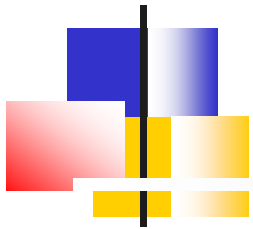


SCC-275 - Ciência de Dados

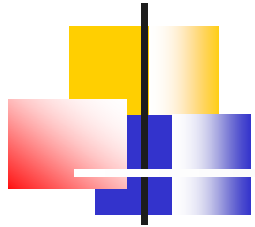
Exploração de dados



Profa. Roseli Ap. Francelin
Romero – SCC

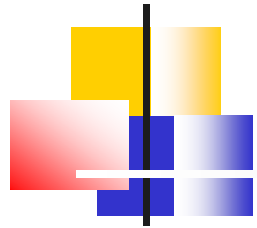
Prof. Dr. André C. P. L. F. de Carvalho
Dr. Isvani Frias-Blanco
ICMC-USP





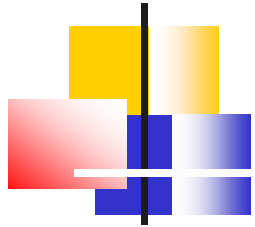
Tópicos

- Dados
- Caracterização de dados
 - Objetos e atributos
 - Tipos de dados
- Exploração de dados
 - Dados univariados
 - Dados multivariados
 - Visualização



Conjuntos de dados

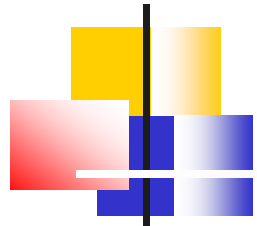
- Estruturados
 - Mais facilmente analisados por técnicas de MD
 - Ex.: Planilhas e tabelas atributo-valor
- Não estruturados
 - Mais facilmente analisados por seres humanos
 - Em DM são geralmente convertidos em dados estruturados
 - Ex.: Sequência de DNA, conteúdo de página na web, emails, vídeos, ...



Conjuntos de dados estruturados

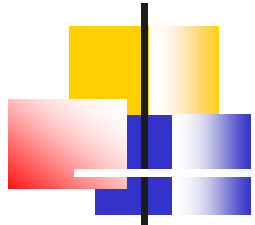
Atributos de entrada (preditivos)

Exemplos (objetos, instâncias)	Atributos de entrada (preditivos)					Diagnóstico
	Nome	Temp.	Idade	Peso	Altura	
	João	37	70	94	190	Saudável
	Maria	38	65	60	172	Doente
	José	39	19	70	185	Doente
	Sílvia	38	25	65	160	Saudável
	Pedro	37	70	90	168	Doente
						Atributo alvo



Tipos de atributos

- Simbólicos ou qualitativos
 - Nominal ou categórico
 - Ex.: cor, código de identificação, profissão
 - Ordinal
 - Ex.: gosto (ruim, médio, bom), dias da semana
- Numéricos, contínuos ou quantitativos
 - Intervalar
 - Ex.: data, temperatura em Celsius
 - Racional
 - Ex.: peso, tamanho, idade

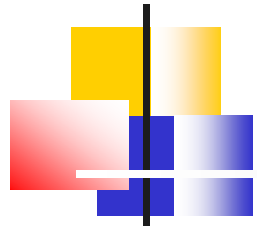


Exemplo

Nome	Temp	Enjão	Batimento	Dor	Salário	Diagnóstico
João	37.7	sim	baixo	sim	1000	doente
Pedro	37.0	não	normal	não	1100	saudável
Maria	38.2	sim	elevado	não	600	saudável
José	39.0	não	baixo	sim	2000	doente
Ana	37.3	não	elevado	sim	1800	saudável
Leila	37.7	não	elevado	sim	900	doente

Nominal Intervalar Ordinal Racional

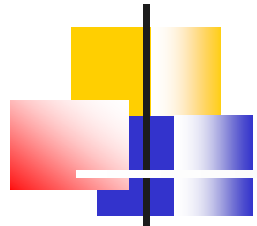
↑ ↙ ↗ ↗



Tipos de atributos

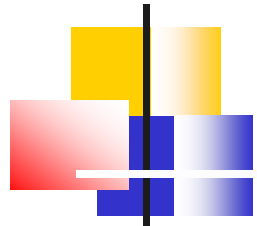
- Nominal ($=$, \neq)
 - Valores são apenas nomes diferentes
- Ordinal ($<$, $>$)
 - Existe uma relação de ordem entre valores
- Intervalar ($+$, $-$)
 - Diferença entre valores faz sentido
- Racional ($*$, $/$)
 - Razão e diferença entre valores fazem sentido





Exercício

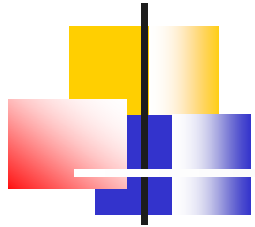
- Definir o tipo dos seguintes atributos:
 - Número de palavras de um texto
 - Fotografia
 - Número de RG
 - Data de nascimento
 - Código de disciplina
 - Posição em uma corrida
 - Expressão de um gene em um tecido
 - Sequência de aminoácidos



Tipos de atributos

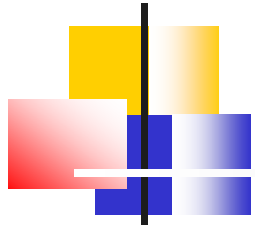
- Nominal ($=$, \neq)
 - Valores são apenas nomes diferentes
- Ordinal ($<$, $>$)
 - Existe uma relação de ordem entre valores
- Intervalar ($+$, $-$)
 - Diferença entre valores faz sentido
- Racional ($*$, $/$)
 - Razão e diferença entre valores fazem sentido





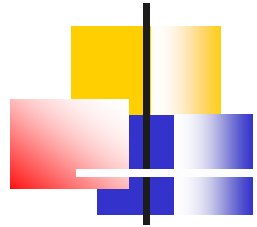
Exercício

- Definir o tipo dos seguintes atributos:
 - Número de palavras de um texto
 - Fotografia
 - Número de RG
 - Data de nascimento
 - Código de disciplina
 - Posição em uma corrida
 - Expressão de um gene em um tecido
 - Sequência de aminoácidos



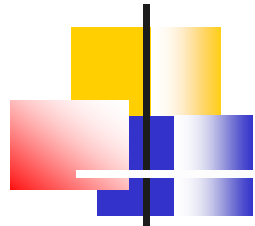
Quantidade de valores

- Atributos também se distinguem pela quantidade de valores
 - Discretos
 - Número finito ou infinito e enumerável de valores, como números naturais
 - Ex.: código postal, contagem (quantidade de algum elemento)
 - Caso especial: valores binários
 - Contínuos
 - Número infinito de valores, como números reais
 - Ex.: temperatura, peso, distância



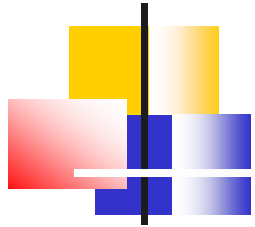
Exploração de dados

- Exploração preliminar dos dados facilita entendimento de suas características
- Principais motivações:
 - Ajudar a selecionar a melhor técnica para pré-processamento e/ou modelagem
- Ferramentas
 - Estatística descritiva
 - Visualização



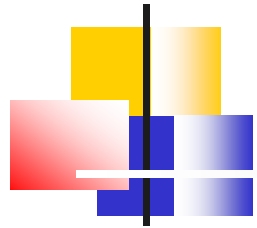
Estatística descritiva

- Descreve propriedades estatísticas de dados
- Produz valores que resumem características de um conjunto de dados
 - Na maioria das vezes por meio de cálculos muito simples



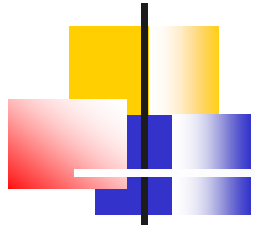
Estatística descritiva

- Pode capturar medidas de:
 - Frequência
 - Localização ou tendência central
 - Ex.: Média
 - Dispersão ou espalhamento
 - Ex.: Desvio padrão
 - Distribuição ou formato



Frequência

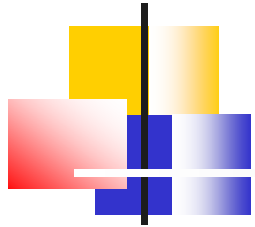
- Proporção de vezes que um atributo assume um dado valor
 - Em um determinado conjunto de dados
 - Muito usada para dados categóricos
 - Ex.: Em um BD de um hospital, 40% dos pacientes é maior de idade



Exemplo

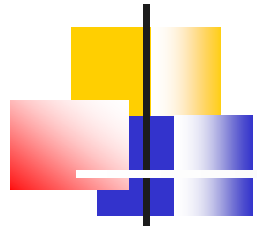
Febre	Idade	Batimento	Dor	Diagnóstico
sim	23	elevado	sim	doente
não	9	normal	não	saudável
sim	61	elevado	não	saudável
sim	32	baixo	sim	doente
sim	21	elevado	sim	saudável
não	48	elevado	sim	doente

66% das medidas de batimento cardíaco encontradas em pacientes são superiores ao normal



Medidas de localidade (centralidade)

- Tendência central
- Valores quantitativos
 - Média
 - Mediana
 - Percentil
- Valores qualitativos ou quantitativos
 - Moda

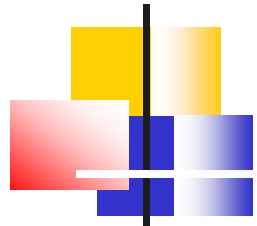


Média

- Pode ser calculada facilmente

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

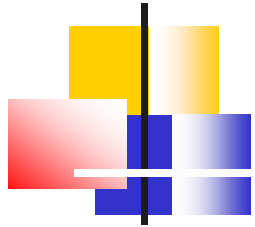
- Problema: sensível a *outliers*



Mediana

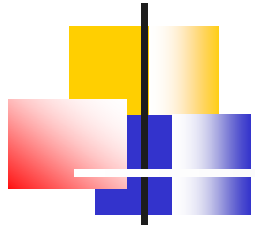
- Menos sensível a *outliers* que média
- Necessário ordenar valores

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



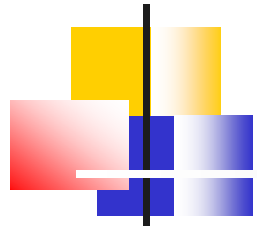
Média *versus* Mediana

- Média é uma boa medida de localização quando os valores estão distribuídos simetricamente
- Mediana indica melhor o centro
 - Se distribuição é oblíqua (assimétrica)
 - *Skewed*
 - Se existem *outliers*



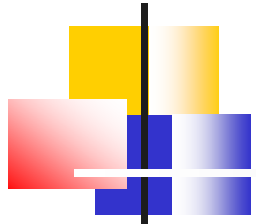
Média Podada

- *Trimmed mean*
- Melhora estimativa da média descartando exemplos nos extremos
 - Define porcentagem p dos exemplos a serem eliminados
 - Ordena os dados
 - Elimina $(p/2)\%$ dos exemplos em cada extremidade



Winsorização

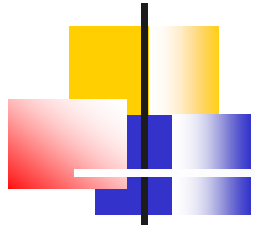
- *Winsorization*
- Semelhante à média podada
 - Valores que passam dos extremos, ao invés de eliminados, são substituídos pelos extremos permitidos
 - Percentis mínimos e máximos
 - Ex.: Winsorização de 80% para os valores
1, 2, 3, 4, 5, 6, 7, 8, 9, 10 =
2, 2, 3, 4, 5, 6, 7, 8, 9, 9



Moda

- Valor mais frequente nos dados
 - Nenhuma moda: Todos os valores são iguais
 - Uma moda: Unimodal
 - Mais de uma moda: Multimodal (Bimodal, Trimodal, ...)
- Indicada quando existem poucos possíveis valores
- Para dados moderadamente assimétricos, moda pode ser estimada por média e mediana

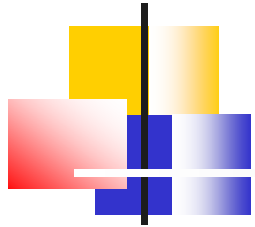
$$moda \approx média - 3 \times (média - mediana)$$



Exemplo

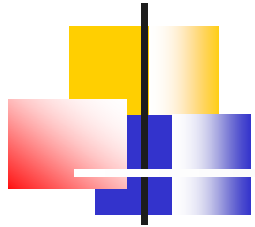
Febre	Idade	Batimento	Dor	Diagnóstico
sim	23	elevado	sim	doente
não	9	normal	não	saudável
sim	61	elevado	não	saudável
sim	32	baixo	sim	doente
sim	23	elevado	sim	saudável
não	48	elevado	sim	doente

Valor da moda para o atributo batimento: elevado
Valor da moda para o atributo idade: 23



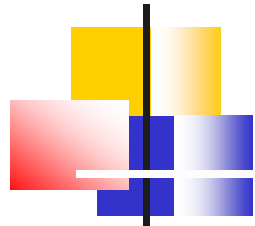
Exercício

- Dado o conjunto de dados $\{2, 2, 3, 4, 5, 80\}$, calcular:
 - Média
 - Mediana
 - Média podada com $p = 33\%$
 - Média com Winsorização de 10%
 - Moda



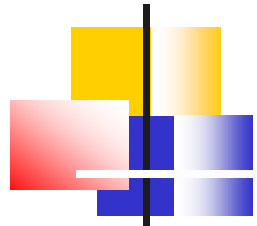
Quartis e Percentis

- Mediana divide os dados ao meio
 - No entanto, pontos de localização diferentes podem ser usados
 - Quartis dividem um conjunto ordenado de dados em quartos
 - Q_1 : Primeiro quartil (quartil inferior)
 - Valor da observação para a qual 25% dos dados do conjunto tem valor menor ou igual
 - Também é o valor do 25º percentil
 - Q_2 : Segundo quartil = mediana
 - Q_3 : Terceiro quartil (quartil superior)
 - 75º percentil



Percentis

- Características do valor do $100p^o$ percentil:
 - Pelo menos $100xp\%$ das observações possuem um valor menor ou igual a ele
 - Pelo menos $100x(1-p)\%$ das observações tem um valor igual ou acima
- Mediana é o $100x0,5^o$ ou 50^o percentil
 - Para cálculo, usar fórmula da mediana

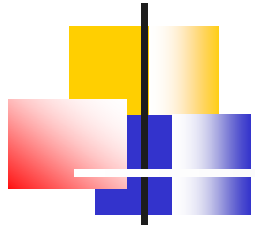


Cálculo dos percentis

- Ordenar os valores
 - Posição do p-percentil:

$$posição = \left\lceil p \times n + \frac{1}{2} \right\rceil$$

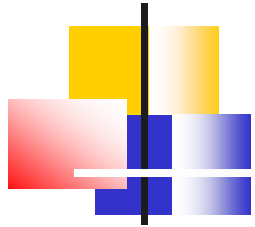
- Arredonda posição para o valor inteiro seguinte (21,5 = 22)
- Retorna o valor nessa posição



Exemplo

- Obter os quartis e o 95º percentil para o conjunto de dados abaixo:

6,2	7,67	8,3	9,0	9,4
9,8	10,5	10,7	11,0	12,3



Exemplo

- Obter os quartis e o 95º percentil para o conjunto de dados abaixo:

6,2	7,67	8,3	9,0	9,4
9,8	10,5	10,7	11,0	12,3

$$Q_1: np = 0,25 \times 10 + 0,5 = 3$$

usar o terceiro valor: $Q_1 = 8,3$

$$Q_2: np = 0,5 \times 10 + 0,5 = 5,5$$

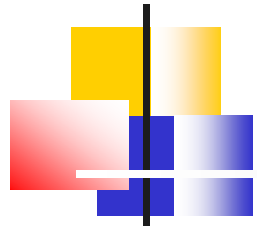
para a mediana, usar a média entre o quinto e o sexto valor: $Q_2 = 9,6$

$$Q_3: np = 0,75 \times 10 + 0,5 = 8$$

usar o oitavo valor: $Q_3 = 10,7$

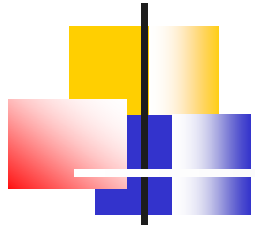
$$P_{0,95}: np = 0,95 \times 10 + 0,5 = 10$$

usar o décimo valor: $P_{0,95} = 12,3$



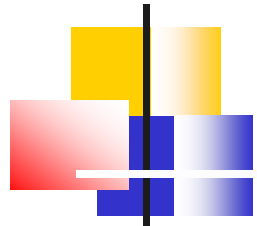
Exercício

- Calcular quartis inferior e superior e o 60º percentil para os valores
 - 16, 25, 4, 18, 11, 13, 20, 8, 11 e 9



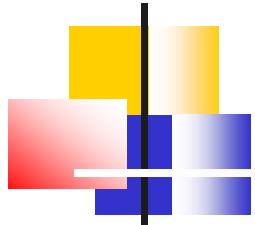
Exercício

- Calcular quartis inferior e superior e o 60º percentil para os valores
 - 16, 25, 4, 18, 11, 13, 20, 8, 11 e 9
 - 4, 8, 9, 11, 11, 13, 16, 18, 20, 25
 - $Q_1 =$
 - $Q_3 =$
 - 60º percentil =



Exercício

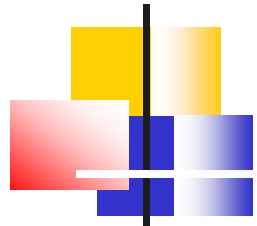
- Calcular a mediana, o primeiro quartil e o segundo quartil
 - 23, 7, 12, 6, 10
 - 23, 7, 12, 6, 10, 7, 10
 - 1, 1, 1, 1, 1, 1, 98



Exercício

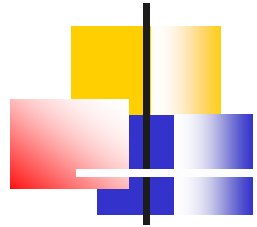
- Representar os dados a seguir usando faces de Chernoff

Febre	Idade	Batimento	Dor	Diagnóstico
sim	23	elevado	sim	doente
não	9	baixo	não	saudável
sim	61	elevado	não	saudável
sim	32	baixo	sim	doente
sim	21	elevado	sim	saudável
não	48	elevado	sim	doente



Considerações Finais

- Caracterização de dados
 - Objetos e atributos
 - Tipos de dados
- Exploração de dados
 - Dados univariados
 - Medidas de localidade, espalhamento e distribuição
 - Dados multivariados
 - Visualização de dados e de resultados



Exercício

- Descrever e explorar os dados utilizados na aula de laboratório



Perguntas

