Encyclopedia of Social Science Research Methods

Correlation

Contributors: Jacques Tacq
Editors: Michael S. Lewis-Beck & Alan Bryman & Tim Futing Liao
Book Title: Encyclopedia of Social Science Research Methods
Chapter Title: "Correlation"
Pub. Date: 2004
Access Date: April 06, 2015
Publishing Company: SAGE Publications, Inc.
City: Thousand Oaks
Print ISBN: 9780761923633
Online ISBN: 9781412950589
DOI: http://dx.doi.org/10.4135/9781412950589.n178
Print pages: 200-204

http://dx.doi.org/10.4135/9781412950589.n178
Generally speaking, the notion of correlation is very close to that of association. It is the statistical association between two variables of interval or ratio measurement level.

To begin with, correlation should not be confused with causal effect. Indeed, statistical research into causal effect for only two variables happens to be impossible, at least in observational research. Even in extreme cases of so-called unicausal effects, such as contamination with the *Treponema* pallidum bacteria and contracting syphilis, there are always other variables that come into play, contributing, modifying, or counteracting the effect. The presence of a causal effect can only be sorted out in a multivariate context and is consequently more complex than correlation analysis.

Let us limit ourselves to two interval variables, denoted as X and Y, and let us leave causality aside. We assume for the moment a linear model. It is important to note that there is a difference between the correlation coefficient, often indicated as r, and the unstandardized regression coefficient b (computer output: B). The latter merely indicates the slope of the regression line and is computed as the tangent of the angle formed by the regression line and the x-axis. With income (X) and consumption (Y) as variables, we now have the consumption quote, which is the additional amount one spends after obtaining one unit of extra income, so the change in Y per additional unit in X is $B = {}_{\delta}Y/{}_{\delta}X$. We will see below that there are, in fact, two such regression coefficients and that the correlation coefficient is the geometrical mean of them.

Five Main Features of a Correlation

Starting from probabilistic correlations, each correlation has five main features:

- 1. nature,
- 2. direction,
- 3. sign,
- 4. strength,
- 5. statistical generalization capacity.

Nature of the Correlation

The nature of the correlation is linear for the simple correlation computation suggested above. This means that through the scatterplot, a linear function of the form $E(Y) = b_0 + b_{y1} X_1$ is estimated. Behind the correlation coefficient of, for example, $r = 0.40$ is a linear model. Many researchers are fixated on the number between 0 and 1 or between 0 and -1, and they tend to forget this background model. Often, they unconsciously use the linear model as a tacit obviousness. They do not seem to realize that nonlinearity occurs frequently.

An example of nonlinearity is the correlation between the percentage of Catholics and the percentage of CDU/CSU voters in former West Germany (Christlich-Demokratische Union/Christlich-Soziale **[p. 200 ↓]** Union). One might expect a linear correlation: the more Catholics, the more voters of CDU/CSU according to a fixed pattern. However, this “the more, the more” pattern only seems to be valid for communes with many Catholics. For communes with few Catholics, the correlation turns out to be fairly negative: The more Catholics, the fewer voters for CDU/CSU. Consequently, the overall scatterplot displays a U pattern. At first, it drops, and from a certain percentage of Catholics onwards, it starts to rise. The quadratic function that describes a parabola therefore shows a better fit with the scatterplot than the linear function.

Many other nonlinear functions describe reality, whether exponential, logistic, discontinuous, or other. The exponential function, for example, was used in the reports by the Club of Rome, in which increased environmental pollution (correlated with time) was described in the 1970s. The logistic function was used by those who reacted

against the Club of Rome, with the objection that the unchecked pollution would reach a saturation level.

The Direction of the Correlation

When dealing with two variables, X_1 and Y, one of them is the INDEPENDENT VARIABLE and the other is the DEPENDENT VARIABLE. In the case of income (X_1) and consumption (Y) the direction $X_1 \rightarrow Y$ is obvious. Here, a distinction between the correlation coefficient and the regression coefficient can already be elucidated. The regression coefficient indicates b_{y1} for the direction $X_1 \rightarrow Y$ and b_{1y} for the direction $Y \rightarrow X_1$. On the other hand, the correlation coefficient, which can be calculated as the geometrical mean of b_{y1} and b_1 (y (= the square root of the product of these coefficients) incorporates both directions and therefore is a nondirected symmetrical coefficient.

A novice might conclude that causal analysis is the privilege of regression coefficients and that correlation coefficients are unsuitable for that purpose. This would be an understandable but premature conclusion. Indeed, the direction in which we calculate

mathematically is not the same as the causality direction. Each function $Y = aX + b$, with Y as a dependent variable, is convertible into its inverted function $X = Y/a - b/a$, with X as a dependent variable. The direction here is of relative value. However, when a stone is thrown into a pond, the expanding ripples are explained by the falling stone, but the opposite process with shrinking ripples would imply an enormous number of ripple generators, the coherence of which would need to be explained. It would also be imperative to demonstrate that they come from one central point (the improbability of an implosion). The direction here is not of relative value.

As was said before, causality cannot be equated to the statistical computation of a regression coefficient. True, it is fortunate that this coefficient represents a direction, and Hubert Blalock (1972) gratefully used this in his CAUSAL MODEL approach. However, this direction in terms of causality is hopelessly relative in statistical computation. This is admitted with perfect honesty by Sewell Wright (1934) in his article “The Method of Path Coefficients” and in earlier articles, when he claimed to have found a method to link the (a priori) causal knowledge with the (a posteriori) empiric correlations in the correlation matrix.

The Sign of the Correlation

A correlation can be positive (the more of X_1, the more of Y) or negative (the more of X_1, the less of Y). In the case of income and consumption, there is a positive correlation (the sign is +): The higher the income, the higher the consumption.

An example of negative correlation is the relationship between social class and number of children at the beginning of the past century. Higher social classes displayed the mechanism that was known as *social capillarity.* This was strikingly described by Friedrich Engels in *Der Ursprung der Familie, des Privateigentums und des Staates.* As soon as a son was born in the higher social classes, the parents started applying birth

control because the estate could be passed on from father to son. In the lower social classes, however, this mechanism was lacking. A family of 10 children was the rule, and sometimes there were as many as 20 children. The sign here is abundantly clear—the higher the social class, the lower the number of children.

The Strength of the Correlation

A correlation can be weak or strong. A correlation coefficient of 0.90 or -0.90 is strong. A coefficient of 0.10 or -0.10 is weak. A coefficient 0 signals lack of correlation. A coefficient 1 or -1 indicates maximal deterministic correlation.

It is convenient to keep the vector model in mind. The vector model is the coordinate system in which the roles of variables and units have been changed: **[p. 201 ↓]** Units are now the axes, and variables are the points forming the scatterplot. These variables are conceived as the end points of vectors starting in the origin—hence the name. In such a vector model, the correlation coefficient between two variables is equal to the cosine of the angle formed by the two vectors. We know that the cosine of 90° equals 0: The vectors then are at perpendicular angles to each other, and the correlation is 0. On the other hand, the cosine of 0° equals 1: The vectors then coincide and are optimally correlated.

Mostly, a strong correlation coefficient (between two variables, not in a multivariate context) will coincide with a strong regression coefficient, but this is not always the case. Many situations are possible, and this is related to the dispersions of the variables, which will be dealt with below.

In addition, a strong correlation will tend to be more significant than a weak correlation, but it does happen that a strong correlation is not significant. Moreover, a significant correlation can be weak. It follows that the strength of a correlation (correlation or regression coefficient) cannot be judged as such because it should be studied in combination with other criteria, the first one being the significance, which is discussed next.

The Statistical Potential for Generalization

A correlation calculated on the basis of a random sample may or may not be significant (i.e., it can or cannot be statistically generalized with respect to the population from which the sample was taken). Here we have to go against the postmodern range of ideas and take a traditional rigorous point of view: Only those correlations that can be statistically generalized (and are therefore statistically significant) are appropriate for interpretation. Nonsignificant correlations are the result of mere randomness and should therefore be considered nonexistent for the population.

Consequently, a strong coefficient (e.g., $r = 0.90$), which is not significant because the sample size is much too small (e.g., $n = 5$), does not stand the test and will be rejected by any social science researcher. A discerning mind will object that a weak correlation (e.g., $r = 0.10$) in a sufficiently large sample (e.g., sample size $n = 5{,}000$) is still significant and that this is only due to the size of the sample. Take a sufficiently large sample and you can prove anything with statistics, so it would appear. However, this argument is only partly correct. When dealing with large numbers, weak correlations or small changes such as changes in fertility can be extremely significant (in the sense of important) for policymaking. Any demographer knows how irreversible such processes are. Besides, for large numbers, it is also still possible that the correlation is not statistically significant. Therefore, it does not hold to pretend that the result is only due to the size of the sample.

Admittedly, there is still a grey area between two extremes: a significant result for a small sample and a nonsignificant one for a large sample. In the latter case, we know with certainty that the correlation is not significant. In the former case, we are certain that a larger sample survey will also yield a significant result. However, in general and certainly in the grey area, nonstatistical criteria, such as social relevance, will always contribute to the interpretation of correlation results.

So far, we have looked at some properties of correlation: nature, direction, sign, strength, and significance. As was already mentioned, one should not be blinded here by the number representing the correlation coefficient; indeed, all these features need to be taken into account. In this sense, major META-ANALYSES in which correlation

coefficients from different studies are all mixed together should not be trusted. Indeed, in each separate study, one should raise many questions. Are we dealing with a linear correlation? If so, is the regression coefficient different from the correlation coefficient, and how large is the intercept? Are the variables standardized? Is the correlation significant? How large is the confidence interval? What is the probability of TYPE I ERROR α? What is the size of the sample? What is the probability of TYPE II ERROR β and the power?

The Dispersion of the Variables: Standardized Coefficients?

The question about the difference between correlation and regression effect actually also refers to the dispersion of the variables. Indeed, we have already mentioned that a strong correlation coefficient does not always coincide with a strong regression coefficient and that this is linked to the dispersion of the variables. This can be derived from the formulas of the coefficients: r_{y1}

(or r_{1y}

) is the covariance divided by the geometrical mean of the variances, b_{y1}

is the covariance divided by the variance of X_{1}

, and $b1y$ is the covariance divided by the variance of Y. In the **[p. 202 ↓]** case of equal dispersions, all three coefficients will be equal. A greater dispersion of Y will have a decreasing effect on the correlation coefficient and on b_{1y}

, but not on b

y1

.

With respect to this, one may wonder if the standardized coefficients (betas) should be preferred. (Standardization is a procedure of subtracting the mean and dividing by the STANDARD DEVIATION.) Dividing by the standard deviation happens to be a rather drastic operation, which causes the variance of the variables to be artificially set equal to 1, with a view to obtaining comparability. But some have argued that a similar difference in education has a different meaning in a society with a great deal of educational inequality compared to a society with a small degree of such inequality, and therefore they advocate nonstandardized coefficients.

All this is not just abstract thinking; it does occur in specific processes in society. Examples may include the processes of efficient birth control, underemployment, the raising of political public opinion, and environmental awareness. From all these examples, we learn that society introduces changes in the dispersion of its features.

Calculation of the Correlation Coefficient

We would like to illustrate this importance of dispersions with a small example in which the correlation coefficient between education (X) and income (Y) is calculated. The formula is $r_{xy} = [\Sigma(X - \bar{X})(Y - \bar{Y})]/[\Sigma(X - \bar{X})^2 \Sigma(Y - \bar{Y})^2]^{1/2}$.

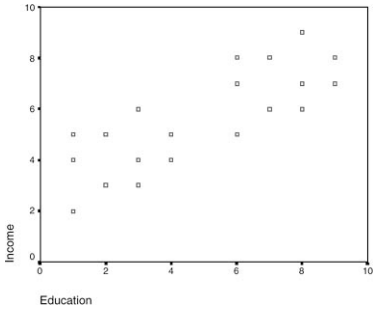
Let us take a nation in which few of its citizens go to school, and those who do happen to be educated have high incomes and the rest are poor. Let the data matrix be as below. Education is measured on a scale from 1 to 9, and income is measured in thousands of Euros.

X	1 1 1 2 2 3 3 3 4 4 6 6 6 7 7 8 8 8 9 9
Y	2 4 5 3 5 3 4 6 4 5 5 7 8 6 8 6 7 9 7 8

The scatterplot is shown in Figure 1.

The standard deviations are 2.81 for education and 1.93 for income. The correlation coefficient is 0.798. The model is linear, the coefficient is symmetric, the relationship is positive, the strength is substantial, and the coefficient is statistically significant ($p<.001$).

Figure 1



Now suppose that in this nation, the more ambitious among the poor demand that their children be admitted to the schools and that the state yields to this pressure and begins to subsidize education. Large numbers now go to school and college, average income rises as education raises the general level of human capability, and the variation of incomes declines. But once subsidized education becomes widespread, the individuals who are more schooled can no longer capture for themselves much of the benefit of that schooling, and the correlation between income and education diminishes.

Now the data matrix is the following.

X	6 6 6 6 6 6 7 7 7 8 8 8 8 8 9 9 9 9
Y	5 7 8 5 7 8 6 8 6 8 6 7 9 6 7 9 7 8 7 8

The scatterplot is shown in Figure 2.

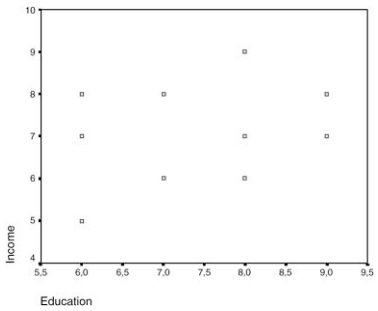
The standard deviations are much smaller: 1.14 for education and 1.17 for income. The correlation coefficient has diminished from 0.798 to 0.285. The relationship is positive but small and no longer statistically significant.

One could also suppose that in this nation, the variation of education remains the same because only some of the citizens go to school, but the variation of incomes has declined because there is a lot of underemployment (people with high diplomas having low-salary jobs). This would result in the following data matrix:

X	1	1	1	2	2	3	3	3	4	4	6	6	6	7	7	8	8	8	9	9
Y	2	4	5	3	5	3	4	6	4	5	2	4	5	3	5	3	4	6	4	5

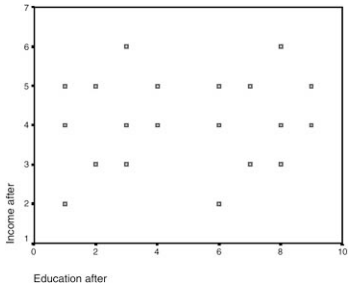
[p. 203 ↓]

Figure 2



The scatterplot is now as follows, in Figure 3:

Figure 3



The standard deviation of education remains the same (2.81), but the standard deviation of income has become very small (1.17). The correlation coefficient is now 0.116. The relationship is positive but small and nonsignificant.

JacquesTacq

http://dx.doi.org/10.4135/9781412950589.n178

References

Blalock, H. M. (1972). Social statistics . Tokyo: McGraw-Hill Kogakusha.

Brownlee, K. (1965). Statistical theory and methodology in science and engineering . New York: John Wiley.

Hays, W. L. (1972). Statistics for the social sciences . New York: Holt.

Kendall, M., & Stuart, A. (1969). The advanced theory of statistics . London: Griffin.

Tacq, J. J. A. (1997). Multivariate analysis techniques in social science research: From problem to analysis . London: Sage.