Introdução à estatística

Aula 9 - Correlação entre as variáveis

Felipe Nunes, Ph.D.

November 26, 2018

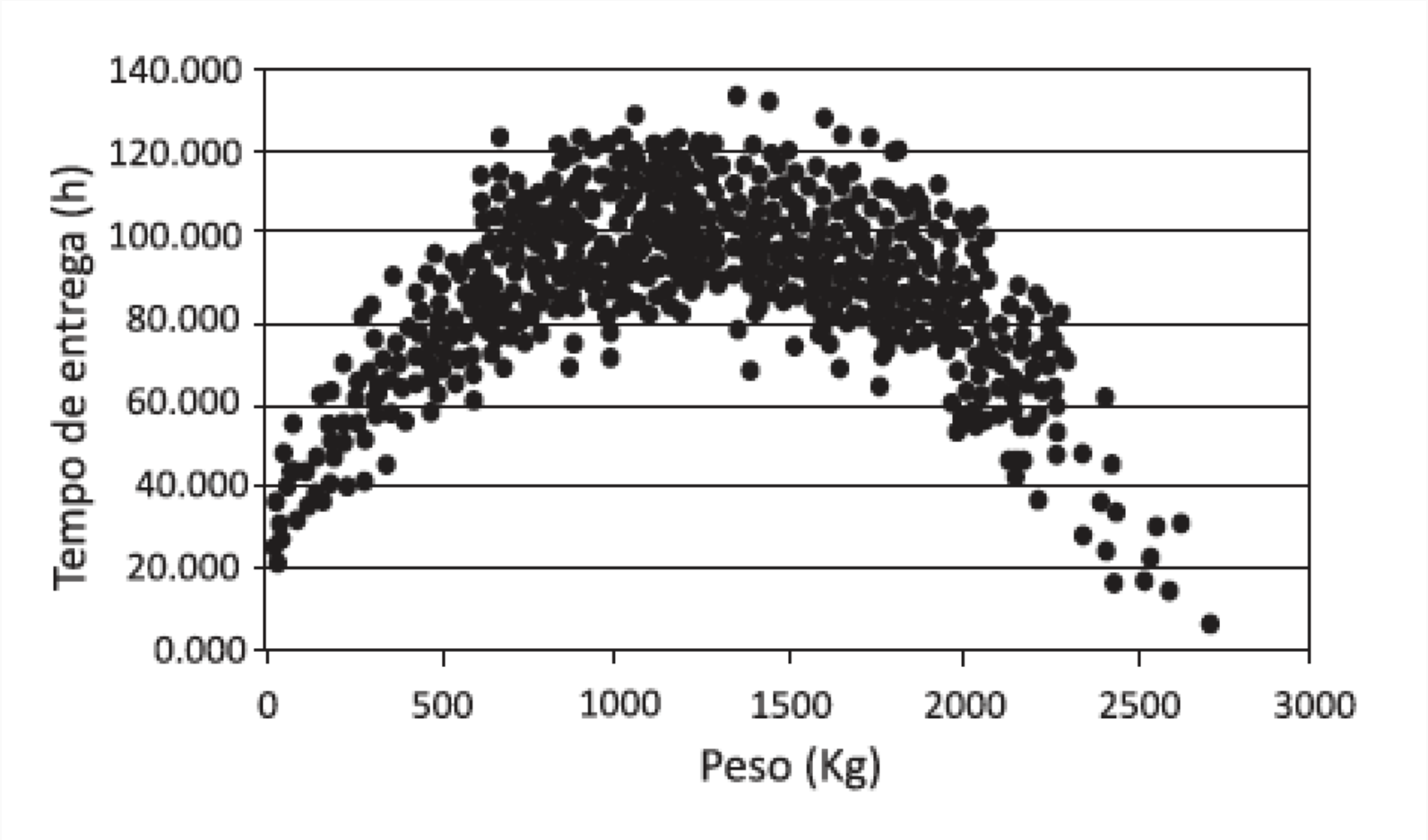
Gestão Pública

Correlação

Correlação

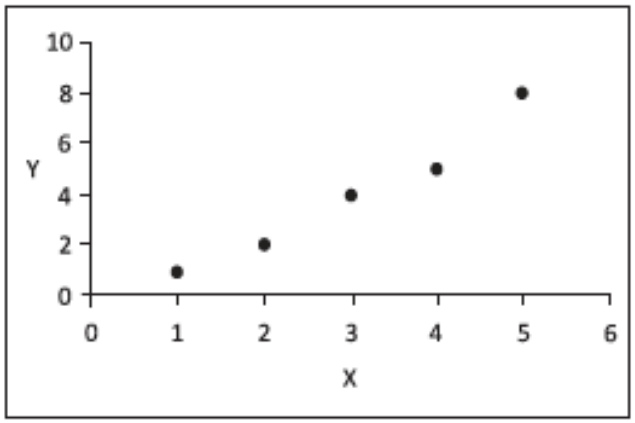
- Para verificar o grau de relacionamento entre duas variáveis, ou seja, o grau de associação entre elas, devemos estudar um coeficiente chamado de coeficiente de correlação.
- Existem vários coeficientes de correlação e, cada um deles, aplicado em casos específicos. Aqui, iremos estudar o coeficiente de correlação de Pearson (r).
- Para que possamos ter uma ideia da associação entre as variáveis que estamos estudando, iremos utilizar um gráfico de dispersão como o apresentado, a seguir, pelo qual podemos constatar a relação entre as variáveis: o peso de um pacote e o seu tempo de entrega.

Correlação

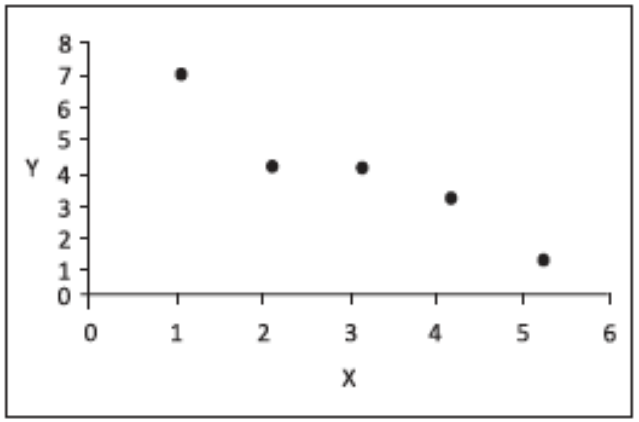

140.000
120.000
100.000
80.000
60.000
40.000
20.000
0.000
Tempo de entrega (h)
0
500
1000
1500
2000
2500
3000
Peso (Kg)

Correlação

- As estimativas de correlação podem ser positivas (à medida que a variável x aumenta a variável y também aumenta) ou negativas (à medida que a variável x aumenta a variável y diminui), como você pode ver a partir dos gráficos a seguir:



Correlação Positiva



Correlação Negativa

Correlação

- O coeficiente de correlação de Pearson (r) nos dá uma ideia da variação conjunta das variáveis analisadas e pode assumir valores de –1 a +1.
- O coeficiente de correlação de Pearson é obtido assim:

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

Correlação

- A ocorrência de um valor de $r = 0$ ou próximo de zero indica apenas que não há correlação linear entre as variáveis, porque pode existir uma forte relação não linear entre as variáveis, como no gráfico de dispersão do peso do pacote e o tempo de entrega, na qual temos uma relação não linear.

Correlação

- Características do coeficiente de correlação de Pearson:

  1. seus valores estão compreendidos entre -1 e 1;
  2. se o coeficiente for positivo, as duas características estudadas tendem a variar no mesmo sentido.
  3. se o sinal for negativo, as duas características estudadas tendem a variar em sentido contrário;
  4. a relação entre duas variáveis é tanto mais estreita quanto mais o coeficiente se aproxima de 1 ou –1;

Correlação

- O valor de r é uma estimativa do parâmetro ρ (rho), da mesma forma que a média $\bar{x}$ é uma estimativa de μ.
- Para testar se o valor de r é estatisticamente igual ao parâmetro de uma população em que $\rho = 0$, podemos empregar o teste t definido por:

$$t_{calculado} = \frac{r - \rho}{\sqrt{1 - r^2}} \times \sqrt{n - 2}$$

onde n é o número total de pares; r^2 é o coeficiente de correlação ao quadrado; e ρ é o parâmetro da correlação populacional (considerado igual a zero).

Correlação

- **Exemplo**: Vamos determinar o coeficiente de correlação entre a porcentagem de aplicação do total de recursos com Educação em uma prefeitura (x) e o grau de conhecimento médio da população da cidade (y). Para isso, foram avaliadas dez cidades.

Correlação

% aplicado na educação	Grau de conhecimento
5	70
10	40
20	27
30	22
40	18
50	16
60	15
70	14
80	13
90	12

Correlação

- Para obtermos a estimativa de correlação, precisamos calcular todos os somatórios presentes na expressão de r:

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

Correlação

- Somatório de todos os valores de x:

$$\sum x_i = x_1 + x_2 + ... + x_{10} = 5 + 10 + ... + 90 = 455$$

- Somatório de todos os valores de x elevados ao quadrado:

$$\sum x_i^2 = x_1^2 + x_2^2 + ... + x_{10}^2 = 5^2 + 10^2 + ... + 90^2 = 28.525$$

- Somatório de todos os valores de y:

$$\sum y_i = y_1 + y_2 + ... + y_{10} = 70 + 40 + ... + 12 = 247$$

- Somatório de todos os valores de y elevados ao quadrado:

$$\sum y_i^2 = y_1^2 + y_2^2 + ... + y_{10}^2 = 70^2 + 40^2 + ... + 12^2 = 9.027$$

Correlação

- Somatório de todos os valores obtidos por meio do produto dos valores de x e y de cada cidade:

$$\begin{aligned} \sum x_i y_i &= x_1 y_1 + x_2 y_2 + ... + x_{10} y_{10} \\ &= 5 \times 70 + 10 \times 40 + ... + 90 \times 12 = 7.470 \end{aligned}$$

Correlação

- Substituindo esses valores na expressão, teremos:

$$
\begin{aligned}
r &= \frac{\sum x_i y_i - \frac{\sum x_i - \sum y_i}{n}}{\sqrt{\sum x_i^2 - \frac{(\sum x_i)^2}{n} \times \sum y_i^2 - \frac{(\sum y_i)^2}{n}}} \\
&= \frac{7470 - \frac{455-247}{10}}{\sqrt{28.525 - \frac{(455)^2}{10} \times 9.027 - \frac{(247)^2}{10}}} \\
&= \frac{-3.768.5}{4.784.3} = -0.788
\end{aligned}
$$

Correlação

- O valor de r = -0,7877 indica que existe uma associação inversa (negativa) e de média magnitude entre a variação da porcentagem de aplicação do total de recursos com educação em uma prefeitura e o grau de conhecimento médio da população da cidade, ou seja, nesta população de cidades, provavelmente os recursos da educação não estão sendo bem empregados, já que a relação foi negativa quando se esperava uma relação positiva.

Correlação

- Para verificarmos se esse resultado é significativo, vamos fazer o seguinte teste de hipótese:

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

- Iremos calcular a estatística por meio da expressão:

$$t_{calculado} = \frac{-0.788 - 0}{\sqrt{1 - 0.788^2}} \times \sqrt{10 - 2} = -1.25 \times 2.82 = -3.525$$

Correlação

- Olhando na tabela de t para 8 graus de liberdade (10-2) e um $\alpha = 0,025$, já que estamos considerando uma significância de 0,05 e o nosso teste é bilateral, teremos um valor tabelado de 2,306.
- Verificamos que o valor calculado de 3,525 está na região de rejeição da hipótese H_0 e, portanto, iremos aceitar a hipótese H_1, ou seja, de que $\rho \neq 0$.
- Então, o resultado encontrado na amostra (r) não foi fruto do acaso, considerando uma significância de 5%.

Correlação

- Devemos ter cuidado na interpretação do coeficiente de correlação, pois este não implica necessariamente uma medida de causa e efeito.
- É mais seguro interpretar o coeficiente de correlação como uma medida de associação.
- Por exemplo, podemos encontrar uma correlação alta entre o aumento dos salários dos professores e o consumo de bebidas alcoólicas através de uma série de anos em uma região.
- Esse valor de r encontrado foi alto apenas porque pode ser que ambas as variáveis tenham sido afetadas por uma causa comum, ou seja, a elevação do padrão de vida de uma região.