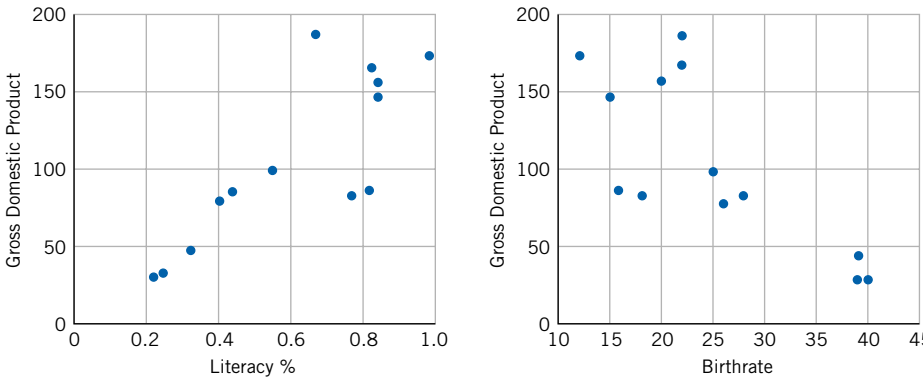
The Correlation Coefficient

Karen Callaghan, Ph.D
The University of Massachusetts-Boston

A correlation coefficient measures the strength and direction of a linear association between two variables. It ranges from −1 to +1. The closer the absolute value is to 1, the stronger the relationship. A correlation of zero indicates that there is no linear relationship between the variables. The coefficient can be either negative or positive. The scatterplots below indicate two linear associations of the same strength but opposite directions.



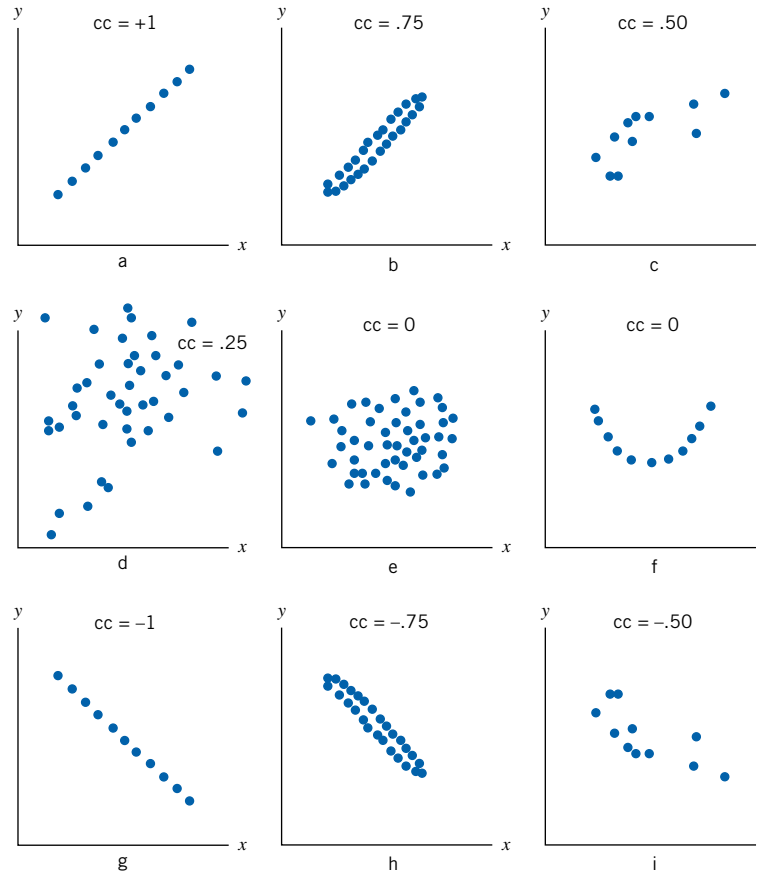
The graph on the left represents the relationship between literacy % and gross domestic product. As literacy % goes up, the gross domestic product goes up, so the correlation coefficient is positive. The graph on the right represents the relationship between birthrate and gross domestic product. As birthrates go up, the gross domestic product goes down, so the coefficient is negative.

The scatterplots on the next page will give you an intuitive grasp of the correlation coefficient, labelled as cc. Figure (a) represents a perfect correlation of +1, all the points fall on a perfectly straight line with a positive slope, an unlikely occurrence in any social science data set. Figure (b) represents a strong correlation where the behavior of one variable is similar, but not identical to the behavior of the other variable (e.g., like miles traveled versus amount of gasoline used).

In Figure (c), a correlation of .50 represents a moderately strong positive relationship. The relationship in Figure (d) is weak, so the coefficient is only .25.

The scatterplot in Figure (e) looks like a shotgun blast so the correlation is zero; the x and y variables are not linearly related.

You might be surprised to learn that the coefficient in Figure (f) is also zero. This is because the correlation coefficient measures a linear association, while the relationship in Figure (f) is curvilinear.



Figures (g), (h), (i) are mirror images of Figures (a), (b), (c). All the correlation coefficients are negative. Figure (g) represents a perfect correlation of -1, all the points fall on a straight line with a negative slope, an unlikely occurence in any social science data set. Figure (h) represents a strong negative correlation. In Figure (i) the correlation is moderately strong.

COMPUTING THE CORRELATION COEFFICIENT (CC.)

While a perfect correlation is easy to decipher, it is difficult to guess the coefficient of weaker correlations. That is why Karl Pearson developed a precise mathematical measure of correlation

known as Pearson's *r*, which is called cc. throughout the text. For those who are interested in knowing how the correlation coefficients are actually calculated, the steps are outlined below.

$$\text{Correlation Coefficient} = \frac{\Sigma(x - \bar{x})\cdot(y - \bar{y})}{\sqrt{\Sigma(x - \bar{x})^2\cdot(y - \bar{y})^2}}$$

The symbol Σ is "sigma," the Greek letter S, which is a mathematical shorthand meaning "sum up." So if x takes on values 2, 3, and 6, then: $\Sigma x = 2 + 3 + 6 = 11$.

Literacy % and Life Expectancy for Six Selected Nations

	(x) Literacy %	(y) Life Expectancy
STEP 1: Start with a set of data, x and y points. Each data point is kept in a separate row.	.29	42
	.92	77
	.52	58
	.55	47
	.40	48
	.66	55
	3.34	327

STEP 2: Find $\bar{x}$, the mean of x; and $\bar{y}$, the mean of y. To do this add the values of x and divide by the number of points; then do the same for y.

$\bar{x} = \Sigma x/n = 3.34/6 = .56$
$\bar{y} = \Sigma y/n = 327/6 = 54.50$

STEP 3: Subtract $\bar{x}$ from each value of x; subtract $\bar{y}$ from each value of y to get a new table of rows.

$x - \bar{x}$	$y - \bar{y}$
−.27	−12.50
.36	22.50
−.04	3.50
−.01	−7.50
−.16	−6.50
.50	.10

STEP 4: Take the products of each row in step 3 and sum them up.

	Products
−.27 × −12.50 =	3.38
.36 × 22.50 =	8.10
−.04 × 3.50 =	−0.14
−.01 × −7.50 =	0.07
−.16 × −6.50 =	1.04
.10 × .50 =	0.05
Sum of products =	12.50

STEP 5:
Take each x value in step 3, square it and sum all the points; then do the same for y

	x squared	y squared
	$-.27^2 = .07$	$-12.50^2 = 156.25$
	$.36^2 = .13$	$22.50^2 = 506.25$
	$-.04^2 = .00$	$3.50^2 = 12.25$
	$-.01^2 = .00$	$-7.50^2 = 56.25$
	$-.16^2 = .03$	$-6.50^2 = 42.25$
	$.10^2 = .01$	$.50^2 = .25$
Sums of Squares	$= .24$	$= 773.50$

STEP 6:
Take the square root of the product of the sums of the squares in step 5.

$$\sqrt{.24 \cdot 773.50} = 13.62$$

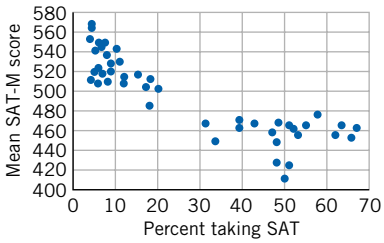
STEP 7:
Divide the sum in step 4 by the value in step 6 to calculate the correlation coefficient.

$$\text{cc.} = \frac{12.50}{13.62} = .92$$

So the correlation coefficient is .92 which says that literacy percent and life expectancy are strongly related.

Suppose we reverse the x and y axis. Now "x," the independent variable, is life expectancy and "y," the dependent variable, is literacy percent. You do not need to retrace steps 1 through 6 again. The correlation is still .92 since the formula does not make a distinction between the x and y variables. You cannot test whether literacy percent affects life expectancy, or life expectancy affects literacy percent. You can only test whether these two variables vary together in a linear relationship.

An example of a negative correlation.



News reports sometimes rate the quality of educational systems by comparing the mean of scores of seniors in each state on college entrance exams. (See discussion p.121 and reading "Verbal SAT Scores.") This method is misleading because the percent of high school seniors who take any particular college entrance test varies greatly from state to state. The figure to the side shows a scatterplot of the mean score on the Scholastic Aptitude Test (SAT) mathematics examination for high school seniors in each state compared with the percent of graduates in each state who took the test.

The negative association between the two variables is evident: SAT scores tend to be lower in states where the percent of students who take the test is higher. Only 3% of the seniors in the three highest scoring states took the SAT.