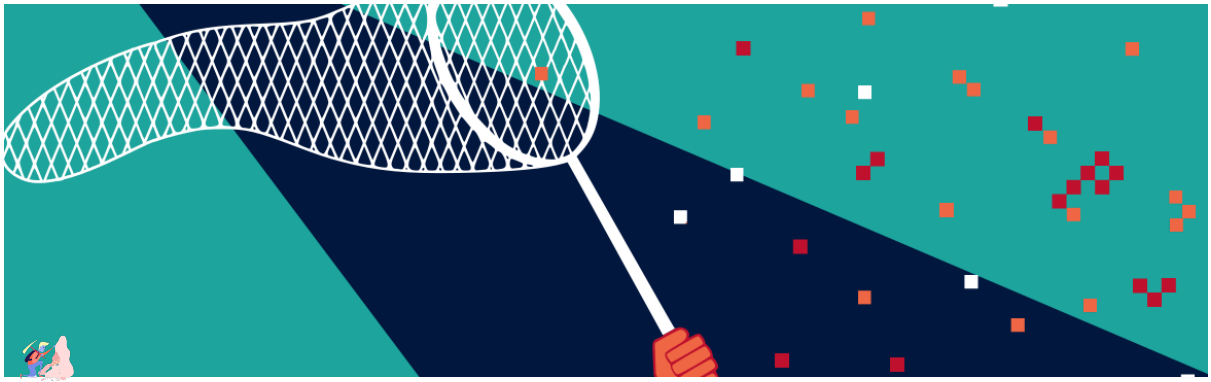


Análise Exploratória de Dados

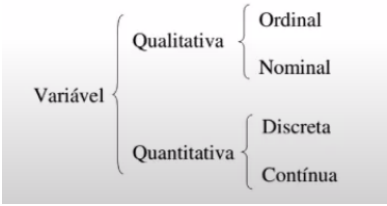
▼ Estatística Descritiva

- A estatística descritiva é um ramo da estatística que aplica várias técnicas para descrever e sumarizar um conjunto de dados. Diferencia-se da estatística inferencial, ou estatística indutiva, pelo objectivo: organizar, sumarizar dados ao invés de usar os dados em aprendizado sobre a população.

▼ Conceitos

▼ **Variável:** É uma característica dos elementos de uma população ou de uma amostra, que pode assumir diferentes valores de interesse ao estudo, sejam eles numéricos ou não.

- As variáveis podem se classificar em:



- **Variável qualitativa:** É aquela que tem como possíveis valores, atributos ou qualidades. Também denominadas variáveis categóricas. **Ex:** `cor dos cabelos`, `marcas de refrigerantes`, `cor dos olhos`, etc.
  - **Qualitativa nominal:** Quando os atributos ou qualidades assumidas não apresentam nenhuma ordem de ocorrência. **Ex:** `cor dos olhos`, `marcas de carros`, etc.
  - **Qualitativa ordinal:** Quando há alguma ordem nos elementos. **Ex:** `grau de escolaridade`, `grau de satisfação de clientes`, etc.
- **Variável Quantitativa:** É aquela cujos possíveis valores são numéricos. **Ex:** `Peso`, `altura`, `número de faltas`, etc.
  - **Discreta:** Assume valores inteiros. **Ex:** `número de filhos`, `número de habitantes`, etc.
  - **Contínua:** Assume qualquer valor em um determinado intervalo. **Ex:** `Altura`, `peso`, `velocidade`, etc.

- **Dado estatístico:** É toda informação devidamente coletada e registrada. Todo dado se refere a uma variável.
- **População:** Conjunto de indivíduos ou objetos que possuem ao menos uma característica em comum.
- **Amostra:** É qualquer subconjunto não vazio de uma população.
- **Parâmetro:** É uma característica numérica estabelecida para toda uma população.
- **Rol:** É a organização dos dados brutos em ordem crescente ou decrescente.
- **Dados Brutos:** É apresentação dos dados na sequência em que foram coletados, isto é, sem nenhuma ordenação numérica.
- **Amplitude:** É o maior valor subtraído do menor valor do conjunto de dados.

▼ **Símbolos**

▼ **Fases do Método Estatístico**

- Quando se pretende empreender um estudo estatístico completo, há várias fases do trabalho que devem ser desenvolvidas para se chegar aos resultados finais do estudo.
- Essas etapas ou operações são chamadas **Fases do Método Estatístico** e são de âmbito da Estatística Descritiva.

- **Definição do problema:** Saber exatamente aquilo que se pretende pesquisar é o mesmo que definir corretamente o problema. Definir população e parâmetros de interesse.
- **Planejamento:** Como levantar informações? Que dados deverão ser obtidos? Cronograma de atividades, Custos envolvidos, etc.
- **Coleta de dados:** Fase operacional. E o registro sistemático de dados, com um objetivo determinado.
  - **Coleta direta:** Quando os dados são obtidos diretamente da fonte. A coleta direta pode ser:
    - **Contínua:** (registros de nascimento, óbitos, casamentos, etc).
    - **Periódica:** (recenseamento demográfico, censo industrial) e ocasional (registros de casos de dengue).
  - **Coleta indireta:** É feita por deduções a partir dos elementos conseguidos pela coleta direta, por analogia, por avaliação, indícios ou proporcionalidade.

- **Apuração dos dados:** Resumo dos dados através de sua contagem e agrupamento. E a condensação e tabulação dos dados.
- **Apresentação dos dados:** Há duas formas de apresentação, que não se excluem mutuamente.

- **Apresentação tabular:** É uma apresentação numérica dos dados em linhas e colunas distribuídas de modo ordenado, segundo regras práticas fixadas.
- **Apresentação gráfica:** Constitui uma apresentação geométrica permitindo uma visão rápida e clara das características de interesse.

- **Análise e interpretação dos dados:** Esta fase está ligada essencialmente ao cálculo de medidas descritivas com a finalidade de descrever o fenômeno (estatística descritiva). E também nesta fase que se realiza o processo de generalização que é de âmbito da Inferência Estatística.

▼ **Modelo determinístico**

- Um modelo é determinístico quando tem um conjunto de entradas conhecido e do qual resultará um único conjunto de saídas.
- Em geral, um sistema determinístico é modelado analiticamente, isto só não ocorre quando o modelo se torna muito complexo envolvendo um grande número de variáveis ou de relações.

▼ **Modelo Não Determinístico / Modelo Empírico**

▼ **Modelo de regressão**

- Inclui somente duas variáveis: uma **independente** e uma **dependente**.
- A variável dependente é aquela que está sendo explicada, enquanto a variável independente é aquela que é utilizada para explicar a variação na variável dependente.

▼ **Classes**

- É uma tabela onde os valores observados são agrupados em intervalos (**classes**) de variação da variável em questão.
- A quantidade de classes é representada por k.
- A escolha do número de classes pode ser arbitrária ou pode seguir algum indicador. Os dois indicadores mais utilizados são:

  1) Regra da raiz quadrada:

  Se o número de observações for menor igual a 25, o k vai ser igual a 5: $k = 5$ para $n \leq 25$.

  Se o número de observações for maior que 25, o k vai ser igual a raiz de n: $k = \sqrt{25}$ para $n > 25$.

  2) Regra de Sturges:

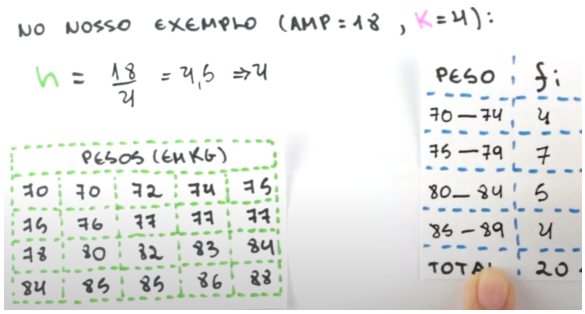
  $$k = 1 + 3,3 \cdot log_{10} n$$

  A desvantagem dessa regra é quando o n for muito grande.

  Obs: Se o número de observações for maior que 100, deve ser usada a Regra da raiz quadrada.

  ▼ **Intervalo de cada classe**

  - Intervalo é a quantidade de valores que cada classes ira ter.
  - A formula é: $h = \frac{AMP}{K}$, o h é o intervalo, o AMP é a amplitude do conjunto de dados, o K é o número de classes que o conjunto de dados irar ter.



  ▼ **Maneiras de expressar as classes**

  - Especialmente no caso de variáveis contínuas, são utilizados intervalos de classes para construção de distribuições de frequências. O agrupamento, acarreta perda de informação uma vez que, os dados originais não podem ser recuperados a partir dos dados agrupados.

- Para se agrupar dados com o máximo de informação e representatividade, alguns critérios devem ser seguidos.
- Existem diversas maneiras de expressar as classes:

a) a ⊢⊣ b	compreende todos os valores entre a e b, incluindo a e b
b) a ⊣ b	compreende todos os valores entre a e b, excluindo a
c) a ⊢ b	compreende todos os valores entre a e b, excluindo b
d) a — b	compreende todos os valores entre a e b, excluindo a e b

▼ **Limites de classe**

- São os valores extremos de cada classe. O menor valor de uma classe i denomina-se limite inferior da classe $i(li)$, e o maior é o limite superior da classe $i(Li)$.

▼ **Amplitude de classe**

- A amplitude de classe é o comprimento da classe, sendo definida como a diferença entre o limite superior e o limite inferior da classe. Notação hi, $h_i = L_i - l_i$. Ou seja é o maior valor subtraído do menor valor da
- Na maioria das aplicações consideramos $h1, h2, ..., hk = h$, ou seja, classes com amplitudes iguais dadas por: $h = \frac{A_T}{k}$. O A_T é a amplitude total e o k é a quantidade de classes.
  - **Exemplo:** Dados: $A_T = 64$ e $k = 7$. Temos, $h = \frac{64}{7} = 9,14$.

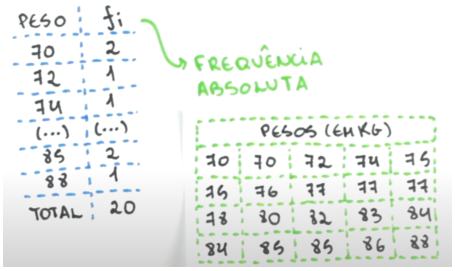
▼ **Frequência**

Símbolo	Sigificado
i	Classe
h	Intervalo de cada classe
k	Número de classes
f_i **ou** n_i	Frequência Absoluta / Simples
F_i **ou** N_i	Frequência Acumulada
fr_i **ou** f_i	Frequência Relativa
Fr_i	Frequência Acumulada Relativa

- Em estatística, a frequência (ou frequência absoluta) de um evento i é o número n_i de vezes que o evento ocorreu em um experimento ou estudo. Essas frequências são normalmente representadas graficamente em histogramas.
- Ou seja, é a quantidade de vezes que um determinado elemento se repete.

▼ **Frequência Absoluta / Simples**

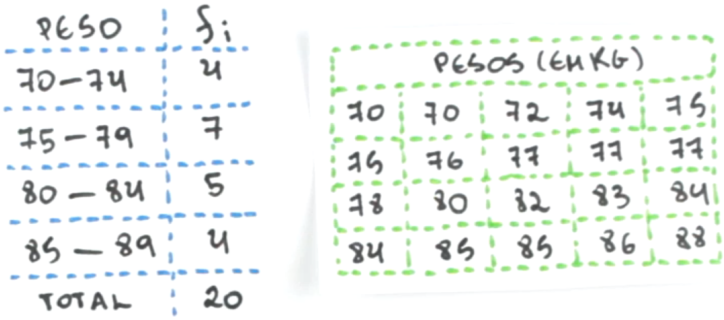
- É o número de vezes que um determinado elemento se repete dentro de uma amostra.
- É representado pelo simbolo f_i.



▼ **Distribuição de frequência por classes:**

- É uma tabela onde os valores observados são agrupados em intervalos (**classes**) de variação da variável em questão.
- Normalmente usado para representar variáveis **contínuas** ou variáveis **discretas** que possuem um grande número de valores observados.

- A frequência ou frequência absoluta não vai ser mais a quantidade de vezes que um determinado elemento se repete, mais sim a quantidade de elementos que tem em cada classe.



PESO	f_i
70 – 74	4
75 – 79	7
80 – 84	5
85 – 89	4
TOTAL	20

PESOS (EM KG)

70	70	72	74	75
76	76	77	77	77
78	80	82	83	84
84	85	85	86	88

- **Obs:** No capítulo anterior (Classes), explicamo quantas classes devemos criar para uma determinada amostra.

▼ **Distribuição de frequência simples:**

- É uma tabela onde os valores da variável analisada aparecem individualmente relacionados as suas frequências.
- Normalmente usada para representar variáveis quantitativas discretas ou qualitativas.
- A diferença dessa para a Distribuição de frequência por classes, é o fato dessa mostra só a frequência de um determinado valor, enquanto que a Distribuição de frequência por classes mostra a frequência de um grupo.

▼ **Frequência Acumulada:**

- É a soma de todas as frequências absolutas anteriores.
- Ela é representada pelo simbolo F_i.

Peso	f_i	F_i
70 - 74	4	4
75 - 79	7	11
80 - 84	5	16
85 - 89	4	20
Total	20	

▼ **Frequência Relativa:**

- Ela é a Frequência absoluta (f_i) da classe dividida pelo valor total de elementos da amostra: $fr_i = \frac{f_i}{n}$.
- Ela é representado pelo símbolo fr_i.

Peso	f_i	fr_i
70 - 74	4	0.2
75 - 79	7	0.35
80 - 84	5	0.25
85 - 89	4	0.2
Total	20	

Obs: Sempre a soma de todos os valores da frequência relativa absoluta tem que ser igual a `1`.

- **Diferença entre Frequência Simples e Relativa:**
  - **Simples:** Diz o número de observações.
  - **Relativa:** Diz a proporção de observações.

▼ **Frequência Acumulada Relativa:**

- É a F_i (frequência acumulada) divido pelo n (número total de elementos da amostra).
- Ela é representada pelo símbolo Fr_i.

Peso	f_i	F_i	Fr_i
70 - 74	4	4	0.2
75 - 79	7	11	0.55
80 - 84	5	16	0.8
85 - 89	4	20	1
Total	20		

▼ **Ex1:**

i	X_i	n_i	f_i	$f_i(\%)$	N_i	F_i	$F_i(\%)$
1	0	4	0,20	20	4	0,20	20
2	1	8	0,40	40	12	0,60	60
3	2	3	0,15	15	15	0,75	75
4	3	3	0,15	15	18	0,90	90
5	4	1	0,05	5	19	0,95	95
6	5	1	0,05	5	20	1,00	100
Total		20	1,00	100	–	–	–

▼ **Ex2:**

- Em uma amostra de 30 milheiros de telhas recebidas pela Construtora DEMA Ltda, constatou-se os seguintes números de unidades defeituosas por milheiro:

  5 − 20 − 10 − 5 − 40 − 30 − 20 − 5 − 10 − 15 − 10 − 30 − 40 − 10 −50 −10 − 30 − 15 − 20 − 40 − 10 − 20 − 20 − 50 − 10 − 40 − 30 − 20 − 0 − 30

- **Obs:** Cada valor apresentado anteriormente, representa a quantidade de telhas defeituosas para cada milheiro de telha.

(a) Apresente os dados em uma distribuições de frequências simples.

i	X_i	n_i	N_i	f_i	F_i
1	0	1	1	0,033	0,033
2	5	3	4	0,100	0,133
3	10	7	11	0,233	0,366
4	15	2	13	0,067	0,433
5	20	6	19	0,200	0,633
6	30	5	24	0,167	0,800
5	40	4	28	0,133	0,933
6	50	2	30	0,067	1,000
	Total	30	–	1,00	–

(b) Qual a porcentagem de milheiros com mais de 30 telhas defeituosas?

$f_7 + f_8 = 0,133 + 0,067 = 0,200 = 20\%$

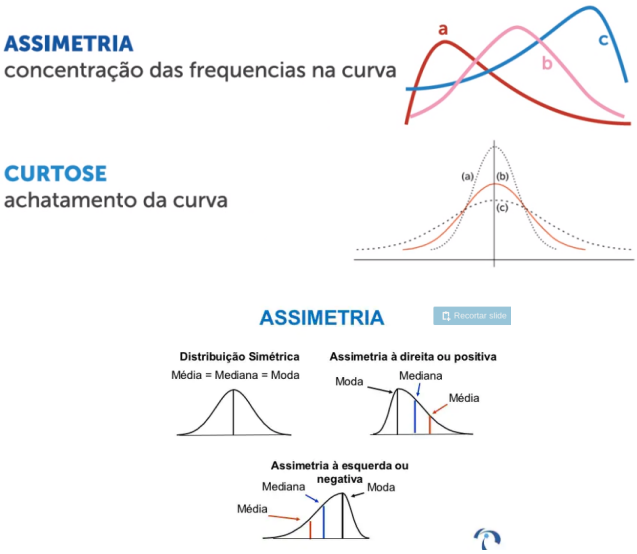
(c) Quantos milheiros tiveram menos de 10 telhas defeituosas?

$N_2 = 4$

(d) Qual a proporção de milheiros com menos de 20 telhas defeituosas?

$F_4 = 0,433$

▼ **Simetria**



O ponto mais alto de um gráfico é a moda.

- **Simétrica:** O lado esquerdo da moda é igual ao lado direito da moda. Também podemos identifica a simetria quando a Média, Moda e Mediana tiverem valores iguais.
- **Assimétrica á Direita ou positiva:** Quando a média se encontra mais a direita.
- **Assimétrica á Esquerda ou negativa:** Quando a média se encontra mais a esquerda.

▼ **Coeficiente de Assimetria de Pearson**

- Definimos o coeficiente de assimetria de Pearson por $CA_p = \frac{m_3}{m_2^{3/2}}$. O m_3 e o m_2 são momentos centrado na média.
- CA_p é uma medida adimensional.
- Valores de CA_p próximos a zero indicam **Simetria**. Valores negativos indicam **Assimetria a Esquerda**, enquanto valores positivos indicam **Assimetria a Direita**.
- **Ex**: Os dados apresentados se referem as notas de 20 alunos submetidos a provas de Estatística Elementar, em três situações diferentes. As provas possuem 4 questões. A cada estudante foi atribuída um nota que pode variar de zero a dez. Determine o coeficiente de assimetria de Pearson desses dados.

Tabela 1: Notas de 20 alunos

Notas			
0,5	2,6	5,1	7,5
1,2	2,7	4,5	7,2
0,8	4,0	5,4	8,2
3,8	5,6	6,3	8,3
2,9	4,5	6,7	9,3

Notas	n_i	x_i	$(n_i \cdot x_i)$	$(n_i \cdot x_i^2)$	$(n_i \cdot x_i^3)$
$[0, 2)$	3	1	3	3	3
$[2, 4)$	4	3	12	36	108
$[4, 6)$	6	5	30	150	750
$[6, 8)$	4	7	28	196	1372
$[8, 10)$	3	9	27	243	2187
Total	20	-	100	628	4420

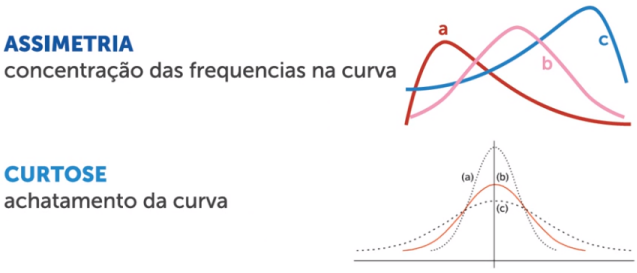
$$
\begin{aligned}
m'_1 &= \frac{\sum_{i=1}^{5} n_i \cdot x_i}{20} = \frac{100}{20} = 5 \\
m'_2 &= \frac{\sum_{i=1}^{5} n_i \cdot x_i^2}{20} = \frac{628}{20} = 31.4 \\
m'_3 &= \frac{\sum_{i=1}^{5} n_i \cdot x_i^3}{20} = \frac{4420}{20} = 221
\end{aligned}
$$

$$
\begin{aligned}
m_3 &= m'_3 - 3m'_1 \cdot m'_2 + 2(m'_1)^3 \\
m_3 &= 221 - 3 \cdot 5 \cdot 31.4 + 2 \cdot 5^3 \\
m_3 &= 221 - 471 + 250 \\
m_3 &= 0
\end{aligned}
$$

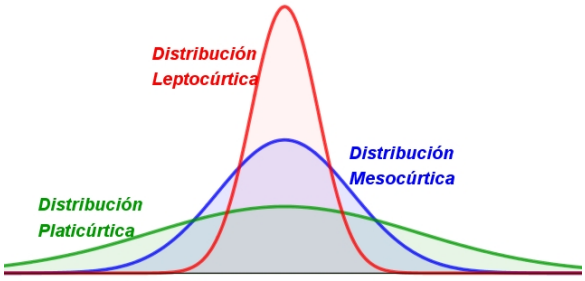
$$
CA_p = \frac{m_3}{(m_2)^{\frac{3}{2}}} = \mathbf{0}
$$

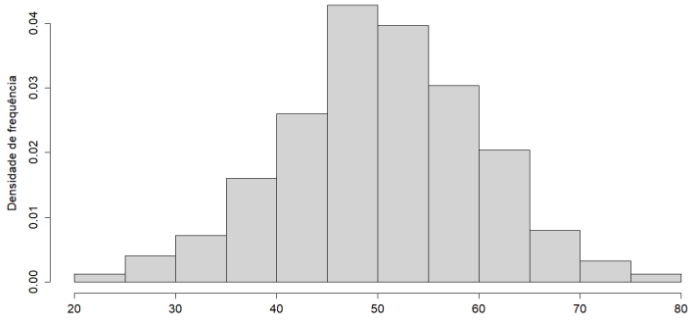
-

▼ **Curtose**

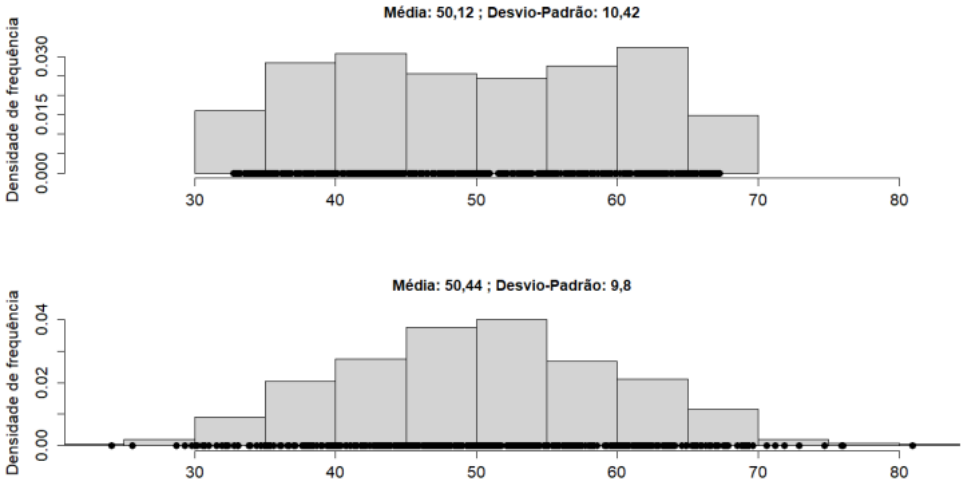


- Assim como a assimetria, a curtose fornece informações importantes a respeito da distribuição dos dados observados.
- Especificamente, a curtose mede o grau de decaimento das caudas da distribuição em relação a uma medida de posição central, que pode ser a média ou a mediana.
- Pode se dizer que as caudas de uma distribuição são os seus extremos.
- Basicamente a curtose vai tenta dizer se a distribuição tem muitas observações nos extremos ou no centro.
- A curtose tenta captar o caimento da frequência, se ele cair abruptamente, ou suave:





- Neste caso, a densidade de frequência dos dados atinge o seu ponto máximo em torno da média ou mediana.
- A densidade de frequência decresce na medida que as observações se distanciam do valor central.
- Se este decaimento for lento, dizemos que a distribuição tem **Caudas Pesadas**, ou seja, a cauda persiste em ter densidade relativamente alta.
- Se o decaimento for rápido, dizemos que a distribuição possui **Caudas Leves**, ou seja, a cauda da distribuição tem densidade baixa.
- **Ex:** A distribuição de cima tem cauda leve em comparação com a distribuição de baixo. Pois a primeira distribuição começa na posição **30** e vai até a posição **70**. Enquanto que a outra distribuição começa antes do **30** e vai até depois do **70**.



- **Coeficiente de Curtose de Pearson**
  - Karl Pearson introduziu um coeficiente para identifica o decaimento nas caudas de uma distribuição de dados, esse coeficiente ficou conhecido como **Coeficiente de Curtose**.
  - Esse coeficiente é definido como: Dado um conjunto de dados $X = \{x_1,\ x_2,\ ...,\ x_n\}$, o coeficiente de curtose de X é definido como o quarto momento centrado na média divido pelo segundo momento centrado na média ao quadrado: $k_p = \frac{m_4}{(m_2)^2}$. Alguns atores usar o desvio padrão elevado a quarta potência no lugar do segundo momento centrado na média ao quadrado: $k_p = \frac{m_4}{S^4_{n,x}}$.
- **Moors** em 1986 mostrou que a curtose pode ser interpretada como uma medida de dispersão entre dois pontos (**Média** $\pm$ **Desvio Padrão**). Ou seja, a curtose méde a dispersão nas caudas da distribuição.
- A justificativa para a tese de Moors foi: Considerando os dados de $\{x_1, \cdots, x_n\}$, dessa distribuição calculamos sua **média** e seu **desvio padrão**. Padronizando os dados: Subtraindo a média de cada observação e dividindo o resultado pelo desvio padrão. Considere também que o desvio padrão é maior que 0. Com essas suposições conseguimos calcular as observações padronizadas:

$$z_1 = \frac{x_1 - \bar{x}}{S_{n,x}},\ z_2 = \frac{x_2 - \bar{x}}{S_{n,x}}, \cdots, z_n = \frac{x_n - \bar{x}}{S_{n,x}}$$

Agora o coeficiente de curtose pode ser reescrito em termos da média das observações padronizadas elevado a quarta potência.

Para os dados $\{z_1, z_2, \cdots, z_n\}$, a média é igual a 0 e a variância é igual a 1. Isso funciona para todos os todos os dados que tenha desvio padrão maior que 0.

Logo o coeficiente de curto é o segundo momento das observações padronizadas ao quadrado:

$$k = \frac{1}{n}\sum_{1=n}^{n} z_i^4$$

Podemos reescrever-lo, em termos da variância e da média de z_1^2, z_2^2 até z_n^2.

$$k = S_{n,z^2}^2 + \left(\frac{1}{n}\sum_{i=1}^{n} z_i^2\right)^2$$

Observe que a média dessas observações ao quadrado é exatamente igual a variância dos dados padronizados, como a variância dos dados padronizados é igual a 1. Logo o coeficiente de curtose vai ser igual a variância das variáveis padronizadas ao quadrado mais 1:

$$k = S_{n,z^2}^2 + 1$$

Concluímos que o coeficiente de curtose sempre vai ser maior ou igual a 1, pois a variância dos dados observados não pode ser negativa.

- **Ex:** Determine o coeficiente de curtose de $x = \{2, 3, 5, 7, 8\}$.

x_i	x_i^2	x_i^3	x_i^4
2	4	8	16
3	9	27	81
5	25	125	625
7	49	343	2401
8	64	512	4096
25	151	1015	7219

$$k_p = \frac{m_4}{(m_2)^2}$$

$$m_1' = \frac{25}{5} = 5\ m_2' = \frac{151}{5} = 30,2\ m_3' = \frac{1015}{5} = 203\ m_4' = \frac{7219}{5} = 1443,8$$

$$m_2 = m_2' - (m_1')^2 = 30,2 - 5^2 = 5,2$$

$$m_4 = m_4' - 4m_1'm_3' + 6(m_1')^2m_2' - 3(m_1')^4$$

$$= 1443,8 - 4 \cdot 5 \cdot 203 + 6 \cdot 5^2 \cdot 30,2 - 3 \cdot 5^4 = 38,8$$

$$k_p = \frac{m_4}{(m_2)^2} = \frac{38,8}{(5,2)^2} \approx 1,435$$

- Distribuições com coeficiente de curtose menor do que 3, são consideradas achatadas no topo. Uma Distribuição desse tipo é denominada **Platicúrtica**.
- Distribuições com coeficiente de curtose maior do que 3, são consideradas com pico acentuado. Uma distribuição com essa característica é denominada **Leptocúrtica**.
- Já distribuições que possuem coeficiente de curtose próximo de 3 são denominadas **Mesocúrticas**.

▼ **Medidas Descritivas**

- Uma outra maneira de resumir os dados de uma variável quantitativa é a partir de medidas descritivas.
- Tais medidas estão relacionadas a aspectos específicos dos dados(tais como, dispersão, localização, posição).
- As medidas descritivas são classificadas como:

▼ **Medidas de tendência central**

- As medidas de tendência central são assim denominadas por indicarem um ponto em torno do qual se concentram os dados.
- Este ponto tende a ser o centro da distribuição dos dados.
- As principais medidas de tendência central são: **média**, **mediana** e **moda**.

▼ **Média Aritmética**

- A média aritmética é a soma de todos os valores observados dividida pelo número total de observações.
- É a medida de tendência central mais utilizada para representar uma massa de dados.
- Seja $(x_1, ..., x_n)$ um conjunto de dados. A média é dada por: $\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$.
- **Obs:** A média aritmética é mais usada quando os intervalos tem a mesma frequência.
- **Ex:** Considere o conjunto de dados {2, 4, 6}. Então: $\bar{x} = \frac{2+4+6}{3} = 4$.
- **Obs:** Caso os dados estejam apresentados segundo uma distribuição de frequência, tem-se: $\bar{X} = \frac{\sum_{i=1}^{k} x_i n_i}{n}$. Em que k é o número de observações distintas ou o número de classes.
- Para dados agrupados em classes, x_i corresponde ao ponto médio da classe i, ou seja $x_i = \frac{l_i + L_i}{2}$.

▼ **Ex:** Os dados a seguir são referentes a notas de alunos. Determine a nota média.

Notas	Nº de alunos
4	1
5	5
6	6
7	5
8	3
Total	20

$$\bar{X} = \frac{\sum_{i=1}^{5} x_i n_i}{n} = \frac{4+5 \cdot 5+6 \cdot 6+7 \cdot 5+8 \cdot 3}{20} = \frac{124}{20} = 6.2$$

▼ **Ex:** Os dados a seguir são referentes ao consumo de energia elétrica de 80 usuários. Determine o consumo médio desses usuários.

Consumo (Kwh)	Número de usuários
[5, 25)	4
[25, 45)	6
[45, 65)	14
[65, 85)	26
[85, 105)	14
[105, 125)	8
[125, 145)	6
[145, 165)	2
Total	80

Precisamos determinar o ponto médio das classes:

Consumo (Kwh)	Número de usuários	x_i
[5, 25)	4	15
[25, 45)	6	35
[45, 65)	14	55
[65, 85)	26	75
[85, 105)	14	95
[105, 125)	8	115
[125, 145)	6	135
[145, 165)	2	155
Total	80	–

$$\bar{X} = \frac{\sum_{i=1}^{8} x_i n_i}{n} = \frac{4 \cdot 15 + 6 \cdot 35 + 14 \cdot 55 + 26 \cdot 75 + 14 \cdot 95 + 8 \cdot 115 + 6 \cdot 135 + 2 \cdot 155}{80} = \frac{6360}{80} = 79.5$$

- **Somando-se** (ou subtraindo-se) um valor constante e arbitrário a cada observação de conjunto de dados, a média aritmética fica somada (ou subtraída) por essa constante.
- **Multiplicando-se** (ou **dividindo-se**) cada observação de um conjunto de dados por um valor constante e arbitrário, a média fica multiplicada (ou dividida) por essa constante.
- A média aritmética é sensível a valores extremos. Por exemplo, compare as médias dos seguintes conjuntos: $X = 4, 7, 5, 4, 6, 2 \approx 4, 7, Y = 4, 7, 5, 4, 6, 100 \approx 21$.

▼ **Média Aritmética Ponderada**

- A média aritmética é considerada ponderada quando os valores do conjunto de dados tiverem "pesos" diferentes. De maneira genérica teríamos: $\bar{X} = \frac{\sum_{i=1}^{n} P_i x_i}{\sum_{i=1}^{n} P_i}$, em que p_i representa o peso atribuído à i-ésima observação.
- Obs: A média aritmética ponderada é mais usada quando os intervalos tem frequências diferentes.
- **Ex:** Suponha que um professor realiza quatro provas atribuindo a cada uma delas os seguintes pesos: 1,2,3,4. Se um aluno tiver recebido as notas 8, 7, 9 e 9, nessa ordem, qual será sua média final? Então: $\bar{X} = \frac{\sum_{i=1}^{4} P_i x_i}{\sum_{i=1}^{n} P_i} = \frac{(1 \cdot 8) + (2 \cdot 7) + (3 \cdot 9) + (4 \cdot 9)}{1+2+3+4} = \frac{8+14+27+36}{10} = 8.5$.

▼ **Moda**

- Representado por M_o. Ela é o valor mais frequente.
- Para variáveis discretas, a moda pode ser determinada imediatamente observando-se o rol ou a frequência absoluta dos dados.
- **Ex:** Qual a moda dos conjuntos $X = (4, 5, 5, 6, 6, 7, 7, 8, 8)$, $Z = (1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6)$, $W = 1, 2, 3, 4, 5$, $M_o(X) = 6$, $M_o(Z) = (2, 5)$, $M_o(W) = \varnothing$.
- No caso de dados agrupados em classes, o procedimento não é imediato, sendo disponível alguns métodos de cálculo distintos.
-

▼ **Mediana**

- A mediana é o valor central da amostra de dados.
- Quando for determinar a mediana de um conjunto de dados precisamos primeiramente determinar o Rol da distribuição para só depois encontra a mediana.
  - Quando o conjunto de dados tiver uma quantidade par de valores, deve pegar os dois valores centrais e realizar a média deles, é claro que antes disso deve ser feito o rol da distribuição.

▼ **Medidas de dispersão**

- Dispersão mostra o quão esticada ou espremida uma distribuição é.
- Exemplos comuns de medidas de dispersão estatística são a **Variância**, o **Desvio Padrão** e a **Amplitude Interquartil**. Dispersão é contrastada (oposto) com **Medidas de posição** ou **Tendência central**, e juntas elas são as propriedades de distribuições mais usadas.
- Ou seja, ela indica o grau de variação no conjunto de dados.
- Quando há muita variabilidade, as medidas de tendência central tornam-se pouco representativas.

▼ **Amplitude Total**

- É a diferença entre o maior e o menor valor da sequência de dados.
- Representado por $A_T = \text{valor máximo} - \text{valor mínimo}$.

▼ **Amplitude Interquartil / Desvio Interquartil**

- O desvio interquartil é a diferença entre o terceiro e o primeiro quartil.

$$d_q = Q_3 - Q_1$$

- Esta medida é mais estável que a amplitude total por não considerar os valores mais extremos.
- Útil para detectar valores discrepantes (outlires).

▼ **Variância**

- A dois tipos de variância a **Populacional** e a **Amostral**.
- **Populacional:** Usada quando se deseja analisar doto o conjunto de dados:

$$s^2 = \frac{\sum_{i=0}^{n}(x^i - \bar{x})^2}{n}$$

$$\bar{x} = Média \quad n = Número\ de\ elementos$$
$$x^i = Valores\ do\ conjunto\ de\ dados$$

- **Amostral:** Usada quando se deseja analisar só uma amostra (parte) do conjunto de dados:

$$s^2 = \frac{\sum_{i=0}^{n}(x^i - \bar{x})^2}{n-1}$$
$$Simplificando\ essa\ formula:$$
$$S^2 = \frac{1}{n-1}[\sum_{i=1}^{k} n_i \cdot x_i^2 - n\bar{x}^2]$$

- **Ex**: Calcule a variância dos dados `8 6 6 12`.=

$$s^2 = \frac{\sum_{i=0}^{n}(x^i - \bar{x})^2}{n} \rightarrow \frac{(8-8)^2 + (6-8)^2 + (6-8)^2 + (12-8)^2}{4} \rightarrow \frac{0+4+4+16}{4} \rightarrow \frac{24}{4} \rightarrow$$

- Caso os dados estejam apresentados segundo uma distribuição de frequência, tem-se:

$$s^2 = \frac{\sum_{i=0}^{k}(x^i - \bar{x})^2 \cdot n^i}{n-1}$$

em que k representa o número de observações distintas (ou número de classes).

▼ **Desvio Padrão**

- É a raiz quadrada da variância $S = \sqrt{s^2}$. Isso acontece pois a variância vai sempre retornar um resultado elevado ao quadrado.
- O desvio padrão vem para solucionar esse problema, ao realizar a raiz quadrada da variância, obteremos o resultado desejado.
- **Obs**: Alguns altores denotam o desvio padrão com o símbolo S ou DP.
- **Obs**: O desvio padrão só pode ser usado para comparar a dispersão entre dois conjuntos de dados, se os conjuntos estiverem com a mesma escala. Para resolver esse problema devemos utilizar o **Coeficiente de Variação**.

▼ **Coeficiente de Variação**

- O coeficiente de variação é uma medida de dispersão relativa definida como a razão entre o desvio-padrão e a média:

$$CV = \frac{S}{\bar{X}} 100.$$

- É utilizado, principalmente, para comparar conjuntos com unidades de medidas distintas.
- Uma desvantagem do coeficiente de variação é que ele deixa de ser útil quando a média está próxima de zero.
- Uma média muito próxima de zero pode inflacionar o CV .
- **Obs**: Quanto maior o coeficiente de variação maior vai ser a dispersão.

Exemplo:
$A = \{10, 20, 30\}$, $\overline{x}_A = 20$, $s_A = 10$.
$B = \{10000, 10010, 10020\}$, $\overline{x}_B = 10010$, $s_B = 10$.
$CV_A = \frac{s_A}{\overline{x}_A} = 0,5$ e $CV_B = \frac{s_B}{\overline{x}_B} \approx 0,0009$.

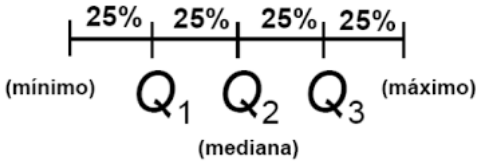
Exemplo:
Prova 1: 0 a 100. Média da turma: 70. Desvio padrão 1.
Prova 2: 0 a 10. Média da turma: 7. Desvio padrão 1.
$CV_1 = \frac{s_1}{\overline{x}_1} = 0,014$ e $CV_2 = \frac{s_2}{\overline{x}_2} \approx 0,14$.

▼ **Propriedades:**

- Somando-se ou subtraindo-se um valor constante e arbitrário `c`, a cada elemento de um conjunto de dados, a **variância** (e consequentemente o **desvio-padrão**) não se altera.
- Multiplicando ou dividindo por um valor constante e arbitrário `c`, a cada elemento de um conjunto de dados, o **desvio-padrão** fica multiplicado ou dividido por `c` e a **variância** fica multiplicada ou dividida por `c²`.

▼ **Medidas de posição (Quantis)**

- Os quantis são a generalização da ideia de mediana. Estes fornecem informações a respeito de posição e forma da distribuição dos dados.
  Alguns quantis importantes são:
  - **Quartis:** Dividem a distribuição de dados em quatro partes iguais. Ou seja, para dividir uma conjunto de dados em quatro partes iguais, precisamos de três quartis.



De maneira geral, o quantil de ordem `p` (denotaremos por Q_p) é um valor que dividi o conjunto de dados ordenados, de modo que, p% das observações estão situadas até Q_p e (1 − p)% situadas acima de Q_p para todo `p` ∈ (0, 1).

Para determinar a posição de um quantil dentro de uma determinada amostra, primeiramente ordenamos a amostra (Rol) e depois usamos a formula $P_k = p(n+1)$, p é a ordem do quantil que deve ser informada no enunciado, n é o número de elementos que tem a amostra.

Obs: A formula mostra acima só deve ser usada para dados que não estejam agrupados por classe, quando for esse caso deve-se usar a formula:

$$Q_p = l_i + \frac{h(p - F_{i-1})}{f_i}$$

Ex: Determine os quantil de ordem p para todo $p \in \{0,50; 0,25; 0,20; 0,95\}$. A amostra de dados é:

13	11	15	24	20	20	19
18	22	22	20	17	25	

R:
$$Pk = 0.50(13 + 1) = 7$$
$$Pk = 0.25(13 + 1) = 3,5$$
$$Pk = 0.20(13 + 1) = 2,8$$
$$Pk = 0.95(13 + 1) = 13,3$$

Obs: O resultado de cada linha é referente a posição do quantil na amostra.

Obs: Se o resultado for $posicao \leq 1$ o quantil vai ser igual ao menor valor da amostra, se for $posicao \geq n$ o quantil vai ser igual ao maior valor da amostra.

Obs: Se o resultado não for um valor inteiro, precisamos encontra dois valores (P_1 e P_2) mais próximo do resultado, ou seja ($P_1 < P_n$ e $P_2 > P_n$), e usar a formula:

$$\text{QUANTIL p} = (\overline{p}_2 - p_n) \times x_{(\overline{p}_1)} + (p_n - \overline{p}_1) \times x_{(\overline{p}_2)}$$

Ex: No exemplo acima, a posição do segundo quartil deu um valor não inteiro. Vamos descobrir qual é o valor desse quantil na posição `3.5`.

Rol:

11	13	15	17	18	19	20	20	20	22	22	24	25

Percebesse que a posição `3.5` fica entre os valores `15` e `17`, logo o $P_1 = 15$ e $P_2 = 17$, substituindo os valores na formula, ira fica:

$$p = (4 - 3.5) \cdot 15 + (3.5 - 3) \cdot 17$$
$$p = 0.5 \cdot 15 + 0.5 \cdot 17$$
$$p = 7.5 + 8.5 = 16$$

Os valores `4`, `3.5`, `3` são referente a posição do P_2, P_n e do P_1 respectivamente.

- **Decis:** Dividem a distribuição de dados em dez partes iguais.
- **Percentis:** Dividem a distribuição de dados em cem partes iguais.

▼ **Medidas de forma**

- As medidas de forma nos auxiliam a entender onde os dados estão concentrados.
- A informação fornecida pelas medidas de forma, em conjunto com as medidas de localização e dispersão, pode nos ajudar a escolher, de maneira apropriada, um modelo estatístico que descreva o comportamento probabilístico dos dados.
- A respeito da forma, a distribuição dos dados pode diferir em relação ao grau de deformação (assimetria) e o grau de achatamento (curtose).

- Para definir medidas de assimetria e curtose, precisamos definir o conceito de **Momento**.

▼ **Momento Natural**

Definição: Seja $X = \{x1, x2, \cdots, xn\}$ um conjunto de dados observados. O momento de ordem r de X é definido quando for:

Dados Brutos: $m´_r = \frac{\sum_{i=1}^{n} x_i^r}{n}$

Dados em distribuição de frequências: $m´_r = \frac{\sum_{i=1}^{k} n_i\, x_i^r}{n}$

Ex: Determine os momentos de primeira, segunda, terceira e quarta ordens de $X = \{2, 3, 5, 7, 8\}$.

x_i	x_i^2	x_i^3	x_i^4
2	4	8	16
3	9	27	81
5	25	125	625
7	49	343	2401
8	64	512	4096
25	151	1015	7219

$$
\begin{aligned}
m'_1 &= \frac{25}{5} = 5 = \bar{x} \\
m'_2 &= \frac{151}{5} = 30,2 \\
m'_3 &= \frac{1015}{5} = 203 \\
m'_4 &= \frac{7219}{5} = 1443,8
\end{aligned}
$$

▼ **Momento Centrado na Média**

Seja $X = \{x1, x2, \cdots, xn\}$ um conjunto de dados observados. O r-ésimo momento centrado na média aritmética é definido por:

Dados Brutos: $m_r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^r}{n}$

Dados em distribuição de frequências: $m_r = \frac{\sum_{i=1}^{k} n_i (x_i - \bar{x})^r}{n}$

Ex: Determine os momentos de primeira, segunda, terceira e quarta ordens de $X = \{2, 3, 5, 7, 8\}$.

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^3$	$(x_i - \bar{x})^4$
2	-3	9	-27	81
3	-2	4	-8	16
5	0	0	0	0
7	2	4	8	16
8	3	9	27	81
	0	26	0	194

$$
\begin{aligned}
m_1 &= 0 \\
m_2 &= \frac{26}{5} = 5,2 \\
m_3 &= 0 \\
m_4 &= \frac{194}{5} = 38,8
\end{aligned}
$$

Obs: Na maioria dos casos só precisamos saber até o quarto momento. O primeiro momento m_1 está relacionado com o a **Média**; O segundo momento m_2 está relacionado com o a **Variância**; O terceiro

momento m_3 está relacionado com o a **Assimetria**; O quarto momento m_4 está relacionado com o a **Curtose**.

▼ **Relação entre Momentos Naturais e Momentos Centrais**

- $m_1 = 0$
- $m_2 = m'_2 - (m'_1)^2$
- $m_3 = m'_3 - 3m'_1m'_2 + 2(m'_1)^3$
- $m_4 = m'_4 - 4m'_1m'_3 + 6(m'_1)^2m'_2 - 3(m'_1)^4$
- $m_r = \sum_{k=0}^{r} \binom{r}{k}(-1)^k(m'_1)^k m'_{r-k}$

▼ **Coeficiente de Assimetria de Pearson**

- Definimos o coeficiente de assimetria de Pearson por $CA_p = \frac{m_3}{m_2^{3/2}}$. O m_3 e o m_2 são momentos centrado na média.
- CA_p é uma medida adimensional.
- Valores de CA_p próximos a zero indicam **Simetria**. Valores negativos indicam **Assimetria a Esquerda**, enquanto valores positivos indicam **Assimetria a Direita**.
- **Ex**: Os dados apresentados se referem as notas de 20 alunos submetidos a provas de Estatística Elementar, em três situações diferentes. As provas possuem 4 questões. A cada estudante foi atribuída um nota que pode variar de zero a dez. Determine o coeficiente de assimetria de Pearson desses dados.

Tabela 1: Notas de 20 alunos

Notas			
0,5	2,6	5,1	7,5
1,2	2,7	4,5	7,2
0,8	4,0	5,4	8,2
3,8	5,6	6,3	8,3
2,9	4,5	6,7	9,3

Notas	n_i	x_i	$(n_i \cdot x_i)$	$(n_i \cdot x_i^2)$	$(n_i \cdot x_i^3)$
$[0, 2)$	3	1	3	3	3
$[2, 4)$	4	3	12	36	108
$[4, 6)$	6	5	30	150	750
$[6, 8)$	4	7	28	196	1372
$[8, 10)$	3	9	27	243	2187
Total	20	-	100	628	4420

O x_i é o ponto médio da classe.

$$
\begin{aligned}
m'_1 &= \frac{\sum_{i=1}^{5} n_i \cdot x_i}{20} = \frac{100}{20} = 5 \\
m'_2 &= \frac{\sum_{i=1}^{5} n_i \cdot x_i^2}{20} = \frac{628}{20} = 31.4 \\
m'_3 &= \frac{\sum_{i=1}^{5} n_i \cdot x_i^3}{20} = \frac{4420}{20} = 221 \\
\\
m_3 &= m'_3 - 3m'_1 \cdot m'_2 + 2(m'_1)^3 \\
m_3 &= 221 - 3 \cdot 5 \cdot 31.4 + 2 \cdot 5^3 \\
m_3 &= 221 - 471 + 250 \\
m_3 &= 0 \\
\\
CA_p &= \frac{m_3}{(m_2)^{\frac{3}{2}}} = \mathbf{0}
\end{aligned}
$$

-

▼ **Coeficiente de Assimetria Bowley**

- O coeficiente de assimetria de Bowley se baseia nas distâncias interquartis, e é definido por $g_b = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1}$.
- O coeficiente de assimetria de Bowley é também adimensional.
- Valores próximos a zero indicam **Simetria**, valores negativos indicam **Assimetria a Esquerda** e valores positivos indicam **Assimetria a Direita**.
- Ao contrário do coeficiente de assimetria de Pearson, o coeficiente de Bowley tende a ser menos sensível a presença de observações discrepantes (outliers).

▼ **Análise Bivariada**

- Frequentemente estamos interessados em analisar o comportamento conjunto de duas ou mais variáveis, no intuito de encontrar possíveis relações entre estas.
- A distribuição de frequências conjunta das variáveis nos fornecem informações sobre uma possíveis associações.
- Considerando duas variáveis podemos ter as seguintes situações:
  - As duas variáveis são **quantitativas**;
  - As duas variáveis são **qualitativas**;
  - Uma variável é **qualitativa** e outra é **quantitativa**.

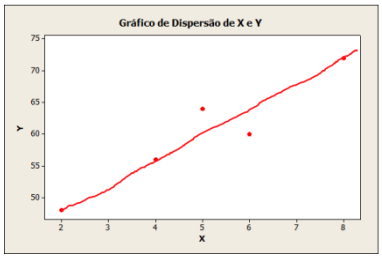
▼ **Análise de Correlação**

- Essa análise se aplica, somente para variáveis **Quantitativas**. Ou seja, quando é Quantitativa X Quantitativa.

- Exemplo 1:

Agente	Anos de Serviço (X)	N^o de Clientes (Y)
A	2	48
B	4	56
C	5	64
D	6	60
E	8	72
Total	25	300

Nessa tabela acima, queremos saber qual é a correlação entre a coluna X (anos de serviço) e o Y (nº de clientes). Para isso vamos fazer um grafico de disperção desses dados:

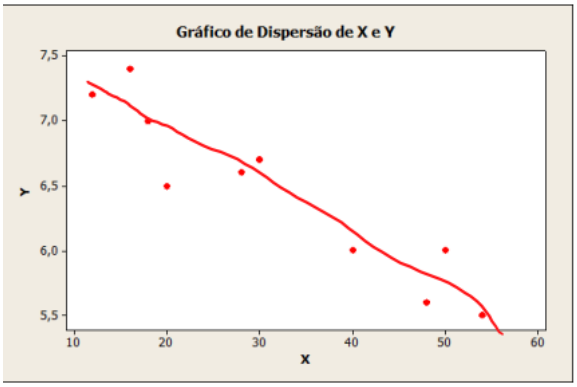


Percebemos que os dados tendem a seguir uma tendência linear.

O gráfico indica uma possível dependência **Linear Positiva** entre as variáveis anos de serviço e número de clientes.

- Exemplo 2:
  - Renda Mensal Bruta (X)
  - % da Renda gasta com Assistência Médica (Y)

Família	X	Y
A	12	7,2
B	16	7,4
C	18	7,0
D	20	6,5
E	28	6,6
F	30	6,7
G	40	6,0
H	48	5,6
I	50	6,0
J	54	5,5



Nesse caso, a dependência entre X e Y parece ser **Linear Negativa**.

- Quando uma variável aumenta a outra também aumenta. É chamado de relação **Linear Positiva**.
- Quando uma variável aumenta a outra diminui. É chamado de relação **Linear Negativa**.

- **Obs**: Quando não houver relação entre duas variáveis, os dados vão se comportar aleatoriamente em um gráfico de dispersão.
- **Coeficiente de Correlação linear de Pearson:** Obter uma medida que permita quantificar a dependência linear que pode existir entre duas variáveis (positiva, negativa, muita ou pouca)

  Dado n pares de observações $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$. Definimos o coeficiente de correlação linear de pearson como:

$$r_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{S_X} \right) \left(\frac{y_i - \bar{y}}{S_X} \right)$$

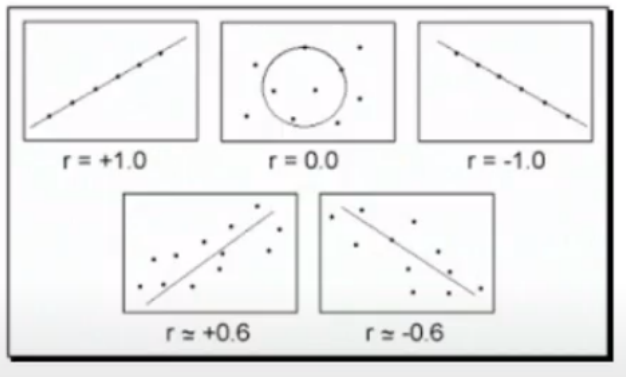
  Basicamente nessa formula temos a soma dos produtos das observações padronizadas de x e de y.

  A formula acima também pode ser expresso como:

$$r_{XY} = \frac{\sum_{i=1}^{n} x_i \, y_i - n \, \bar{x} \, \bar{y}}{\sqrt{(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2) \cdot (\sum_{i=1}^{n} y_i^2 - n\bar{y}^2)}}$$

  Propriedades:

  - $-1 \leq r_{XY} \leq 1$
  - r_{XY} estiver próxima de 1: X e Y estão **Positivamente** correlacionados e o tipo de correlação entre as variáveis é linear.
  - r_{XY} estiver próxima de -1: X e Y estão **Negativamente** correlacionados e o tipo de correlação entre as variáveis é linear.
  - r_{XY} próximo de zero, indica que X e Y não estão correlacionados.



  Exemplo: Retomando o Exemplo 1, quanto vale a correlação entre Anos de Serviço (X) e Nº de Clientes (Y)?

x	y	x^2	y^2	xy
2	48	4	2304	96
4	56	16	3196	224
5	64	25	4096	320
6	60	36	3600	360
8	72	64	5184	576
25	300	145	18320	1576

$$\bar{x} = \frac{25}{5} = 5 \quad \bar{y} = \frac{300}{5} = 60$$

$$r_{XY} = \frac{\sum_{i=1}^{n} x_i \, y_i - n \, \bar{x} \, \bar{y}}{\sqrt{(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2) \cdot (\sum_{i=1}^{n} y_i^2 - n\bar{y}^2)}} = \frac{1576 - 5 \cdot 5 \cdot 60}{\sqrt{(145 - 5 \cdot 5^2)(18320 - 5 \cdot 60^2)}} = \frac{76}{\sqrt{20 \cdot 320}} = 0,$$

▼ Análise de Associação

- Essa análise se aplica, somente para variáveis **Qualitativas**. Ou seja, quando é Qualitativa X Qualitativa.
- No caso de frequências relativas, existem três possibilidades de expressarmos a proporção de cada casela:

  Em relação ao total geral;
  Em relação ao total de cada linha;
  Em relação ao total de cada coluna;

- Como avaliar se as proporções estão realmente próximas, fornecendo indícios de que não há associação?
- De maneira geral suponha que temos duas variáveis qualitativas X e Y , classificadas em r categorias $\{A1,\ A2, \cdots,\ Ar\}$ para X e s categorias $\{B1,\ B2, \cdots,\ Bs\}$, para Y.

  O X é referente as linhas e o Y é referente as colunas.

X \ Y	B_1	B_2	...	B_j	...	B_s	Total
A_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1s}	$n_{1.}$
A_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2s}	$n_{2.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
A_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{is}	$n_{i.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
A_r	n_{r1}	n_{r2}	...	n_{rj}	...	n_{rs}	$n_{r.}$
Total	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.s}$	$n..$

Figura 1: Notação para tabelas de contingência. (Bussab & Morettin, 2010)

- A imagem acima tem vários n seguindo de dois índices n_{rs}, o primeiro índice r é referente a linha, ou seja qual linha o elemento n_{rs} está localizado. O segundo índice s é referente a coluna, ou seja qual coluna o elemento n_{rs} está localizado.
- n_{ij} = número de observações pertencentes à i-ésima categoria de X e j-ésima categoria de Y.
- $n_{i.} = \sum_{j=1}^{s} n_{ij}$ = número de observações da i-ésima categoria de X.
- $n_{.j} = \sum_{i=1}^{r} n_{ij}$ = número de observações da j-ésima categoria de Y.
- $n_{..} = n = \sum_{i=1}^{r} \sum_{j=1}^{s} n_{ij}$ = número total de observações.
- A medida que avaliar a Associação é a χ^2, o nome dela é chi, se pronuncia qui-quadro. Ela informa se a associação é fraca, moderada ou forte. Ela é baseada na diferencia entre as frequências esperadas e_{ij} e as frequências observadas n_{ij}.
- A frequência esperada e_{ij} é a frequência para caso não haja associação.
- Se essas frequências forem próximas ($e_{ij} \approx n_{ij}$), a associação é fraca.
- Se essas frequências se distanciam, se essa distância for grande a associação é forte, a magnitude da associação é dada por essa distância. Para avaliar essa distância usamos o χ^2.
- A **formula** de χ^2 é:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

- Frequência esperada é definida por: $e_{ij} = \frac{n_{i.} n_{.j}}{n}$ é a frequência esperada da categoria i da variável X e categoria j da variável Y. Ou seja é o total nas linhas multiplicado pelo total nas colunas dividido pelo número total de observações.
- A χ^2 não pode ser menor que 0.

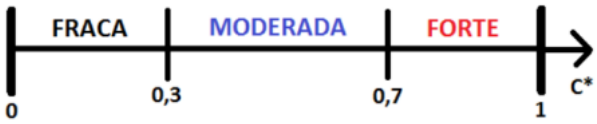
- Note que $\chi^2 = 0$ somente quando todas as frequências observadas são iguais as esperadas. Portanto valores de χ^2 próximos de zero indicam não associação.
- Por outro lado χ^2 pode variar de 0 (zero) a ∞ (infinito) tornando difícil avaliar, a partir dessa medida, a magnitude da associação.
- Outras medidas foram propostas com o intuito de fornecer informações mais precisas a respeito do grau de associação. Por exemplo o **Coeficiente de Contingência de Pearson**:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

- E possível mostrar que o maior valor assumido por C é dado por $\sqrt{(t-1)/t}$, em que $t = mín(s, r)$, ou seja, o mínimo entre o número de linhas e colunas da tabela de contingência.
- Coeficiente de **Contingência Modificado** é definido por:

$$C^* = \frac{C}{\sqrt{(t-1)/t}}$$

- Neste caso $0 \leq C^* \leq 1$ de modo que:



- **Ex:** Avalie se existe associação entre grau de instrução e região de procedência por meio do Coeficiente de Contingência modificado.

- Grau de Instrução: X
- Região de Procedência: Y

	Ensino Fundamental	Ensino Médio	Ensino Superior	Total
Capital	4	5	2	11
Interior	3	7	2	12
Outra	5	6	2	13
Total	12	18	6	36

Para calcular a frequência esperada usamos a formula $e_{ij} = \frac{n_j n_i}{n}$. A primeira frequência vai ser $e_{1\,1} = \frac{12 \cdot 11}{36} = 3,67$, o segundo valor da primeira coluna $e_{2\,1} = \frac{12 \cdot 12}{36} = 4,00$, e assim por diante com todos os valores:

Y \ X	Fundamental	Medio	Superior	Total
Capital	4 (3,67)	5 (5,50)	2 (1,83)	11
Interior	3 (4,00)	7 (6,00)	2 (2,00)	12
Outra	5 (4,33)	6 (6,50)	2 (2,17)	13
Total	12	18	6	36

As frequências esperadas são os valores que estão entre parênteses.

Logo o χ^2 dessa amostra de dados vai ser:

$$\chi^2 = \sum_{i=1}^{3}\sum_{j=1}^{3} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} =$$

$$\frac{(4-3,67)^2}{3,67} + \frac{(3-4,00)^2}{4,00} + \frac{(5-4,33)^2}{4,33} + \frac{(5-5,50)^2}{5,50} + \frac{(7-6,00)^2}{6,00} + \frac{(6-6,50)^2}{6,50} + \frac{(2-1,83}{1,83}$$

O Coeficiente de Contingência de Pearson dessa distribuição é:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{0,67}{0,67 + 36}} = 0,135$$

Coeficiente de Contingência Modificado é:

$$C^* = \frac{C}{\sqrt{(t-1)/t}} = \frac{0,135}{\sqrt{(3-1)/3}} = 0,165$$

O t da formula acima é o mínimo entre o número de linhas e colunas, que nesse casso é o 3, pois o mínimo de 3 e 3 é o próprio 3. Se por exemplo fosse 2 linhas e 3 colunas, o t dessa distribuição seria 2.

Concluímos que a distribuição tem uma associação fraca, para chegar a essa conclusão basta pega o C^* e ver onde ela se encaixa na image acima que classifica se a distribuição tem a associação fraca, moderada ou forte.

▼ **Variável Quantitativa X Qualitativa**

- Nesta situação é comum analisarmos como se comporta a variável quantitativa em cada nível da variável qualitativa.
- Essa análise pode ser conduzida por meio de medidas descritivas, **histogramas**, **box plots**, etc.
- Ou seja, analisamos a variável quantitativa para cada nível da variável qualitativa.

Tabela 1: Distribuição do salário de funcionários de uma determinada empresa, segundo o grau de instrução (em R$ /h).

	Grau de instrução			
Salário	fundamental	médio	superior	Total
$4,00 \mid - 6,86$	4	1	0	5
$6,86 \mid - 9,71$	4	6	0	10
$9,71 \mid - 12,60$	2	4	1	7
$12,60 \mid - 15,40$	1	4	1	6
$15,40 \mid - 18,30$	0	2	2	4
$18,30 \mid - 21,10$	0	1	1	2
$21,10 \mid - 24,00$	0	0	1	1
Total	11	18	6	35

Tabela 2: Medidas descritivas do salário conforme o grau de instrução

Medidas	Grau de instrução		
	fundamental	médio	superior
Mín.	4,00	5,73	10,53
1^o quartil	6,01	8,84	13,65
Mediana	7,12	10,91	16,74
Média	7,84	11,53	16,48
3^o quartil	9,16	14,42	18,38
Máx.	13,85	19,40	23,30
Desvio-padrão	2,96	3,72	4,50

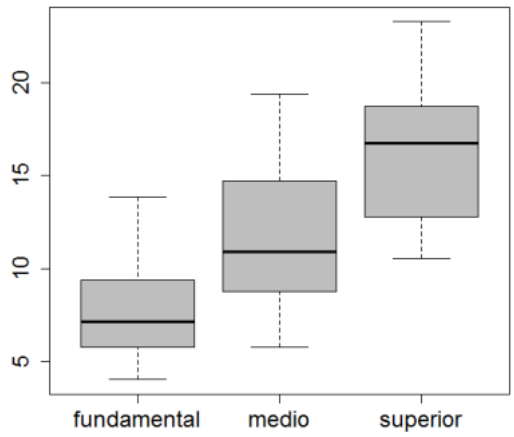


Figura 2: Salários em termos do grau de escolaridade

▼ **Exemplo 1:**

Os dados a seguir são de 20 observações relativas ao índice pluviométrico em determinados municípios do estado do Ceará.

144	152	159	160
160	151	157	146
154	145	141	150
142	146	142	141
141	150	143	158

Apresente estes dados em uma distribuição de frequências.

Obs: Para fazer essa distribuição, eu preciso primeiro descobrir o número de classes, para isso eu faço Regra da raiz quadrada: $k = \sqrt{20} = 4.47$, o valor inteiro mais próximo de 4.47 é 5, então o número de classes vai ser igual a 5.

Obs: Agora precisamos saber qual vai ser o tamanho da classe. Para isso eu defino a Amplitude Total: $A_T = 160 - 141 = 19$, depois eu defino a Amplitude de classe:$h = \frac{19}{5} = 3,8 \approx 4$. Agora é só determinar o primeiro e ultimo elemento da classe.

i	$Classes$	n_i
1	[141, 145)	7
2	[145, 149)	3
3	[149, 153)	4
4	[153, 157)	1
5	[157, 161)	5
	Total	20

▼ **Gráfico**

- Gráfico á um recurso visual da Estatística utilizado para representar um fenômeno.
- É importante ressaltar que a relação entre medida e escala deve ser rigorosamente respeitada.
- Há três tipos de gráficos, classificados quanto ao critério da forma:

  Diagramas: São gráficos geométricos dispostos em duas dimensões. São os mais usados na representação de séries estatísticas.

Cartogramas: São ilustrações relativas a cartas geográficas, largamente difundidas em Geografia, História e Demografia.

Estereogramas: Representam volumes e são apresentados em três dimensões.

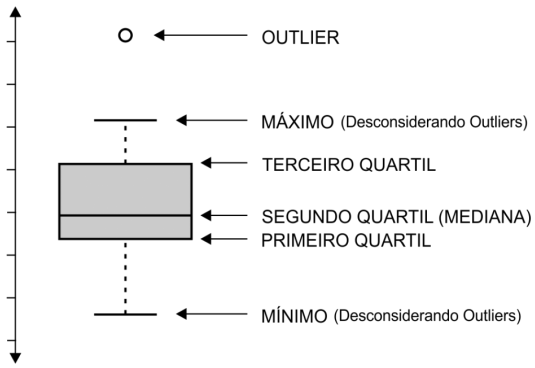
▼ Gráfico de Ramos e Folhas

- Usado principalmente para variáveis contínuas, pode auxiliar na determinação do número de classes;
- Permite visualizar a forma da distribuição dos dados;
- Característica: Não perde informação sobre os dados;
- Cada informação é dividida em duas partes: A primeira (parte inteira do número) é colocada à esquerda da linha vertical. E a segunda (parte decimal do número) à direita.
- **Ex:**

4	00	56		
5	25	73		
6	26	66	86	
7	39	44	59	
8	12	46	74	95
9	13	35	77	80
10	53	76		
11	06	59		
12	00	79		
13	23	60	85	
14	69	71		
15	99			
16	22	61		
17	26			
18	75			
19	40			
20				
21				
22				
23	30			

  - Os valores a direita, antes da reta, são os valores inteiros.
  - As colunas a esquerda depois da reta, são as partes decimais dos valores que estão antes da reta.
  - Cada coluna depois da reta, significa o valor decimal do valor que esta antes da reta, para cada vez que ele se repetir.
  - **Ex:** Na primeira linha temos $4 \mid 00\ 56$, como tem duas colunas depois da reta, significa que o 4 vai se repetir duas vezes: $4,00$ e $4,56$.

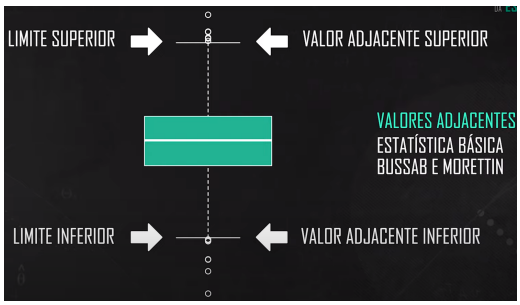
▼ Gráfico Box-Plot



- O Box-Plot é um esquema gráfico que utiliza cinco medidas descritivas: **Máximo**, **Mínimo**, **Mediana**, **Primeiro e terceiro quartil**.

- Dessa forma, temos em um só gráfico informações sobre **posição**, **dispersão**, forma da distribuição de dados e observações discrepantes.
- A **posição central** é dada pela mediana e a dispersão pelo desvio interquartílico.
- As posições relativas de Q1 , Q2 e Q3 dão uma noção da forma da distribuição.
- Os comprimentos das caudas são dados pelas linhas que vão do retângulo aos valores atípicos.
- Um outlier ou valor discrepante é um valor que se localiza distante de quase todos os outros pontos da distribuição.
- A distância a partir da qual um valor é considerado discrepante é aquela que supera $1.5 \times d_q$, o d_q é o desvio interquartil.
- De maneira geral, são considerados outliers todos os valores inferiores a $Q1-1.5 \times d_q$ ou superiores a $Q3 + 1.5 \times d_q$.
- **Limite superior:** $Q3 + 1.5 \times d_q$

  Limite inferior: $Q1-1.5 \times d_q$



- **Ex**: População, em 1000 habitantes, de cada unidade da federação.

RR	325	MS	2079	PB	3444	PR	9564
AP	478	MT	2505	GO	5004	RS	10188
AC	558	RN	2777	SC	5357	BA	13071
TO	1158	AM	2813	MA	5652	RJ	14392
RO	1380	AL	2823	PA	6193	MG	17892
SE	1785	PI	2844	CE	7431	SP	37033
DF	2052	ES	3098	PE	7919		

Mediana: $x_{14} = 3098 (ES)$

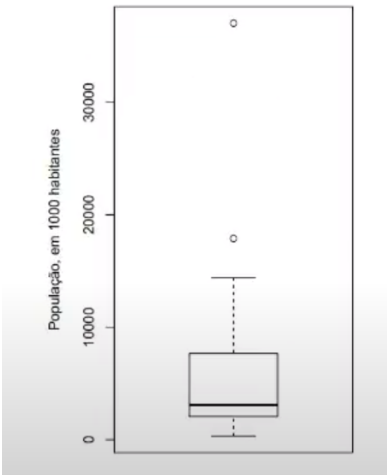
Primeiro quartil: $Q_1 = x_7 = 2052 (DF)$

Terceiro quartil: $Q_3 = x_{21} = 7919 (PE)$

Desvio Interquartil: $d_q = Q_3 - Q_1 = 7919 - 2052 = 5867$

Ponto de corte inferior: $Q_1 - 1.5 \times d_q = -6748.5$

Ponto de corte superior:$Q_3 + 1.5 \times d_q = 16720$


População, em 1000 habitantes
0
10000
20000
30000