

MODELAGEM E PREPARAÇÃO DE DADOS PARA APRENDIZADO DE MÁQUINA: Preparação de dados

Professor:
Luis E. Zárate

Tipos conceituais de variáveis

- **Variável Escalar**

- Valor da medida+escala da medida (Exemplo, velocidade 50 m/s). Podem ser Quantitativas ou Qualitativas.

- **Variável Vectorial**

- Precisa de mais de um valor para defini-la (Exemplo, vetor de localização [10°latitude, 30°longitude]). Cada valor da variável possui sentido individual e em conjunto.



Tipos de variáveis

- **Variáveis Quantitativas**

- **Variáveis discretas:** Podem assumir apenas um número finito ou infinito contável de valores entre dois valores. Correspondem a dados inteiros. Exemplos: Número de filhos, número de vezes que exercita por mês, número de bactérias no leite, etc.
- **Variáveis contínuas:** características que assumem valores infinitos entre dois valores numa escala contínua, onde valores fracionais são relevantes. Exemplo: peso, altura, idade, tempo, etc.

Tipos de variáveis

- **Variáveis Qualitativas (ou categóricas) Politómicas**
 - **Variáveis nominais:** não existe uma ordem natural entre as categorias. Exemplo: sexo, estado civil, religião, profissão, raça, cor de pele, etc.
 - **Variáveis ordinais:** existe uma ordenação natural entre as categorias. Exemplo: nível de escolaridade, mês de observação.
- **Variáveis Qualitativas (ou categóricas) Dicotómicas**
 - Variáveis que podem adotar somente dois valores. Exemplo: Sexo {M,F}.

Variáveis qualitativas nominais

- Refere-se aos nomes das “coisas” que nominalmente é imposto por nós.
- Exemplo:
 - Alimentos: café, aceitona, palmito;
 - Vestimenta: calça, camisa, blusa;
 - Artigos de Escritório: caneta, papel, lápis, etc.
- Essas medidas não possuem uma ordem inerente podendo ser codificadas de formas arbitrárias como {1,2,3...}, {A, B, C,...} ou pelos próprios nomes atribuídos dos valores.
- Cada medida (nome) possui diferenças claras na sua identidade, mas essa diferença não pode ser manipulada matematicamente. A sequência dos números é somente uma conveniência, não implicam numa ordem que quantifique a diferença.

Variável qualitativa ordinal

“Pense na medição da temperatura do seu café”

- A medição pode estar limitada a: “*quente*” ou “*frio*”
- A medição anterior possui menos informação que: “*fervendo*”, “*muito quente*”, “*quente*”, “*morno*”, “*frio*”, “*congelado*”.
- Dependendo da aplicação pode ser necessário aumentar ou diminuir o tamanho do grão. Grãos maiores proporcionam resultados mais genéricos. Grãos menores proporcionam resultados mais específicos.

Variável qualitativa ordinal

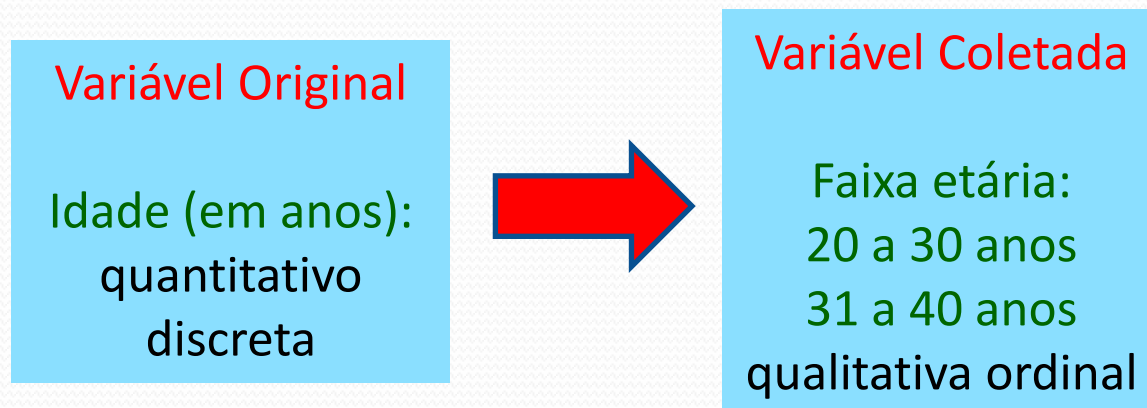
- Para as medidas pertencerem a uma escala ordinal é necessário cumprir a condição de *TRANSITIVIDADE*.

Se $(A > B)$ e $(B > C)$ então $(A > C)$

- Caso seja necessário criar uma nova codificação para a variável ordinal é necessário cumprir a condição de *TRANSITIVIDADE*.

Convertendo variável quantitativa em qualitativa

- Uma variável originariamente quantitativa pode ser coletada, transformada e armazenada de forma qualitativa.
- Essas medidas não possuem uma ordem inerente.
- Uma variável quantitativa contínua pode ser transformada em variável qualitativa (categórica) ordinal.



Convertendo variável nominal em ordinal

- A escala nominal não traz informação relevante.
- A escala ordinal traz muita informação mas não traz a magnitude da diferença entre as categorias.
- É possível incorporar informação da magnitude de diferença para uma variável nominal transformando a variável para variável quantitativa ordinal.

Nome do Vinho (Nominal)	Preferência %
Marca 1 (Vinho Cabernet)	51
Marca 2 (Vinho Merlot)	49
Marca 3 (Vinho Shiraz)	0

SE (Cabernet é-pref-a Merlot) **E**

(Merlot é-pref-a Shiraz) **ENTÃO**

(Cabernet é-pref-a Shiraz)

SE (Cabernet não-é-dispon) **ENTÃO**

(Merlot é-pref-a Shiraz)

Alguns Comentários

- O principal objetivo da mineração de dados é transformar a informação contida no conjunto de dados na forma que possa ser diretamente usada e entendida.
- O entendimento da informação contida nas variáveis permite melhores previsões do comportamento de alguns aspectos do mundo.
- O entendimento da informação de uma variável usualmente não é perfeita, existindo incertezas, portanto o conhecimento extraído não é perfeito.



Observação de Variáveis

- **Remoção de Variáveis**

- A informação básica de uma variável compreende o número de valores distintos e a frequência de cada valor.

Valor da Variável	Frequencia de ocorrência
A	1
B	2
C	15
D	2
E	1

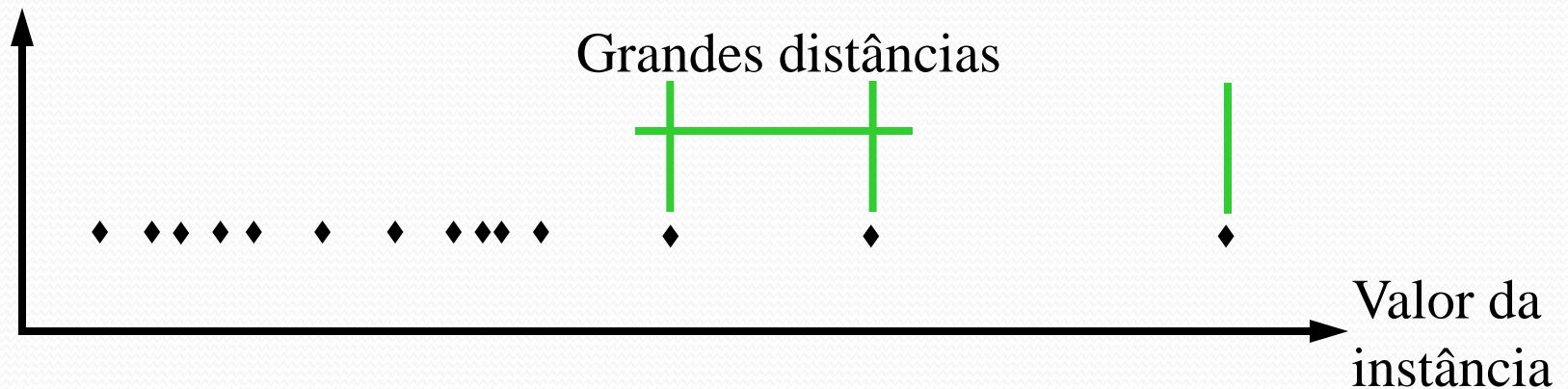


A variável deve ser uma constante ou a amostra está polarizada.

Observação de Variáveis

- **Esparcidade de Variáveis**

- As variáveis esparças podem ter insignificante valor, portanto poderiam ser eliminadas.
- Pode ser aplicada a chamada “redução dimensional” ou “Colapso dimensional”
- O minerador deverá analisar para tomar essa decisão



Observação de Variáveis

- **Incremento Dimensional de uma Variável**

- Existem algumas circunstâncias onde a dimensão de uma variável requer ser incrementada.

Código CEP



Longitude

Latitude

Observação de Variáveis

- **Numerando variáveis Categóricas**

- Numerar variáveis categóricas é uma atividade que exige cuidado para não destruir informações ao listar números para as categorias.
- Deve ser procurado manter a “*ordem natural*” de forma que a distância entre os números contenha informação.

Nunca	0	1
Casou		
Solteiro	0.1	4
Divorciado	0.15	3
Viúvo	0.65	5
Casado	1	2



**Ordem
Natural**



**Destroi
Informação**

Variabilidade nos dados

Está relacionado à análise dos valores que uma variável pode adotar

Considere a seguinte amostra representativa de uma população

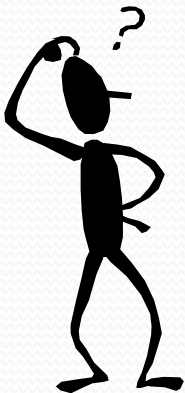
49	63	44	25	16	34	62	55	40	31	44	37	48	65	83	53	39	15	25	52
68	35	64	71	43	76	39	61	51	30	32	74	28	64	46	31	79	69	38	69
53	32	69	39	32	67	17	52	64	64	25	28	64	65	70	44	43	72	37	31
67	69	64	74	32	25	65	39	75	36	26	59	28	23	40	56	77	68	46	48

Existe algum padrão evidente??



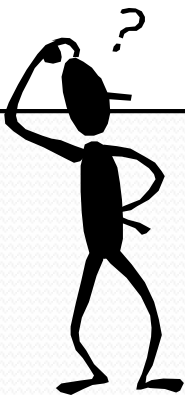
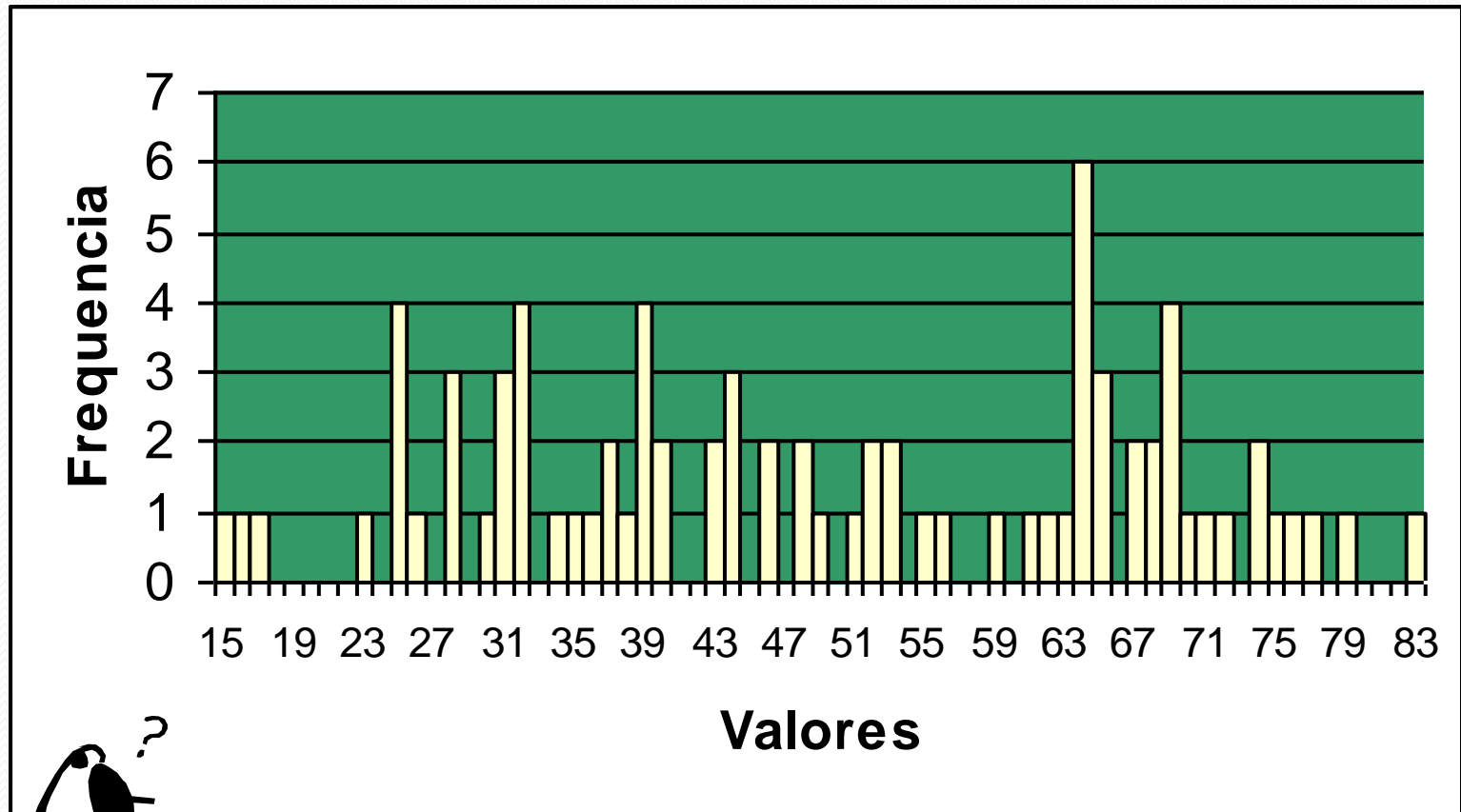
Considere que a amostra está ordenada

15	16	17	23	25	25	25	25	26	28	28	28	30	31	31	31	32	32	32	32
34	35	36	37	37	38	39	39	39	39	40	40	43	43	44	44	44	46	46	48
48	49	51	52	52	53	53	55	56	59	61	62	63	64	64	64	64	64	64	65
65	65	67	67	68	68	69	69	69	69	70	71	72	74	74	75	76	77	79	83



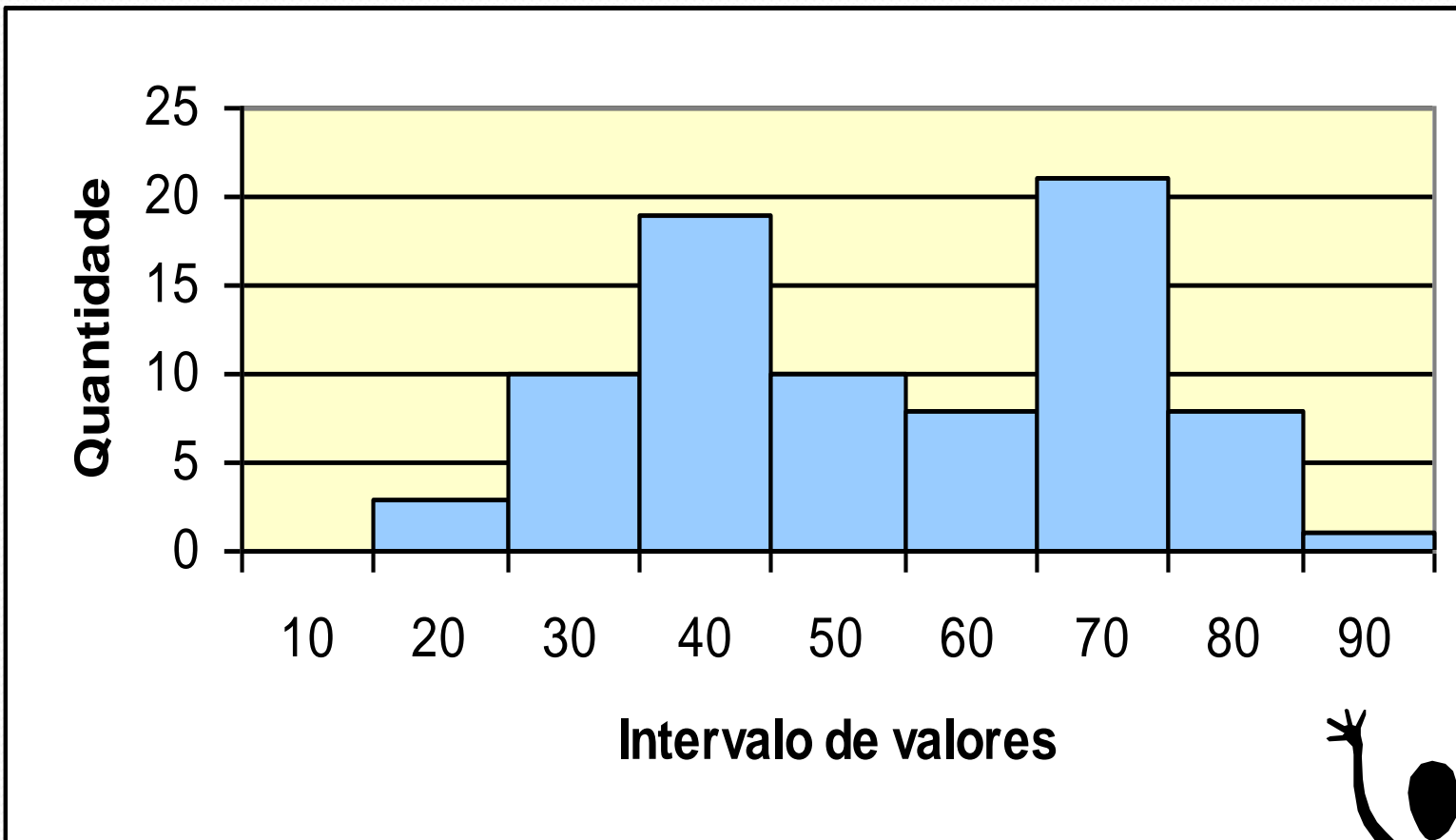
**Pode existir algum padrão -
Embora é difícil descreve-o**

Considerando o histograma de frequências

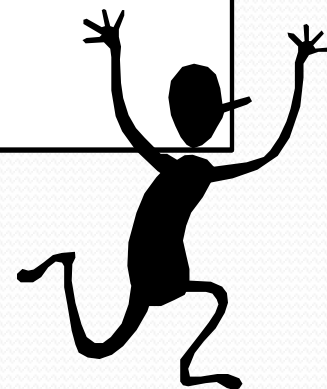


Ainda é difícil detectar algum padrão!!!

Considerando o histograma por intervalos



É possível observar algum padrão !!!



Estatística descritiva de variáveis – Medidas de tendência central

Seja o conjunto de dados relativos ao peso de 80 pessoas:

81,80	87,10	82,70	79,80	81,30	79,50	88,50	75,90
81,60	73,90	84,50	87,10	82,00	79,30	82,50	87,10
83,00	87,30	79,70	82,00	83,60	84,50	80,40	78,10
86,40	76,70	83,70	78,40	76,00	80,90	80,20	78,90
77,40	78,50	82,90	81,90	80,70	78,40	78,00	81,40
84,60	79,50	82,30	80,50	80,70	79,00	90,00	79,90
86,80	80,10	83,20	78,20	80,40	85,50	85,50	79,30
83,00	78,10	83,40	83,60	85,70	86,80	86,50	83,80
86,80	83,50	79,90	76,60	84,30	78,50	74,40	71,80
79,10	82,10	84,50	78,40	80,70	70,70	78,50	85,20

Media:

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

Mediana:


$$\tilde{x} = x_{([n+1]/2)}$$

para "n" impar

$$\tilde{x} = \frac{x_{(n/2)} + x_{(n/2+1)}}{2}$$

para "n" par

Desvio Padrão:

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


n-1: para amostra; n: para população

Variância:

$$Var(x) = S^2(x)$$

Medidas de Têndencia Central

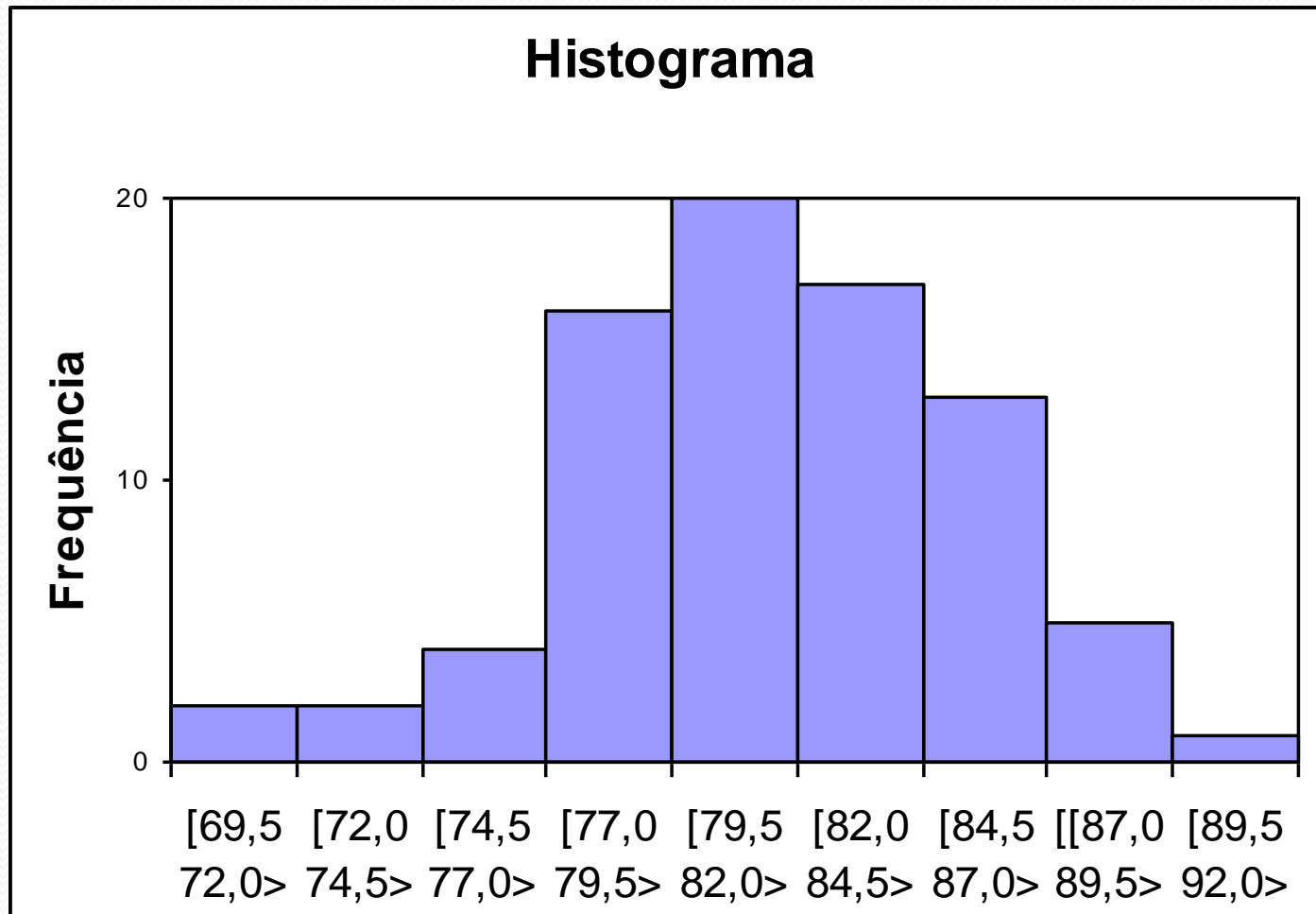
Média	81,44
Mediana	81,35
D. Pad	3,79
Variância	14,36

Moda: Valor mais frequente (popular) de uma amostra. Se aplica para dados discretos e categóricos.

Estatística descritiva de variáveis – Histograma de frequências

Intervalo i	Limites	Ponto Médio	Freq. Simples	Freq. Relativa
1	[69,5 72,0>	70,75	2,00	0,0250
2	[72,0 74,5>	73,25	2,00	0,0250
3	[74,5 77,0>	75,75	4,00	0,0500
4	[77,0 79,5>	78,25	16,00	0,2000
5	[79,5 82,0>	80,75	20,00	0,2500
6	[82,0 84,5>	83,25	17,00	0,2125
7	[84,5 87,0>	85,75	13,00	0,1625
8	[87,0 89,5>	88,25	5,00	0,0625
9	[89,5 92,0>	90,75	1,00	0,0125
		Total:	80,00	1,0000

Estatística descritiva de variáveis – Histograma de frequências



Distribuição Normal

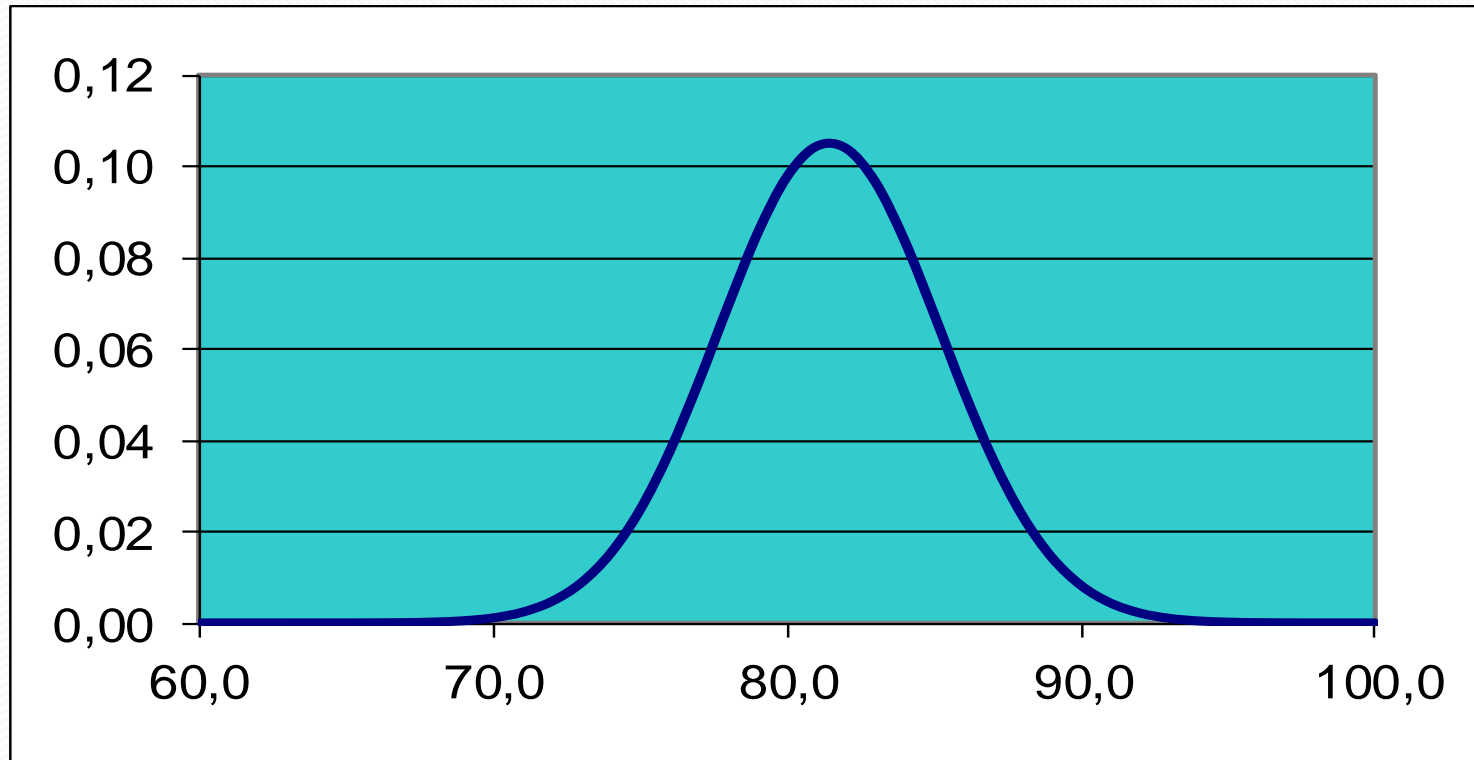
A distribuição normal é um modelo estatístico que fornece uma base teórica para o estudo do padrão de ocorrência dos elementos de uma população.

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

$\mu(x)$ média da população (ou da amostra)

$\sigma(x)$ desvio padrão da população (ou da amostra)

Distribuição Normal



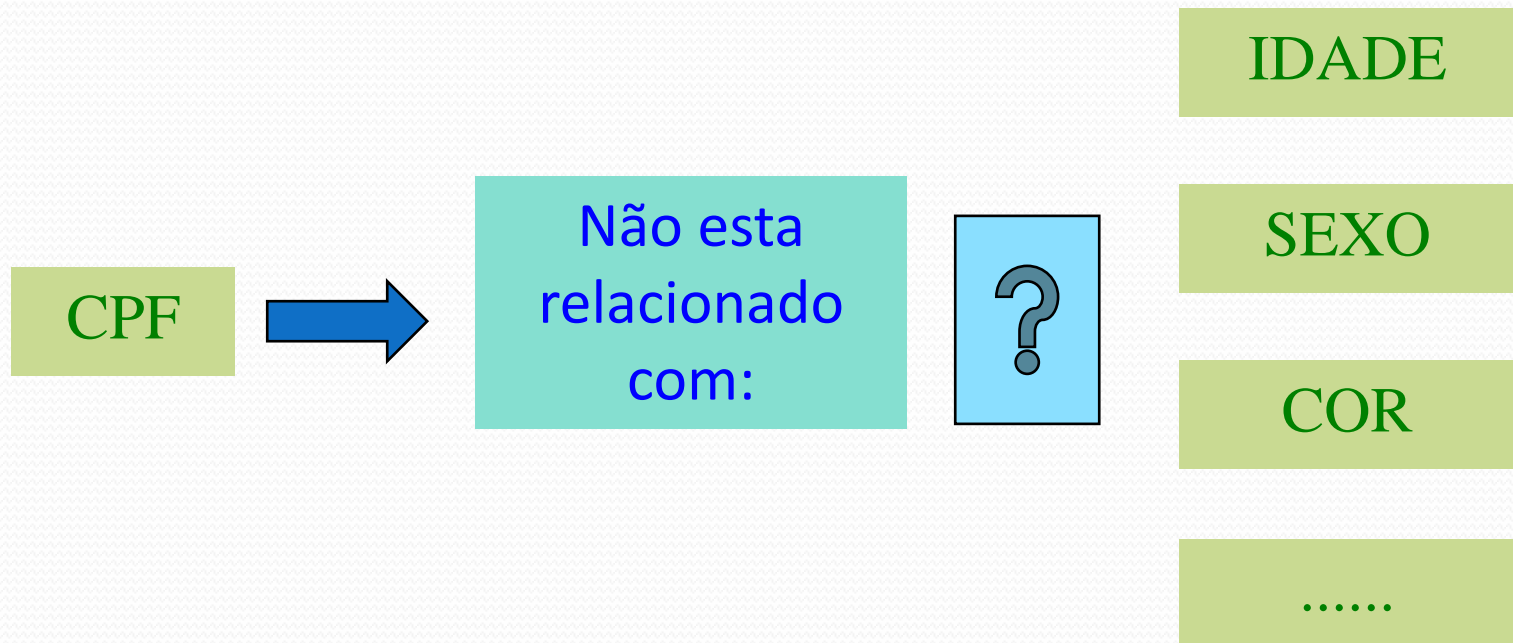
Medidas de Tendência Central	
Média	81,44
Mediana	81,35
D. Pad	3,79
Variância	14,36

Intervalo	Probabilidade	
	Interna	Externa
$\mu \pm 1\sigma$	68,2%	31,74%
$\mu \pm 2\sigma$	95,46%	4,54%
$\mu \pm 3\sigma$	99,73%	0,27%

$N(\mu, \sigma)$

Preparação dos dados

- DADOS IRRELEVANTES
- É necessário retirar dados irrelevantes que podem trazer conhecimento falso ou aumentar o tempo de processamento dos algoritmos de Data Mining.



- GRANULARIDADE

Granularidade = nível (detalhes/Agregação)

Dados detalhados é preferível a dados agregados

Para um determinado produto

Vendas
Diárias

Vendas
Semanais

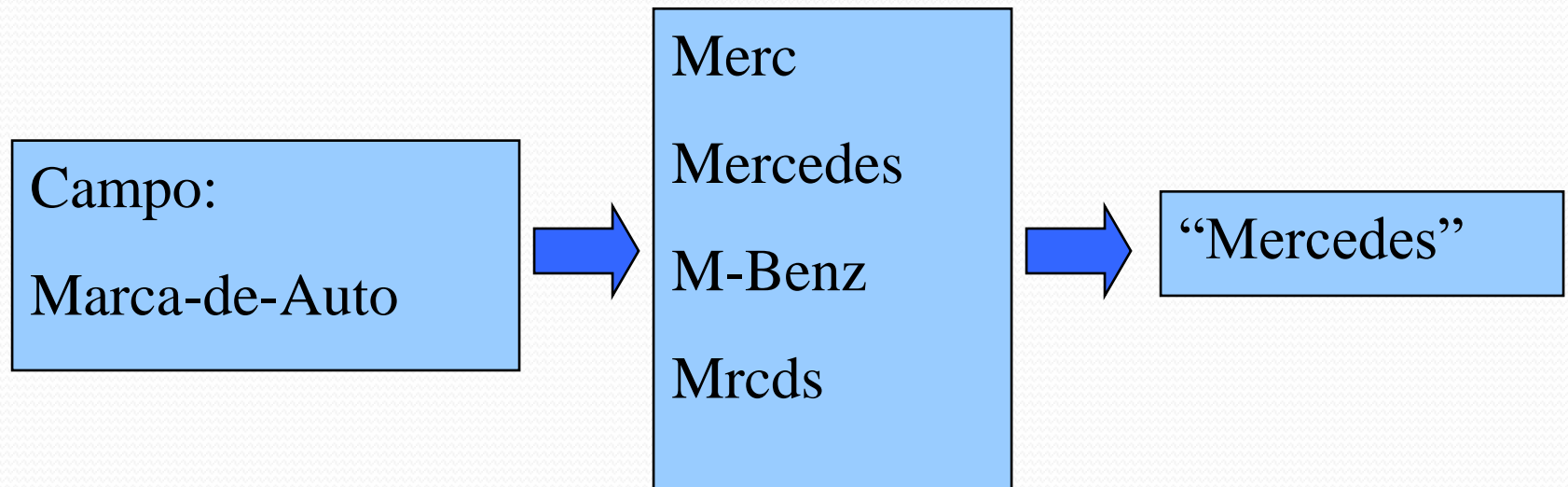
Vendas
Mensais

DETALHE

AGREGAÇÃO

- CONSISTÊNCIA e INCONSISTÊNCIA

A mesma “*coisa*” representada por diferentes nomes em diferentes sistemas.



- POLUIÇÃO

Os campos podem conter espaços em branco, estar incompletos, inexatos, inconsistentes ou não identificáveis.

Pessoa Física	EC	Data de Nasc.	Idade	Dependente	Escola/salário	Telefone
Maria da Silva	C	28/02/03	15		30%	(xxx)4567890

↑ Inconsistente Não ↑ identificável ↑ Ausente ↑ Inexato ↑ Incompleto

- RELAÇÕES

É importante observar e analisar a consistência das instâncias dos objetos da estrutura problema.

Maria da Silva; 15 anos; comprou Toyota Prius



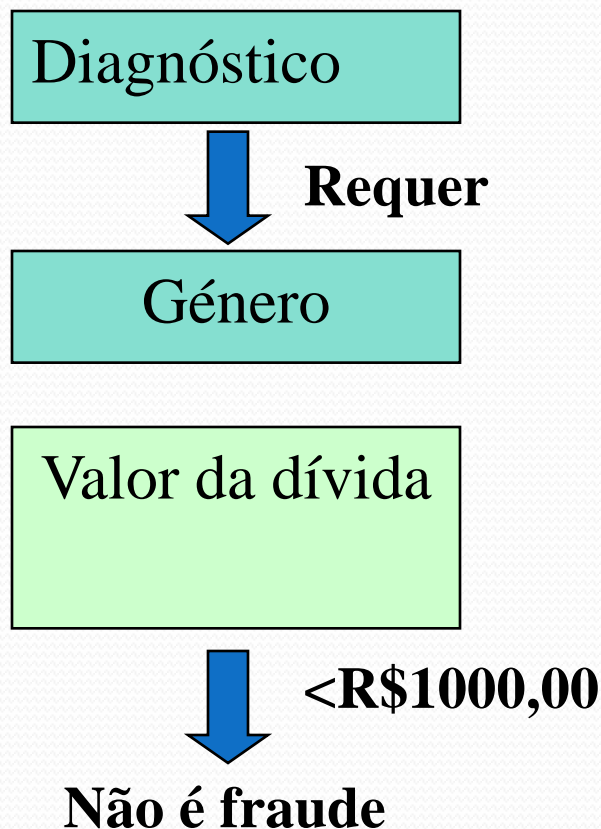
Inconsistência

• DOMÍNIOS

Cada variável possui um domínio particular ou faixa de valores dentro de um contexto.

Domínios Condicionais:

Domínios baseados em regras:



• DEFAULTS

- Alguns campos podem ter valores padrões que devem ser tomados em conta.
- Um valor Default pode ser condicional dependendo da instância que o contém.
- O valor Default condicional pode representar um padrão e deve ser analisado.

CAMPO=FRAUDE	
Registros com dívidas < 1000,00	Não (N)
Registros com dívidas >= 1000,00	SIM (S)

• DUPLICAÇÕES OU REDUNDÂNCIAS

- Ocorre principalmente quando as instâncias dependem de diferentes fluxos de dados

Data de Nascimento => Idade Atual

*Preço Unitário * Quantidade => Preço total*

- As variáveis Duplicadas ou Redundantes exigem maior esforço computacional e dependendo do caso podem ser reduzidas.

Obrigado

Professor:
Luis E. Zárate