Correlação

Prof. Lorí Viali, Dr.

viali@mat.ufrgs.br

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

5/5

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.

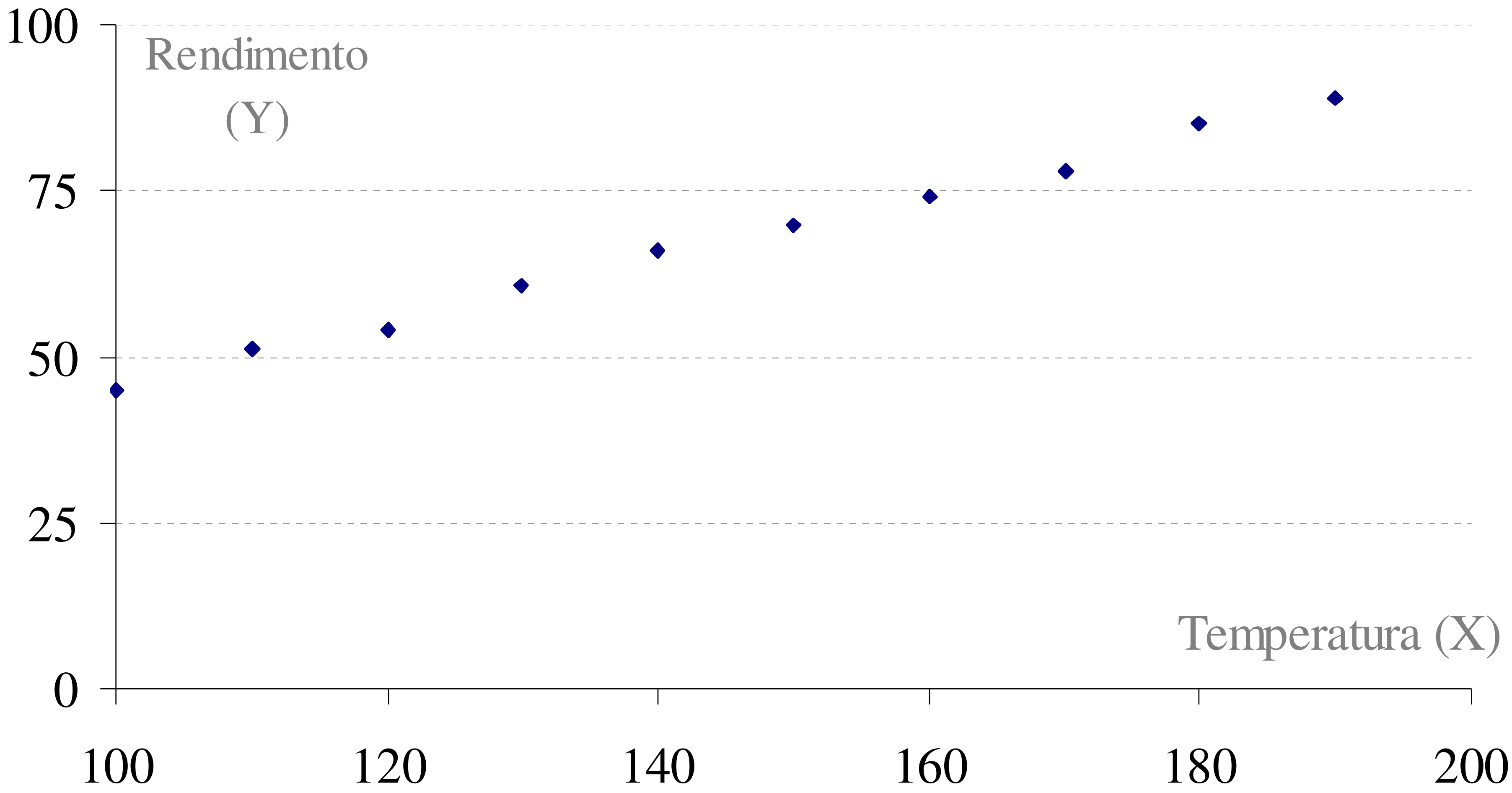
Indicadores de Associação

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

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

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

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

$$= \sum X_iY_i - n\overline{X}\overline{Y} - n\overline{X}\overline{Y} + n\overline{X}\overline{Y} =$$

$$= \sum X_iY_i - n\overline{X}\overline{Y}$$

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 $-\infty$ a $+\infty$. 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: $r = \dfrac{Cov(X,Y)}{S_X S_Y} =$

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

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

F
a
z
e
n
d
o

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

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

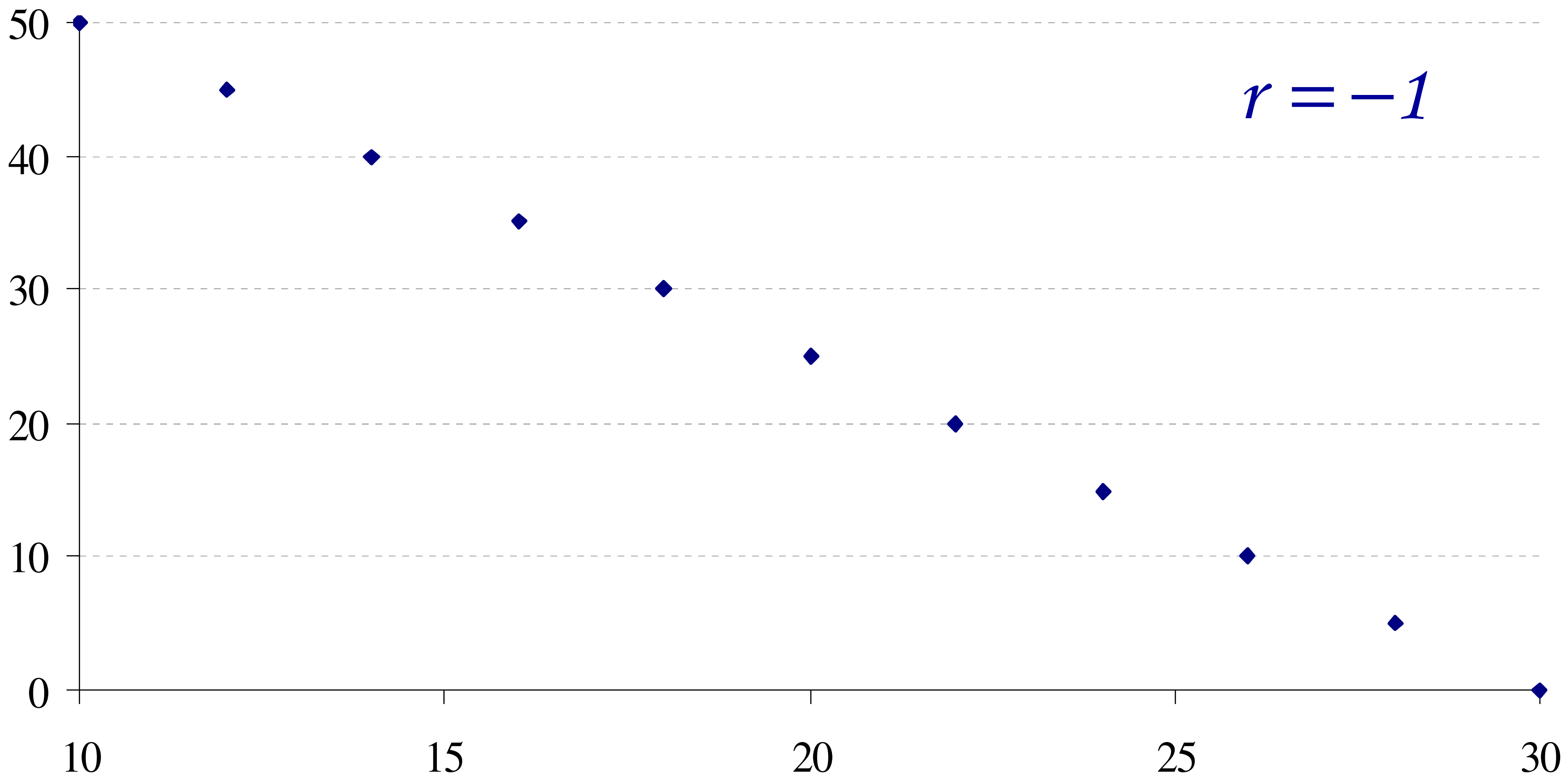
$$S_{YY} = \sum Y_i^2 - n\overline{Y}^2$$

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

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

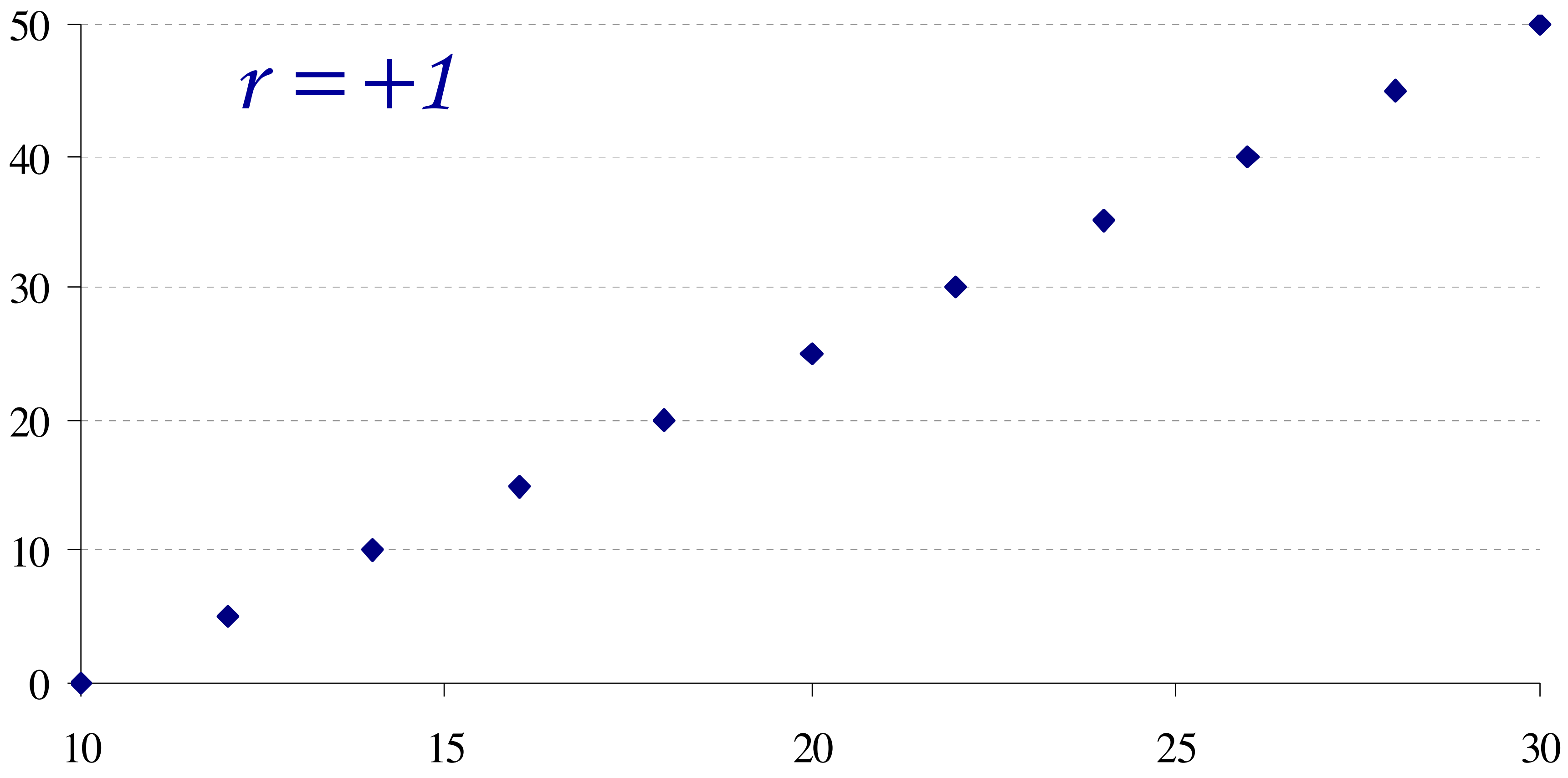
Assim se r = -1, temos uma relacionamento linear negativo perfeito, isto é, os pontos estão todos alinhados e quando X aumenta Y decresce e vice-versa.

Correlação perfeita e negativa



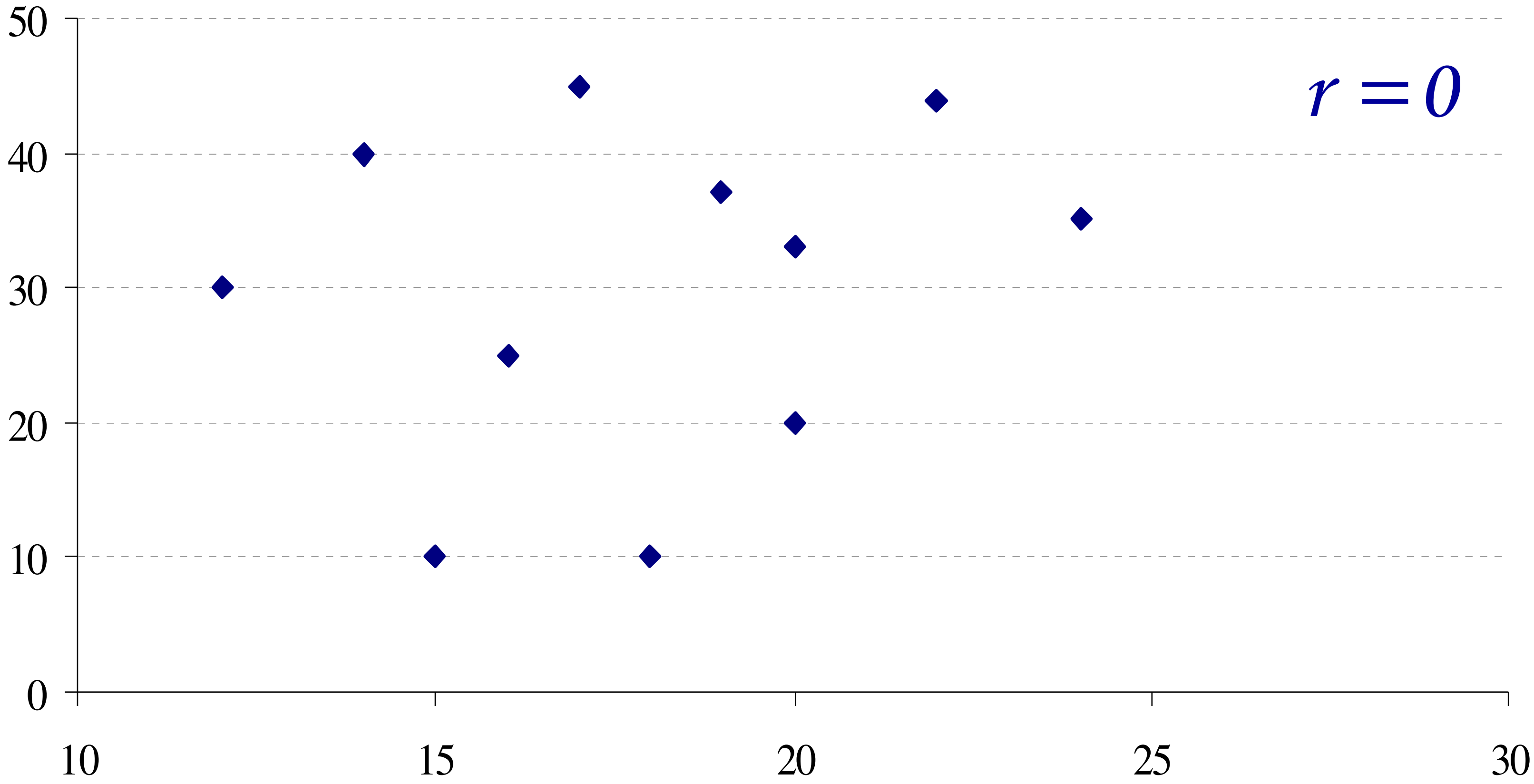
Se $r = +1$, temos uma relacionamento linear positivo perfeito, isto é, os pontos estão todos alinhados e quando X aumenta Y também aumenta.

Correlação perfeita e positiva



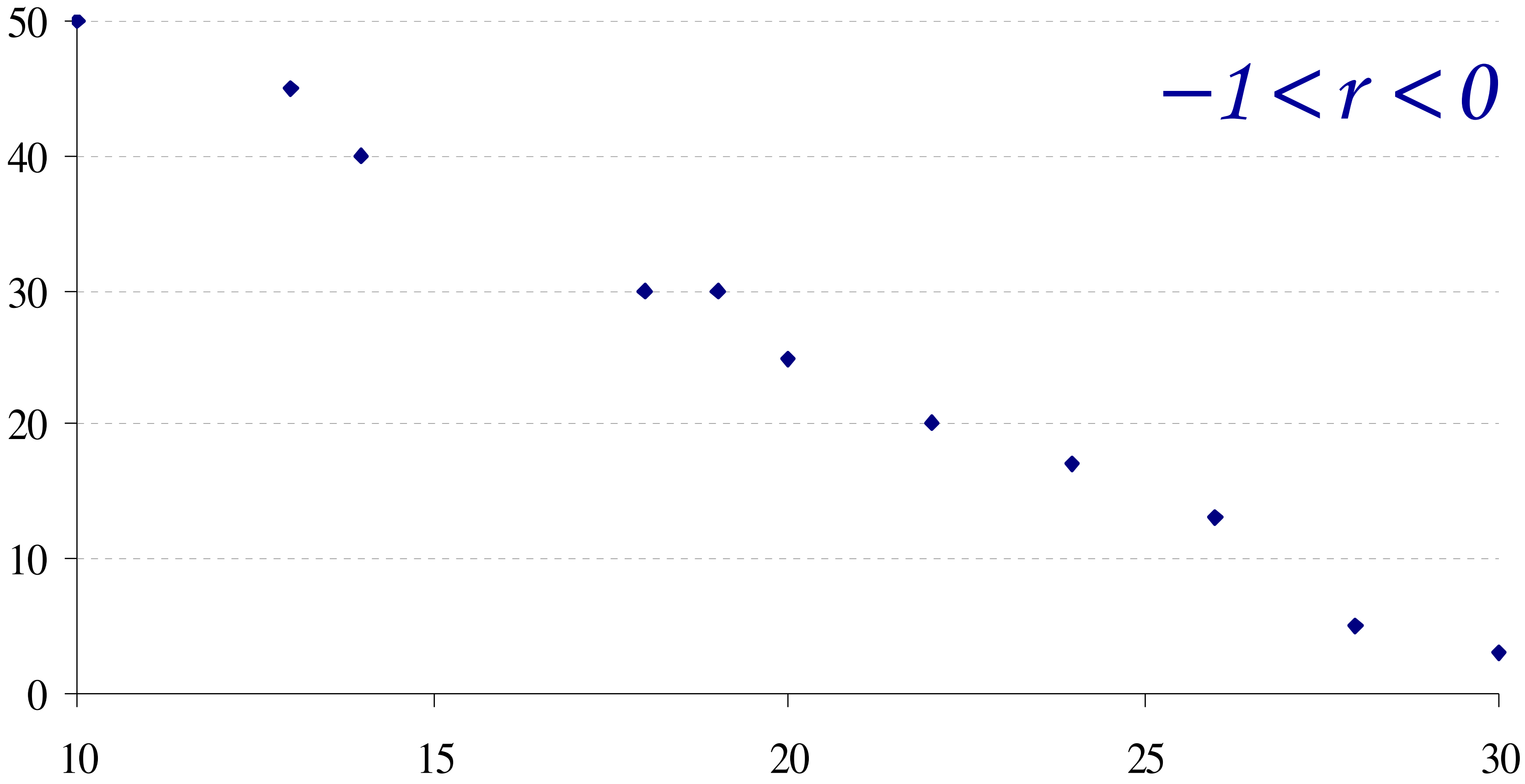
Assim se $r = 0$, temos uma ausência de relacionamento linear, isto é, os pontos não mostram "alinhamento".

Correlação nula



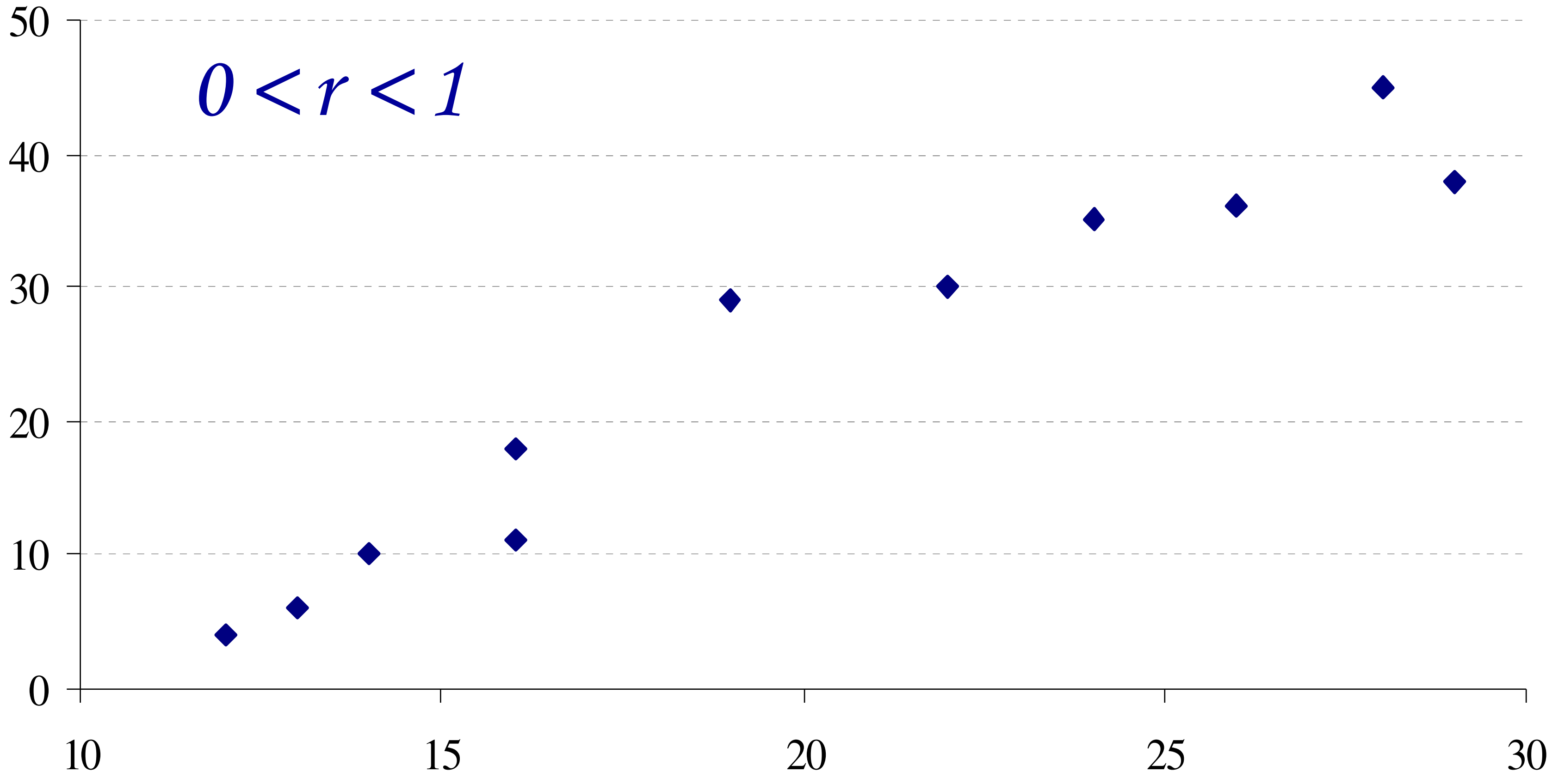
Assim se $-1 < r < 0$, temos uma relacionamento linear negativo, isto é, os pontos estão mais ou menos alinhados e quando X aumenta Y decresce e vice-versa.

Correlação negativa



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



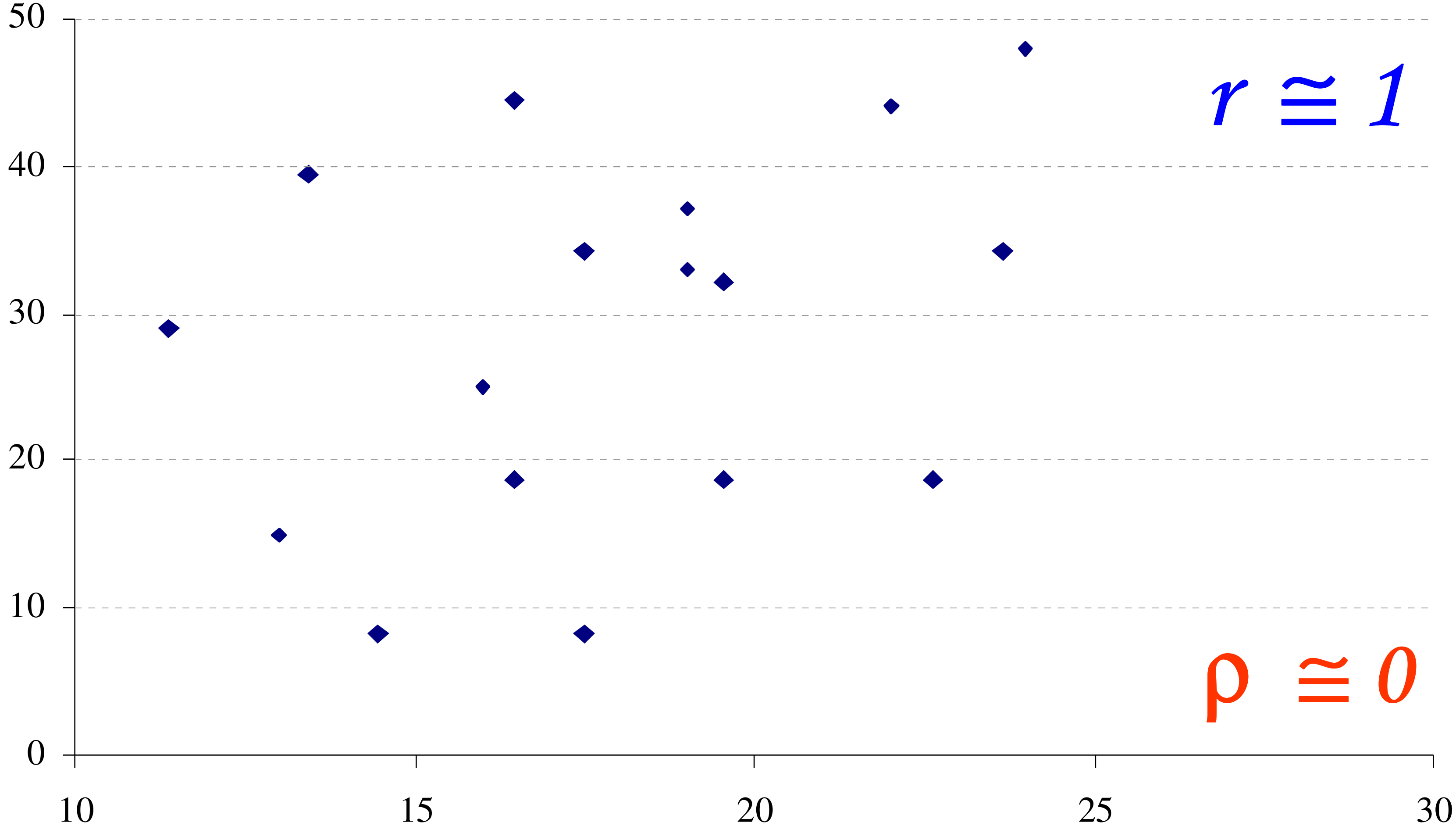
Observação:

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 = nota em Português Y = nota em Matemática, de 20 candidatos em um concurso vestibular com 30 questões, conforme tabela, na próxima lâmina.

	Português (X)	Matemática (Y)	X^2	Y^2	XY
1	23	26	529	676	598
2	13	13	169	169	169
3	15	21	225	441	315
4	4	6	16	36	24
5	12	17	144	289	204
6	17	31	289	961	527
7	11	11	121	121	121
8	16	27	256	729	432
9	12	22	144	484	264
10	16	29	256	841	464
11	11	14	121	196	154
12	20	29	400	841	580
13	14	4	196	16	56
14	19	29	361	841	551
15	6	8	36	64	48
16	19	25	361	625	475
17	9	11	81	121	99
18	15	13	225	169	195
19	16	22	256	484	352
20	16	13	256	169	208
Total	284	371	4442	8273	5836

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

Tem-se: $n = 20 \quad \sum X = 284 \quad \sum Y = 371$

$$\bar{X} = 14{,}20 \quad \bar{Y} = 18{,}55 \quad \sum XY = 5836$$

$$\sum X^2 = 4442 \quad \sum Y^2 = 8273$$

Então: $S_{XY} = \sum X_i Y_i \quad n\bar{X}\bar{Y} =$

$$= 5836 - 20.14{,}20.18{,}55 =$$

$$= 567{,}80.$$

$$S_{XX} = \sum X_i^2 - n\,\overline{X}^2 =$$
$$= 4442 - 20.14{,}20^2 =$$
$$= 409{,}20$$

$$S_{YY} = \sum Y_i^2 - n\,\overline{Y}^2 =$$
$$= 8273 - 20.18{,}55^2 =$$
$$= 1390{,}95.$$

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

$$= \frac{567,80}{\sqrt{409,20.1390,95}} =$$

$$= 0,7526.$$

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