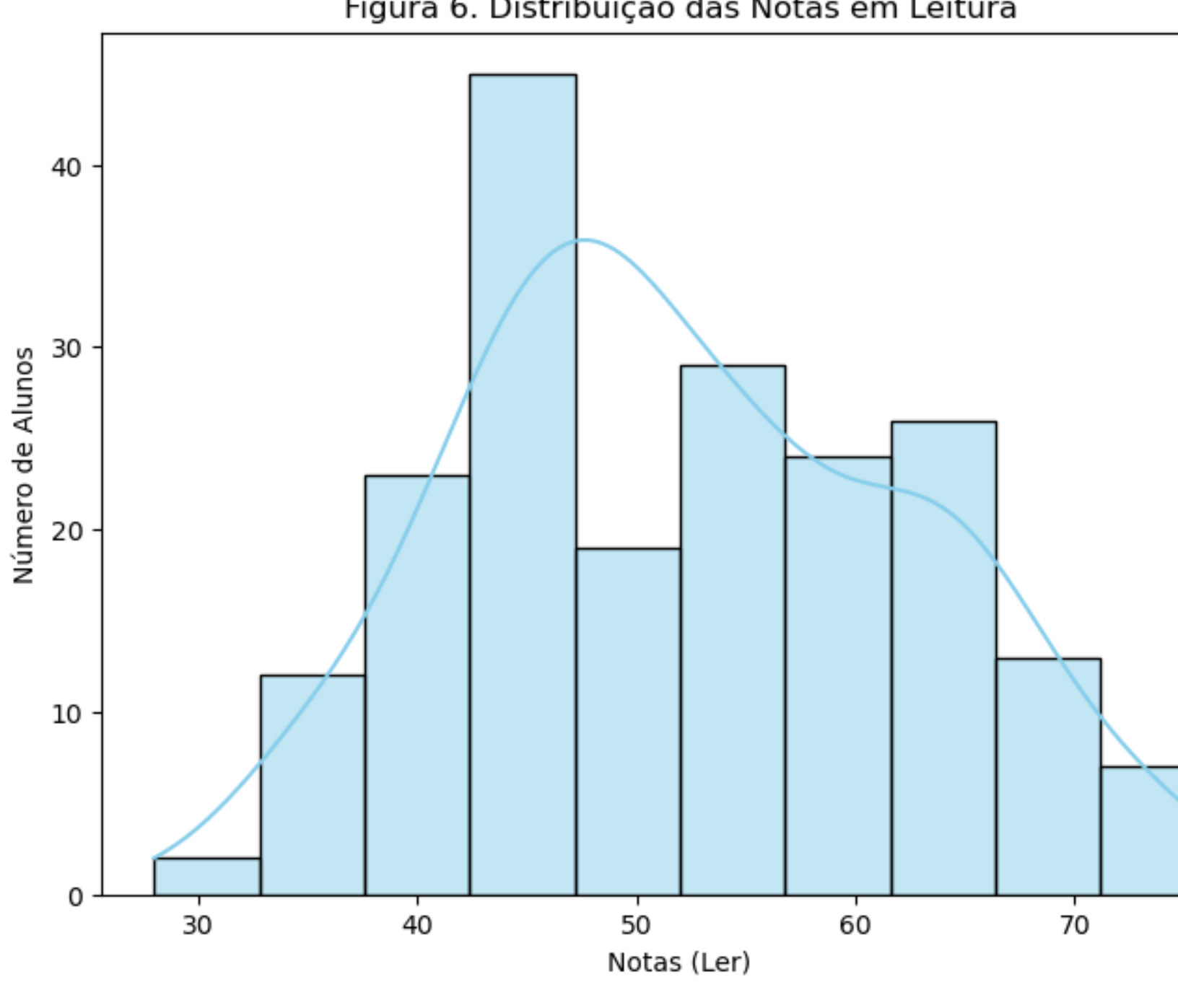
De acord com Pol Mayer (s/d), variáveis quantitativas são aquelas que podem ser medidas numericamente e expressam uma quantidade ou magnitude. Esse tipo de variável pode ser *contínua*, quando assumem valores em um intervalo contínuo (por exemplo: 1.2, 0.5, -3.1), ou **discretas**, quando assumem apenas valores inteiros (por exemplo: 1, -5, 7). Neste trabalho as variáveis *ler*, *escrever*,*matemática*,*ciências* e *estsociais* constituem-se de notas obtidas pelos alunos em cada uma dessas disciplinas e são *variáveis contínuas discretas*.

5.2.1. Análise Descritiva da variável leitura

O gráfico da **Figura 6** mostra a distribuição das notas dos alunos em Leitura.

In [58]:
```
plot_numerical(data, 'ler', 'Figura 6. Distribuição das Notas em Leitura')
```

Figura 6. Distribuição das Notas em Leitura

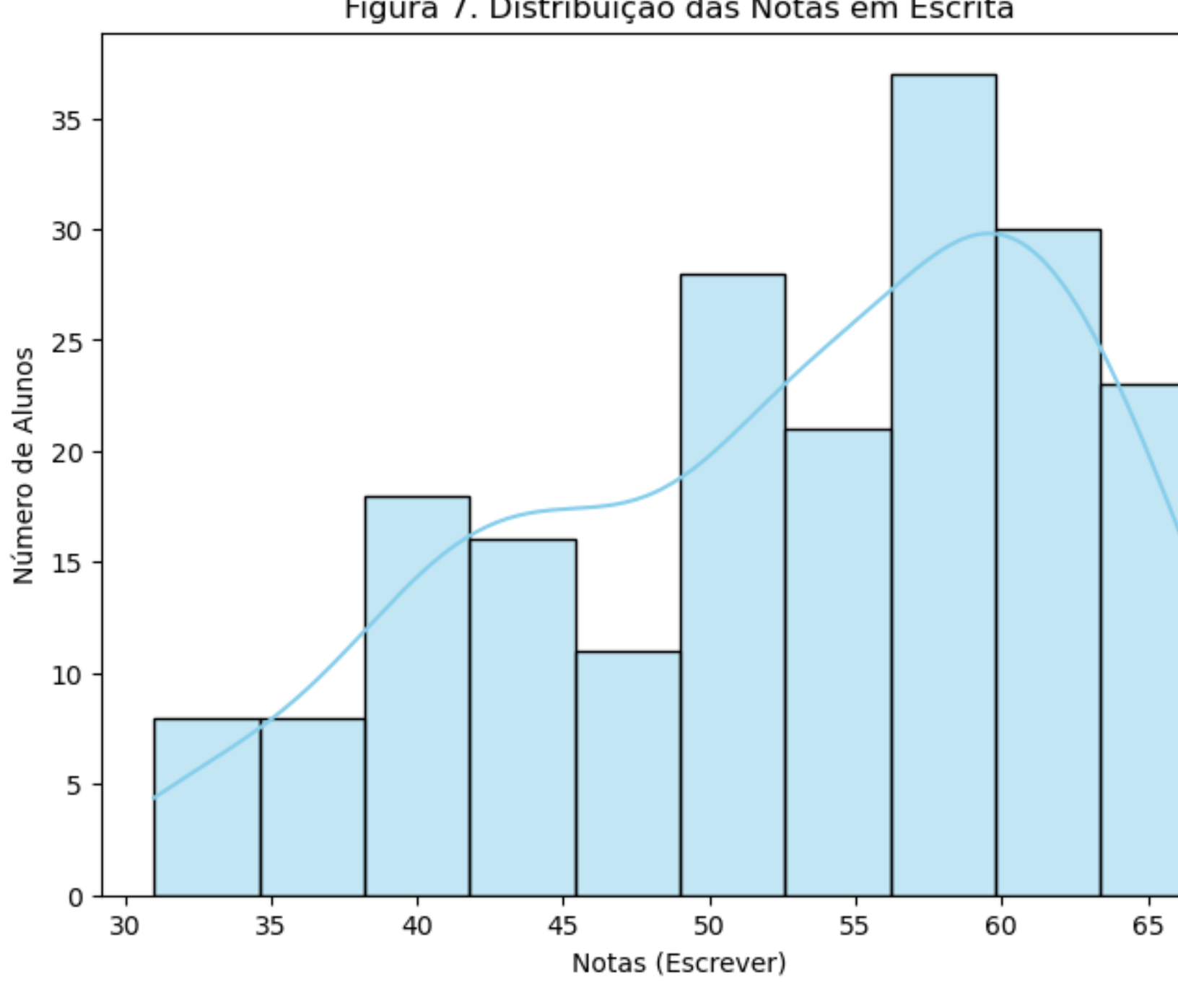


5.2.2. Análise Descritiva da variável escrever

O gráfico da **Figura 7** mostra a distribuição das notas dos alunos em Leitura.

In [59]:
```
plot_numerical(data, 'escrever', 'Figura 7. Distribuição das Notas em Escrita')
```

Figura 7. Distribuição das Notas em Escrita

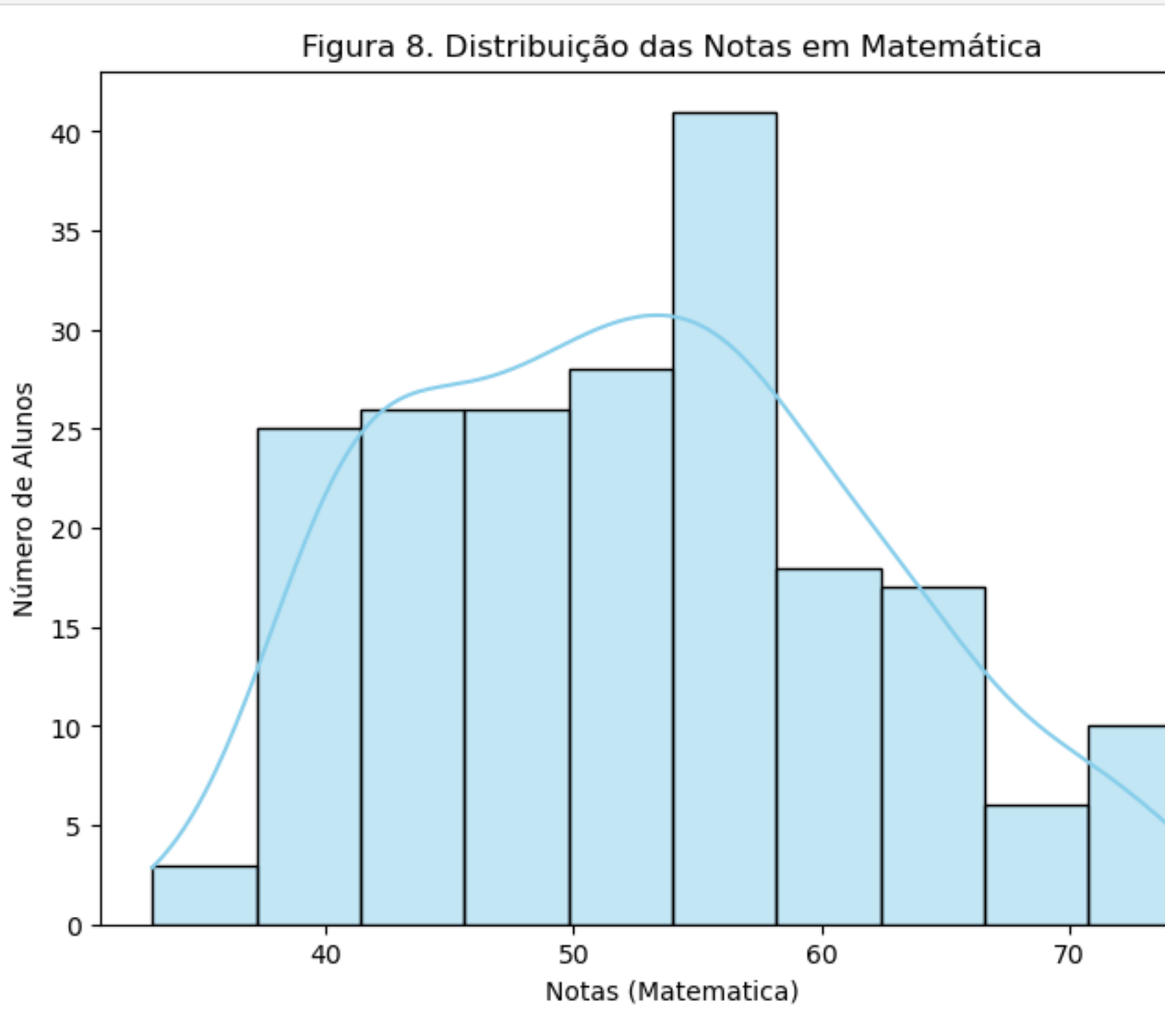


5.2.3. Análise Descritiva da variável matematica

O gráfico da **Figura 8** mostra a distribuição das notas dos alunos em Matemática.

In [60]:
```
plot_numerical(data, 'matematica', 'Figura 8. Distribuição das Notas em Matemática')
```

Figura 8. Distribuição das Notas em Matemática

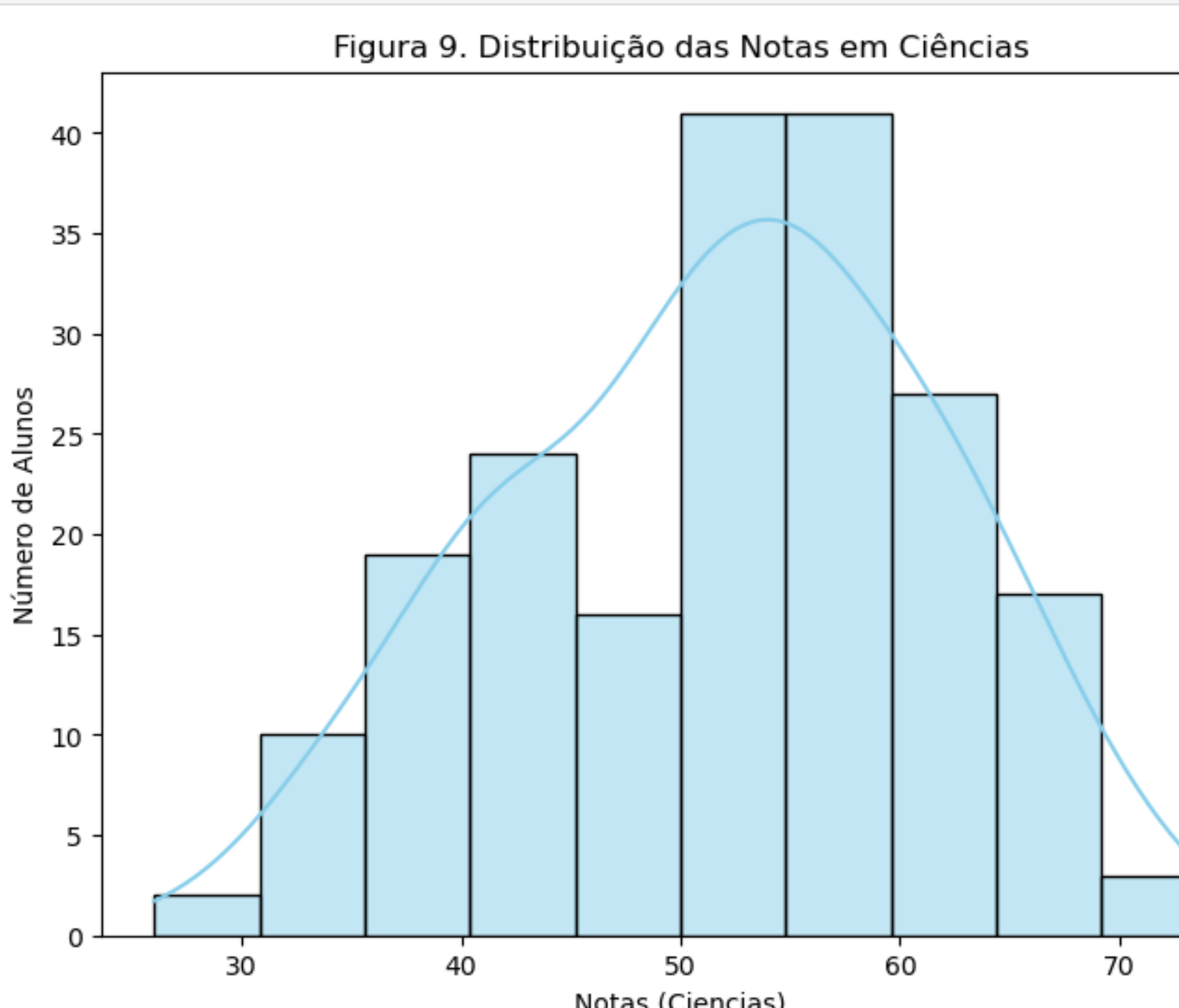


5.2.4. Análise Descritiva da variável ciencias

O gráfico da **Figura 9** mostra a distribuição das notas dos alunos em Ciências.

In [61]:
```
plot_numerical(data, 'ciencias', 'Figura 9. Distribuição das Notas em Ciências')
```

Figura 9. Distribuição das Notas em Ciências

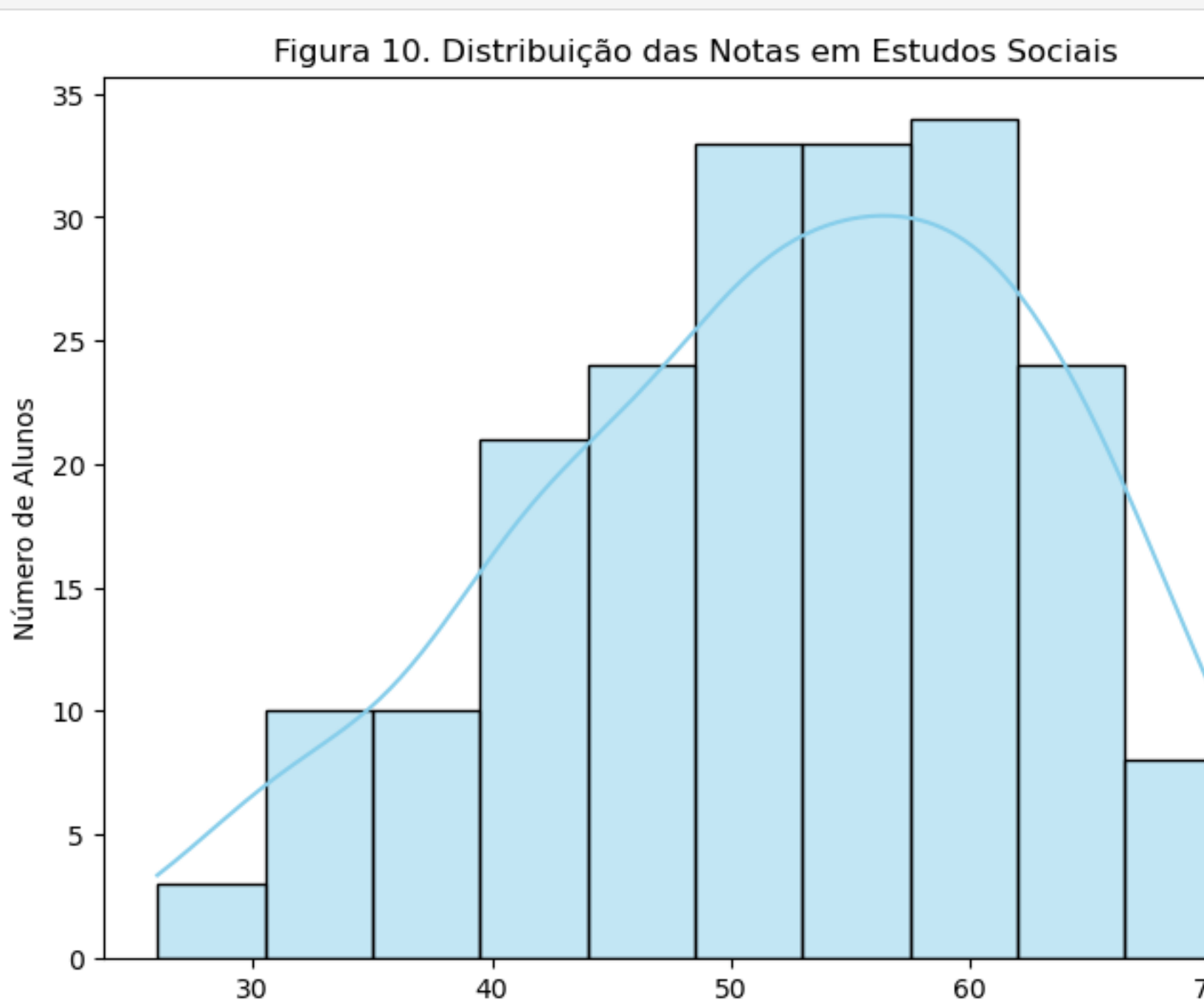


5.2.5. Análise Descritiva da variável estsociais

O gráfico da **Figura 10** mostra a distribuição das notas dos alunos em Estudos Sociais.

In [62]:
```
plot_numerical(data, 'estsociais', 'Figura 10. Distribuição das Notas em Estudos Sociais')
```

Figura 10. Distribuição das Notas em Estudos Sociais



6. Análise de Correlação e Causalidade

A análise de correlação na análise exploratória auxilia o entendimento de como uma variável pode *prever* ou estar *associada* a outra. Mais especificamente, a correlação mostra a direção e a força dessa relação. As correlações podem ser positivas, negativas ou inexistentes. À medida que uma variável aumenta, a outra também tende a aumentar. Em resumo, a correlação é fundamental em análises de dados pois permite identificar relações entre variáveis e, desta forma, compreender como as variáveis interagem pode ajudar a construir modelos preditivos mais precisos.

7. Considerações finais

Referências

ALVES, Ana. Estatística Aplicada: Análise de Dados. Editora Aprender Estatística Fácil, 2022. Mayer, Fernando de Pol. Análise exploratória de dados. Probabilidade e Estatística para Engenharias utilizando o R (RStudio), Universidade Federal de Santa Catarina. Disponível em: https://www.inf.ufsc.br/~andre.zibetti/probabilidade/aed.html. Acesso em: 26 nov. 2024. MÉTRICAS BOSS. Os 4 tipos de análise de dados e como fazê-los. Blog de Web Analytics. 2023. Disponível em: https://metricasboss.com.br/artigos/os-4-tipos-de-analise-de-dados-e-como-faze-los. Acesso em: 26 nov. 2024. OPENINTRO. High School and Beyond survey. s/d. Disponível em: https://www.openintro.org/data/index.php?data=hsb2. Acesso em: 26 nov. 2024. NCES. High School & Beyond, National Center for Educational Studies,US Department of Education. Disponivel em: https://nces.ed.gov/surveys/hsb/surveydesign.asp. Acesso em: 26 nov. 2024. QUESTIONPRO, O que é uma investigação longitudinal? Blog do Software de pesquisa QuestionPro. 2024. Disponível em: https://www.questionpro.com/blog/pt-br/investigacao-longitudinal/#:~:text=O%20que%20%C3%A9%20uma%20investiga%C3%A7%C3%A3o,tempo%2C%20geralmente%20anos%20ou%20d%C3%A9cadas. Acesso em: 24 nov. 2024. SIRIUS. Entenda o que é análise descritiva, quais são os tipos e o passo a passo para fazer uma! Blog Sirius Educação. 07 set. 2022. Disponível em: https://blog.sirius.education/analise-descritiva/#:~:text=A%20an%C3%A1lise%20descritiva%20%C3%A9%20usada,algum%20per%C3%ADodo%20ou%20evento%20espec%C3%ADfico. Acesso em: 26 nov. 2024. TATSUOKA, Maurice M. Análise multivariada: técnicas para pesquisa educacional e psicológica (2ª ed.) Nova York: Macmillan, Apêndice F, pp: 430-442, 1988. UNITED STATES DEPARTMENT OF EDUCATION. Institute of Education Sciences. National Center for Education Statistics. High School and Beyond, 1980: A Longitudinal Survey of Students in the United States. Inter-university Consortium for Political and Social Research, 2006-01-12. Disponível em: https://doi.org/10.3886/ICPSR07896.v2. Acesso em 26 nov. 2024.

In [ ]: