

# **Ciência dos Dados**

## **Aula 03**

### **Análise Exploratória dos Dados**

# Objetivos de Aprendizagem

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

- Construir tabelas de frequências e interpretar resultados considerando uma variável quantitativa e de forma cruzada com uma outra variável qualitativa.
- Explicar vantagens e desvantagens sobre o uso da frequência ou da densidade na construção de um histograma e saber interpretá-lo.
- Desenvolver contas para obtenção de média, mediana e moda e associar ordenação dessas medidas de acordo com assimetria dos dados.

# Construção de tabelas para variáveis quantitativas

## Tabela de frequências para variável quantitativa:

```
dados = pd.read_excel('EmpresaTV.xlsx')
```

```
#Selecionando variável renda  
renda = dados.iloc[:, -1]
```

```
renda.value_counts().head(15)
```

4.9	3
0.8	2
5.3	2
10.7	2
12.9	2
11.2	2
7.4	2
6.0	2
2.5	2
4.7	2
13.2	2
5.5	2
5.4	2
10.2	1
18.5	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

## 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.

**Regra empírica** para se ter um ponto de partida:

Para uma amostra de tamanho  $n$ , sugere-se utilizar  $\sqrt{n}$  classes. Apenas cuidado que, para grandes amostras, esta regra pode levar a um número exagerado de classes.

# Construção de tabelas para variáveis quantitativas

## Determinação do número e da amplitude das classes:

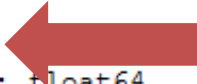
- ❑ Número de classes: aproximadamente  $\sqrt{n}$
- ❑ Obter valores Mínimo e Máximo do conjunto de dados
- ❑ Amplitude dos dados:  $\Delta = \text{Máximo} - \text{Mínimo}$
- ❑ Amplitude sugerida das classes: aproximadamente  $\frac{\Delta}{\sqrt{n}}$
- ❑ **Opte por construir faixas com valores mais fáceis de interpretar, ou seja, valores mais inteiros. A primeira faixa não precisa começar necessariamente com o valor mínimo do conjunto de dados; assim como a última faixa não precisa terminar no valor máximo.**

# Construção de tabelas para variáveis quantitativas

## Determinação do número e da amplitude das classes:


```
#So PLANO A
#Selecionando variável renda
rendaA = dados.RENDA[dados.PLANO=='A']
rendaA.describe()
```

```
count    46.000000
mean     10.421739
std       4.465568
min       0.700000
25%       7.475000
50%      10.350000
75%      13.200000
max      21.400000
Name: RENDA, dtype: float64
```



```
#So PLANO B
#Selecionando variável renda
rendaB = dados.RENDA[dados.PLANO=='B']
rendaB.describe()
```

```
count    36.000000
mean      5.688889
std       3.293437
min       0.600000
25%       4.150000
50%       5.150000
75%       6.375000
max      19.200000
Name: RENDA, dtype: float64
```



# 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*



# 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
```

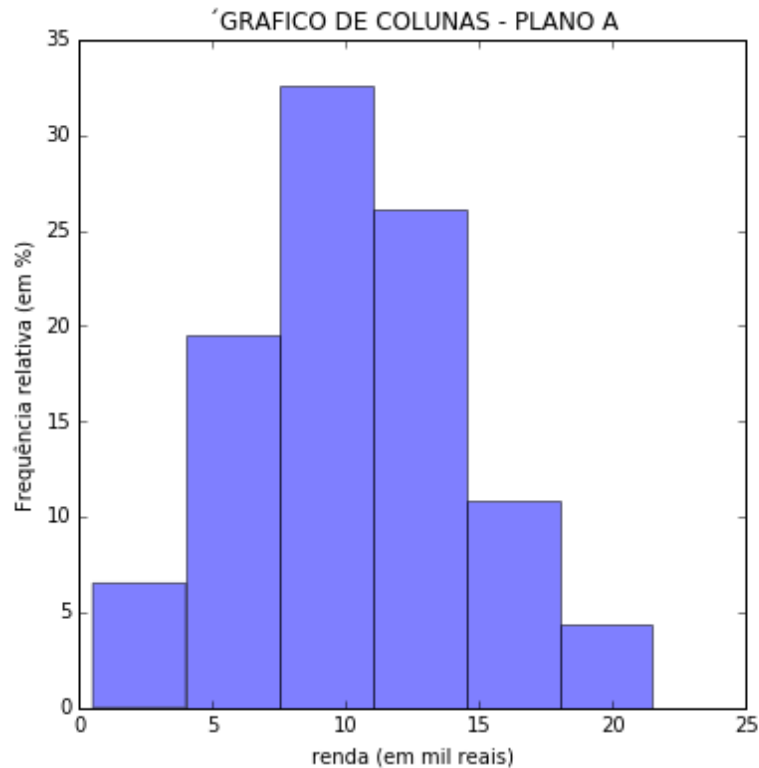
```
faixa = arange(start, stop, step) ou faixa = range(start, stop, step)
```

```
variávelCateg= pd.cut(variávelQuant, faixa)
```

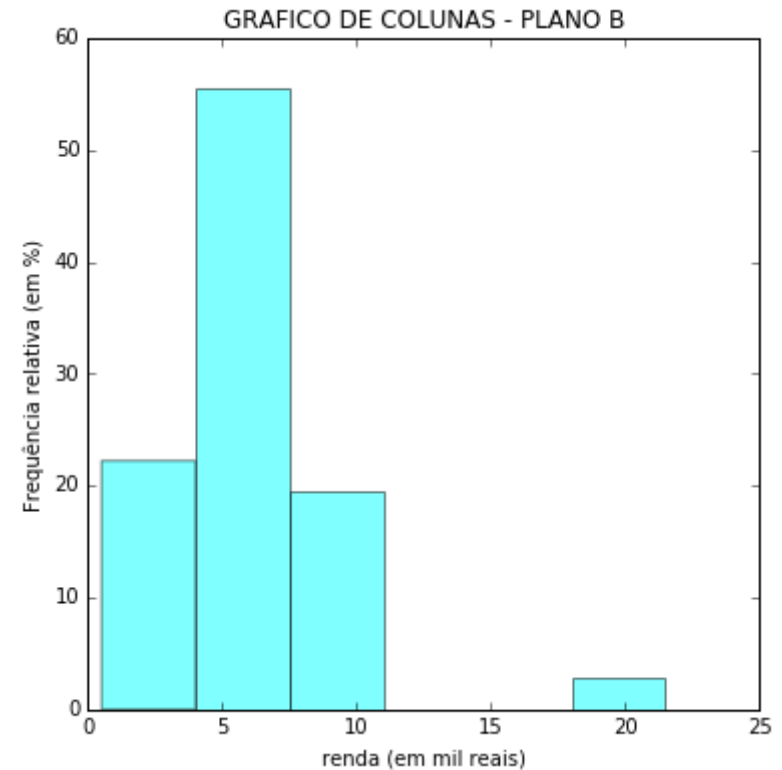
```
variávelCateg.value_counts()
```

# Gráfico de colunas para RENDA

## Plano A



## Plano B



**Comando Python:**

```
plot = variavelQuant.plot.hist(bins=faixa)
```

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

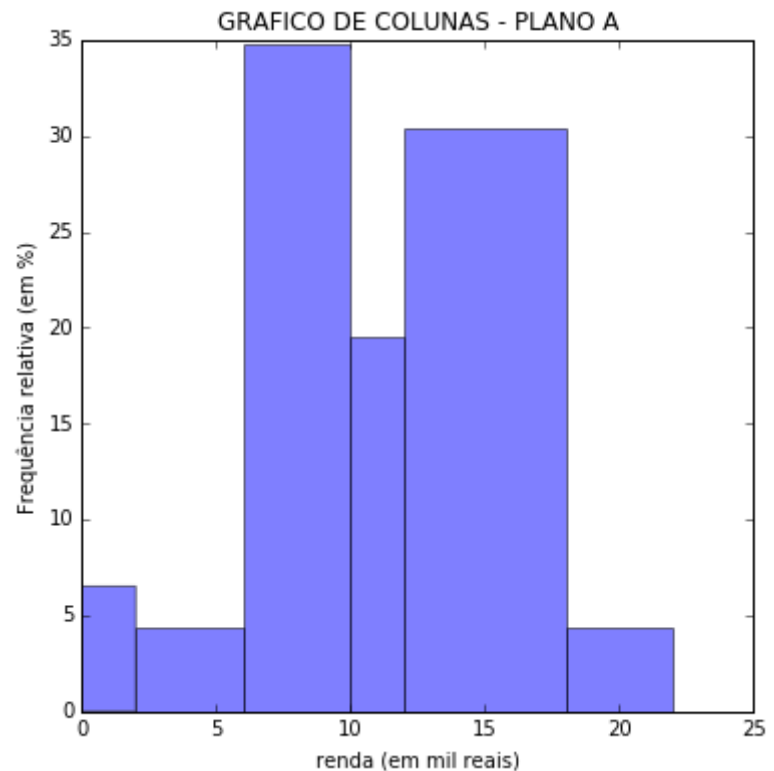
## Plano A

Frequências relativas:

[0, 2)	6.5
[2, 6)	4.3
[6, 10)	34.8
[10, 12)	19.6
[12, 18)	30.4
[18, 22)	4.3

Name: RENDA, dtype: float64

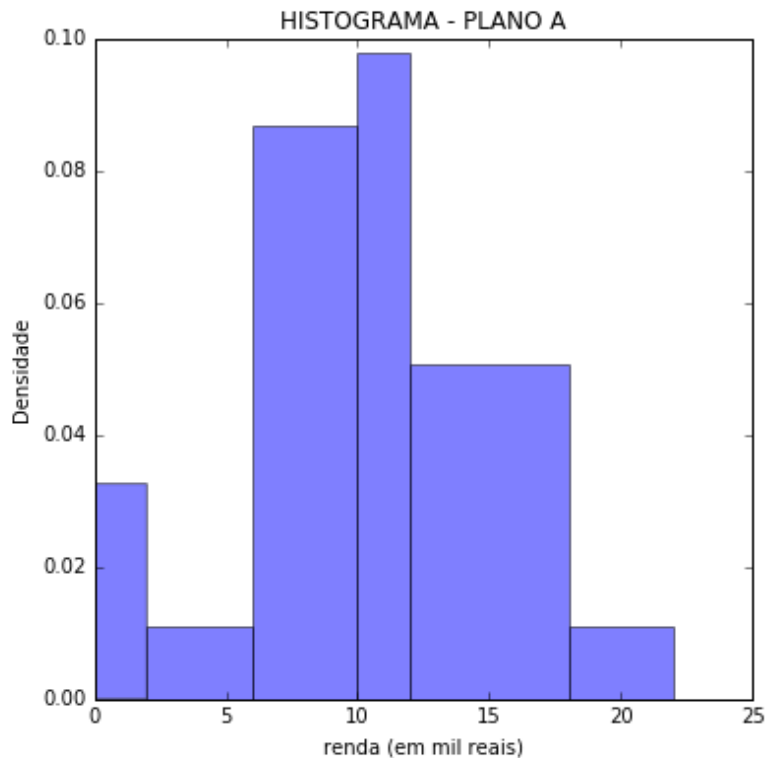
?



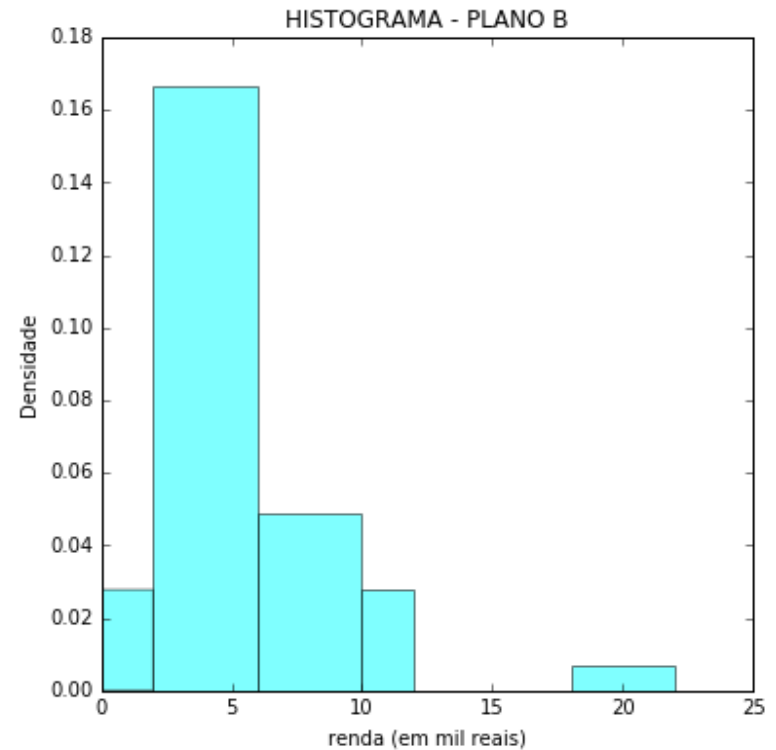
# HISTOGRAMA para RENDA

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

**Plano A**



**Plano B**



**Comando Python:**

```
plot = variavelQuant.plot.hist(bins=faixa, normed=1)
```

# 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:

$$\text{Densidade} = \text{frequência relativa} / \text{amplitude da classe}$$

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.

# **Média, Mediana e Moda**

**via base de dados**

**via tabela de frequências**

# Notação

Amostra de  $n$  observações da variável  $X$ :

$$x_1, x_2, \dots, x_n$$

Amostra **ordenada** de  $n$  observações da variável  $X$ :

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$$

$$\text{Mínimo} = x_{(1)}$$

$$\text{Máximo} = x_{(n)}$$

# Média aritmética

Amostra de  $n$  observações da variável  $X$ :

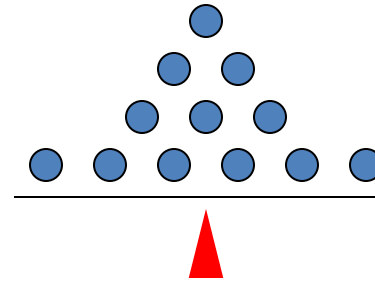
$$x_1, x_2, \dots, x_n$$

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$



# Média aritmética

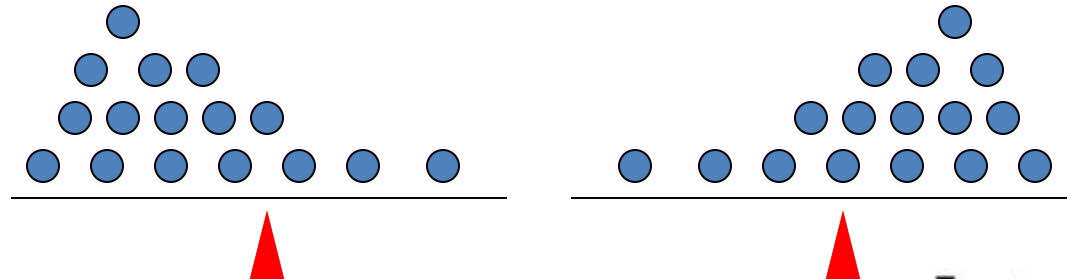
$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$



Valores aberrantes



Assimetrias



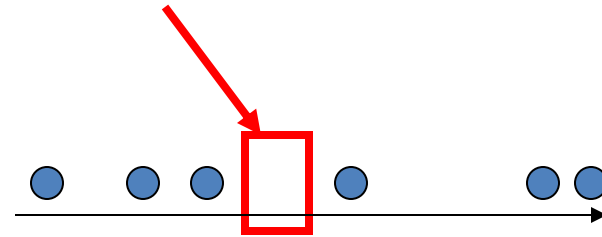
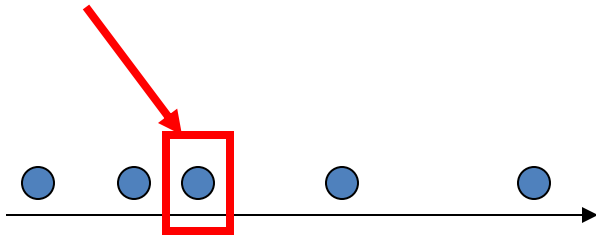
# Mediana

**Mediana** é o valor que divide um conjunto de dados ordenados ao meio. Em outras palavras, é um valor tal que tenha igual quantidade de valores menores e maiores do que ele.

Uma característica importante da mediana é que ela não é afetada por dados extremos, como acontece com a média.

# Mediana

$$\text{md}(X) = \begin{cases} x_{\left(\frac{n+1}{2}\right)}; \text{sen é ímpar} \\ \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}; \text{sen é par} \end{cases}$$



# Moda

## Para variável discreta:

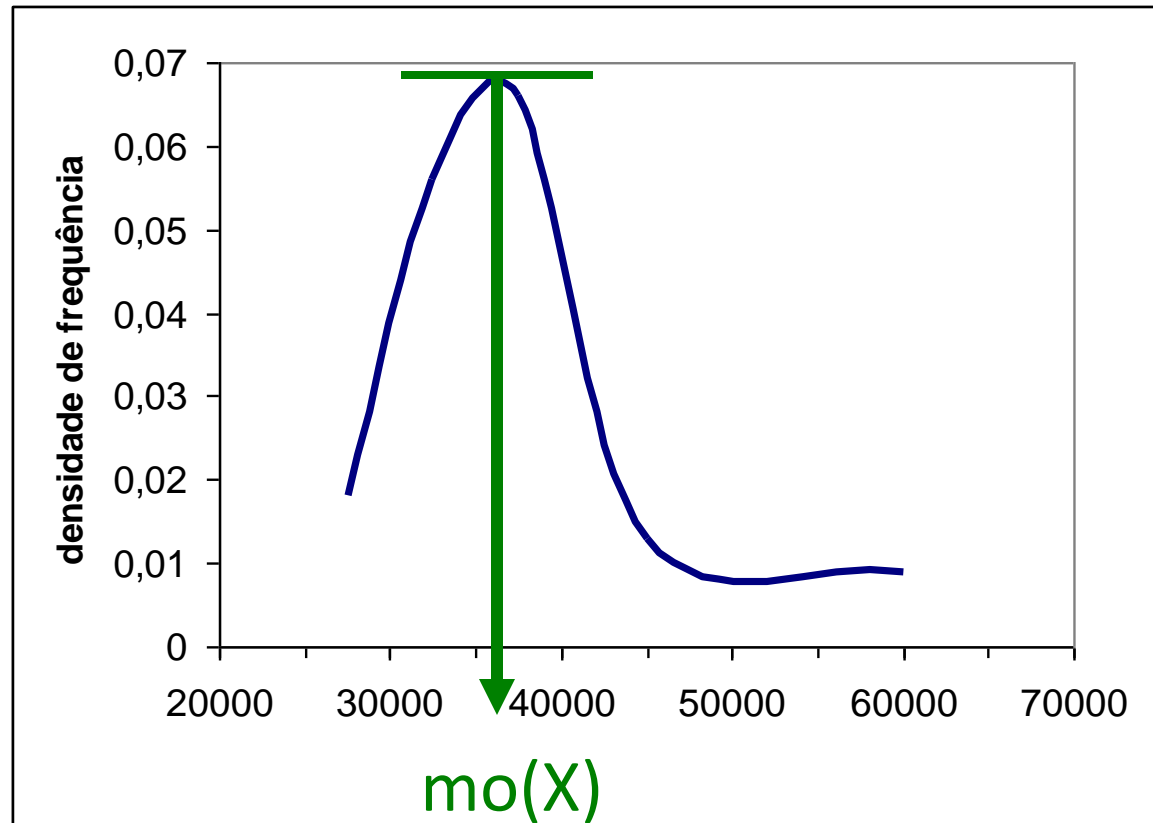
**Moda** de um conjunto de dados, representado por **mo(x)**, é o valor que ocorre com maior frequência.

- Quando dois valores ocorrem com a mesma maior frequência, cada um é uma moda e o conjunto de dados é chamado de **bimodal**.
- Quando mais de dois valores ocorrem com a mesma maior frequência, cada um é uma moda e o conjunto de dados é **multimodal**.

# Moda (escolhendo a classe modal)

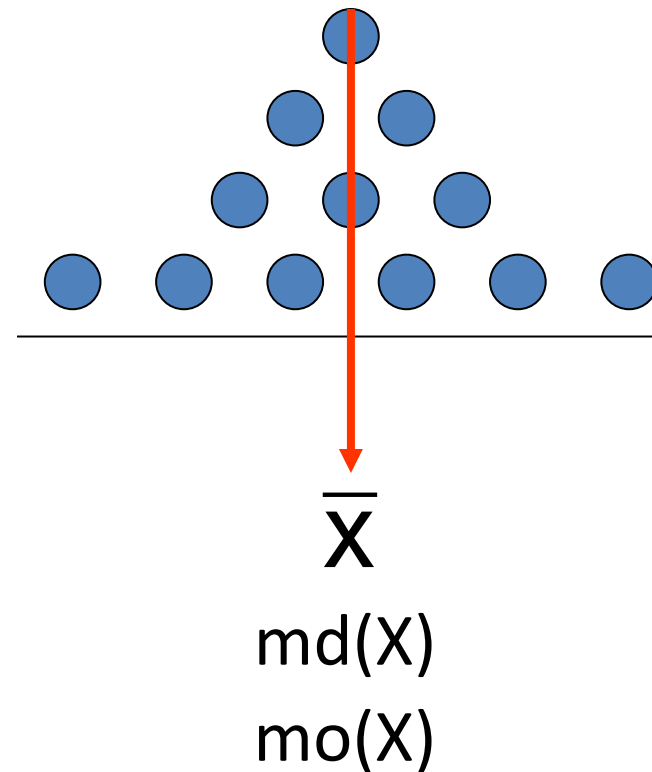
$\text{mo}(X)$  = classe com maior densidade

Para variáveis  
contínuas =



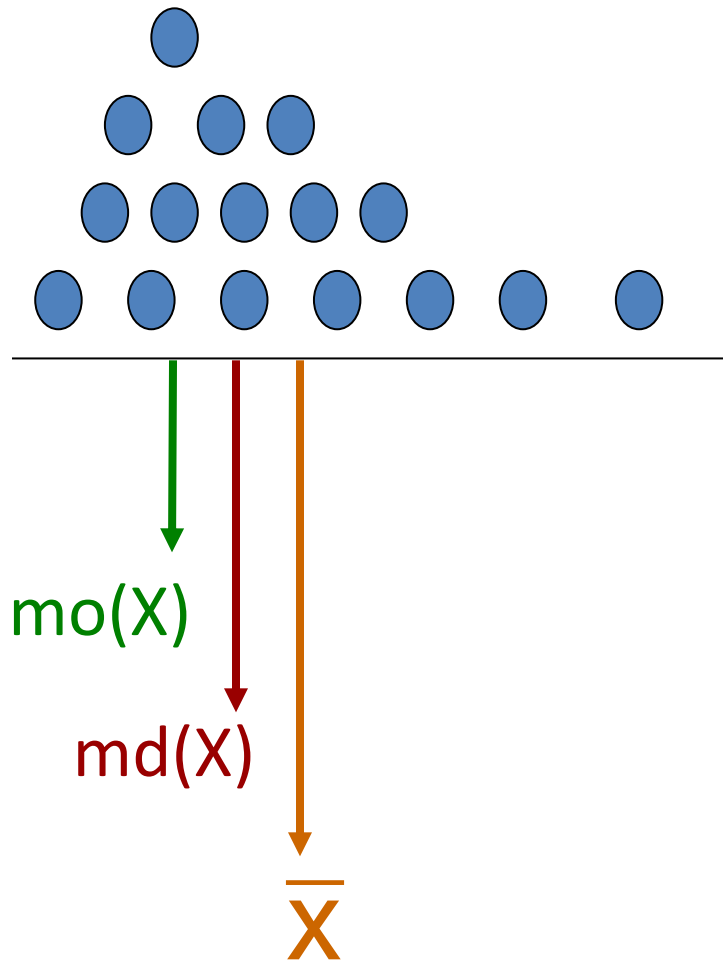
# Posição relativa

Distribuições  
Simétricas

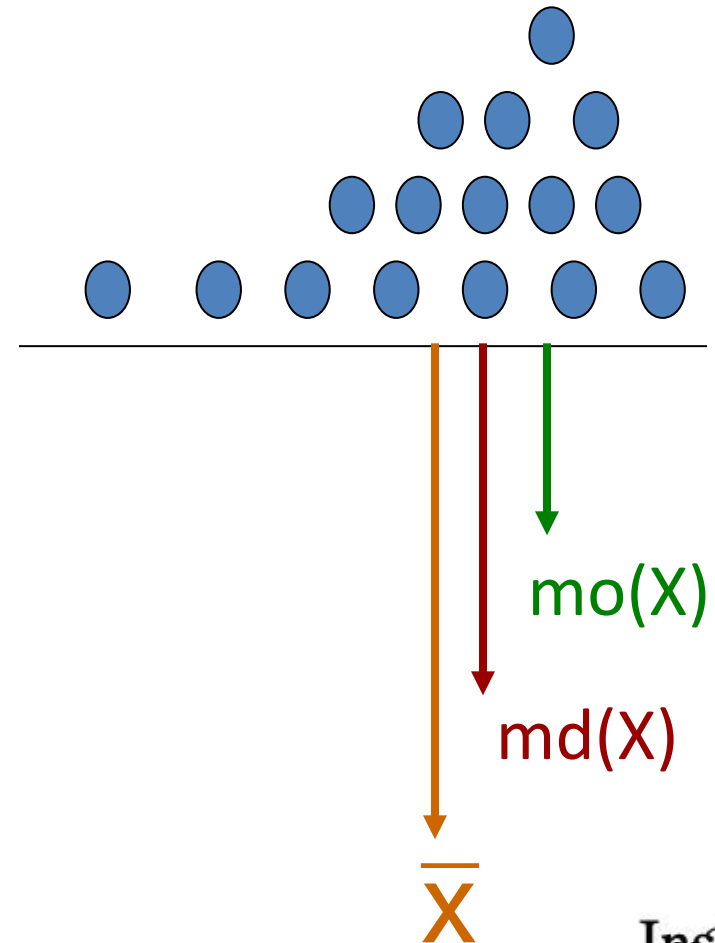


# Posição relativa

Assimetria à direita ou positiva



Assimetria à esquerda ou negativa



# Obtenção da **MÉDIA** via tabela de frequência

## Plano A

Frequências relativas:

[0.5, 4)            6.5

[4, 7.5)           19.6

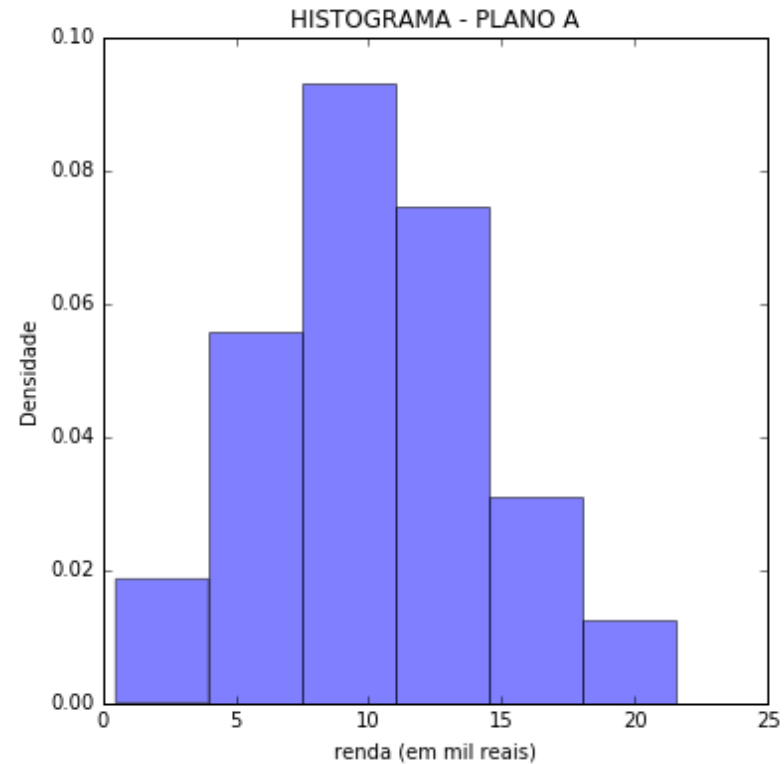
[7.5, 11)          32.6

[11, 14.5)        26.1

[14.5, 18)        10.9

[18, 21.5)        4.3

Name: RENDA, dtype: float64



**Considerar ponto médio de cada classe e ponderar pela frequência relativa (sem percentual).**

$$\begin{aligned}\bar{x} &= 2,25 * 0,065 + 5,75 * 0,196 + \dots + 19,75 * 0,043 = \\ &= 10,237\end{aligned}$$



# Obtenção da **MEDIANA** via tabela de frequência

## 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

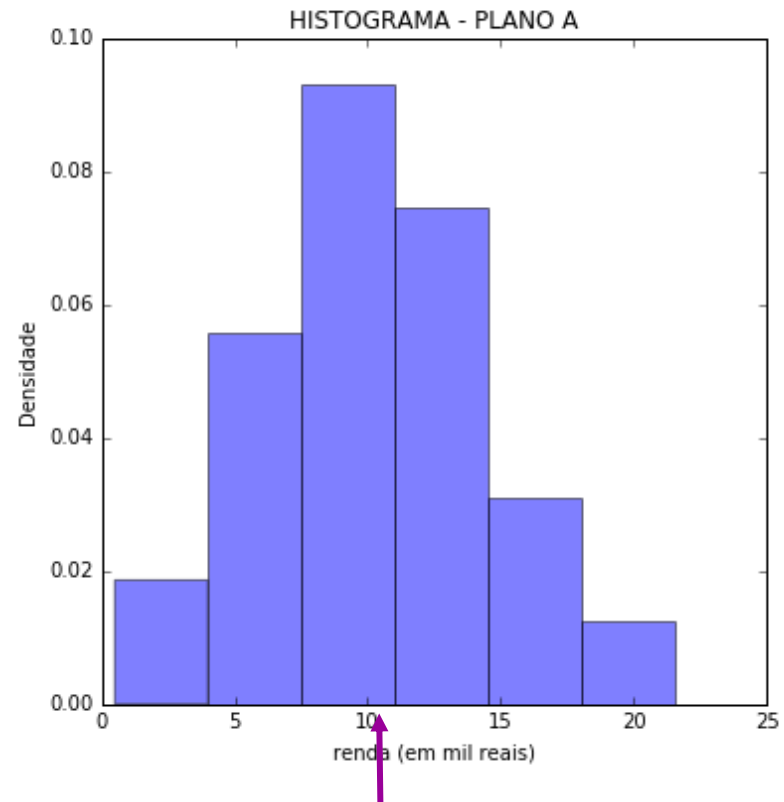
[18, 21.5)    4.3

Name: RENDA, dtype: float64

$11 - 7,5 \rightarrow 32,6\%$

$med - 7,5 \rightarrow 23,9\%$

$$med = 7,5 + 3,5 \cdot \frac{23,9}{32,6} = 10,066 \text{ mil reais}$$



Qual RENDA que divide os 50% que ganham menos dos 50% que ganham mais do plano A?

# Obtenção de uma frequência relativa via histograma

## 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

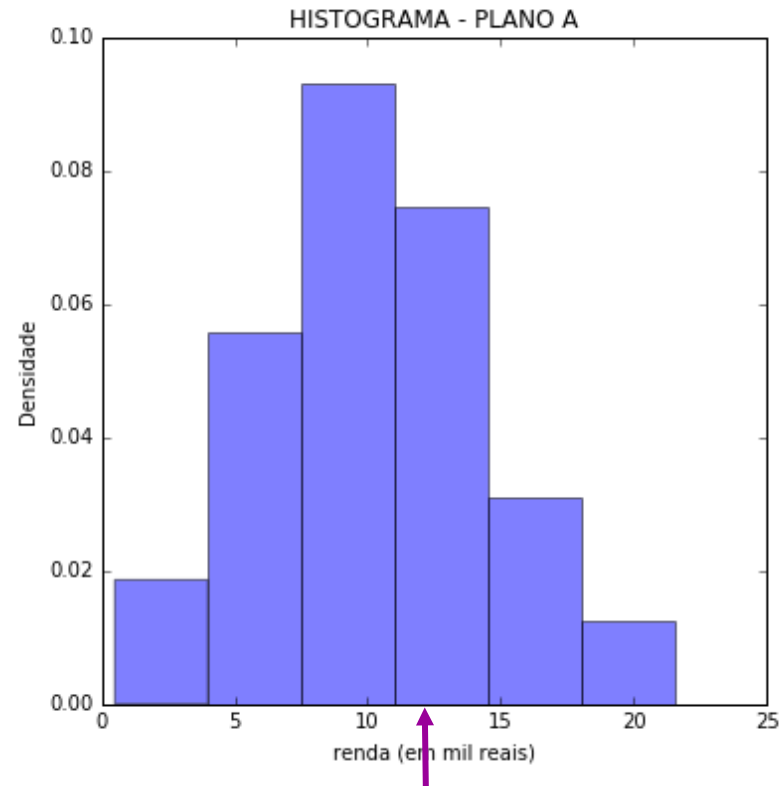
Name: RENDA, dtype: float64

14,5 - 11 → 26,1%

12 - 11 →  $x$

$$x = 26,1 \cdot \frac{1}{3,5} = 7,5\%$$

$$6,5\% + 19,6\% + 32,6\% + x = 66,2\%$$



Qual % de clientes do plano A que ganham até 12 mil reais?

# Atividade com Expectativa de Vida (variável quantitativa)

**?? minutos:**

**Análise descritiva da expectativa de vida de diversos países do mundo em três anos: 1800, 1926 e 2000.**

**Arquivo:**

Aula03	Atividade	Variáveis
Quantitativas	com	Expectativa
Vida.ipynb		de

# Preparo para próxima aula

Os alunos devem se preparar com:

1. Leitura prévia necessária: Magalhães e Lima (7ª. Edição): pág. 18 a 23 e pág. 114 a 117.
2. Python.