

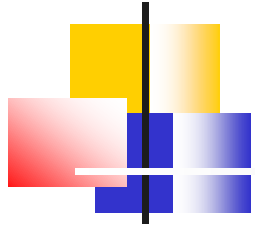
# SCC-275 - Ciência de Dados

## Exploração de dados PARTE II

Profa. Roseli Ap. Francelin  
Romero – SCC

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

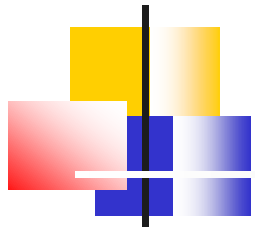




# Tópicos

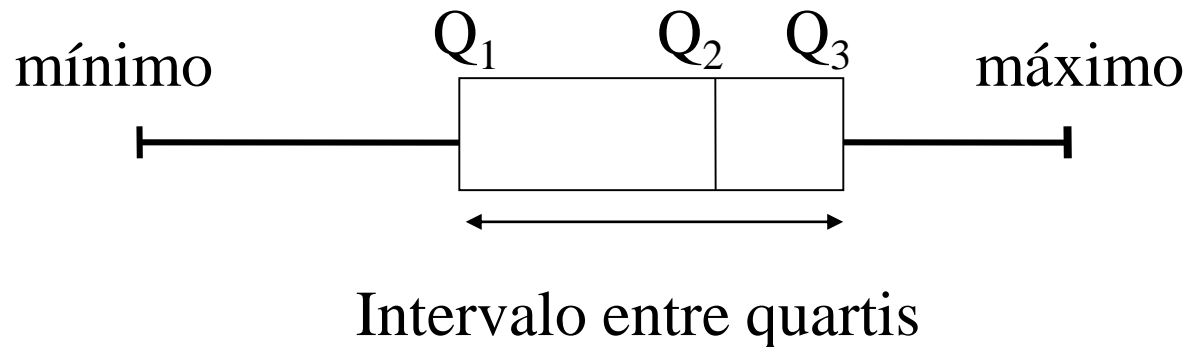
---

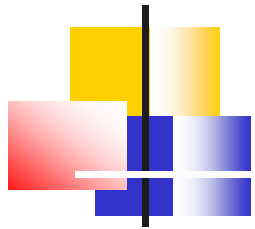
- Dados
- Caracterização de dados
  - Objetos e atributos
  - Tipos de dados
- Exploração de dados
  - Dados univariados
  - Dados multivariados
  - Visualização



# Boxplot

- Gráfico que resume informações dos quartis

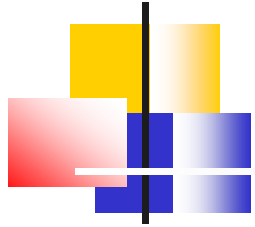




# Boxplot modificado

---

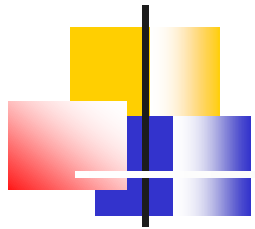
- Identifica *outliers* e reduz seu efeito no formato do boxplot
  - Tolerância =  $1,5 \times$  intervalo entre quartis
  - Verificar se  $\text{máximo} - Q_3$  ( $Q_1 - \text{mínimo}$ )  $>$  tolerância
    - Valor fora do intervalo é considerado *outlier*
    - Define novo mínimo e/ou máximo



# Medidas de espalhamento


---

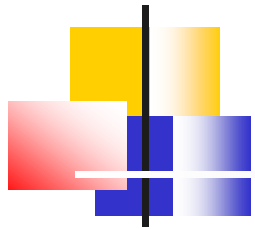
- Medem variabilidade, dispersão ou espalhamento de um conjunto de valores
- Indicam se os dados estão:
  - Amplamente espalhados ou
  - Relativamente concentrados em torno de um ponto (ex. média)
- Medidas comuns
  - Intervalo ou amplitude
  - Variância
  - Desvio padrão



# Intervalo

---

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

- 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

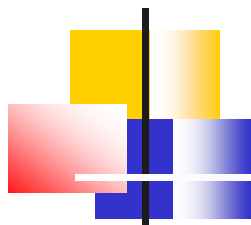


# VARIÂNCIA

---

- "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:

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

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

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

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

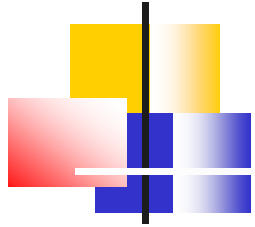
$$\text{Var(D)} = 6,0$$



# Variância e Desvio Padrão

---

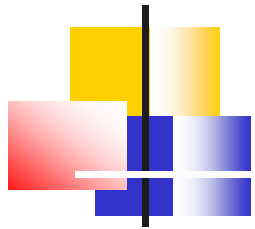
- **$dp(A) \approx 1,41$**
- **$dp(B) \approx 2,32$**
- **$dp(C) \approx 1,36$**
- **$dp(D) \approx 2,45$**
- **Funcionário A:  $10,0 \pm 1,41$  peças por dia**  
**Funcionário B:  $12,8 \pm 2,32$  peças por dia**  
**Funcionário C:  $10,4 \pm 1,36$  peças por dia**  
**Funcionário D:  $11,0 \pm 2,45$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



# 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
  - ...



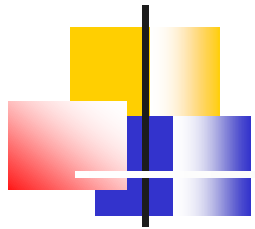
# Momento original

---

- 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

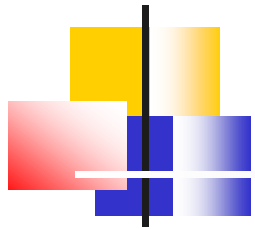
---

- 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

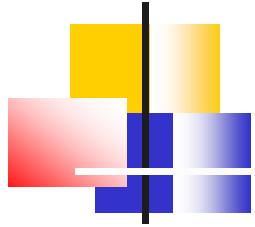


# Momento padronizado

---

- 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

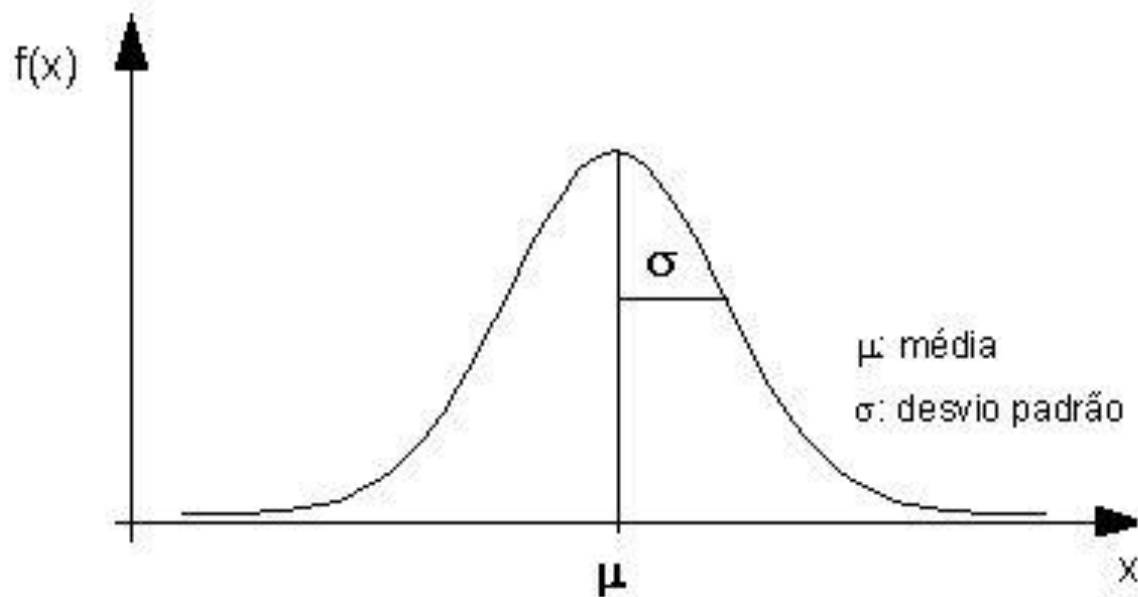
---

- 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

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

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

# Distribuição normal





# 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

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

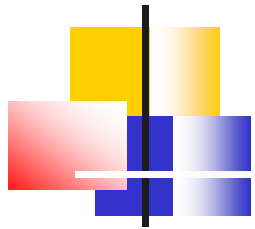


# Curtose

---

- 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:

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

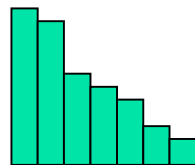


# Histograma

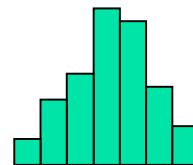
---

- Melhor forma para verificar graficamente curtose e obliquidade

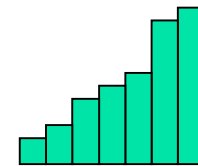
Obliquidade



Positiva

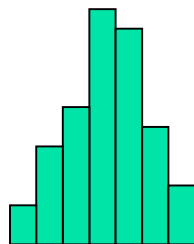


Zero (simétrica)

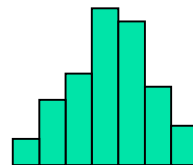


Negativa

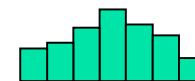
Curtose



Positiva

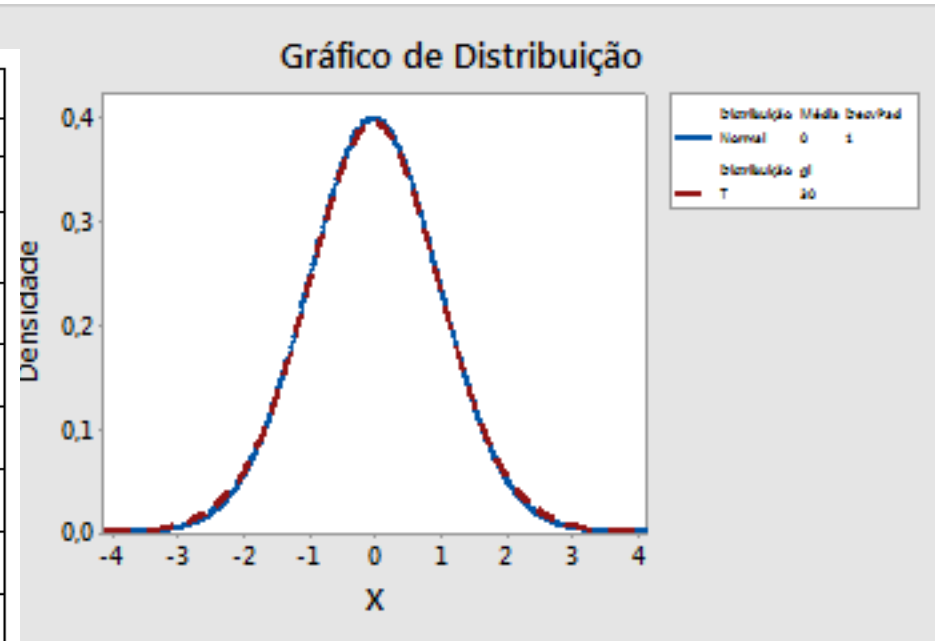
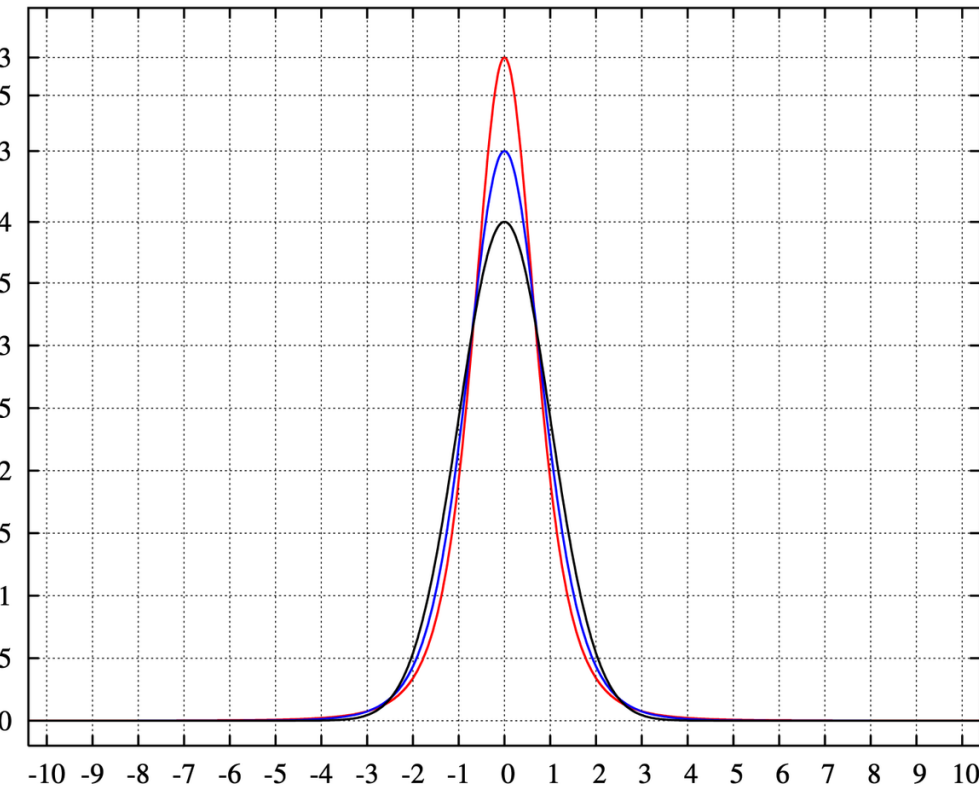


Zero (normal)

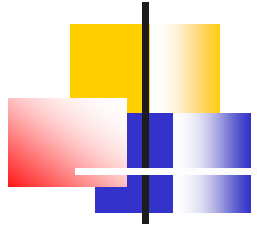


Negativa

# Curtose faz a diferença



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



# Exercício

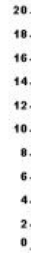
---

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

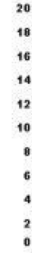
1, 3, 5, 6, 8, 10, 15

# Boxplot e Estatística Descritiva

## ■ Centralidade

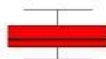
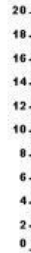


Média = 7

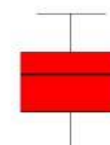
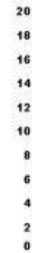


Média = 12

## ■ Espalhamento



Desvio padrão = 3

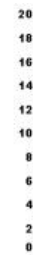
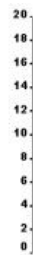
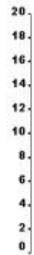


Desvio padrão = 7

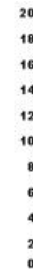
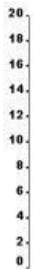
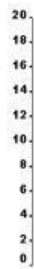


# Boxplot e Estatística Descritiva

- Obliquidade (simetria)

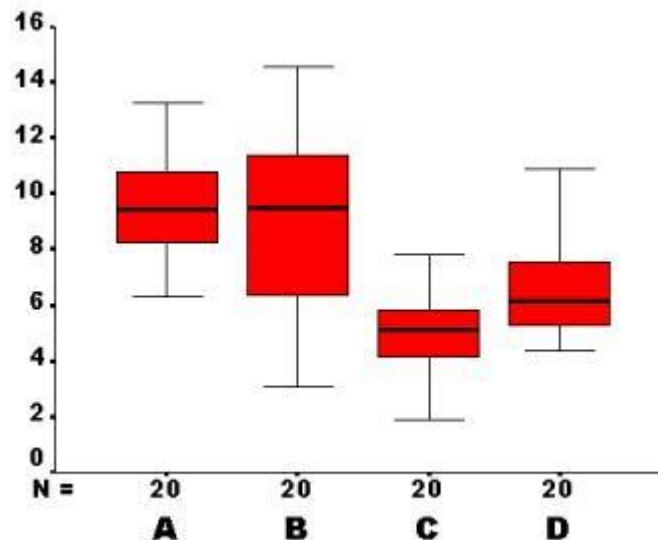


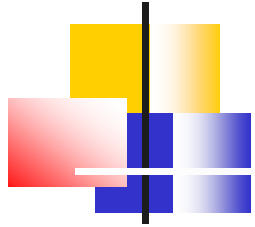
- Curtose (achatamento)



# Boxplot e Estatística Descritiva

- Análise da distribuição de dados para 4 atributos preditivos:

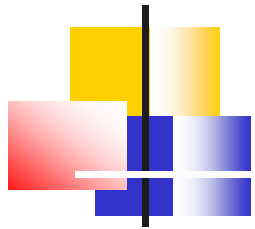




# Dados multivariados

---

- Possuem mais de um atributo
  - Cada atributo é uma variável
- Medidas de localização (tendência central)
  - Podem ser obtidas calculando medida de localização de cada atributo separadamente
  - Ex.: média, mediana, ...
    - Média dos objetos de um conjunto de dados com  $m$  atributos é dada por:  $\bar{x} = (\bar{x}_1, \dots, \bar{x}_m)$



# Dados multivariados

---

- Medidas de espalhamento (dispersão)
  - Podem ser calculadas para cada atributo independentemente dos demais
    - Usando qualquer medida de espalhamento
      - Intervalo, variância, desvio padrão
  - Para dados multivariados numéricos é melhor usar uma matriz de covariância
    - Cada elemento da matriz é a covariância entre dois atributos



# Dados multivariados

---

- Cálculo de cada elemento  $s_{ij}$  de uma matriz de covariância  $S$  para um conjunto de  $n$  objetos

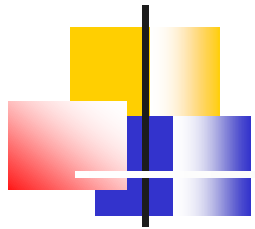
$$s_{ij} = \text{covariância}(x_i, x_j) = \frac{1}{n-1} \sum_{k=1}^n (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$$

Onde:

$\bar{x}_i$  : Valor médio do i-ésimo atributo

$x_{ki}$  : Valor do i-ésimo atributo para o k-ésimo objeto

- Obs: covariância  $(x_i, x_i) = \text{variância}(x_i)$ 
  - Matriz de covariância tem em sua diagonal as variâncias dos atributos



# Exercício

---

- Calcular a matriz de covariância para o conjunto de dados:

Peso	Altura	Temperatura
73	170	37
67	165	38
90	190	34
49	152	31



# Dados multivariados

---

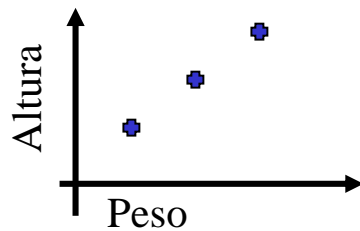
- Covariância de dois atributos
  - Mede o grau com que os atributos variam juntos (linearmente)
    - Valor próximo de 0:
      - Atributos não têm um relacionamento linear
    - Valor positivo:
      - Atributos diretamente relacionados
        - Quando o valor de um atributo aumenta, o do outro também aumenta
      - Valor negativo:
        - Atributos inversamente relacionados
    - Valor depende da magnitude dos atributos



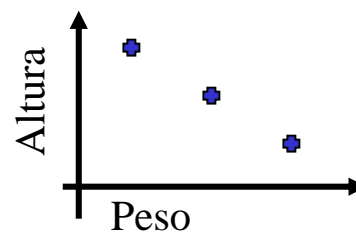
# Exemplo

---

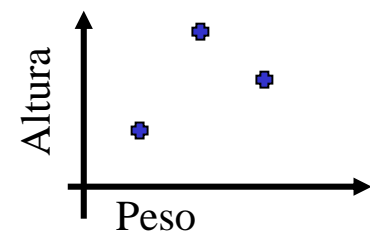
Peso	Altura
60	170
70	180
80	190



Peso	Altura
60	190
70	180
80	170



Peso	Altura
60	170
70	190
80	180



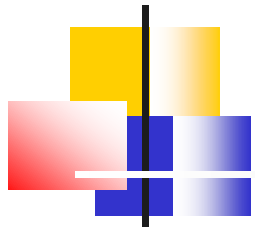




# Dados multivariados

---

- Covariância de dois atributos
  - É difícil avaliar o relacionamento entre dois atributos olhando apenas a covariância
    - Sofre influência da faixa de valores dos atributos
    - Correlação linear entre dois atributos ilustra mais claramente a força da relação linear entre eles
      - Mais popular que covariância
      - Elimina influência da faixa de valores



# Perguntas

---

