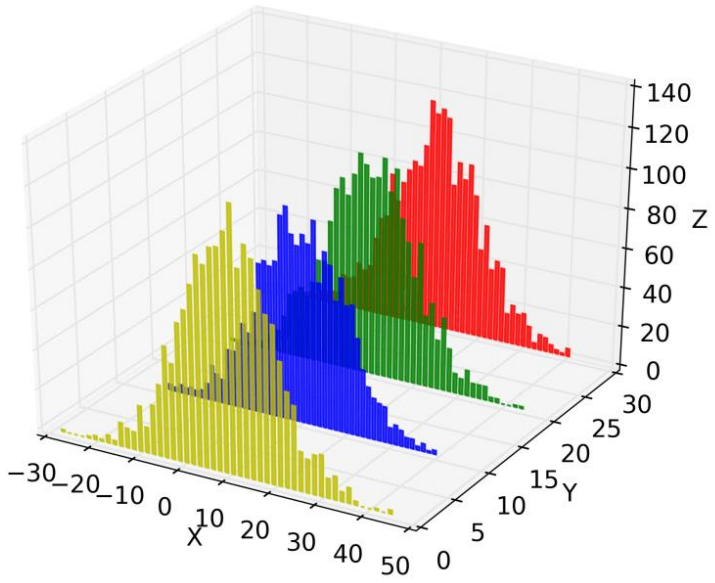
Unichristus
Centro Universitário Christus



Prof. José Belo Aragão Júnior

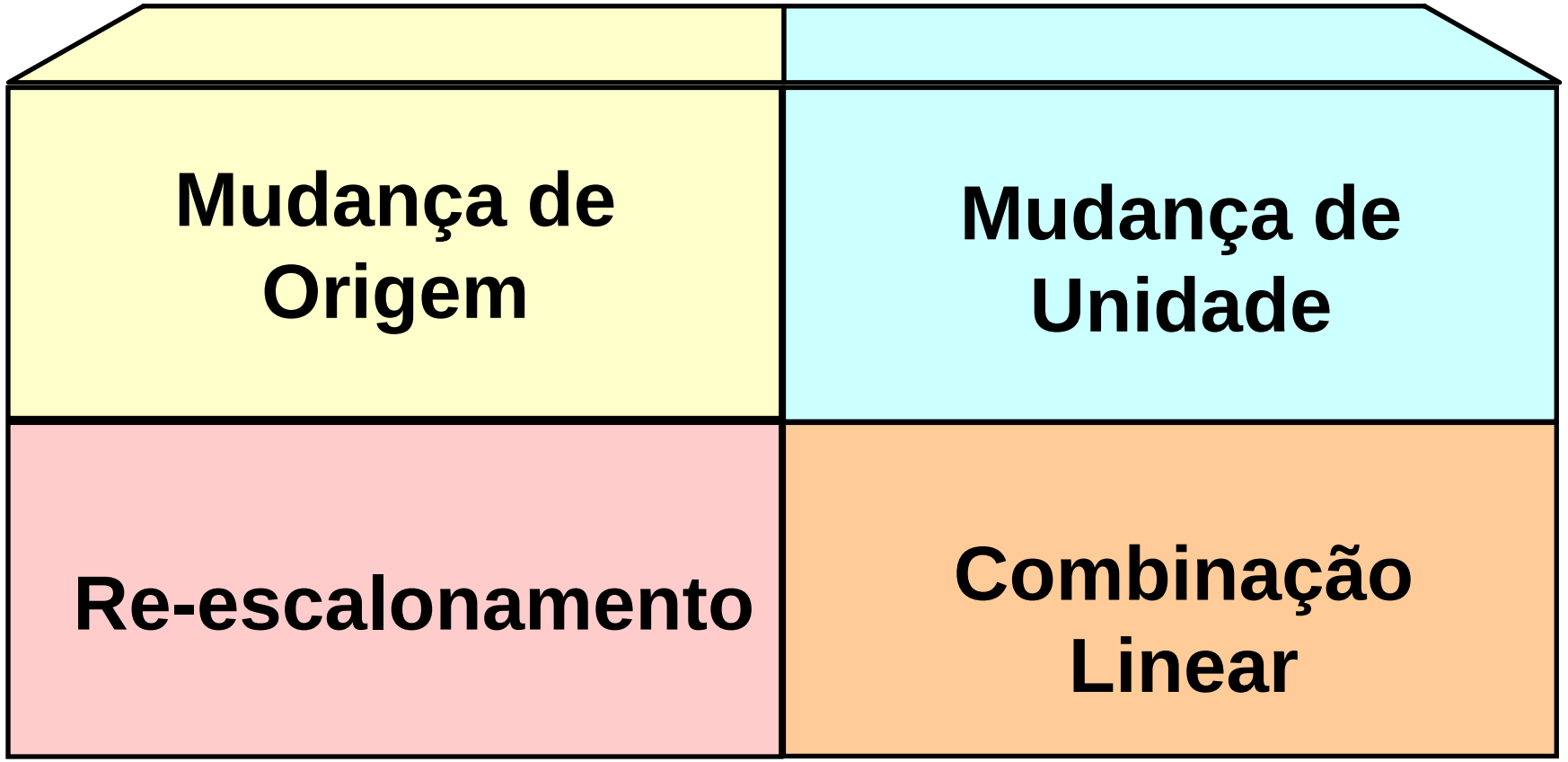
Análise Exploratória de Dados

Unichristus
Centro Universitário Christus

Transformação de Variáveis

Unichristus
Centro Universitário Christus

Transformações de Variáveis



Problema Conceitual 1

- $X = \{x_1, x_2, \ldots, x_N\}$

- $Y = \{y_1, y_2, \ldots, y_N\}$, onde $y_i = x_i + n$

  - $\mu_Y = \mu_X + n$

  - $\sigma_Y = \sigma_X$

Problema Conceitual 2

- $X = \{x_1, x_2, \ldots, x_N\}$

- $W = \{w_1, w_2, \ldots, w_N\}$, onde $w_i = m.\ x_i$

  - $\mu_W = m.\mu_X$

  - $\sigma_W = |m|.\sigma_X$

Unichristus
Centro Universitário Christus

Problema Conceitual 3

- $X = \{x_1, x_2, \ldots, x_N\}$

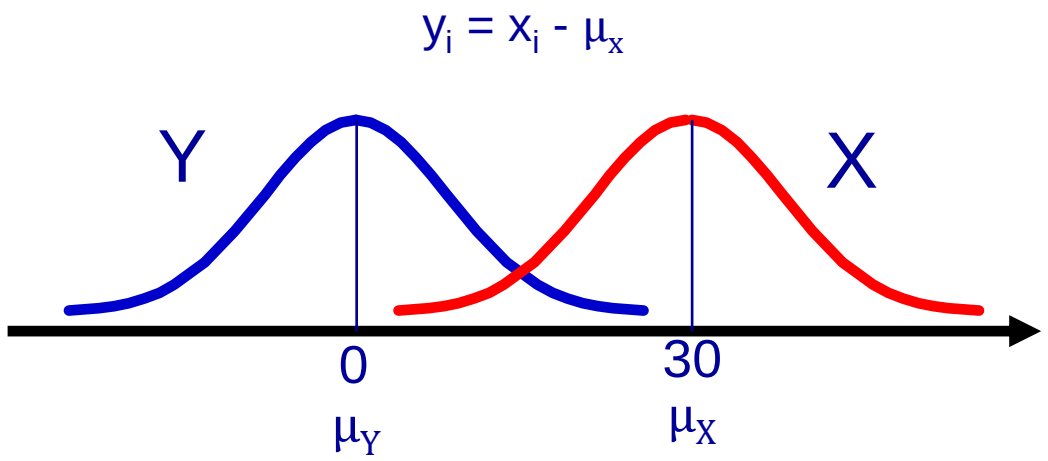
- $Q = \{q_1, q_2, \ldots, q_N\}$, onde $q_i = m.(x_i + n)$

  - $\mu_Q = m.(\mu_X + n)$

  - $\sigma_Q = |m|.\sigma_X$

Mudança de Origem

- Ao diminuir cada observação pela média, a nova distribuição se desloca para um novo centro (origem): Zero



- As variáveis mantêm suas próprias unidades.

Mudança de Origem - Exemplo

- Considere os escores de 5 alunos nas provas de Português e Matemática. Note que as notas estão em escalas diferentes.

Aluno / Prova	1	2	3	4	5
Português	37	36	46	39	42
Matemática	8	6	4	7	5

  - Como comparar o desempenho dos alunos nas duas provas?
  - Como classificar os alunos pelo desempenho nas duas provas?

Mudança de Origem

- Média em Português: $\mu_P = 40$
- Média em Matemática: $\mu_M = 6$
- Escores com a mudança de origem. ($y_i = x_i - \mu_x$)

Prova \ Aluno	1	2	3	4	5
Português	-3	-4	6	-1	2
Matemática	2	0	-2	1	-1

As unidades ainda estão expressas na escala original de cada prova.

Mudança de Origem

- Com a mudança de origem, quem teve desempenho acima da média ficou com nota positiva e quem teve desempenho abaixo da média ficou com nota negativa.

Aluno / Prova	1	2	3	4	5
Português	-3	-4	6	-1	2
Matemática	2	0	-2	1	-1

- Ainda não podemos comparar os desempenhos. As unidades permanecem diferentes.

Unichristus
Centro Universitário Christus

Mudança de Unidade

- Ao dividir o valor de cada afastamento em relação à média pelo desvio padrão, a nova variável, z, fica expressa em número de desvios padrão em torno da média.

$$z_i = \frac{y_i}{\sigma_x} = \frac{x_i - \mu_x}{\sigma_x}$$

- A esse procedimento chamamos PADRONIZAÇÃO.
- A média e o desvio padrão das distribuições na forma padronizada são 0 e 1, respectivamente.

Mudança de Unidade - Exemplo

Prova \ Aluno	1	2	3	4	5
Português (x_P)	-3	-4	6	-1	2
Matemática (x_M)	2	0	-2	1	-1

$\bar{x}_P = 40$ $s_P = 3{,}6332$

$\bar{x}_M = 6$ $s_M = 1{,}4142$

$$z_i = \frac{y_i}{s} = \frac{x_i - \bar{x}}{s}$$

Prova \ Aluno	1	2	3	4	5
Português (z_P)	-0,75	-1	1,5	-0,25	0,5
Matemática (z_M)	1,25	0	-1,25	0,63	-0,63

Reescalonamento dos Escores

- Mudanças da origem e/ou unidade padronizada a fim de se obter melhor representação dos valores;
- Mantém a ordenação das unidades;
- Usado para evitar valores negativos ou colocar os escores em uma escala conveniente;
- O procedimento deve ser comum a todas as variáveis sob análise.

Reescalonamento dos Escores

- Nova média = 100 (arbitrária)
- Novo desvio padrão = 10 (arbitrário)
- Escores na nova escala: $w_i = 10.z_i + 100$

Prova \ Aluno	1	2	3	4	5
Português Esc.	92,5	90	115	97,5	105
Matemática Esc.	112,5	100	87,5	106,3	93,7

A média e o desvio padrão das notas das provas na nova escala são 100 e 10, respectivamente.

Unichristus
Centro Universitário Christus

Combinação Linear

- Permite ordenar os indivíduos utilizando mais de uma dimensão
- Exemplo

  Calcular a média ponderada dos escores das provas de Português (peso 1) e Matemática (peso 2) para cada aluno a fim de classificá-los.

Combinação Linear

$$\text{Nota Final} = \frac{(\text{Português} \times 1) + (\text{Matemática} \times 2)}{3}$$

Aluno / Prova	1	2	3	4	5
Nota Final	106,7	96,3	96,1	103,8	97,1
Classificação	1º	4º	5º	2º	3º

Unichristus
Centro Universitário Christus

Covariância e Correlação

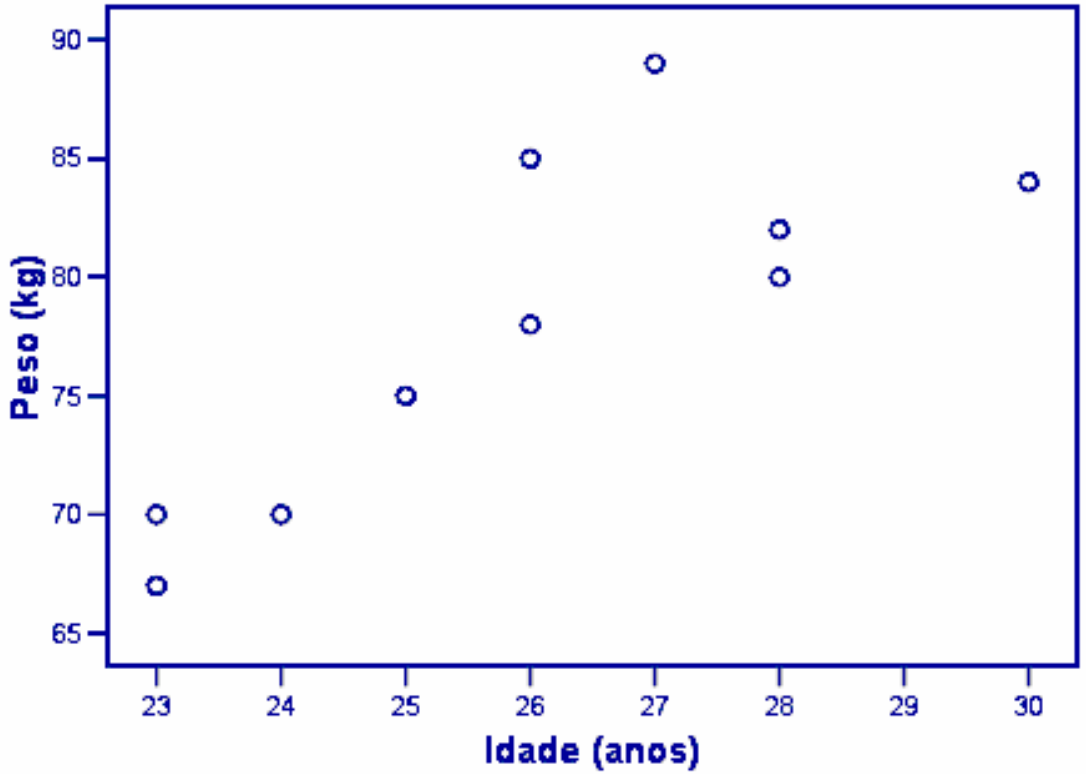
Covariância

- Utilizada para avaliar o tipo da relação **LINEAR** entre dois fenômenos.
- Considere a tabela abaixo contendo os dados sobre as idades (X) e os pesos (Y) de 10 alunos.

Idades (X)	Pesos (Y)
25	75
27	89
24	70
28	82
23	70
26	85
30	84
28	80
26	78
23	67

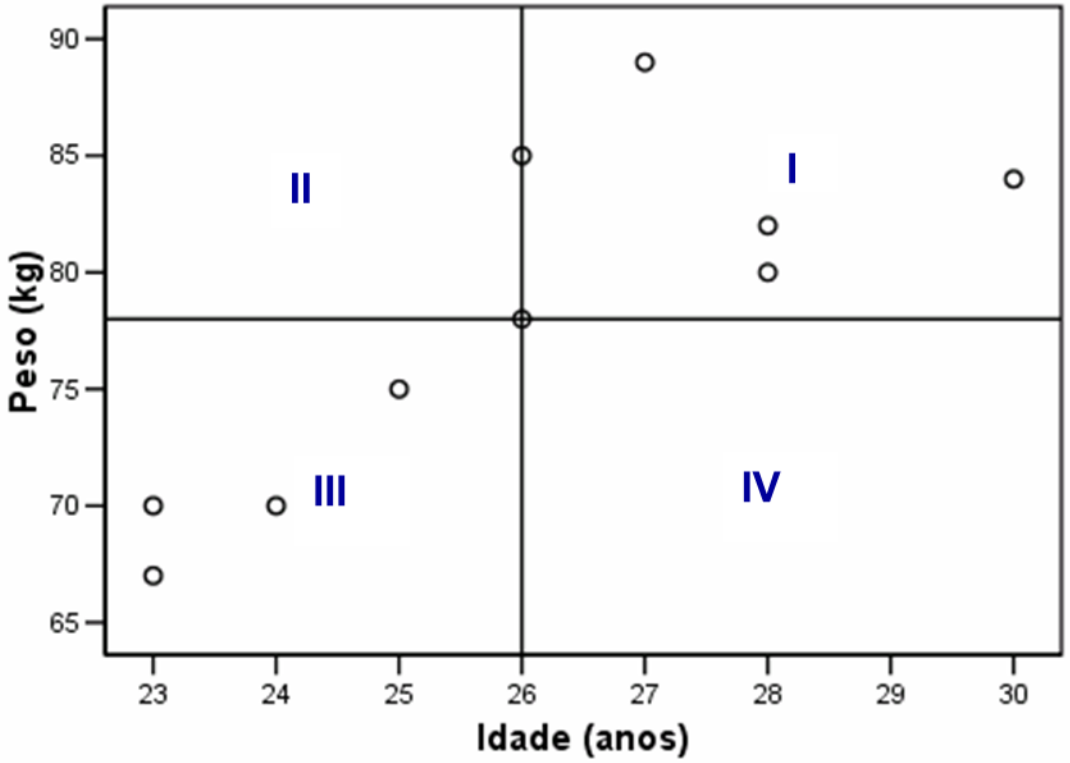
Covariância

Podemos dispor esses dados em um gráfico denominado **Diagrama de Dispersão**:



Covariância

- Subtraindo cada observação da média de seu respectivo conjunto (mudança de origem) obtemos o Diagrama de Dispersão dividido em 4 quadrantes com novos eixos definidos.



Covariância

- Multiplicando as coordenadas de cada ponto na nova escala, podemos posicionar o ponto em cada um dos 4 quadrantes.

  - Quando o produto for positivo o ponto estará no quadrante I ou III.

  - Quando o produto for negativo o ponto estará no quadrante II ou IV.

Covariância

- Considerando todos os pontos, a soma $\sum[(x_i - \bar{x}).(y_i - \bar{y})]$ será:
  - Positiva, quando a maior parte dos pontos estiver concentrada em torno dos quadrantes I e III, indicando uma relação linear DIRETA entre X e Y.
  - Negativa, quando a maior parte dos pontos estiver concentrada em torno dos quadrantes II e IV indicando uma relação linear INVERSA entre X e Y.
  - Nula, indicando falta de relação linear entre X e Y.

Unichristus
Centro Universitário Christus

Covariância

- Se dividirmos $\sum[(x_i - \bar{x}).(y_i - \bar{y})]$ pelo tamanho da amostra teremos uma medida de variação conjunta média entre X e Y, denominada Covariância: COV(X,Y).
- Essa medida indica o tipo de relação linear entre duas variáveis e é definida como:
  - Amostra: $COV(X,Y) = s_{XY} = \frac{\sum_{i=1}^{n}[(x_i - \bar{x}).(y_i - \bar{y})]}{n-1}$
  - População: $COV(X,Y) = \sigma_{XY} = \frac{\sum_{i=1}^{N}[(x_i - \mu_X).(y_i - \mu_Y)]}{N}$

Covariância

$$\sigma_{XY} = \frac{\sum_{i=1}^{N}[(x_i - \mu_X).(y_i - \mu_Y)]}{N} = \frac{\sum_{i=1}^{N}(x_i y_i - x_i \mu_Y - y_i \mu_X + \mu_X \mu_Y)}{N}$$

$$\sigma_{XY} = \frac{\sum_{i=1}^{N} x_i y_i}{N} - \mu_Y \frac{\sum_{i=1}^{N} x_i}{N} - \mu_X \frac{\sum_{i=1}^{N} y_i}{N} + \frac{N \mu_X \mu_Y}{N}$$

$$\sigma_{XY} = \frac{\sum_{i=1}^{N} x_i y_i}{N} - \mu_Y \mu_X - \mu_X \mu_Y + \mu_X \mu_Y$$

$$\sigma_{XY} = \frac{\sum_{i=1}^{N} x_i y_i}{N} - \mu_X \mu_Y$$

Covariância - Exemplo

- Considere a tabela abaixo contendo os dados sobre as idades (X) e os pesos (Y) de 10 alunos.

Idades (X)	Pesos (Y)
25	75
27	89
24	70
28	82
23	70
26	85
30	84
28	80
26	78
23	67

$$n = 10,\ \bar{x} = 26 \text{ e } \bar{y} = 78$$

$$s_{XY} = \frac{\sum_{i=1}^{n}[(x_i - \bar{x}).(y_i - \bar{y})]}{n - 1}$$

$$s_{XY} = \frac{(25 - 26).(75 - 78) + \cdots + (23 - 26).(67 - 78)}{9}$$

$$s_{XY} = 13{,}67$$

Unichristus
Centro Universitário Christus

Covariância

- Como a covariância depende das unidades das variáveis, seu valor não pode ser usado para avaliar **o grau** da relação linear entre elas.

- Quando as variáveis são padronizadas, a covariância entre elas define o coeficiente de correlação.

Coeficiente de Correlação

- Utilizado para avaliar o **grau** da relação LINEAR entre 2 variáveis.
- É uma medida adimensional. Varia entre -1 e 1, inclusive.
- r = 0 não significa ausência de correlação, mas falta de relação **linear**.
- É expresso por:
  - Amostra: $r_{XY} = \frac{s_{XY}}{s_X s_Y}$
  - População: $\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$

Unichristus
Centro Universitário Christus

Coeficiente de Correlação da Amostra

$$r_{XY} = \frac{s_{XY}}{s_X s_Y} = \frac{\frac{\sum_{i=1}^{n}[(x_i - \bar{x}).(y_i - \bar{y})]}{n-1}}{\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}\sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}} = \frac{\sum_{i=1}^{n}[(x_i - \bar{x}).(y_i - \bar{y})]}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$r_{XY} = \frac{s_{XY}}{s_X s_Y} = \frac{\frac{\sum_{i=1}^{n}[(x_i - \bar{x}).(y_i - \bar{y})]}{n-1}}{\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}\sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}} = \frac{\sum_{i=1}^{n}\left[\frac{(x_i - \bar{x})}{\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}} \cdot \frac{(y_i - \bar{y})}{\sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}}\right]}{n-1}$$

$$r_{XY} = \frac{\sum_{i=1}^{n}\left[\frac{(x_i - \bar{x})}{s_X} \cdot \frac{(y_i - \bar{y})}{s_Y}\right]}{n-1}$$

Unichristus
Centro Universitário Christus

Coeficiente de Correlação da População

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\frac{\sum_{i=1}^{N}[(x_i - \mu_X).(y_i - \mu_Y)]}{N}}{\sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu_X)^2}{N}}\sqrt{\frac{\sum_{i=1}^{N}(y_i - \mu_Y)^2}{N}}} = \frac{\sum_{i=1}^{N}[(x_i - \mu_X).(y_i - \mu_Y)]}{\sqrt{\sum_{i=1}^{N}(x_i - \mu_X)^2 \sum_{i=1}^{N}(y_i - \mu_Y)^2}}$$

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\frac{\sum_{i=1}^{N}[(x_i - \mu_X).(y_i - \mu_Y)]}{N}}{\sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu_X)^2}{N}}\sqrt{\frac{\sum_{i=1}^{N}(y_i - \mu_Y)^2}{N}}} = \frac{\sum_{i=1}^{N}\left[\frac{(x_i - \mu_X)}{\sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu_X)^2}{N}}} \cdot \frac{(y_i - \mu_Y)}{\sqrt{\frac{\sum_{i=1}^{N}(y_i - \mu_Y)^2}{N}}}\right]}{N}$$

$$\rho_{XY} = \frac{\sum_{i=1}^{N}\left[\frac{(x_i - \mu_X)}{\sigma_X} \cdot \frac{(y_i - \mu_Y)}{\sigma_Y}\right]}{N}$$

Coeficiente de Correlação - Curiosidade

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\frac{\sum_{i=1}^{N}[(x_i - \mu_X).(y_i - \mu_Y)]}{N}}{\sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu_X)^2}{N}}\sqrt{\frac{\sum_{i=1}^{N}(y_i - \mu_Y)^2}{N}}} = \frac{\sum_{i=1}^{N}[(x_i - \mu_X).(y_i - \mu_Y)]}{\sqrt{\sum_{i=1}^{N}(x_i - \mu_X)^2}.\sqrt{\sum_{i=1}^{N}(y_i - \mu_Y)^2}}$$

Sejam $\vec{u}$ e $\vec{v}$ vetores tais que: $\vec{u} = (x_1 - \mu_X, \dots, x_N - \mu_X)$ e $\vec{v} = (y_1 - \mu_Y, \dots, y_N - \mu_y)$.

O cosseno do ângulo θ formado pelos vetores $\vec{u}$ e $\vec{v}$ é obtido por:

$$\cos\theta = \frac{\vec{u}.\vec{v}}{|\vec{u}|.|\vec{v}|} = \frac{\sum_{i=1}^{N}[(x_i - \mu_X).(y_i - \mu_Y)]}{\sqrt{\sum_{i=1}^{N}(x_i - \mu_X)^2}.\sqrt{\sum_{i=1}^{N}(y_i - \mu_Y)^2}} = \rho_{XY}$$

Como o coeficiente de correlação é o cosseno de um ângulo: $-1 \leq \rho_{XY} \leq 1$

Correlação - Exemplo

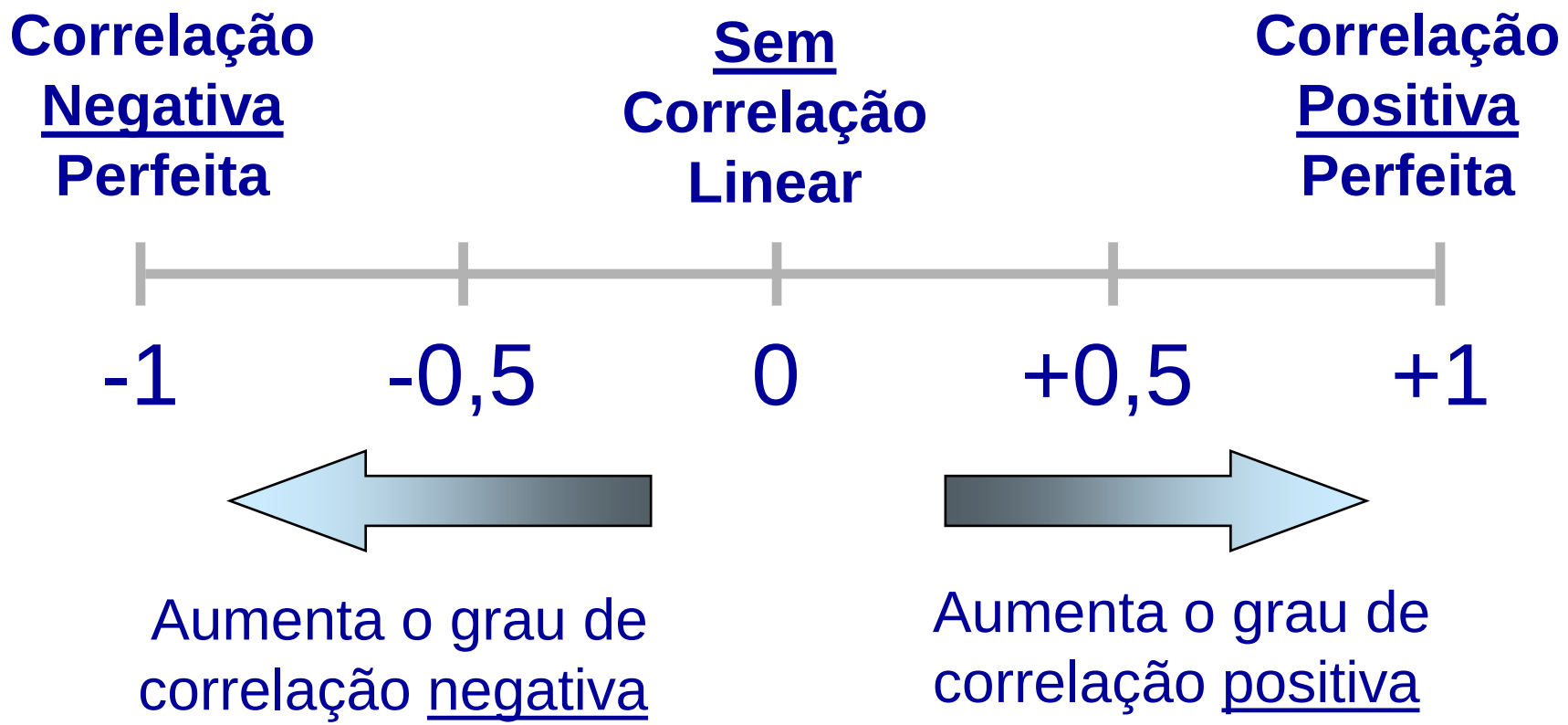
- Considere a tabela abaixo contendo os dados sobre as idades (X) e os pesos (Y) de 10 alunos.

Idades (X)	Pesos (Y)
25	75
27	89
24	70
28	82
23	70
26	85
30	84
28	80
26	78
23	67

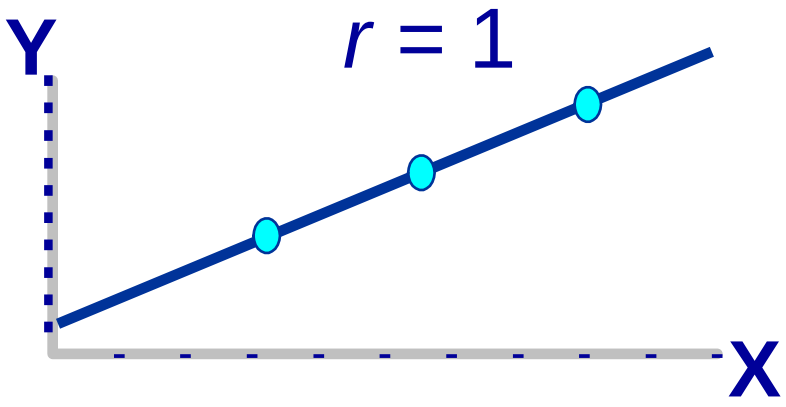
$$n = 10, \bar{x} = 26, s_X = 2{,}31, \bar{y} = 78, s_Y = 7{,}33, s_{XY} = 13{,}67$$

$$r_{XY} = \frac{s_{XY}}{s_X s_Y} = \frac{13{,}67}{2{,}31.7{,}33} = \frac{13{,}67}{16{,}93} \cong 0{,}81$$

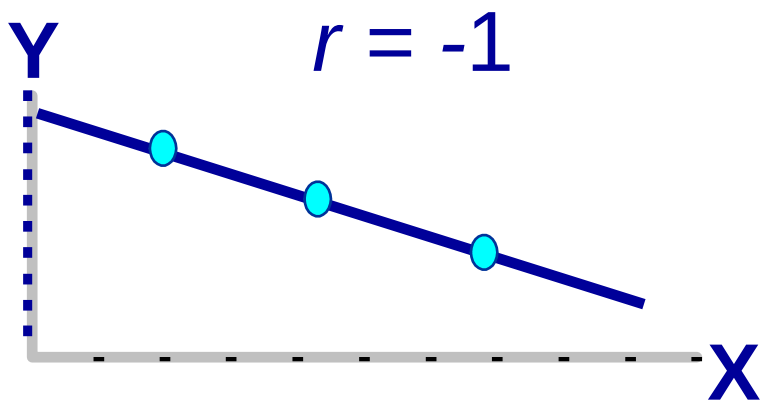
Coeficiente de Correlação



Coeficiente de Correlação (r)

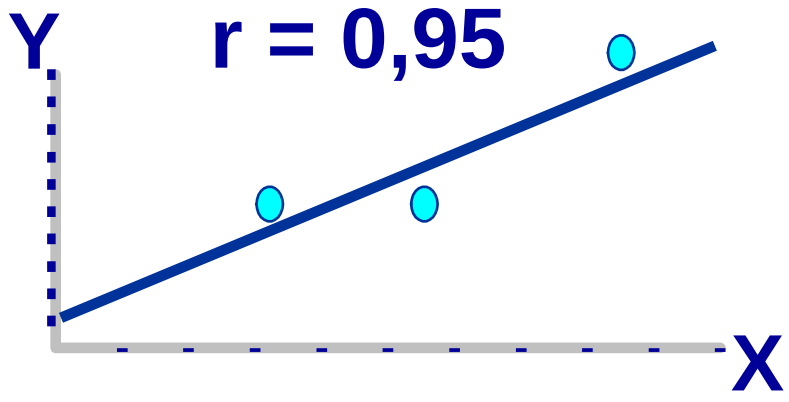


Correlação perfeita e direta (positiva)

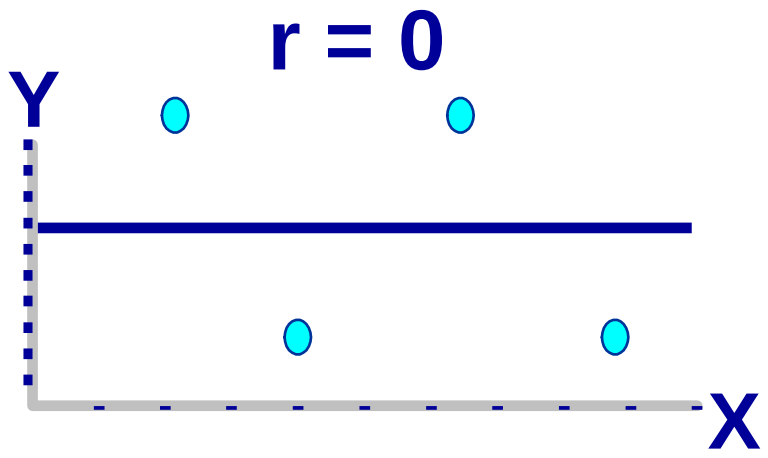


Correlação perfeita e inversa (negativa)

Coeficiente de Correlação (r)



Correlação alta e direta



Correlação nula: não há relação linear

Unichristus
Centro Universitário Christus

Exercício