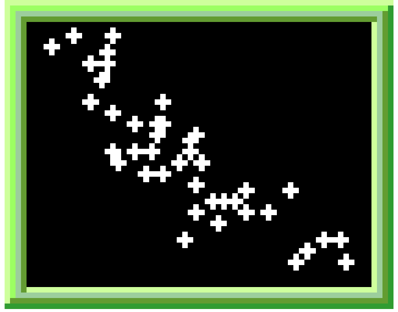


Correlação

Prof. Lorí Viali, Dr.

vialli@mat.ufrgs.br

http://www.mat.ufrgs.br/~vialli/

É o grau de associação entre duas ou mais variáveis. Pode ser:

correlacional

ou

experimental.

Numa relação experimental os valores de uma das variáveis são controlados.

No relacionamento correlacional, por outro lado, não se tem nenhum controle sobre as variáveis sendo estudadas.

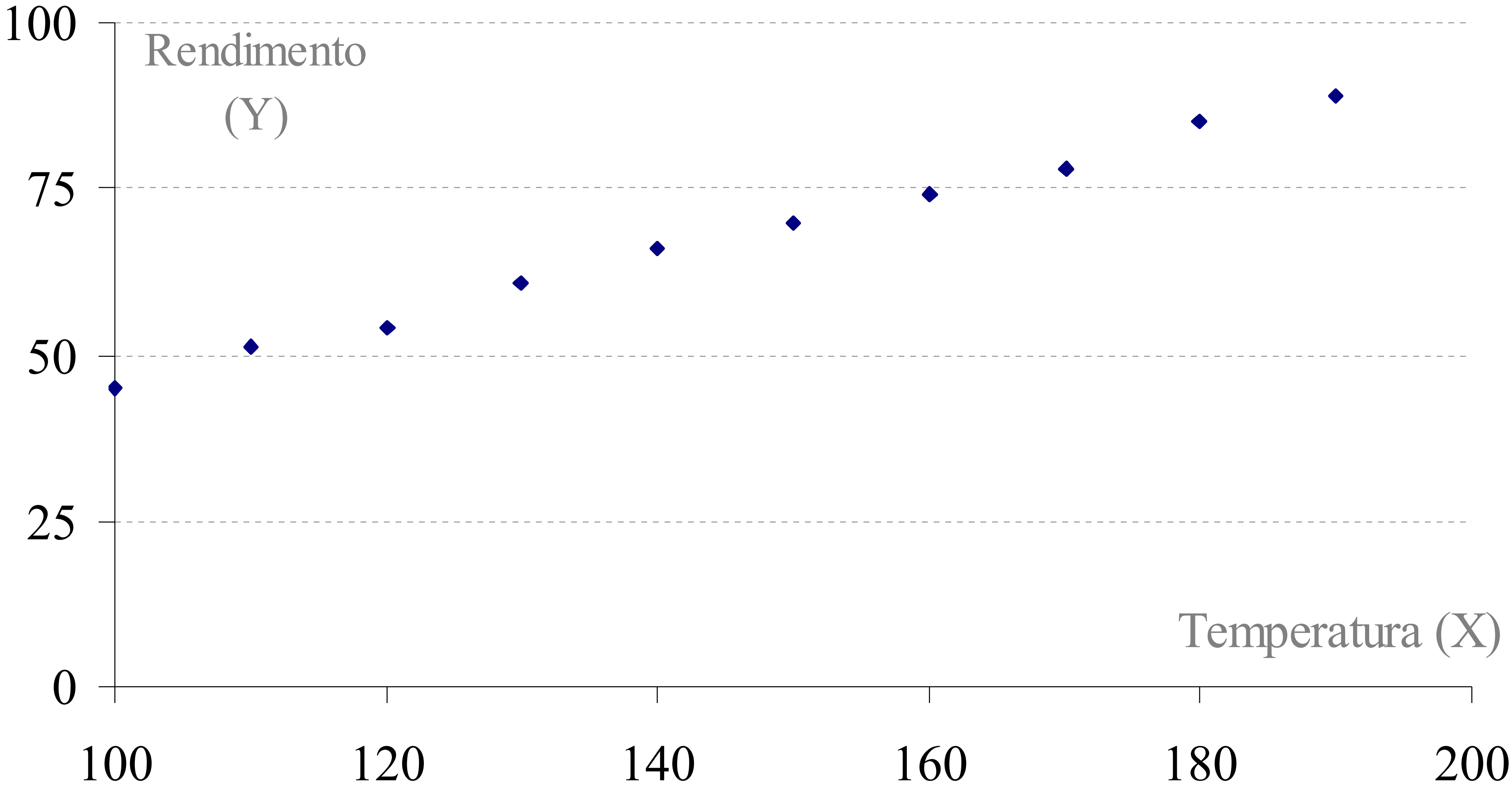
Indicadores de Associação

Um engenheiro químico está investigando o efeito da temperatura de operação do processo no rendimento do produto. O estudo resultou nos dados da tabela seguinte:

Temperatura, C^0 (X)	Rendimento (Y)
100	45
110	51
120	54
130	61
140	66
150	70
160	74
170	78
180	85
190	89

O primeiro passo para determinar se existe relacionamento entre as duas variáveis é obter o diagrama de dispersão (*scatter diagram*).

Diagrama de Dispersão



O diagrama de dispersão fornece uma idéia do tipo de relacionamento entre as duas variáveis. Neste caso, percebe-se que existe um relacionamento linear.

Quando o relacionamento entre duas variáveis quantitativas for do tipo linear, ele pode ser medido através do:

Coeficiente de Correlação

Observado um relacionamento linear entre as duas variáveis é possível determinar a intensidade deste relacionamento. O coeficiente que mede este relacionamento é denominado de Coeficiente de Correlação (linear).

Quando se está trabalhando com amostras o coeficiente de correlação é indicado pela letra "r" e é uma estimativa do coeficiente de correlação populacional que é representado por "ρ" (rho).

Determinação do Coeficiente de Correlação

Para determinar o coeficiente de correlação (grau de relacionamento linear entre duas variáveis) vamos determinar inicialmente a variação conjunta entre elas, isto é, a covariância.

A covariância entre duas variáveis X e Y, é representada por "Cov(X; Y)" e calculada por:

$$Cov(X,Y) = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

Mas

$$\sum (X_i - \overline{X})(Y_i - \overline{Y}) =$$

$$= \sum [X_i Y_i - \overline{X} Y_i - X_i \overline{Y} + \overline{X}\,\overline{Y}] =$$

$$= \sum X_i Y_i - \sum \overline{X} Y_i - \sum X_i \overline{Y} + \sum \overline{X}\,\overline{Y} =$$

$$= \sum X_i Y_i - \overline{X} \sum Y_i - \overline{Y} \sum X_i + \sum \overline{X}\,\overline{Y} =$$

$$= \sum X_i Y_i - n\overline{X}\,\overline{Y} - n\overline{X}\,\overline{Y} + n\overline{X}\,\overline{Y} =$$

$$= \sum X_i Y_i - n\overline{X}\,\overline{Y}$$

Então:

$$Cov(X,Y) = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{n-1} =$$

$$= \frac{\sum X_i Y_i - n\overline{X}\overline{Y}}{n-1}$$

A covariância poderia ser utilizada para medir o grau e o sinal do relacionamento entre as duas variáveis, mas ela é difícil de interpretar por variar de -∞ a +∞. Assim vamos utilizar o coeficiente de correlação linear de Pearson.

O coeficiente de correlação linear (de Pearson) é definido por:

$$r = \frac{Cov(X,Y)}{S_X S_Y}$$

Onde:

$$Cov(X,Y) = \frac{\sum X_i Y_i - n\overline{X}\overline{Y}}{n-1}$$

$$S_X = \sqrt{\frac{\sum X_i^2 - n\,\overline{X}^2}{n-1}}$$

$$S_Y = \sqrt{\frac{\sum Y_i^2 - n\,\overline{Y}^2}{n-1}}$$

Esta expressão não é muito prática para calcular manualmente o coeficiente de correlação. Pode-se obter uma expressão mais conveniente para o cálculo manual e o cálculo de outras medidas necessárias mais tarde.

Tem-se:

$$r = \frac{Cov(X,Y)}{S_X S_Y} =$$

$$= \frac{\dfrac{\sum X_i Y_i - n\overline{X}\,\overline{Y}}{n-1}}{\sqrt{\dfrac{\sum X_i^2 - n\overline{X}^2}{n-1}}\sqrt{\dfrac{\sum Y_i^2 - n\overline{Y}^2}{n-1}}} =$$

$$= \frac{\sum X_i Y_i - n\overline{X}\,\overline{Y}}{\sqrt{\left(\sum X_i^2 - n\overline{X}^2\right)\left(\sum Y_i^2 - n\overline{Y}^2\right)}}$$

F
a
z
e
n
d
o

$$S_{XY} = \sum X_i Y_i - n\overline{X}\overline{Y}$$

$$S_{XX} = \sum X_i^2 - n\,\overline{X}^2$$

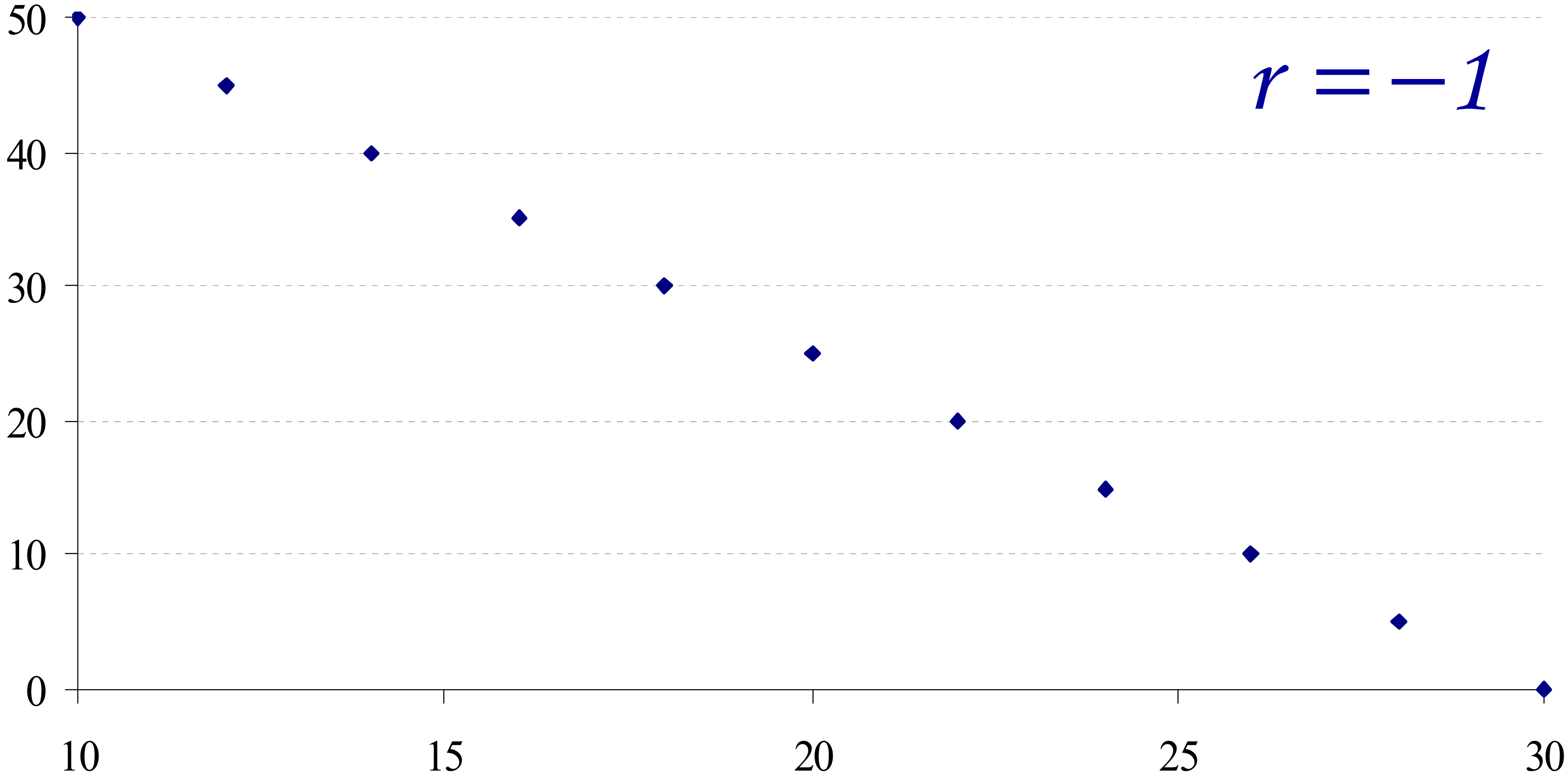
$$S_{YY} = \sum Y_i^2 - n\,\overline{Y}^2$$

$$Tem-se: \quad r = \frac{S_{XY}}{\sqrt{S_{XX} \cdot S_{YY}}}$$

A vantagem do coeficiente de correlação (de Pearson) é ser adimensional e variar de – 1 a + 1, que o torna de fácil interpretação.

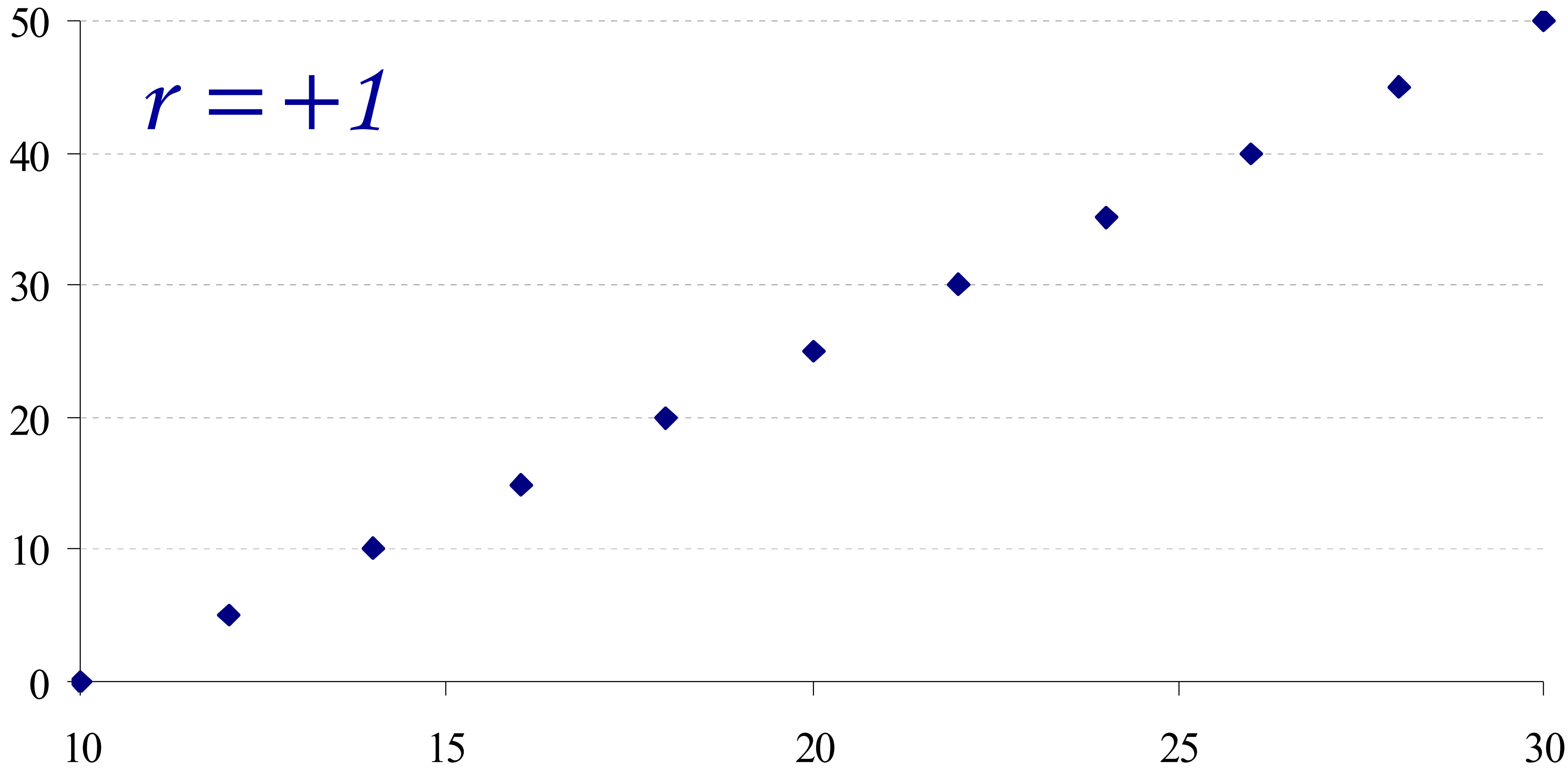
Assim se $r = -1$, temos uma relacionamento linear negativo perfeito, isto é, os pontos estão todos alinhados e quando X aumenta Y decresce e vice-versa.

Correlação perfeita e negativa



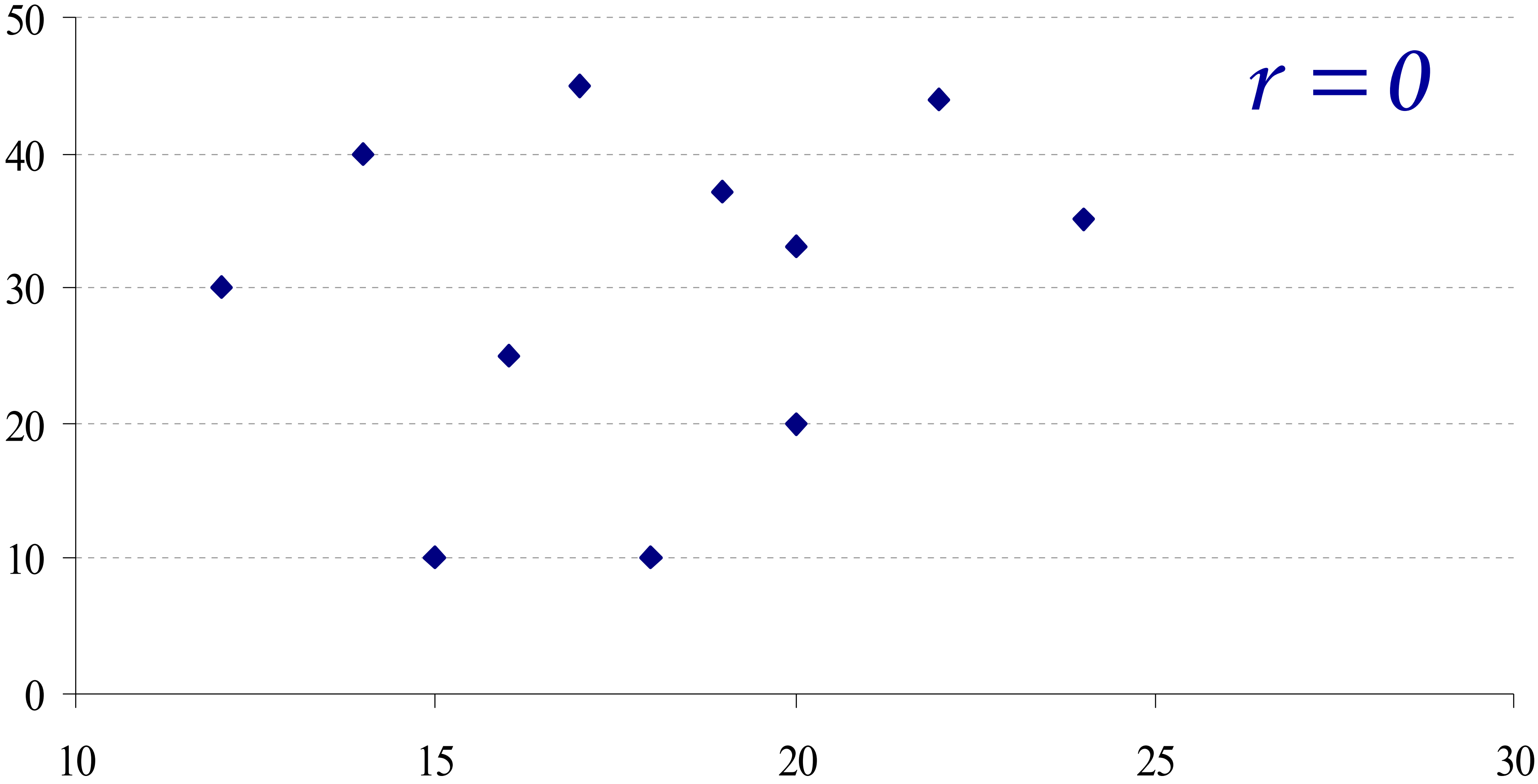
Se $r = +1$, temos uma relacionamento linear positivo perfeito, isto é, os pontos estão todos alinhados e quando X aumenta Y também aumenta.

Correlação perfeita e positiva



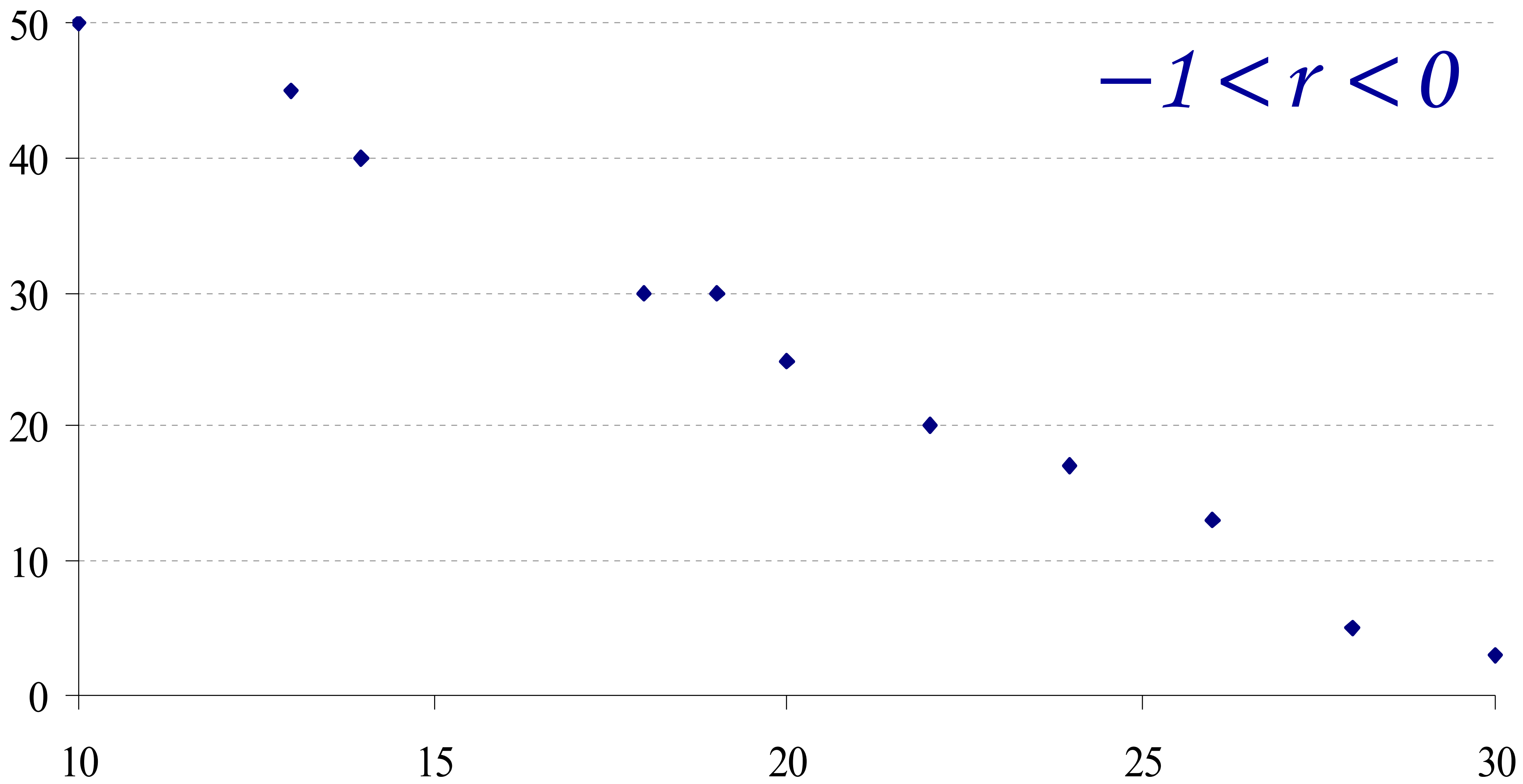
Assim se r = 0, temos uma ausência de relacionamento linear, isto é, os pontos não mostram "alinhamento".

Correlação nula



Assim se $-1 < r < 0$, temos uma relacionamento linear negativo, isto é, os pontos estão mais ou menos alinhados e quando X aumenta Y decresce e vice-versa.

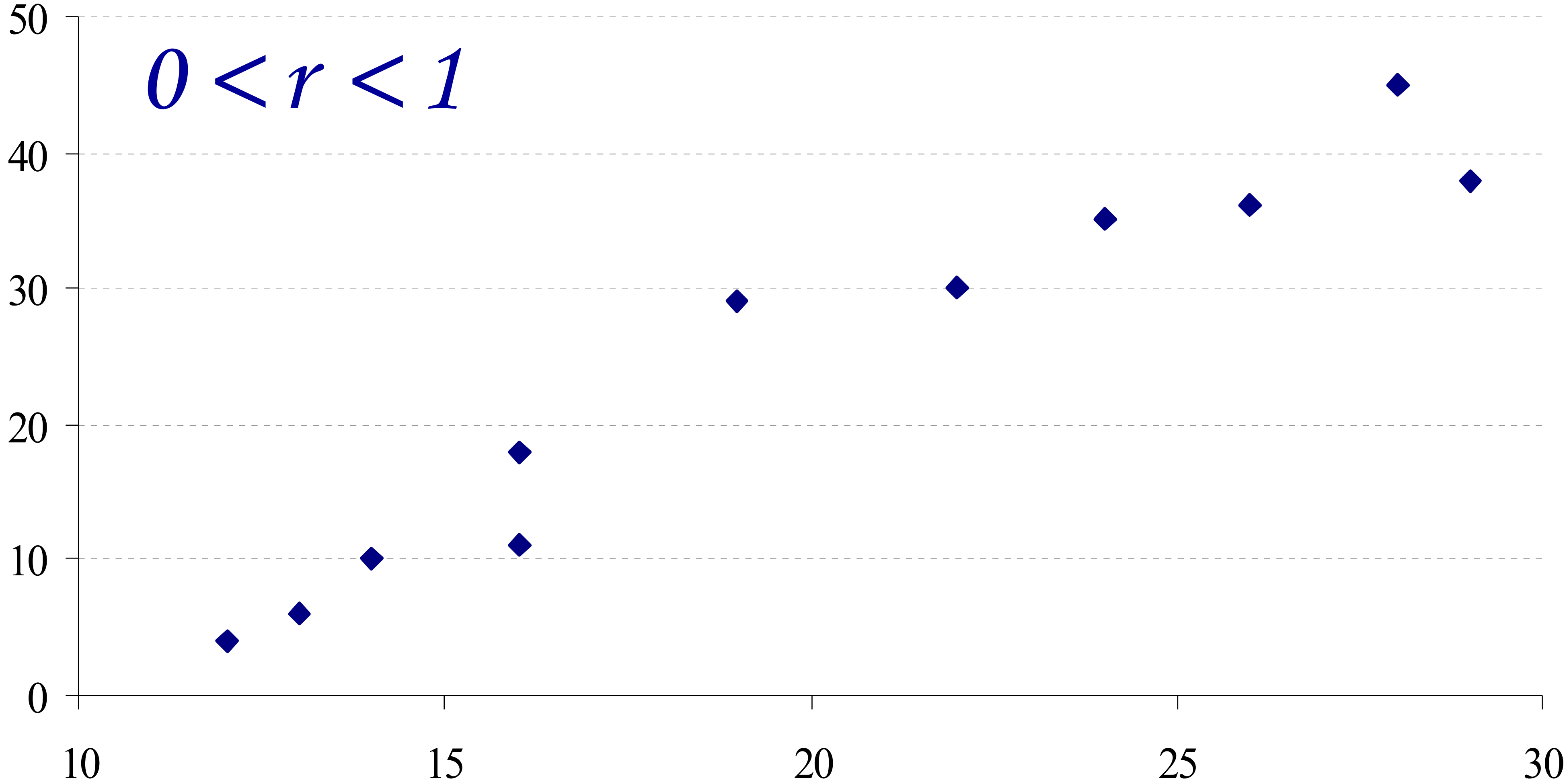
Correlação negativa



$$-1 < r < 0$$

Assim se $0 < r < 1$, temos uma relacionamento linear positivo, isto é, os pontos estão mais ou menos alinhados e quando X aumenta Y também aumenta.

Correlação positiva



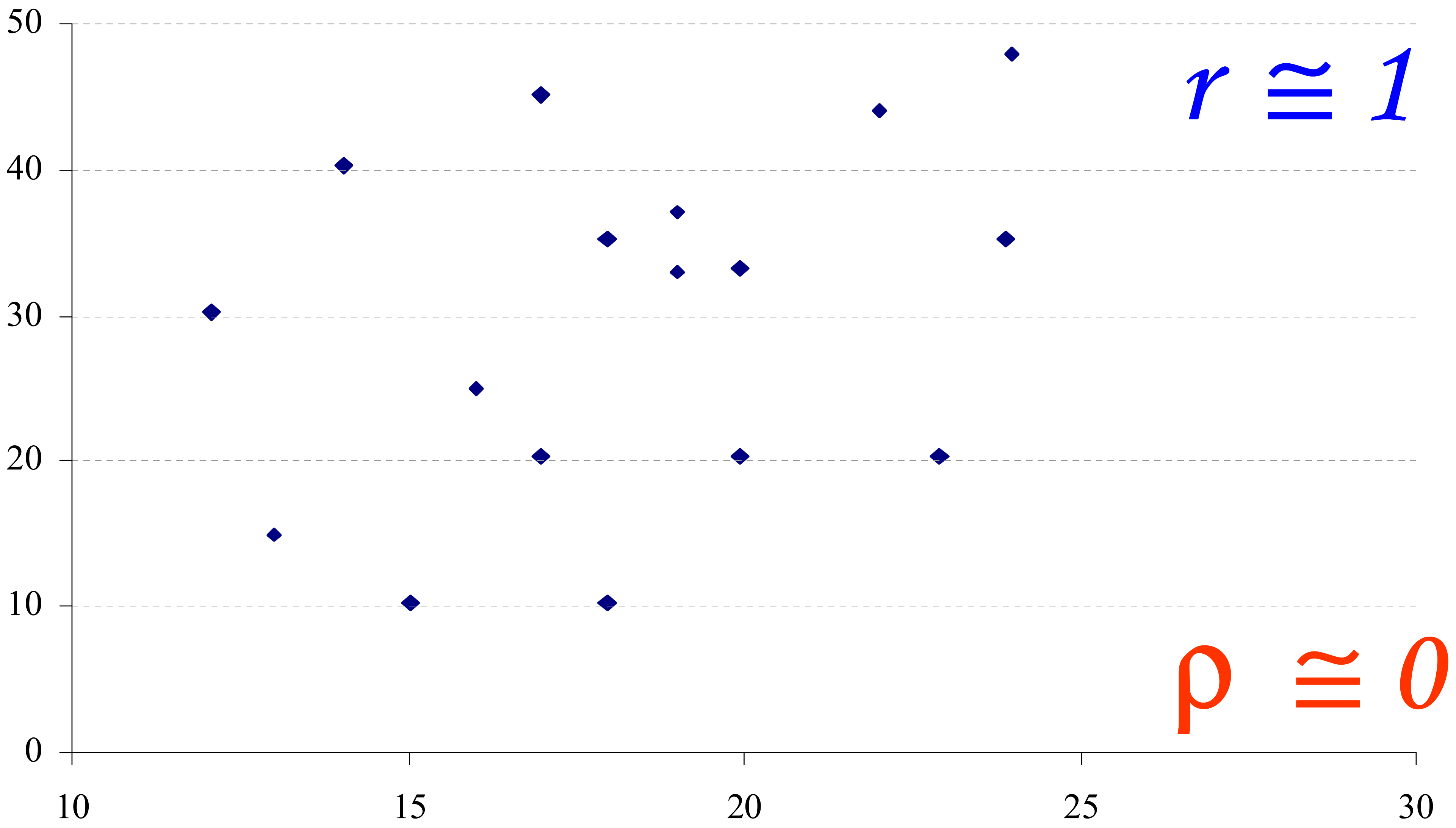
Observação:

Uma correlação amostral não significa necessariamente uma correlação populacional e vice-versa. É necessário testar o coeficiente de correlação para verificar se a correlação amostral é também populacional.

Ilustração

Observada uma amostra de seis pares, pode-se perceber que a correlação é quase um, isto é, $r \cong 1$. No entanto, observe o que ocorre quando mais pontos são acrescentados, isto é, quando se observa a população!

Correlação amostral X populacional



Exemplo

Determinar o "grau de relacionamento linear" entre as variáveis X = temperatura de operação do processo versus Y = rendimento do produto, conforme tabela.

X	Y	XY	X^2	Y^2
100	45	4500	10000	2025
110	51	5610	12100	2601
120	54	6480	14400	2916
130	61	7930	16900	3721
140	66	9240	19600	4356
150	70	10500	22500	4900
160	74	11840	25600	5476
170	78	13260	28900	6084
180	85	15300	32400	7225
190	89	16910	36100	7921
1450	673	101570	218500	47225

Vamos calcular "r" utilizando a expressão em destaque vista anteriormente, isto é, através das quantidades, S_{XY}, S_{XX} e S_{YY}.

Tem-se:

$$n = 10 \quad \sum X = 1450 \quad \sum Y = 673$$

$$\overline{X} = 145 \quad \overline{Y} = 67{,}3 \quad \sum XY = 101570$$

$$\sum X^2 = 218500 \qquad \sum Y^2 = 47225$$

Então:

$$S_{XY} = \sum X_i Y_i - n\overline{X}\overline{Y} =$$

$$= 101570 - 10.145.67{,}3 =$$

$$= 3985$$

$$S_{XX} = \sum X_i^2 - n\overline{X}^2 =$$

$$= 218500 - 10.145^2 =$$

$$= 8250$$

$$S_{YY} = \sum Y_i^2 - n\overline{Y}^2 =$$

$$= 47225 - 10.67{,}3^2 =$$

$$= 1932{,}10$$

$$r = \frac{S_{XY}}{\sqrt{S_{XX} \cdot S_{YY}}} =$$

$$= \frac{3985}{\sqrt{8250.1932,10}} =$$

$$= 0,9981$$

Apesar de "r" ser um valor adimensional, ele não é uma taxa. Assim o resultado não deve ser expresso em percentagem.