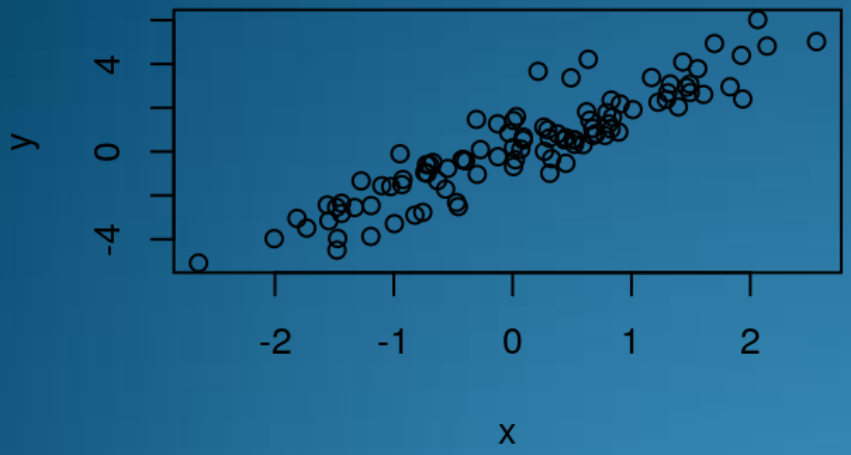
Correlação Linear Simples

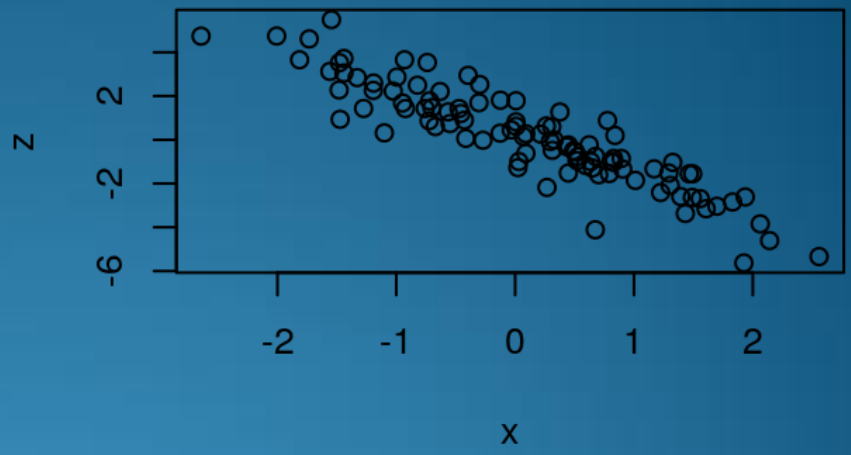
Nas áreas biológicas, em algumas situações, o pesquisador está interessado em estudar a maneira como duas variáveis *X* e *Y* estão associadas e, mais ainda, medir o seu grau de associação. Por exemplo, posso estar interessado em avaliar se existe associação entre o peso de vagem com semente e a largura da vagem ou entre o teor de cálcio no solo e a porcentagem de tubérculos maduros.

Para estudar a associação entre duas variáveis quantitativas, uma amostra aleatória é selecionada e as duas variáveis são observadas simultaneamente para cada indivíduo, animal ou planta. Uma maneira de descrever os dados conjuntamente é através do diagrama de dispersão, que é a representação gráfica dos pares de valores num sistema cartesiano.

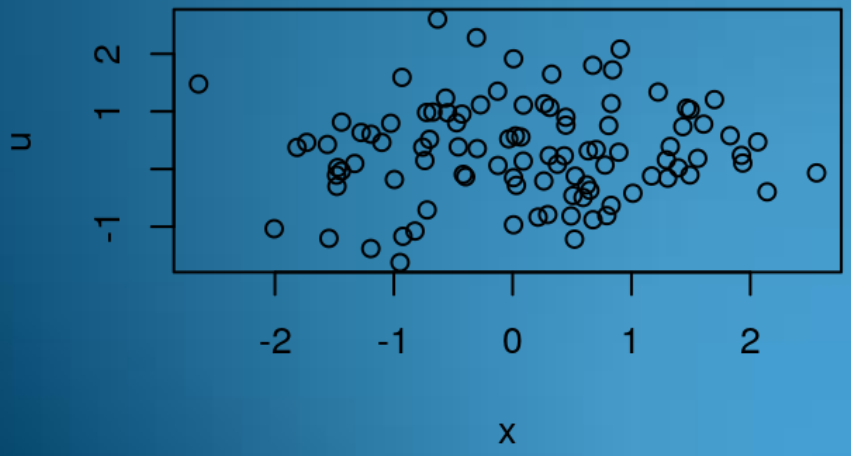
Relação linear positiva



Relação linear negativa

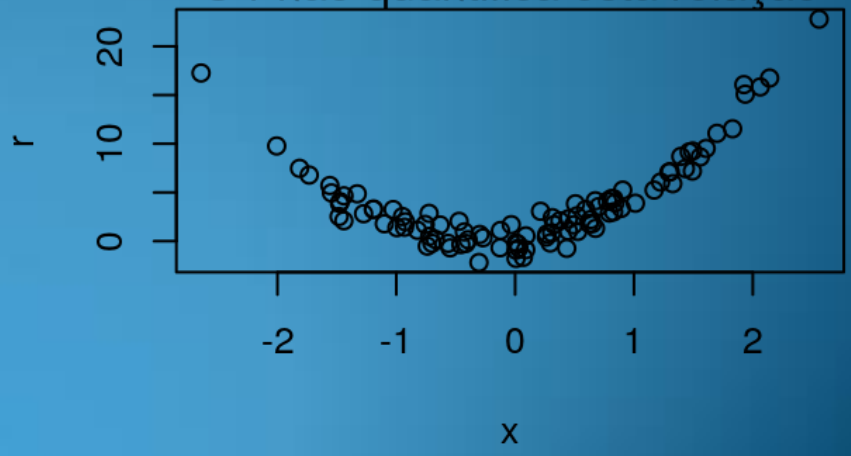


Ausência de relação



Relação não-linear

O r não quantifica esta relação



O objetivo do diagrama de dispersão é possibilitar a visualização da relação existente entre as variáveis *X* e *Y*. Se os pontos estiverem localizados na vizinhança de uma reta imaginária, há indicação de correlação. Se *X* e *Y* crescem no mesmo sentido, a indicação é no sentido de correlação positiva. Caso a variação aconteça no sentido oposto, existe correlação negativa entre as variáveis.

A inspeção visual no diagrama de dispersão mostra, de maneira subjetiva, a associação dos dados e por isso precisa ser quantificada. A força de uma associação pode ser medida pelo coeficiente de correlação de Pearson.

Esse coeficiente de correlação de Pearson, denotado por *r*, mede a intensidade de associação linear existente entre duas variáveis quantitativas.

O coeficiente de correlação de Pearson varia entre −1 e 1. Sua fórmula é dada por:

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

Não havendo relação linear alguma entre X e Y, $r = 0$. Se $r = -1$, existe uma correlação linear perfeita negativa (ou seja, todos os pontos estão sobre uma linha reta decrescente). Se $r = 1$, existe uma correlação linear perfeita positiva (isto é, todos os pontos estão sobre uma linha reta crescente).

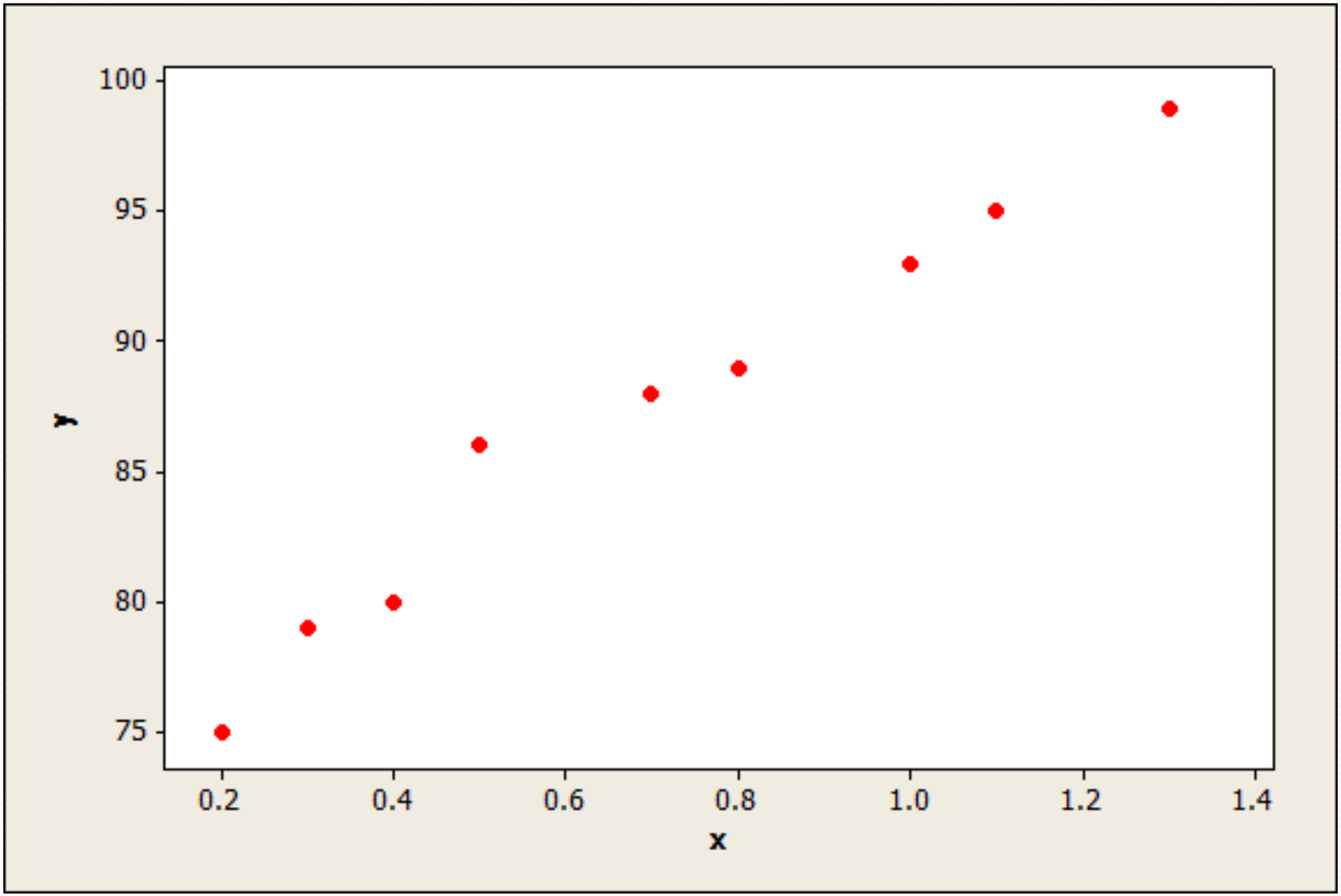
Se $0 < |r| < 0{,}3$, dizemos que há uma correlação fraca. Se $0{,}3 \leq |r| < 0{,}6$, dizemos que há uma correlação regular. Se $0{,}6 \leq |r| < 0{,}9$, dizemos que há uma correlação forte e se $0{,}9 \leq |r| < 1$, dizemos que há uma correlação muito forte.

Obs: Podemos calcular correlações para qualquer par de variáveis, mas sempre devemos ter cuidado ao assumir que uma causa variação na outra.

Exemplo: Em um experimento foram obtidos os resultados para teor de cálcio no solo (X), em *meq*/100 cm^3 e a porcentagem de tubérculos maduros (Y). Faça o diagrama de dispersão e calcule o coeficiente de correlação de Pearson.

X	0,2	0,3	0,4	0,5	0,7	0,8	1,0	1,1	1,3
Y	75	79	80	86	88	89	93	95	99

Diagrama de Dispersão



Cálculo de r:

n = 9 $\quad \sum x_i = 6{,}3 \quad \sum y_i = 784$

$\sum x_i y_i = 572{,}7 \quad \sum x_i^2 = 5{,}57 \quad \sum y_i^2 = 68802$

$$r = \frac{9 \text{ x } 572{,}7 - 6{,}3 \text{ x } 784}{\sqrt{[9 \text{ x } 5{,}57 - (6{,}3)^2] \text{x} [9 \text{ x } 68802 - (784)^2]}}$$

$$r = \frac{5154{,}3 - 4939{,}2}{\sqrt{10{,}44 \text{ x } 4562}} = \frac{215{,}1}{218{,}24} = 0{,}9856$$

Quando se calcula o coeficiente de correlação *r* em uma amostra, é necessário ter em mente que se está, na realidade, estimando a associação verdadeira entre *X* e *Y* existente na população. A correlação na população é designada por ρ. Para avaliar a significância do coeficiente de correlação, geralmente testa-se as hipóteses:

H_0: $\rho = 0$ (não existe correlação)

H_1: $\rho \neq 0$ (existe correlação)

Para realizar o teste de hipóteses, tanto a variável X quanto a variável Y devem ter distribuição normal e a relação entre X e Y deve ser linear.

A estatística de teste é dada por:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

Rejeito a hipótese Ho se $|t| \geq t_{\frac{\alpha}{2};\, n-2}$.

No exemplo, temos:

$$t = 0{,}9856\sqrt{\frac{9-2}{1-(0{,}9856)^2}} = 15{,}42$$

Tabela da distribuição t de Student

	ALFA					
n	0,25	0,10	0,05	0,025	0,010	0,005
1	1,0000	3,0777	6,3137	12,7062	31,8210	63,6559
2	0,8165	1,8856	2,9200	4,3027	6,9645	9,9250
3	0,7649	1,6377	2,3534	3,1824	4,5407	5,8408
4	0,7407	1,5332	2,1318	2,7765	3,7469	4,6041
5	0,7267	1,4759	2,0150	2,5706	3,3649	4,0321
6	0,7176	1,4398	1,9432	2,4469	3,1427	3,7074
7	0,7111	1,4149	1,8946	2,3646	2,9979	3,4995
8	0,7064	1,3968	1,8595	2,3060	2,8965	3,3554
9	0,7027	1,3830	1,8331	2,2622	2,8214	3,2498
10	0,6998	1,3722	1,8125	2,2281	2,7638	3,1693
11	0,6974	1,3634	1,7959	2,2010	2,7181	3,1058
12	0,6955	1,3562	1,7823	2,1788	2,6810	3,0545
13	0,6938	1,3502	1,7709	2,1604	2,6503	3,0123
14	0,6924	1,3450	1,7613	2,1448	2,6245	2,9768
15	0,6912	1,3406	1,7531	2,1315	2,6025	2,9467

Ao nível de 5% de significância, temos que o valor na tabela *t de Student* é $t_{2,5\%;\ 7} = 2{,}3646$.

Como o valor obtido $t = 15{,}42$ é maior que o valor tabelado $t_{2,5\%;\ 7} = 2{,}3646$, rejeitamos H_0 e concluímos que existe uma correlação entre o teor de cálcio no solo e a porcentagem de tubérculos maduros. De acordo com a classificação do r dada anteriormente, essa correlação ($r = 0{,}9856$) é muito forte.