Associação entre Variáveis Quantitativas

Um dispositivo bastante útil para se verificar a associação entre duas variáveis quantitativas, ou entre dois conjuntos de dados, é o *gráfico de dispersão*, que vamos introduzir por meio de exemplos.

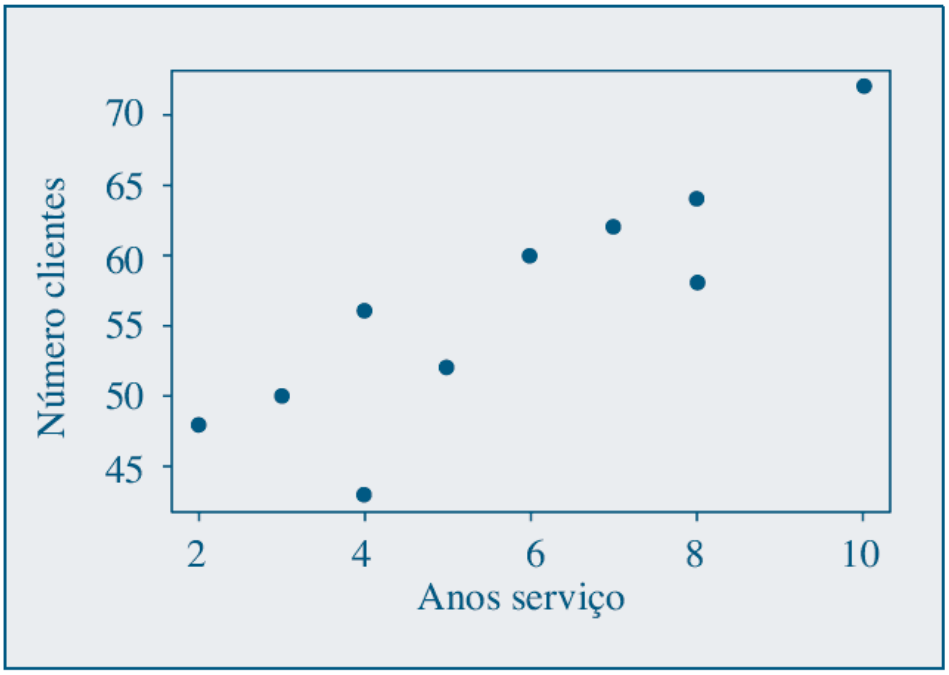
Exemplo 4.4 Na Figura 4.2, temos o gráfico de dispersão das variáveis *X* e *Y* da Tabela 4.12. Nesse tipo de gráfico, temos os possíveis pares de valores (*x*, *y*), na ordem que aparecem.Para o exemplo, vemos que parece haver uma associação entre as variáveis, porque no conjunto, a medida que aumenta o tempo de serviço, aumenta o número de clientes.

Tabela 4.12 Número de anos de serviço (X) por número de clientes (Y) de agentes de uma companhia de seguros.

Agente	Anos de serviço (X)	Número de clientes (Y)
A	2	48
B	3	50
C	4	56
D	5	52
E	4	43
F	6	60
G	7	62
H	8	58
I	8	64
J	10	72

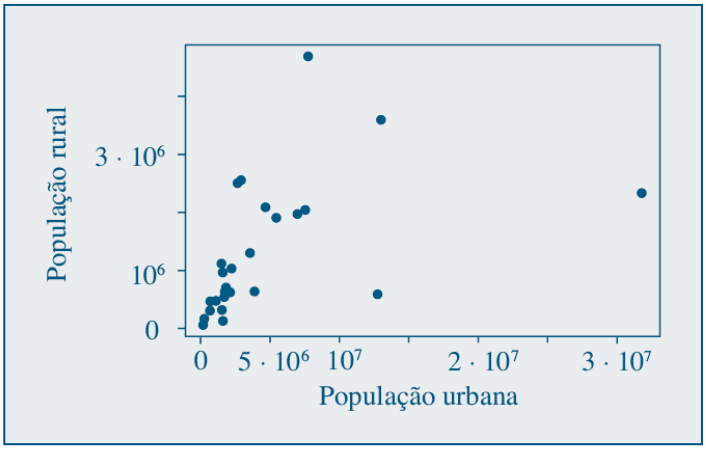
Fonte: Dados hipotéticos.

Figura 4.2 Gráfico de dispersão para as variáveis X: anos de serviço e Y: número de clientes.



Exemplo 4.5 Consideremos os dados das variáveis X: população urbana e Y: população rural, no Brasil, em 1996. O gráfico de dispersão está na Figura 4.3. Vemos que parece não haver associação entre as variáveis, pois os pontos não apresentam nenhuma tendência particular.

Figura 4.3 Gráfico de dispersão para as variáveis X: população urbana e Y: população rural.



Exemplo 4.6 Consideremos agora as duas situações abaixo e os respectivos gráficos de dispersão.

(a) Numa pesquisa feita com dez famílias com renda bruta mensal entre 10 e 60 salários mínimos, mediram-se:

X: renda bruta mensal (expressa em número de salários mínimos).

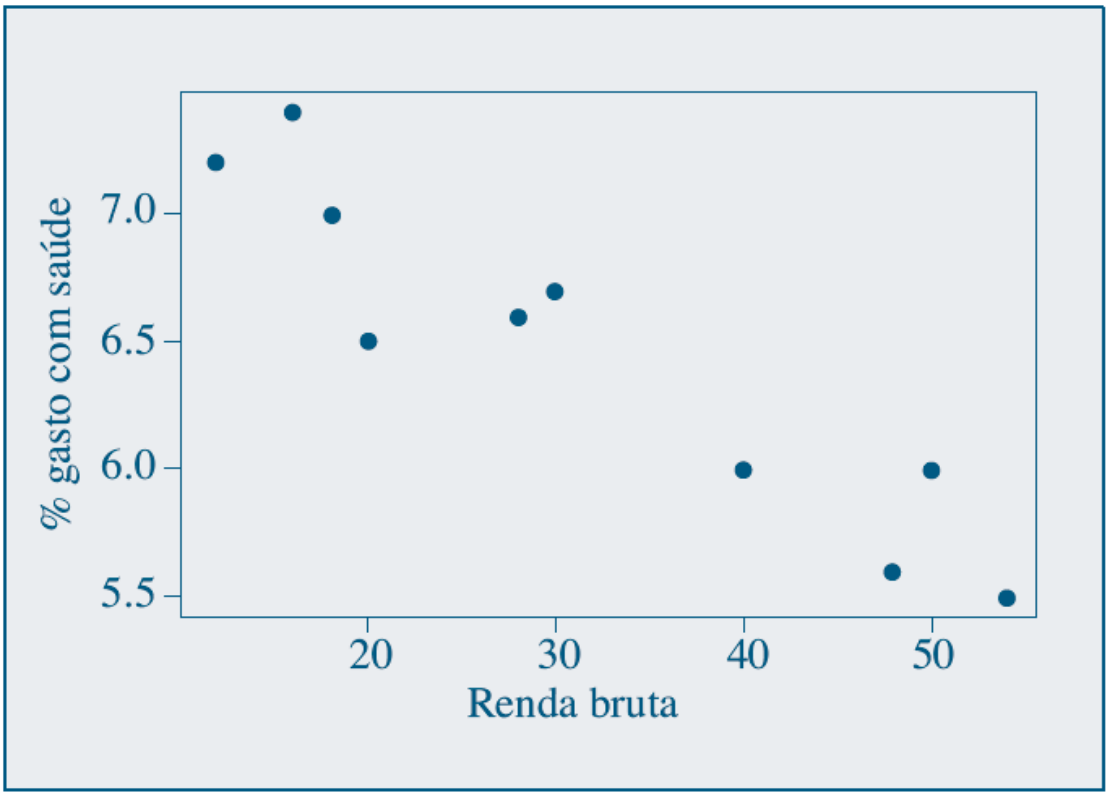
Y: a porcentagem da renda bruta anual gasta com assistência médica; os dados estão na Tabela 4.13. Observando o gráfico de dispersão (Figura 4.4), vemos que existe uma associação "inversa", isto é, aumentando a renda bruta, diminui a porcentagem sobre ela gasta em assistência médica.

Tabela 4.13 Renda bruta mensal (X) e porcentagem da renda gasta em saúde (Y) para um conjunto de famílias.

Família	X	Y
A	12	7,2
B	16	7,4
C	18	7,0
D	20	6,5
E	28	6,6
F	30	6,7
G	40	6,0
H	48	5,6
I	50	6,0
J	54	5,5

Fonte: Dados hipotéticos.

Figura 4.4 Gráfico de dispersão para as variáveis X: renda bruta e Y: % renda gasta com saúde.



(b) Oito indivíduos foram submetidos a um teste sobre conhecimento de língua estrangeira e, em seguida, mediu-se o tempo gasto para cada um aprender a operar uma determinada máquina. As variáveis medidas foram:

X: resultado obtido no teste (máximo = 100 pontos);

Y: tempo, em minutos, necessário para operar a máquina satisfatoriamente.

Tabela 4.14 Resultado de um teste (X) e tempo de operação de máquina (Y) para oito indivíduos.

Indivíduo	X	Y
A	45	343
B	52	368
C	61	355
D	70	334
E	74	337
F	76	381
G	80	345
H	90	375

Fonte: Dados hipotéticos.

Figura 4.5 Gráfico de dispersão para as variáveis X: resultado no teste e Y: tempo de operação.

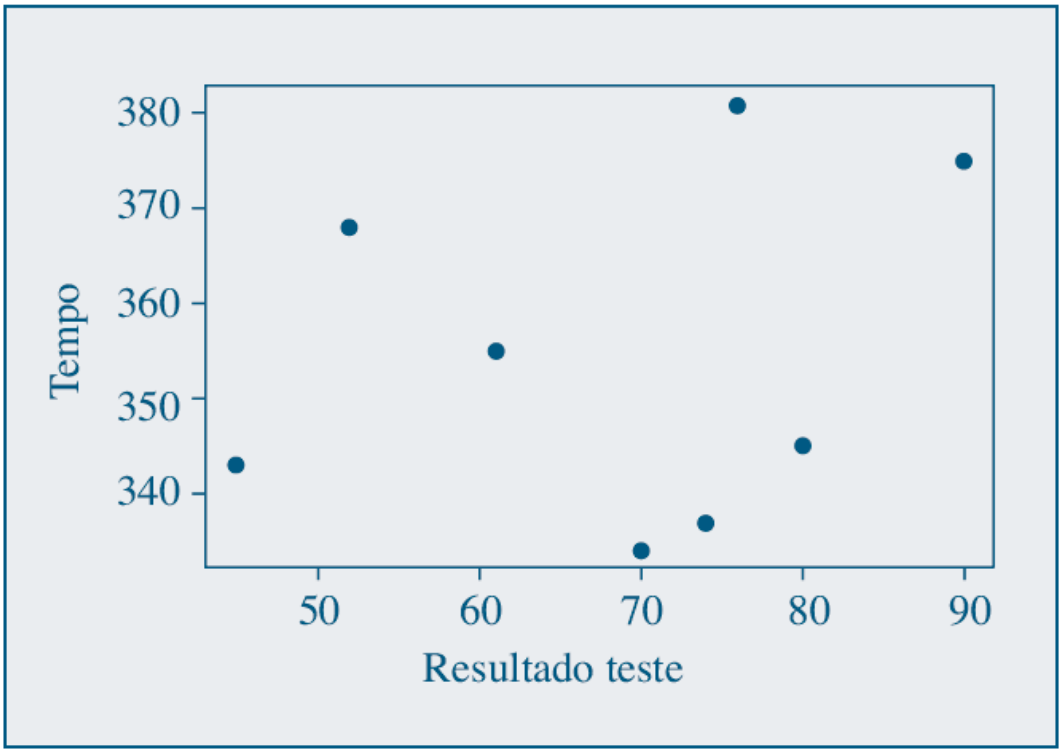


Figura 4.6 Tipos de associações entre duas variáveis.

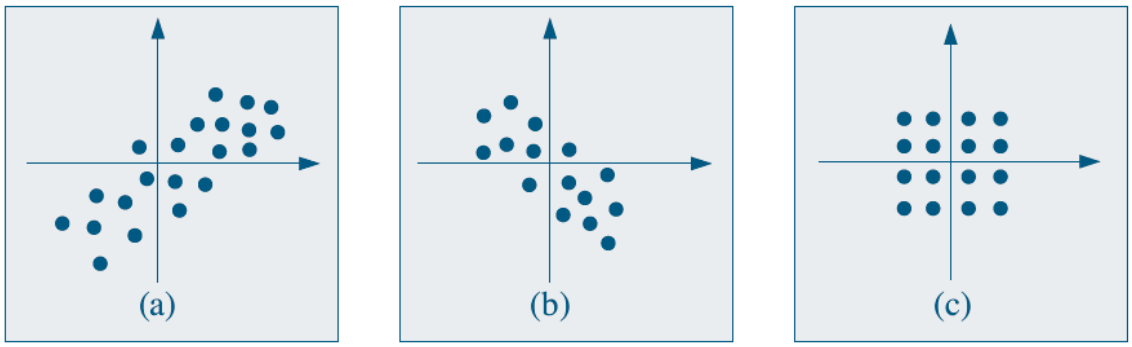


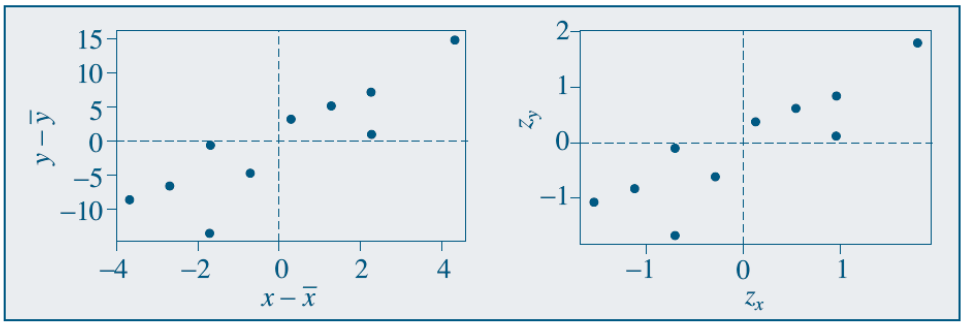
Tabela 4.15 Cálculo do coeficiente de correlação.

Agente	Anos x	Clientes y	$x-\bar{x}$	$y-\bar{y}$	$\frac{x-\bar{x}}{dp(x)}=z_x$	$\frac{y-\bar{y}}{dp(y)}=z_y$	$z_x \cdot z_y$
A	2	48	−3,7	−8,5	−1,54	−1,05	1,617
B	3	50	−2,7	−6,5	−1,12	−0,80	0,846
C	4	56	−1,7	−0,5	−0,71	−0,06	0,043
D	5	52	−0,7	−4,5	−0,29	−0,55	0,160
E	4	43	−1,7	−13,5	−0,71	−1,66	1,179
F	6	60	0,3	3,5	0,12	0,43	0,052
G	7	62	1,3	5,5	0,54	0,68	0,367
H	8	58	2,3	1,5	0,95	0,19	0,181
I	8	64	2,3	7,5	0,95	0,92	0,874
J	10	72	4,3	15,5	1,78	1,91	3,400
Total	57	565	0	0			8,769

$$\bar{x} = 5{,}7, \qquad dp(X) = 2{,}41, \qquad \bar{y} = 56{,}5, \qquad dp(Y) = 8{,}11$$

Portanto, para esse exemplo, o grau de associação linear está quantificado por 87,7%.

Figura 4.7 Mudança de escalas para o cálculo do coeficiente de correlação.



Da discussão feita até aqui, podemos definir o coeficiente de correlação do seguinte modo.

Definição Dados n pares de valores $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, chamaremos de coeficiente de correlação entre as duas variáveis X e Y a

$$\text{corr}(X, Y) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{dp(X)}\right)\left(\frac{y_i - \bar{y}}{dp(Y)}\right), \tag{4.7}$$

ou seja, a média dos produtos dos valores padronizados das variáveis.

Não é difícil provar que o coeficiente de correlação satisfaz

$$-1 \leq \text{corr}\,(X, Y) \leq 1. \tag{4.8}$$

A definição acima pode ser operacionalizada de modo mais conveniente pelas seguintes fórmulas:

$$\text{corr}(X,Y) = \frac{1}{n}\sum\left(\frac{x_i - \overline{x}}{dp(X)}\right)\left(\frac{y_i - \overline{y}}{dp(Y)}\right) = \frac{\sum x_i y_i - n\overline{x}\,\overline{y}}{\sqrt{\left(\sum x_i^2 - n\overline{x}^2\right)\left(\sum y_i^2 - n\overline{y}^2\right)}}. \tag{4.9}$$

O numerador da expressão acima, que mede o total da concentração dos pontos pelos quatro quadrantes, dá origem a uma medida bastante usada e que definimos a seguir.

Definição Dados n pares de valores (x_1, y_1), ..., (x_n, y_n), chamaremos de *covariância* entre as duas variáveis X e Y a

$$\text{cov}(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n}, \tag{4.10}$$

ou seja, a média dos produtos dos valores centrados das variáveis.

Com essa definição, o coeficiente de correlação pode ser escrito como

$$\text{corr}(X,Y) = \frac{\text{cov}(X,Y)}{dp(X)\cdot dp(Y)}. \tag{4.11}$$