Big Data Real-Time Analytics com Python e Spark

Seja muito bem-vindo(a)!

Análise Estatística de Dados

Análise Estatística de Dados

Parte 1

Parte 2

Definindo Estatística

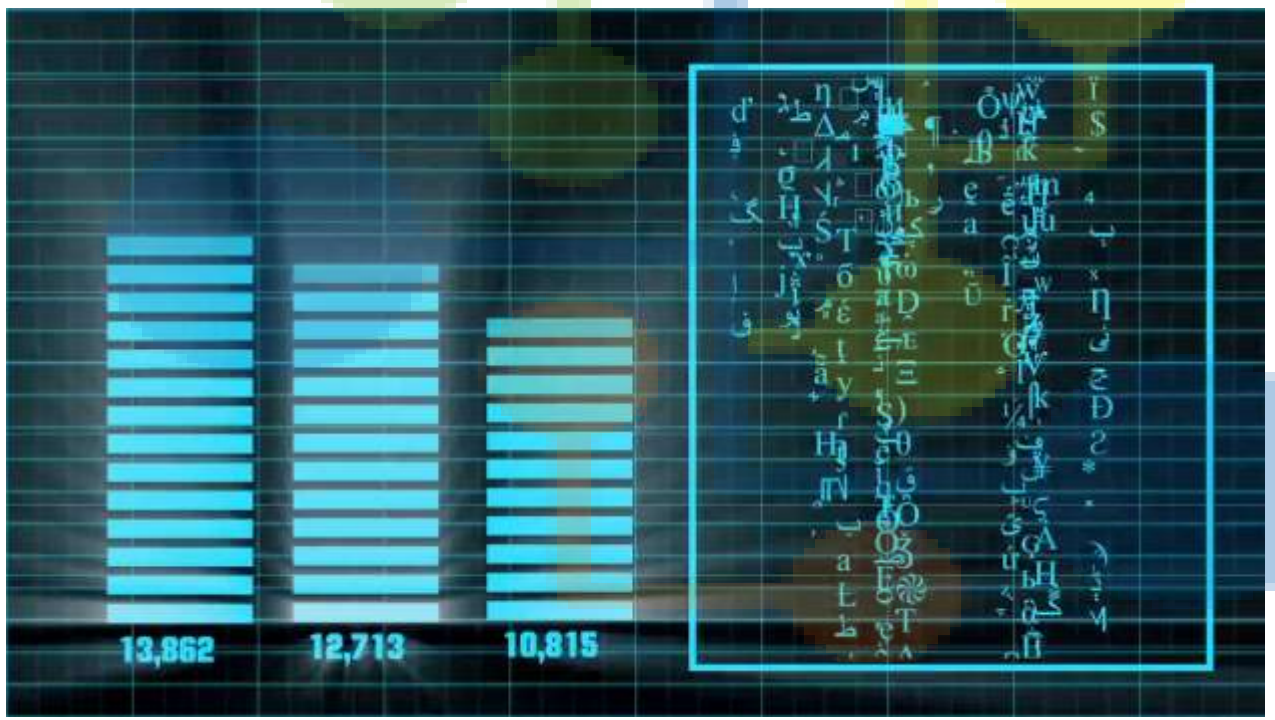
Definindo Estatística

Os números constituem a única verdade universal.

Nathanael West

Definindo Estatística

Análise Estatística de Dados



Definindo Estatística

O que é Estatística?

Definindo Estatística

É a ciência que nos permite aprender a partir dos dados.

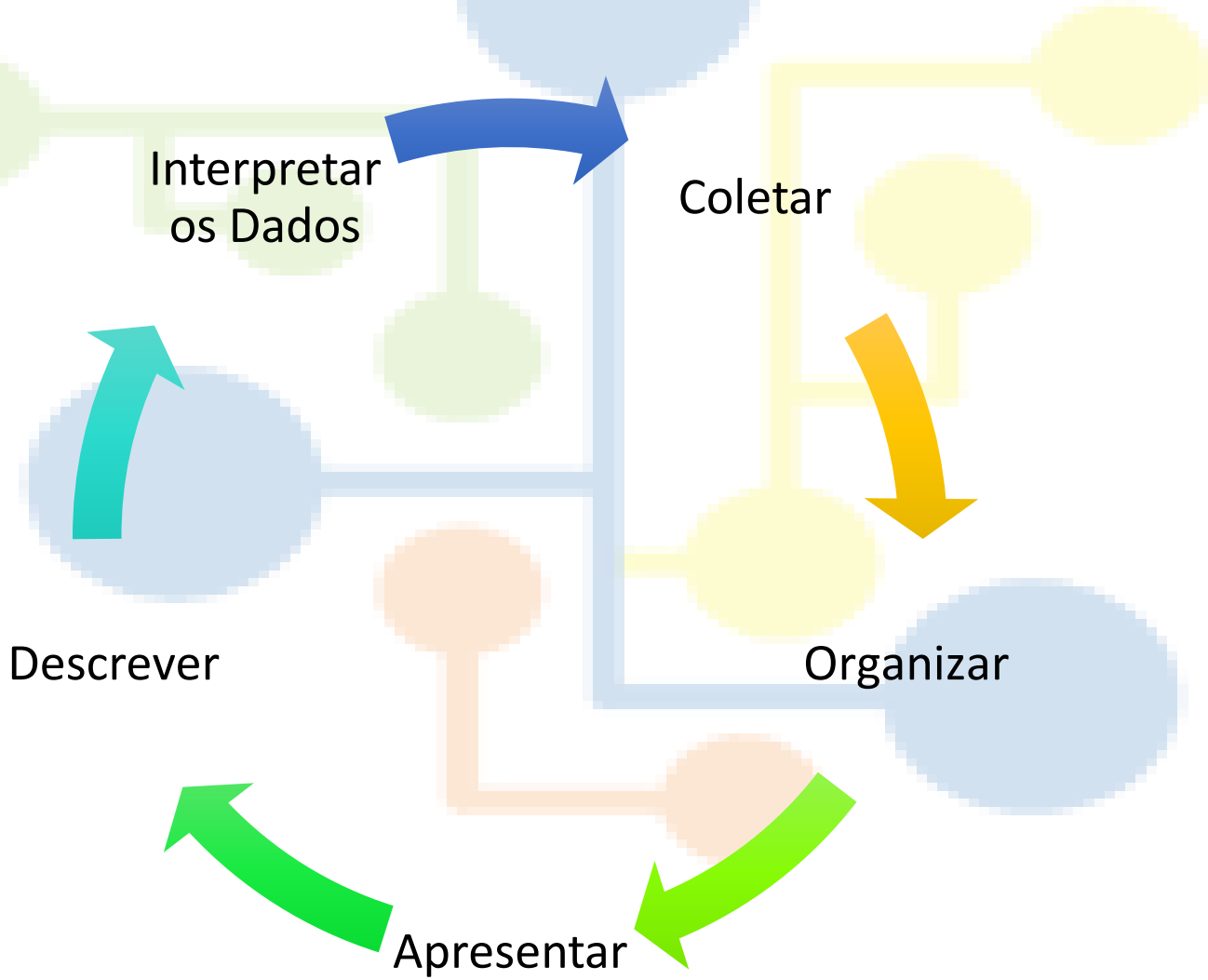
Definindo Estatística

Com a Estatística nós podemos:

Definindo Estatística

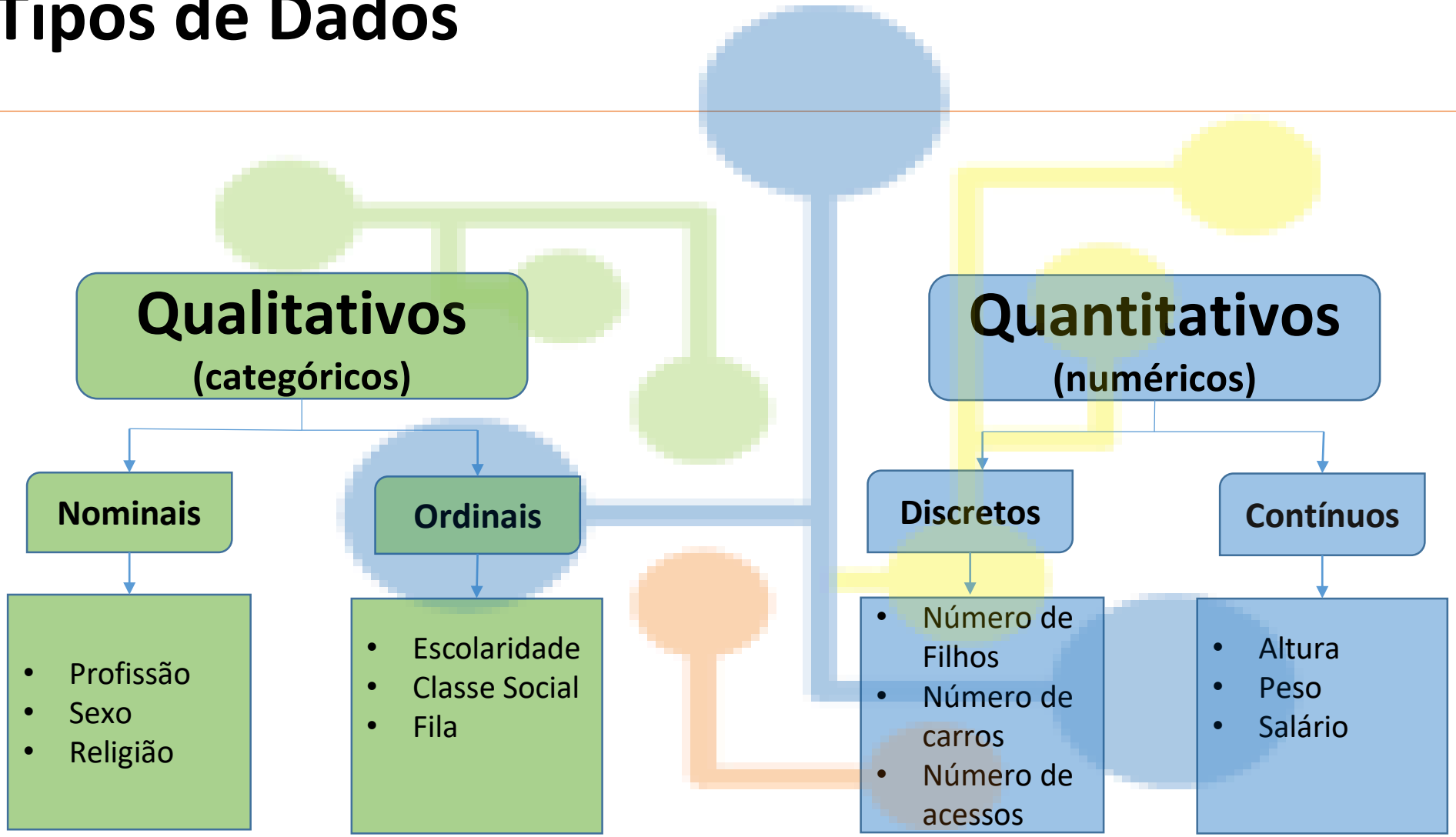
Coletar

Definindo Estatística



Tipos de Dados

Tipos de Dados

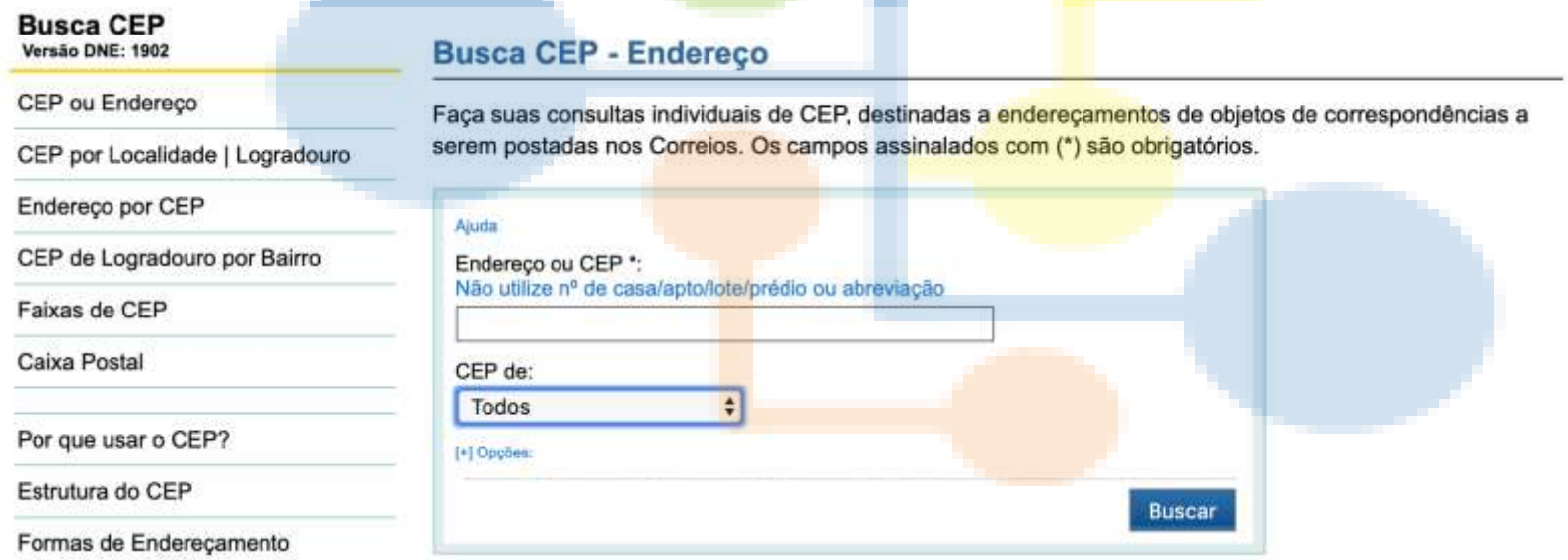


Tipos de Dados

Exemplo

Tipos de Dados

Dados Qualitativos Nominais – representam descrições para os dados e não permitem ranqueamento. Exemplo: CEP (70.098-080).



Tipos de Dados

Dados Qualitativos Ordinais - existe uma ordenação entre as categorias (ranqueamento) e os dados podem ser medidos.

Tipos de Dados

Dados Quantitativos Discretos – valores baseados em observações que podem ser contados, normalmente representados por valores inteiros.

Tipos de Dados

Dados Quantitativos Contínuos – valores baseados em observações que podem ser medidas e normalmente representados por valores decimais.

Observação x Experimentação

Observação x Experimentação

Há dois tipos de estudos estatísticos:

- Observacional
- Experimental

Observação x Experimentação

Em um estudo de **observação**, os dados e as características específicas são recolhidos e observados, entretanto, não há iniciativa de modificar os estudos que estão sendo realizados.

Observação x Experimentação

Em um estudo **experimental**, cada indivíduo é aleatoriamente atribuído a um grupo de tratamento, em seguida, os dados e as características específicas são observados e coletados.

Observação x Experimentação

A Análise de Dados é o meio através do qual utilizamos a estatística para apresentar e demonstrar os resultados dos dados que foram avaliados.

Observação x Experimentação

Estatística não tem sido usada apenas por técnicos, mas também por gestores de todos os níveis.

Para onde se olha, se vê **Estatística** sendo aplicada, desde o planejamento corporativo, até decisões simples do dia a dia.

Principais Áreas da Estatística

Principais Áreas da Estatística

Estatística Descritiva

Estatística Descritiva

É um conjunto de métodos estatísticos utilizados para descrever as principais características dos dados.

Estatística Descritiva

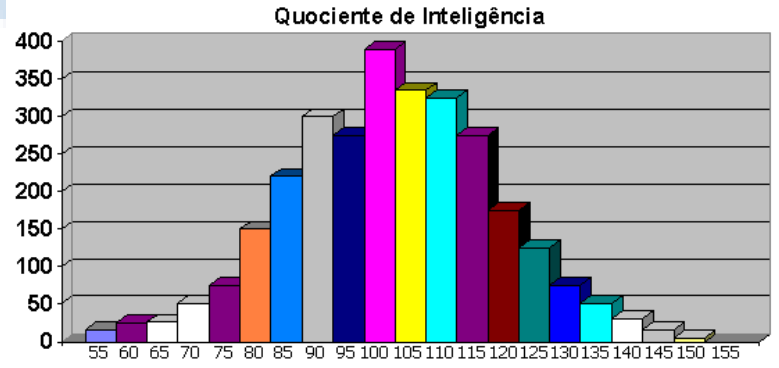
O principal propósito de métodos gráficos é organizar e apresentar os dados de forma gerencial e ágil.

Estatística Descritiva

A **Estatística Descritiva** tem por objetivo sumarizar e mostrar os dados, de forma que se possa rapidamente obter uma visão geral da informação que está sendo analisada.

Estatística Descritiva

Por meio da Estatística Descritiva entendemos melhor um conjunto de dados através de suas características.



Estatística Descritiva

As três principais características são:

- Um valor representativo do conjunto de dados. Ex.: a média.
- Uma medida de dispersão ou variação. Ex: variância, desvio padrão.
- A natureza ou forma da distribuição dos dados: sino, uniforme ou assimétrica.

Tabela de Frequência

Tabela de Frequência

Um dos meios mais simples de descrever dados é através de **tabelas de frequência**, que refletem as observações feitas nos **dados**.

Tabela de Frequência

Cada linha em uma **tabela de frequência** corresponde a uma **classe.**

Número de tablets vendidos por dia	Frequência
0	5
1	8
2	14
3	13
4	6

Classe

Tabela de Frequência

Cada **classe** corresponde a uma **categoria** em uma **tabela de frequência.**

Número de tablets vendidos por dia	Frequência
0	5
1	8
2	14
3	13
4	6

Classe

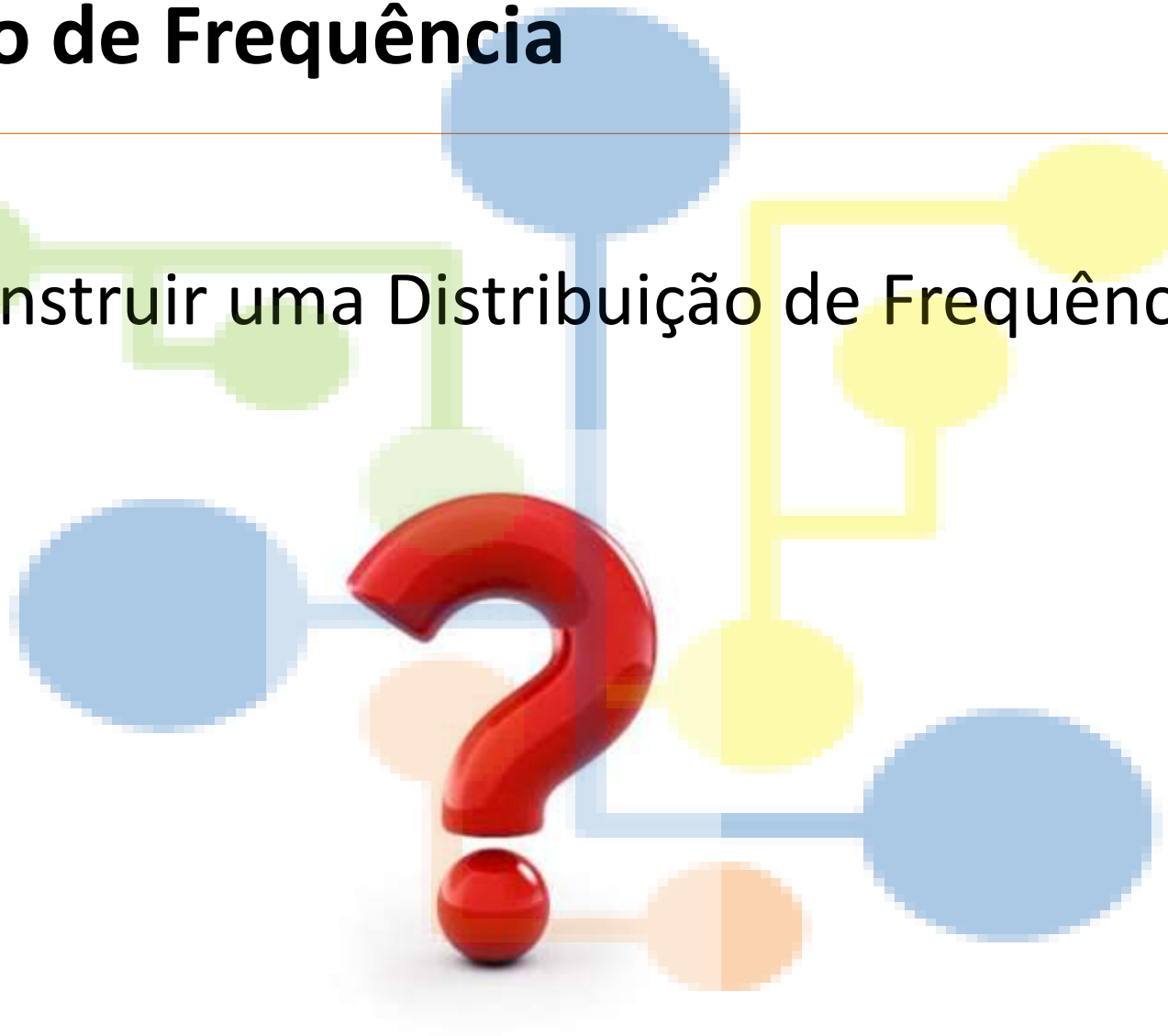
Distribuição de Frequência

Distribuição de Frequência

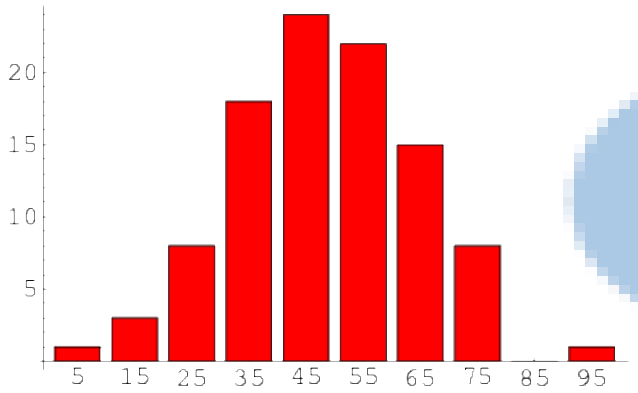
Uma Distribuição de Frequência mostra o número de observações de dados que estão em um intervalo específico.

Distribuição de Frequência

Como construir uma Distribuição de Frequência?



Distribuição de Frequência



1. Criar o Rol
2. Definir a Amplitude
3. Determinar o Número de Classes
4. Determinar o Tamanho do Intervalo de Classes
5. Construir a Distribuição de Frequência

Distribuição de Frequência

Ferramentas Oferecidas Pela Estatística Descritiva

Ferramentas Oferecidas Pela Estatística Descritiva

Quais as principais ferramentas e/ou elementos usados na Estatística Descritiva?

Ferramentas Oferecidas Pela Estatística Descritiva

- Tabela de Frequência
- Tabela de Contingência
- Gráfico de Pizza
- Gráfico de Barras
- Gráfico de Linha
- Gráfico de Dispersão
- Gráfico de Pareto
- Caule e Folha
- Histograma

Ferramentas Oferecidas Pela Estatística Descritiva

Ferramentas Oferecidas Pela Estatística Descritiva

Análise Univariada

Tabela de Frequência
Gráfico de Barras
Gráfico de Pareto
Gráfico de Pizza
Gráfico de Linha
Caule e Folha
Histograma

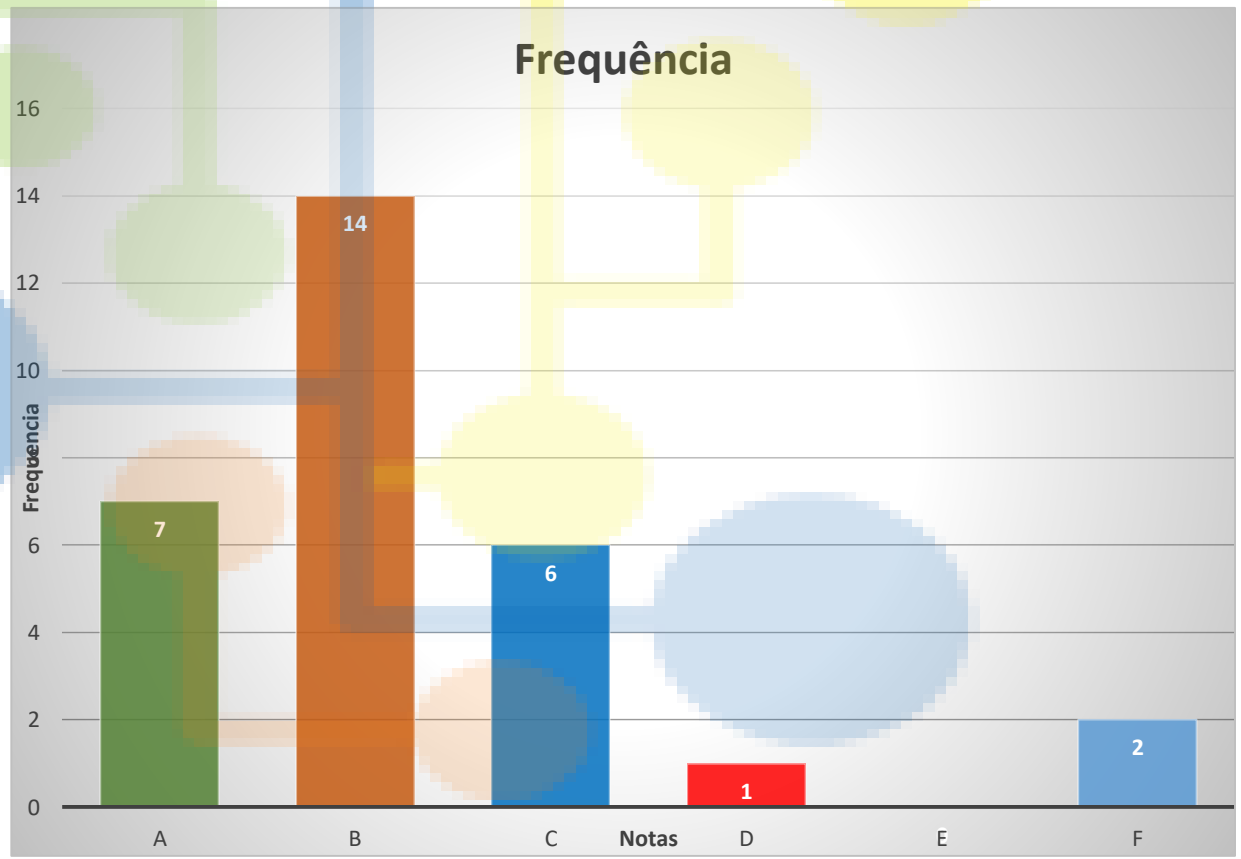
Análise Bivariada

Tabela de Contingência
Gráfico de Dispersão

Ferramentas Oferecidas Pela Estatística Descritiva

Gráfico de Barras

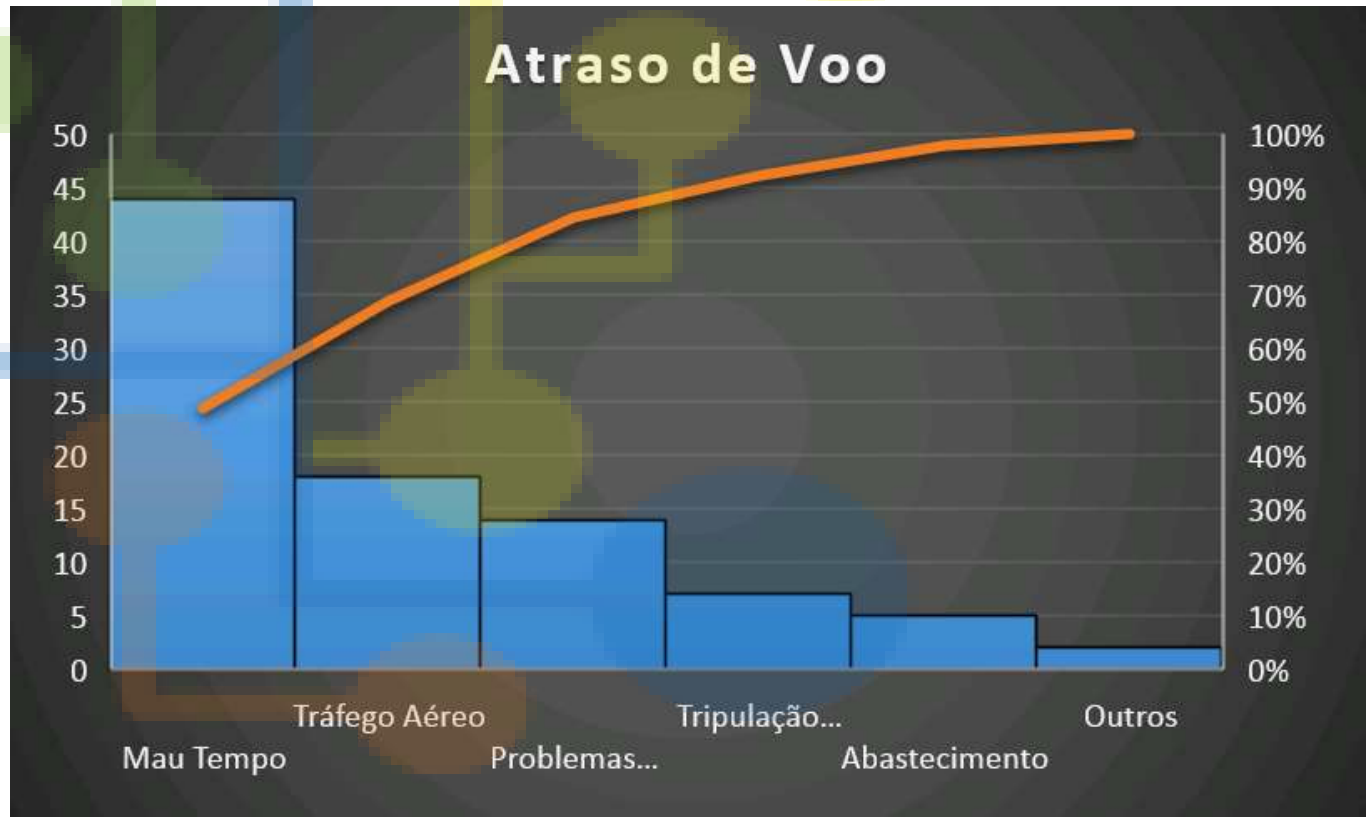
Notas	Frequência
A	7
B	14
C	6
D	1
E	0
F	2



Ferramentas Oferecidas Pela Estatística Descritiva

Gráfico de Pareto

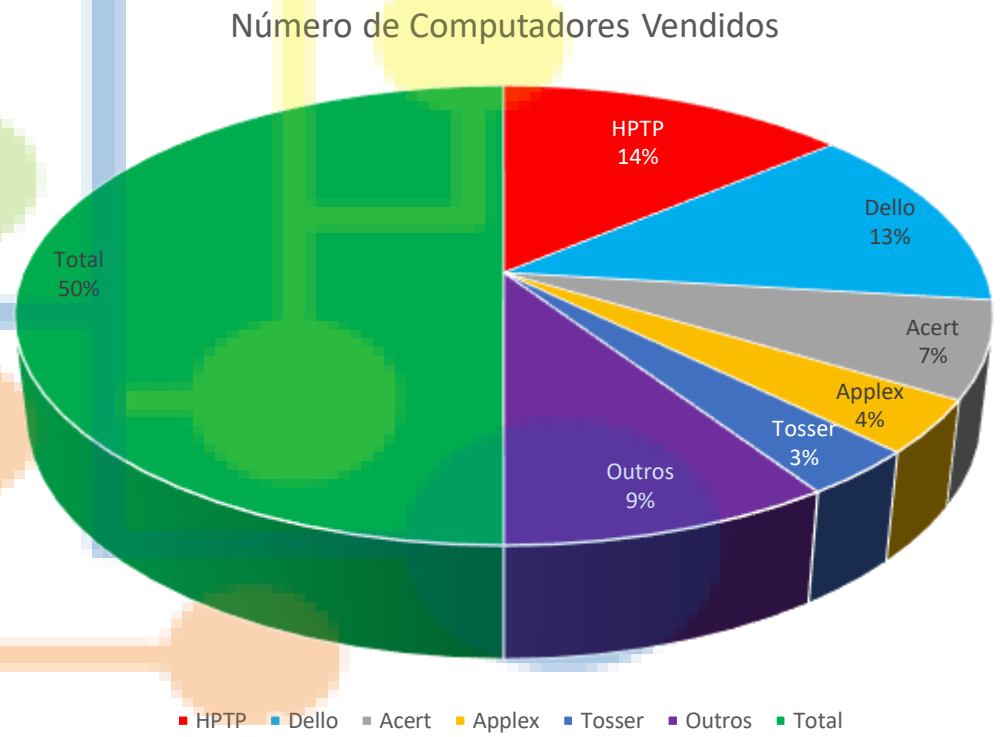
Razões de Atraso de Voo	Frequência (fi)	Frequência Relativa (fri)	Frequencia Relativa Acumulada (Fri)
Mau Tempo	44	0.489	0.489
Tráfego Aéreo	18	0.200	0.689
Problemas Mecanicos	14	0.156	0.844
Tripulação Reduzida	7	0.078	0.922
Abastecimento	5	0.056	0.978
Outros	2	0.022	**1.000**
Total	**90**	**1.000**	



Ferramentas Oferecidas Pela Estatística Descritiva

Gráfico de Pizza

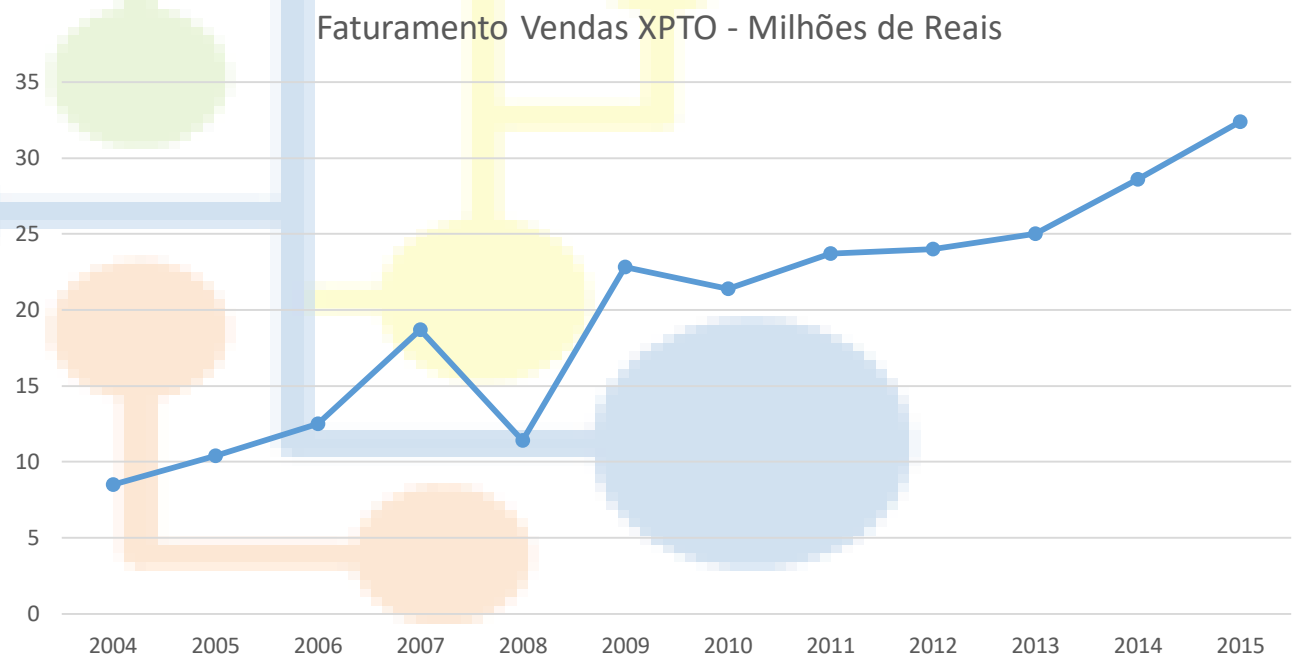
Empresa	Número de Computadores Vendidos
HPTP	4228
Dello	3996
Acert	2076
Applex	1135
Tosser	1005
Outros	2837
Total	**15277**



Ferramentas Oferecidas Pela Estatística Descritiva

Gráfico de Linha

Ano	Faturamento Vendas XPTO - Milhões de Reais
2004	8.5
2005	10.4
2006	12.5
2007	18.7
2008	11.4
2009	22.8
2010	21.4
2011	23.7
2012	24
2013	25
2014	28.6
2015	32.4



Ferramentas Oferecidas Pela Estatística Descritiva

Caule e Folha

O Gráfico Caule e Folha, divide os dados em duas partes:

O caule (ramo) representa os valores maiores e ficam à esquerda do traço vertical.

Diâmetros abdominais de 40 indivíduos

Ramo (dezena)	Folhas (unidades)
5	7 9
6	0 0 2 3 3 3 4 6 6 8 9 9
7	0 0 1 2 2 3 4 5 5 7 8
8	1 3 5 6 6 7 8 8 9
9	1 4 5
10	1 7
11	9

Ferramentas Oferecidas Pela Estatística Descritiva

Caule e Folha

O Gráfico Caule e Folha, divide os dados em duas partes:

As folhas são os menores valores, ficam à direita do traço vertical. Listando todas folhas à direita de cada caule, podemos graficamente descrever como os dados estão distribuídos.

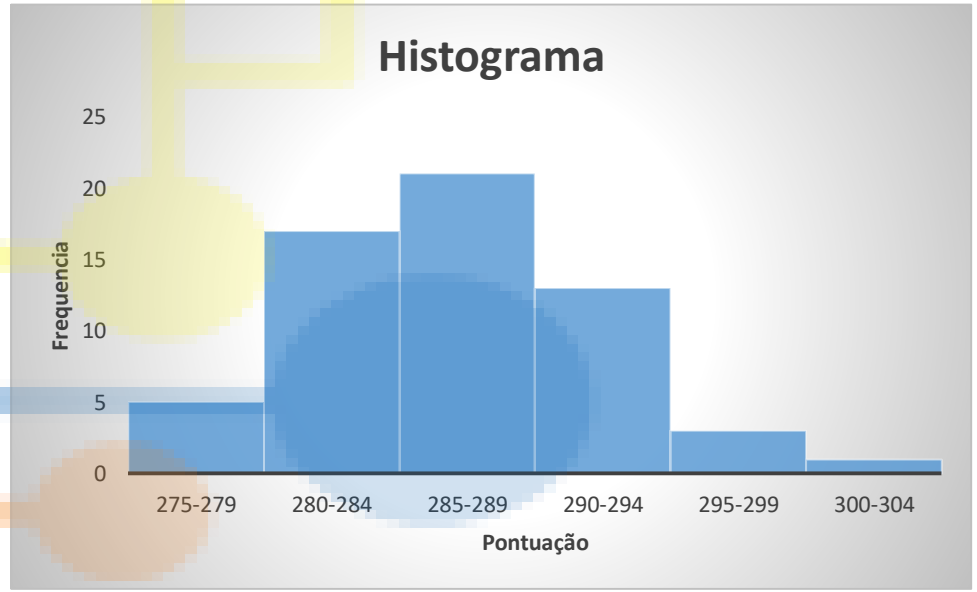
Diâmetros abdominais de 40 indivíduos

Ramo (dezena)	Folhas (unidades)
5	7 9
6	0 0 2 3 3 3 4 6 6 8 9 9
7	0 0 1 2 2 3 4 5 5 7 8
8	1 3 5 6 6 7 8 8 9
9	1 4 5
10	1 7
11	9

Ferramentas Oferecidas Pela Estatística Descritiva

Histograma

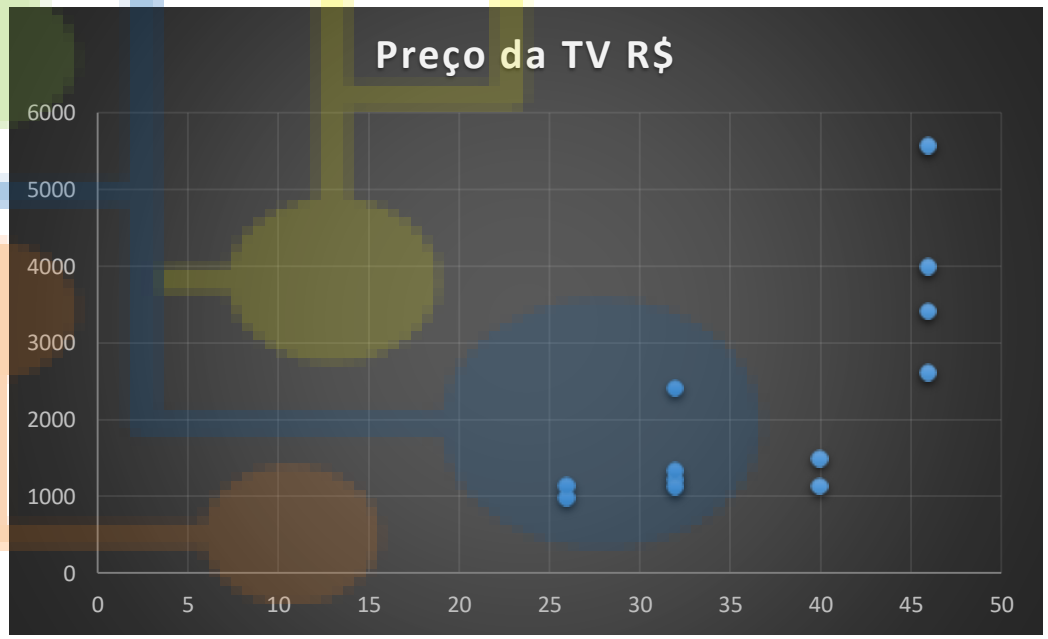
Pontuação Campeonato Golf	Frequência	Frequência Relativa	Frequência Relativa Acumulada
275-279	5	0.083	0.083
280-284	17	0.283	0.367
285-289	21	0.350	0.717
290-294	13	0.217	0.933
295-299	3	0.050	0.983
300-304	1	0.017	**1.000**
Total	**60**	**1.000**	



Ferramentas Oferecidas Pela Estatística Descritiva

Gráfico de Dispersão

Tamanho da TV LED	Preço da TV R$
46	2600
46	3980
32	1200
40	1480
26	970
32	1115
46	3400
46	5560
32	2400
40	1120
26	1130
32	1320



Ferramentas Oferecidas Pela Estatística Descritiva

Tabela de Contingência

Cliente	Sexo	Condição de Pagamento
1	Feminino	Dinheiro
2	Masculino	Cartão
3	Masculino	Dinheiro
4	Masculino	Dinheiro
5	Feminino	Cartão
6	Feminino	Cartão
7	Masculino	Dinheiro
8	Feminino	Cartão
9	Masculino	Cartão
10	Feminino	Dinheiro
11	Masculino	Cartão
12	Feminino	Cartão
13	Masculino	Dinheiro
14	Feminino	Cartão
15	Feminino	Dinheiro

Soma de Cliente	Rótulos de Coluna		
Rótulos de Linha	Cartão	Dinheiro	Total Geral
Feminino	45	26	71
Masculino	22	27	49
Total Geral	**67**	**53**	**120**

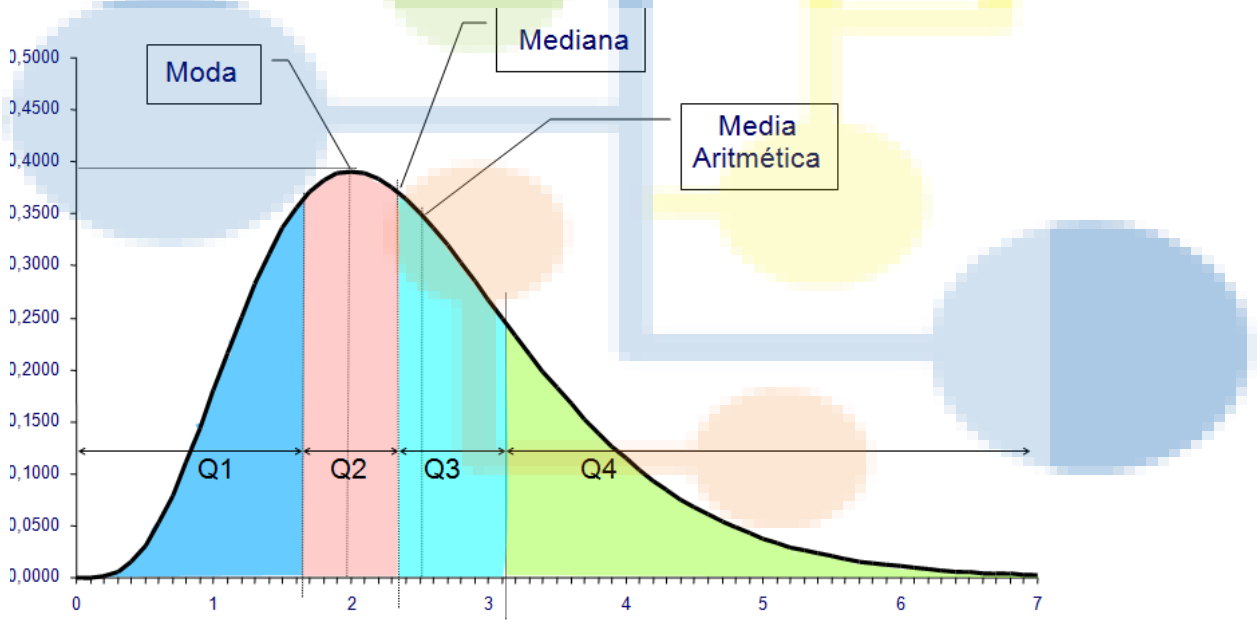
Ferramentas Oferecidas Pela Estatística Descritiva

Cabe a você escolher a ferramenta adequada para cada etapa do processo de análise de dados.

Medidas de Tendência Central

Medidas de Tendência Central

Estas são as principais medidas de tendência central utilizadas em Estatística Descritiva:



Medidas de Tendência Central

Média (Mean ou Average em inglês) é uma medida de tendência central dos dados, ou seja, um número em torno do qual um dataset inteiro está distribuído. É um número único que pode estimar o valor do conjunto de dados completo.

Vamos calcular a média do conjunto de dados com 8 inteiros.

$$x = \frac{12+24+41+51+67+67+85+99}{8} = 55.75$$

Medidas de Tendência Central

Médias são as formas mais simples de identificar tendências em um conjunto de dados. Entretanto, **médias** podem trazer armadilhas que levam a conclusões distorcidas.

Medidas de Tendência Central

Mediana é o valor que divide os dados em 2 partes iguais, ou seja, o número de termos no lado direito é igual ao número de termos no lado esquerdo quando os dados são organizados em ordem crescente ou decrescente.

A **Mediana** será um elemento do meio da distribuição, se o número de termos for ímpar.

A **Mediana** será a média de 2 elementos do meio da distribuição, se o número de termos for par.

Medidas de Tendência Central

A **Moda** é o termo que aparece mais vezes no conjunto de dados, ou seja, o termo que tem a frequência mais alta.

Mas pode haver um conjunto de dados em que não há nenhuma **Moda**, pois todos os valores aparecem o mesmo número de vezes.

Se dois valores aparecerem ao mesmo tempo e mais do que o resto dos valores, o conjunto de dados será **bimodal**. Se três valores aparecerem no mesmo tempo e mais do que o resto dos valores, o conjunto de dados é **trimodal** e, para n modas, esse conjunto de dados é **multimodal**.

Medidas de Tendência Central

O que usar?	Vantagens	Desvantagens
Média	• Relativamente fácil de calcular • Fácil de compreender seu significado	• Pode ser muito afetada por valores extremos
Mediana	• Não é afetada por valores extremos	• Requer mais esforço para ser determinada que a Média
Moda	• Pode ser usada com dados descritivos	• Pode não existir em um conjunto de dados • Pode não ser única (pode existir mais de uma moda)

Medidas de Dispersão

Medidas de Dispersão

Medidas de Dispersão referem-se à variabilidade dentro do conjunto de dados.

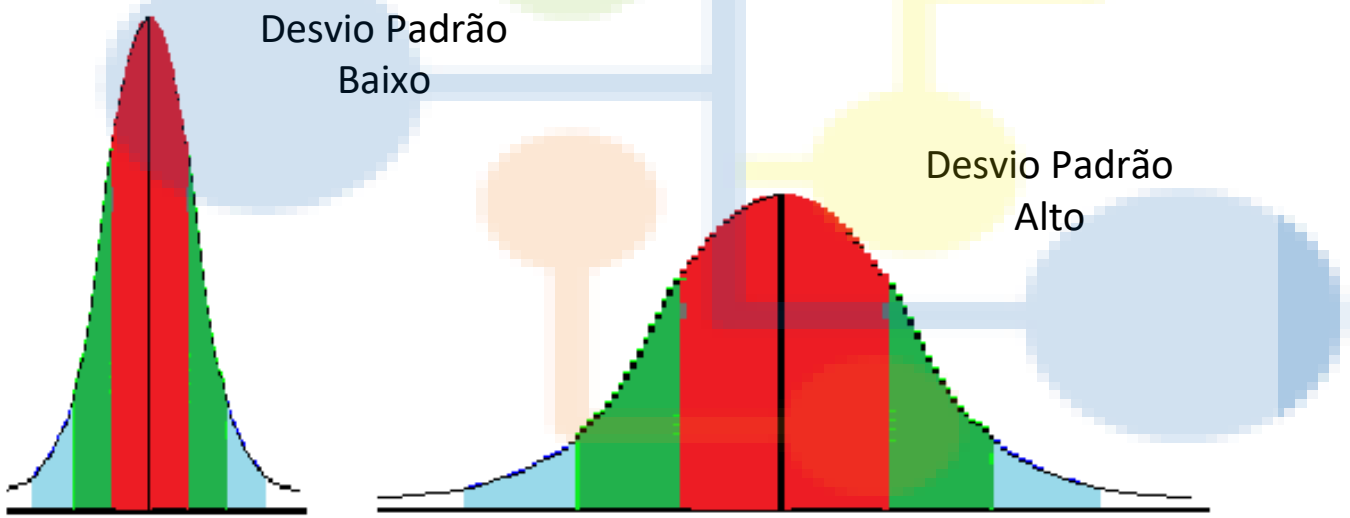
Medidas de Dispersão

Desvio Padrão (Standard Deviation)

O desvio padrão é a medida da distância média entre cada elemento e a média. Isto é, como os dados são distribuídos a partir da média. Um desvio padrão baixo indica que os pontos de dados tendem a estar próximos da média do conjunto de dados, enquanto um desvio padrão alto indica que os pontos de dados estão espalhados em uma faixa mais ampla de valores.

Medidas de Dispersão

Desvio Padrão (Standard Deviation)



Medidas de Dispersão

Desvio Padrão (Standard Deviation)

$$\text{S.D.} = \sqrt{\frac{1}{n-1}\sum_{i=0}^{n}(x-\bar{x})^2}$$

Desvio Padrão da Amostra

$$\text{S.D.} = \sqrt{\frac{1}{n}\sum_{i=0}^{n}(x-\mu)^2}$$

Desvio Padrão da População

Medidas de Dispersão

Variância (Variance)

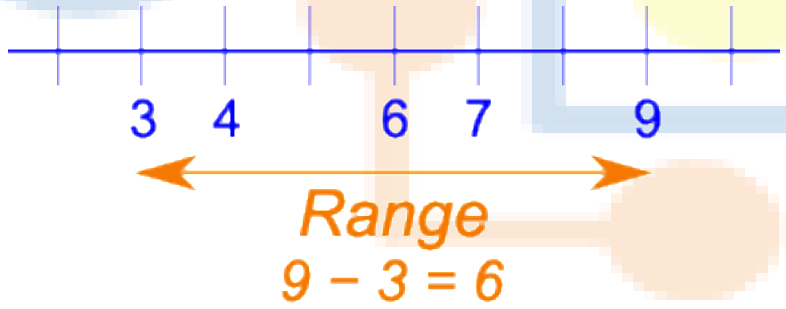
A variância é o quadrado do desvio padrão.

$$\textbf{Variance} = (S.D.)^2$$

Medidas de Dispersão

Intervalo (Range)

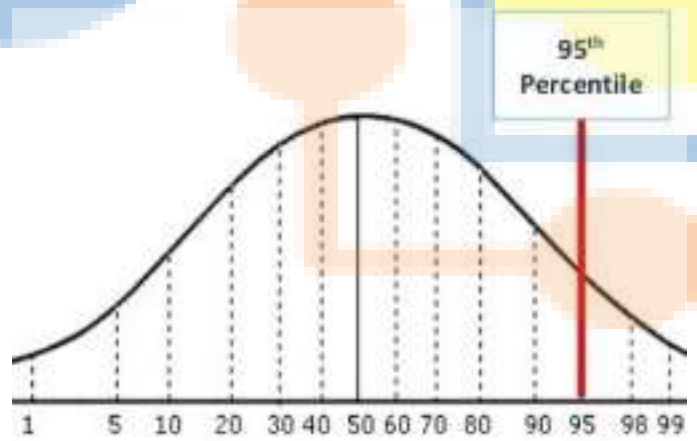
Intervalo é uma das técnicas mais simples de estatística descritiva. É a diferença entre o menor e o maior valor do conjunto de dados.



Medidas de Dispersão

Percentil

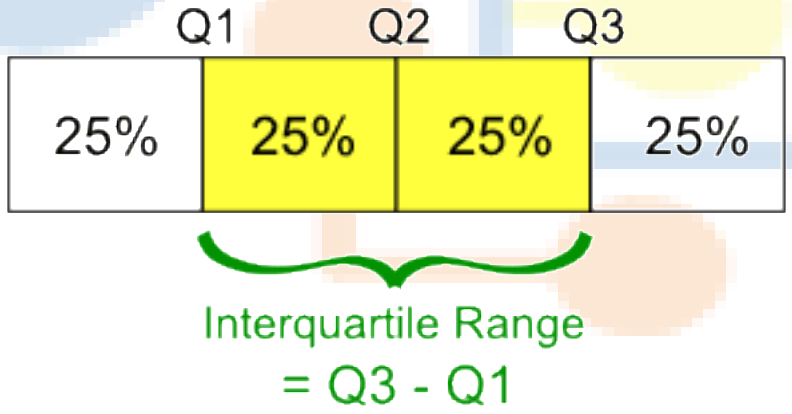
O percentil é uma maneira de representar a posição de um valor no conjunto de dados. Para calcular o percentil, os valores no conjunto de dados devem estar sempre em ordem crescente.



Medidas de Dispersão

Quartil

Os quartis são valores que dividem os dados em *quarters*, desde que os dados sejam classificados em ordem crescente.



Medidas de Forma
Skewness e kurtosis

Medidas de Forma - Skewness e kurtosis

As medidas de assimetria (skewness) e curtose (kurtosis) caracterizam a forma da distribuição de elementos em torno da média.

Medidas de Forma - Skewness e kurtosis

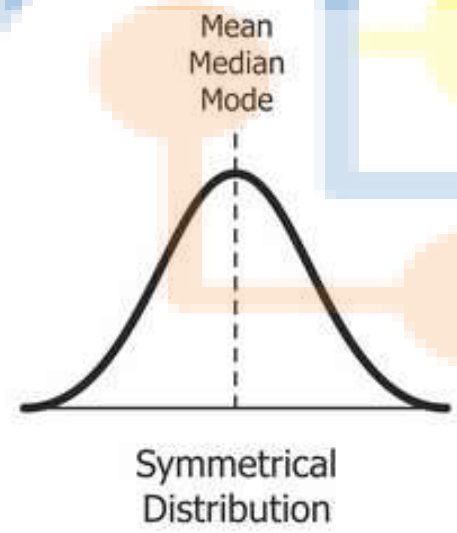
Assimetria (Skewness)

Skewness é uma medida da assimetria da distribuição de probabilidade de uma variável aleatória de valor real sobre sua média. O valor da assimetria pode ser positivo, negativo ou indefinido.

Medidas de Forma - Skewness e kurtosis

Assimetria (Skewness)

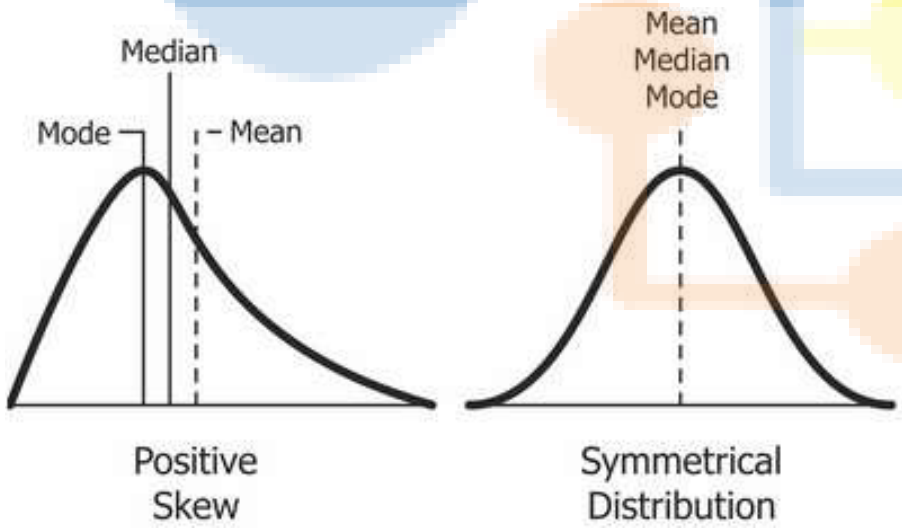
Em uma distribuição normal perfeita, as caudas de cada lado da curva são imagens espelhadas exatas uma da outra.



Medidas de Forma - Skewness e kurtosis

Assimetria (Skewness)

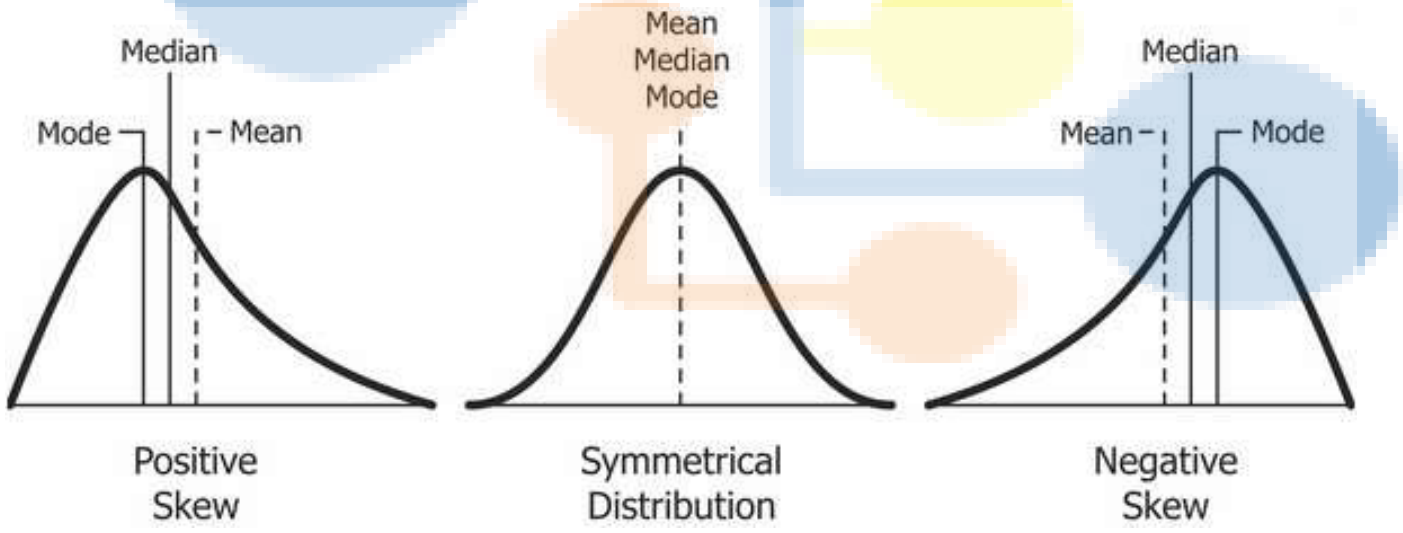
Quando uma distribuição é inclinada para a direita, a cauda no lado direito da curva é maior que a cauda no lado esquerdo, e a média é maior que a moda. Essa situação também é chamada de assimetria positiva.



Medidas de Forma - Skewness e kurtosis

Assimetria (Skewness)

Quando uma distribuição é inclinada para a esquerda, a cauda do lado esquerdo da curva é maior que a cauda do lado direito e a média é menor que a moda. Essa situação também é chamada de assimetria negativa.



Medidas de Forma - Skewness e kurtosis

Assimetria (Skewness)

Para calcular o coeficiente de assimetria, usamos:

$$\frac{\text{Mean - Mode}}{\text{Standard Deviation}}$$

First Coefficient of Skewness
(Mode skewness)

$$\frac{\text{3 (Mean - Median)}}{\text{Standard Deviation}}$$

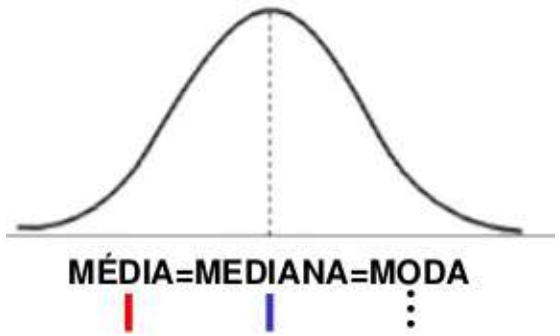
Second Coefficient of Skewness
(Median skewness)

Medidas de Forma - Skewness e kurtosis

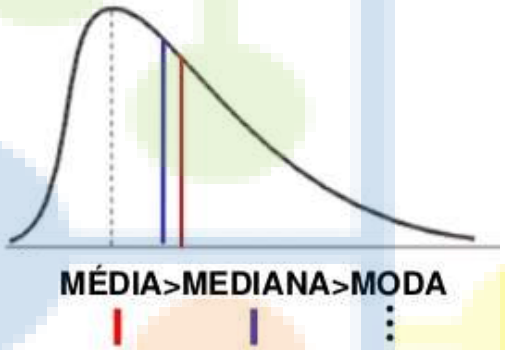
Assimetria (Skewness)

- A direção da assimetria é dada pelo sinal. Um zero significa nenhuma assimetria.
- Um valor negativo significa que a distribuição é negativamente assimétrica. Um valor positivo significa que a distribuição está positivamente assimétrica.
- O coeficiente compara a distribuição da amostra com uma distribuição normal. Quanto maior o valor, mais a distribuição difere de uma distribuição normal.

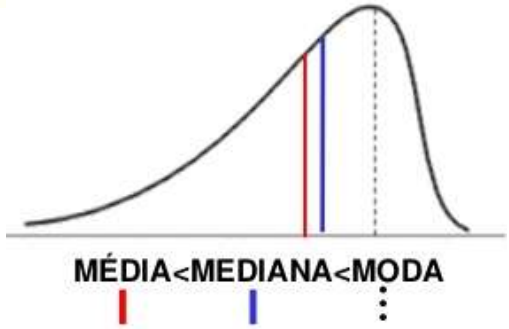
Medidas de Forma - Skewness e kurtosis



Distribuição Simétrica



Distribuição Assimétrica Positiva ou à direita



Distribuição Assimétrica Negativa ou à esquerda

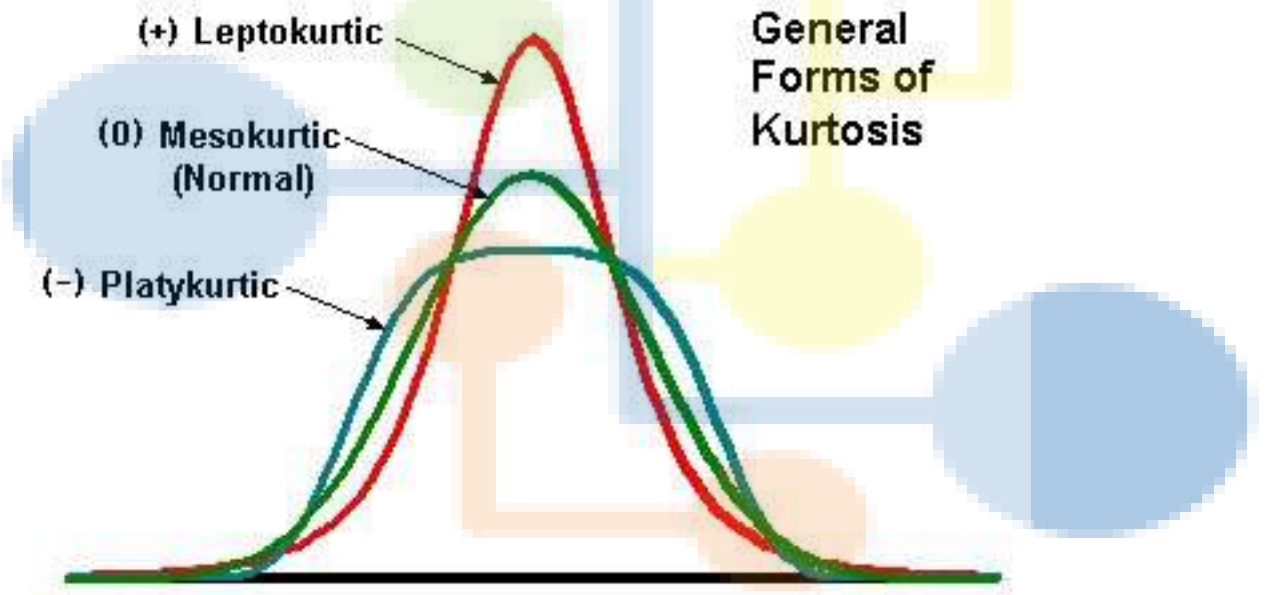
Medidas de Forma - Skewness e kurtosis

Curtose (Kurtosis)

Um dos coeficientes mais utilizados para medir o grau de achatamento ou curtose de uma distribuição é o coeficiente percentílico de curtose, ou simplesmente coeficiente de curtose (k), calculado a partir do intervalo interquartil dos percentis de ordem 10 e 90.

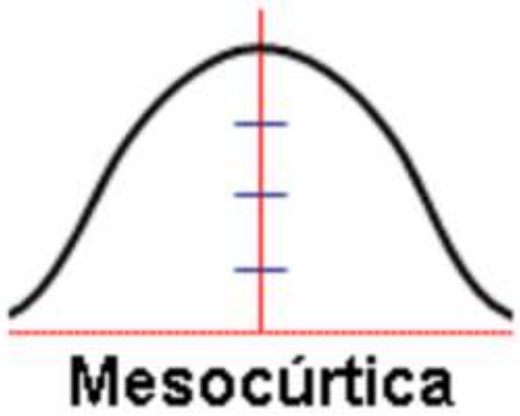
Medidas de Forma - Skewness e kurtosis

Curtose (Kurtosis)



Medidas de Forma - Skewness e kurtosis

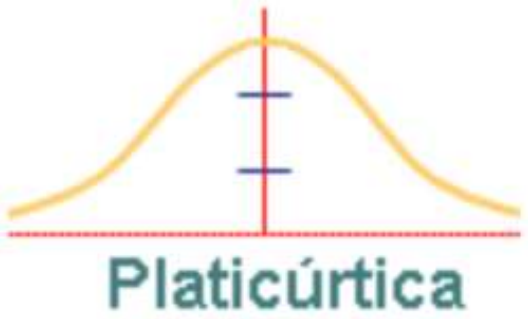
Curtose (Kurtosis)



Quando a forma da distribuição não é nem muito achatada e nem muito alongada, com uma aparência semelhante à da curva normal, é denominada mesocúrtica.

Medidas de Forma - Skewness e kurtosis

Curtose (Kurtosis)

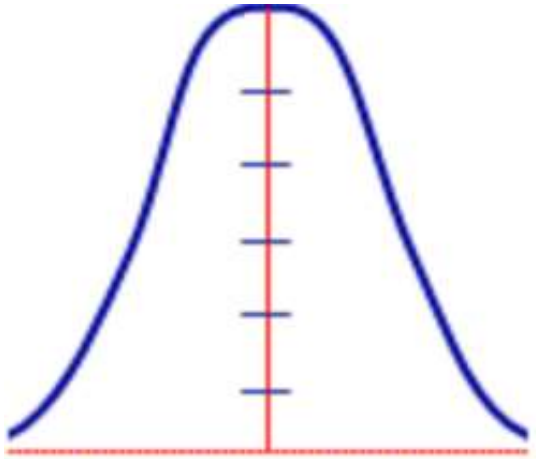


Por outro lado, quando a distribuição apresenta uma curva de frequências mais achatada que a curva normal é denominada platicúrtica.

Apresenta uma medida de curtose menor que a da distribuição normal.

Medidas de Forma - Skewness e kurtosis

Curtose (Kurtosis)



Ou ainda, quando a distribuição apresenta uma curva de frequências mais alongada que a curva normal é denominada leptocúrtica.

Apresenta uma medida de curtose maior que a da distribuição normal.

Medidas de Forma - Skewness e kurtosis

Curtose (Kurtosis)

$$K = \frac{\frac{1}{2}(Q_3 - Q_1)}{P_{90} - P_{10}} \rightarrow 0{,}263$$

Se k = 0,263 → dizemos que a distribuição é mesocúrtica
Se k > 0,263 → dizemos que a distribuição é platicúrtica
Se k < 0,263 → dizemos que a distribuição é leptocúrtica

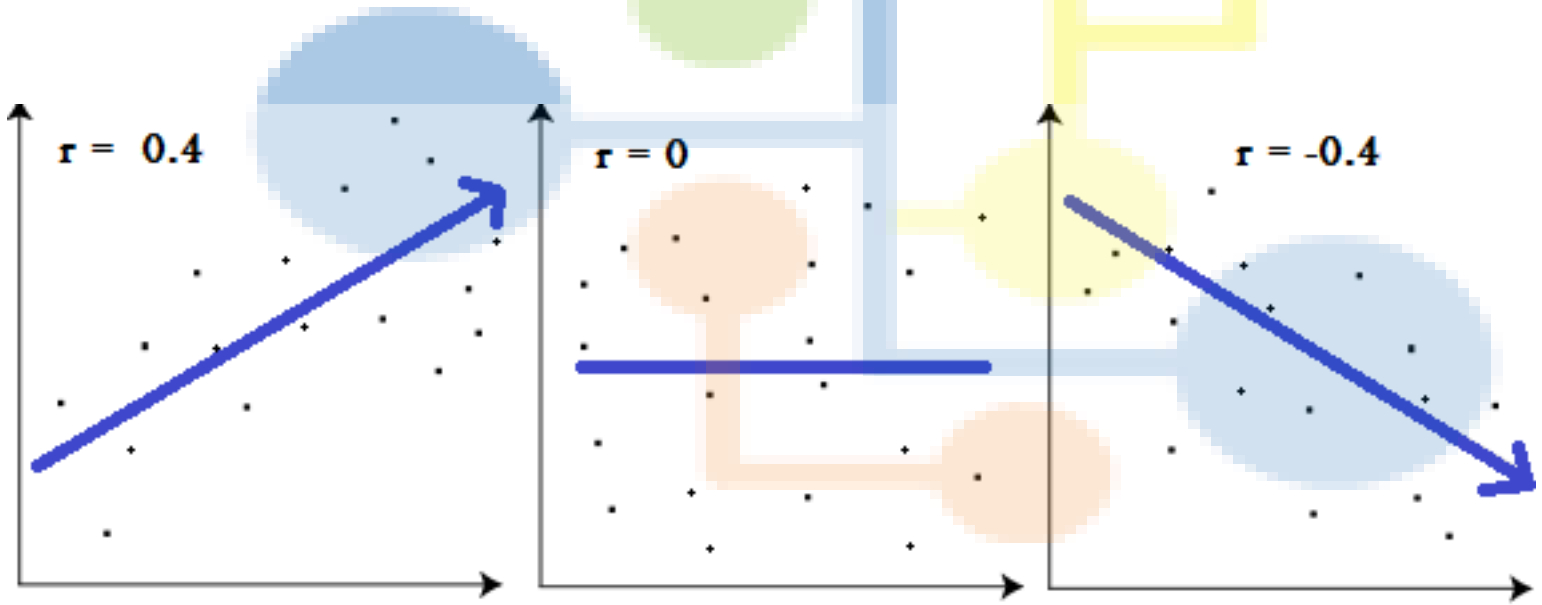
Coeficiente de Correlação

Coeficiente de Correlação

A Correlação permite determinar quão fortemente os pares de variáveis estão relacionados.

Coeficiente de Correlação

O principal resultado de uma correlação é chamado de **coeficiente de correlação** (ou “r”). Varia de -1.0 a +1.0. Quanto mais próximo r for +1 ou -1, mais próximas as duas variáveis estarão relacionadas.



Tenha uma Excelente Jornada de Aprendizagem.

Muito Obrigado por Participar!

Equipe Data Science Academy