



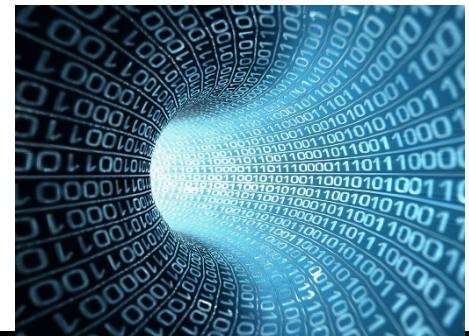
# Aprendizado de máquina

## Cap 2: Analise de dados

**UNIVERSIDADE FEDERAL DO PARÁ  
INSTITUTO DE CIÊNCIAS EXATAS E NATURAIS  
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO**  
**Prof. Jefferson Morais**  
**Email: jeffersonmorais@gmail.com**

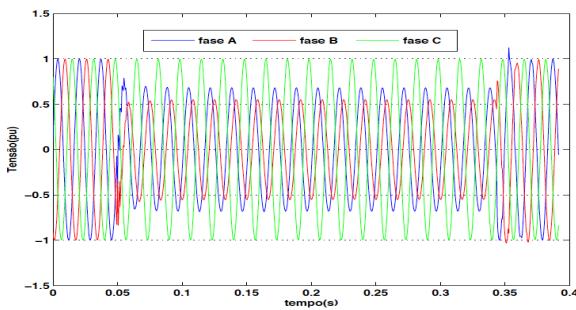
# Dados

- A cada dia, uma enorme quantidade de dados é gerada
- Motivo: Avanços recentes nas tecnologias de aquisição, transmissão e armazenamento de dados
- Estima-se que a quantidade de dados em bases de dados mundiais dobra a cada 20 meses
- Crescimento tem ocorrido em várias áreas
  - Transações bancárias, utilização de cartões de crédito, dados governamentais, medições ambientais, dados clínicos, projetos genoma, informações disponíveis na Web, etc



# Dados

- Podem ter diferentes formatos



Séries Temporais



Páginas Web

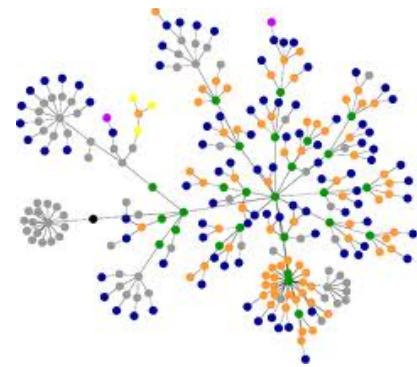


Áudios

1. Os mitólogos costumam chamar de imagens de mundo certas estruturas simbólicas pelas quais, em todas as épocas, as diferentes sociedades humanas fundamentaram, tanto coletiva quanto individualmente, a experiência do existir. Ao longo da história, essas constelações de idéias foram geradas quer pelas tradições étnicas, locais, de cada povo, quer pelos grandes sistemas religiosos. No Ocidente, contudo, desde os últimos três séculos uma outra prática de pensamento veio se acrescentar a estes
5. 10. modos tradicionais na função de elaborar as bases de nossas experiências concretas de vida: a ciência. Com efeito, a partir da revolução científica do Renascimento as ciências naturais passaram a contribuir de modo cada vez mais decisivo para a formulação das categorias que a cultura ocidental empregará para compreender a realidade e agir sobre ela.
- 15.

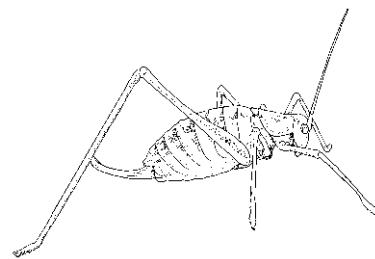


Vídeos

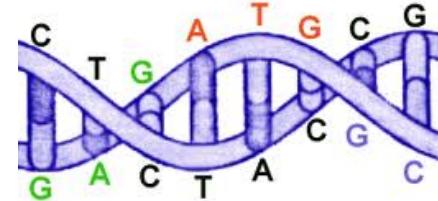


Grafos

Textos



Imagens



Sequênciais

# Representação de dados

- Os dados são geralmente transformados para o formato atributo-valor
  - Cada objeto (instância ou exemplo) corresponde a uma ocorrência dos dados
  - Cada instância é descrita por um conjunto de atributos (parâmetros) de entrada

**Dados**

**Sintomas (parâmetros)**

**Parâmetro de saída**

Temperatura	Dor	...	Pressão	Doente
40°C	sim	...	14	SIM
38°C	sim	...	12.9	SIM
			.	
			.	
			.	
36°C	não	...	12.3	NÃO

# Representação de dados

- Formalmente, os dados podem ser representados por uma matriz de objetos

$$\mathbf{X}_{n \times d}$$

$n$  = número de objetos

$d$  = número de parâmetros (excluindo o de saída)

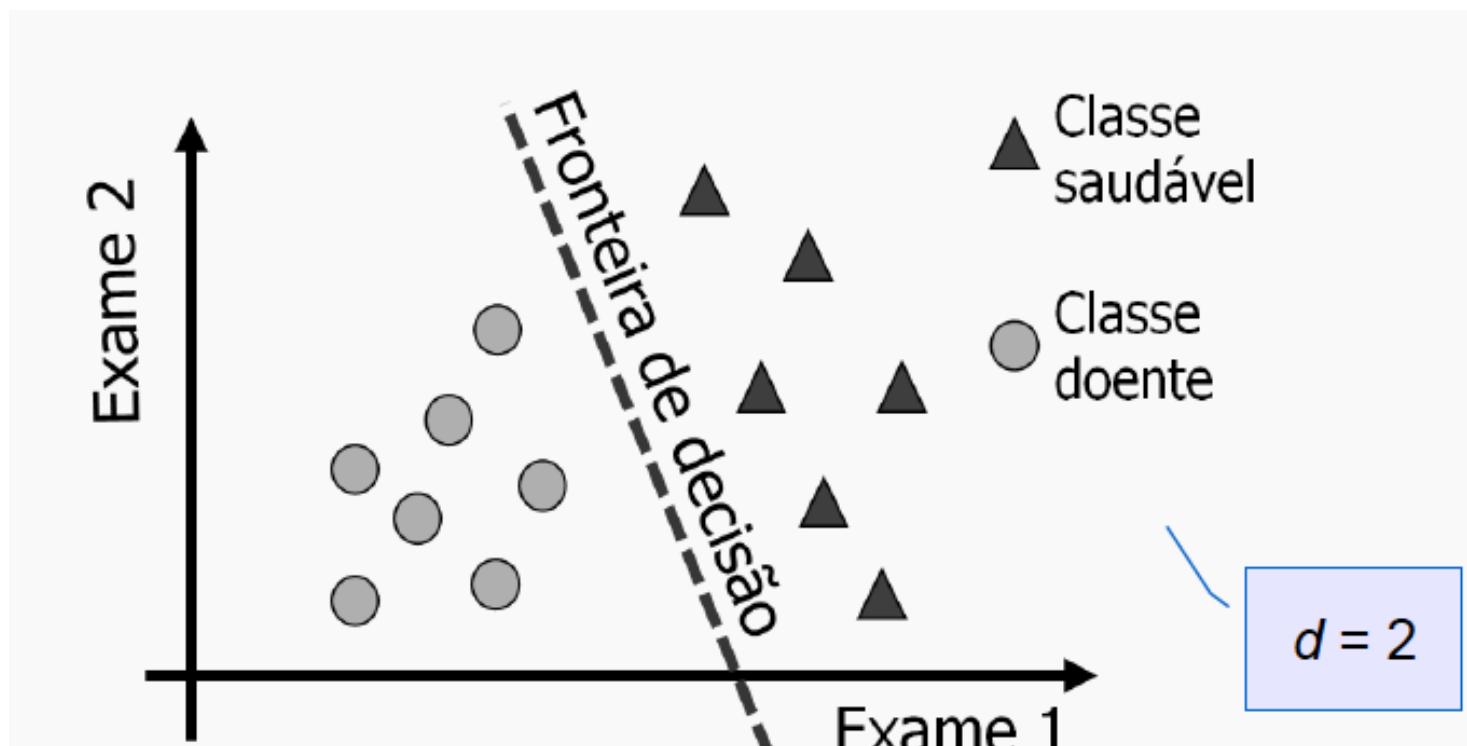
O valor de  $d$  define a dimensionalidade dos objetos

Do espaço de objetos (de entradas/de parâmetros)

Cada elemento  $x_i^j$  (ou  $x_{ij}$ )  $\Rightarrow$  valor do  $j$ -ésimo parâmetro para o objeto  $i$

# Exemplo de Espaço de Objetos

- A posição de cada objeto é definido pelos valores de dois atributos (Exame 1 e Exame 2) de entrada



# Técnicas de pré-processamento

- São frequentemente utilizadas para tornar os conjuntos de dados mais adequados para o uso de algoritmos de AM
- Podem ser agrupadas em
  - Eliminação manual de atributos
  - Integração de dados
  - Amostragem de dados
  - Balanceamento de dados
  - Limpeza de dados
  - Redução de dimensionalidade
  - Transformação de dados

# Análise de dados

- Análise das características de um conjunto de dados
  - Permite a descoberta de padrões e tendências que podem fornecer informações valiosas
  - Muitas podem ser obtidas por fórmulas estatísticas simples
    - Estatística descritiva
  - Uso de técnicas de visualização também é importante

# Análise de dados

- Caracterização de dados
  - Instâncias e Atributos
  - Tipos de Dados
- Exploração de Dados
  - Dados univariados
  - Medidas de localidade, espalhamento e distribuição
  - Dados multivariados
  - Visualização

# Análise de dados

- Valores que um atributo pode assumir podem ser definidos de diferentes formas
  - Tipo
    - Grau de quantização nos dados
  - Escala
    - Significância relativa dos valores
- Conhecer o tipo/escala dos atributos auxilia a identificar a forma adequada de preparar os dados e posteriormente de modelá-los

# Tipos de Atributos

- Quantitativo (numérico)
  - Representa quantidades
  - Valores podem ser tanto ordenados quanto utilizados em operações aritméticas
  - Pode ser **contínuos** e **discretos**
  - Possuem unidade associada
- Qualitativo (simbólico ou categórico)
  - Representa qualidades
  - Valores podem ser associados a categorias
  - Apesar de poder ter valores ordenados, operações aritméticas não são aplicáveis
  - Ex: {pequeno, médio, grande} e {matemática, física, química}

# Tipos de Atributos

## ● Atributos Quantitativos

### ■ Contínuos

- Podem assumir um número infinito de valores
- Geralmente resultados de medidas
- Frequentemente são representados por números reais
- Ex: peso, tamanho, distância

### ■ Discretos

- Número finito ou infinito contável de valores
- Um caso especial são os atributos binários (ou booleanos)
- Ex: 0/1, sim/não, {12, 23, 45}

# Tipos de Atributos

		Qualitativo	Quantitativo contínuo	Quantitativo contínuo				
Id.	Nome	Idade	Sexo	Peso	Manchas	Temp	# Int.	Est. Diagnóstico
4201	João	28	M	79	Concentradas	38,0	2	SP Doente
3217	Maria	18	F	67	Inexistentes	39,5	4	MG Doente
4039	Luiz	49	M	92	Espalhadas	38,0	2	RS Saudável
1920	José	18	M	43	Inexistentes	38,5	8	MG Doente
4340	Cláudia	21	F	52	Uniformes	37,6	1	PE Saudável
2301	Ana	22	F	72	Inexistentes	38,0	3	RJ Doente
1322	Marta	19	F	87	Espalhadas	39,0	6	AM Doente
3027	Paulo	34	M	67	Uniformes	38,4	2	GO Saudável

Qualitativo

Quantitativo discreto

Qualitativo

Quantitativo

Quantitativo discreto

Qualitativo

Alguns atributos qualitativos são representados por números, mas não faz sentido a utilização de operadores aritméticos sobre seus valores

# Escala de atributos

- Define operações que podem ser realizadas sobre os valores dos atributos

- Nominais
  - Ordinais
  - Intervalares
  - Racionais
- 
- The diagram illustrates the classification of attributes. It shows four categories listed vertically: 'Nominais', 'Ordinais', 'Intervalares', and 'Racionais'. A black curly brace on the right side groups 'Nominais' and 'Ordinais' under the heading 'Qualitativos'. A red curly brace on the right side groups 'Intervalares' and 'Racionais' under the heading 'Quantitativos'.
- Qualitativos
- Quantitativos

# Escala de Atributos

- Escala Nominal

- Valores são nomes diferentes e carregam a menor quantidade de informação
- Não existe relação de ordem entre os valores
- Operações que podem ser aplicadas:  $=$ ,  $\neq$
- Ex.: número de conta corrente, cores, sexo

# Escala de Atributos

- Escala ordinal

- Valores refletem ordem das categorias representadas
- Operações aplicáveis:  $=$ ,  $\neq$ ,  $<$ ,  $>$ ,  $\leq$ ,  $\geq$
- Ex: avaliações qualitativas de temperatura {quente, morno, frio}

# Escala de Atributos

- Escala intervalar

- Números que variam em um intervalo
- É possível definir tanto a ordem quanto a diferença em magnitude entre dois valores
  - A diferença em magnitude indica a distância que separa dois valores no intervalo possível de valores
- Origem da escala definida de maneira arbitrária
  - O valor zero não tem o mesmo significado do zero utilizado em operações aritméticas
- Operações aplicáveis:  $=$ ,  $\neq$ ,  $<$ ,  $>$ ,  $\leq$ ,  $\geq$ ,  $+$ ,  $-$
- Ex: Temperatura em  $^{\circ}\text{C}$  ou  $^{\circ}\text{F}$ , datas

# Escala de Atributos

- Escala racional

- Carregam mais informações
- Os números têm significado absoluto
  - Existe um zero absoluto junto com a unidade de medida, de forma que a razão tenha significado
- Operações aplicáveis:  $=, \neq, <, >, \leq, \geq, +, -, *, /$
- Ex: tamanho, distância, salário, saldo em conta

# Escala de Atributos

<b>Id.</b>	<b>Nome</b>	<b>Idade</b>	<b>Sexo</b>	<b>Peso</b>	<b>Manchas</b>	<b>Temp.</b>	<b># Int.</b>	<b>Est.</b>	<b>Diagnóstico</b>
4201	João	28	M	79	Grandes	38,0	2	SP	Doente
3217	Maria	18	F	67	Pequenas	39,5	4	MG	Doente
4039	Luiz	49	M	92	Grandes	38,0	2	RS	Saudável
1920	José	18	M	43	Grandes	38,5	20	MG	Doente
4340	Cláudia	21	F	52	Médias	37,6	1	PE	Saudável
2301	Ana	22	F	72	Pequenas	38,0	3	RJ	Doente
1322	Marta	19	F	87	Grandes	39,0	6	AM	Doente
3027	Paulo	34	M	67	Médias	38,4	2	GO	Saudável

**Nominal**

**Ordinal**

**Intervalar**

**Racional**

Fonte: Profa. Ana Lorena

# Exercício de Fixação

- Definir o tipo e escala dos seguintes atributos:
  - Renda Mensal: **Quantitativo racional**
  - Número de matrícula **Qualitativo nominal**
  - Data de Nascimento **Quantitativo intervalar**
  - Número de palavras de um texto **Quantitativo racional**
  - Código postal **Qualitativo nominal**
  - Posição em uma corrida **Qualitativo ordinal**

# Exploração de Dados

- Grande quantidade de informações úteis pode ser extraída de um conjunto de dados por meio de sua análise ou exploração
  - Ajuda na seleção de técnicas de pré-processamento e aprendizado de máquina
- Forma simples de exploração: estatística descritiva
  - Resumo quantitativo das principais características de um conjunto de dados
  - Muitas dessas medidas são calculadas rapidamente
  - Ex: Idade média dos pacientes e a porcentagem de pacientes do sexo masculino
  - **Captura de informações como: frequência, localização ou tendência central, dispersão ou espalhamento, distribuição ou formato**

# Exploração de Dados

- Frequência

- Medida mais simples
- Mede a proporção de vezes que um atributo assume um dado valor em um determinado conjunto de dados
- Pode ser aplicada a valores numéricos e simbólicos
- Ex: Em um conjunto de dados médicos, 40% dos pacientes têm febre

# Exploração de Dados

- Frequência

- Medida mais simples
- Mede a proporção de vezes que um atributo assume um dado valor em um determinado conjunto de dados
- Pode ser aplicada a valores numéricos e simbólicos
- Ex: Em um conjunto de dados médicos, 40% dos pacientes têm febre

# Exploração de Dados

- Localização, dispersão e distribuição
  - Diferem para os casos em que os dados apresentam apenas um atributo (dados **univariados**) ou mais de um atributo (dados **multivariados**)
    - Maioria dos dados em AM é multivariado, mas análises em cada atributo podem fornecer informações valiosas
  - Geralmente aplicados a valores numéricos

# Exploração de Dados

- Localização, dispersão e distribuição
  - Diferem para os casos em que os dados apresentam apenas um atributo (dados **univariados**) ou mais de um atributo (dados **multivariados**)
    - Maioria dos dados em AM é multivariado, mas análises em cada atributo podem fornecer informações valiosas
  - Geralmente aplicados a valores numéricos

# Dados univariados

- Dados com apenas um atributo

Conjunto com  $n$  dados  $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$

**Observação:** O termo conjunto não tem o mesmo significado do termo utilizado em teoria dos conjuntos. Em conjunto de dados um mesmo valor pode aparecer mais de uma vez em atributo

# Dados univariados: medidas de localidade

- Definem pontos de **referência** nos dados e variam para dados numéricos e símbolicos
- Para dados símbolicos
  - Utiliza-se geralmente a **moda**: valor encontrado com maior frequencia para um atributo
- Para dados numéricos
  - Média, Mediana e percentil

# Moda

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Concentradas	38,0	2	SP	Doente
3217	Maria	18	F	67	Inexistentes	39,5	4	MG	Doente
4039	Luiz	49	M	92	Espalhadas	38,0	2	RS	Saudável
1920	José	18	M	43	Inexistentes	38,5	20	MG	Doente
4340	Cláudia	21	F	52	Uniformes	37,6	1	PE	Saudável
2301	Ana	22	F	72	Inexistentes	38,0	3	RJ	Doente
1322	Marta	19	F	87	Espalhadas	39,0	6	AM	Doente
3027	Paulo	34	M	67	Uniformes	38,4	2	GO	Saudável



Moda: Inexistentes

# Média

- Equação

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

**Observação:** Problema da média é que ela é sensível à presença de *outliers*. Bom indicador do meio de um conjunto de valores apenas se os valores estão distribuídos simetricamente

# Mediana

- O valor que ocupa a **posição central** em um conjunto de dados **ordenados**
- Consequência: tem a propriedade de dividir os dados em duas partes iguais quanto ao número de elementos
  - 50% dos elementos são maiores e 50% são menores do que a mediana
- Não é influenciada por valores extremos (*outliers*)
  - Ex.: A = {4, 9, 1000} e B={4, 9, 10} em ambos a mediana é 9

# Mediana

- Passos

- Ordenar os valores de forma crescente
- Calcular a equação

$$mediana(\mathbf{x}) = \begin{cases} \frac{1}{2}(x_r + x_{r+1}) & \text{se } n \text{ for par } (n = 2r) \\ x_{r+1} & \text{se } n \text{ for ímpar } (n = 2r + 1) \end{cases}$$

- Facilita observar se a distribuição é assimétrica ou se existem *outliers*

# Mediana

- Exemplos

- {17, 4, 8, 21, 4 }
  - Ordenado (rol): {4, 4, 8, 17, 21}
  - Número ímpar de elementos => mediana = 8
- {17, 4, 8, 21, 4, 15, 13, 9}
  - Ordenado: 4, 4, 8, 9, 13, 15, 17, 21
  - Número par de elementos => mediana =  $(9+13)/2 = 11$

# Exemplo 1 Media e Mediana

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Concentradas	38,0	2	SP	Doente
3217	Maria	18	F	67	Inexistentes	39,5	4	MG	Doente
4039	Luiz	49	M	92	Espalhadas	38,0	2	RS	Saudável
1920	José	18	M	43	Inexistentes	38,5	20	MG	Doente
4340	Cláudia	21	F	52	Uniformes	37,6	1	PE	Saudável
2301	Ana	22	F	72	Inexistentes	38,0	3	RJ	Doente
1322	Marta	19	F	87	Espalhadas	39,0	6	AM	Doente
3027	Paulo	34	M	67	Uniformes	38,4	2	GO	Saudável

Média: 26,1

Mediana: 21,5

Média: 5

Mediana: 2,5

# Exemplo 2: Média e Mediana

- Calcular a média e a mediana em relação ao resultado de um teste objetivo de conhecimento gerais aplicado a um grupo de alunos, cujas pontuações foram: 5, 8, 6, 3, 7, 5, 9
  - Média = 6,14
  - Mediana = 6

# Média truncada

- Descarta elementos extremos da sequência ordenada de valores
  - Minimizar problemas da média
  - Necessário definir um porcentagem de descarte
- Passos
  - Definir a porcentagem  $p$
  - Ordenar os valores
  - Descartar  $(p/2)\%$  de valores de cada extremo
  - Calcular a média dos exemplos restantes

# Média truncada

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Concentradas	38,0	2	SP	Doente
3217	Maria	18	F	67	Inexistentes	39,5	4	MG	Doente
4039	Luiz	49	M	92	Espalhadas	38,0	2	RS	Saudável
1920	José	18	M	43	Inexistentes	38,5	20	MG	Doente
4340	Cláudia	21	F	52	Uniformes	37,6	1	PE	Saudável
2301	Ana	22	F	72	Inexistentes	38,0	3	RJ	Doente
1322	Marta	19	F	87	Espalhadas	39,0	6	AM	Doente
3027	Paulo	34	M	67	Uniformes	38,4	2	GO	Saudável



Média: 26,1

Mediana: 21,5

Média truncada (p =25%): 23,7

Média: 5

Mediana: 2,5

Média truncada (p=25%: 3,2

# Exercício de Fixação

- Dado o conjunto de dados  $\{4, 2, 3, 7, 5, 30\}$   
calcular
  - Média = **8,5**
  - Mediana= **4,5**
  - Média truncada ( $p=25\%$ )= **4,75**

# Quartis e percentis

- Mediana divide dados ordenados ao meio
  - Quartis e percentis usam pontos de divisão diferentes
- Quartis
  - Divide em quartos
  - 1º quartil ( $Q1$ ) => valor que tem 25% dos demais valores abaixo dele
  - 2º quartil = mediana
- Percentil
  - Para  $p$  entre 0 e 100
  - $p$ -ésimo percentil =  $P_p \Rightarrow x_i$ 
    - $P_{25} = Q1$
    - $P_{50} = Q2 = \text{mediana}$

# Percentil

## Algoritmo para cálculo do percentil

Entrada: n valores e percentil p

Saída: valor do percentil

- Ordenar os n valores de maneira crescente
- Calcular  $k = n * p$
- Se k não for inteiro então
  - Arredondar para o próximo inteiro
  - Retornar o valor dessa posição na sequência
- Senão
  - Retornar média entre os valores nas posições k e k+1

# Quartil e Percentil

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Concentradas	38,0	2	SP	Doente
3217	Maria	18	F	67	Inexistentes	39,5	4	MG	Doente
4039	Luiz	49	M	92	Espalhadas	38,0	2	RS	Saudável
1920	José	18	M	43	Inexistentes	38,5	20	MG	Doente
4340	Cláudia	21	F	52	Uniformes	37,6	1	PE	Saudável
2301	Ana	22	F	72	Inexistentes	38,0	3	RJ	Doente
1322	Marta	19	F	87	Espalhadas	39,0	6	AM	Doente
3027	Paulo	34	M	67	Uniformes	38,4	2	GO	Saudável



Média=26,1

Mediana=21,5

Média truncada ( $p = 25\%$ )=23,7

Q1= 18,5; Q2 = 21,5; Q3:31 P40: 21



Média = 5

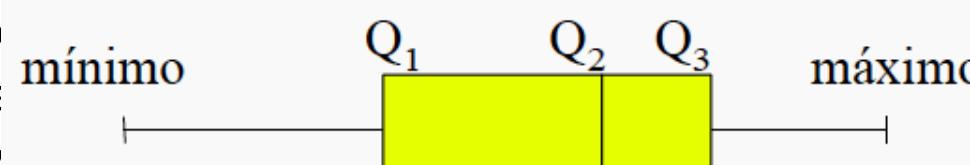
Mediana = 2,5

Média truncada ( $p=25\%$ ) = 3,2

Q1= 2; Q2 =2,5; Q3:5 P40: 2

# Boxplots

- Também chamado de diagramas de Box e Whisker
- Ferramenta para localizar e analisar a variação de uma variável dentre diferentes grupos de dados
- Forma gráfica de visualizar quartis
  - Usa quartis ( $Q_1$ ,  $Q_2$  e  $Q_3$ ) e valores máximo e mínimo

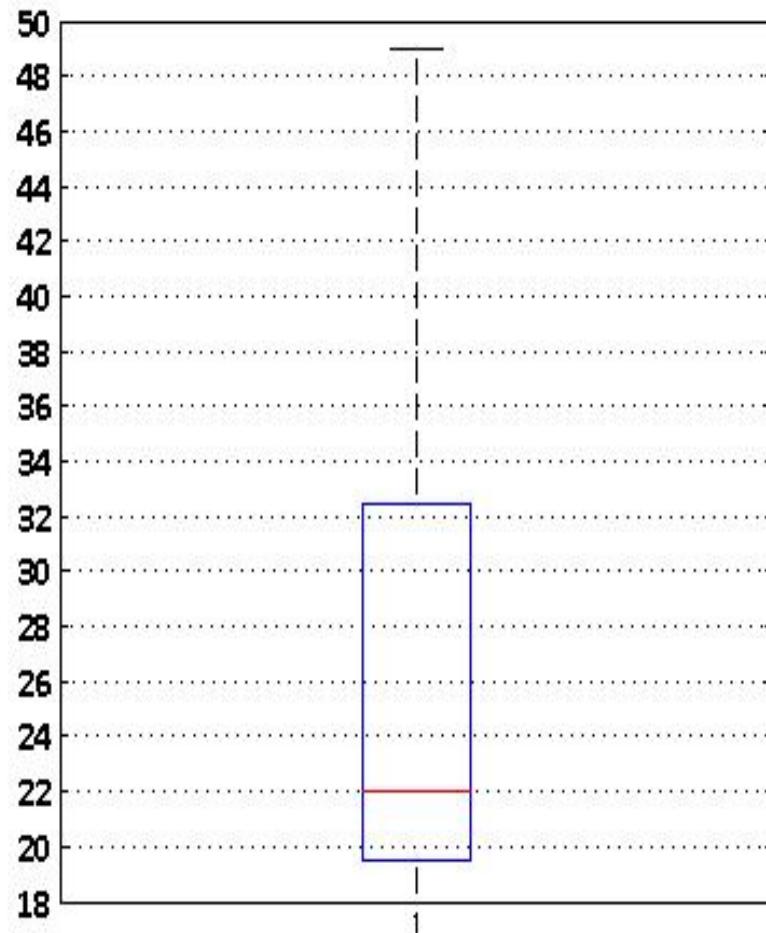
- Boxplot  
apenas  
intervalo  
*outliers*

mínimo                          Q<sub>1</sub>                          Q<sub>2</sub>                          Q<sub>3</sub>                          máximo

maior/menor valor  
quartil (até 1,5 \*  
desvio padrão) são considerados  
outliers

# Boxplot

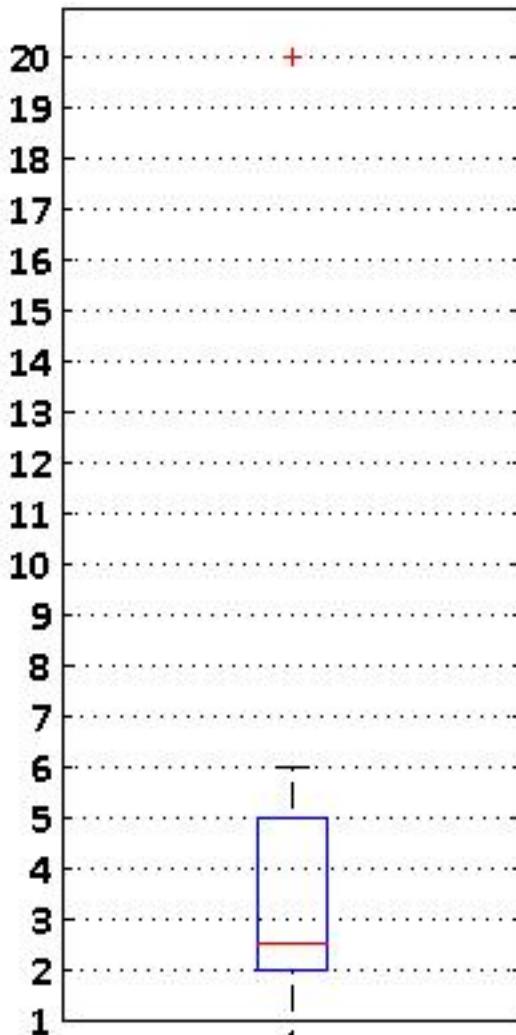
Id.	Nome	Idade
4201	João	28
3217	Maria	18
4039	Luiz	49
1920	José	18
4340	Cláudia	21
2301	Ana	22
1322	Marta	19
3027	Paulo	34



Est.	Diagnóstico
SP	Doente
MG	Doente
RS	Saudável
MG	Doente
PE	Saudável
RJ	Doente
AM	Doente
GO	Saudável

# Boxplot

*Outlier*

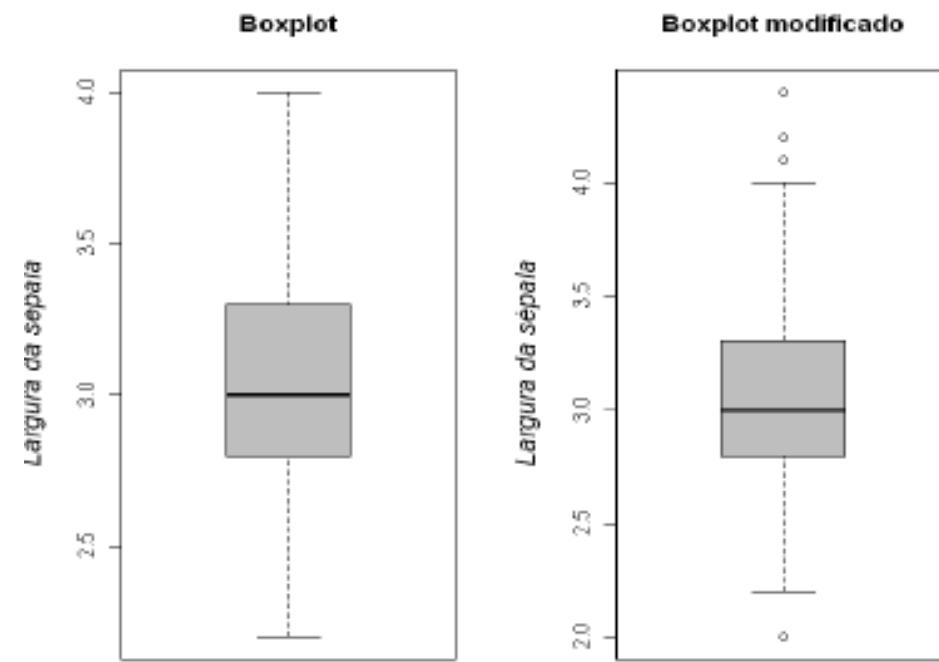


Id.	Nome	Idade	Sexo
4201	João	28	M
3217	Maria	18	F
4039	Luiz	49	M
1920	José	18	M
4340	Cláudia	21	F
2301	Ana	22	F
1322	Marta	19	F
3027	Paulo	34	M

# Int.	Est.	Diagnóstico
2	SP	Doente
4	MG	Doente
2	RS	Saudável
20	MG	Doente
1	PE	Saudável
3	RJ	Doente
6	AM	Doente
2	GO	Saudável

# Boxplot

- Ex: conjunto de dados iris
  - 150 instâncias
  - 4 atributos de entrada (contínuos)
    - Tamanho pétala
    - Tamanho sépala
    - Largura pétala
    - Largura sépala
  - 3 classes (espécies de íris)
    - Íris vírginica
    - Íris setosa
    - Íris versicolor



# Dados univariados: medidas de espalhamento

- Medem dispersão ou espalhamento de um conjunto de valores
  - Permitem observar se valores estão
    - Espalhados (dispersos)
    - Concentrados em torno de um valor (ex. Média)
- Medidas comuns
  - Intervalo
  - Variância
  - Desvio padrão

# Intervalo

- Mostra a dispersão máxima entre os valores
  - Medida simples
- Problema: não é uma medida boa se a maioria dos valores estão próximos de um ponto, com um pequeno número de valores extremos

$$\text{intervalo}(\mathbf{x}) = \max_{i=1,\dots,n}(x_i) - \min_{i=1,\dots,n}(x_i)$$

# Intervalo

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Grandes	38,0	2	SP	Doente
3217	Maria	18	F	67	Pequenas	39,5	4	MG	Doente
4039	Luiz	49	M	92	Grandes	38,0	2	RS	Saudável
1920	José	18	M	43	Grandes	38,5	20	MG	Doente
4340	Cláudia	21	F	52	Médias	37,6	1	PE	Saudável
2301	Ana	22	F	72	Pequenas	38,0	3	RJ	Doente
1322	Marta	19	F	87	Grandes	39,0	6	AM	Doente
3027	Paulo	34	M	67	Médias	38,4	2	GO	Saudável



Intervalo = 31



Intervalo = 19

# Valor Esperado

- Busca resumir o comportamento de uma variável aleatória encontrando a tendência central da variável aleatória
- Exemplos
  - Quando pedimos nossa comida em um restaurante e o garçom informa quanto tempo leva para ficar pronto
  - Quando estamos em um ponto de ônibus e perguntamos para pessoa ao lado, quando tempo leva até que o próximo ônibus venha

# Valor esperado de variáveis aleatórias discretas

- Seja  $X$  uma variável aleatória discreta, com valores possíveis  $x_1, \dots, x_n, \dots$ . Seja  $p(x_i) = P[X = x_i]$ . Então, a esperança de  $X$ , também chamada de valor esperado de  $X$

$$E(X) = \mu_X = \sum_{i=1}^{\infty} x_i P[X = x_i].$$

- Este número também é denominado valor médio ou expectância de  $X$
- Lembrando que a esperança só está bem definida se, e somente se, a série acima for absolutamente convergente

$$E(X) = \sum_{i=1}^{\infty} |x_i| P[X = x_i].$$

# Valor esperado de variáveis aleatórias discretas

- Observação: Se  $X$  tomar apenas um número finito de valores o valor esperado assume a seguinte expressão

$$E(X) = \sum_{i=1}^n x_i P[X = x_i]$$

- Isto pode ser considerado como uma média ponderada dos possíveis valores  $x_1, \dots, x_n$ .
- Se todos esses valores possíveis forem igualmente prováveis, então temos a média aritmética

$$E(X) = \frac{1}{n} \sum_{i=1}^n x_i$$

# Valor esperado de variáveis aleatórias discretas

- Exemplo: Considere o lançamento equilibrado de um dado de 6 faces. A variável aleatória  $X$  = “número da face voltada para cima”. Calcular o valor esperado de  $X$ .
- Os valores possíveis de  $X$  são  $\{1, 2, 3, 4, 5, 6\}$
- Esses valores são equiprováveis

$$E(X) = \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = \frac{7}{2}.$$

- Obs: Este exemplo ilustra claramente que  $E(X)$  não é o resultado que podemos esperar quando  $X$  for observado uma única vez.  $E(X) = 7/2$  nem é um possível valor de  $X$ . Esse valor na verdade significa que se jogássemos o dado um grande número de vezes e depois calculássemos a média aritmética dos vários resultados, esperaríamos que essa média ficasse próxima de  $7/2$  e quanto maior fosse o número de vezes que o dado fosse lançado, mais a média aritmética se aproximaria de  $7/2$ .

# Valor esperado de variáveis aleatórias discretas

- Seja  $X$  uma variável aleatória com valores finitos e  $g$  uma função, então

$$E[g(X)] = \sum_{i=1}^n g(x_i)P(x_i)$$

- Ex: Os possíveis valores de  $X$  são  $\{-1, 5\}$ . Além disso,  $P[x = -1] = 0,1$  e  $P[x = 5] = 0,9$ . Portanto, o valor esperado é dado por

$$E(X) = -1(0.1) + 5(0.9) = 4.4$$

# Valor esperado de variáveis aleatórias discretas

- Seja X uma variável aleatória tal que

X	-2	-1	0	1	2
P[X=x]	1/5	1/5	1/5	1/5	1/5

- Seja  $g(x) = X^2$ , calculemos a esperança de X e de g(X)

$$E[X] = \sum_{i=1}^5 x_i P(x_i) = -2\frac{1}{5} - 1\frac{1}{5} + 0\frac{1}{5} + 1\frac{1}{5} + 2\frac{1}{5} = 0.$$

$$E[g(X)] = \sum_{i=1}^5 g(x_i) P(x_i) = (-2)^2\frac{1}{5} - 1^2\frac{1}{5} + 0^2\frac{1}{5} + 1^2\frac{1}{5} + 2^2\frac{1}{5} = 2.$$

# Valor esperado de variáveis aleatórias discretas

- Exercício de fixação
- Uma indústria alimentícia participou de três licitações públicas as quais lhe proporcionaram lucros de 30, 50 e 60 mil reais, respectivamente. Se a probabilidade de que essa indústria vença a licitação forem de 0,3; 0,7 e 0,2 respectivamente qual o valor esperado de lucros desta indústria?

$$E(X) = 0,3 \cdot 30 + 0,7 \cdot 50 + 0,2 \cdot 60 = 56$$

# Valor esperado de variáveis aleatórias contínuas

- Seja  $X$  uma variável aleatória contínua com **função densidade de probabilidade** (fdp)  $f$ . Definimos o valor esperado ou esperança matemática ou média  $X$

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx,$$

- Temos que  $E(X)$  vai estar bem definida se a integral for finita

$$\int_{-\infty}^0 xf(x)dx < \infty \quad \text{e} \quad \int_0^{\infty} xf(x)dx < \infty.$$

# Valor esperado de variáveis aleatórias contínuas

Exemplo: Seja  $X$  o tempo (em minutos) durante o qual um equipamento elétrico é utilizado em carga máxima, em um certo período de tempo especificado. Então,  $X$  é uma variável aleatória contínua e sua fdp é dada por

$$f(x) = \begin{cases} \frac{1}{(1500)^2}x, & \text{se } 0 \leq x \leq 1500; \\ \frac{-1}{(1500)^2}(x - 3000), & \text{se } 1500 \leq x \leq 3000; \\ 0, & \text{para quaisquer outros valores.} \end{cases}$$

Calcular a esperança de  $X$ .

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx = \frac{1}{(1500)(1500)} \left[ \int_0^{1500} x^2 dx - \int_{1500}^{3000} x(x - 3000) dx \right] = 1500 \text{ minutos.}$$

# Propriedades do Valor Esperado

- Se  $X=c$ , então  $E(X)=c$
- Seja  $C$  uma constante e  $X$  uma variável aleatória. Então

$$E(CX) = CE(X)$$

- Sejam  $X$  e  $Y$  duas variáveis aleatórias quaisquer. Então,

$$E(X + Y) = E(X) + E(Y)$$

- Sejam  $n$  variáveis aleatórias  $X_1, \dots, X_n$ . Então

$$E(X_1 + X_2 + \dots + X_n) = E(X_1) + E(X_2) + \dots + E(X_n)$$

- Sejam  $X$  e  $Y$  variáveis aleatórias independentes. Então

$$E(XY) = E(X)E(Y)$$

# Momentos

- Se  $X$  é uma variável aleatória, então  $E(X^r)$  é chamado de r-ésimo momento de  $X$  caso exista, ou seja, se  $E(X^r) < \infty$
- O r-ésimo momento em torno de  $\alpha$  é definido como sendo  $E[(X-\alpha)]^r$ , caso exista
- Se  $\alpha = E[X]$  então ele é chamado de o r-ésimo momento central
- Notem que , ou seja, a  $Var(x)$  é definida como sendo o segundo momento central

$$Var(X) = E[(X - E(X))^2]$$

# Variância de variáveis aleatórias

- Suponhamos que, para uma variável aleatória  $X$ , verificamos que  $E(X) = 2$ . Qual o significado disso?
- Significa: se um grande número de determinações de  $X$ , digamos  $x_1, \dots, x_n$ , ao calcularmos a média desses valores de  $X$  ela estará próxima de 2, se  $n$  for grande
- Suponhamos, por exemplo, que  $X$  representa a duração de vida de lâmpadas que estão sendo recebidas de um fabricante, e que  $E(X) = 1000$  horas
  - Isto pode significar que a maioria das lâmpadas deve durar um período de tempo compreendido entre 900 horas e 1100 horas

# Variância de variáveis aleatórias

- Seja  $X$  uma variável aleatória. Definimos a variância de  $X$ , denotada por  $\text{Var}(X)$  ou  $\sigma^2$  por

$$\text{Var}(X) = E[X - E(X)]^2.$$

- A raiz quadrada positiva de  $\text{Var}(X)$  é denominada o desvio-padrão de  $X$ , denotado por  $\sigma$
- Obs: O número  $\text{Var}(X)$  é expresso por unidades quadradas de  $X$ .

# Variância e desvio padrão

- Mais utilizados

Valor esperado

$$E[\mathbf{X}] = \mu = \sum_{i=1}^n p_i x_i$$

$$\text{Var}(\mathbf{X}) = \sigma^2 = E[(\mathbf{X} - \mu)^2]$$

$$\text{Desvio padrão}(\mathbf{X}) = \sigma = \sqrt{\sigma^2}$$

- Problema: também são distorcidas pela presença de outliers

Prove que:

$$\text{Var}(\mathbf{X}) = E[X^2] - E[X]^2$$

# Desvio padrão

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Grandes	38,0	2	SP	Doente
3217	Maria	18	F	67	Pequenas	39,5	4	MG	Doente
4039	Luiz	49	M	92	Grandes	38,0	2	RS	Saudável
1920	José	18	M	43	Grandes	38,5	20	MG	Doente
4340	Cláudia	21	F	52	Médias	37,6	1	PE	Saudável
2301	Ana	22	F	72	Pequenas	38,0	3	RJ	Doente
1322	Marta	19	F	87	Grandes	39,0	6	AM	Doente
3027	Paulo	34	M	67	Médias	38,4	2	GO	Saudável

Desvio\_padrão= 10,8

Desvio\_padrão = 6,3

# Outras medidas

- Pesquisar sobre
  - Desvio médio absoluto
  - Desvio mediano absoluto
  - Intervalo interquartil