UNIVERSIDADE FEDERAL
DE SERGIPE

ESTATÍSTICA APLICADA

GUILHERME MENEZES DE AZEVEDO

RELATÓRIO

Análise estatístico do sangue de pacientes com Câncer

Aracaju, Sergipe

15/01/2024

1. Introdução

O câncer é uma condição complexa e multifacetada que continua a desafiar a comunidade médica em sua compreensão, diagnóstico e tratamento. Este relatório apresenta uma análise detalhada de um banco de dados específico [1], contendo informações relevantes sobre pacientes diagnosticados com câncer. Os dados aqui reunidos representam um esforço para compreender melhor a relação entre diferentes parâmetros e as características dos pacientes afetados por essa condição.

Neste estudo, são analisados diversos marcadores bioquímicos e características demográficas dos pacientes, como idade, níveis séricos de Espectro químico da análise do sangue-alkalino phosphatose (AKP), Concentração de fosfato no sangue (P), Enzima lactante de-hydrogenase (LDH), Albumina (ALB), bem como informações relacionadas Nitrogênio na uréia (N) e dados glicêmicos (GL). A análise desses parâmetros busca identificar possíveis padrões, correlações ou insights relevantes que possam contribuir para uma compreensão mais abrangente do cenário oncológico.

A busca por padrões e associações dentro desses dados pode oferecer uma base sólida para aprimorar estratégias de diagnóstico precoce, tratamento personalizado e prognósticos mais precisos para pacientes afetados por diferentes formas de câncer. [2]

Este relatório visa fornecer uma análise inicial dos dados, explorando tendências e relações entre variáveis. Desta forma buscamos neste trabalho avaliar o banco de dados utilizando as técnicas expostas em sala de aula. Assim, nesta atividade proposta como avaliação da Unidade I da disciplina de Estatística aplicada será avaliado o banco de dados de Câncer.

2. Objetivo

O propósito deste relatório é apresentar uma análise estatística do banco de dados selecionado referente à distribuição da análise sanguínea dos pacientes. Os grupos com valores qualitativos são definidos da seguinte maneira: 1 para Falso-negativo, representando diagnósticos incorretos de ausência da doença quando, de fato, os pacientes a possuíam; 2 para negativo, indicando diagnósticos corretos de ausência da doença; 3 para positivo, representando diagnósticos corretos de presença da doença; e 4 para Falso-positivo, indicando diagnósticos incorretos de presença da doença quando, na verdade, os pacientes não a tinham.

Com base nisso, será possível construir uma Distribuição de Frequência para a Variável Qualitativa presente no banco de dados, assim como para a variável quantitativa. Além disso, serão calculadas Medidas Descritivas de Posição, como Média, Mediana, Moda, bem como Medidas de Dispersão, como Variância, Desvio Padrão e Coeficiente de Variação. Também serão gerados gráficos para representar visualmente esses dados e realizar uma análise dos dados. [3]

3. Metodologia

A análise estatística será conduzida por meio da elaboração de algoritmos em Python, utilizando bibliotecas específicas como Pandas e Numpy. Esses algoritmos serão empregados na construção e análise da distribuição de frequências de valores qualitativos e quantitativos. Além disso, o Excel será utilizado para criar tabelas e gráficos que complementarão a visualização e compreensão dos resultados obtidos.

4. Distribuição de Frequência da Variável Qualitativa Grupos

Para relembrar, neste banco de dados apresenta-se as seguintes Classes, 1 para **falso-negativo**, representando diagnósticos incorretos de ausência da doença quando, de fato, os pacientes a possuíam; 2 para **negativo**, indicando diagnósticos corretos de ausência da doença; 3 para **positivo**, representando diagnósticos corretos de presença da doença; e 4 para **falso-positivo**, indicando diagnósticos incorretos de presença da doença quando, na verdade, os pacientes não a tinham.

Utilizamos um código em Python usando o Pandas e Numpy para isolar a coluna dos Grupos com suas 362 linhas respectivas. (Imagem do código abaixo)

Imagem 1: Código em Python

```python
import pandas as pd
import numpy as np

# Importa o banco de dados para ser lido pelo Pandas
caminho_arquivo = "C:\\Users\\guilh\\OneDrive\\Documents\\4_Periodo_UFS\\ESTATISTICA\\cancer_formatados.csv"

# O Pandas analisa todo o arquivo csv e retira os espaços em branco.
base = pd.read_csv(caminho_arquivo, sep=';', converters={
                   'Grupo': lambda x: str(x).replace('\u200b', '')})

# Printa o shape: (362,9)
print(base.shape)

# Separa a coluna desejada das demais
grupo_array = base['Grupo'].values

# Printa o shape da coluna Grupo: (362,)
print(grupo_array.shape)

# Faz um reshape para reconhecer ela como apenas uma coluna
grupo_array_reshape = grupo_array.reshape(-1, 1)
# Printa o shape corrigido: (362,1)
print(grupo_array_reshape.shape)

print(grupo_array_reshape)

print("\n")
```

Imagem 2: Código em Python

```python
# Supondo que 'grupo_array' contenha seus valores
grupo_array_flat = np.array(
    [item for sublist in grupo_array_reshape for item in sublist])

# Encontrando os valores únicos e suas contagens
valores, contagens = np.unique(grupo_array_flat, return_counts=True)

# Exibindo a contagem de cada valor
for valor, contagem in zip(valores, contagens):
    print(f"{valor}: tem quantidade {contagem}")
```

Com a ajuda do algoritmo feito em Python foi desenvolvido uma tabela de frequência e gráficos no Excel para detalhar os dados analisados.

Tabela 1: Distribuição das frequências nos Grupos

Classes	Fr. Abs	Fr. Relativa (%)	Fr. Ac	Fr. Ac. Rel (%)
1	56	15,47	56	15,47
2	146	40,33	202	55,8
3	95	26,24	297	82,04
4	65	17,96	362	100
Total:	362	100	724	

Gráfico 1: Distribuição dos pacientes por Grupos

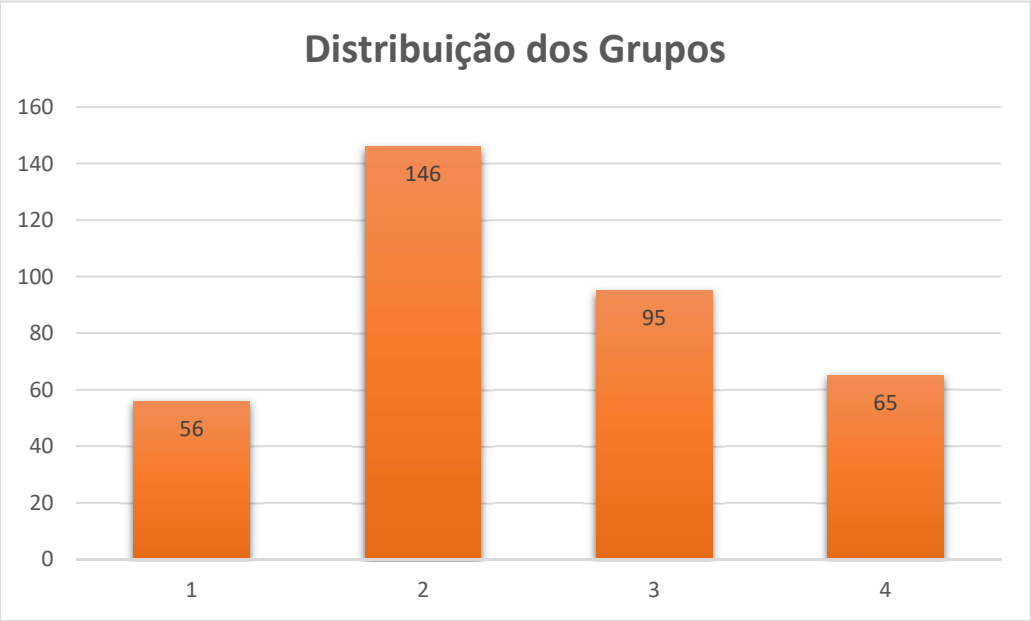
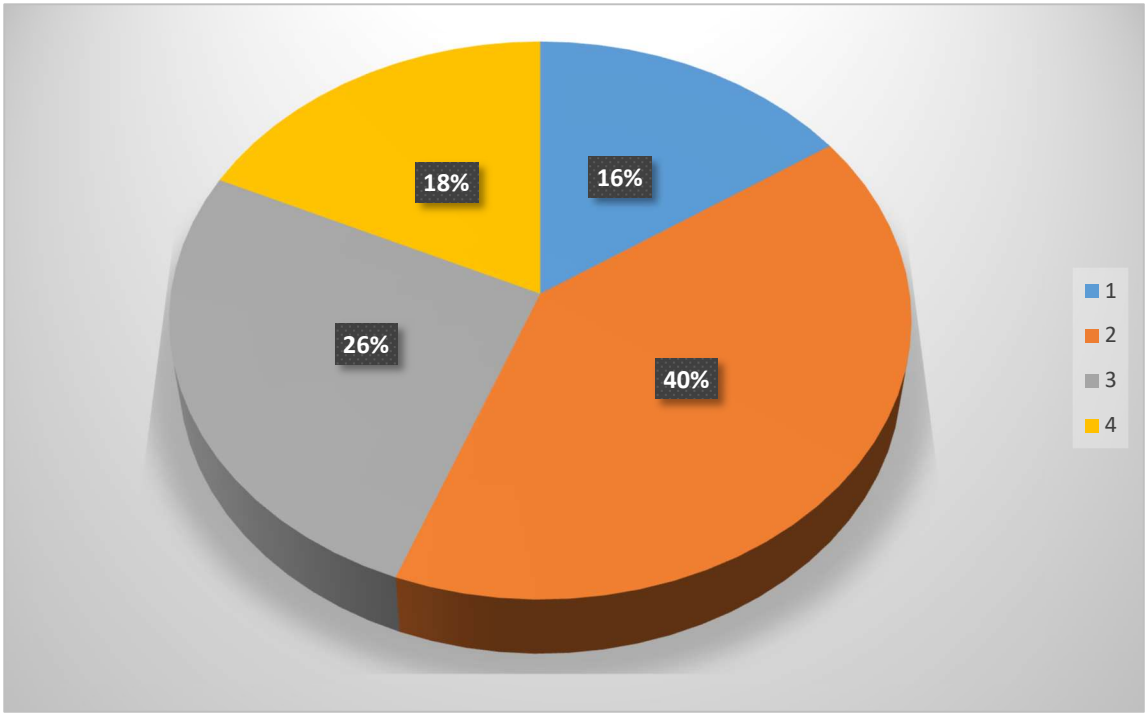


Gráfico 2: Distribuição Relativa dos Grupos



5. Distribuição de Frequência da Variável Quantitativa Idade

Será desenvolvido uma classificação das frequências das idades seguindo alguns passos que serão citados:

Passo 1: Colocar todos os dados brutos em ROL (Ordem Crescente).

Passo 2: Construir a tabela com os seguintes dados: Intervalos, Ponto Médio, o número de elementos que pertence ao intervalo (frequência absoluta), frequências relativas, frequências acumuladas das frequências absolutas e frequências acumuladas relativas.

Passo 3: Determinar o número de classes para a distribuição de frequência. Existem diferentes critérios para isso, como o critério da raiz, de Sturges e citado por Milone.

Passo 4: Calcular o intervalo de classe usando o critério escolhido.

Passo 5: Construir os limites dos intervalos de classe, tomando o menor valor como limite inferior e calculando os limites superiores para cada classe.

Passo 6: Calcular os pontos médios de cada classe.

Passo 7: Determinar as frequências absolutas para cada intervalo.

Passo 8: Calcular as frequências acumuladas das frequências absolutas.

Passo 9: Calcular as frequências relativas e as frequências acumuladas relativas.

Passo 10: Montar a tabela de distribuição de frequência com todas as colunas preenchidas.

Para colocar os dados em ROL foi desenvolvido um algoritmo em Python para colocar em ordem crescente os valores. (Segue abaixo imagem do algoritmo)

Imagem 3: Código em Python

```python
import pandas as pd
import numpy as np

# Importa o banco de dados para ser lido pelo Pandas
caminho_arquivo = "C:\\Users\\guilh\\OneDrive\\Documents\\4_Periodo_UFS\\ESTATISTICA\\cancer_formatados.csv"

# O Pandas analisa todo o arquivo csv e retira os espaços em branco.
base = pd.read_csv(caminho_arquivo, sep=';', converters={
                    'Grupo': lambda x: str(x).replace('\u200b', '')})

# Printa o shape: (362,9)
print(base.shape)

# Separa a coluna desejada das demais
Idade_array = base['Idade'].values

# Printa o shape da coluna Grupo: (362,)
print(Idade_array.shape)

# Faz um reshape para reconhecer ela como apenas uma coluna
Idade_array_reshape = Idade_array.reshape(-1, 1)
# Printa o shape corrigido: (362,1)
print(Idade_array_reshape.shape)

print(Idade_array_reshape)

print("\n")
```

Imagem 4: Código em Python

```
print("Array Idades em Rol\n")
Idade_array_reshape = np.sort(Idade_array_reshape, axis=0)
print(Idade_array_reshape)
print("\n")
print(Idade_array_reshape.shape)
print("\n")
```

Com isso, o número de classes para a distribuição de frequência terá que ser definido escolhendo entre os 3 métodos sendo: o critério da raiz, de Sturges e citado por Milone.

Temos atualmente 362 amostras para serem calculadas na divisão das classes, sendo n o número de amostras e Nc o número de classes.

Utilizando o Critério da Raiz temos: $Nc = \sqrt{n} \Rightarrow Nc = \sqrt{362} \Rightarrow Nc \approx 19$

Utilizando o Critério de Sturges temos: $Nc = 1 + 3{,}3 * \log_{10}(n) \Rightarrow$

$Nc = 1 + 3{,}3 * \log_{10}(362) \Rightarrow Nc = 1 + 3{,}3 * (2{,}5587) \Rightarrow Nc \approx 9$

Utilizando o Critério de Milone temos: $Nc = -1 + 2 * \ln(n) \Rightarrow$

$Nc = -1 + 2 * \ln(362) \Rightarrow Nc \approx 11$

Vamos escolher o critério de Sturges que gerou 9 classes e assim iremos determinar o intervalo das classes seguindo a seguinte fórmula de Sturges: $Ic = \frac{x_n - x_1}{N_c}$, sendo Ic o intervalo das classes, x_n é a maior idade analisada e x_1 é $a\ menor\ idade$ do banco de dados analisado. Calculando temos: $Ic = \frac{103-9}{9} \Rightarrow Ic \approx 10.$

Com os intervalos das classes começamos da seguinte forma: pegamos a menor idade como sendo o limite inferior e somamos ao intervalo de classes até atingir o valor máximo desejado que é 103.

Vamos descobrir a frequência absoluta dos dados em cada classe sendo: [9,19), [19,29), [29,39), [39,49), [49,59), [59,69), [69,79), [79,89), [89,99), [99,109).

Para realizar essa contagem iremos utilizar um algoritmo em Python que dirá as ocorrências em cada uma dessas classes. (Segue imagem do algoritmo abaixo)

Imagem 5: Código em Python

```
# Criação das faixas de idade
faixas_idade = [9, 19, 29, 39, 49, 59, 69, 79, 89, 99, 109]

# Contagem de ocorrências de cada faixa de idade
contagem_por_faixa = np.histogram(Idade_array, bins=faixas_idade)[0]

# Exibindo a contagem para cada faixa
for i, faixa in enumerate(faixas_idade[:-1]):
    print(f'Faixa de idade: [{faixa}, {faixas_idade[i+1]}) - Ocorrências: {contagem_por_faixa[i]}')
```

Teremos o seguinte retorno:

Faixa de idade: [9, 19) - Ocorrências: 15

Faixa de idade: [19, 29) - Ocorrências: 47

Faixa de idade: [29, 39) - Ocorrências: 35

Faixa de idade: [39, 49) - Ocorrências: 53

Faixa de idade: [49, 59) - Ocorrências: 71

Faixa de idade: [59, 69) - Ocorrências: 74

Faixa de idade: [69, 79) - Ocorrências: 41

Faixa de idade: [79, 89) - Ocorrências: 21

Faixa de idade: [89, 99) - Ocorrências: 2

Faixa de idade: [99, 109) - Ocorrências: 3

A frequência Relativa das Frequências relativas será encontrada pegando cada ocorrência de cada classe e dividindo pela soma das frequências absolutas.

Vamos descobrir o ponto médio em cada classe sendo: [9,19), [19,29), [29,39), [39,49), [49,59), [59,69), [69,79), [79,89), [89,99), [99,109).

Para realizar essa contagem iremos utilizar um algoritmo em Python que dirá as ocorrências em cada uma dessas classes. (Segue imagem do algoritmo abaixo)

```
# Calculando Ponto Medio:
print("Calculando Ponto Medio")
for i, faixa in enumerate(faixas_idade[:-1]):
    ponto_medio = 0
    ponto_medio = (faixa + faixas_idade[i+1])/2
    print(
        f'Faixa de idade: [{faixa}, {faixas_idade[i+1]}) - Ponto Médio: {ponto_medio}')
```

Obteve-se o seguinte retorno:

Faixa de idade: [9, 19) - Ponto Médio: 14.0

Faixa de idade: [19, 29) - Ponto Médio: 24.0

Faixa de idade: [29, 39) - Ponto Médio: 34.0

Faixa de idade: [39, 49) - Ponto Médio: 44.0

Faixa de idade: [49, 59) - Ponto Médio: 54.0

Faixa de idade: [59, 69) - Ponto Médio: 64.0

Faixa de idade: [69, 79) - Ponto Médio: 74.0

Faixa de idade: [79, 89) - Ponto Médio: 84.0

Faixa de idade: [89, 99) - Ponto Médio: 94.0

Faixa de idade: [99, 109) - Ponto Médio: 104.0

Para achar as frequências acumuladas pegamos cada frequência absoluta e somamos com a posterior até a última classe e depois calculamos as frequências relativas dividindo os valores encontrados pela soma das frequências absolutas.

Tabela 2: Distribuição das frequências separadas por classes

Intervalo de Classes	Ponto Médio	Fr. Abs	Fr. Rel (%)	Fr. Ac	Fr. Ac. Rel (%)
9 |-- 19	14	15	4,14	15	4,14
19 |-- 29	24	47	12,98	62	17,13
29 |-- 39	34	35	9,67	97	26,8
39 |-- 49	44	53	14,64	150	41,44
49 |-- 59	54	71	19,61	221	61,05
59 |-- 69	64	74	20,44	295	81,49
69 |-- 79	74	41	11,33	336	92,82
79 |-- 89	84	21	5,8	357	98,62
89 |-- 99	94	2	0,55	359	99,17
99 |-- 109	104	3	0,83	362	100
Total:		362	99,99	2254	

Gráfico 3: Frequência Absoluta das Idades por Classes

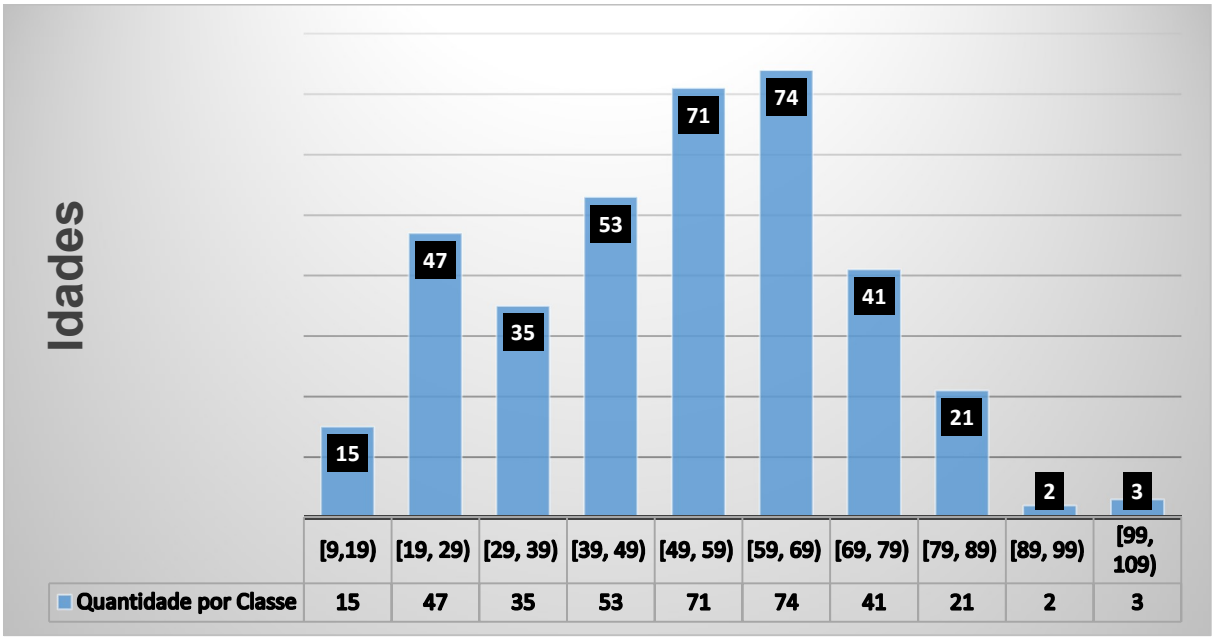
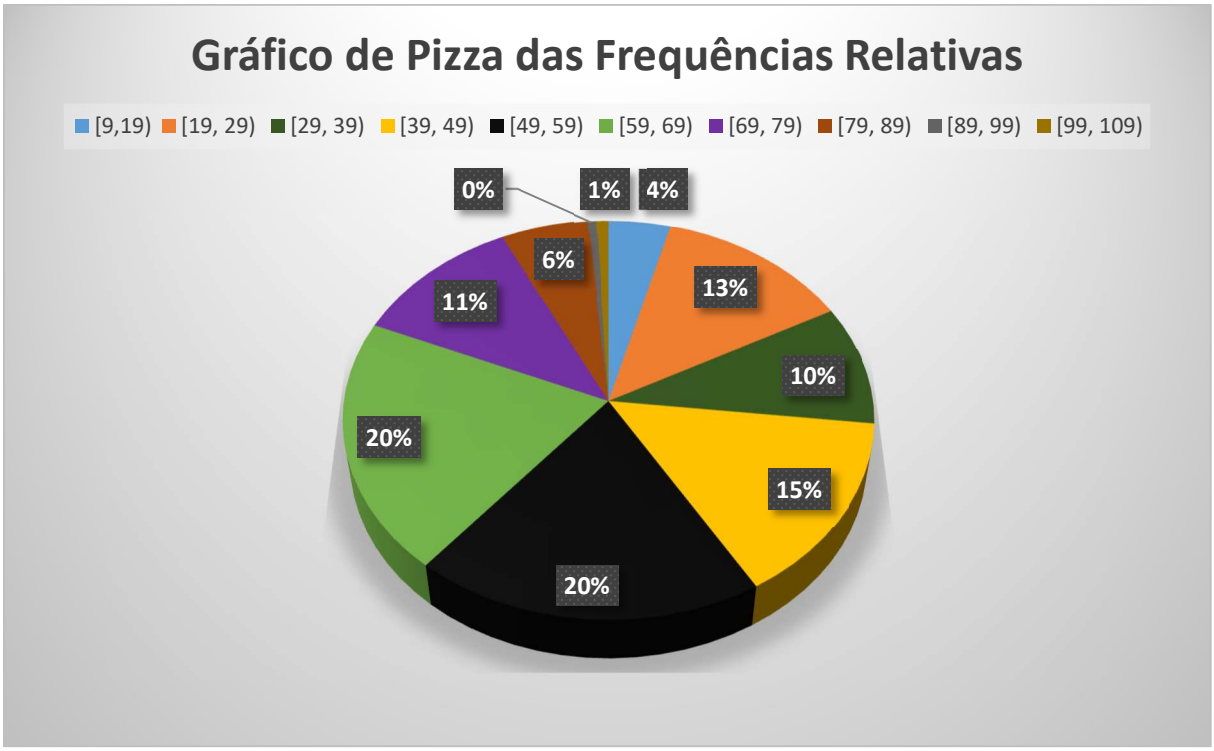


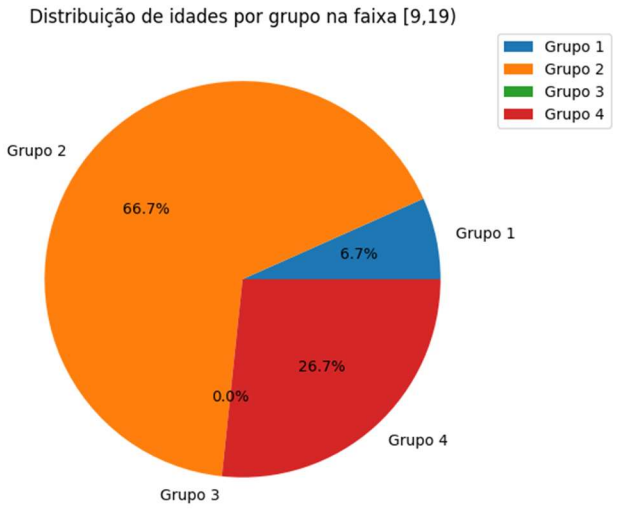
Gráfico 4: Frequências Relativas em gráfico de Pizza

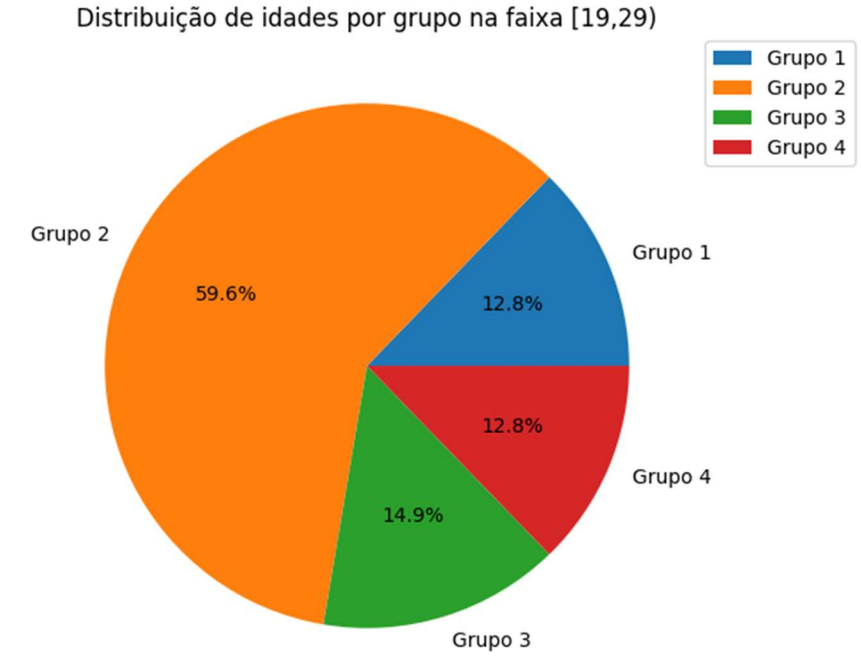


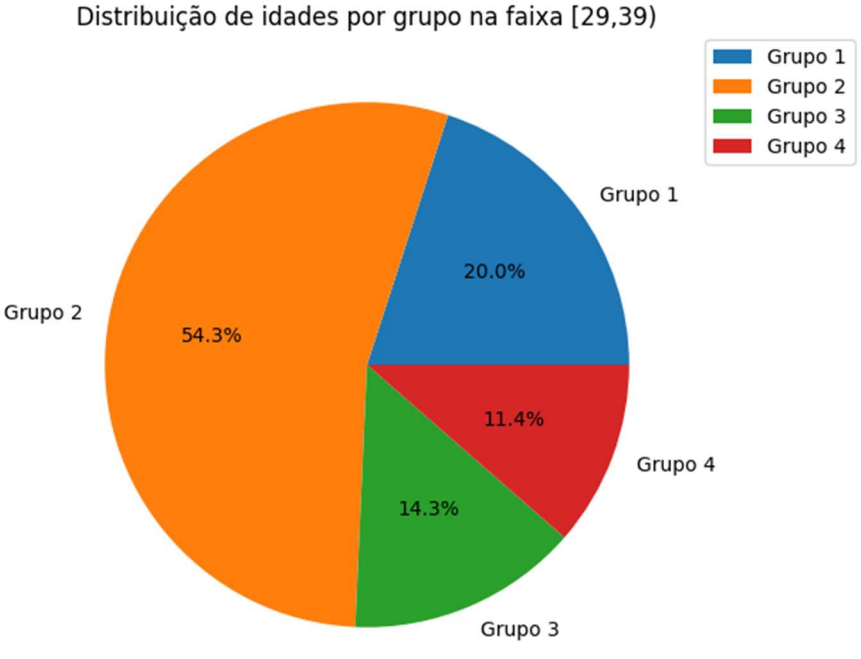
Com base nisso, pode-se demonstrar uma referência capaz de mostrar a quantidade de valores Qualitativos (Grupos) nos valores quantitativos (Idades). Esse processo foi realizado por meio de um algoritmo em Python que estará a seguir:

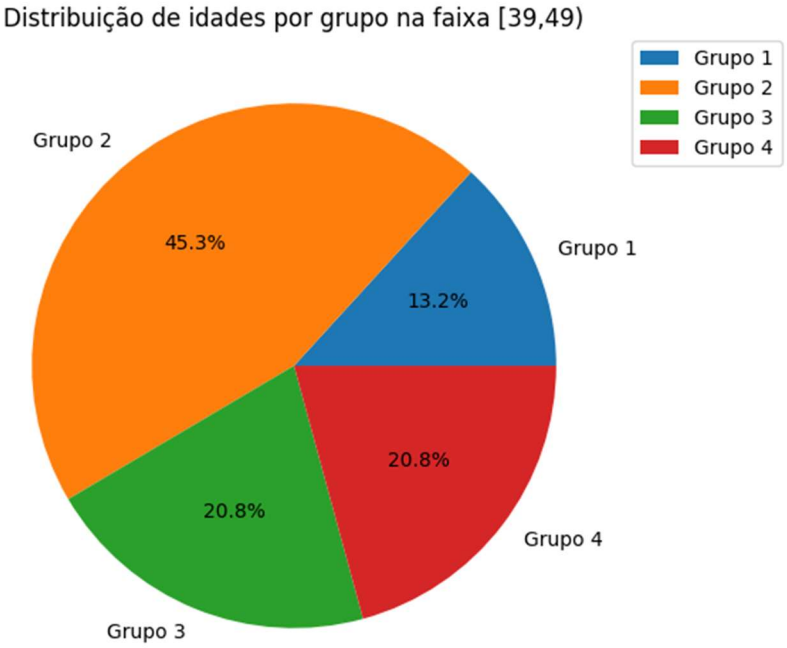
E assim, foi realizado a construção de gráficos de pizza dessa análise de dados por meio do código em Python, utilizando a biblioteca matplotlib.

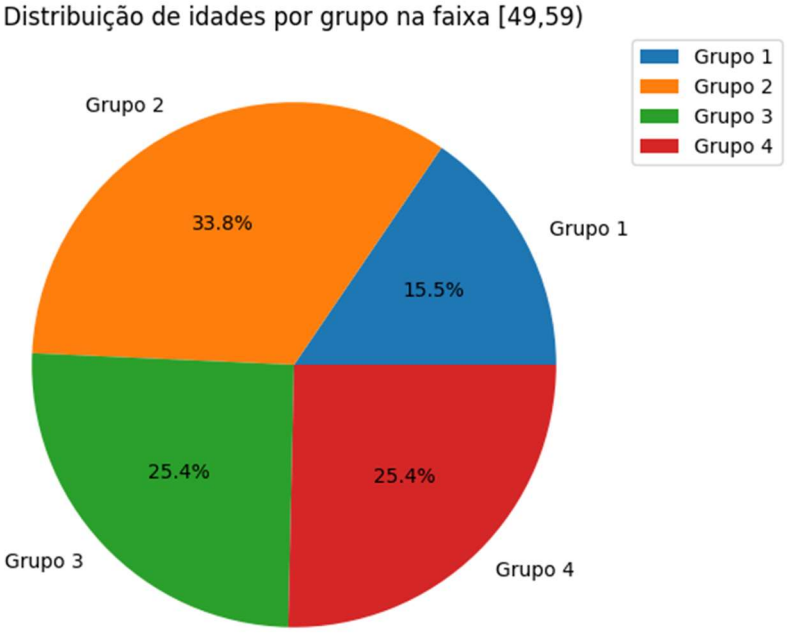
Dessa maneira, temos como resultado os seguintes gráficos por faixa de idade correspondentes aos grupos. (Segue imagens abaixo)

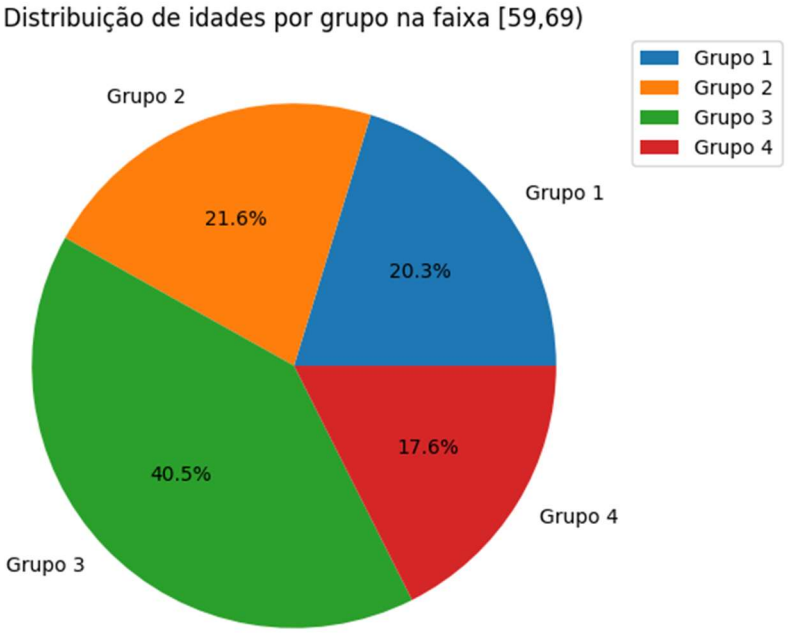


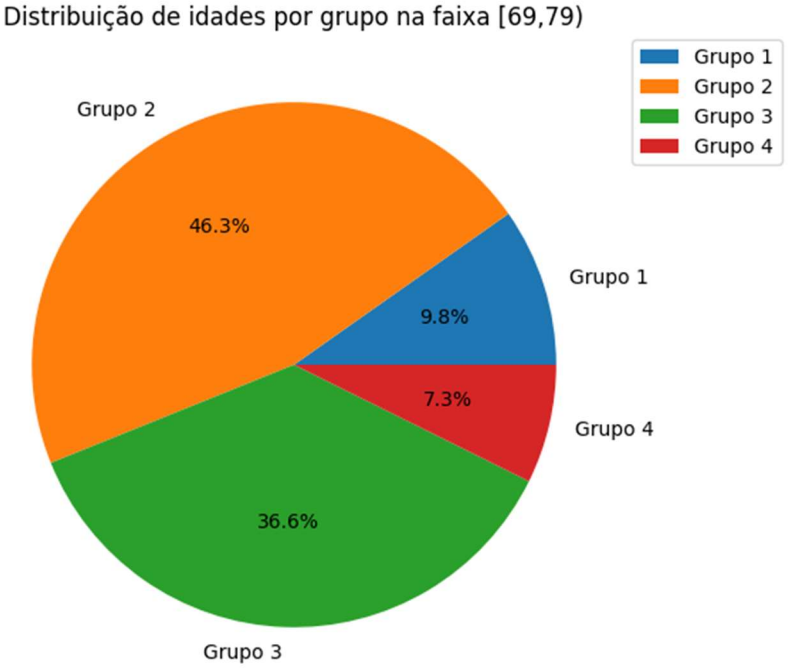

Distribuição de idades por grupo na faixa [19,29)
Grupo 1
Grupo 2
Grupo 3
Grupo 4
59.6%
12.8%
12.8%
14.9%

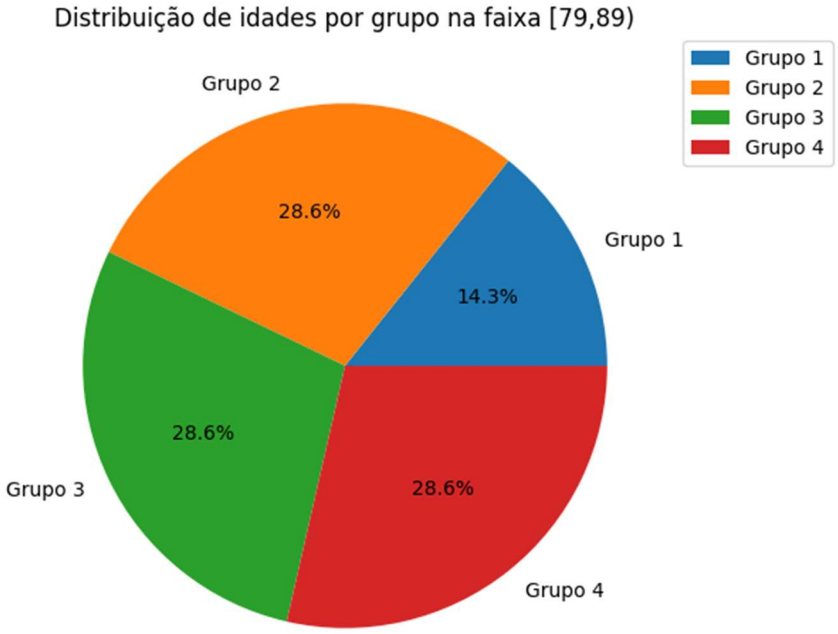

Distribuição de idades por grupo na faixa [29,39)
Grupo 1
Grupo 2
Grupo 3
Grupo 4
20.0%
54.3%
11.4%
14.3%

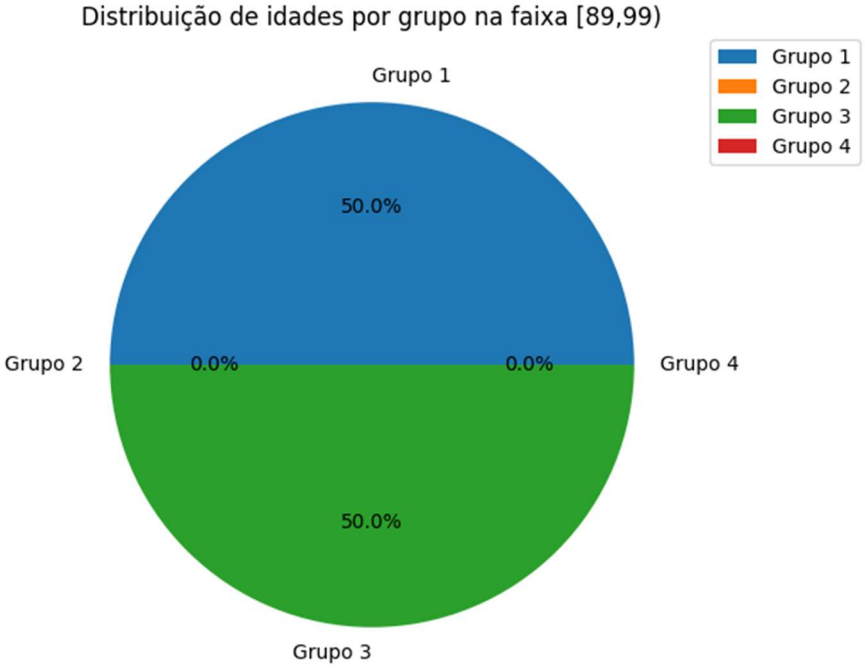

Distribuição de idades por grupo na faixa [39,49)
Grupo 1
Grupo 2
Grupo 3
Grupo 4
Grupo 2
45.3%
Grupo 1
13.2%
20.8%
20.8%
Grupo 4
Grupo 3

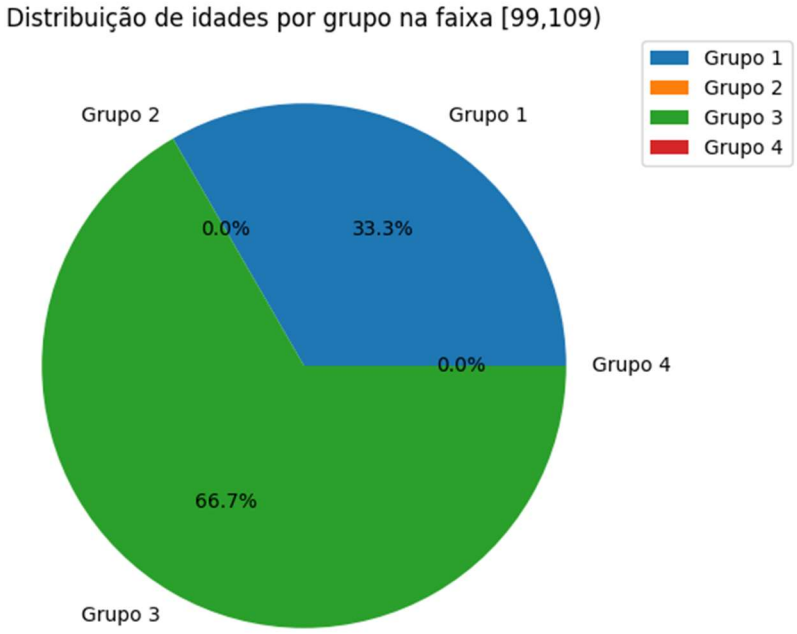

Distribuição de idades por grupo na faixa [49,59)
Grupo 1
Grupo 2
Grupo 3
Grupo 4
Grupo 2
33.8%
Grupo 1
15.5%
25.4%
25.4%
Grupo 3
Grupo 4


Distribuição de idades por grupo na faixa [59,69)
Grupo 1
Grupo 2
Grupo 3
Grupo 4
Grupo 2
Grupo 1
21.6%
20.3%
17.6%
40.5%
Grupo 4
Grupo 3


Distribuição de idades por grupo na faixa [69,79)
Grupo 1
Grupo 2
Grupo 3
Grupo 4
Grupo 2
46.3%
Grupo 1
9.8%
7.3%
Grupo 4
36.6%
Grupo 3


Distribuição de idades por grupo na faixa [79,89)
Grupo 1
Grupo 2
Grupo 3
Grupo 4
28.6%
14.3%
28.6%
28.6%


Distribuição de idades por grupo na faixa [89,99)
Grupo 1
Grupo 2
Grupo 3
Grupo 4
50.0%
0.0%
0.0%
50.0%



6. Medidas Descritivas de Posição

6.1. Medidas de Posição e Medidas Descritivas

Será colocado o conjunto Idade em Rol para analisar a média, mediana e moda das idades que estão no conjunto de dados.

A média será calculado usando a função mean do Numpy nas colunas da idade dos pacientes, dos níveis de AKP no sangue dos pacientes, da concentração de fosfato, LDH, Albumina, Nitrogênio na ureia e glicose. Para a moda foi feito um código em Python para registrar os valores que mais se repetem nas colunas analisadas. Para a mediana foi usado a função usando a função median da biblioteca Numpy. Para a variância, desvio padrão será utilizado as funções var() e std() do Numpy para calcular seus valores respectivos e o coeficiente de variação será calculado na divisão entre o desvio padrão e a média dos valores, expressa em porcentagem. Segue abaixo tabela com os valores encontrados.

Tabela 3: Distribuição das Medidas de Posição e Medidas Descritivas

Coluna1	Idade	AKP	P	LDH	ALB	N	GL
Média	51,21	9,67	3,19	18,26	59,65	14,9	104,32
Mediana	54	8,6	3,2	15,15	60	14	99
Moda	54	5,4 / 6,8 / 7,1	3,2	15,5	59	15	91 / 93
Variância	364,35	25,32	0,34	137,91	45,62	38,99	613,84
Desvio Padrão	19,09	5,03	0,59	11,74	6,75	6,24	24,78
Coef. Variação (%)	37,27	52,03	18,38	64,31	11,32	41,9	23,75
Máximo	103	50	5,6	99,9	76	58	298
Mínimo	9	1,7	1,2	0	20	3	0

7. Conclusão

Com base na tabela fornecida, é possível extrair informações cruciais sobre as variáveis analisadas no estudo. A média de idade dos pacientes é de 51 anos, com uma faixa considerável entre 9 e 103 anos. A fosfatase alcalina (AKP) tem uma média de 9,67, com valores variando entre 1,7 e 50. Observa-se uma variação considerável nos parâmetros Idade e GL, indicada pelos altos valores de variância e desvio padrão. Por exemplo, a variação de GL é significativamente alta, sugerindo uma dispersão maior em torno da média. A mediana da maioria das variáveis está próxima das respectivas médias, sugerindo distribuições aproximadamente simétricas. No entanto, em alguns casos, como AKP, a mediana é menor que a média, indicando uma possível assimetria à direita. O coeficiente de variação (%) indica a consistência dos dados em relação à média. Valores mais baixos, como na ALB, sugerem menor dispersão dos dados.

8. Referências

[1] https://www.ime.usp.br/~noproest/doku.php?id=dados .: Dados de Incidência de Câncer

[2] https://www.gov.br/inca/pt-br/assuntos/cancer/o-que-e-cancer

[3] Esdras Adriano, and Esdras Adriano Santos. “Notas de Aula de Estatística.” Notas de aula exposta ao alunos das discplinas de serviço da UFS. Notas de Aula de Estatística, 2009, São Cristóvão, Sergipe, Brasil.