AULA 2:
MEDIDAS DESCRITIVAS
PROBABILIDADE/TEOREMA BAYES
DISTRIBUIÇÕES DE PROBABILIDADE

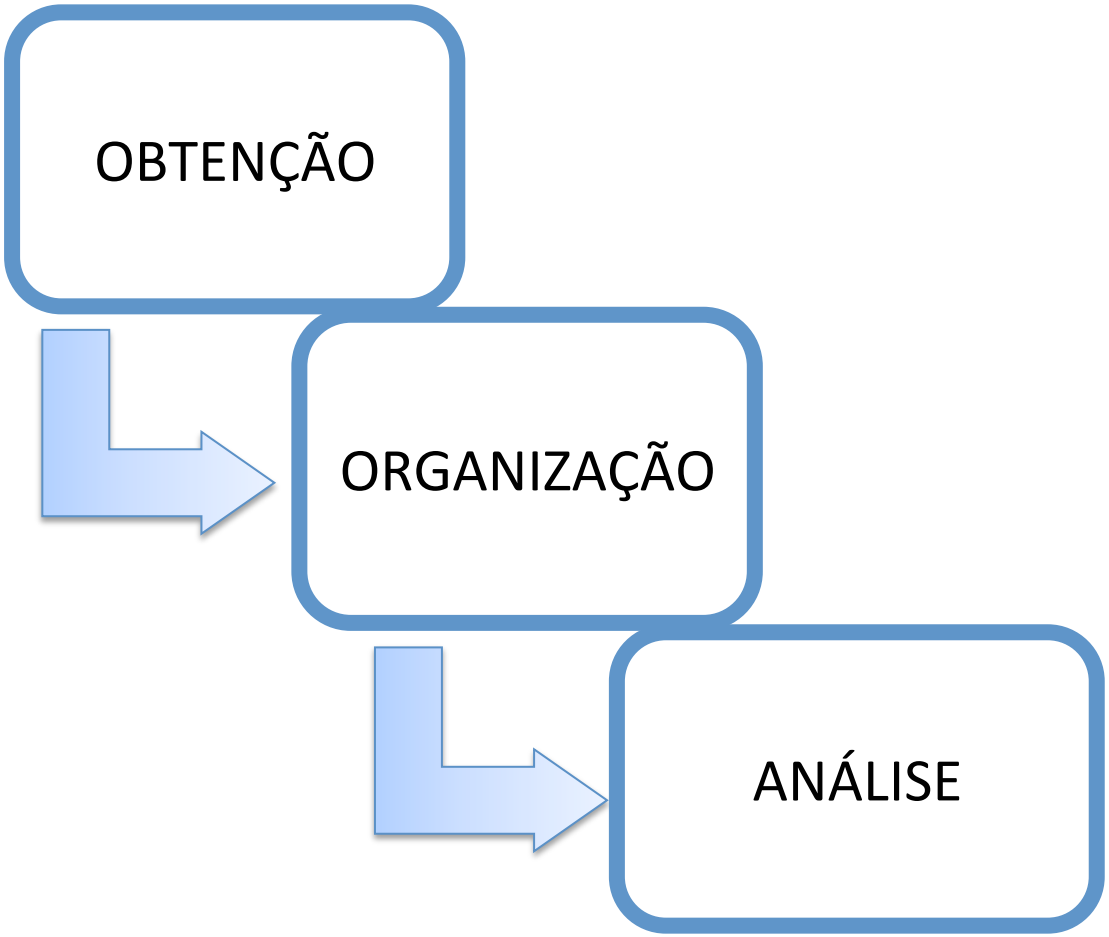
Gleici Castro Perdoná

pgleici@fmrp.usp.br

AVALIAÇÃO 1 - (20 min)

- ENTRAR NO STOA E RESPONDER 5 QUESTÕES

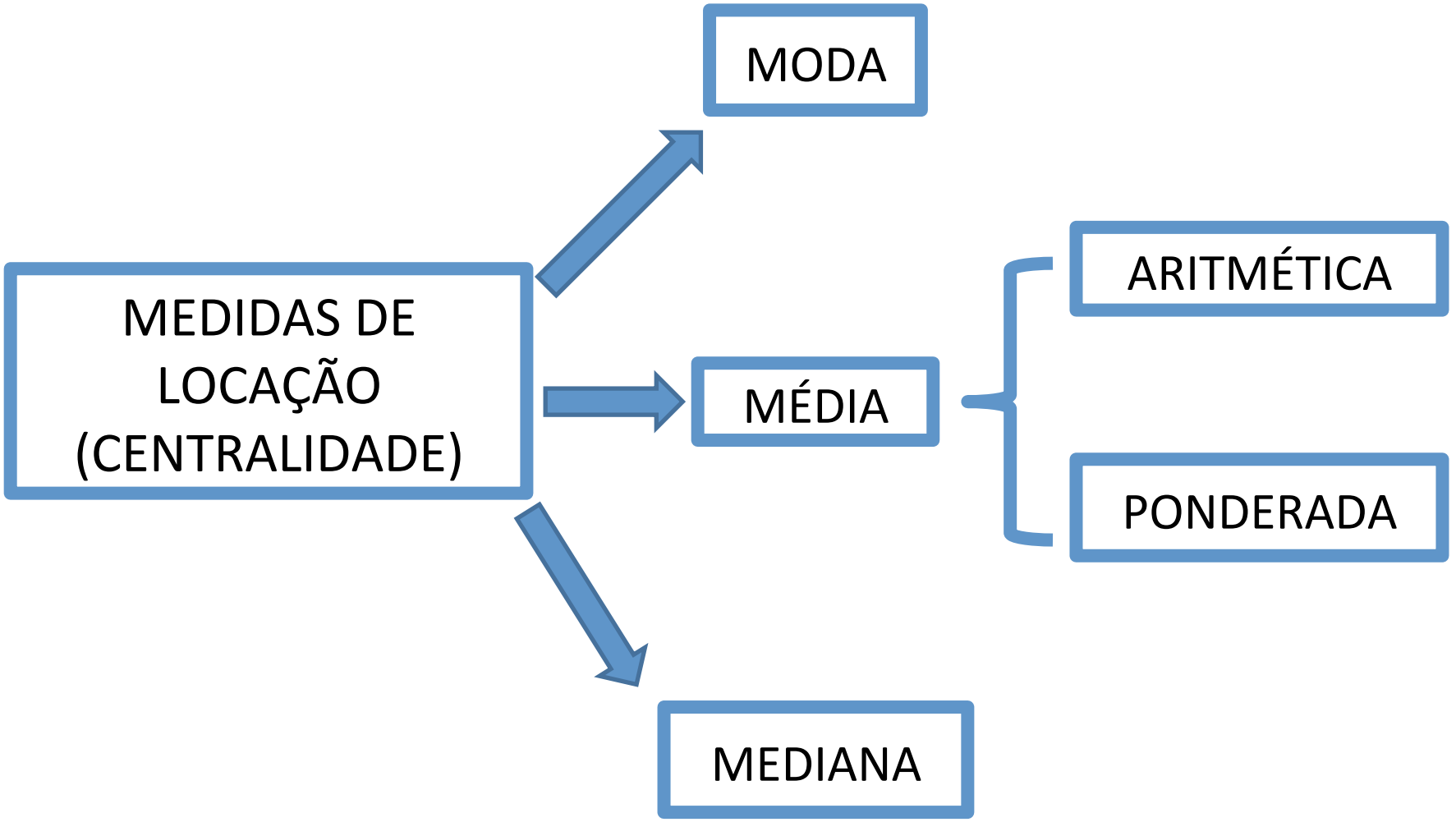
Qual objetivo das medidas descritivas?

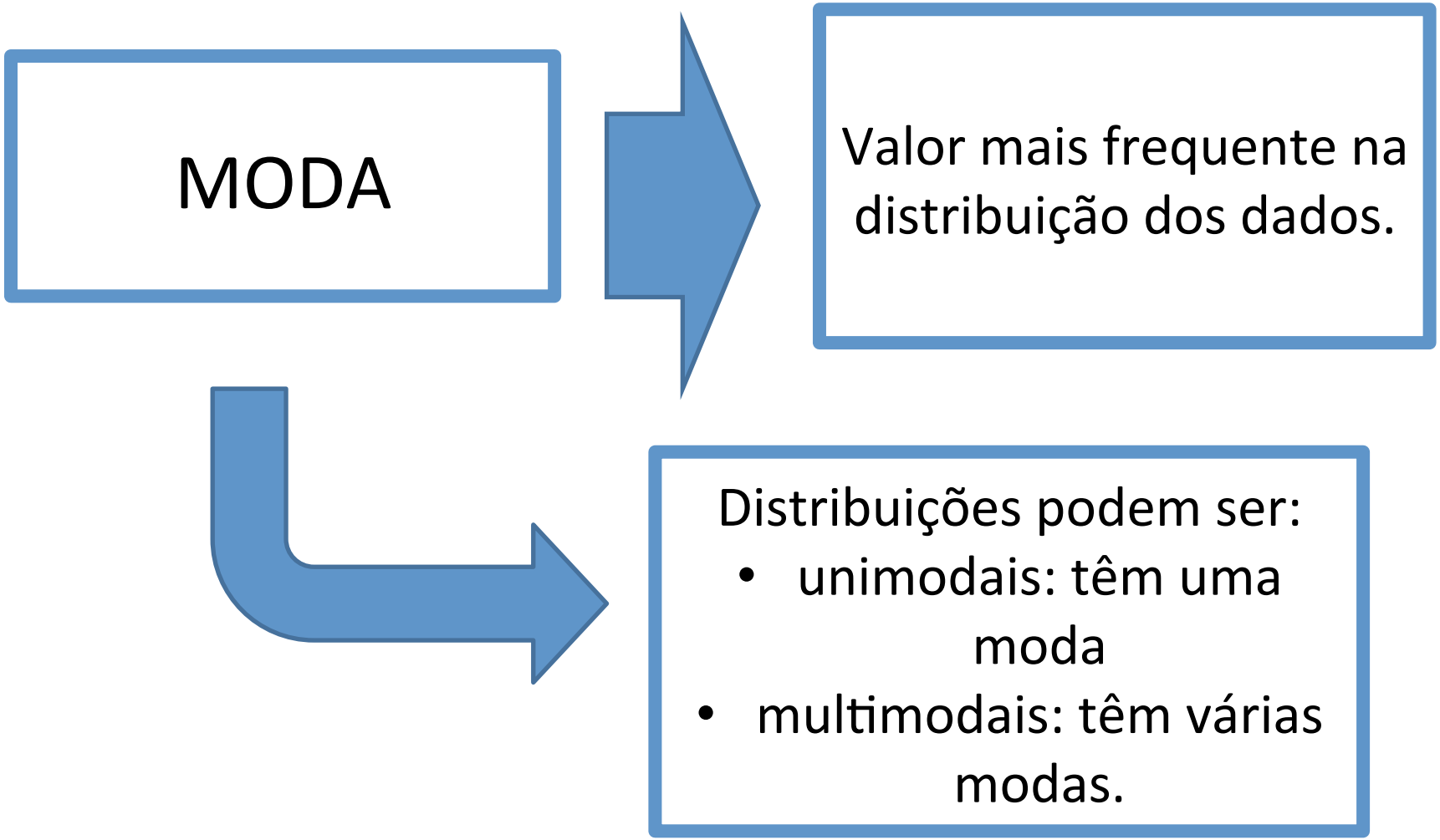


Qual objetivo das medidas descritivas?

- Resumir rapidamente um conjunto de dados.
- Podemos utilizar medidas de localização (posição) ou dispersão.
- Isto significa resumir uma conjunto de dados , por exemplo (amostra = 1000 individuos) em 2 valores.

Vamos começar com as medidas de posição




MODA
Valor mais frequente na distribuição dos dados.
Distribuições podem ser:
• unimodais: têm uma moda
• multimodais: têm várias modas.

MÉDIA

Média Aritmética

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

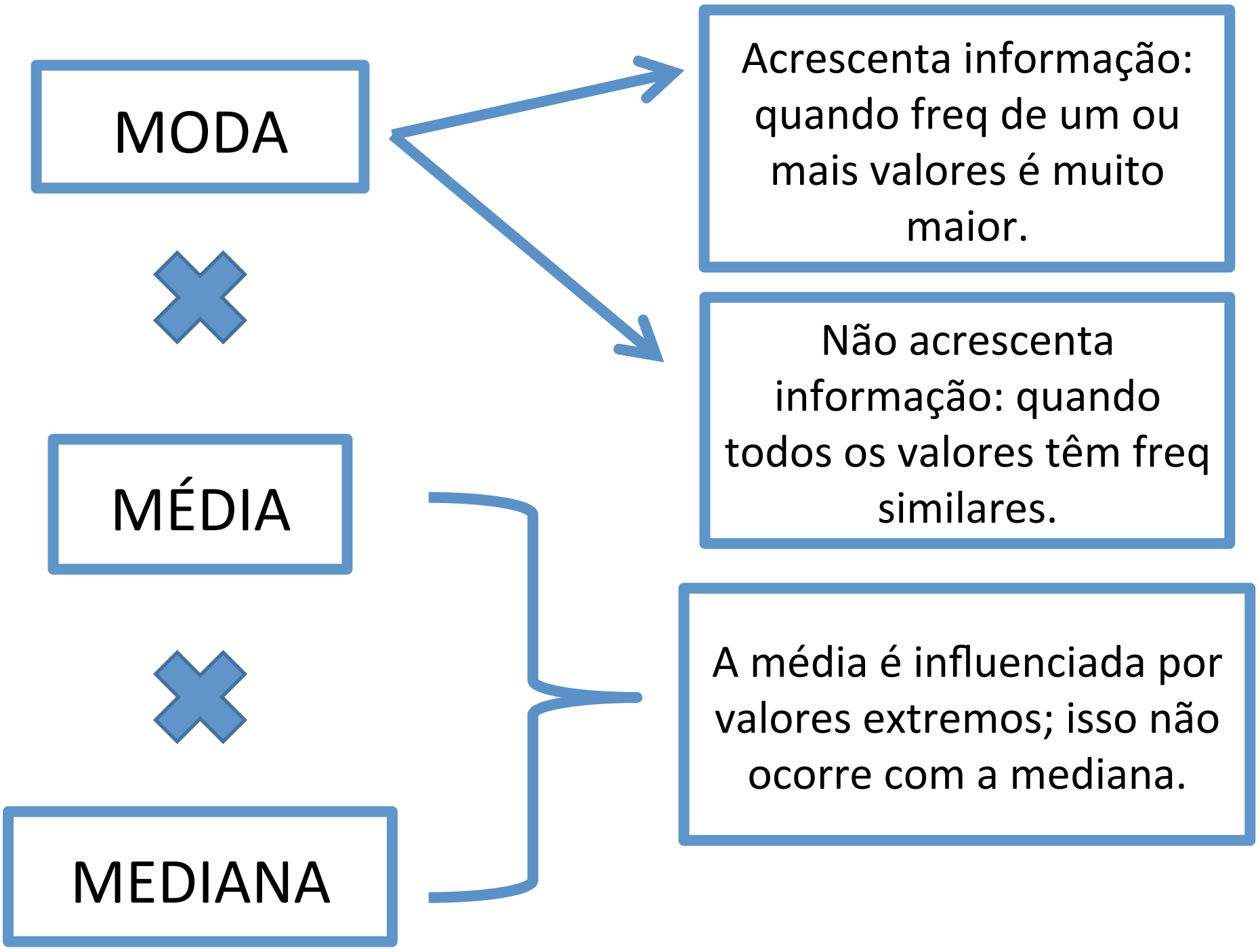
Média Ponderada

$$\bar{x}_p = \frac{f_1 x_1 + f_3 x_2 + f_3 x_3 + \cdots + f_r x_r}{\sum f_r}$$

- Amostras ímpares: mediana é o valor central.
- Amostras pares: mediana é a média dos valores centrais.

MEDIANA

Valor que separa 50% das observações à sua esquerda e 50% à sua direita quando os dados estão em ordem (crescente ou decrescente).


MODA
Acrescenta informação: quando freq de um ou mais valores é muito maior.
Não acrescenta informação: quando todos os valores têm freq similares.
MÉDIA
MEDIANA
A média é influenciada por valores extremos; isso não ocorre com a mediana.

Exemplo : moda x mediana

- Conjunto de dados I:

  2 4 6 8 10

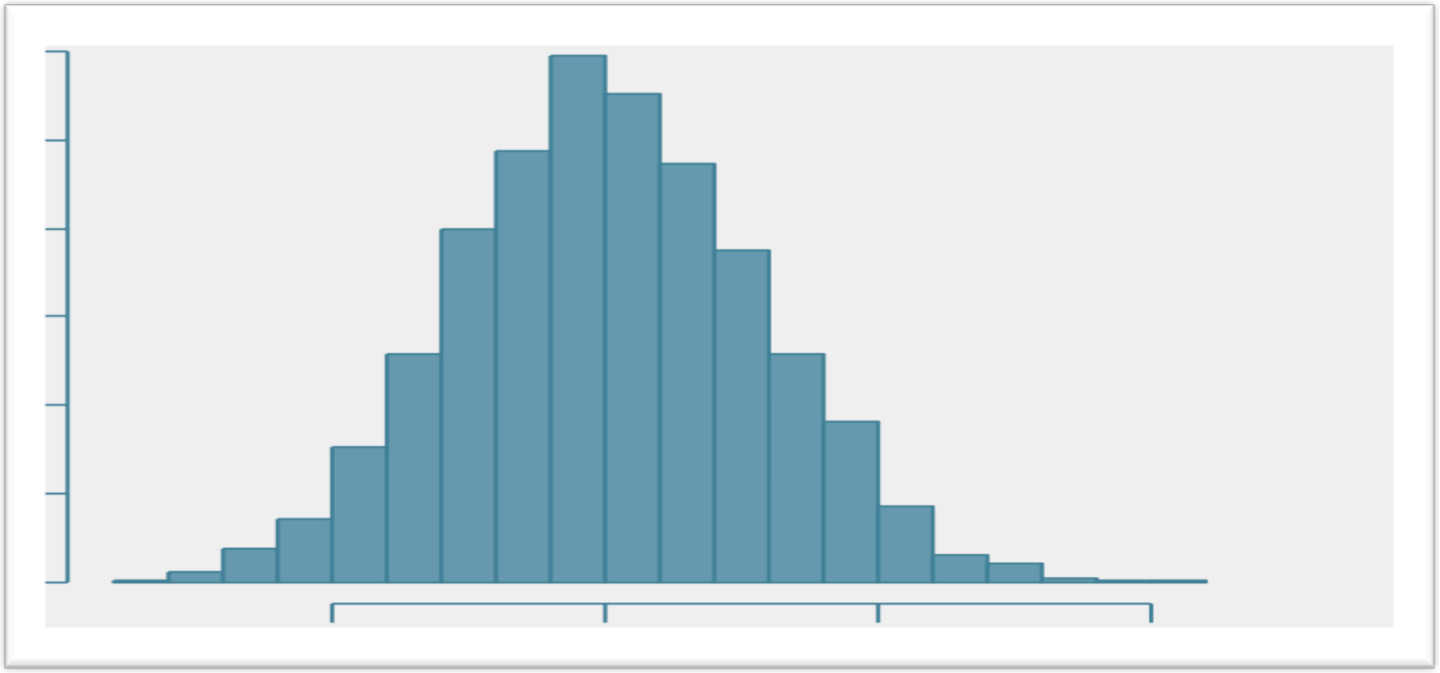
  Média = 6
  Mediana = 6

- Conjunto de dados

  2 4 6 8 100

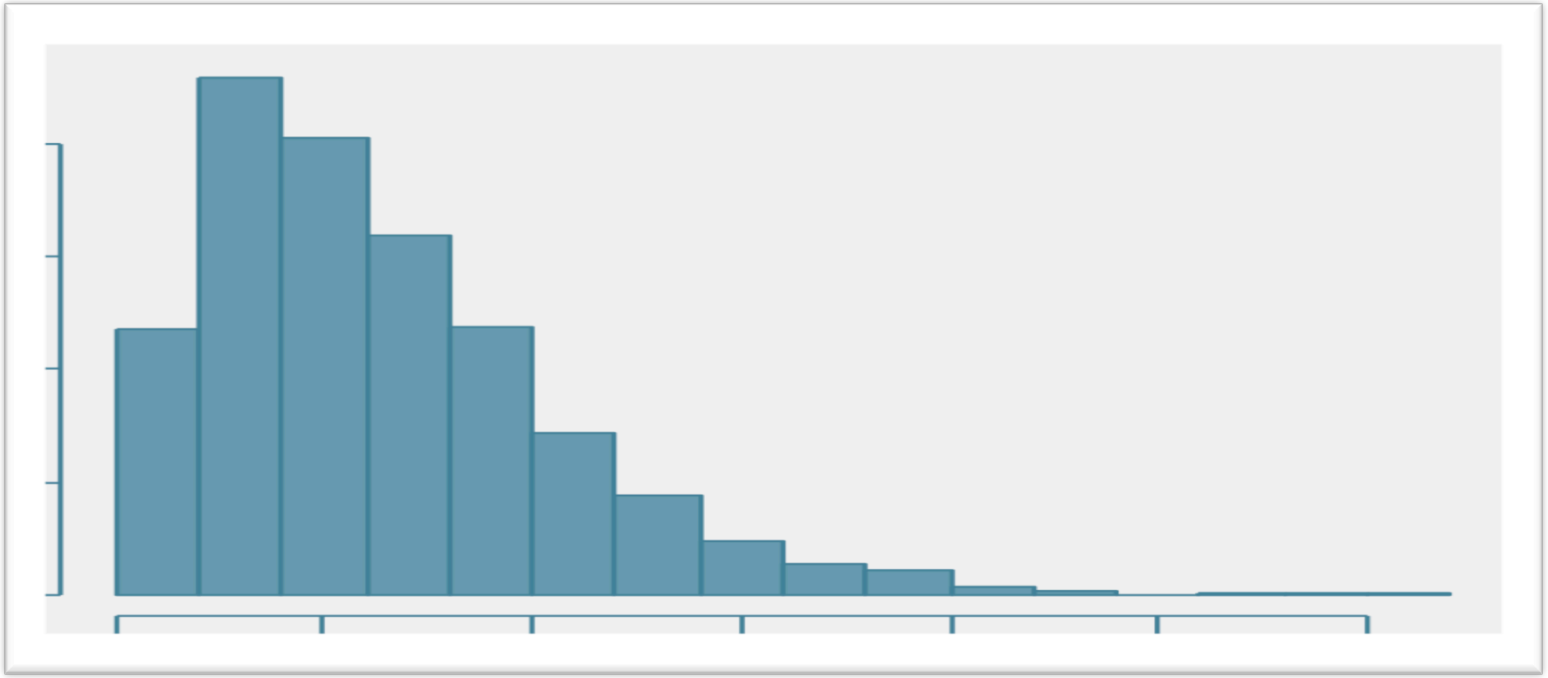
  Média = 24
  Mediana = 6

MEDIDAS DE LOCAÇÃO X TIPOS DE DISTRIBUIÇÃO



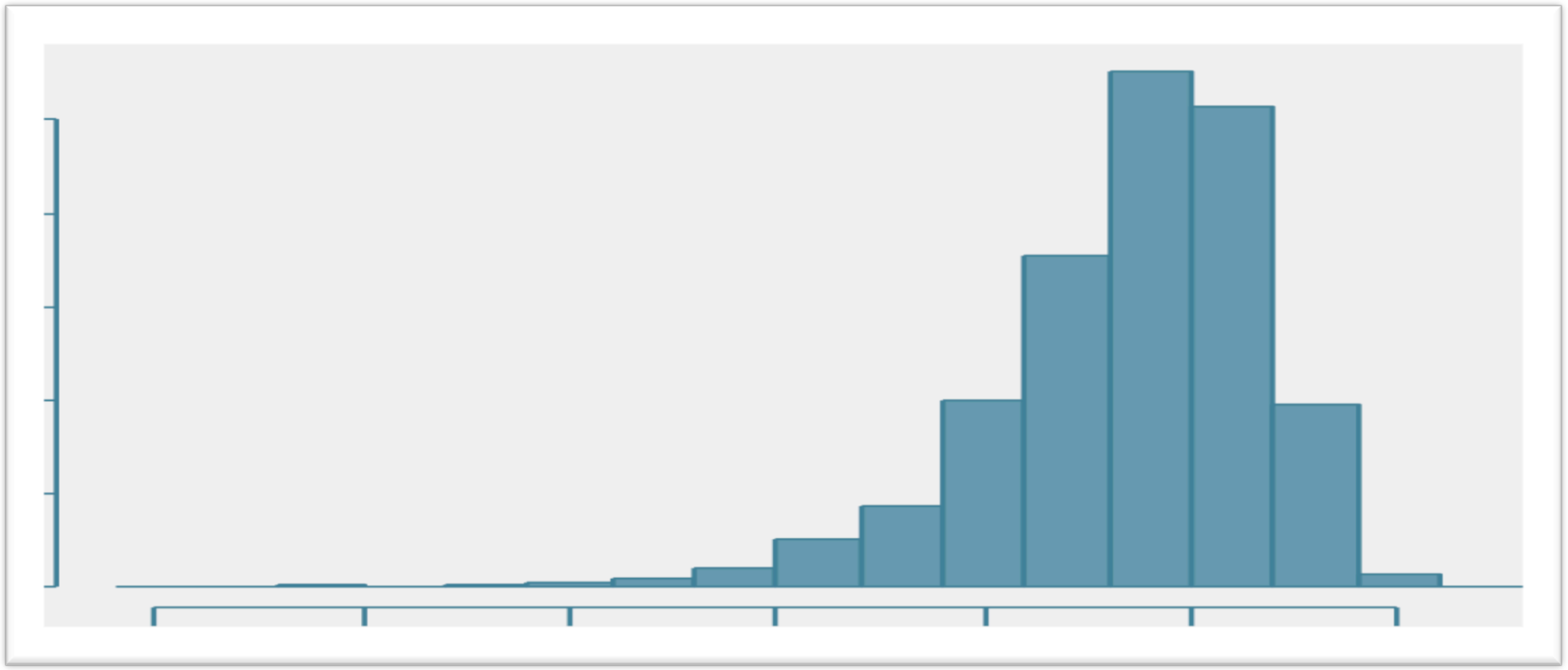
DISTRIBUIÇÃO SIMÉTRICA: Me = Mo =Md

MEDIDAS DE LOCAÇÃO X TIPOS DE DISTRIBUIÇÃO



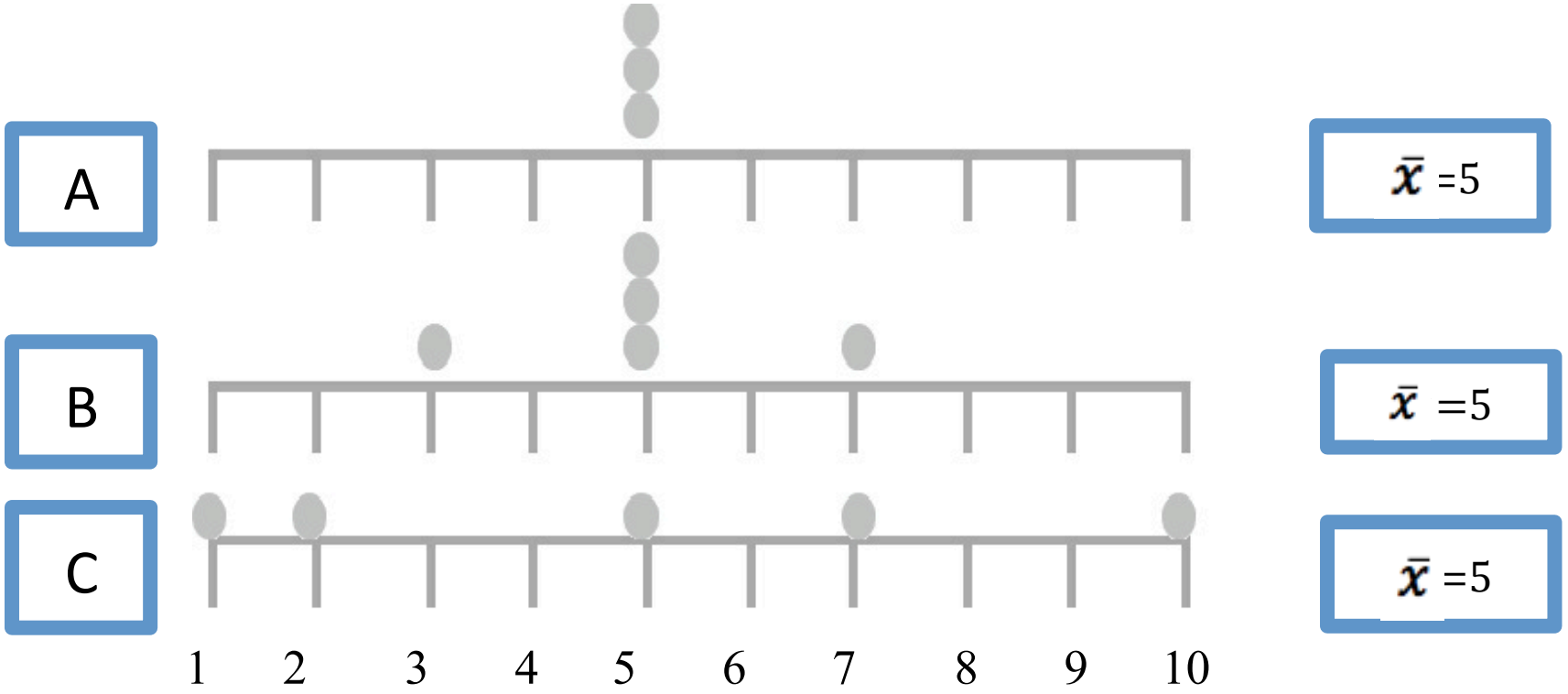
DISTRIBUIÇÃO ASSIMÉTRICA À DIREITA: Me > Md > Mo

MEDIDAS DE LOCAÇÃO X TIPOS DE DISTRIBUIÇÃO

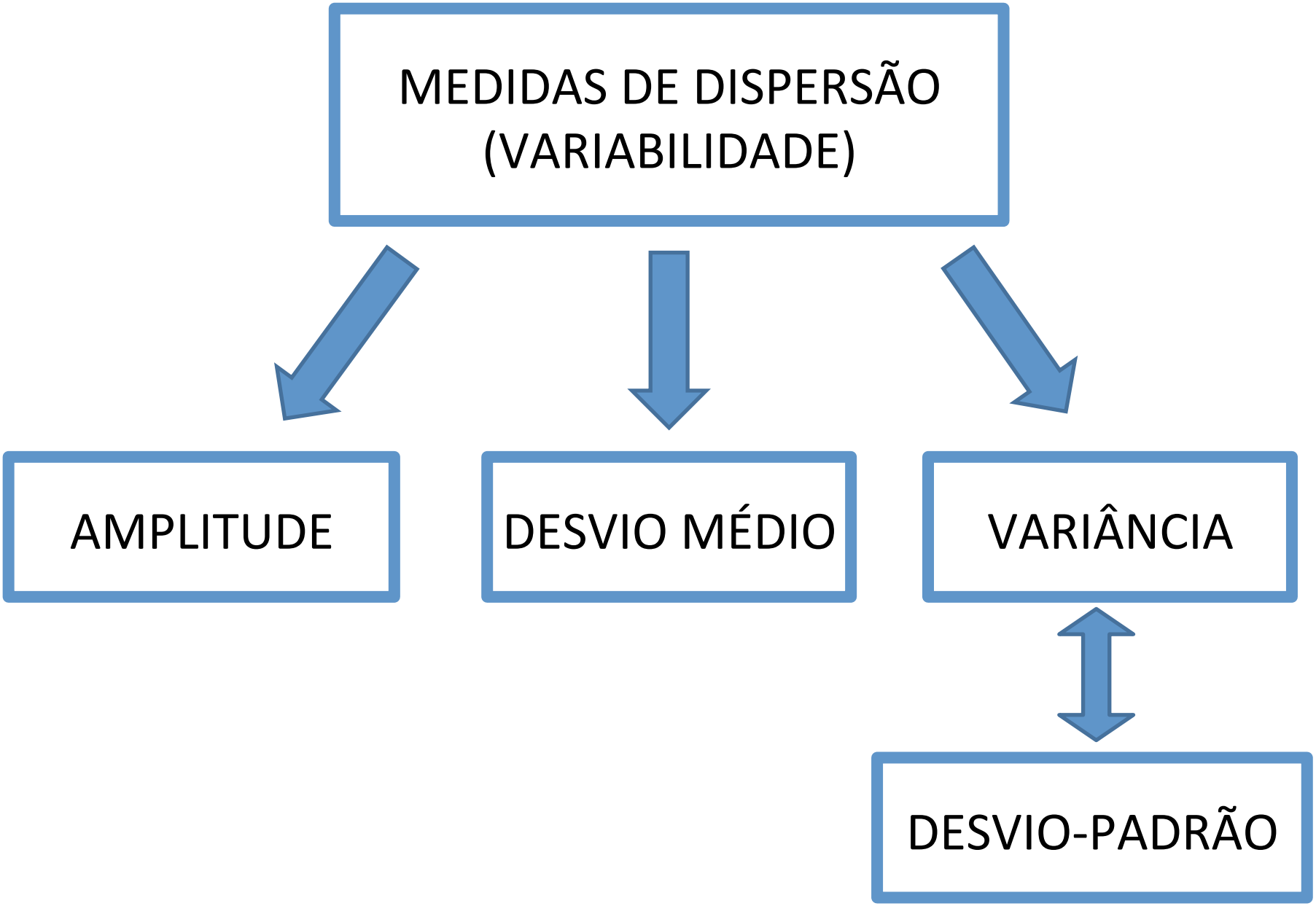


DISTRIBUIÇÃO ASSIMÉTRICA À ESQUERDA: Me < Md < Mo

Somente medidas de locação descrevem um conjunto de dados?



- A média nada informa sobre a variação dos dados.
- **SAÍDA**: criar uma medida para quantificar a variabilidade.


MEDIDAS DE DISPERSÃO
(VARIABILIDADE)
AMPLITUDE
DESVIO MÉDIO
VARIÂNCIA
DESVIO-PADRÃO

AMPLITUDE

Δ = maior valor – menor valor

- Problema: só leva em conta valores extremos.
- **SAÍDA**: obter uma medida que considere a variabilidade em relação à média.

DESVIO MÉDIO

$$Desvio = x_i - \bar{x}$$

Problema: somatória dos desvios é zero.

SAÍDA: usar o módulo ou o quadrado dos desvios.

$$\overline{DM} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

Quando se consideram os quadrados dos desvios no cálculo do DM, obtém-se a VARIÂNCIA.

- Para Amostra: $$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$
- Para População: $$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

Problema: como a unidade da variância é a da variável ao quadrado, precisa-se retornar à unidade original.

SAÍDA: cálculo do desvio-padrão

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{s^2}$$

A partir do desvio-padrão, define-se o coeficiente de variação:

$$CV = s/\bar{x}$$

Calcule as medidas descritivas (20 min)

- Numero crises mensais em asmaticos

2, 6, 7, 4, 1, 1, 3, 1, 7, 7, 3, 2, 3, 3, 1, 1, 4, 3, 1, 4, 2, 1, 4, 6

- Peso gestacional de fetos com 20 semanas(IG)

364 ,243, 408, 443, 232, 222, 368, 362, 346, 407, 342, 347, 266, 243 ,243, 254, 298, 437, 301, 240

MEDIDAS DE LOCAÇÃO
MEDIDAS DE DISPERSÃO
INTERVALOS DE VARIAÇÃO

LEI DE CHEBYCHEV

Para qualquer conjunto de dados o intervalo

$$(\overline{x} - k.\mathrm{s} \;\; ; \;\; \overline{x} + k.\mathrm{s})$$

contém, pelo menos, $\left(1 - \frac{1}{k^2}\right)$ das observações $k \geq 1$.

EXEMPLOS

Se k = 2	$(\bar{x} - 2s; \bar{x} + 2s)$	CONTÉM PELO MENOS	75% DAS AMOSTRAS
Se k = 3	$(\bar{x} - 3s; \bar{x} + 3s)$		89%
Se k = 4	$(\bar{x} - 4s; \bar{x} + 4s)$		94%

APLICAÇÃO DA LEI DE CHEBYCHEV

Uma indústria farmacêutica faz medicamentos contendo, em média, 510g de vitamina C, com desvio-padrão de 3g.

Pela Lei de CHEBYCHEV, sabe-se que pelo menos 75% dos comprimidos têm entre 504 g e 516 g de vitamina C.

REGRA EMPÍRICA BASEADA NA LEI DE CHEBYCHEV

Se a distribuição de frequencia de uma variável é simética e unimodal

67%	DOS DADOS ESTÃO DENTRO DO INTERVALO	$(\overline{x} - \mathrm{s}\ ;\ \overline{x} + \mathrm{s})$
95%		$(\overline{x} - 2\mathrm{s};\ \overline{x} + 2\mathrm{s})$
99%		$(\overline{x} - 3\mathrm{s};\ \overline{x} + 3\mathrm{s})$

ASSIMETRIA

Afastamento em relação ao centro

$$As = \frac{\bar{x} - Mo}{s}$$

CURTOSE

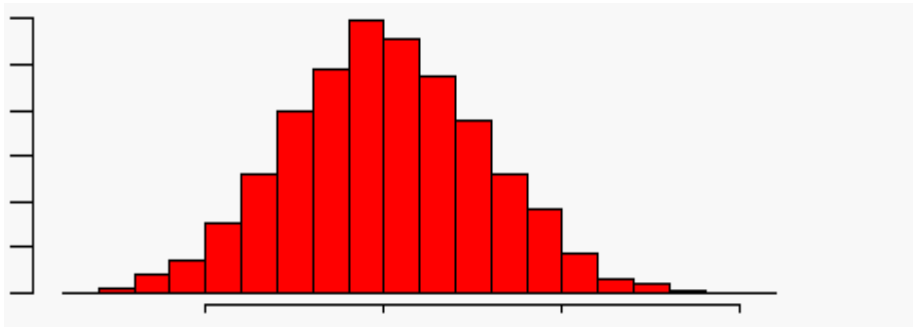
Achatamento da distribuição

$$C = \frac{Q_3 - Q_1}{2(P_{90} - P_{10})}$$

VISUALIZANDO A ASSIMETRIA

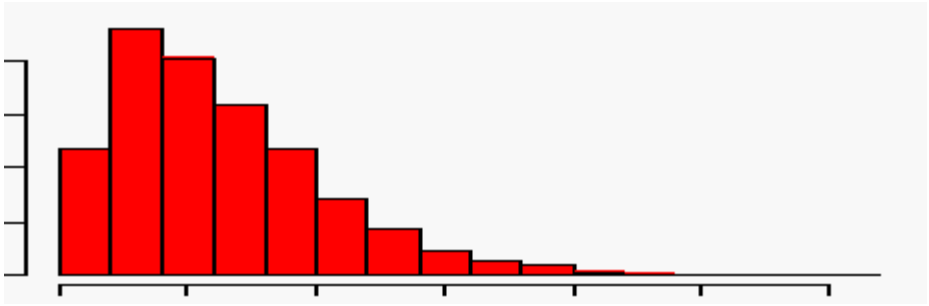
As = 0

(distribuição simétrica)



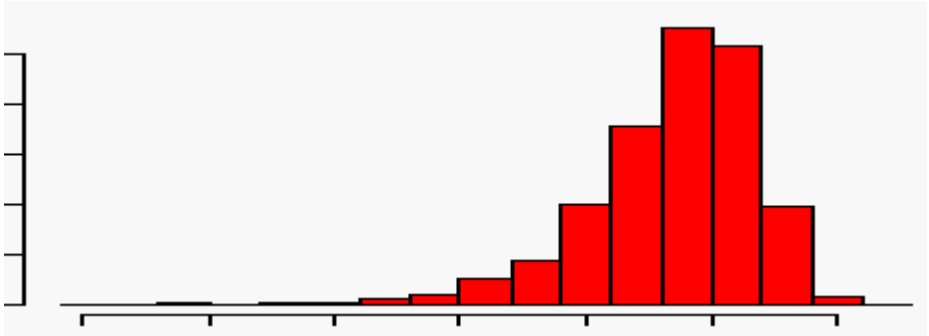
As > 0

(distribuição assimétrica à direita)



As < 0

(distribuição assimétrica à esquerda)

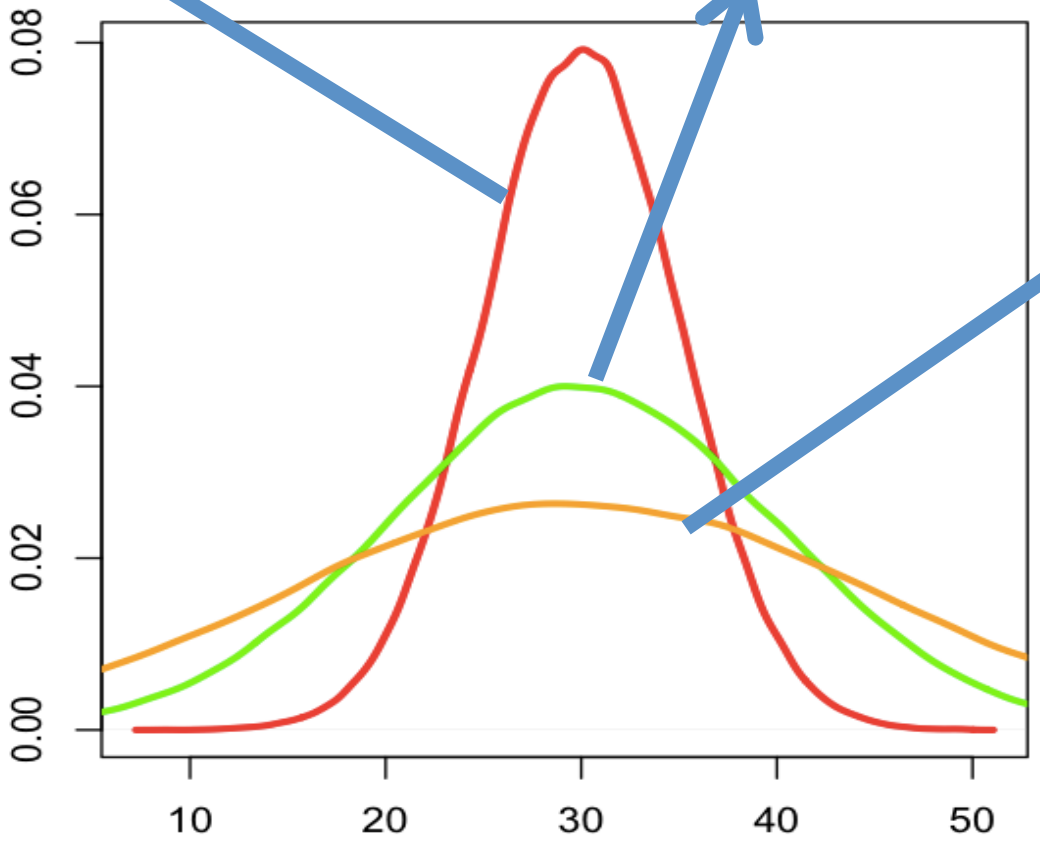


VISUALIZANDO A CURTOSE

C < 0,263
(distribuição
leptocúrtica)

C = 0,263
(distribuição
mesocúrtica)

C > 0,263
(distribuição
platicúrtica)



COMPLEMENTO : MÉDIA GEOMÉTRICA

MÉDIA GEOMÉTRICA

$$\bar{X}g = \sqrt[n]{X1 \,.\, X2 \,.\, \ldots \,.\, Xn}$$

$$\bar{X}g = (X1 \,.\, X2. \,\ldots\, .\, Xn)^{1/n}$$

MÉDIA GEOMÉTRICA PONDERADA

$$\bar{X}gp = \sqrt[\sum fi]{X1^{f1} . X2^{f2} . \ldots . Xi^{fi}}$$

$$\bar{X}gp = (X1^{f1} . X2^{f2} . \ldots . Xi^{fi})^{\frac{1}{\sum fi}}$$

REVISÃO

1. Medidas descritivas

- Medidas de locação – moda, média (aritmética e ponderada), mediana – e sua relação com a simetria da distribuição;
- Medidas de dispersão – amplitude, desvio médio, variância e desvio-padrão;
- Intervalo de variação: combinação entre medidas de locação e dispersão, lei de CHEBYCHEV.

2. Visualizando a variação dos dados

- Quantis e percentis;
- Boxplot;
- Medida de assimetria;
- Medida de curtose.

Encontre o intervalo empirico (5 min)

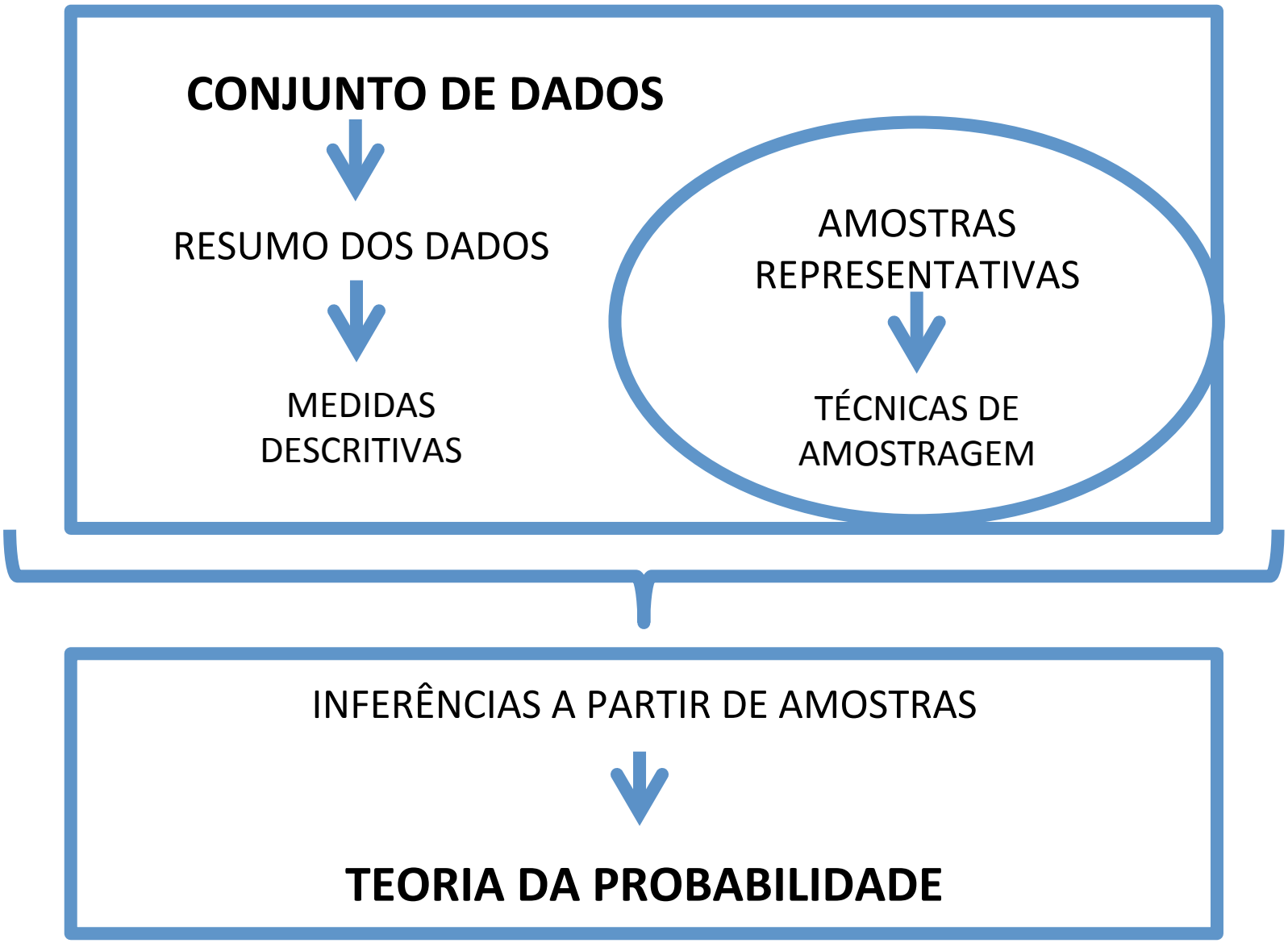
- Numero crises mensais em asmaticos

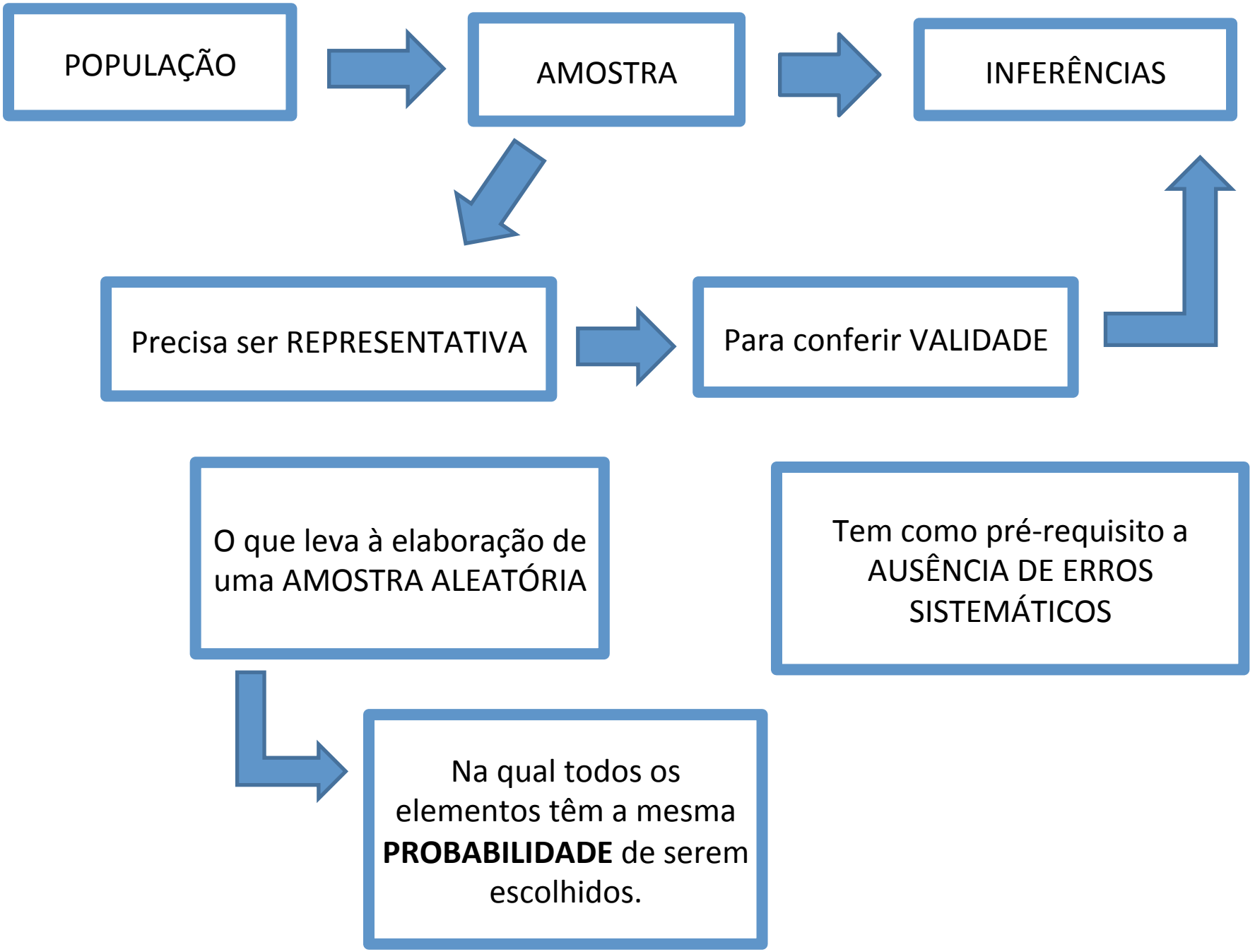
2, 6, 7, 4, 1, 1, 3, 1, 7, 7, 3, 2, 3, 3, 1, 1, 4, 3, 1, 4, 2, 1, 4, 6

- Peso gestacional de fetos com 20 semanas(IG)

364 ,243, 408, 443, 232, 222, 368, 362, 346, 407, 342, 347, 266, 243 ,243, 254, 298, 437, 301, 240

Probabilidade e Teste de Diagnóstico




POPULAÇÃO
AMOSTRA
INFERÊNCIAS
Precisa ser REPRESENTATIVA
Para conferir VALIDADE
O que leva à elaboração de uma AMOSTRA ALEATÓRIA
Tem como pré-requisito a AUSÊNCIA DE ERROS SISTEMÁTICOS
Na qual todos os elementos têm a mesma **PROBABILIDADE** de serem escolhidos.

DEFINIÇÃO FREQUENTISTA DE PROBABILIDADE

$$\text{Probabilidade} = \frac{\text{nº de vezes que o evento ocorre}}{\text{nº total de ensaios}}$$

DEFINIÇÃO MATEMÁTICA DE PROBABILIDADE

Uma função P(X) é denominada probabilidade se satisfaz as seguintes condições:

1) $0 \leq P(A) \leq 1, \ \forall A \subset \Omega;$

2) $P(\Omega) = 1;$

3) $P(\bigcup_{j=1}^{n} A_j) = \sum_{j=1}^{n} P(A_j)$

UNIÃO DE PROBABILIDADES

- Para eventos mutuamente excludentes:

$$P(A \cup B) = P(A) + P(B)$$

- Para eventos não-mutuamente excludentes:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

INTERSECÇÃO DE PROBABILIDADES

- Para eventos independentes:

$$P(A \cap B) = P(A).P(B)$$

A e B são independentes se a informação da ocorrência ou não de B não altera a probabilidade da ocorrência de A.

- Para eventos não-independentes (probabilidade condicional):

$$P(A|B) = P(A \cap B)/P(B) \rightarrow P(A \cap B) = PAB\ .P(B)$$

EXEMPLOS

Distribuição do tipo sanguíneo por sexo

	Probabilidades		
Tipo sanguíneo	Homens	Mulheres	Total
O	0,210	0,210	0,420
A	0,215	0,215	0,430
B	0,055	0,055	0,110
AB	0,020	0,020	0,040
Total	0,500	0,500	1,000

Fonte: Dawson, B. e Trapp, R. G., "Bioestatística Básica e Clínica", terceira edição.

Questões:

1) P(tipo sanguíneo A)?
2) P(tipo O ou A)?
3) P(homem e sangue tipo O)?
4) P(homem ou tipo O)?

Respostas:

1) 0,43
2) 0,85
3) 0,50 . 0,42 = 0,21
4) 0,50 + 0,42 – 0,21 = 0,71

Local da infecção dos casos de sorogrupo B, Oregon, 1996

		Intervalo de tempo			
		Pré-epidemia 1987 – 1992	Início da epidemia 1993 – 1994	Epidemia recente 1995 – 1996	Total
Local da infecção	Sepse	66	45	40	151
	Meningite	39	32	39	110
	Ambos	39	32	34	105
	Desconhecido	6	8	28	42
Total		150	117	141	408

Fonte: Dawson, B. e Trapp, R. G., "Bioestatística Básica e Clínica", terceira edição.

Questões e respostas

1) P(apenas meningite)? 110/408 = 0,27

2) P(complemento de ter meningite sabidamente conhecida)?

P(meningite desconhecida) = P(sepse ou desconhecido) = (151 + 42)/408 = 0,473
ou 1 – P(meningite conhecida) = 1 – (110+105)/408 = 0,473

Local da infecção dos casos de sorogrupo B, Oregon, 1996

		Intervalo de tempo			
		Pré-epidemia 1987 – 1992	Início da epidemia 1993 – 1994	Epidemia recente 1995 – 1996	Total
Local da infecção	Sepse	66	45	40	151
	Meningite	39	32	39	110
	Ambos	39	32	34	105
	Desconhecido	6	8	28	42
Total		150	117	141	408

Fonte: Dawson, B. e Trapp, R. G., "Bioestatística Básica e Clínica", terceira edição.

3) P(local desconhecido e pré-epidêmico)? E o produto das probabilidades marginais?

6/408 = 0,015

P(local desconhecido) . P(pré-epidêmico) = 42/408 . 150/408 = 0,103 . 0,368 = 0,038

4) P(meningite conhecida|epidemia recente)?

(39 + 34)/141 = 0,518

5) P(meningite conhecida e epidemia recente)?

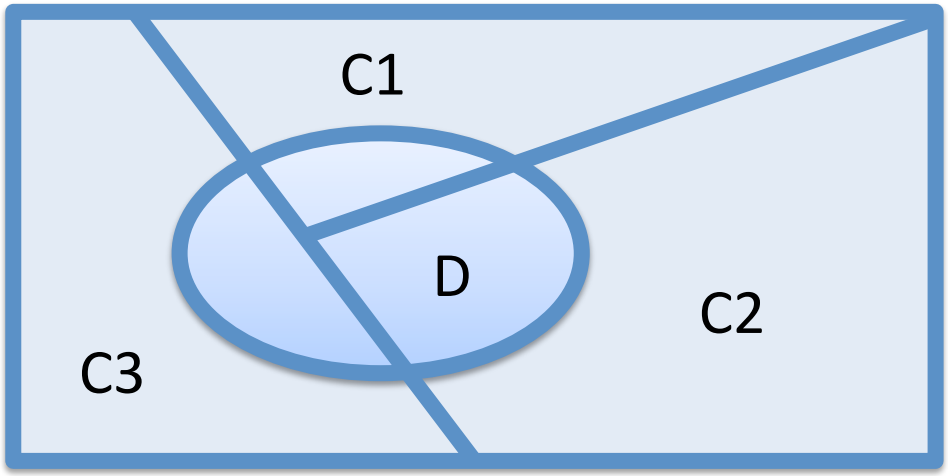
P(meningite conhecida|epidemia recente) . P(epidemia recente) =
(39 + 34)/141 . 141/408 = 0,518 . 0,346 = 0,179

Intervalo (15 minutos)

Teorema de Bayes

- Thomas Bayes foi um reverendo presbiteriano que viveu no início do século 18 (1701-1761) na Inglaterra.
- Dois anos após sua morte, um amigo, encontrou entre os papéis do reverendo, com o 'Ensaio buscando resolver um problema na doutrina das probabilidades'. Nesse artigo estava a demonstração do famoso teorema de Bayes.

TEOREMA DE BAYES



Sendo C1, C2 e C3 eventos que formam uma partição de Ω, cujas probabilidades são conhecidas, e D um evento conhecido, então:

$$P(C_j \mid D) = \frac{P(D \mid C_j)P(C_j)}{\sum_{i=1}^{k} P(D \mid C_i)P(C_i)}, j = 1,2,..k$$

EXEMPLO

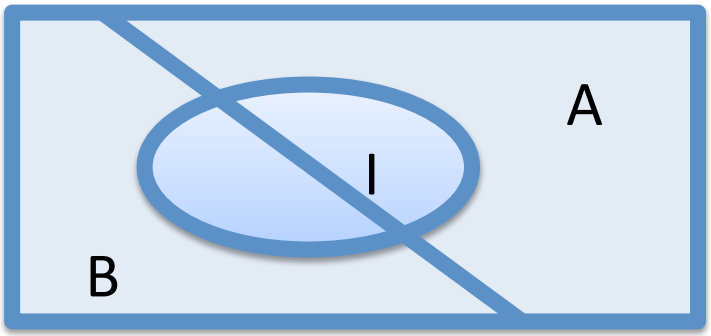
Dois tipos de vacina foram aplicados em uma população, de tal forma que 60% das pessoas receberam vacina do tipo A e as 40% restantes receberam vacina do tipo B. Sabendo que a vacina do tipo A fornece 70% de imunização, e a B fornece 80%, determine a probabilidade de que uma pessoa, escolhida ao acaso:

I) esteja imunizada, dado que foi vacinada com por A;

II) esteja imunizada;

III) tenha sido vacinada por A, dado que não esteja imunizada.

	Imunização		
Tipo	I	NI	Total
A	0,7 . 0,6 = 0,42	0,3 . 0,6 = 0,18	0,6
B	0,8 . 0,4 = 0,32	0,2 . 0,4 = 0,08	0,4
Total	0,74	0,26	1,0



I) esteja imunizada, dado que foi vacinada com por A;
II) esteja imunizada;
III) tenha sido vacinada por A, dado que não esteja imunizada.

I) $P(I|A) = 70\%$ (dado do enunciado).

II) $P(I) = P(A \cap I) + P(B \cap I) =$
$P(I|A) \cdot P(A) + P(I|B) \cdot P(B) =$
$0,7 \cdot 0,6 + 0,8 \cdot 0,4 = 0,42 + 0,32 = 0,74 = 74\%$

III) $P(A|NI) = P(NI \cap A) / [P(NI \cap A) + P(NI \cap B)] =$
$[P(NI|A) \cdot P(A)] / [P(NI|A) \cdot P(A) + P(NI|B) \cdot P(B)] =$
$(0,3 \cdot 0,6) / (0,3 \cdot 0,6 + 0,2 \cdot 0,4) = 0,18 / (0,18 + 0,08)$
$= 0,18 / 0,26 = 0,69 = 69\%$

	Imunização		
Tipo	I	NI	Total
A	0,7 . 0,6 = 0,42	0,3 . 0,6 = 0,18	0,6
B	0,8 . 0,4 = 0,32	0,2 . 0,4 = 0,08	0,4
Total	0,74	0,26	1,0

O Que é Teste de diagnóstico

- Denomina-se teste de diagnóstico exames de laboratórios entretanto um conjunto de sinais e sintomas também pode ser visto como TD.
- Portanto é um instrumento capaz de diagnosticar a doença com determinada precisão. Para cada teste diagnóstico existe um valor de referência.
- A prática médica moderna utiliza-se das leis de probabilidade como um importante auxiliar na interpretação e avaliação dos TD

Possíveis resultados de um Teste de Diagnóstico para identificar uma doença

		DOENÇA Padrão Ouro		
		PRESENTE	AUSENTE	Total
TESTE	POSITIVO	a	b	a+b
		V +	F+	
	NEGATIVO	c	d	c+d
		F-	V-	
	Total	a+c	b+d	a+b+c+d

Quantidade de interesse nos testes de diagnostico

- Sensibilidade
- Especificidade
- Valor Preditivo Positivo
- Valor Preditivo Negativo
- Prevalencia
- Razão de Verossimilhança positiva e negativa
- Curva ROC

Sensibilidade e Especificidade

	D	ND
T+	a	b
T-	c	d

$$\text{Sensibilidade} = \frac{a}{a + c}$$

$$\text{Especificidade} = \frac{d}{b + d}$$

- Sensibilidade: Probabilidade de um teste ser positivo, dado que existe a doença.
- Especificidade: Probabilidade de um teste ser negativo, dado que não existe a doença.

Valor Preditivo Positivo e Negativo

	D	ND
T+	a	b
T-	c	d

$$VPP + = \frac{a}{a + b}$$

$$VPN - = \frac{d}{c + d}$$

- Valor Preditivo Positivo: Probabilidade de existir a doença, dado que o teste foi positivo.
- Valor Preditivo Negativo: Probabilidade de não existir a doença, dado que o teste foi negativo.

Razão de Verossimilhança Positiva e Negativa

	D	ND
T +	a	b
T -	c	d

RVP = a prob do resultado + nos doentes / a prob do resultado + nos sadios

RVP = a prob do resultado - nos doentes / a prob do resultado - nos sadios

$$RVP = \frac{\frac{a}{a+c}}{\frac{b}{b+d}}$$

$$RVN = \frac{\frac{c}{a+c}}{\frac{d}{b+d}}$$

Razão de Verossimilhança Positiva:

$$\frac{\text{Sensibilidade}}{\text{1-Especificidade}}$$

Razão de Verossimilhança Negativa:

$$\frac{\text{1-Sensibilidade}}{\text{Especificidade}}$$

Razão de Verossimilhança Positiva:

- Expressa quantas vezes é mais provável encontrar um resultado positivo em pessoas doentes quando comparado com os não doentes. RVP elevada melhor é o desempenho do teste

Razão de Verossimilhança Negativa:

- Expressa quantas vezes é mais provável encontrar um resultado negativo em pessoas doentes quando comparado com pessoas não doentes. RVN baixa/menor melhor é o desempenho do teste.

Intervalo (5 minutos)

Como uso isso?

Um teste de laboratório para detectar microalbuminúria na manhã em amostras de urina. Uma pessoa foi considerada como tendo microalbuminúria se a excreção de albumina urinária determinada por nefelometria excedesse 2,3 mg / l.

	Padrao Ouro		
microalbuminúria	Doente	Nao doente	Total
Positivo	130	357	487
Negativo	23	2017	2040
Total	153	0,26	2527

KJ van Stralen et al.: Diagnostic methods I
Kidney International (2009) 75, 1257–1263

Encontre as medidas

- Sensibilidade
- Especificidade
- Valor Preditivo Positivo
- Valor Preditivo Negativo
- Prevalência
- Razão de Verossimilhança positiva e negativa
- Curva ROC

RESPOSTAS

- Sensibilidade 85%
- Especificidade 85%
- Valor Preditivo Positivo 27%
- Valor Preditivo Negativo 99%
- Prevalência
- Razão de Verossimilhança positiva e negativa
- 5.7 e 0.18
- Percentual de concordância (130+2017)/2527 = 85%
- Prevalência – Probabilidade da doença (130+23)/2527

Algumas interpretações

- Neste estudo avaliamos a precisão diagnóstica do teste de microalbuminúria
- Dos 2374 indivíduos sem microalbuminúria, 2017 tiveram um teste de urina negativo, resultando em uma especificidade de 85%.
- Em pacientes com microalbuminúria, um teste positivo é encontrado 5,7 vezes mais frequentemente que em pacientes sem microalbuminúria.
- Mas…se sou o paciente posso estar mais interessado no VPP ou VPN.
- Isto é, na probabilidade de estar ou não doente dado o resultado.

Temos mais

- Pré teste em termos de odds

Prevalencia/(1-prevalencia) = 0.064

Representa a chance antes do teste

- Pos teste

Depois que deu postivo = pré x RVP= 0.064x5.7=0.37

Depois que deu negativo = pré x RVN=0.064x0.18 = 0.012

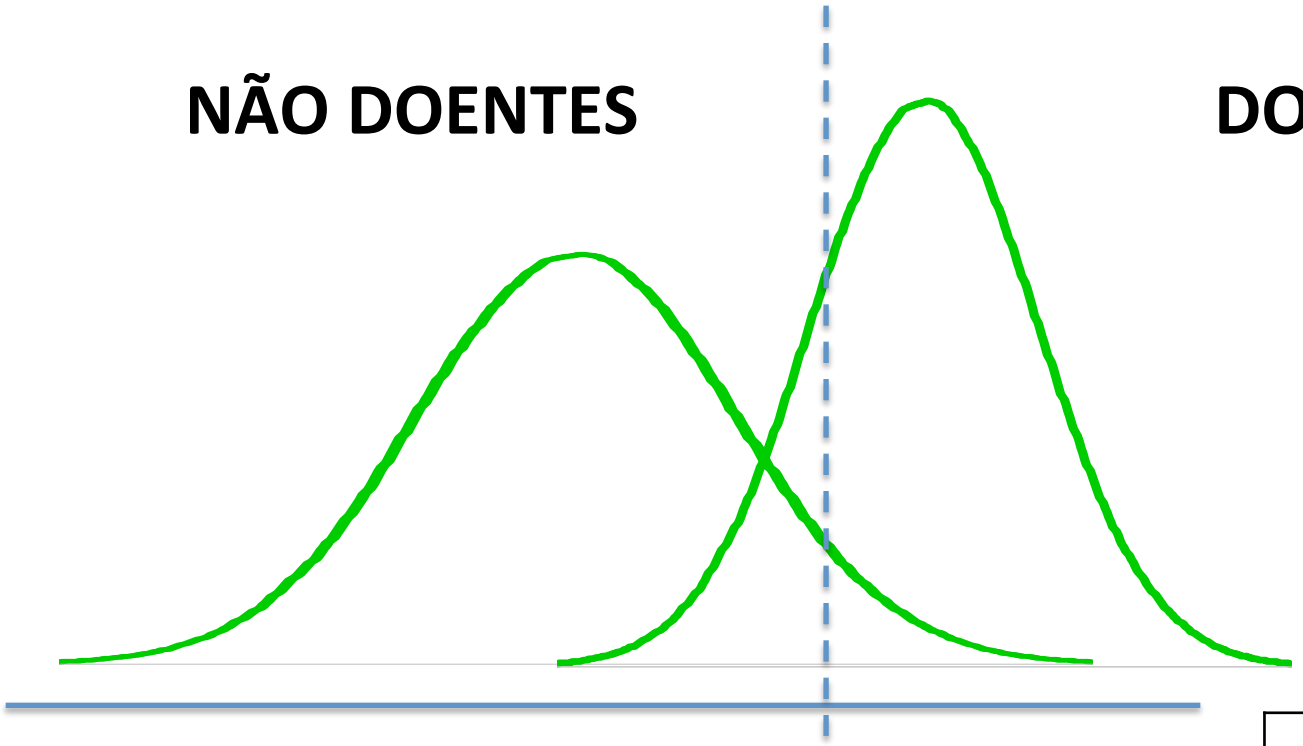
✓ Chance da doença depois do resultado do teste - Teorema de Bayes

Em termos de probabilidade

Chance /(1+chance) 0.37/(1+0.37)=26.9% (+) e

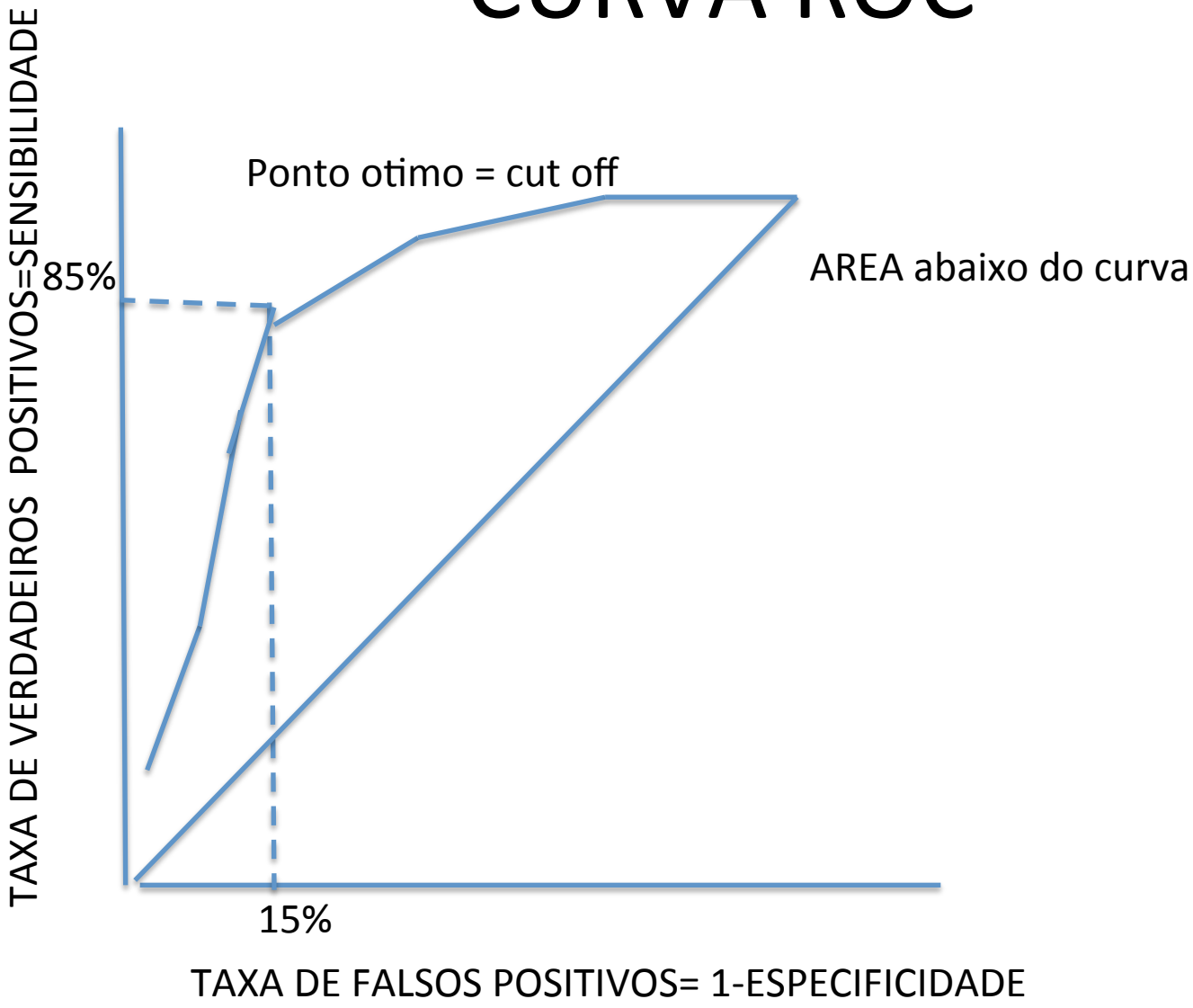
0.012/(1+0.012)= 1.14%(-)

UM PONTO NA CURVA ROC



	Padrao Ouro		
microalbuminúria	Doente	Nao doente	Total
Positivo	130	357	487
Negativo	23	2017	2040
Total	153	0,26	2527

CURVA ROC


TAXA DE VERDADEIROS POSITIVOS=SENSIBILIDADE
Ponto otimo = cut off
85%
AREA abaixo do curva
15%
TAXA DE FALSOS POSITIVOS= 1-ESPECIFICIDADE

O QUE É ?

Receiver Operating Characteristic

A CURVA ROC é uma medida da capacidade discriminativa de um teste, isto é, a capacidade de um teste classificar corretamente aqueles com e sem a doença.

BIBLIOGRAFIA RECOMENDADA

BUSSAB, W.O.; MORETTIN, P**. Estatística básica.** 4 ed. São Paulo, Atual, 1987.

PAGANO, M. e GAUVREAU, K. Princípios de Bioestatística - Tradução da 2ª Edição Norte Americana, Pioneira Thonpson Learning, São Paulo, SP,2004.

MEDRONHO R; CARVALHO DM; BlOCH KV; LUIZ RR; WERNECK GL. Epidemiologia.
Atheneu, 2 ed. São Paulo, 2008

ROSNER, B.Fundamentos de bioestatistica**.** 8ª Edição Norte Americana, Cengage Learning, 2016.

Aula usando R

- Precisa baixar uma biblioteca epiR
- Carregar um o pacote

  epi.2by2

```
dat <- as.table(matrix(c(130,357,23,2017), nrow = 2,
byrow = TRUE))
 colnames(dat) <- c("Dis+","Dis-")
 rownames(dat) <- c("Test+","Test-")
 rval <- epi.tests(dat, conf.level = 0.95)
 print(rval)
summary(rval)
```

PROXIMA AULA

- Distribuições de probabilidade
- Pbinom()
- Qbinom()

- Rnorm()
- Pnorm()
- Qnorm()

OBRIGADA