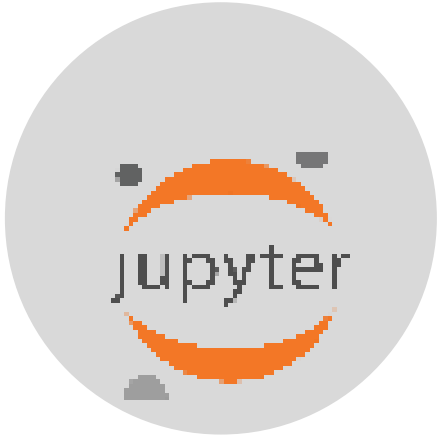
Aula de hoje



- **Histograma com uso de densidade**
- **Alguns possíveis formatos do histograma**
- **Formato do histograma *versus* Medidas de posição**

1º: uso do notebook:

Aula04_Atividade

2º: uso do notebook:

Aula04_Exercício

Insper

Ciência dos dados

Histograma

Objetivos de Aprendizado

Ao final desta aula, o aluno deve ser capaz de:

- Calcular medidas que representem os percentis (ou quantis) de uma particular amostra e interpreta-los.
- Construir um HISTOGRAMA útil para visualização gráfica de variáveis quantitativas.
- Explicar vantagens sobre o uso da **densidade** na construção de um histograma e saber interpretá-lo.
- Discutir alguns formatos de um histograma e suas relações com as medidas de posição: média, mediana e moda

Tabela de frequências para variável quantitativa

Construção de tabelas para variáveis quantitativas

Tabela de frequências para variável quantitativa:

```
In [3]: dados = pd.read_excel('EmpresaTV.xlsx')

In [6]: dados.RENDA.value_counts().head(15)

Out[6]: 4.9     3
        5.4     2
        2.5     2
        13.2    2
        0.8     2
        12.9    2
        7.4     2
        10.7    2
        5.5     2
        5.3     2
        6.0     2
        4.7     2
        11.2    2
        3.9     1
        0.6     1
        Name: RENDA, dtype: int64
```

?

Construção de tabelas para variáveis quantitativas

Tabela de frequências para variável quantitativa:

A construção de tabelas de frequências para variáveis quantitativas necessita de alguns cuidados.

Se construirmos uma tabela de frequências para a variável RENDA, por exemplo, usando função .value_counts(), essa tabela não resumirá as observações num grupo menor, pois não existem ou existem poucos valores iguais. Certamente, dificultará na interpretação!

A solução empregada é agrupar os dados por faixa de renda as quais podem ter amplitudes iguais ou desiguais.

Construção de tabelas para variáveis quantitativas

Tabela de frequências para variável quantitativa:

- Dividir os dados em classes
- Contar quantas observações há em cada classe:

  Frequência Absoluta
- Dividir pelo número total de observações:

  Frequência Relativa

Construção de tabelas para variáveis quantitativas

Determinação do número e da amplitude das classes:

O número de classes não deve ser tão grande a ponto de se ter classes com muito poucas observações e nem tão pequeno a ponto de mascarar o comportamento dos dados.

Tabela de frequências relativas para RENDA

Plano A

```
Frequências relativas:
[0.5, 4)        6.5
[4, 7.5)       19.6
[7.5, 11)      32.6
[11, 14.5)     26.1
[14.5, 18)     10.9
[18, 21.5)      4.3
Name: RENDA, dtype: float64
```

Plano B

```
Frequências relativas:
[0.5, 4)       22.2
[4, 7.5)       55.6
[7.5, 11)      19.4
[11, 14.5)      0.0
[14.5, 18)      0.0
[18, 21.5)      2.8
Name: RENDA, dtype: float64
```

Comando Python:

from numpy import arange

faixa = arange(start, stop, step) ou faixa = range(start, stop, step)

variávelCateg = **pd.cut(**variávelQuant, bins=faixa, right=False**)**

variávelCateg.value_counts()

Histograma

Uso de densidade no eixo y

Gráfico de colunas para RENDA – com **amplitudes desiguais**

Plano A

```
Frequências relativas:
[0.5, 4.0)       6.5
[4.0, 7.5)      19.6
[7.5, 11.0)     32.6
[11.0, 14.5)    26.1
[14.5, 21.5)    15.2
Name: RENDA, dtype: float64
```

?

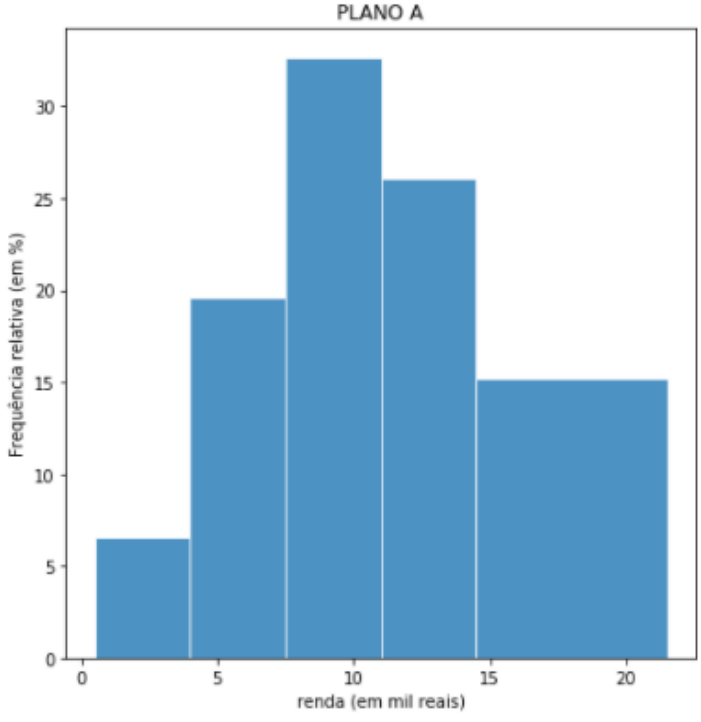
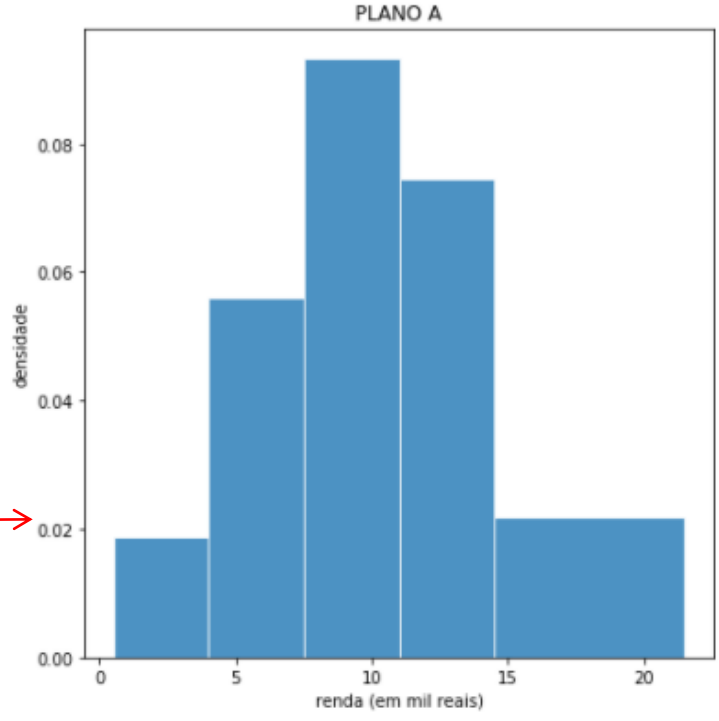


Gráfico de colunas para RENDA – com **amplitudes desiguais**

Plano A

```
Frequências relativas:
[0.5, 4.0)       6.5
[4.0, 7.5)      19.6
[7.5, 11.0)     32.6
[11.0, 14.5)    26.1
[14.5, 21.5)    15.2
Name: RENDA, dtype: float64
```



Usar densidade no eixo y para forçar área do histograma igual a 1!!

Gráfico de colunas para RENDA – com **amplitudes desiguais**

OUTRO EXEMPLO

Plano A

```
Frequência relativas:
(0, 4]        6.5
(4, 8]       23.9
(8, 11]      28.3
(11, 15]     28.3
(15, 22]     13.0
Name: RENDA, dtype: float64
```

?

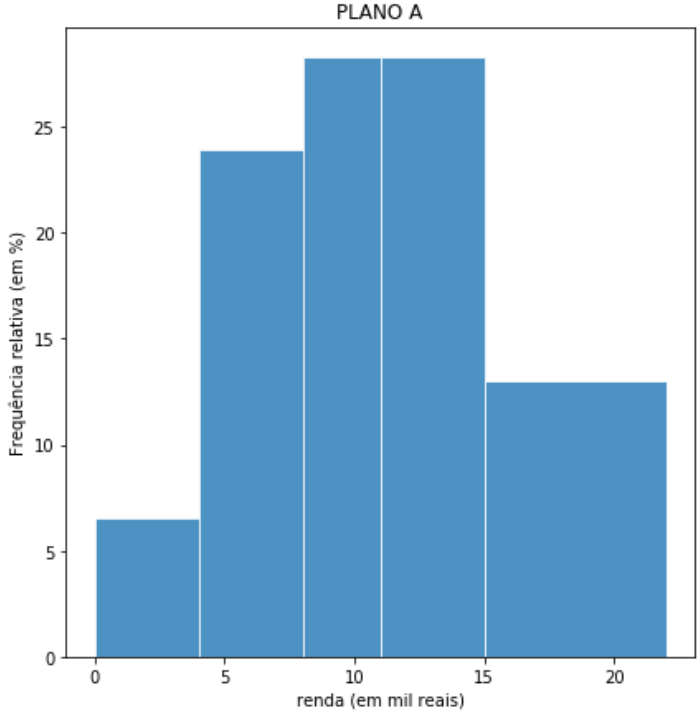
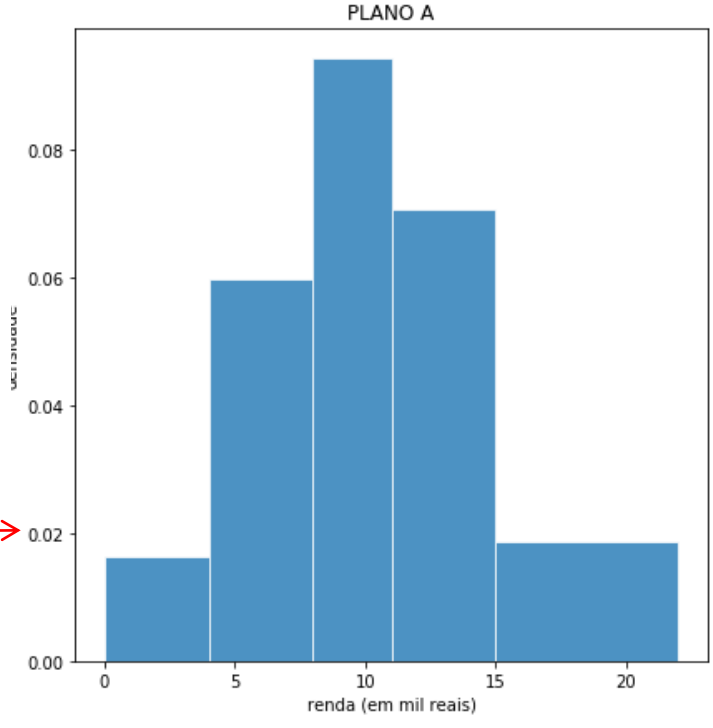


Gráfico de colunas para RENDA – com **amplitudes desiguais**

OUTRO EXEMPLO

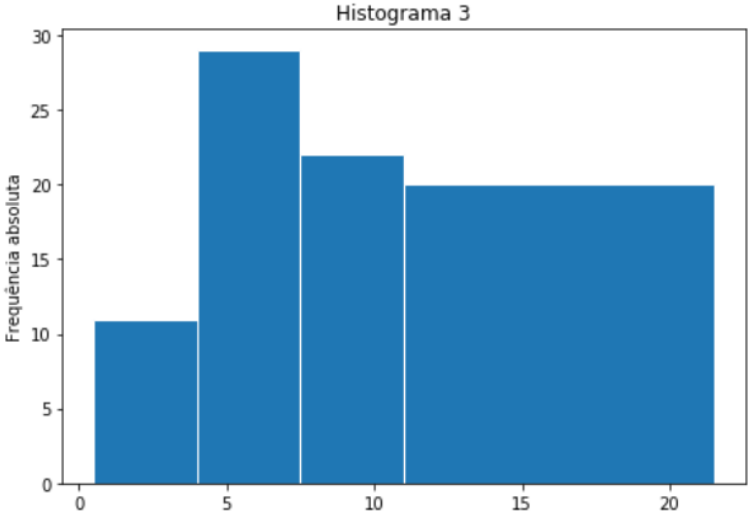
Plano A

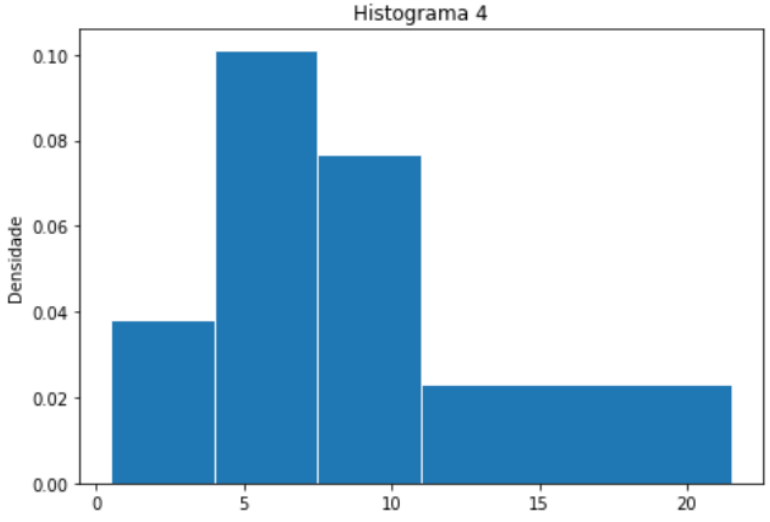
```
Frequência relativas:
(0, 4]       6.5
(4, 8]      23.9
(8, 11]     28.3
(11, 15]    28.3
(15, 22]    13.0
Name: RENDA, dtype: float64
```



Usar densidade no eixo y para forçar área do histograma igual a 1!!

Exercício 1 – da Aula04_Atividade



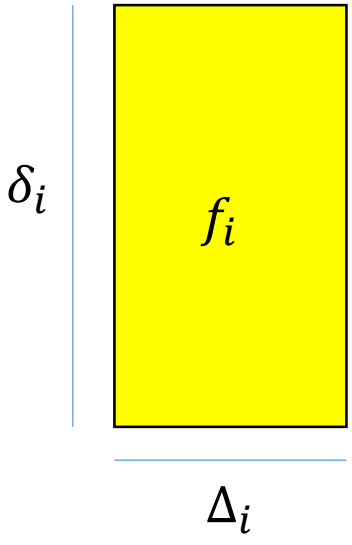



```
dados.RENDA.describe().to_frame().transpose()
```

	count	mean	std	min	25%	50%	75%	max
RENDA	82.0	8.343902	4.620622	0.6	4.925	7.75	10.775	21.4

Na aula passada, já discutimos que o Histograma 3 está errado.
Porém, como calcular a densidade?

Como calcular densidade



Δ_i: amplitude (largura) da classe i

δ_i: altura da classe i

f_i **:** área da classe i

A área é o que chama a atenção no gráfico e queremos representar a frequência relativa com que cada classe aparece.

Como determinar δ_i (medida para eixo y)?

Sabemos que $Area = base \times altura$

Logo,

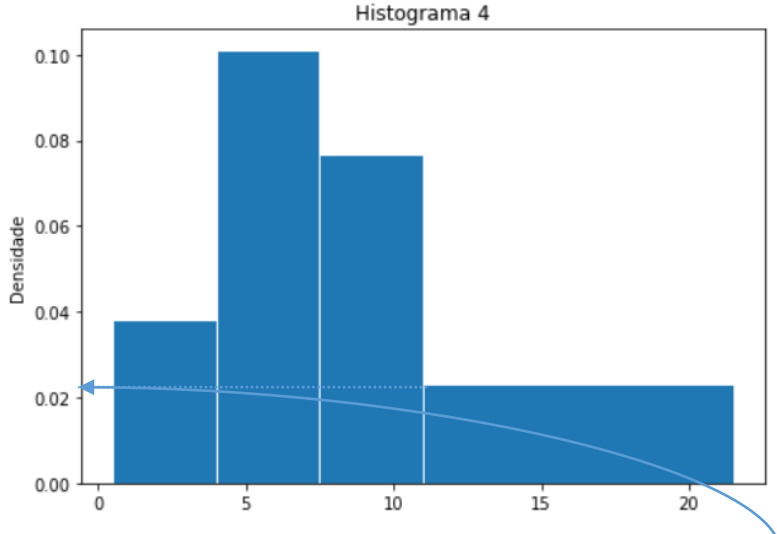
$$f_i = \Delta_i \delta_i \quad \Rightarrow \quad \delta_i = \frac{f_i}{\Delta_i}$$

O que é δ_i?

Exercício 2 - Aula03_Atividade

Frequência relativas:

	RENDA
[0.5, 4.0)	0.134
[4.0, 7.5)	0.354
[7.5, 11.0)	0.268
[11.0, 21.5)	0.244



Como calcular a densidade?

Considerando a quarta classe, por exemplo, temos:

$$f_i = 0{,}244$$

$$\Delta_i = (21{,}5 - 11) = 10{,}5 \quad \Rightarrow \quad \delta_i = \frac{f_i}{\Delta_i} = \frac{0{,}244}{10{,}5} 0{,}0232$$

Construção de histograma

Determinação da densidade:

O nome densidade é dado para distribuições cuja área total sob a curva é igual a 1.

Ou seja, **Área total na soma de todos os retângulos formados no histograma deve ser igual a 1**.

Com isso, a densidade para classe é obtida a partir da conta:

Densidade = frequência relativa / amplitude da classe

Dessa forma, frequência relativa de uma classe está refletida na área de sua respectiva caixa formada no histograma.

É possível construir um histograma com classes de tamanhos diferentes?

Sim. Entretanto, é necessário ter cuidado na interpretação do histograma.

Notebook Atividade – sala

Explorando base de dados reais:

- Download pelo Github:

  https://github.com/Insper/CD22-2

- Fazer individual e discutir em sala

Notebook Exercício

Explorando base de dados reais:

- Download pelo Github:

  https://github.com/Insper/CD22-2

- Fazer individual e discutir na mesa

Próxima aula...

Leitura prévia necessária:

- Tutorial de Pandas via Jupyter
- Montgomery & Runger, Seç. 6.6 e Seç. 11.2.
- Magalhães & Lima, Cap. 1. e Cap. 4.
- Grus, Cap. 5
- Jogar no http://guessthecorrelation.com/