Correlation Coefficient

Can you measure how strong or how weak the relationship is between two variables? The *correlation coefficient* measures the degree of linear relationship between two variables.

[redacted]

The population correlation coefficient is denoted by ρ. Its value ranges from –1 (perfect negative correlation) to +1 (perfect positive correlation). The closer the correlation coefficient is to either –1 or +1, the stronger the linear relationship is. The closer the correlation coefficient to 0, the weaker the linear relationship is.

[redacted]

A correlation coefficient of 0 means there is no linear relationship between the two variables, but it does not indicate that there is no association. There may exist a relationship between the two variables that is not linear when the correlation coefficient is 0. Figure 6.4 shows the scatter plots with the corresponding descriptions of the linear associations between the two variables.

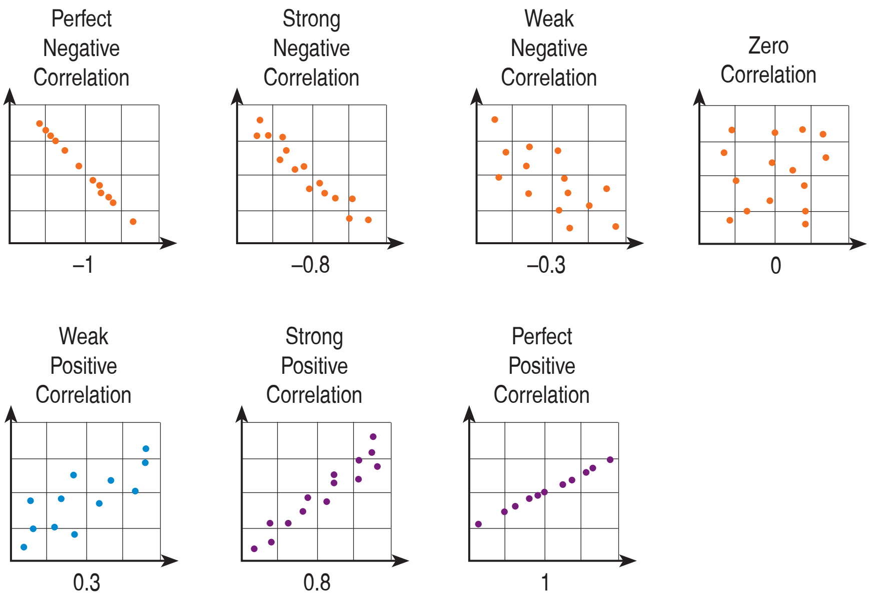

Figure 6.4 Scatter plots with correlation descriptions

[redacted]

The Pearson product-moment correlation coefficient or sample correlation coefficient r can be computed by

$$r = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}}$$

where

$$s_{xy} = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i Y_i - \frac{\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n}\right],$$

$$s_x^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^{\,2} - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}\right], \text{ and}$$

$$s_y^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}\right]$$

Substituting these formulas to the formula for r, and eventually cancelling the factor $\frac{1}{n-1}$, you can derive the following formula.

$$r = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sqrt{\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]\left[\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right]}}$$

Simplifying further, you get

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left[n\sum_{i=1}^{n} x_i^2 - \left(n\sum_{i=1}^{n} x_i\right)^2\right]\left[n\sum_{i=1}^{n} y_i^2 - \left(n\sum_{i=1}^{n} y_i\right)^2\right]}}.$$

where x_i is the value of the independent variable x for the ith observation,
y_i is the value of the dependent variable y for the ith observation, and
n is the sample size on number of paired observations.

[redacted]

[redacted]

Example 6.4

In example 6.1, what is the degree of linear relationship between height and arm span? Interpret.

Height	*Arm Span*
63	61
59	62
62	63
67	64
62	61
71	72
67	66
59	57
72	72
68	66

Solution.

Find the values of n, Σx, Σy, Σxy, Σx^2, and Σy^2, and then substitute these in the sample correlation coefficient formula.

	x	y	x^2	y^2	xy
1	63	61	3969	3721	3843
2	59	62	3481	3844	3658
3	62	63	3844	3969	3906
4	67	64	4489	4096	4288
5	62	61	3844	3721	3782
6	71	72	5041	5184	5112
7	67	66	4489	4356	4422
8	59	57	3481	3249	3363
9	72	72	5184	5184	5184
10	68	66	4624	4356	4488
Sum	650	644	42 446	41 680	42 046

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left[n\sum_{i=1}^{2} x_i^2 - \left(\sum_{i=1}^{2} x_i\right)^2\right]\left[n\sum_{i=1}^{2} y_i^2 - \left(\sum_{i=1}^{2} y_i\right)^2\right]}}$$

$$= \frac{10(42\,046) - (650)(644)}{\sqrt{\left[10(42\,446) - (650)^2\right]\left[10(41\,680) - (644)^2\right]}}$$

$$= 0.925.$$

The degree of linear relationship between height and arm span is 0.925. Thus, there is a very strong positive direct linear relationship between height and arm span.

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left[n\sum_{i=1}^{2} x_i^2 - \left(\sum_{i=1}^{2} x_i\right)^2\right]\left[n\sum_{i=1}^{2} y_i^2 - \left(\sum_{i=1}^{2} y_i\right)^2\right]}}$$

$$= \frac{9(209\,350)-(441)(4600)}{\sqrt{\left[9(25\,395)-(441)^2\right]\left[9(2\,420\,000)-(4600)^2\right]}}$$

$$= -0.994.$$

There is a very strong negative correlation between mileage and the price of the used vans at –0.994. This implies that as mileage increases, the price of the used vans decreases.

[redacted]