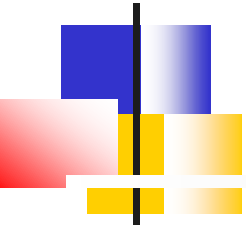


Ciência de Dados

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)

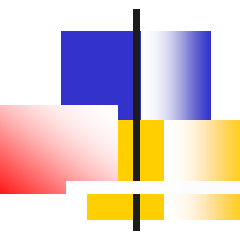
Nome Temp. Idade Peso Altura Diagnóstico

Exemplos
(objetos,
instâncias)

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

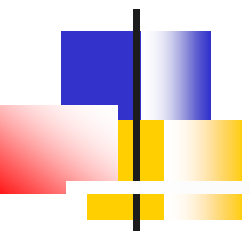
Nome	Temp	Enjão	Batimento	Dor	Salário	Diagnóstico
João	37.7	sim	baixo	sim	1000	doente
Pedro	37.0	não	normal	não	1100	saudável
Maria	38.2	sim	elevado	não	600	saudável
José	39.0	não	baixo	sim	2000	doente
Ana	37.3	não	elevado	sim	1800	saudável
Leila	37.7	não	elevado	sim	900	doente

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

Quantidade de valores

- Atributos também se distinguem pela quantidade de valores

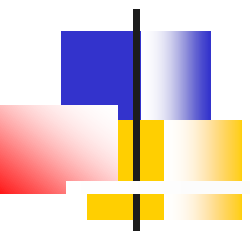
- Discretos

- Número finito ou infinito e enumerável de valores, como números naturais
 - Ex.: código postal, contagem (quantidade de algum elemento)
- Caso especial: valores binários

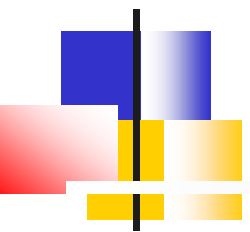
- Contínuos

- Número infinito de valores, como números reais
 - Ex.: temperatura, peso, distância

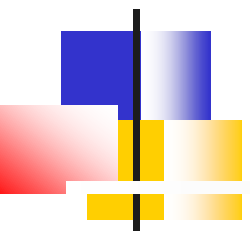
Exploração de dados

- 
- 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

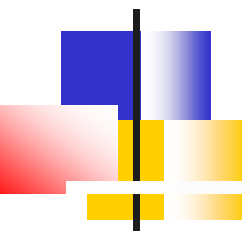
Estatística descritiva

- 
- 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
-

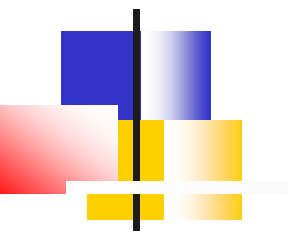
Estatística descritiva

- 
- 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

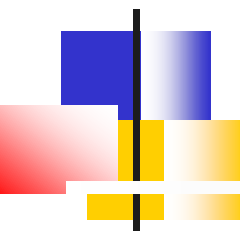
Exemplo



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

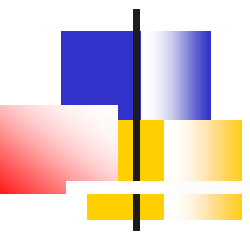
66% das medidas de batimento cardíaco encontradas em pacientes são superiores ao normal

Medidas de localidade (centralidade)

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

Média

- Pode ser calculada facilmente


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

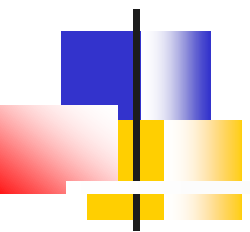
- Problema: sensível a *outliers*

Mediana

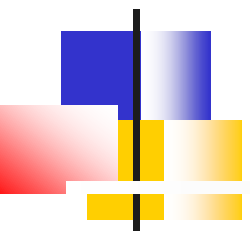
- Menos sensível a *outliers* que média
 - Necessário ordenar valores
-

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

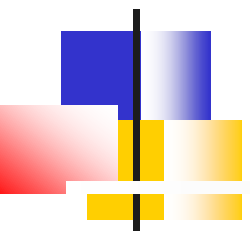
Média *versus* Mediana

- 
- 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*
-

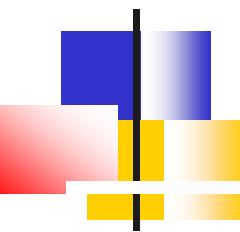
Média Podada

- 
- *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

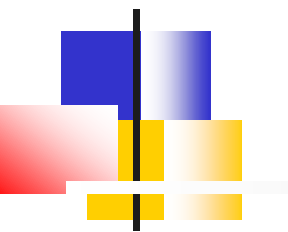
- 
- *Winsorization*
 - Semelhante à média podada
 - Valores que passam dos extremos, ao invés de eliminados, são substituídos pelos extremos permitidos
 - Percentis mínimos e máximos
 - Ex.: Winsorização de 80% para os valores
1, 2, 3, 4, 5, 6, 7, 8, 9, 10 =
2, 2, 3, 4, 5, 6, 7, 8, 9, 9

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

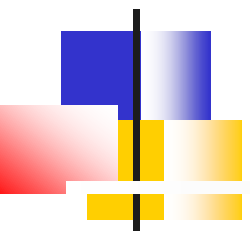


Febre	Idade	Batimento	Dor	Diagnóstico
sim	23	elevado	sim	doente
não	9	normal	não	saudável
sim	61	elevado	não	saudável
sim	32	baixo	sim	doente
sim	23	elevado	sim	saudável
não	48	elevado	sim	doente

Valor da moda para o atributo batimento: elevado

Valor da moda para o atributo idade: 23

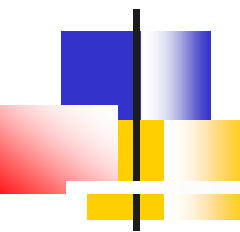
Exercício

- 
- 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

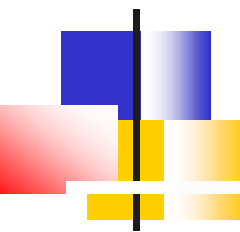
- Obter os quartis e o 95º percentil para o conjunto de dados abaixo:



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: $Q_3 = 10,7$

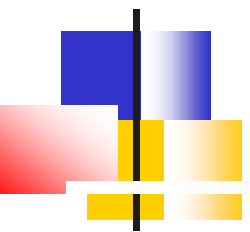
$$P_{0,95}: np = 0,95 \times 10 + 0,5 = 10$$

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

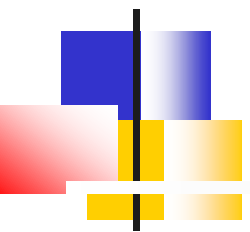
Exercício

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

Exercício

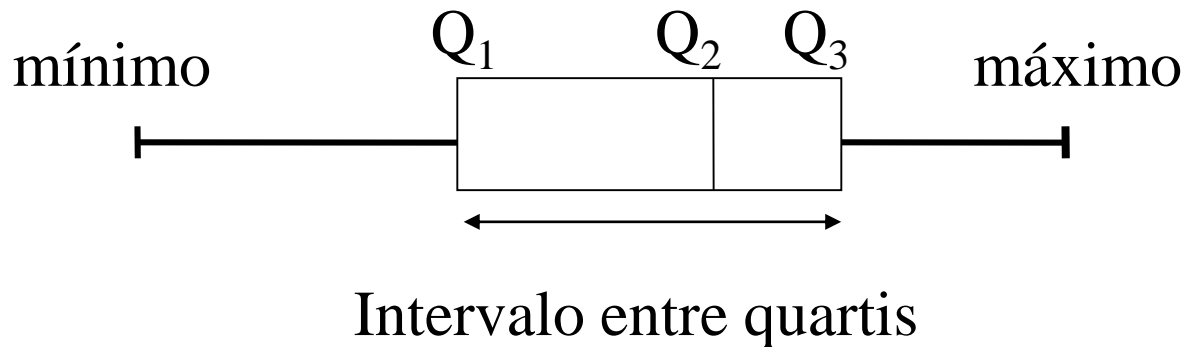
- 
- Calcular quartis inferior e superior e o 60º percentil para os valores
 - 16, 25, 4, 18, 11, 13, 20, 8, 11 e 9
 - 4, 8, 9, 11, 11, 13, 16, 18, 20, 25
 - $Q_1 =$
 - $Q_3 =$
 - 60º percentil =

Exercício

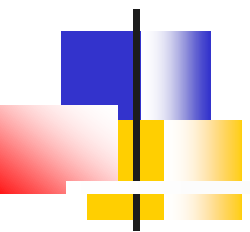
- 
- Calcular a mediana, o primeiro quartil e o segundo quartil
 - 23, 7, 12, 6, 10
 - 23, 7, 12, 6, 10, 7, 10
 - 1, 1, 1, 1, 1, 1, 98
-

Boxplot

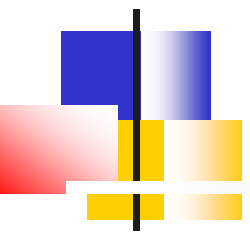
- Gráfico que resume informações dos quartis



Boxplot modificado

- 
- Identifica *outliers* e reduz seu efeito no formato do boxplot
 - Tolerância = $1,5 \times \text{intervalo entre quartis}$
 - Verificar se $\text{máximo} - Q_3$ ($Q_1 - \text{mínimo}$) $>$ tolerância
 - Valor fora do intervalo é considerado *outlier*
 - Define novo mínimo e/ou máximo

INTERQUARTIL (IQR)

- 
- $IQR = Q3 - Q1$
 - Representa 50% dos dados do conjunto
 - Ajuda a encontrar outliers

Cerca Inferior(LF) = $Q1 - 1.5 IQR$

Cerca Superior(UF) = $Q3 + 1.5 IQR$

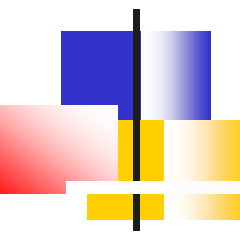
INTERQUARTIL (IQR)

- Exercício: Considere a lista:

(10 12 23 23 25 35 37 45 46 55 56 67 70)

- Montar o BoxPlot correspondente e determinar LF e UF.

Medidas de espalhamento

- 
- Medem variabilidade, dispersão ou espalhamento de um conjunto de valores
 - Indicam se os dados estão:
 - Amplamente espalhados ou
 - Relativamente concentrados em torno de um ponto (ex. média)
 - Medidas comuns
 - Intervalo ou amplitude
 - Variância
 - Desvio padrão

Intervalo

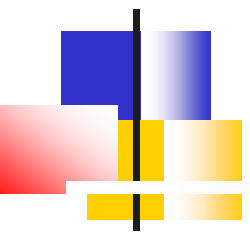
- Medida mais simples
 - Mostra espalhamento máximo
 - Usada em controle de qualidade
- Sejam $\{x_1, \dots, x_n\}$ n valores para um atributo x

$$x_{(1)} = \min\{x_1, \dots, x_n\}$$

- Pode não ser uma boa medida
 - Maioria dos valores próximos de um ponto e poucos valores próximos aos extremos

Variância

- Medida mais utilizada para analisar espalhamento de valores


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

- Denominador $n-1$: correção de Bessel, usada para uma melhor estimativa da variância verdadeira
 - Amostra (estimada) e população (verdadeira)
- Desvio padrão: raiz quadrada da variância
- Um dos momentos de uma distribuição de probabilidade

VARIÂNCIA

- "o quão longe" em geral os seus valores se encontram do valor esperado (média) da variável aleatória X .
- Desvio Padrão indica qual é o "erro" se quiséssemos substituir um dos valores coletados pelo **valor da média**.

Funcionários	Quantidade de peças produzidas por dia				
	Segunda	Terça	Quarta	Quinta	Sexta
A	10	9	11	12	8
B	15	12	16	10	11
C	11	10	8	11	12
D	8	12	15	9	11

Funcionários	Média Aritmética (\bar{x})	
A	$\bar{X}_A = \frac{10 + 9 + 11 + 12 + 8}{5} = \frac{50}{5}$	$\bar{X}_A = 10,0$
B	$\bar{X}_B = \frac{15 + 12 + 16 + 10 + 11}{5} = \frac{64}{5}$	$\bar{X}_B = 12,8$
C	$\bar{X}_C = \frac{11 + 10 + 8 + 11 + 12}{5} = \frac{52}{5}$	$\bar{X}_C = 10,4$
D	$\bar{X}_D = \frac{8 + 12 + 15 + 9 + 11}{5} = \frac{55}{5}$	$\bar{X}_D = 11,0$

Variância → Funcionário A:

$$\text{var (A)} = \frac{(10 - 10)^2 + (9 - 10)^2 + (11 - 10)^2 + (12 - 10)^2 + (8 - 10)^2}{5}$$


$$\text{var (A)} = \frac{10}{5} = 2,0$$

$$\text{Var(B)} = 5,36$$

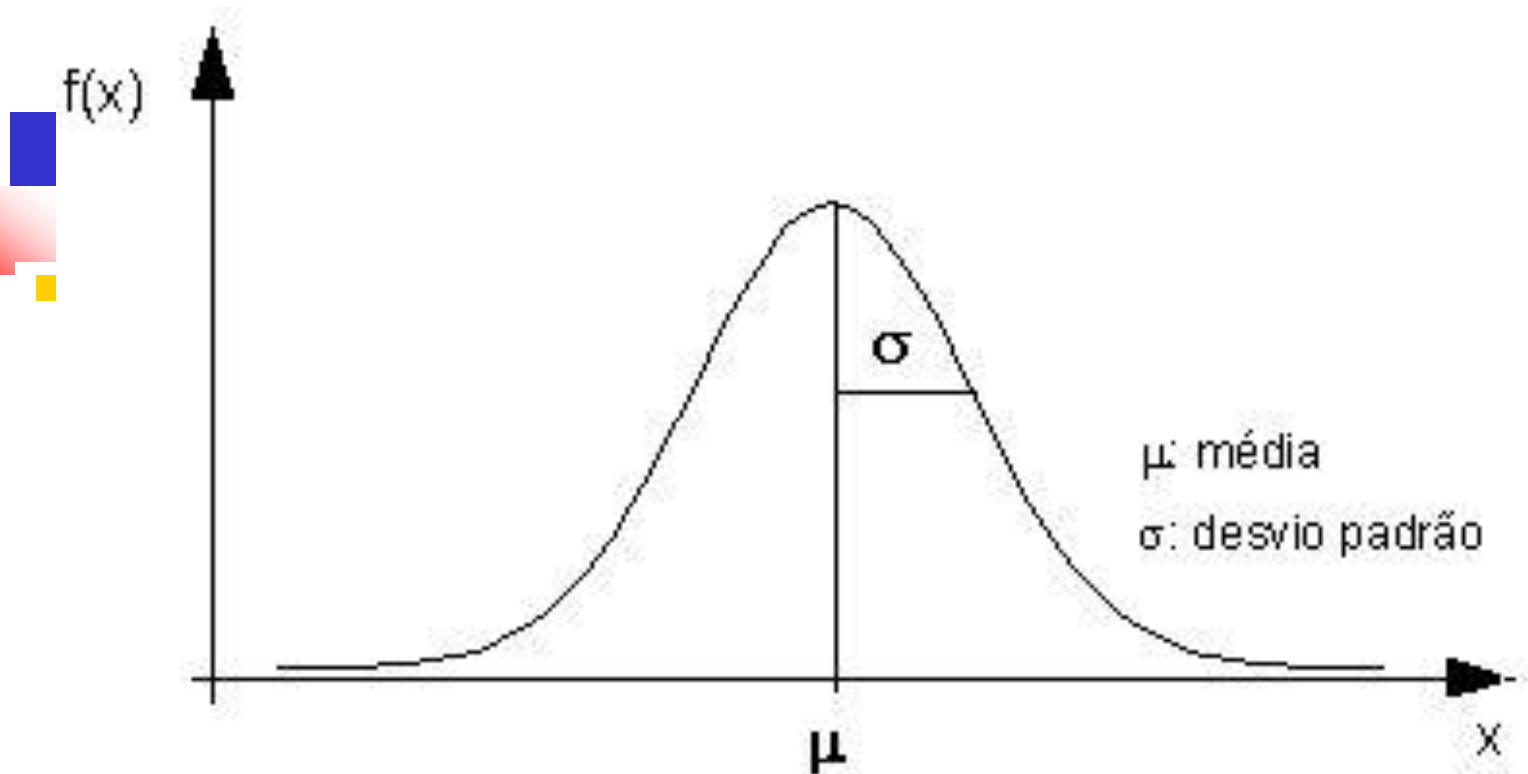
$$\text{Var(C)} = 1,84$$

$$\text{Var(D)} = 6,0$$

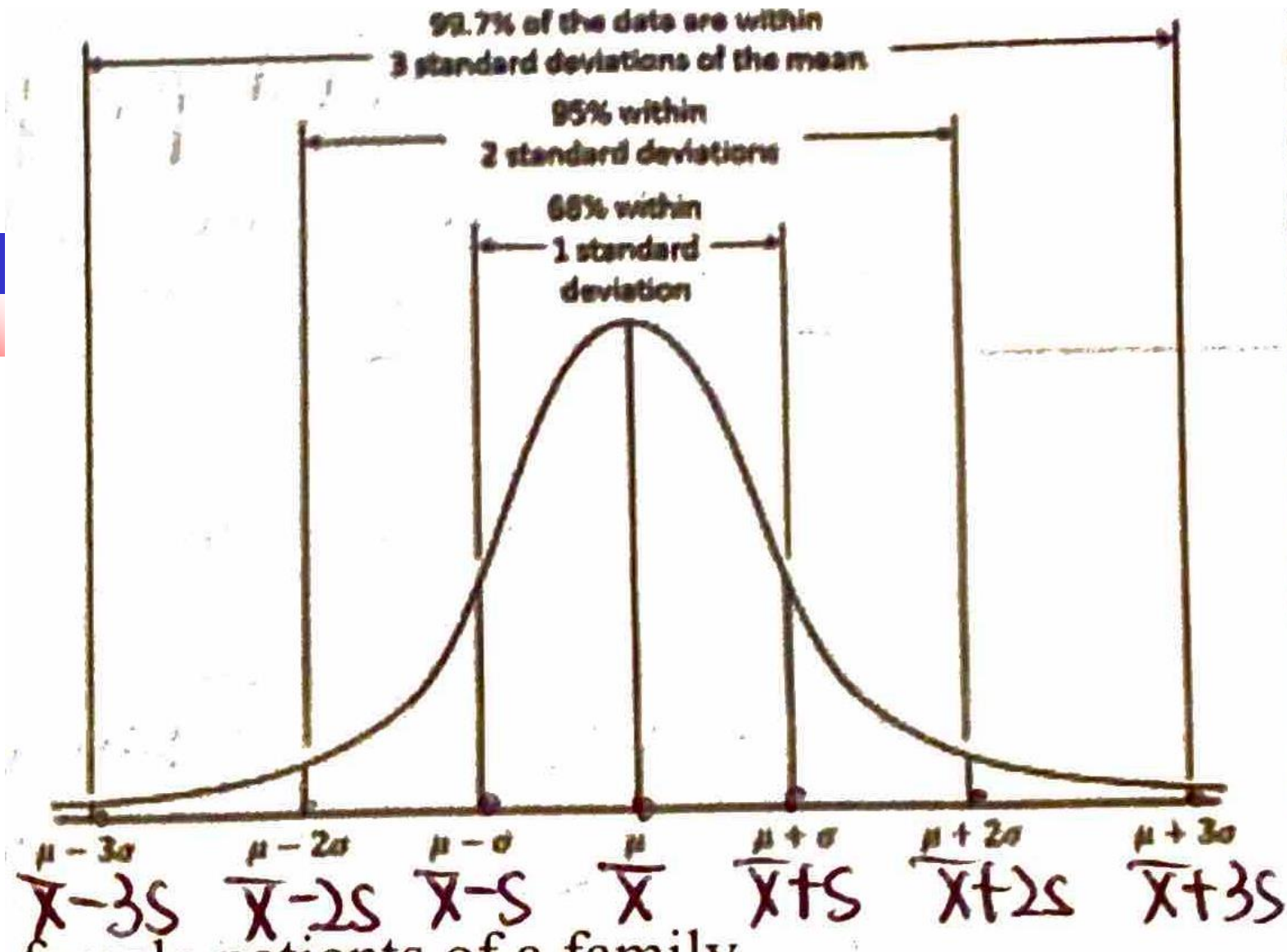
Variância e Desvio Padrão

- 
- **$dp(A) \approx 1,41$**
 - **$dp(B) \approx 2,32$**
 - **$dp(C) \approx 1,36$**
 - **$dp(D) \approx 2,45$**
 - **Funcionário A: $10,0 \pm 1,41$ peças por dia**
Funcionário B: $12,8 \pm 2,32$ peças por dia
Funcionário C: $10,4 \pm 1,36$ peças por dia
Funcionário D: $11,0 \pm 2,45$ peças por dia
-

Distribuição normal



Normal – Regra Empírica



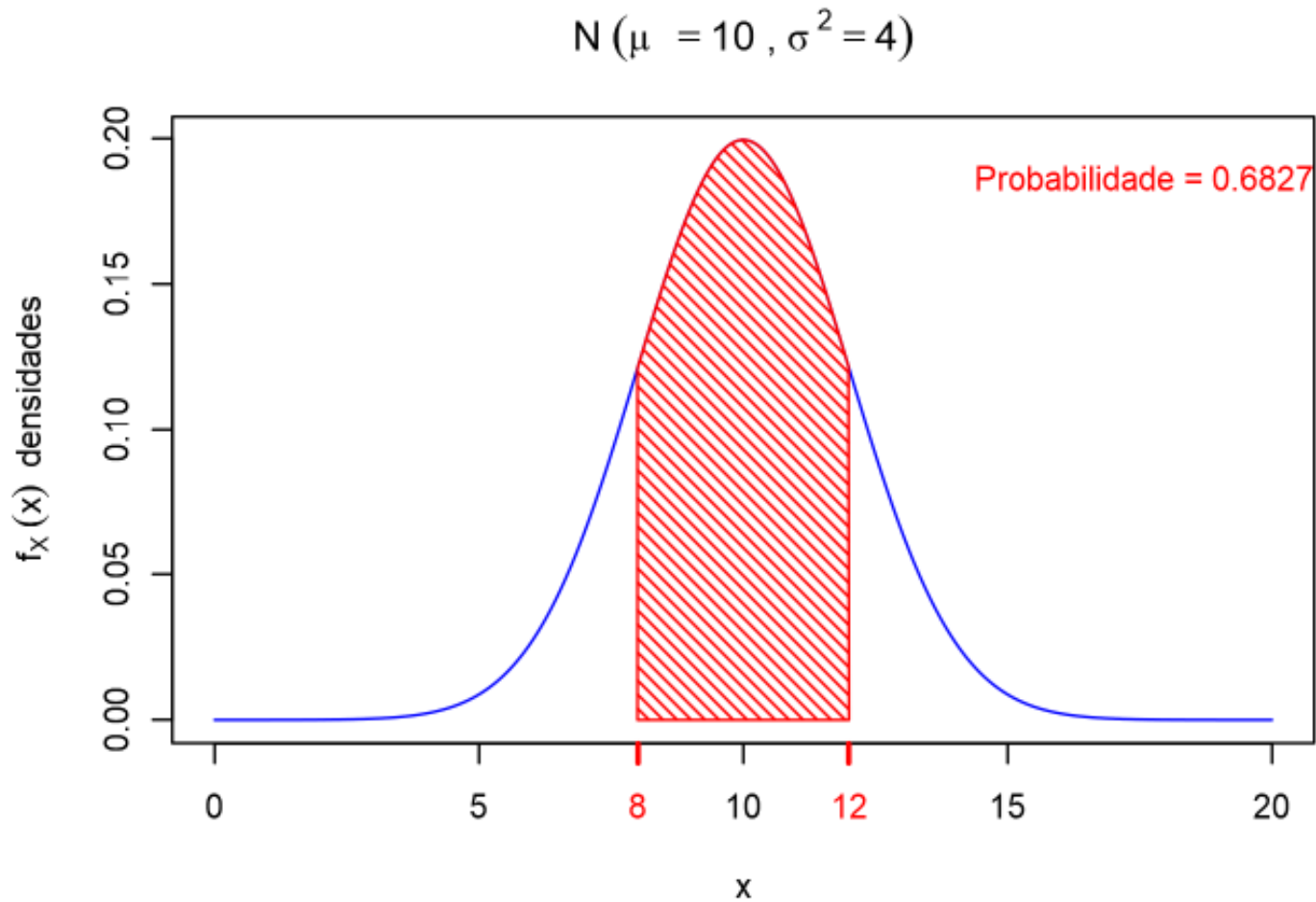
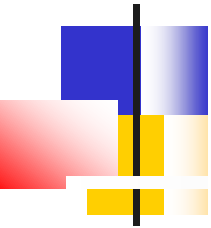
Interessante na distr. Normal


$$P(\mu - \sigma < X < \mu + \sigma) = 0.68$$

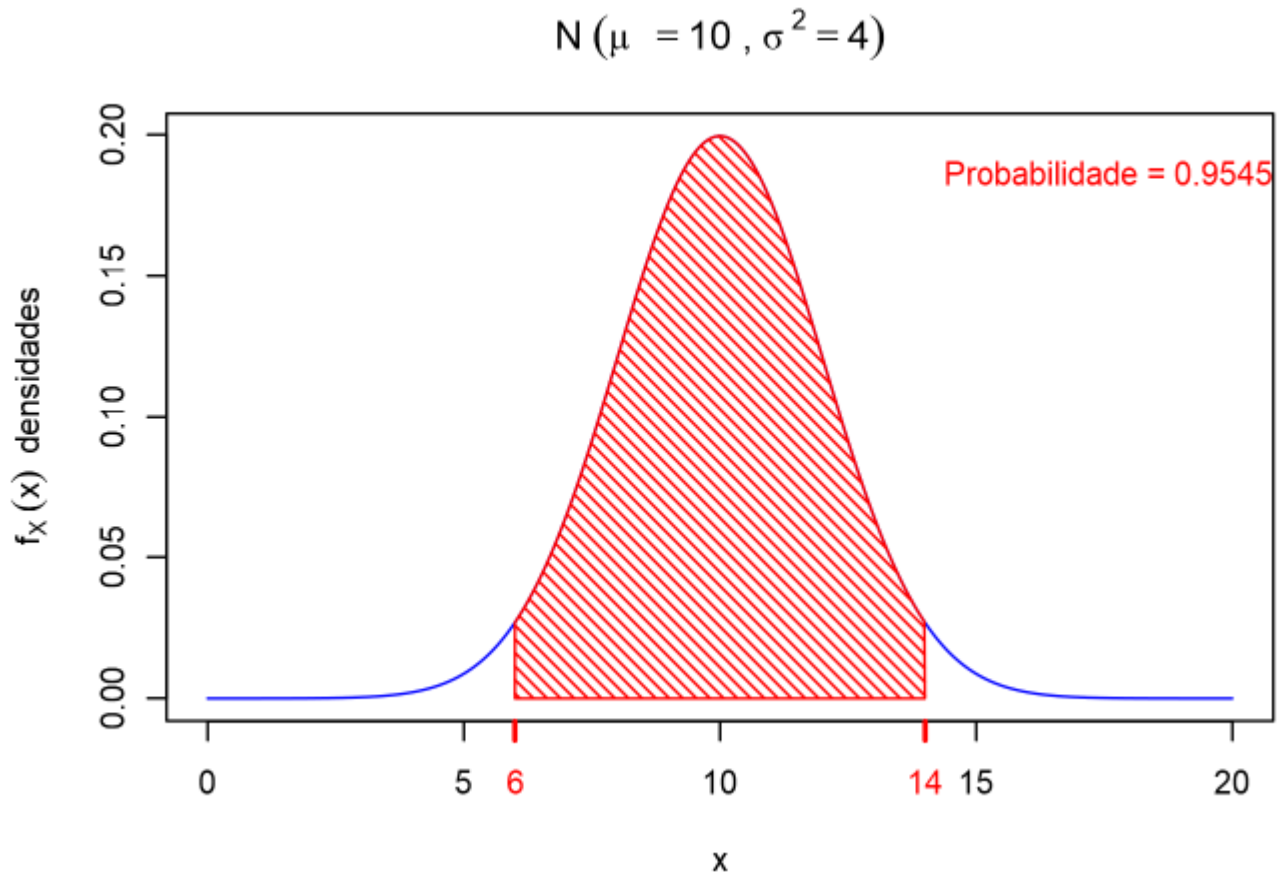
$$P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.95$$

$$P(\mu - 3\sigma < X < \mu + 3\sigma) = 0.99$$

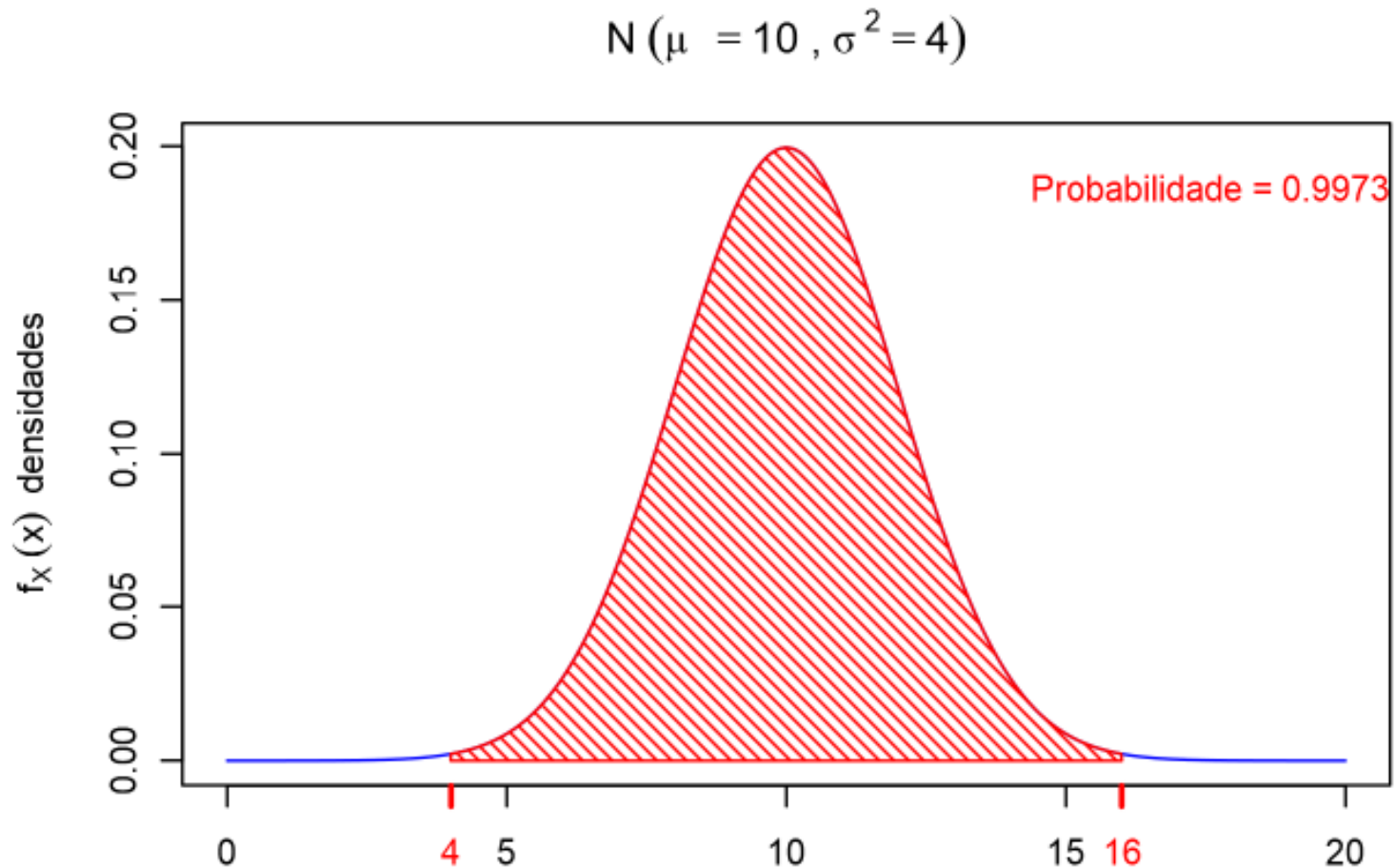
Exemplos de distr. Normal



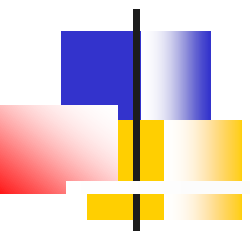
Exemplos de distr. Normal



Exemplos de distr. Normal



Medidas de distribuição

- 
- Definem como os valores de uma variável (atributo) estão distribuídos
 - Calculada por meio de momentos
 - Medida quantitativa usada na estatística e na mecânica
 - Captura o formato da distribuição de um conjunto de valores

Momentos

- Usados para caracterizar a distribuição de valores de variáveis aleatórias
 - Estimam medidas de uma população de valores usando uma amostra dela
- Vários cálculos de momento
 - Cálculo de momento original
 - Cálculo de momento central
 - Cálculo de momento padronizado
 - ...

Momento original

- Momento em torno da origem

$$\mu_k = E(x^k) = \sum_{i=1}^n x_i^k p(x_i) = \sum_{i=1}^n x_i^k f(x_i)$$

- Valor de k define qual é a medida de momento estimada
 - Em geral, apenas primeiro momento (k = 1) é usado: média

Momento central

- Centralizado ou centrado

- K=1: média = 0 (primeiro momento em torno da média = primeiro momento central)
- K=2: variância (segundo momento central)
- K=3: obliquidade (terceiro momento central)
- K=4: curtose (quarto momento central)

$$\mu_k = E[x - E(x)]^k = \sum_{i=1}^n (x_i - \bar{x})^k p(x_i) = \sum_{i=1}^n (x_i - \bar{x})^k f(x_i)$$

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

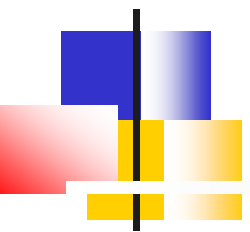
Assumindo cada x_i aparece com a mesma frequência

Momento padronizado

- Fornece informações mais claras sobre a distribuição dos dados
 - Utiliza distribuição normal padrão
 - Normaliza o k-ésimo momento pelo desvio padrão elevado a k
 - Torna a medida independente de escala

$$\mu'_k = \frac{\mu_k}{\sigma^k} \quad \text{Em torno da média}$$

Momento padronizado

- 
- Primeiro momento (K=1):
 - Média = 0
 - Segundo momento (K=2):
 - Variância = 1
-

$$\mu_2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^k}{(n-1)\sigma^2}$$

Obliquidade

- Terceiro momento (*Skewness*)
 - Mede a simetria da distribuição dos dados em torno da média
 - Distribuição simétrica tem a mesma aparência à direita e à esquerda do ponto central

$$Obl = \mu_3 = \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{(n-1)\sigma^3}$$

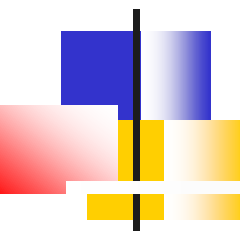
$$\mu_3 = \frac{1}{\sigma_3} \sum_{i=1}^n (x_i - \bar{x})^3 p(x_i) = \frac{1}{\sigma_3} \sum_{i=1}^n (x_i - \bar{x})^3 f(x_i)$$

Curtose

- Quarto momento (*Kurtosis*)
 - Medida de dispersão que captura o achatamento da função de distribuição
 - Verifica se os dados apresentam um pico elevado ou são achatados em relação a uma distribuição normal

$$Curt = \mu_4 = \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{(n-1)\sigma^4}$$

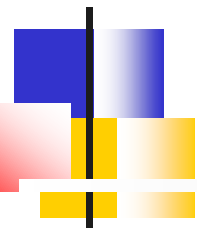
Curtose (Kurtosis)

- 
- Para uma distribuição normal padrão (média = 0 e desv. pad. = 1), $Curt = 3$
 - Para que a distribuição normal padrão tenha curtose = 0, usa-se a correção:

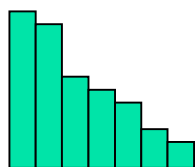
$$Curt = \mu_4 = \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{(n-1)\sigma^4} - 3$$

Obliquidade x Curtose

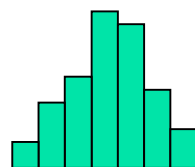
- Melhor forma para verificar graficamente curtose e obliquidade



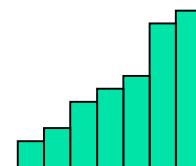
Obliquidade



Positiva

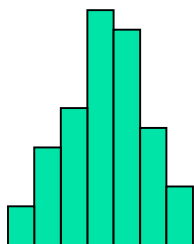


Zero (simétrica)

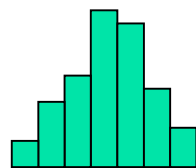


Negativa

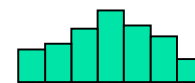
Curtose



Positiva

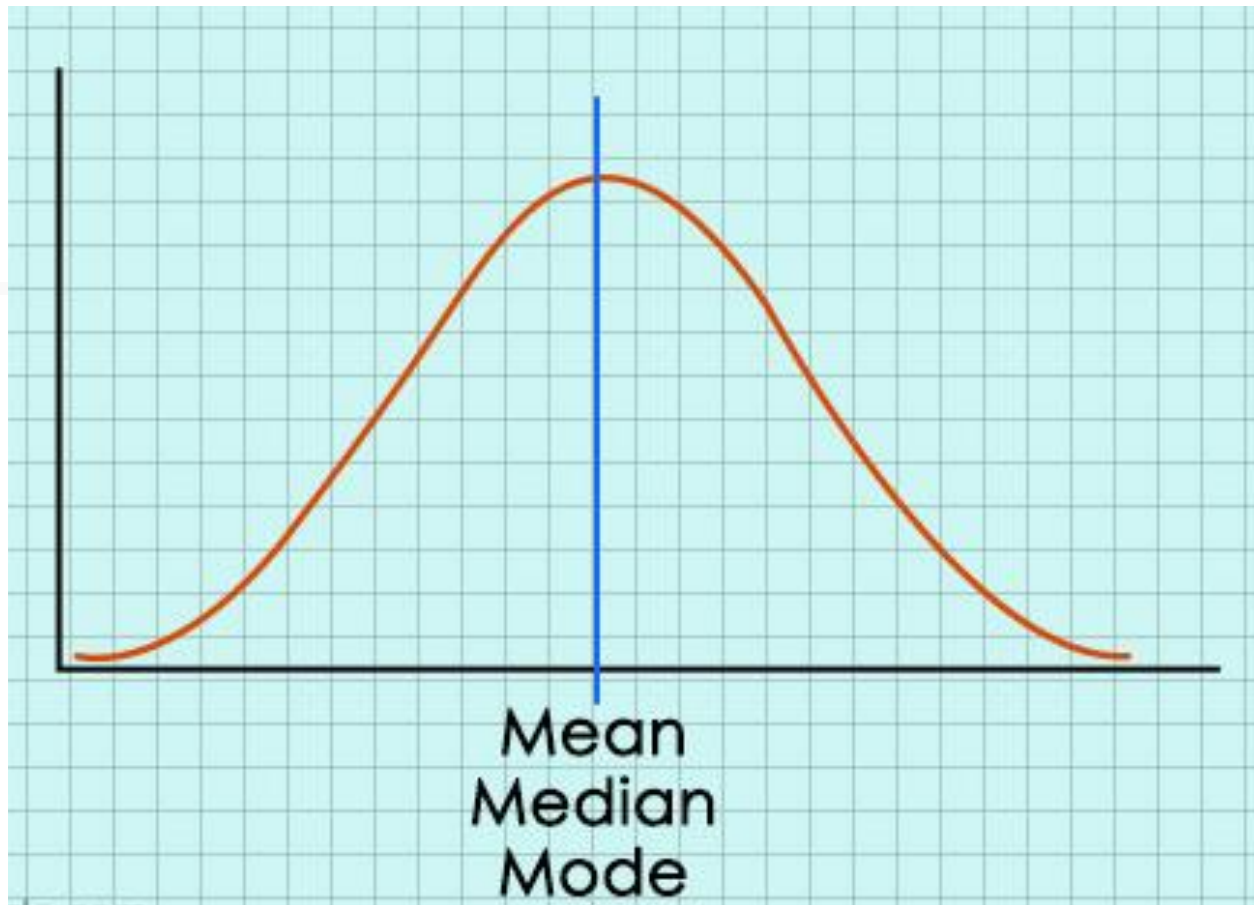


Zero (normal)

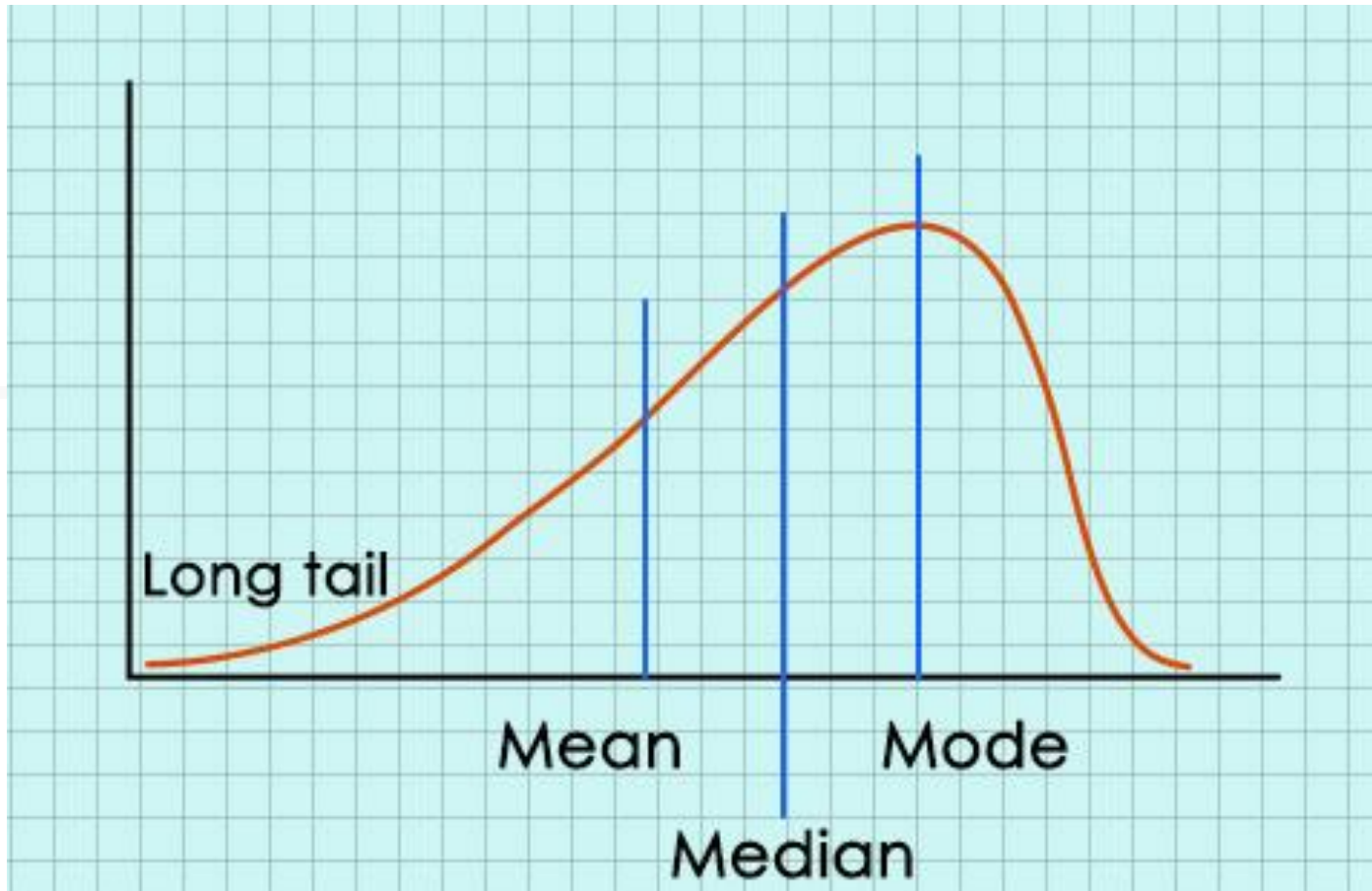


Negativa

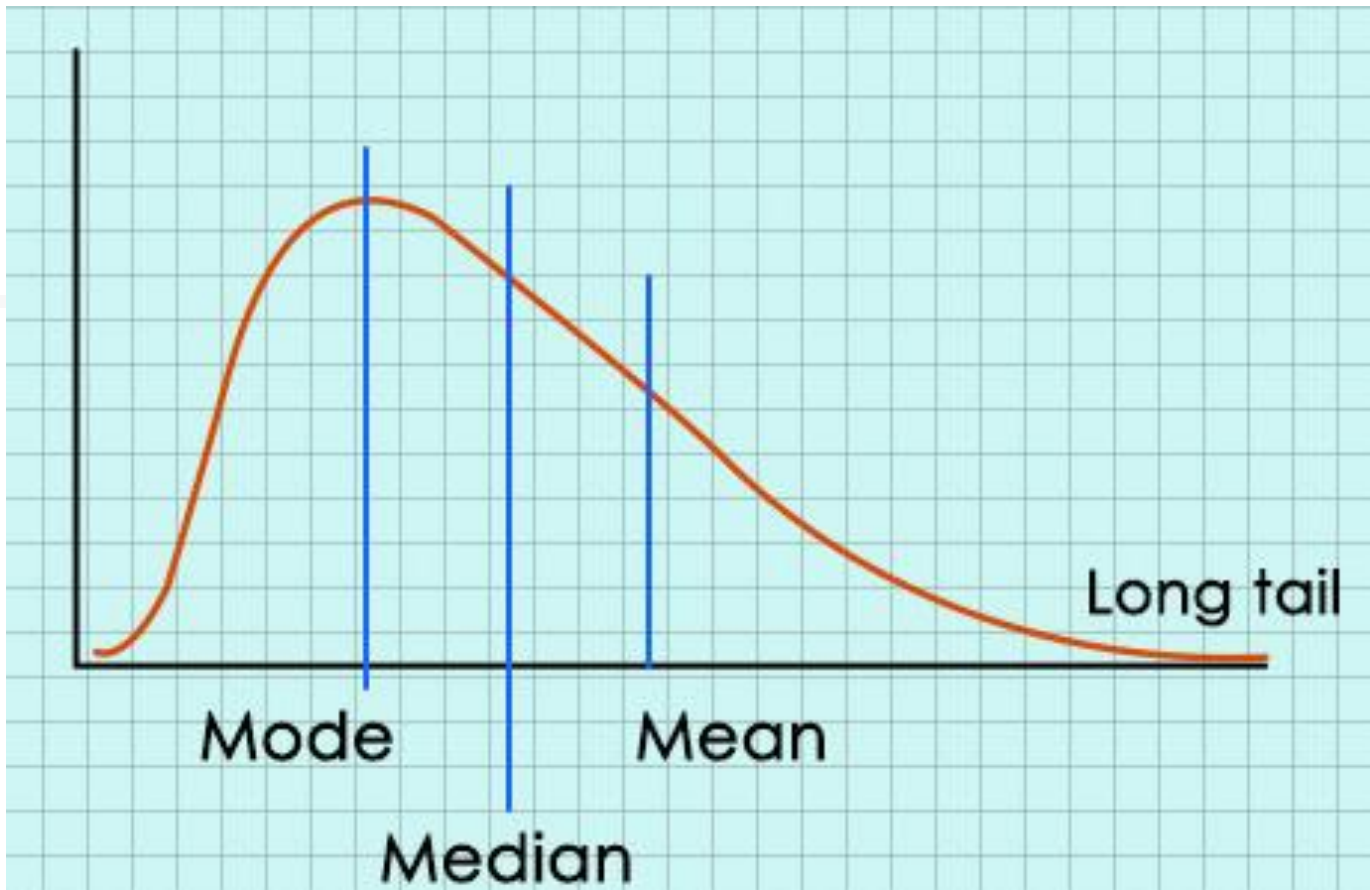
Distribuição Normal



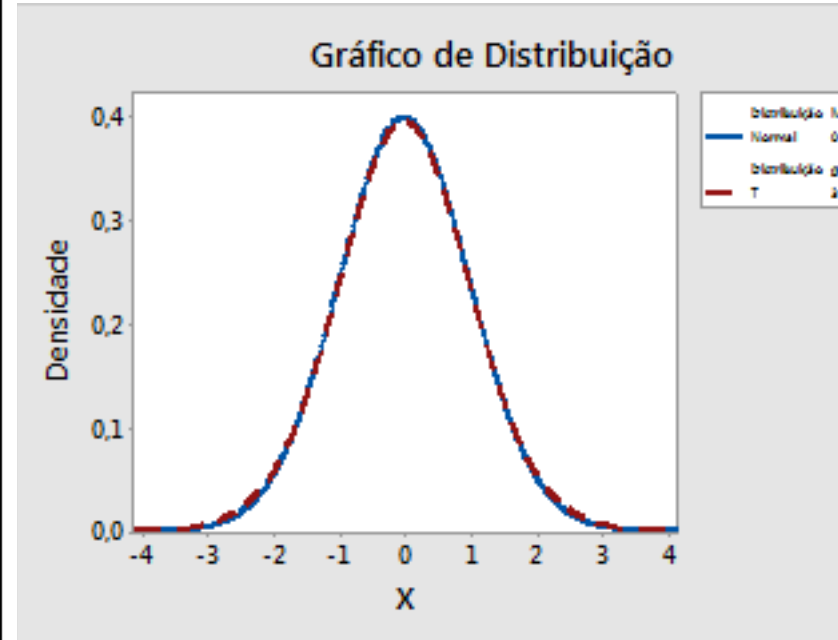
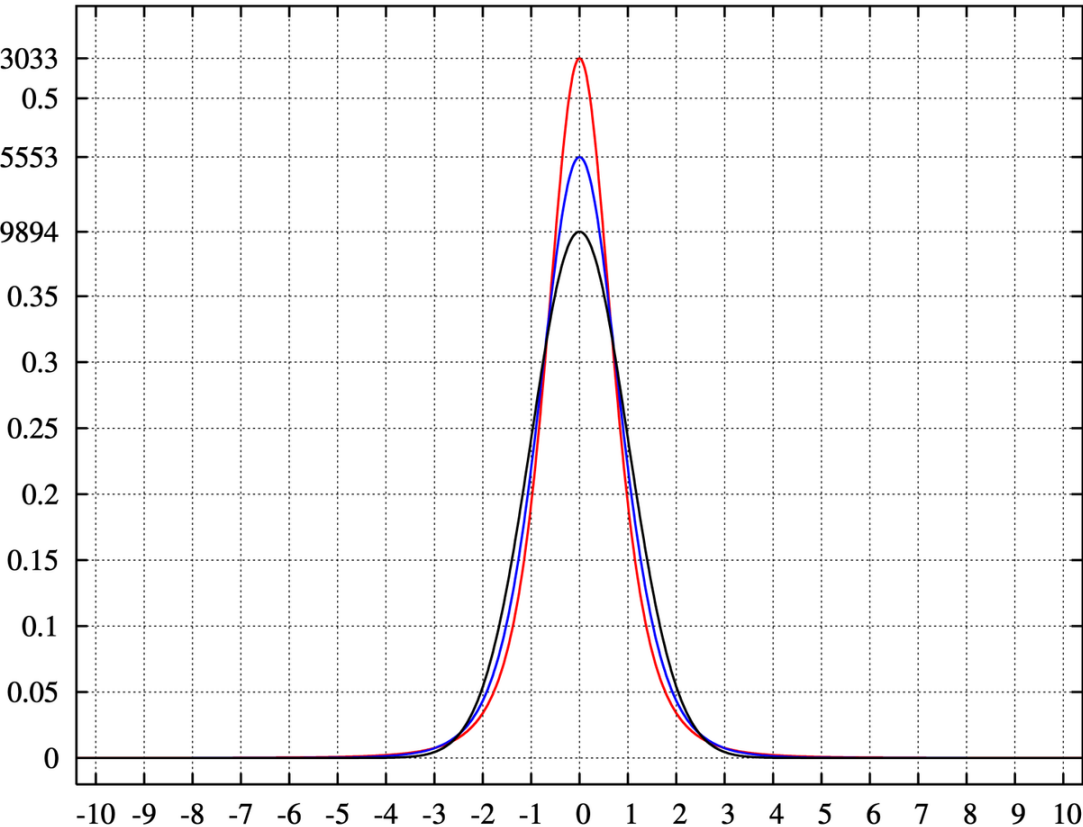
Obliquidade (negativa)



Obliquidade (Positiva)

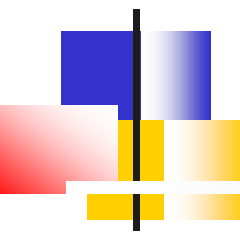


Curtose faz a diferenca



Todas tem media zero
e variância 1
São diferentes!!!

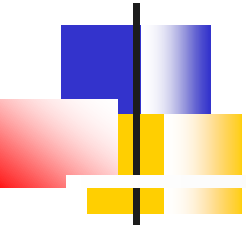
Exercício

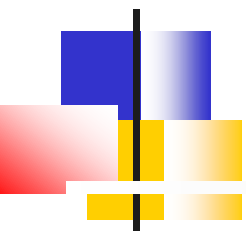
- 
- Obter o valor dos 4 primeiros momentos padronizados para os valores:
-

1, 3, 5, 6, 8, 10, 15

Exercício

- Descrever e explorar os dados utilizados na aula de laboratório
-





sklearn's OrdinalEncoder method

sklearn's StandardScaler method.

Perguntas

