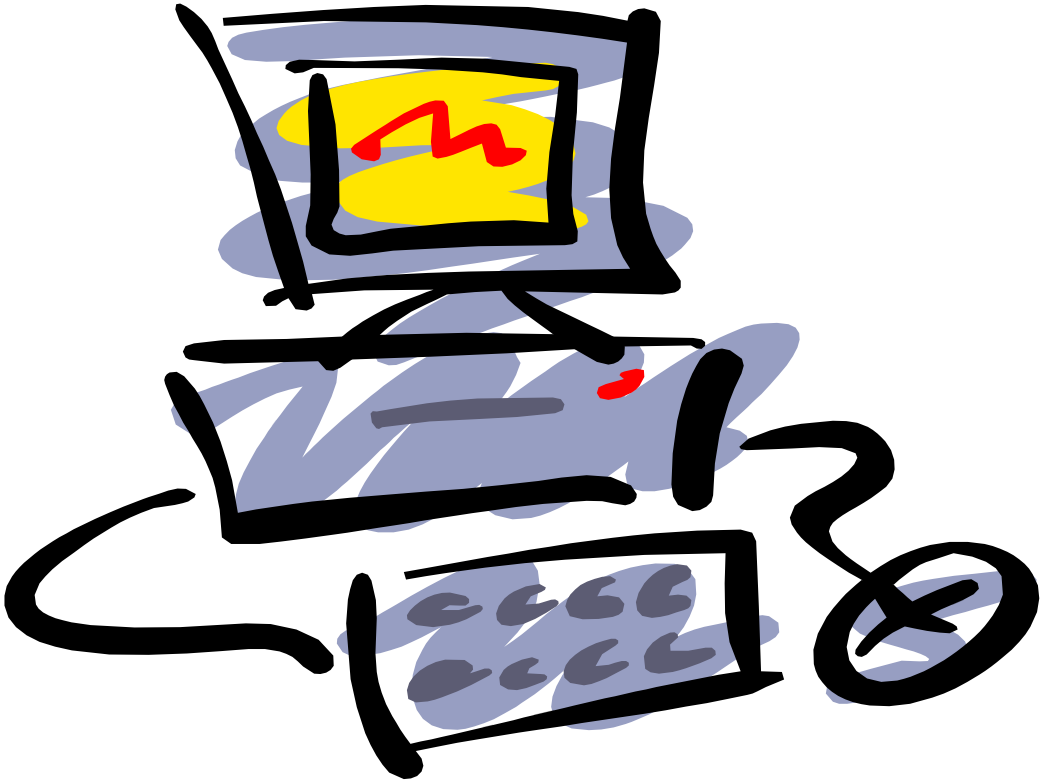
Inteligência nos Negócios (Business Inteligente)

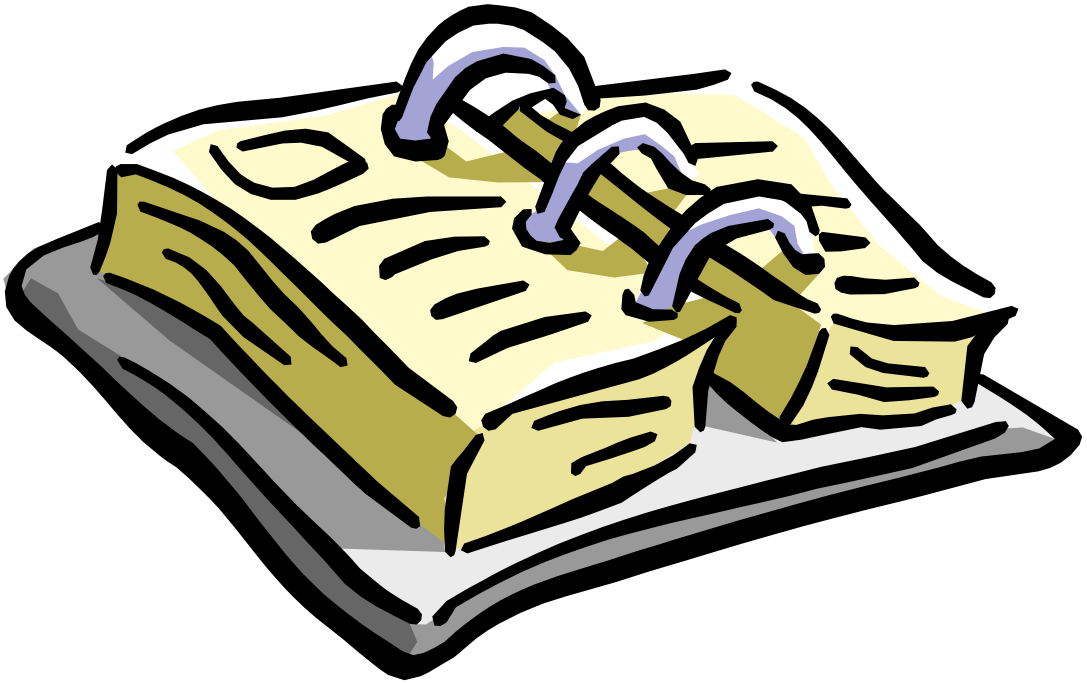


Sistemas de Informação – Sistemas de Apoio a Decisão

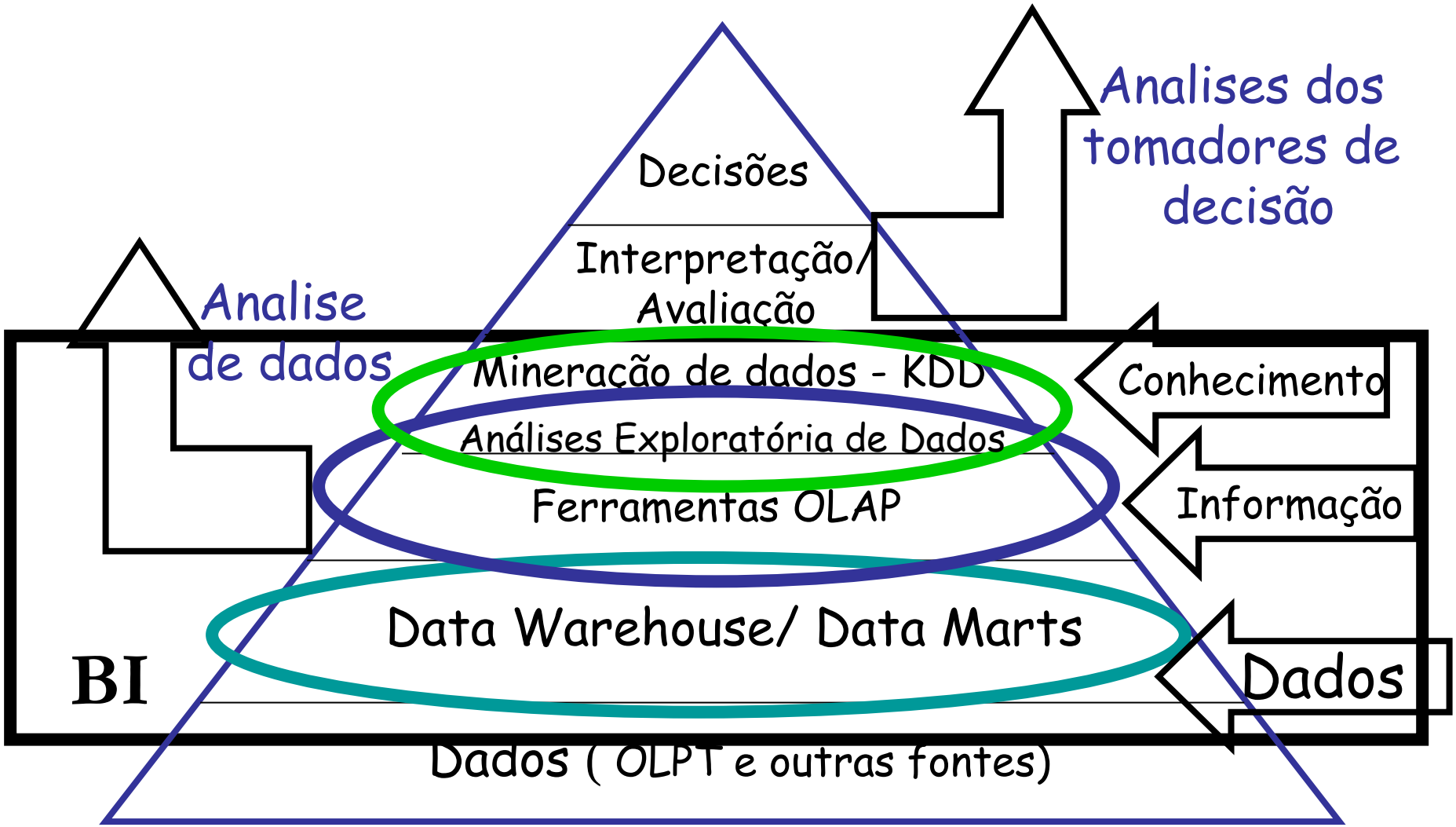
Aran Bey Tcholakian Morales, Dr. Eng.

(Apostila 6)

3. Interpretação dos dados e elaboração de informações



Fundamentação da disciplina



3.2 Análise Exploratória de Dados (AED)

Análise Exploratória de Dados (AED)

AED consiste em **ORGANIZAR** e **RESUMIR** os dados coletados por meio de tabelas, gráficos ou medidas numéricas (técnicas de **estatística descritiva**) e, a partir dos dados resumidos, procurar identificar padrões, comportamentos, relações e dependências. O objetivo final é tornar mais clara a descrição dos dados a fim de ajudar o analista a desenvolver algumas hipóteses sobre o assunto e modelos apropriados para tais dados, isto é auxiliar na **INTERPRETAÇÃO** dos dados.

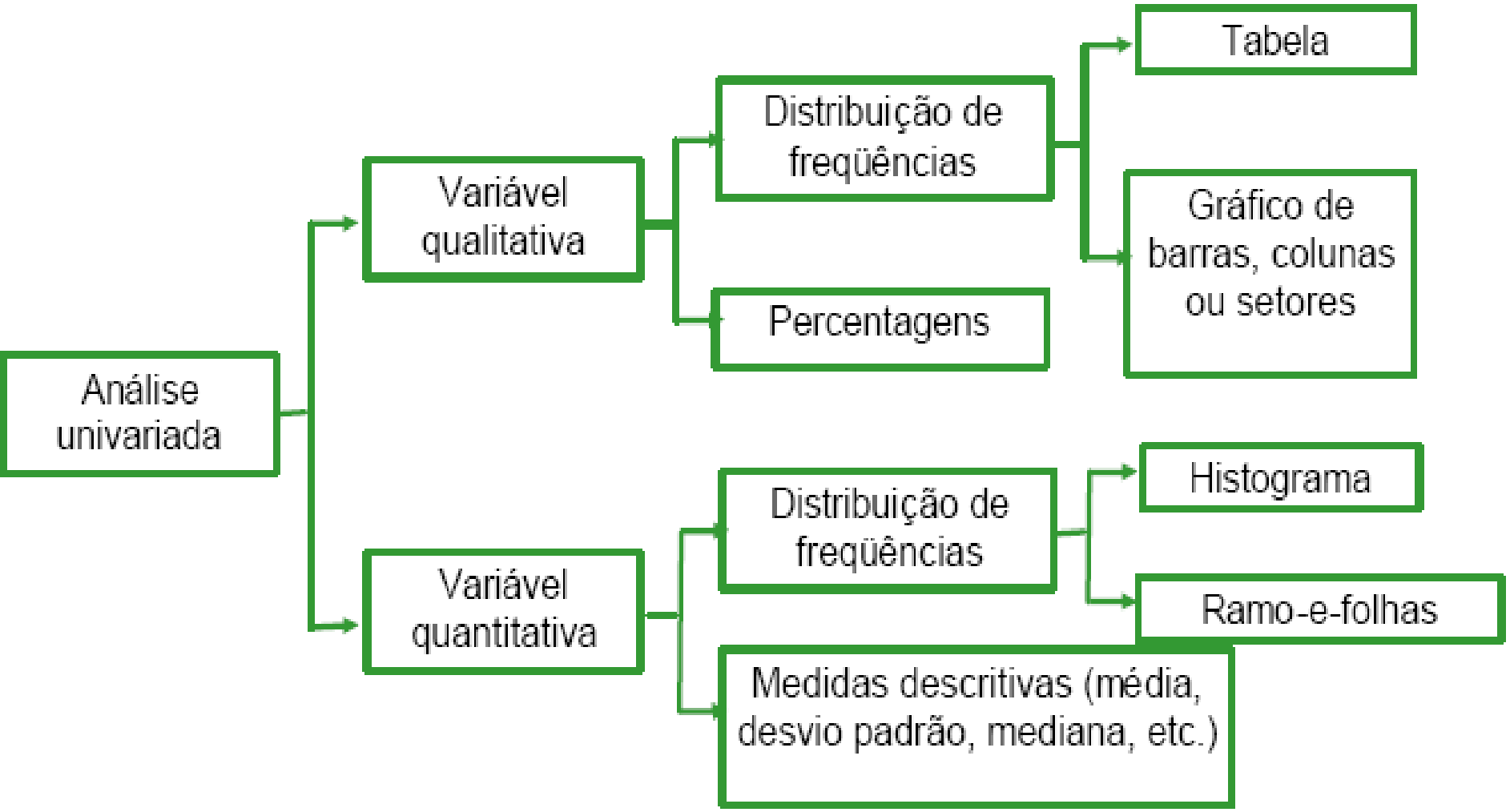
Análise Exploratória de Dados (AED)

A **AED** inclui um conjunto de ferramentas gráficas e descritivas para explorar os dados, como pré-requisito para uma análise de dados mais formal (Predição, Previsão, Estimação, Classificação e Testes de Hipóteses) e como parte da construção de modelos.

A **AED** emprega técnicas estatísticas descritivas e gráficas para estudar um conjunto de dados, detectando agrupamento de dados, medidas de tendência central, de ordenação, de dispersão e de correlação entre variáveis.

A **estatística descritiva** é um conjunto de técnicas que permite, de forma sistemática, organizar, descrever, analisar e interpretar dados oriundos de estudos ou experimentos por meio do uso de certas medidas-síntese que tornem possível a interpretação de resultados.

Análise Exploratória de Dados



Análise Exploratória de Dados

Descrição de dados com tabelas e gráficos:

A **freqüência absoluta** de uma variável é uma função formada pelos valores da variável e por suas freqüências (número de repetições do valor de cada variável).

A **freqüência relativa** do valor de uma variável é o resultado de dividir a sua freqüência absoluta pelo tamanho da amostra.

A **freqüência acumulada** do valor de uma variável é a soma das freqüências absolutas ou relativas.

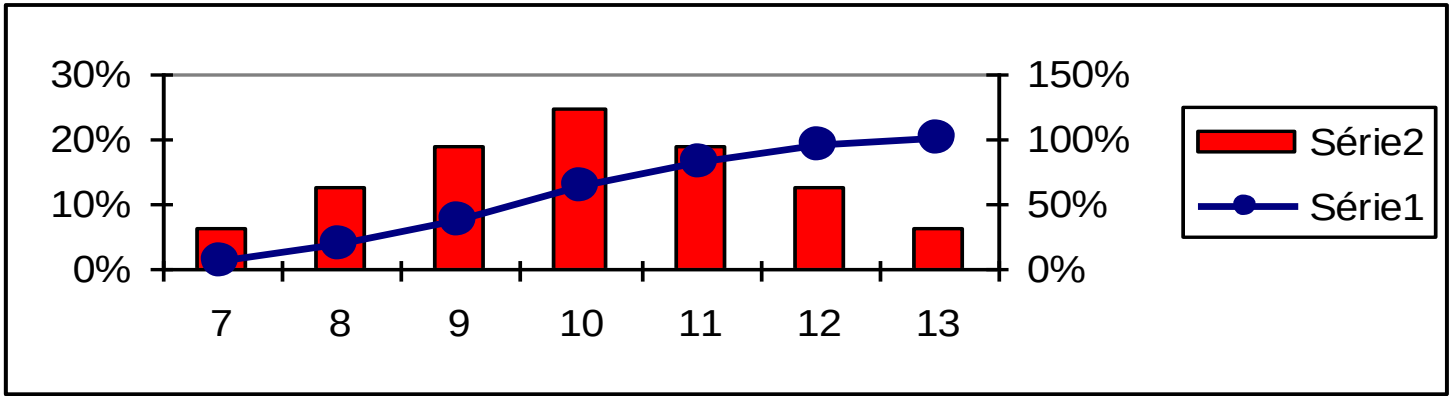
O **histograma** é uma forma gráfica de apresentar as freqüências de uma variável.

Análise Exploratória de Dados

Exemplo (LAPPONI, 2000): dada a amostra a seguir, construir a tabela de freqüências e o histograma.

10	9	9	9	10	11	8	8
10	11	12	13	10	11	12	7

	Freqüências		
Dados	Absolutas	Relativas	Acumuladas
7	1	6,3%	6,3%
8	2	12,5%	18,8%
9	3	18,8%	37,5%
10	4	25,0%	62,5%
11	3	18,8%	81,3%
12	2	12,5%	93,8%
13	1	6,3%	100,0%



Análise Exploratória de Dados

Medidas de tendência central:

- **Média aritmética**: somam-se todos os valores do conjunto e divide-se o resultado pelo número total.

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

- **Mediana**: divide o conjunto de dados em duas partes iguais, metade dos dados é ***menor*** do que a mediana e a outra metade é ***maior*** do que a mediana. Primeiramente é preciso obter a posição da mediana (a partir do conjunto ordenado):

Md elemento da posição $(n+1)/2$ para n impar

Md soma dos elementos da posição $n/2$ e $(n/2)+1$, dividido entre 2 para n par

Análise Exploratória de Dados

Medidas de tendência central

Moda é o valor mais freqüente do conjunto de dados. Teoricamente é o valor mais provável.

Um conjunto pode ter uma única moda, várias modas (dois ou mais valores ocorrem igual número de vezes) ou nenhuma moda (nenhum valor se repete).

Costuma ser utilizada em conjunção com média e mediana para avaliar a simetria do conjunto de dados.

Análise Exploratória de Dados

Exemplo:

Deseja-se estudar o número de falhas no envio de mensagens,

Considerando-se três algoritmos diferentes para o envio dos pacotes:

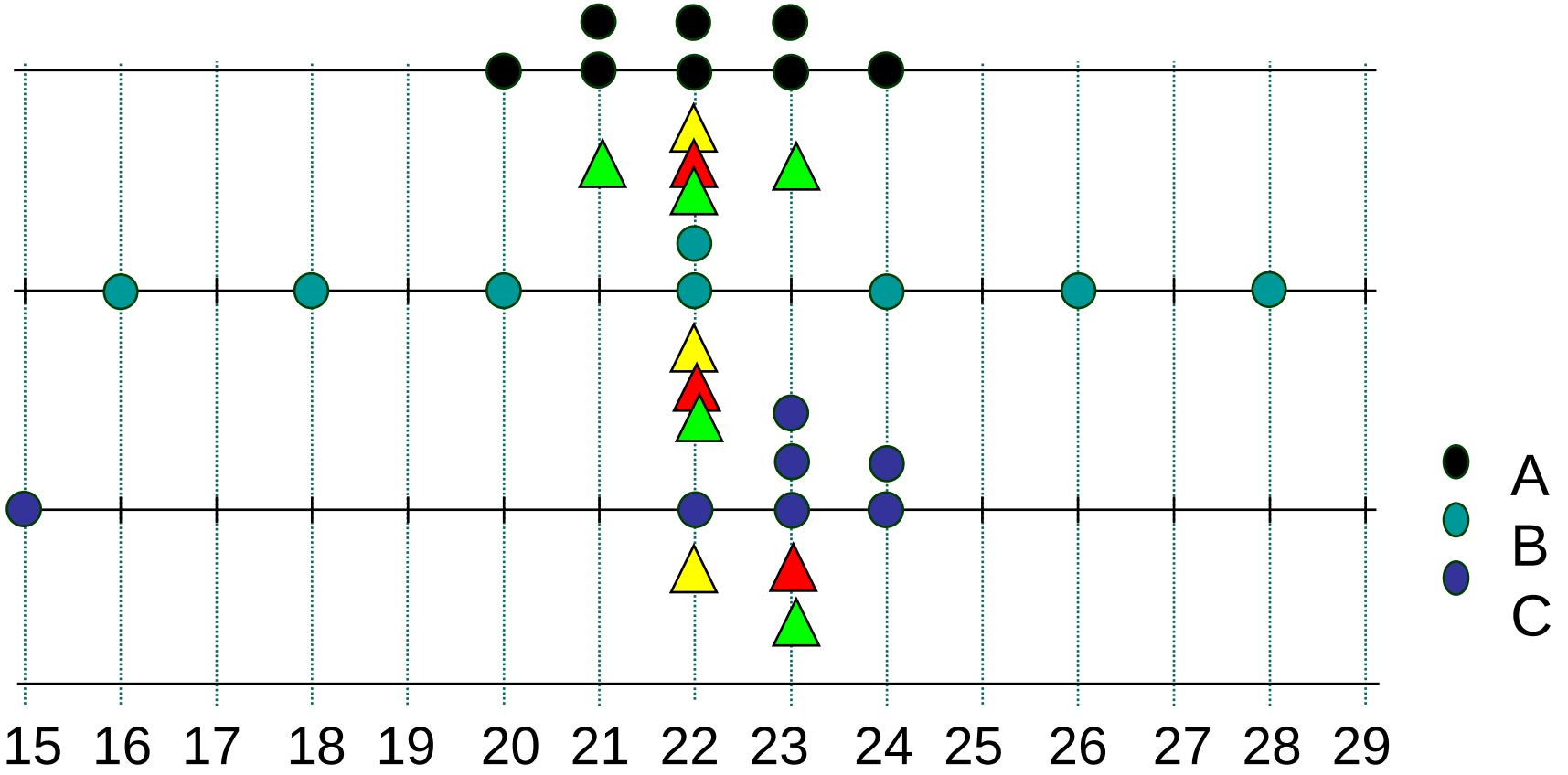
Algoritmo A (8 observações)
Algoritmo B (8 observações)
Algoritmo C (7 observações)

Número de falhas a cada 10.000 mensagens enviadas.

A:	20	21	21	22	22	23	23	24
B:	16	18	20	22	22	24	26	28
C:	15	22	23	23	23	24	24	

Encontrar a média, a mediana e a moda e observar a simetria das três amostras de dados.

Análise Exploratória de Dados



Número de falhas

Média Mediana Moda

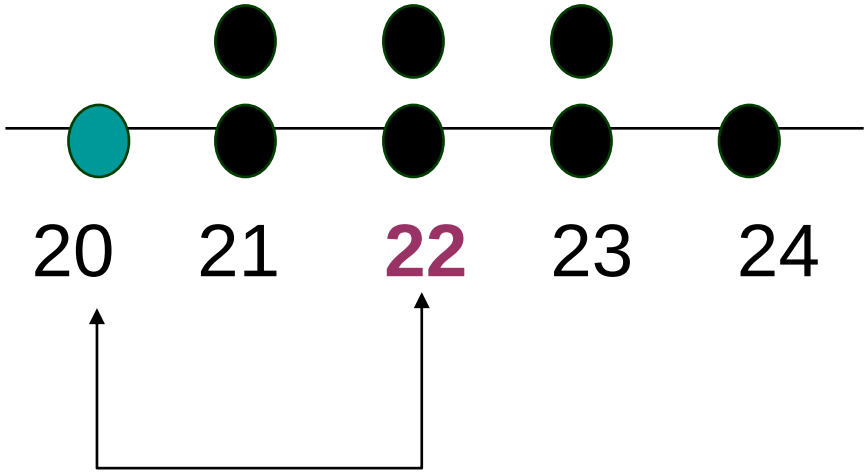
Análise Exploratória de Dados

Medidas de dispersão

As medidas de tendência central são insuficientes para identificar uma Variável, sendo necessária uma média que mostra a **dispersão** dos dados.

Como medir a dispersão?

Exemplo: A (20 21 21 22 22 23 23 24)



distância (desvio) em relação à média

Análise Exploratória de Dados

Valores	20 21 21 22 22 23 23 24	176
Média	22	
Desvios	-2 -1 -1 0 0 1 1 2	0
Desvios quadráticos	4 1 1 0 0 1 1 4	12

Os **desvios** de uma variável podem medir a **dispersão** dos valores da variável. Como a soma dos desvios da variável sempre dá zero, para manter o desvio como medida de variabilidade utilizamos o **desvio quadrático**.

Para medir a dispersão de uma variável, é considerada então a média da soma dos desvios quadráticos (**variância**).

Análise Exploratória de Dados

A **variância** ($\mathbf{S^2}$) é uma média dos desvios quadráticos.

Por conveniência, usa-se **n-1** no denominador em vez de **n**.

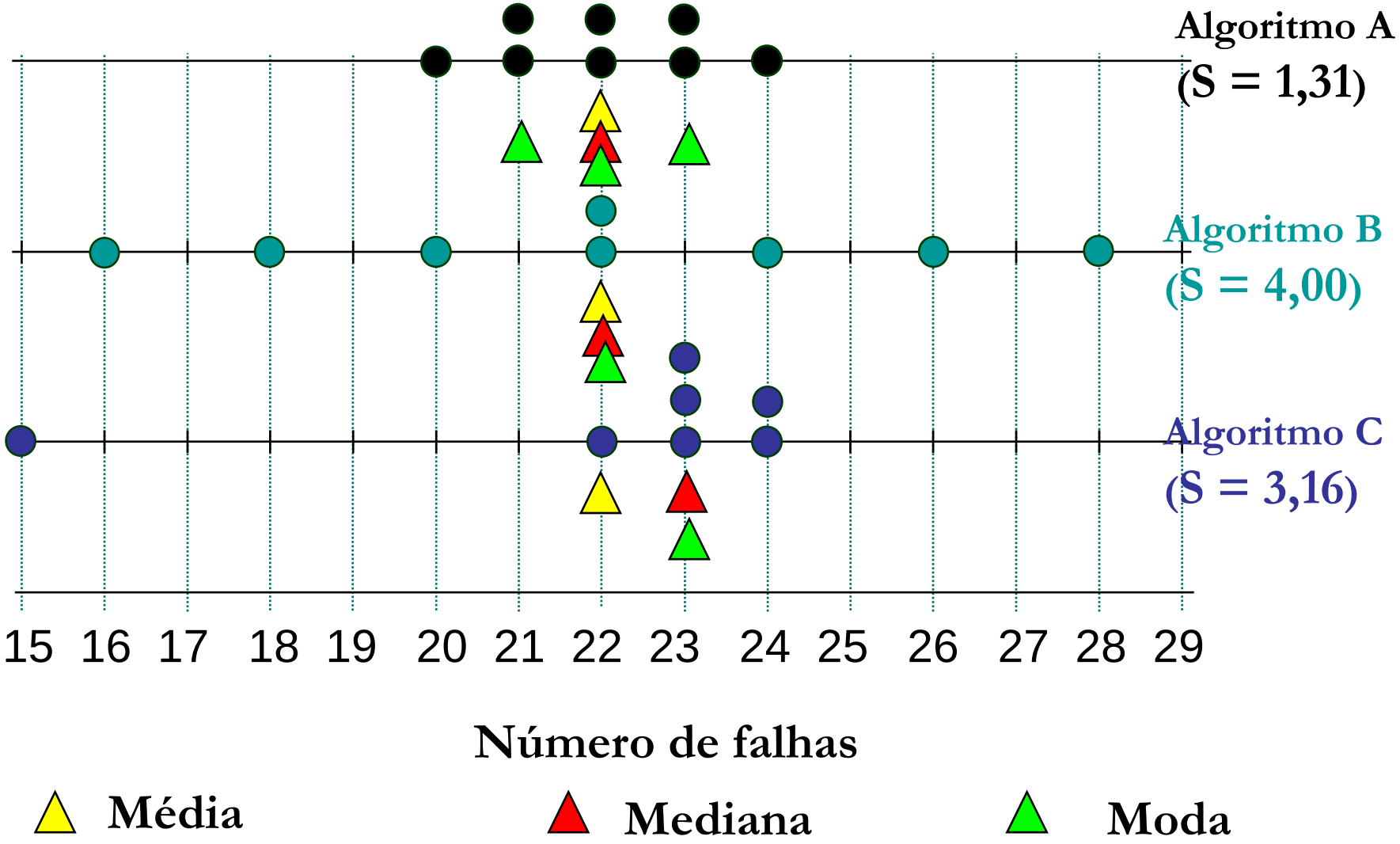
$$S^2 = \frac{\sum \left(X - \overline{X}\right)^2}{n-1}$$

No exemplo apresentado (algoritmo A), a variância é: 12/7 = 1,71

O **desvio padrão** (**S**) é a raiz quadrada da variância, no caso do algoritmo A, o S = 1,31.

Quanto menor o S, mais os valores de aproximam da média.

Análise Exploratória de Dados



Análise Exploratória de Dados

Medidas de correlação

Os retornos anuais durante os últimos anos da Ação A e da Ação B negociadas na bolsa de valores estão registrados na tabela, assim como a análises estatística dos retornos das duas ações (LAPPONI, 2000).

Ano	Ação A	Ação B
1994	9,0%	12,0%
1995	10,0%	10,5%
1996	12,0%	9,0%
1997	10,5%	11,0%
1998	9,5%	12,5%
1999	11,0%	10,0%
Mediana	**10,3%**	**10,8%**
Média	**10,3%**	**10,8%**
DP	**1,1%**	**1,3%**
CV	**10,5%**	**11,9%**
Inclinação	**0,46**	**-0,08**

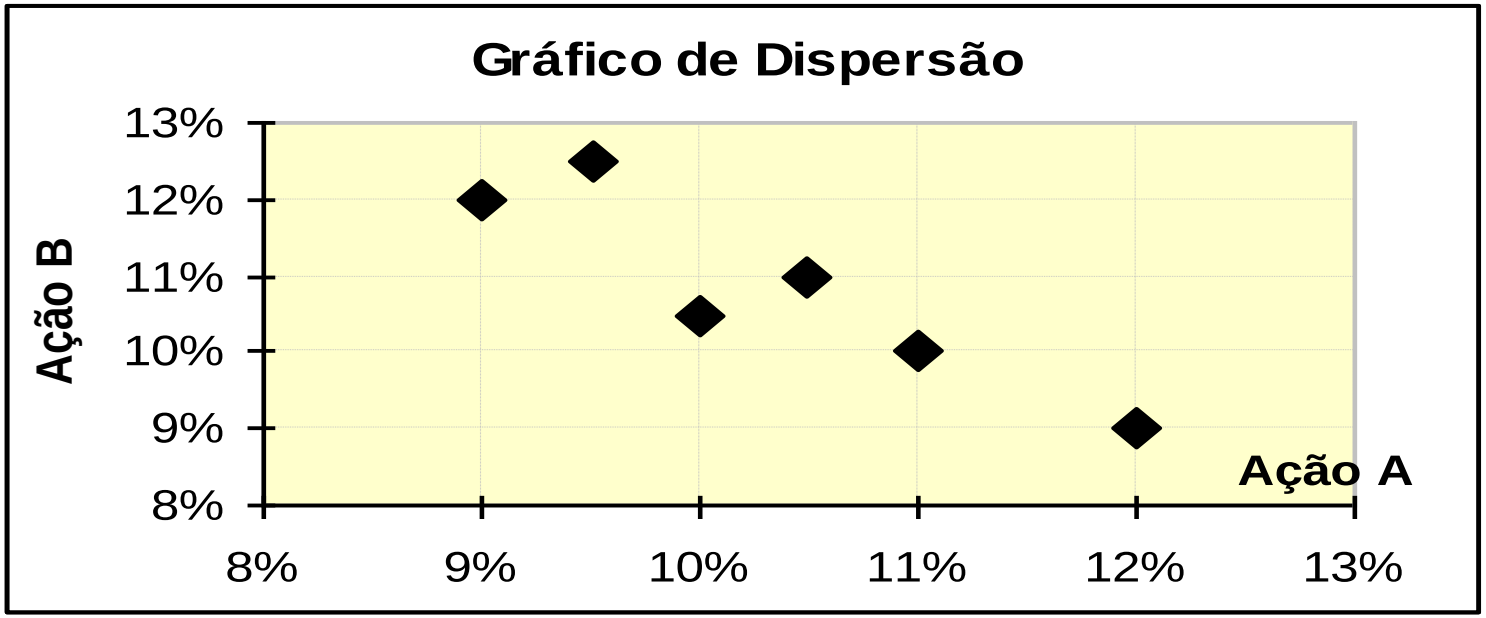
Em qual ação você aplicaria?

Existem particularidades que as medidas de tendência central e dispersão não capturam.

Qual é a diferença entre os retornos das duas ações?

Análise Exploratória de Dados

Vemos no gráfico de dispersão que os retornos têm tendências opostas (quando uma aumenta, a outra diminui). A **covariância** e o **coeficiente de correlação** medem a **tendência** e a **força da relação linear** entre duas variáveis.



Análise Exploratória de Dados

A **covariância** é a média dos produtos dos desvios.

$$S_{xy} = \frac{\sum (Xi - \bar{X})(Yi - \bar{Y})}{n-1}$$

A unidade de medida da **covariância** é o produto das unidades de medidas das duas variáveis.

Por exemplo: faturamento mensal em R$ e retorno mensal em %, tem como unidade de **covariância** R$%, o que não tem significado.

Para facilitar a relação, foi definido o **coeficiente de correlação**.

Análise Exploratória de Dados

O **coeficiente de correlação** de duas variáveis assume valores no intervalo [-1, +1] e é definido como :

$$r_{xy} = \frac{S_{xy}}{S_x \,.\, S_y}$$

- $r = +1$, correlação positiva perfeita, existe uma relação linear perfeita entre as duas variáveis
- r próximo de +1, forte correlação positiva
- $r = 0$, não há correlação, não existe tendência dos dados
- r próximo de -1 , forte correlação negativa
- $r = -1$, correlação negativa perfeita, existe uma relação linear perfeita entre as duas variáveis