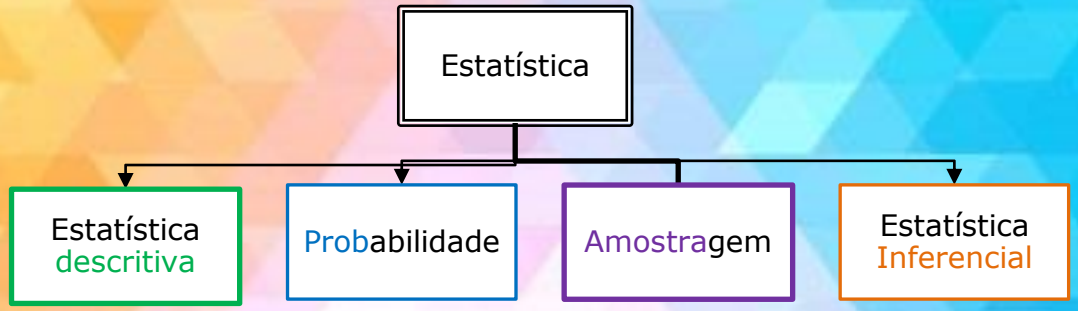
Introdução a Estatística

Prof. Thiago Marques

O que é Estatística?

- A **estatística** é um **conjunto de técnicas** que permite de forma **siste mática organizar, descrever, analisar e interpretar** dados advindos de diversas origens, a fim de **extrair deles conclusões.**

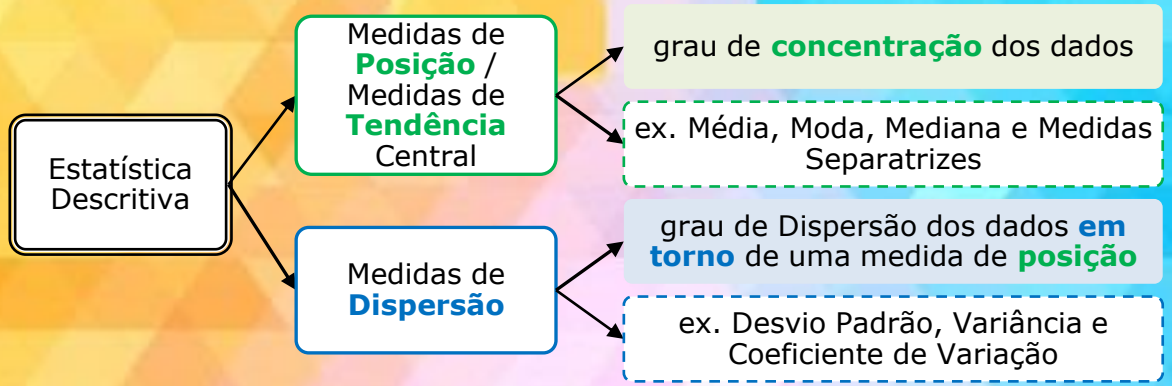
- Pode ser **subdivida** em **quatro grandes áreas**



Estatística Descritiva

- É o **ramo da estatística** que se ocupa em **organizar e descrever** os dados, que podem ser **expressos em tabelas e gráficos.**

- Pode ser dividida em **dois Grupos**:



Probabilidade

- Nos **permite descrever** os **fenômenos aleatórios**, ou seja, **aqueles** em que está **presente a incerteza.**

Amostragem

- Conjunto de **técnicas** para **selecionar** uma **amostra** da população, com o **objetivo** de obter **informações** de **uma ou mais** características de **interesse**, as quais permitam chegar a **conclusões** a respeito dos **parâmetros.**

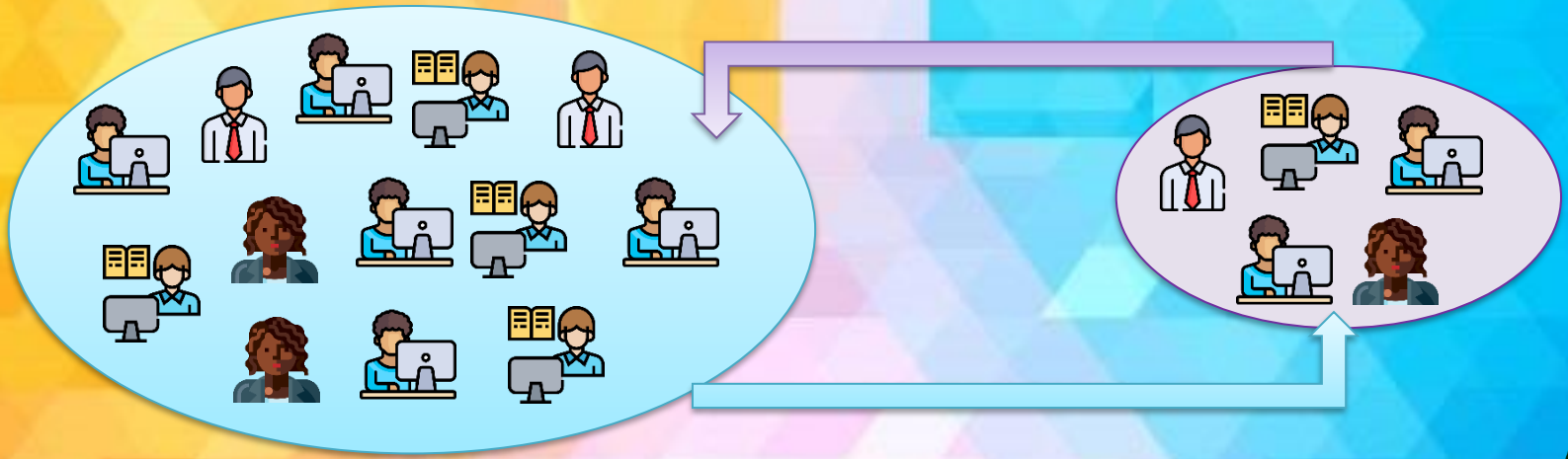
- **População**: É a **coleção de todos** os indivíduos que possuem **deter minadas características,** as quais estamos **interessados em estu dar.**
- Representamos por: N = "**Tamanho Populacional**".

- **Amostra**: É um **Subconjunto da população**, uma **parte** dos indiví duos que possuem **determinadas características**.
- Representamos por: n = "**Tamanho Amostral**".

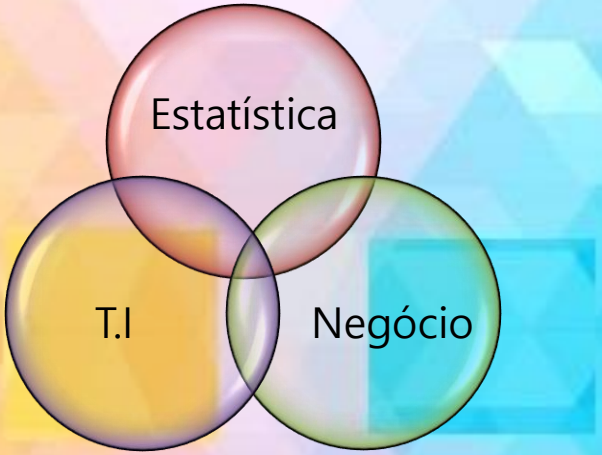
Importante: Sempre que falarmos em **Amostra**, usaremos **letras Minúsculas e População, por sua vez, Maiúsculas.**

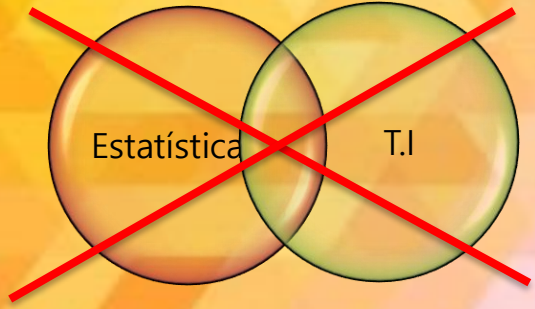
Inferência Estatística

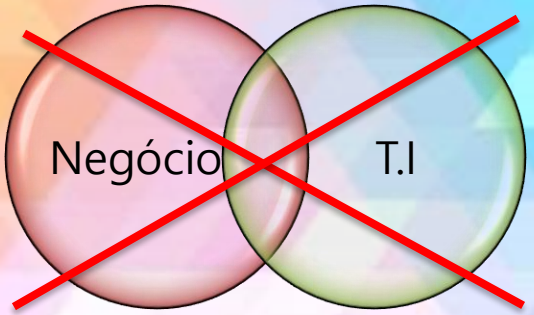
- É o **estudo de técnicas** que **possibilitam** a **extrapolação**, a um **grande conjunto de dados**, das **informações e conclusões** obtidas a partir da **amostra.**

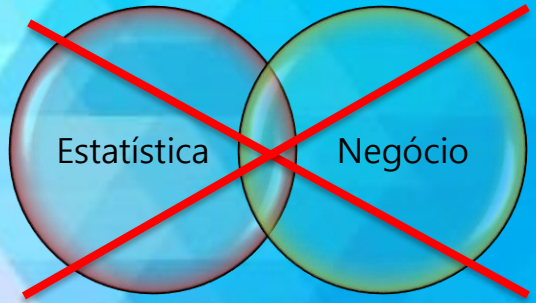


O que é Data Science?









Etapas da Análise Estatística

População
Características

Amostra

Informações Contidas nos dados

Conclusões sobre as características da População

Conceitos Básicos

- **Censo:** Exame de **todos os Elementos da população**.

- **Dados Brutos**: São dados na sua **forma mais primitiva, desprovidos de ordenação**, assim que coletados.

- **Rol Estatístico:** São os **dados brutos já ordenados**, em ordem **crescente ou decrescente**.

Dados Tabelados

- Os dados podem ser **expressos em tabelas de frequência,** tanto em Frequências Absolutas Simples (f_i) ou Frequências Absolutas Acumuladas (f_{ac})**,** podendo ser **subdividas** em **Frequências Relativas Simples** (fr) e **Frequências Relativas Acumuladas** (fr_{ac}).

- **Frequência Absoluta Simples** (f_i): É a **contagem simples** de Elementos.

Idade(Anos)	f_i
10	4
30	8
50	4
70	3
90	1

Dados Tabelados

- **Frequência Absoluta Acumulada (f_{ac}): É a contagem acumulada até a classe de interesse (Inclusive).**

Idade(Anos)	f_i	f_{ac}
10	4	4
30	8	(8+f0)=12
50	4	(4 + f1 +f2) =16
70	3	(3+ f1 + f2 + f3) =19
90	1	(1 + f1 + f2 + f3 + f4) =20

- **Frequência Relativa Simples (fr)**: É a **contagem simples** de Elementos, divididos pela **soma das frequências simples**, ou seja, representa a **proporção ou o percen tual de observações**.

Idade(Anos)	f_i	fr
10	4	4/20=0,2 ou 20%
30	8	8/20=0,4 ou 40%
50	4	4/20=0,2 ou 20%
70	3	3/20=0,15 ou 15%
90	1	1/20=0,05 ou 5%

Soma das f_i =20

Dados Tabelados

- **Frequência Relativa Acumulada** (fr_{ac}): É a **contagem acumulada até a classe de in teresse, divididos pela soma das frequências simples**.

Idade(Anos)	f_i	fr_{ac}
10	4	4/20=0,2 ou 20%
30	8	12/20=0,6 ou 60%
50	4	16/20=0,8 ou 80%
70	3	19/20 ou 0,95 ou 95%
90	1	20/20=1, ou 100%

- **Distribuições de Frequência em Classes:** Quando possuímos um **grande conjunto de dados**, se **agruparmos em classes**, teremos uma **boa ideia do comportamento** dos dados.

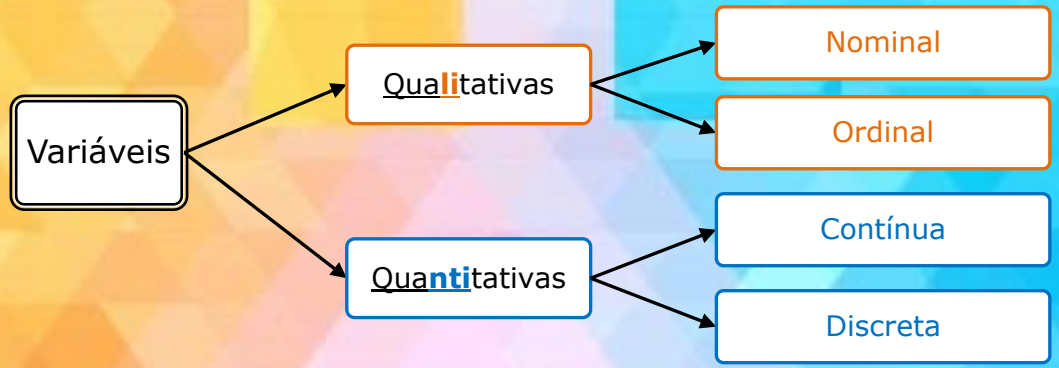
Idade(Anos)	f_i
2 ⊣ 7	4
7 ⊣ 12	8
12 ⊣ 17	4
17 ⊣ 22	3
22 ⊣ 27	1

Essa **notação significa** que o **7, que é o limite superior da primeira classe, está contido** no **intervalo, já o 2**, que é o **limite inferior da primeira classe, não está contido.**

O que são variáveis e quais são os seus Tipos?

- Qualquer **característica** associada a **uma população.**

- **Podem ser classificadas em:**



O que são variáveis e quais são os seus Tipos?

- **Variáveis Quantitativas: Podem ser divididas em dois grupos**, **discretas e contínuas**, **o primeiro**, quando for **finita e enumerável(contagem)** e o último quando os **resultados possíveis**, pertencerem a um **intervalo de números reais** e **resultados de mensuração.**

Exemplos:

Variáveis Quantitativas **Discretas:** Número de filhos, Número de carros e número de cigarros fumados por dia.

Variáveis Quantitativas **Contínuas:** Peso, altura e salário.

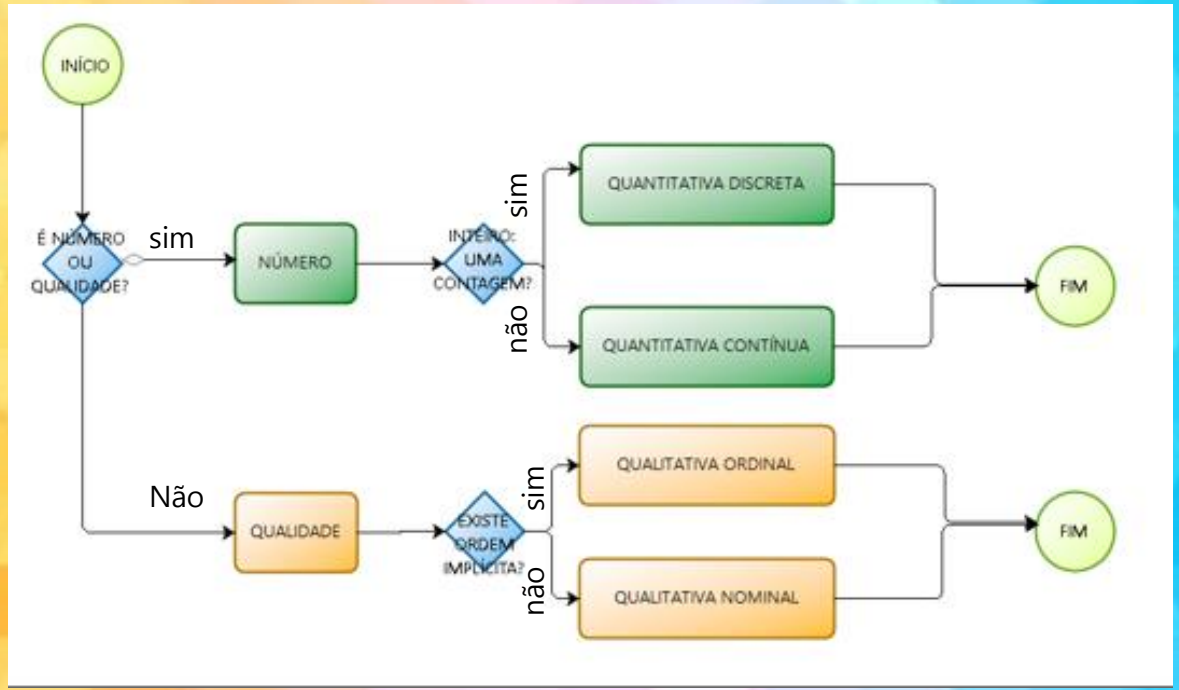
- **Variáveis Qualitativas: Representam atributos, qualidades**, que podem ser **divididas em dois grupos**, **ordinais e nominais, o primeiro**, quando existir **uma ordem implícita** e o último, quando **não existir uma ordem implícita**.

Exemplos:

Variáveis Qualitativas **Nominais:** Sexo, cor dos olhos, fumante/não fumante e doente/sadio.

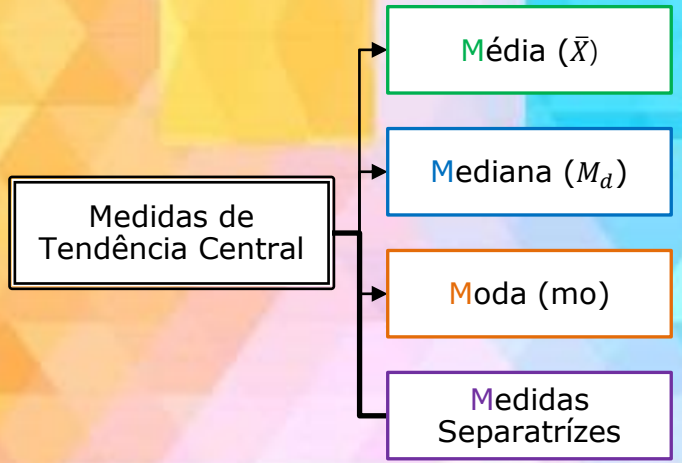
Variáveis Qualitativas **Ordinais:** Classe social, grau de instrução e estágio da doença.

Algoritmo do tipo de variáveis



Medidas de tendência central

- **Possibilitam** saber o **grau de concentração dos dados**, uma forma de **resumir** os seus dados **por meio de valores, representa tivos** do conjunto de dados.



Tipos de Médias

- **Média Aritmética** (MA)

É a **soma de todos** os elementos do conjunto, **divididos** pelo **número de elementos** que compõe o conjunto, essa nós estamos acostumados, sempre usamos para **auferir nossos resultados** no colégio.

Sua **fórmula** é dada por:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- **Média Geométrica** (MG)

É a **raíz n-ésima** do **produto de todos os elementos** que compõe o conjunto.

Sua **fórmula** é dada por:

$$MG = \sqrt[n]{x_1 . x_2 . \ldots . x_n}$$

Obs: **n** é o **número** de **elementos** que compõe o conjunto.

Tipos de Médias

- **Média Harmônica** (MH)

É o **número de elementos**, divididos pela soma dos **inversos** dos **elementos** que compõe o conjunto.

Sua **fórmula** é dada por:

$$MH = \frac{n}{\left(\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}\right)}$$

Relação entre as **médias**: $\mathbf{MA \geq MG \geq MH.}$

Importante: **Único caso** em que a **MA=MG=MH** é o caso onde **todos** os elementos possuem o **mesmo valor** no conjunto de dados!

Tipos de Médias

- **Exemplo didático para fixação do conteúdo:**

- Calcule a **média aritmética, média geométrica e a média harmônica**, para o seguinte conjunto de dados: **{1,2,5,3,4)**

Média **A**ritmética (MA)

- $MA = \bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \therefore \frac{(1+2+5+3+4)}{5} = 3{,}0$

Média **G**eométrica (MG)

- $MG = \sqrt[n]{x_1 . x_2 . \ldots . x_n} \therefore \sqrt[5]{1.2.5.3.4} \approx 2{,}605.$

Média **H**armônica (MH)

- $MH = \frac{n}{\left(\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}\right)} \therefore \frac{5}{\left(\frac{1}{1} + \frac{1}{2} + \frac{1}{5} + \frac{1}{3} + \frac{1}{4}\right)} \approx 2{,}19.$

- Como já sabíamos: **MA> MG >MH.**

Mediana

- É o **valor da variável** que **divide** os dados **ordenados** em **duas partes** de **igual frequência.**

- **Mediana** em dados **não divididos em intervalo de Classe:**

- **Primeiro Passo:**

  Colocar os **dados em rol** (**Ordenar** os dados de forma **crescente ou decrescente**)

- **Segundo Passo:**

  Observar a **paridade do n**, pois o cálculo da mediana **difere para n par e n ímpar.**

Mediana

- **Se n é ímpar:**

- ✓ Temos uma **posição central única**, dada por **P.C (Posição Central)= $\frac{n+1}{2}$**

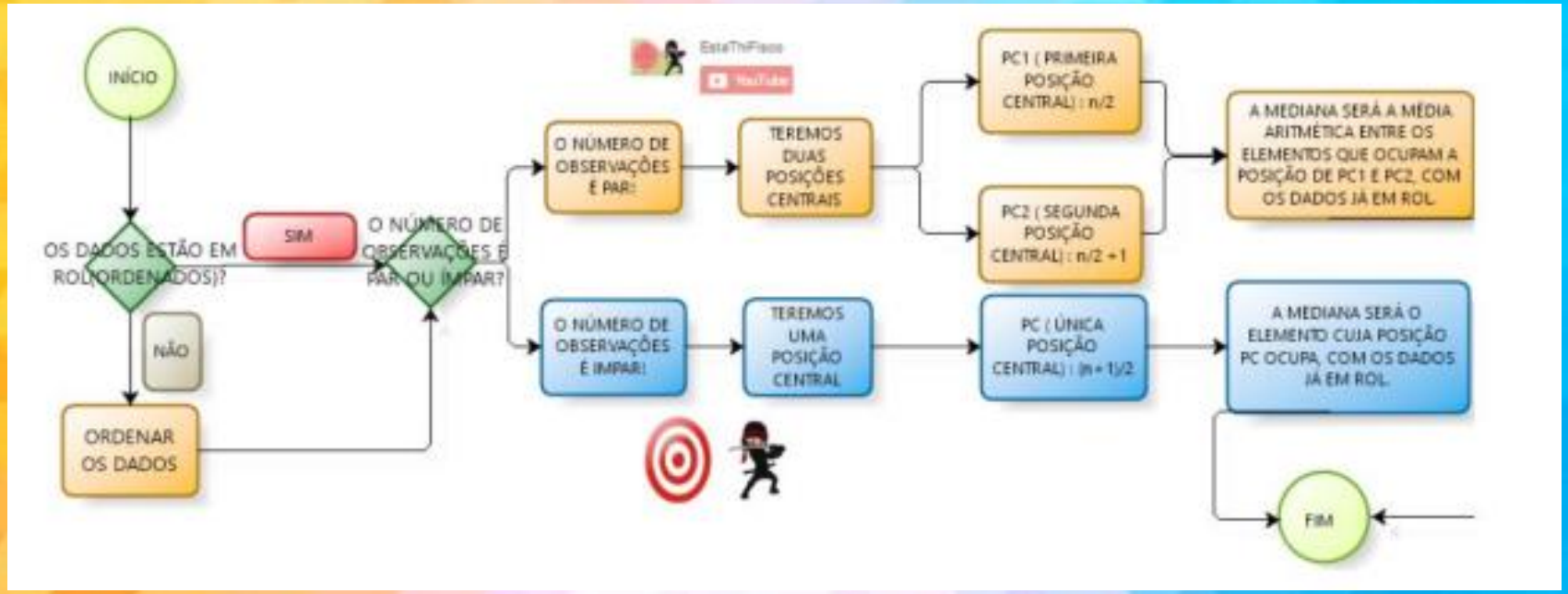
- ✓ **Após** calcularmos P.C , a **mediana** será o **valor** que **ocupa a posição central.**

- **Se n é par:** calcularemos **duas posições centrais,** quais sejam: **P.C1 (Posição Central1)= $\frac{n}{2}$ e P.C2(Posição Central2) = $\frac{n}{2}$ +1.**

- ✓ **Após** calcularmos **P.C1 e P.C2 , a mediana** será a **média aritmética de P.C1 E PC.2.**

Mediana

- O **processo** do **cálculo da mediana**, pode ser visto em forma de algoritmo no Diagrama a seguir:



Mediana

- **Exemplo didático para fixação do conteúdo:**

- **Calcule a mediana das observações: {7,1,5,2,3,1,6}**

- **Ordenar os dados :** {1,1,2,3,5,6,7}

Calcular P.C

n=7, temos somente **um P.C**:

$$\textbf{P.C} = \frac{n+1}{2} \therefore \textbf{P.C} = \frac{7+1}{2} = 4$$

- **Concluímos** então que a **mediana** é a **observação cuja posição**, com os dados ordenados (em rol), é a **quarta posição**, obtemos então que **Md=3.**

Mediana

- **Exemplo didático para fixação do conteúdo:**

- **Calcule a mediana das observações: {1,2,1,1,4,5,3,6}**

Ordenar os dados: **{1,1,1,2,3,4,5,6}**

Calcular P.C:
n=8, logo **par**, teremos **dois P.C´s**:

$$\textbf{P.C1=}\frac{\textbf{n}}{2}\textbf{=}\frac{\textbf{8}}{2}\textbf{=4 e P.C2=}\frac{\textbf{n}}{2}\textbf{+1= 5.}$$

- Logo, a nossa **mediana** será a **média aritmética** da **quarta observação e a quinta observação.**

$$\textbf{Md=}\frac{\textbf{2+3}}{2}\textbf{= 2,5.}$$

Moda

- É o valor que possui a **maior frequência simples no conjunto de dados**, consequentemente o de **maior probabilidade de ocorrência** em um conjunto de **dados não agrupados em classes.**

- **Exemplo didático para fixação do conteúdo:**

- **Calcule a moda do conjunto {4,5,4,6,5,8,4}**
- Vamos realizar a **tabela de frequências** para facilitar a nossa visualização:

x	f_i
4	3
5	2
6	1
8	1

- Percebemos pela **tabela de frequências** que a **moda é 4**, pois possui a maior **frequência simples** do conjunto, logo a **moda é única** e chamada de **unimodal.**

Moda

- **Calcule a moda do conjunto {4,5,4,6,5,8,4,4,5,5}**

x	f_i
4	4
5	4
6	1
8	1

- Percebemos pela **tabela de frequências** que a **moda possui dois valores**, pois **duas observações** do conjunto **se repetem 4 vezes**, portanto as **maiores frequência simples** do conjunto, logo a **moda é 4 e 5** e é chamada de **bimodal.**

Moda

- **Calcule a moda do conjunto {1,2,3,4,5}:**

x	f_i
1	1
2	1
3	1
4	1
5	1

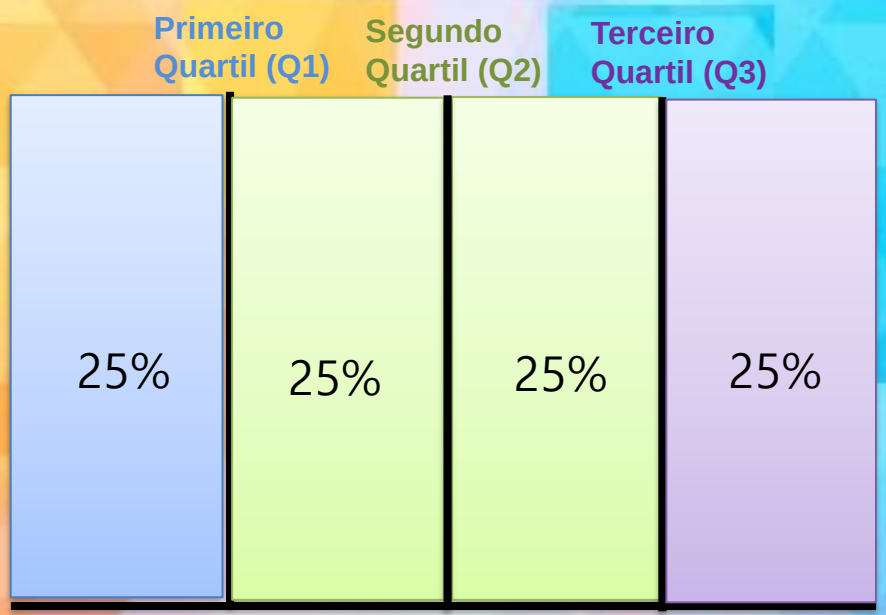
- Percebemos pela **tabela de frequências** que a **moda não possui valor**, pois **todas as observações do conjunto se repetem nenhuma vez,** portanto é chamada de **amodal.**

Importante: Um conjunto de dados que possui **duas modas** é chamado **bimodal** , **mais de duas, multimodal** e se **não possuir moda**, é um conjunto **amodal.**

Medidas separatrizes

- Tem como **objetivo dividir** o conjunto de dados em **n partes de igual frequência**, os mais utilizados são os **quartis e os percentis.**

- **Quartis: Dividem** o conjunto em **quatro partes Iguais.**



Medidas separatrizes

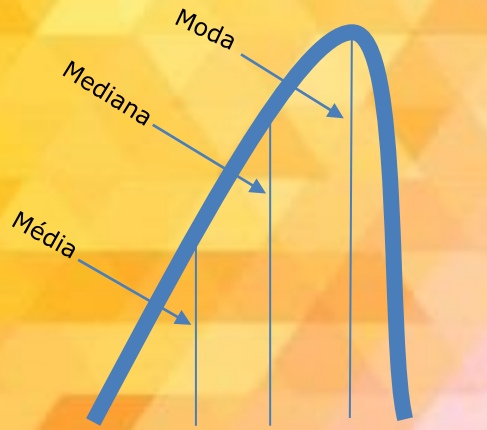
- **Percentis:** Dividem o conjunto em **100 partes iguais.**

- Percentil **Vinte e Cinco (P25=Q1).**
- Percentil **Cinquenta (P50=Q2=Md).**
- Percentil **Setenta e Cinco (P75=Q3).**

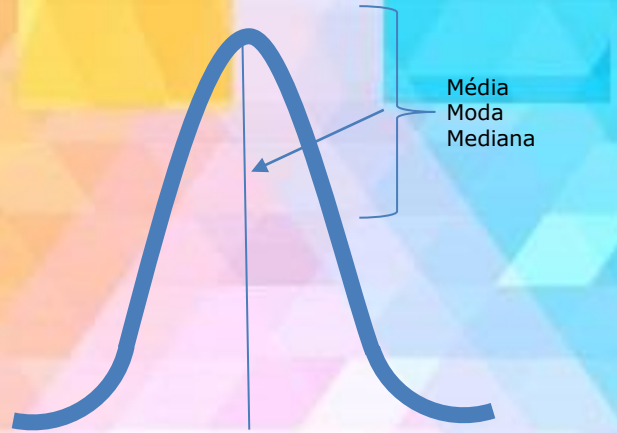
Medidas de Assimetria

- Possibilitam **analisar uma distribuição em relação a sua moda, mediana e média.**

Pense no **conceito de simetria**, fazendo a **analogia a um espelho**, se traçarmos um **eixo vertical no meio da curva**, e **enxergarmos o mesmo de um lado, o que você vê do outro**, significa que seus **dados são simétricos.**



Assimétrica negativa / à esquerda
(Média < Mediana < Moda)



Simétrica
(Média=Moda=Mediana)

Assimétrica à direita
(Média > Mediana > Moda)

Coeficiente de Assimetria

- Primeiro **Coeficiente de Pearson**:

$$AS = \frac{Média - Moda}{Desvio\ Padrão}$$

➢ O **desvio padrão** é sempre positivo (**>0**).

➢ Note que quando a **média** for **igual** a **moda** -> (**AS=0**), a distribuição será **simétrica**, terá **ausência** de assimetria.

➢ Note que quando a **média** for **menor** que a **moda** -> (**AS<0**), a distribuição será **assimétrica** à **esquerda** ou **negativa.**

➢ Note que quando a **média** for **maior** que a **moda** -> (**AS>0**), a distribuição será **assimétrica** à **direita** ou **positiva.**

Coeficiente de Curtose

- $K = \frac{Q3-Q1}{2.(P90-P10)}$

- $Q3 = valor\ do\ 3^{o}\ Quartil$; $Q1 = valor\ do\ 1^{o}\ Quartil$; $P90 = valor\ do\ 90^{o}\ Percentil$; $P10 = valor\ do\ 10^{o}\ Percentil$

- O coeficiente de **curtose** da distribuição normal é **aproximadamente 3**, e utilizamos como base de comparação:

- Se k=3, então chamamos de **Mesocúrtica** (Grau de achatamento da curva **normal**).

- Se k>3, então chamamos de **Leptocúrtica** (Mais alongada (**Pontiaguda**)).

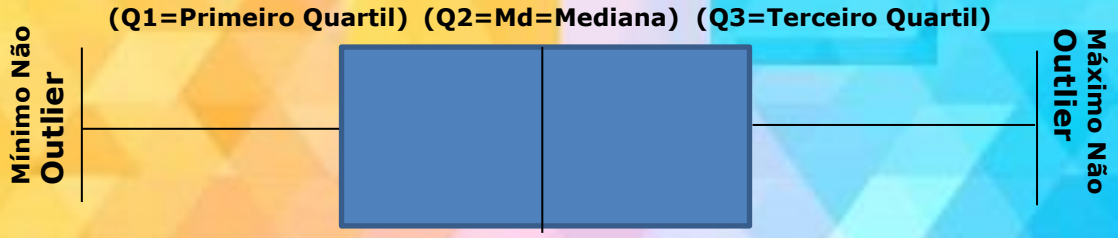
- Se k<3, então chamamos de **Platicúrtica** (Mais achatada (**platô**)).

Importante:
no **software R**:

K=0, Mesocúrtica.
K<0, Platicúrtica.
K>0, Leptocúrtica.

Diagrama de Box-Plot ou Diagrama de Caixas

- É uma **representação gráfica** da distribuição dos dados, nos dá informação da **assimetria da distribuição, presença de outliers** (Valores Atípicos) e da **Variabilidade** dos dados, por meio da **ampli tude** (Máx-Min).



Limite Inferior = Q1 – 1,5ΔQ, sendo ΔQ=Q3-Q1

Limite Superior = Q3 + 1,5ΔQ, sendo ΔQ=Q3-Q1

Medidas de Variação/Dispersão

- Nos permitem saber o **grau** de **dispersão** dos dados, em relação a uma medida de tendência central, geralmente a média.

Exemplos: **Amplitude**, **Variância**, **D**esvio **P**adrão, **C**oeficiente de **V**ariação.

População

- **Amplitude** Populacional**:** É a diferença entre o maior e o menor valor da População:
  - H = (Máximo - Mínimo)

- **V**ariância **P**opulacional:
  - $\sigma^2 = \frac{(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \cdots + (x_N-\bar{x})^2}{N} = \frac{\sum_{i=1}^{N}(x_i-\bar{x})^2}{N}$

- **D**esvio **P**adrão Populacional:
  - $\sigma = \sqrt{\sigma^2}$

- **C**oeficiente de **V**ariação Populacional:
  - $CV_x = \frac{\sigma}{\mu}$

Medidas de Variação/Dispersão

Amostra

- Amplitude Amostral – É a **diferença** entre o **maior e o menor** valor da **amostra**.

h = máximo – mínimo

- **Variância:**
  - $s2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$

- **D**esvio **P**adrão Amostral
  - $s = \sqrt{s^2}$

- **C**oeficiente de **V**ariação
  - $cv_x = \frac{s}{\bar{x}}$

Importante: O **cv** é a **única medida de variação adimensional** (Não possui unidade de medida).
Em geral, consideramos um **coeficiente de variação < 25%**, um **bom indicador de homogeneidade** dos dados!

Algoritmo da Variância

➢O **processo** do **cálculo da variância**, pode ser visto em forma de algoritmo no Diagrama a seguir:

Medidas de Variação/Dispersão

- **Exemplo didático para fixação do conteúdo:**

Considere a amostra :
{3,4,5,6,12}
Calcule: Amplitude, variância, desvio padrão e coeficiente de variação:

Amplitude
h = (máximo - mínimo) = 12 – 3 = 9.

Variância:

$$s^2=\frac{(3-6)^2+(4-6)^2+(5-6)^2+(6-6)^2+(12-6)^2}{5-1}=\frac{50}{4}=\mathbf{12,5.}$$

Desvio Padrão

$s=\sqrt{12,5}\approx$ **3,53.**

Coeficiente **de V**ariação

$cv_x=\frac{s}{\bar{x}}$

$cv_x=\frac{3,53}{6}$ **=** 0,5883 ou **58,83%**