Instituto Federal de Educação, Ciência e Tecnologia de São Paulo (IFSP)

Relatório de Análises Estatísticas, de Probabilidade e de Inferência

Alunos:

- **Igor Ferreira Franco - SP3099695**
- **Gustavo Butigeli Silva - SP3094596**
- **Thiago Marcio Barone - SP3110826**
- **Vitor Soares Coelho - SP3093174**

SÃO PAULO, 17 DE JUNHO DE 2024

Análises da base de dados

Sistema de Informação sobre Mortalidade – SIM

O Sistema de Informação sobre Mortalidade (SIM), desenvolvido pelo Ministério da Saúde em 1975, é produto da unificação de mais de quarenta modelos de Declaração de Óbito utilizados ao longo dos anos, para coletar dados sobre mortalidade no país.

Com sua longa série temporal, o SIM é um patrimônio nacional, visto que possui informações fundamentais para que possamos conhecer os aspectos referentes à mortalidade no Brasil e às causas de adoecimento que levaram ao óbito. É, ainda, um dos principais instrumentos para apoiar a elaboração de políticas públicas de saúde e seguridade social mais efetivas visando à prevenção, promoção e cuidado em saúde.

Mortalidade Geral 2023 - prévia
link para a base de dados: Sistema de Informação sobre Mortalidade – SIM - Mortalidade Geral 2023 - OPEN DATASUS
GitHub: https://github.com/Igor-Franco/ESP1A5

O arquivo CSV tem 1.485.653 linhas e 86 colunas.

Nomes das colunas:
Index(['contador', 'ORIGEM', 'TIPOBITO', 'DTOBITO', 'HORAOBITO', 'NATURAL', 'CODMUNNATU', 'DTNASC', 'IDADE', 'SEXO', 'RACACOR', 'ESTCIV', 'ESC', 'ESC2010', 'SERIESCFAL', 'OCUP', 'CODMUNRES', 'LOCOCOR', 'CODESTAB', 'CODMUNOCOR', 'IDADEMAE', 'ESCMAE', 'ESCMAE2010', 'SERIESCMAE' 'OCUPMAE', 'QTDFILVIVO', 'QTDFILMORT', 'GRAVIDEZ', 'SEMAGESTAC', 'GESTACAO', 'PARTO', 'OBITOPARTO', 'PESO', 'TPMORTEOCO', 'OBITOGRAV', 'OBITOPUERP', 'ASSISTMED', 'EXAME', 'CIRURGIA', 'NECROPSIA', 'LINHAA', 'LINHAB', 'LINHAC', 'LINHAD', 'LINHAII', 'CAUSABAS', 'CB_PRE', 'COMUNSVOIM', 'DTATESTADO', 'CIRCOBITO', 'ACIDTRAB', 'FONTE', 'NUMEROLOTE', 'DTINVESTIG', 'DTCADASTRO', 'ATESTANTE', 'STCODIFICA', 'CODIFICADO', 'VERSAOSIST', 'VERSAOSCB', 'FONTEINV', 'DTRECEBIM' 'ATESTADO', 'DTRECORIGA', 'OPOR_DO', 'CAUSAMAT', 'ESCMAEAGR1', 'ESCFALAGR1', 'STDOEPIDEM', 'STDONOVA', 'DIFDATA', 'NUDIASOBCO', 'DTCADINV', 'TPOBITOCOR', 'DTCONINV', 'FONTES', 'TPRESGINFO', 'TPNIVELINV', 'DTCADINF', 'MORTEPARTO', 'DTCONCASO', 'ALTCAUSA', 'CAUSABAS_O', 'TPPOS', 'TP_ALTERA', 'CB_ALT'],

Ferramentas para análises dos Óbitos em 2023

Para os óbitos em 2023 foi utilizado análises em python e seguindo a estrutura SIM disponibilizada no site opendadasus. Foram utilizadas uma série de bibliotecas com funções específicas para análise de dados que serão descritas a seguir:

1. Pandas

Principais Funções:

pd.DataFrame(): Cria um DataFrame.
pd.read_csv(): Lê um arquivo CSV.
df.head(): Mostra as primeiras linhas do DataFrame.
df.describe(): Gera estatísticas descritivas.
df.groupby(): Agrupa dados.
df.merge(): Mescla DataFrames.
df.pivot_table(): Cria tabelas dinâmicas.

2. NumPy

Principais Funções:

np.array(): Cria arrays.
np.mean(): Calcula a média dos elementos.
np.median(): Calcula a mediana.
np.std(): Calcula o desvio padrão.
np.sum(): Soma dos elementos.
np.arange(): Cria uma sequência de números.
np.linspace(): Cria uma sequência de números espaçados uniformemente.

3. Matplotlib

Principais Funções:

plt.plot(): Cria gráficos de linha.
plt.scatter(): Cria gráficos de dispersão.
plt.bar(): Cria gráficos de barras.
plt.hist(): Cria histogramas.
plt.xlabel(): Define o rótulo do eixo x.
plt.ylabel(): Define o rótulo do eixo y.
plt.title(): Define o título do gráfico.

4. SciPy

Principais Funções:

scipy.stats.ttest_ind(): Teste t para a média de duas amostras independentes.
scipy.optimize.minimize(): Otimização de funções.
scipy.integrate.quad(): Integração de funções.
scipy.linalg.inv(): Inversão de matrizes.
scipy.fft.fft(): Transformada de Fourier.
scipy.stats.shapiro(): Teste de Shapiro-Wilk para normalidade.
scipy.stats.kstest(): Teste de Kolmogorov-Smirnov para uma amostra.

Para a análises utilizando a IDADE foi usada a seguinte estrutura para determinar o idade em anos:

- a idade do falecido em minutos, horas, dias, meses ou anos. (Idade: composto de dois subcampos. - O primeiro, de 1 dígito, indica a unidade da idade (se 1 = minuto, se 2 = hora, se 3 = mês, se 4 = ano, se = 5 idade maior que 100 anos). - O segundo, de dois dígitos, indica a quantidade de unidades: Idade menor de 1 hora: subcampo varia de 01 e 59 (minutos); De 1 a 23 Horas: subcampo varia de 01 a 23 (horas); De 24 horas e 29 dias: subcampo varia de 01 a 29 (dias); De 1 a menos de 12 meses

completos: subcampo varia de 01 a 11 (meses); Anos - subcampo varia de 00 a 99; -9 - ignorado)

Análises Descritivas

Média da idade: 66.69 anos

- A média indica que a idade média dos óbitos é de aproximadamente 66.69 anos. Isso é consistente com a observação de que a maioria das mortes ocorre em idades mais avançadas.

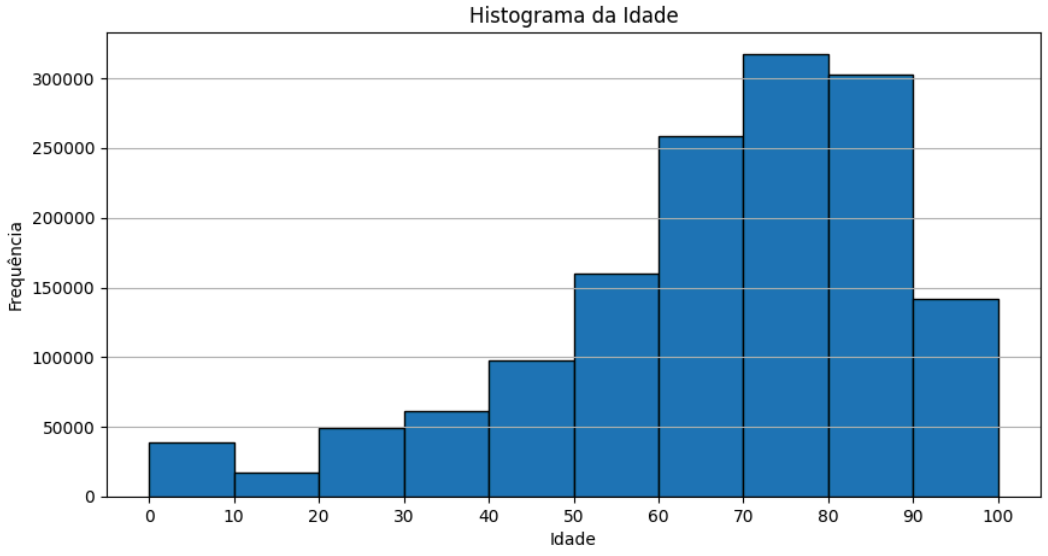
Mediana da idade: 71.00 anos

- A mediana, sendo maior que a média, confirma que a distribuição é assimétrica à direita. Isso significa que há mais óbitos em idades avançadas, mas com uma cauda longa de idades menores.

Desvio padrão da idade: 21.42 anos

- O desvio padrão é relativamente grande, indicando uma ampla dispersão das idades dos óbitos em torno da média. Isso também se reflete no boxplot, onde podemos ver uma distribuição ampla das idades.

Idade mínima: 0.00 anos
Idade máxima: 123.00 anos



Distribuição Assimétrica:

- A distribuição das idades dos óbitos é assimétrica à direita. Isso é consistente com a média sendo menor que a mediana.

Pico de Frequência:

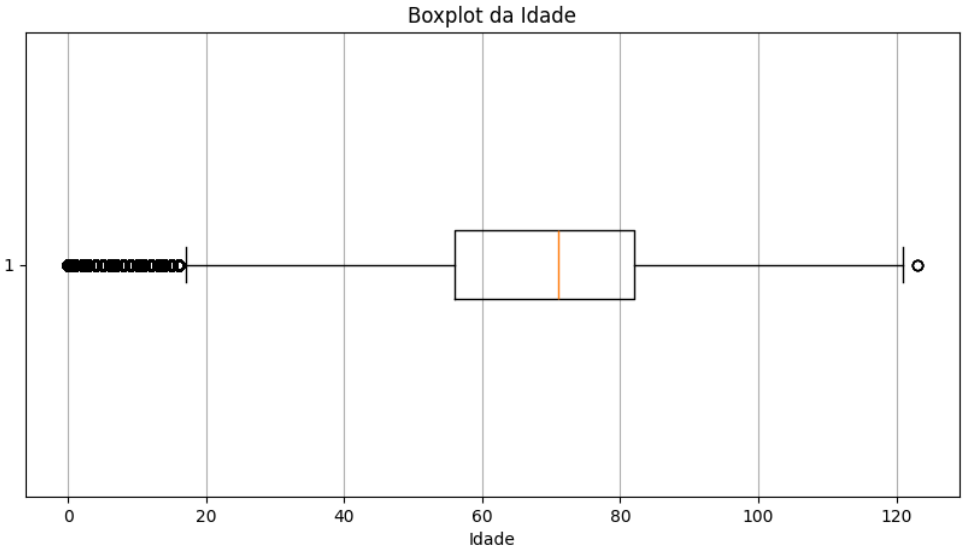
- O pico de frequência de óbitos está entre 70 e 80 anos, indicando que a maioria dos óbitos ocorrem nessa faixa etária.

Óbitos em Idades Jovens:

- As faixas etárias mais jovens (0-20 anos) têm uma frequência menor de óbitos, mas ainda há uma quantidade notável de mortes. Esses dados são consistentes com os outliers no boxplot.

Idades Avançadas:

- Um número considerável de óbitos ocorre em idades superiores a 90 anos, embora a frequência diminua em idades extremamente avançadas (100+ anos).



Quartis:

- A mediana (linha central do boxplot) está em 71 anos, o que está alinhado com a mediana fornecida.
- O primeiro quartil (Q1) e o terceiro quartil (Q3) parecem estar em torno de 50 e 80 anos, respectivamente, o que indica que 50% dos dados estão entre essas idades.

Intervalo Interquartil (IQR):

- O IQR (Q3 - Q1) é de aproximadamente 30 anos (80 - 50). Isso representa a faixa central de 50% das idades dos óbitos.

Outliers:

- Há muitos outliers abaixo de 50 anos, indicando uma ocorrência significativa de mortes em idades muito jovens.

- Há também alguns outliers acima de 100 anos, indicando casos de longevidade extrema.

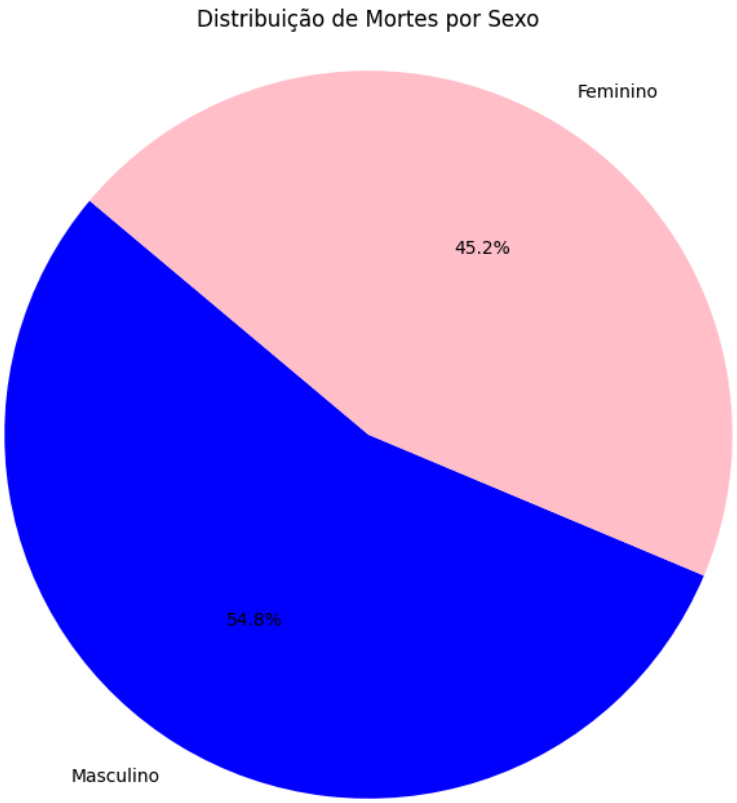
Bigodes:

- Os "bigodes" do boxplot se estendem desde os limites do IQR até os valores máximos e mínimos dentro de 1.5 vezes o IQR. Os valores fora deste intervalo são considerados outliers.

Análise Comparativa

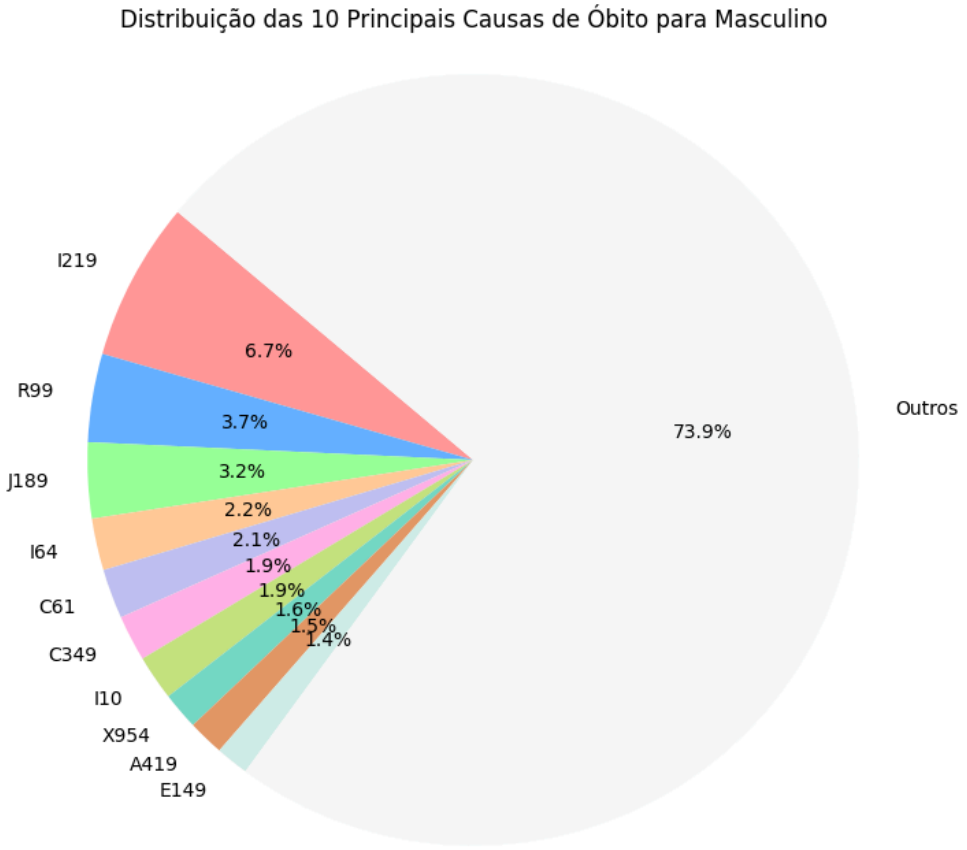
Distribuição de Mortes pelo Sexo

- Número total de registros: 1.485.653
- Contagem de mortes por sexo: SEXO
- Masculino: 813.123
- Feminino: 671.195



Valores absolutos das 10 principais causas de óbito para Masculino:

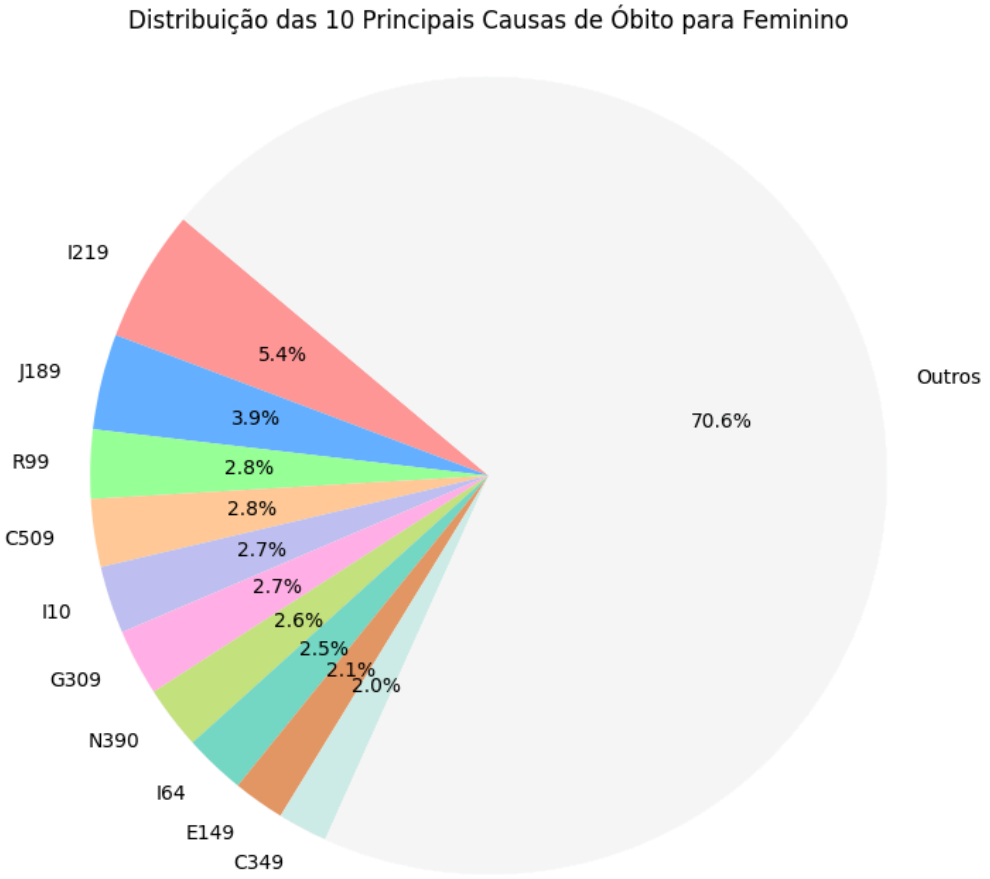
1. **I219** (Infarto Agudo do Miocárdio): 54.272
2. **R99** (Outras Causas de Morte Mal Especificadas): 30.074
3. **J189** (Pneumonia, não especificada): 25.783
4. **I64** (Acidente Vascular Cerebral): 17.688
5. **C61** (Neoplasia Maligna da Próstata): 16.897
6. **C349** (Neoplasia Maligna dos Brônquios e do Pulmão): 15.661
7. **I10** (Hipertensão Essencial [Primária]): 15.321
8. **X954** (Agressão por Arma de Fogo): 12.698
9. **A419** (Septicemia, não especificada): 12.268
10. **E149** (Diabetes Mellitus, não especificado): 11.209
11. **Outros**: 601.252



Valores absolutos das 10 principais causas de óbito para Feminino:

1. **I219** (Infarto Agudo do Miocárdio): 36.092
2. **J189** (Pneumonia, não especificada): 25.998

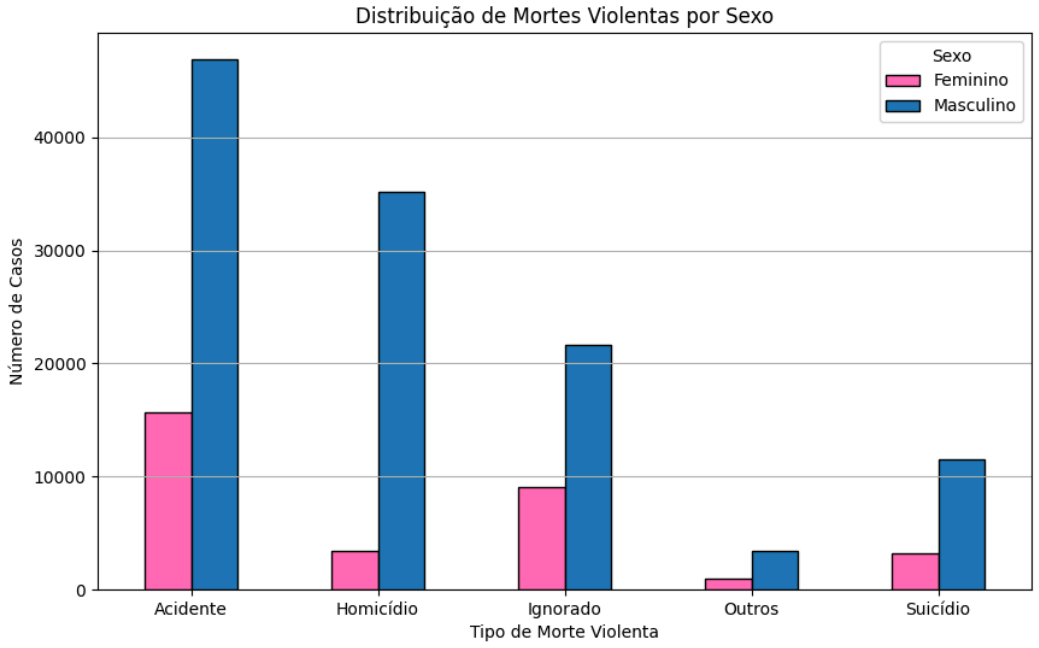
3. **R99** (Outras Causas de Morte Mal Especificadas): 18.734
4. **C509** (Neoplasia Maligna da Mama): 18.480
5. **I10** (Hipertensão Essencial [Primária]): 18.341
6. **G309** (Doença de Alzheimer, não especificada): 18.125
7. **N390** (Infecção do Trato Urinário, não especificada): 17.136
8. **I64** (Acidente Vascular Cerebral): 16.736
9. **E149** (Diabetes Mellitus, não especificado): 14.029
10. **C349** (Neoplasia Maligna dos Brônquios e do Pulmão): 13.527
11. **Outros**: 473.997



Valores absolutos das mortes violentas (CIRCOBITO) por sexo:

Tipo de Óbito	Feminino	Masculino

Acidente	15.697	46.902
Homicídio	3.399	35.204
Ignorado	9.026	21.632
Outros	974	3.445
Suicídio	3.173	11.484
Total	32.269	118.667

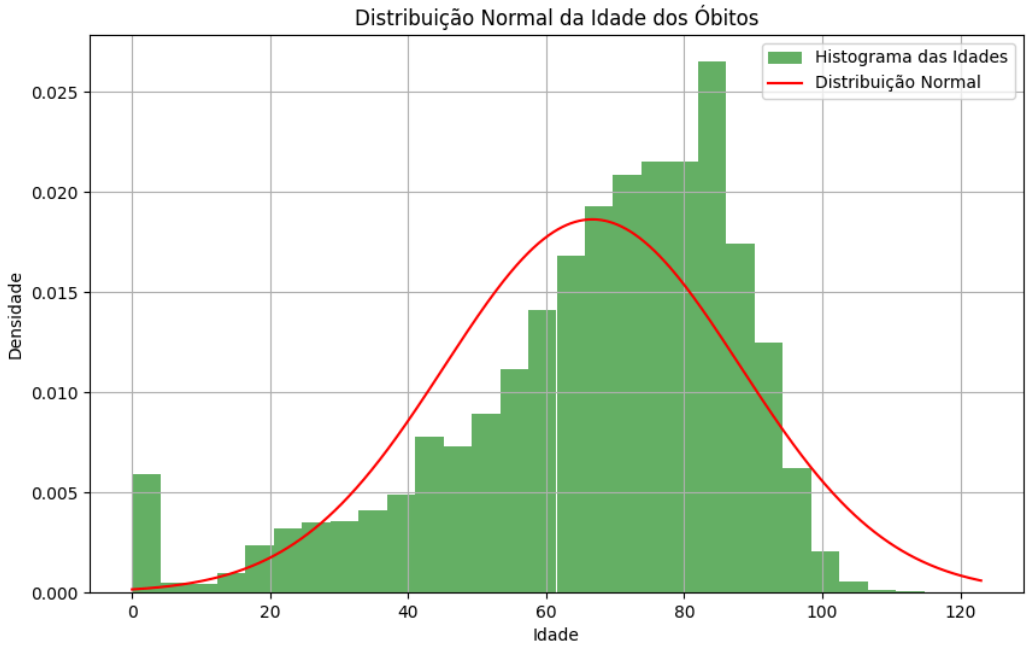


Análises de Probabilidade

- Probabilidade de ter falecido de Infarto Agudo do Miocárdio (I219) dado que é homem: **0.0681**
- Probabilidade de ter falecido de Infarto Agudo do Miocárdio (I219) dado que é mulher: **0.0549**
- Probabilidade de ter falecido de Pneumonia (J189) dado que é mulher: **0.0395**
- Probabilidade de ter falecido de Pneumonia (J189) dado que é homem: **0.0323**
- Probabilidade de ter falecido de Hipertensão Essencial (Primária) (I10) dado que é mulher: **0.0279**
- Probabilidade de ter falecido de Hipertensão Essencial (Primária) (I10) dado que é homem: **0.0192**

- Probabilidade de ter falecido de Neoplasia Maligna da Próstata (C61) dado que é mulher: **0.0000**
- Probabilidade de ter falecido de Neoplasia Maligna da Próstata (C61) dado que é homem: **0.0212**
- Probabilidade de ter falecido de Neoplasia Maligna dos Brônquios e do Pulmão (C349) dado que é mulher: **0.0206**
- Probabilidade de ter falecido de Neoplasia Maligna dos Brônquios e do Pulmão (C349) dado que é homem: **0.0197**
- Probabilidade de uma pessoa de 60 anos ou mais ter falecido de Pneumonia (J189): **0.0427**

Distribuição de Probabilidade - Normal



Teste de Shapiro-Wilk:

- estatística=0.9247798953905576,
- p-valor=1.5381085258569183e-44

Teste de Kolmogorov-Smirnov:

- estatística=0.09802184524061208,
- p-valor=2.861755255746507e-42

Interpretação:

No caso do Shapiro-Wilk:

O p-valor é extremamente pequeno (muito menor que 0.05) Isso sugere que os dados não seguem uma distribuição normal.

No caso do Kolmogorov-Smirnov:

O p-valor é novamente extremamente pequeno (muito menor que 0.05), o que nos leva a rejeitar a hipótese nula. Isso sugere que os dados não seguem uma distribuição normal.

Ambos os testes de normalidade (Shapiro-Wilk e Kolmogorov-Smirnov) indicam que a idade dos óbitos não segue uma distribuição normal. O gráfico (histograma com curva de densidade) também pode fornecer uma confirmação visual dessa conclusão.Portanto, baseado nesses resultados, podemos afirmar que a distribuição das idades dos óbitos não é normal.

Inferência Estatística - Intervalo de Confiança e Teste de Hipóteses

Foi selecionada uma amostra de 5000 dados para analisar a idade dos óbitos. O desvio padrão das idades na amostra foi calculado como 21.41 anos, indicando uma variação significativa das idades em relação à média.

Hipótese de Teste

Foi formulada a seguinte hipótese para o teste t para uma amostra:

- **Hipótese Nula (H0)**: A média das idades dos óbitos é 70 anos.
- **Hipótese Alternativa (H1)**: A média das idades dos óbitos não é 70 anos.

Resultados dos Testes

Valor Crítico para um Intervalo de Confiança de 95%:

- O valor crítico para um intervalo de confiança de 95% foi calculado como 1.9604386466615242.

Intervalo de Confiança de 95% para a Média da Idade (Amostra):

- O intervalo de confiança para a média das idades na amostra foi [65.80734330208722, 66.99473896956843]. Isso indica que estamos 95% confiantes de que a média real da idade dos óbitos na população está entre 65.81 e 66.99 anos.

Teste de Shapiro-Wilk:

- Estatística: 0.9247798953905576
- p-valor: 1.5381085258569183e-44

O p-valor extremamente baixo indica que rejeitamos a hipótese nula de que os dados seguem uma distribuição normal. Isso sugere fortemente que a distribuição das idades dos óbitos na amostra não é normal.

Teste de Kolmogorov-Smirnov:

- Estatística: 0.10331721342568079
- p-valor: 6.36734421429003e-47

O p-valor muito baixo confirma ainda mais que a distribuição das idades dos óbitos na amostra não é normal.

Teste t para uma Amostra:

- Estatística t: -11.884055565125767
- p-valor: 3.866604648514155e-32

O p-valor extremamente baixo indica que rejeitamos a hipótese nula de que a média das idades dos óbitos é 70 anos. A estatística **t** negativa sugere que a média observada na amostra (66.40 anos) é significativamente menor do que 70 anos.

Conclusão

Os resultados dos testes estatísticos indicam que a média das idades dos óbitos na amostra é significativamente diferente de 70 anos e, na verdade, é menor do que 70 anos. Além disso, os testes de normalidade mostram que a distribuição das idades dos óbitos na amostra não segue uma distribuição normal. Essas conclusões são baseadas em uma amostra representativa de 5000 dados, proporcionando uma inferência estatisticamente robusta sobre a população maior.

Apêndice

README

Projeto de ESP1A5 do IFSP

Alunos: Igor Ferreira Franco - SP3099695
Gustavo Butigeli Silva - SP3094596
Thiago Marcio Barone - SP3110826
Vitor Soares Coelho - SP3093174

Clone o repositório: git clone https://github.com/Igor-Franco/ESP1A5.git

link para a base de dados: Sistema de Informação sobre Mortalidade – SIM - Mortalidade Geral 2023 - OPEN DATASUS

Bibliotecas utilizadas

1. Pandas

Instalação: pip install pandas

Principais Funções:
- pd.DataFrame(): Cria um DataFrame.
- pd.read_csv(): Lê um arquivo CSV.
- df.head(): Mostra as primeiras linhas do DataFrame.
- df.describe(): Gera estatísticas descritivas.
- df.groupby(): Agrupa dados.
- df.merge(): Mescla DataFrames.
- df.pivot_table(): Cria tabelas dinâmicas.

2. NumPy

Instalação: pip install numpy

Principais Funções:
- np.array(): Cria arrays.
- np.mean(): Calcula a média dos elementos.
- np.median(): Calcula a mediana.
- np.std(): Calcula o desvio padrão.
- np.sum(): Soma dos elementos.
- np.arange(): Cria uma sequência de números.
- np.linspace(): Cria uma sequência de números espaçados uniformemente.

3. Matplotlib

Instalação: pip install matplotlib

Principais Funções:
- plt.plot(): Cria gráficos de linha.
- plt.scatter(): Cria gráficos de dispersão.
- plt.bar(): Cria gráficos de barras.
- plt.hist(): Cria histogramas.
- plt.xlabel(): Define o rótulo do eixo x.
- plt.ylabel(): Define o rótulo do eixo y.
- plt.title(): Define o título do gráfico.

4. SciPy

Instalação: pip install scipy

Principais Funções:
- scipy.stats.ttest_ind(): Teste t para a média de duas amostras independentes.
- scipy.optimize.minimize(): Otimização de funções.
- scipy.integrate.quad(): Integração de funções.
- scipy.linalg.inv(): Inversão de matrizes.
- scipy.fft.fft(): Transformada de Fourier.
- scipy.stats.shapiro(): Teste de Shapiro-Wilk para normalidade.
- scipy.stats.kstest(): Teste de Kolmogorov-Smirnov para uma amostra.