Probabilidade e Estatística

Correlação e Regressão Linear

Correlação

- Existe uma correlação entre duas variáveis quando uma delas está, de alguma forma, relacionada com a outra.
- Gráfico ou Diagrama de Dispersão é o método gráfico feito sobre dois eixos, 'x' e 'y', que representa a correção entre as variáveis.

Diagramas de Dispersão

- Um diagrama de dispersão mostra a relação entre duas variáveis quantitativas, medidas sobre os mesmos indivíduos.
- Os valores de uma variável aparecem no eixo horizontal, e os da outra, no eixo vertical.
- Cada indivíduo aparece como o ponto do gráfico definido pelos valores de ambas as variáveis para aquele indivíduo

Variáveis

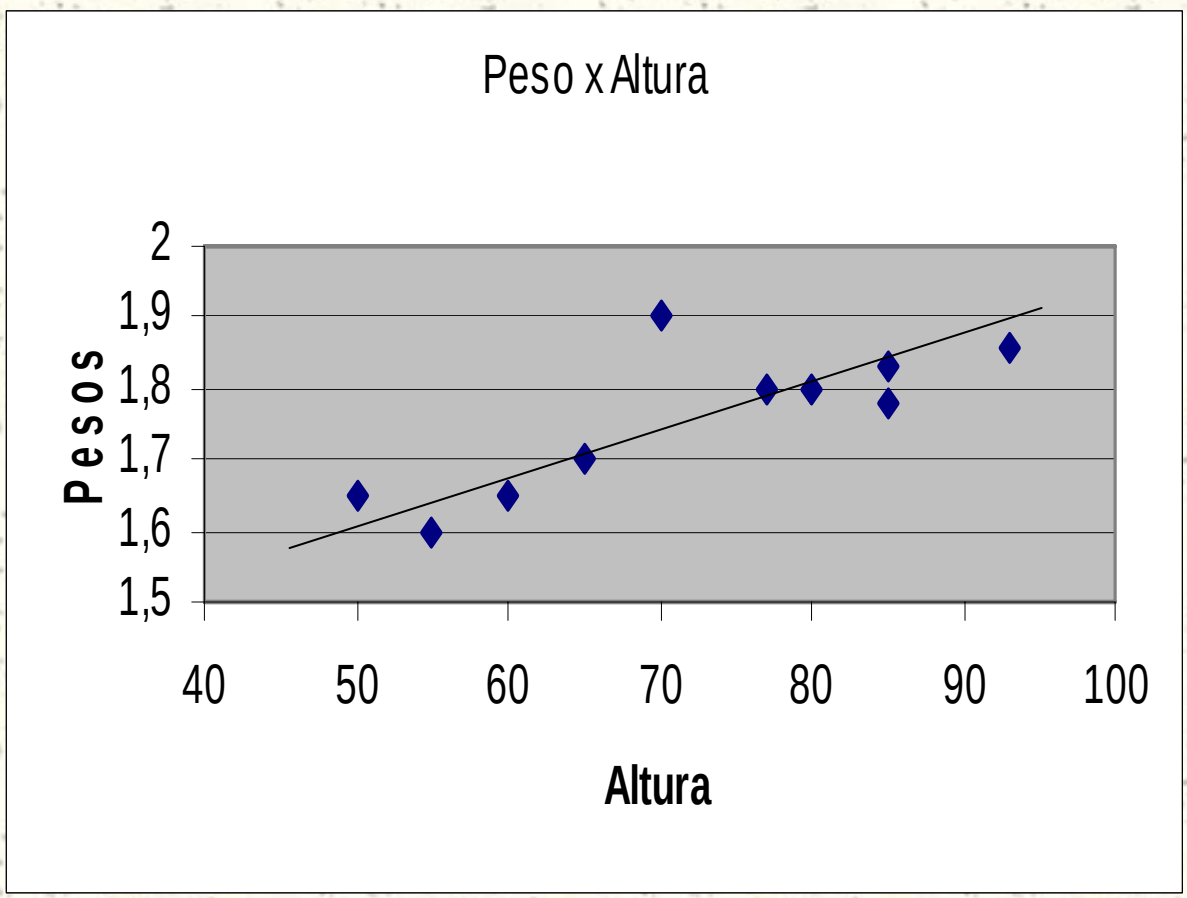
- Variável: características ou itens de interesse de cada elemento de uma população ou amostra
  - Também chamada parâmetro, posicionamento, condição...
- Duas variáveis estão relacionadas se a mudança de uma provoca a mudança na outra.
    - Exemplo: velocidade x consumo combustível
- O eixo x geralmente é um parâmetro.

Exemplos

- **Fabricação**
  - Número de peças produzidas e número de peças defeituosas
- **Construção**
  - Número de falhas em uma obra e a satisfação média dos produtivos
  - Dias de atraso de entrega x número de dias chuvosos
- **Financeiro**
  - Média de tempo de atraso de pagamento e número de erros de fatura
- **Vendas**
  - % de imóveis vendidos na data de entrega da obra x satisfação média dos clientes nos últimos 10 empreendimentos.

Exemplo - Peso x altura

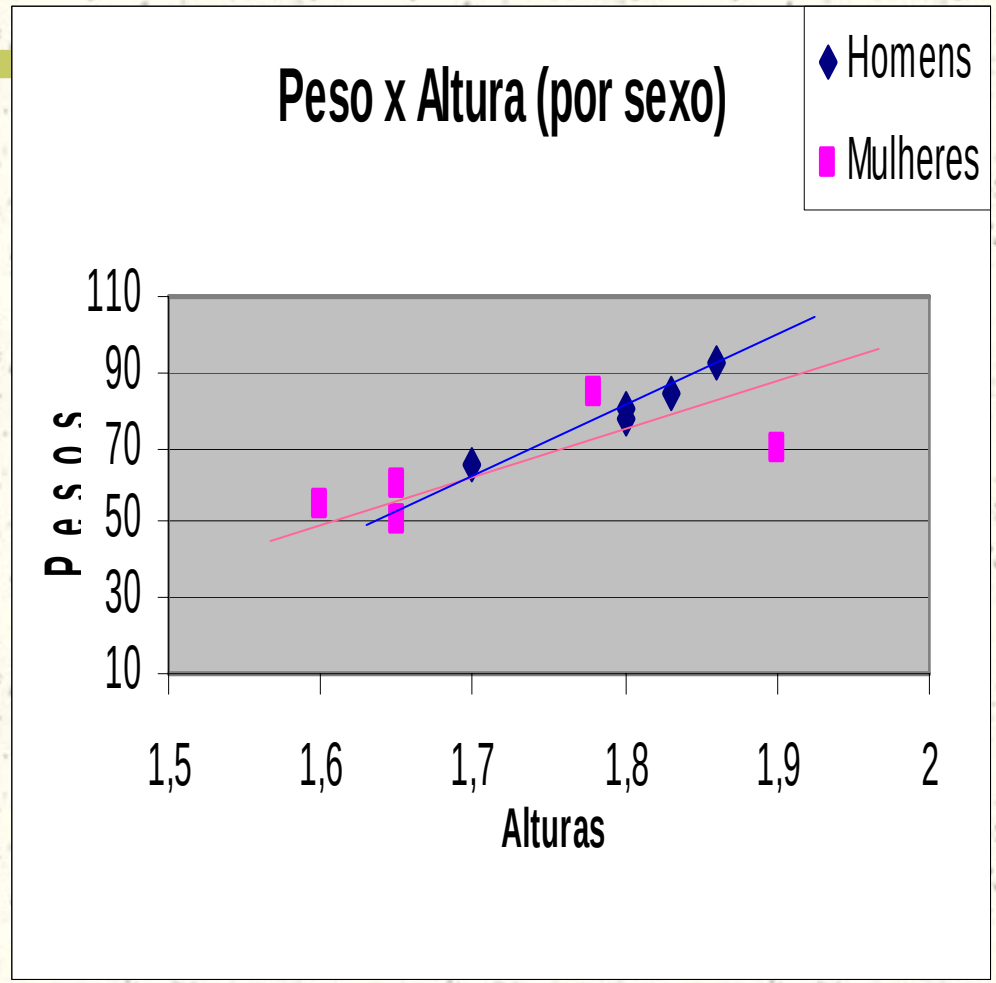
Peso (kg)	Altura (m)
80	1,80
85	1,83
50	1,65
70	1,90
55	1,60
77	1,80
85	1,78
93	1,86
65	1,70
60	1,65



Exemplo – Peso x Altura

Estratificando

Peso (kg)	Altura homens (m)	Altura Mulheres (m)
80	1,80	---
85	1,83	---
50	---	1,65
70	---	1,90
55	---	1,60
77	1,80	---
85	---	1,78
93	1,86	---
65	1,70	---
60	---	1,65



Dicas

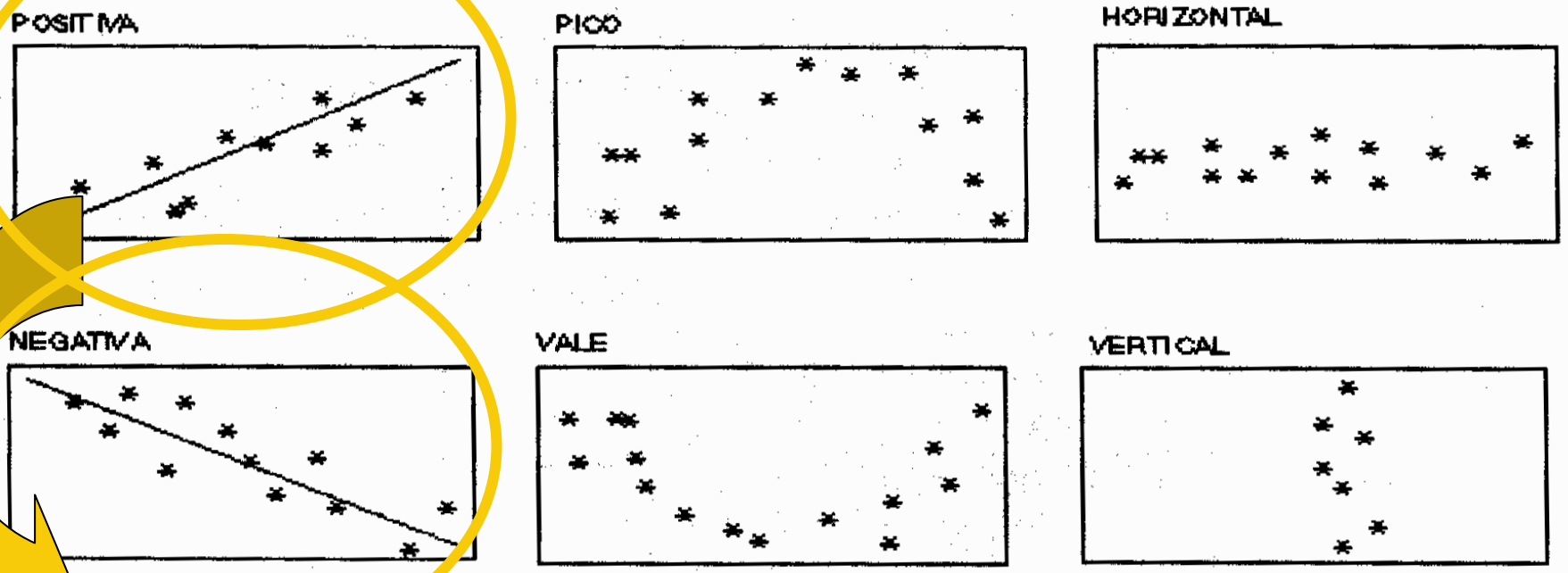
- Eixo ´x´
  - Variável que é alterada por uma modificação no processo (variável independente)
  - Geralmente uma possível causa de um problema
- Eixo ´y´
  - Variável que pode mudar de acordo com a mudança da variável em ´x´ (variável dependente)
  - Geralmente um indicador de qualidade ou efeito gerado por uma causa.

Analisando Diagramas de Dispersão

- Os aspectos abaixo são relevantes na análise dos Diagramas:
  - DIREÇÃO
  - FORMA (linear, não-linear, aglomerados)
  - PONTOS DISCREPANTES

Interpretando

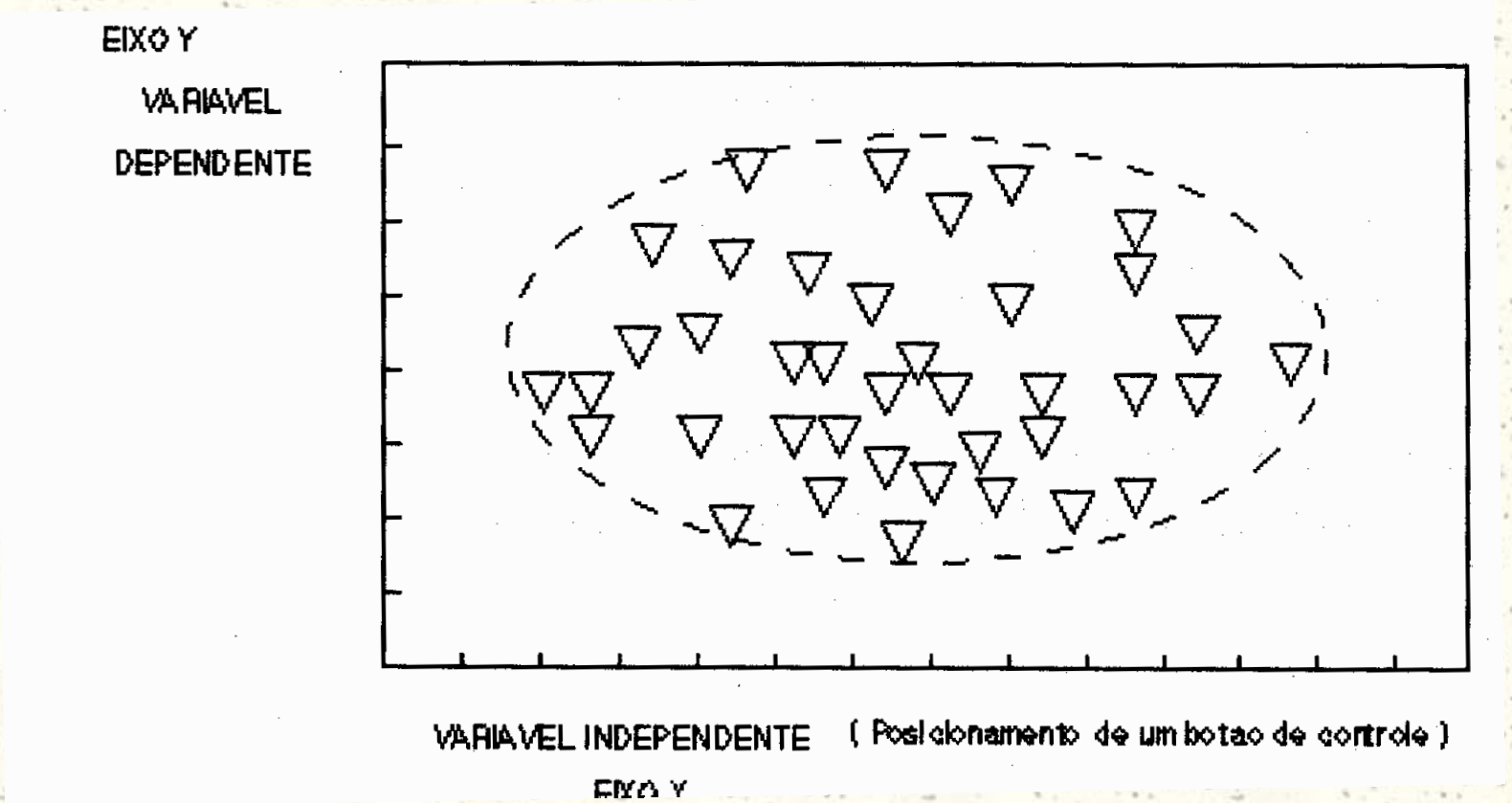
Padrões de Dispersão



Quanto maior a correlação, mais próxima de uma reta a 45º ou 135º será a distribuição.

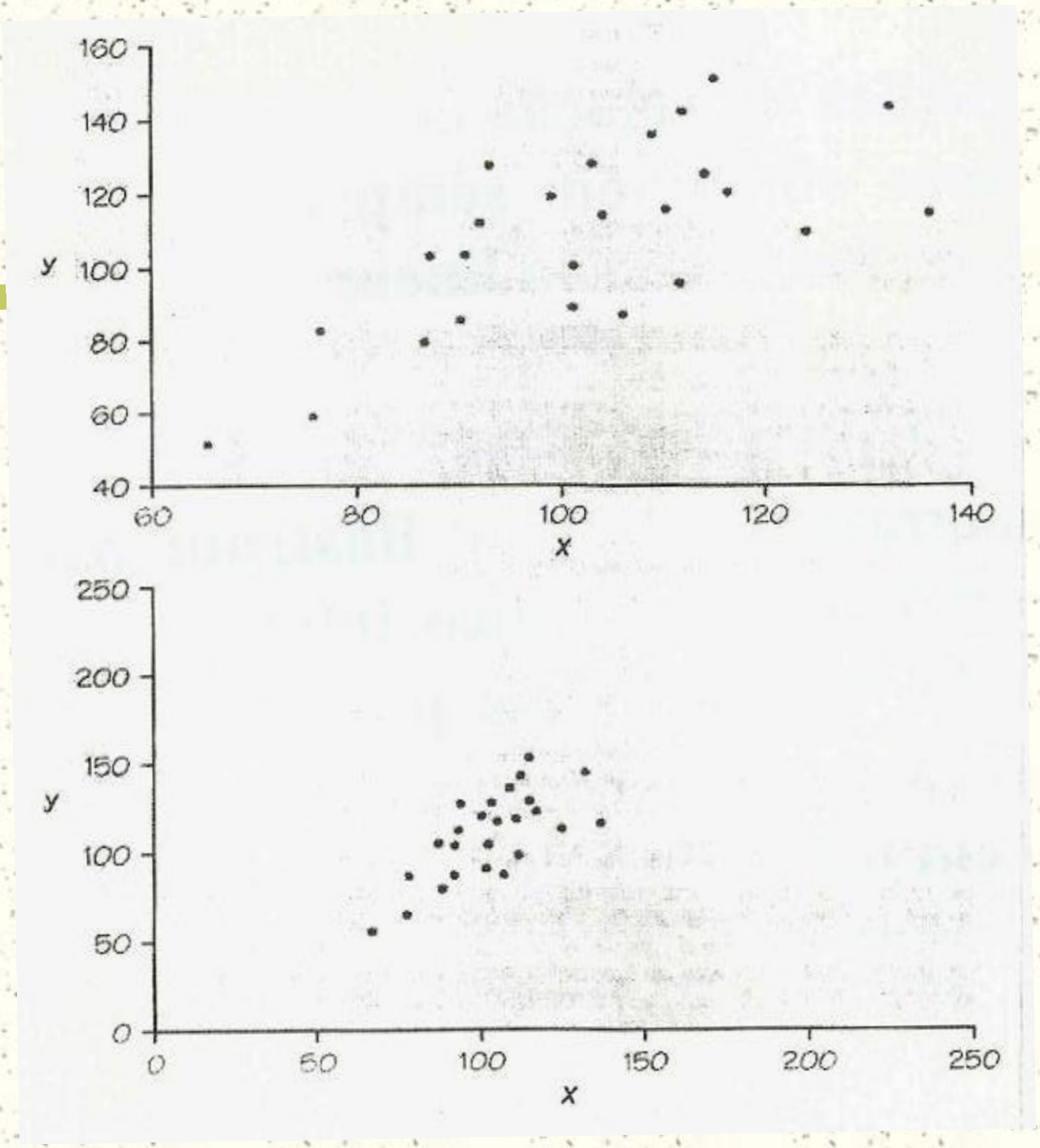
Interpretando

Grau de Relacionamento



Problemas da Análise Gráfica

- A análise gráfica da relação entre variáveis é importante, mas os olhos nem sempre são um bom juiz da *intensidade* de uma relação linear.
- Os diagramas a seguir ilustram precisamente os mesmos dados, mas o gráfico inferior é menor em um campo mais amplo.


160
140
120
y 100
80
60
40
60
80
100
120
140
x
250
200
150
y 100
50
0
0
50
100
150
200
250
x

Problemas da Análise Gráfica

- Nossos olhos podem ser enganados por uma mudança de escalas, ou pela quantidade de espaço em branco em torno do aglomerado dos pontos.
- Deve-se, então, utilizar uma *medida numérica* para suplementar o gráfico.
  - Coeficiente de Correlação Linear (*r*)

Coeficiente de Correlação Linear

- *r* → mede o grau de relacionamento linear entre valores emparelhados *x* e *y* em uma *amostra*.
- Mede a *intensidade* e a *direção* da relação linear entre duas variáveis quantitativas
- Chamado também de Coeficiente de Correlação de Pearson (Karl Pearson, 1857-1936).

Coeficiente de Correção Linear ou Coeficiente de Pearson

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \Rightarrow S_{xx} = n(\sum x_i^2) - (\sum x_i)^2$$

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \Rightarrow S_{yy} = n(\sum y_i^2) - (\sum y_i)^2$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \Rightarrow S_{xy} = n\sum x_i.y_i - (\sum x_i)(\sum y_i)$$

$$r = \frac{Sxy}{\sqrt{S_{xx}.S_{yy}}}$$

$-1 \leq r \leq 1$

Coeficiente de Correção Linear ou Coeficiente de Pearson

$$r = \frac{n\sum(x_i \cdot y_i) - \left(\sum x_i\right)\left(\sum y_i\right)}{\sqrt{n\sum x_i^2 - \left(\sum x_i\right)^2} \cdot \sqrt{n\sum y_i^2 - \left(\sum y_i\right)^2}}$$

$$-1 \leq r \leq 1$$

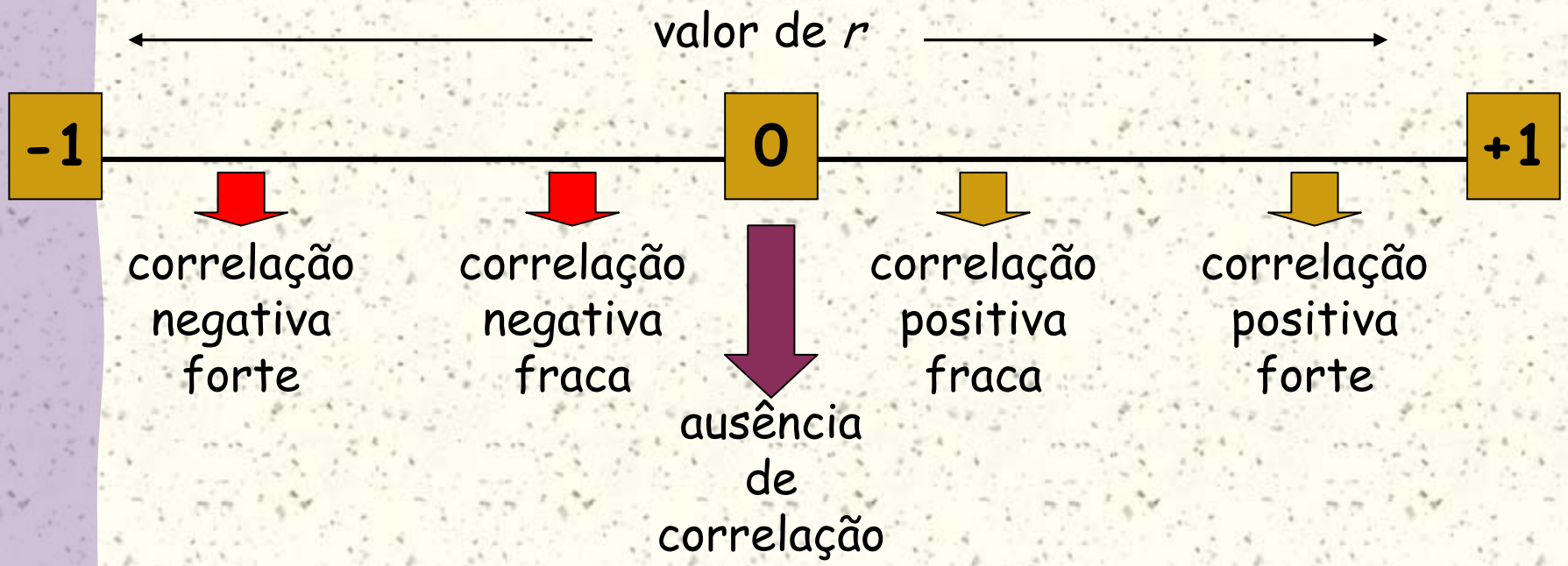
Interpretando o Coeficiente de Correlação Linear

'r' sempre será um valor entre

$$-1 \leq r \leq 1$$

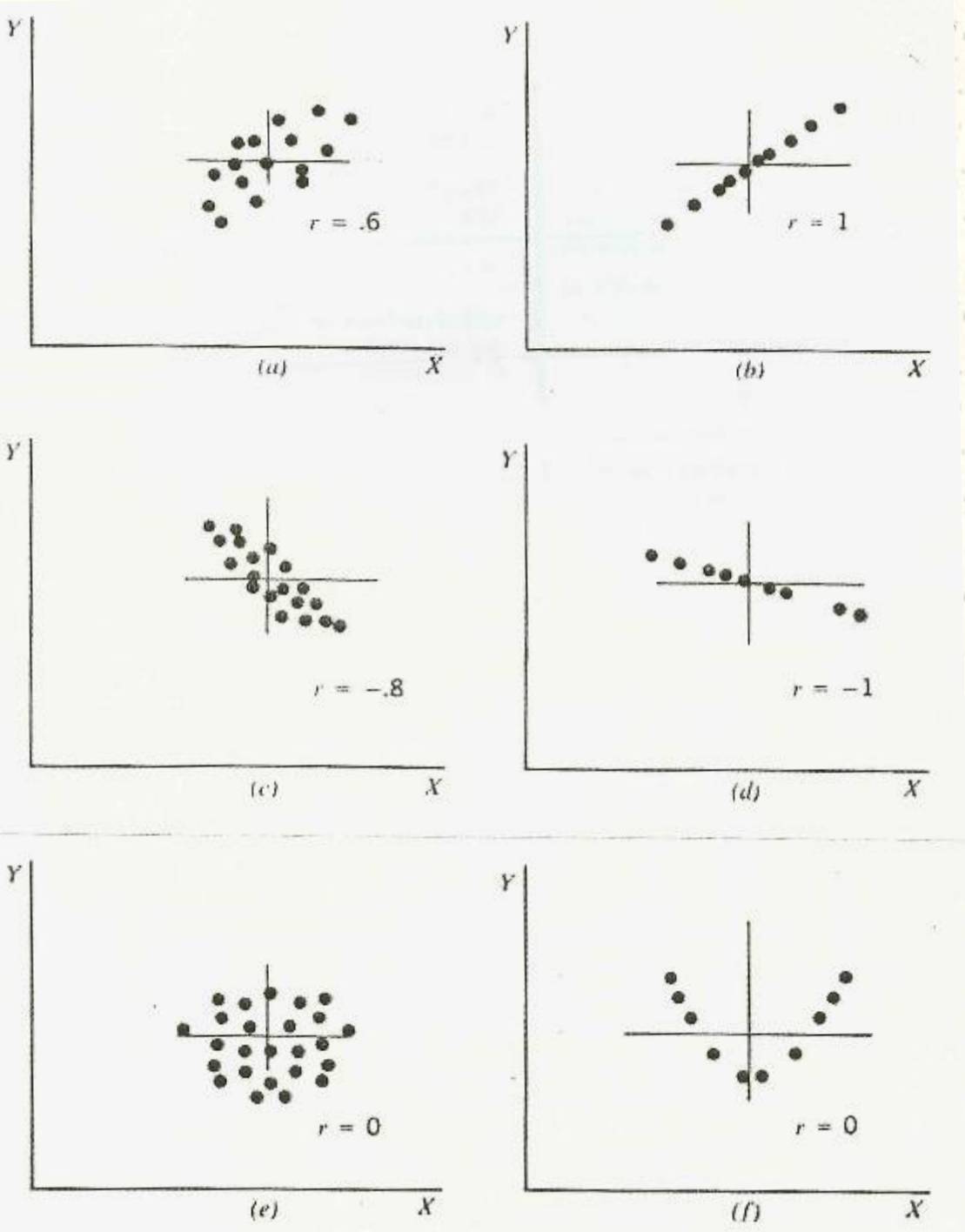
- Quanto mais próximo de –1: maior correlação negativa
- Quanto mais próximo de 1: maior correlação positiva
- Quanto mais próximo de 0: menor a correlação linear

Interpretação do Valor de *r*



Propriedades do Coeficiente de Correlação de Pearson

- $-1 \leq r \leq +1$
- O valor de *r* não varia se todos os valores de qualquer uma das variáveis são convertidos para uma escala diferente.
- O valor de *r* não é afetado pela escolha de *x* ou *y*. Permutando *x* e *y*, *r* permanece inalterado.
- *r* só mede a intensidade ou grau de relacionamentos lineares. Não serve para medir intensidade de relacionamentos não-lineares.


Y
r = .6
(a)
X
Y
r = 1
(b)
X
Y
r = −.8
(c)
X
Y
r = −1
(d)
X
Y
r = 0
(e)
X
Y
r = 0
(f)
X

Ex.: Alturas e Pesos de Ursos Siberianos

Comprimento (pol.)		Peso (lb.)			
	x	y	x.y	x2	y2
	53,0	80	4.240	2.809,00	6.400
	67,5	344	23.220	4.556,25	118.336
	72,0	416	29.952	5.184,00	173.056
	72,0	348	25.056	5.184,00	121.104
	73,5	262	19.257	5.402,25	68.644
	68,5	360	24.660	4.692,25	129.600
	73,0	332	24.236	5.329,00	110.224
	37,0	34	1.258	1.369,00	1.156
Totais	517	2.176	151.879	34.525,75	728.520

Ex.: Alturas e Pesos de Ursos Siberianos

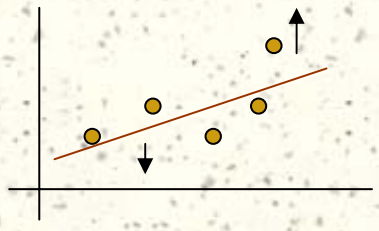
$$r = \frac{n\sum(x_i \cdot y_i) - \left(\sum x_i\right)\left(\sum y_i\right)}{\sqrt{n\sum x_i^2 - \left(\sum x_i\right)^2} \cdot \sqrt{n\sum y_i^2 - \left(\sum y_i\right)^2}} \therefore$$

$$r = \frac{8(151.879) - (516{,}5)(2.176)}{\sqrt{8(34.525{,}75) - (516{,}5)^2}\sqrt{8(728.520) - (2.176)^2}} =$$

$$= \frac{91.128}{\sqrt{9433{,}75 \cdot 1.093.184}} = 0{,}897$$

Reta de Regressão Linear

- Diferentes retas podem ser traçadas, a olho nu, e um diagrama de dispersão
  - Cada pessoa terá uma tendência diferente
- Nenhuma reta passará exatamente por todos os pontos (se a correlação não for máxima)
- Precisamos encontrar uma reta que esteja tão próxima dos pontos quanto possível
- Os erros de predição para a reta são erros em *y* (direção vertical)



Reta de Regressão Linear

- Se um diagrama de dispersão sugere uma relação linear, é de interesse representar este padrão através de uma *reta*
- Usa-se o *método dos mínimos quadrados* para ajustar uma *reta de regressão* ao conjunto de pontos do diagrama
- A *reta de regressão* descreve como uma variável resposta (*dependente*) *y* varia em relação a uma variável explanatória (*independente*) *x*

Variáveis

- Variável <u>resposta *(y) (dependente)*</u>
  - Mede um resultado em um estudo
- Variável <u>explanatória *(x) (independente)*</u>
  - Procura explicar os resultados observados

Variável independente (x)	**Variável dependente (y)**
Temperatura do forno (°C)	Resistência mecânica da cerâmica (MPa)
Quantidade de aditivo (%)	Octanagem da gasolina
Renda (R$)	Consumo (R$)
Memória RAM (GB)	Tempo de resposta do sistema (s)

Definição

- Dada uma coleção de dados amostrais emparelhados, a seguinte ***equação de regressão*** descreve a relação entre as duas variáveis

$$\hat{y} = a + bx$$

- O gráfico da equação é chamado ***reta de regressão*** (ou reta de melhor ajuste, ou reta de mínimos quadrados)

Definição

$$\hat{y} = a + bx$$

$$b = \frac{n\left(\sum x_i y_i\right) - \left(\sum x_i\right)\left(\sum y_i\right)}{n\left(\sum x_i^2\right) - \left(\sum x_i\right)^2}$$

$$a = \frac{\left(\sum y_i\right)\left(\sum x_i^2\right) - \left(\sum x_i\right)\left(\sum x_i y_i\right)}{n\left(\sum x_i^2\right) - \left(\sum x_i\right)^2}$$

$$a = \frac{\sum y_i - b\sum x_i}{n}$$

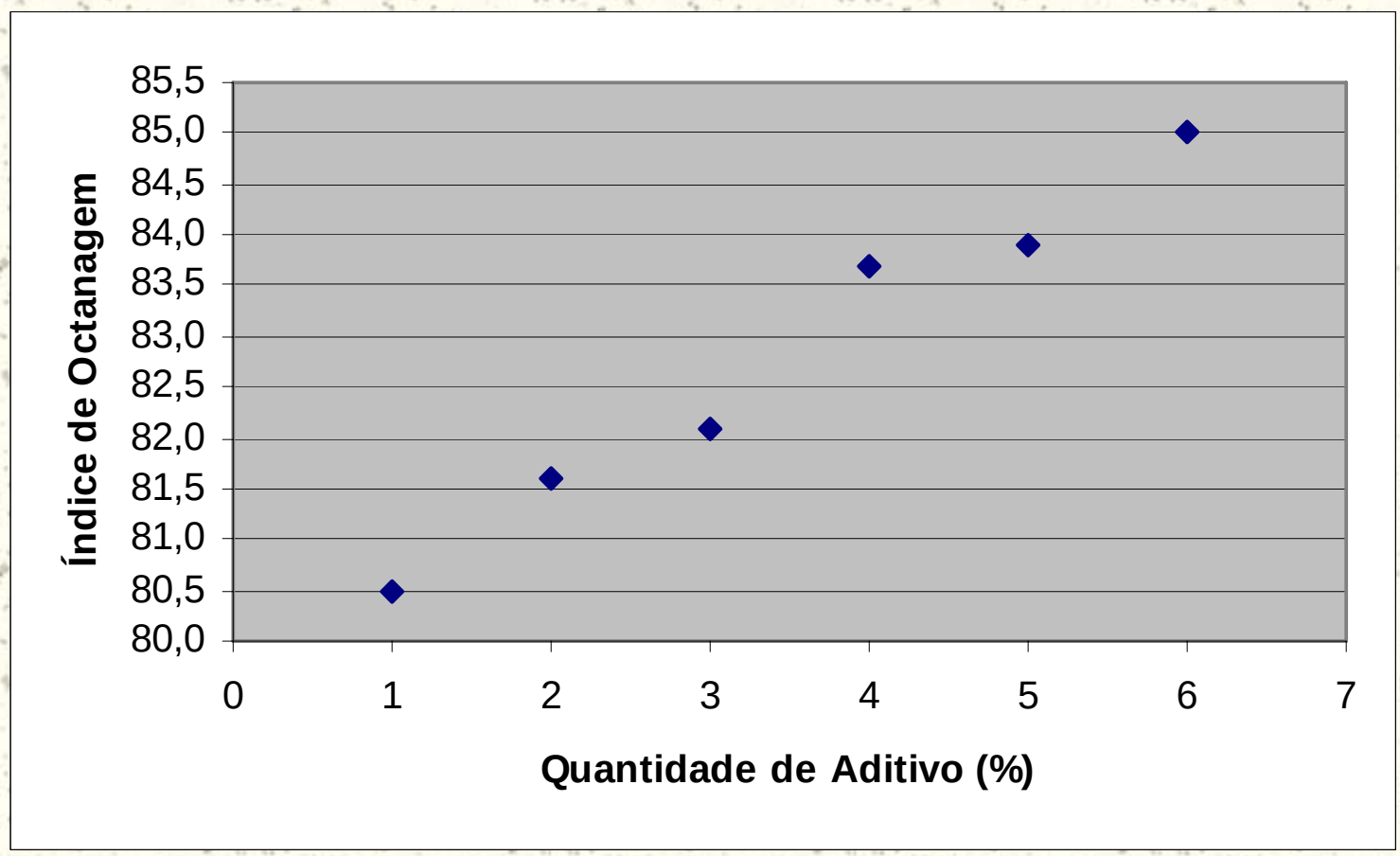
- b: coeficiente angular
- a: ponto onde a reta intercepta eixo y

Exemplo

- Considere um experimento em que se analisa a octanagem da gasolina (Y) em função da adição de um aditivo (X). Para isto, foram realizados ensaios com os percentuais de 1, 2, 3, 4, 5 e 6% de aditivo. Os resultados seguem.

Exemplo

X	Y
1	80,5
2	81,6
3	82,1
4	83,7
5	83,9
6	85,0



Exemplo

Calculando a equação de regressão...

	x_i	y_i	x_i^2	$x_i y_i$
	1	80,5	1	80,5
	2	81,6	4	163,2
	3	82,1	9	246,3
	4	83,7	16	334,8
	5	83,9	25	419,5
	6	85,0	36	510,0
Soma	21	496,8	91	1.754,3

$$b = \frac{6(1754,3) - (21)(496,8)}{6(91) - (21)^2} = \frac{93}{105} = 0,886$$

$$a = \frac{496,8 - (0,886)(21)}{6} = 79,7$$

$$\therefore \hat{y} = 79,7 + 0,886x$$

Exemplo

$$\hat{y} = 79{,}7 + 0{,}886x$$