The Journal of Educational Research

ISSN: 0022-0671 (Print) 1940-0675 (Online) Journal homepage: <http://www.tandfonline.com/loi/vjer20>

The Correlation Ratio

Leonard P. Ayres

To cite this article: Leonard P. Ayres (1920) The Correlation Ratio, The Journal of Educational Research, 2:1, 452-456, DOI: [10.1080/00220671.1920.10879073](https://doi.org/10.1080/00220671.1920.10879073)

To link to this article: <http://dx.doi.org/10.1080/00220671.1920.10879073>

Published online: 15 Dec 2014.

Submit your article to this journal [↗](#)

Article views: 1

View related articles [↗](#)

Citing articles: 1 View citing articles [↗](#)

THE CORRELATION RATIO

LEONARD P. AYRES
Russell Sage Foundation

In the March, April, and May numbers of the JOURNAL OF EDUCATIONAL RESEARCH there have appeared articles by the writer describing methods for computing the coefficients of correlation and regression that are shorter and simpler than those that have heretofore been in use.

The purpose of this article is to describe a short method for computing the correlation ratio which is a measure proposed by Pearson for computing correlation in distributions in which the relationship between the two series is a curve relation rather than rectilinear. Such relationships are frequent.

For example there may be a real relationship between the amount of money spent on municipal sanitation and the decline in the death rate. In extreme cases, such as have recently existed in tropical America, conditions have been found such that slight expenditures for sanitation resulted in marked reductions in sickness. As improvements go on, however, it is found that the expenditures must be greatly increased to effect even small decreases in the death rate. Similar conditions are common in the field of education as, for example, in the relation between the amount of time devoted to the practice of penmanship and the quality of the results obtained.

The coefficient of correlation will not accurately measure these curvilinear relations and the correlation ratio is the device that is proposed to obtain a trustworthy measurement of them.

The correlation ratio can only be computed for tables of distribution. It is the ratio between two standard deviations. One of these is the standard deviation of all the values of one series—say the subject series. The other standard deviation is that of the averages of the columns of the subject values from the average of all the subject values in the table. If the relationship of the two series is rectilinear the correlation ratio is equal to the coefficient of correlation. If it is curvilinear the ratio will be greater than the coefficient of correlation. It is always positive and never exceeds 1. Its symbol is η (eta). The best description of

its derivation and use is found in Rugg's *Statistical Methods Applied to Education*.

The computation of this measure may be illustrated by finding both the coefficient of correlation and the correlation ratio for a simple distribution table in which the relationship is roughly curvilinear. The coefficient of correlation is found by the formula:

$$r = \frac{\text{Sum of products of subject and relative items} - \text{av. of subject items} \times \text{tot. of relative items}}{\sqrt{\left(\frac{\text{Sum of squares of sub. items} - \text{av.} \times \text{total of these items}}{\text{av.} \times \text{total of these items}} \right) \times \left(\frac{\text{Sum of squares of rel. items} - \text{av.} \times \text{total of these items}}{\text{av.} \times \text{total of these items}} \right)}}$$

The computations are shown in the accompanying tabulation in which "S," "R," and "T" stand for "subject," "relative," and "total" respectively. The letter "f" means "frequencies" and refers to the entries in the body of the table. The figures in the columns and rows at the right and bottom of the table give the operations necessary for finding the coefficient of correlation, with the exception of the two lowest rows at the bottom which are separated from the others and which represent the operations necessary for finding the correlation ratio.

As the tabulation shows, the coefficient of correlation is 0.73. The next step is to find the correlation ratio. This is the ratio between two standard deviations of which the first is that of all the items of the subject series. Since the standard deviation is the square root of the average of the squares of the deviations of the numbers of a series away from their own average, the first step is to find the sum of the squares of the deviations in the subject series.

This is 62 and it has already been found in computing the coefficient of correlation. It is equal to the sum of the squares of the subject series (678) minus the total times the average of that series [678- (136×4.53)=62].

The next step is to find the sum of the squares of the deviations of the averages of each column from the average of the entire subject series. This work can be greatly shortened by following the method utilized in finding the coefficient of correlation. That

Subject Values	Relative Values					T	ST	SST
	1	2	3	4	5			
2	3	1				4	8	16
3	3	1				4	12	36
4	3	2				5	20	80
5	1	3	2			6	30	150
6		2	3	3	3	11	66	396
T	10	9	5	3	3	30	136	678
RT	10	18	15	12	15	70	Av. of sub. items, $\frac{136}{30} = 4.53$	
RRT	10	36	45	48	75	214		
ΣSf	32	40	28	18	18		Av. of rel. items, $\frac{70}{30} = 2.33$	
R(ΣSf)	32	80	84	72	90	358		
$(\Sigma Sf)^2$	1,024	1,600	784	324	324			
$\frac{(\Sigma Sf)^2}{T}$	102.4	177.8	156.8	108	108	653		

$$70 \times 4.53 = 317$$

$$136 \times 4.53 = 616$$

$$70 \times 2.33 = 163$$

$$358 - 317 = 41$$

$$678 - 616 = 62$$

$$214 - 163 = 51$$

$$r = \frac{41}{\sqrt{62 \times 51}} = \frac{41}{\sqrt{3,162}} = \frac{41}{56.23} = 0.73$$

method is based on the principle that the sum of the squares of the deviations of a series away from its average is equal to the sum of the squares of the numbers in the series, minus the product of the total of the series and its average.

Following the same principle we find the sum of the squares of the deviations of the averages of the columns away from the average of the whole subject series. The first step is to find for each column the square of its average multiplied by, or weighted

by, the number of cases in the column. In symbols the desired figure for each column is equal to $\frac{\Sigma Sf}{n} \times \frac{\Sigma Sf}{n} \times \frac{n}{I}$, which is equal to $\frac{(\Sigma Sf)^2}{n}$.

In finding the coefficient of correlation we have already found ΣSf for each column, so the new operations may be performed by squaring these numbers and dividing the result by the number of cases. This is done in the two added rows below the table. Thus, in the third column of the table the entries are 2 and 3 and they correspond to the values 5 and 6. The sum of the subject values times their frequencies (ΣSf) is then (2×5) plus (3×6) or 28. In the next to the last row this number is squared and in the last it is divided by the number of cases (5) giving a result of 156.8, which is the square of the average of the values of the column weighted by the number of cases. This may be verified by noting that the sum of the values of the column is (2×5) plus $(3 \times 6) = 28$ and their average $28/5$. This squared is 31.36, which weighted by the number of cases (5) equals 156.8.

The total of this last row of weighted squares of averages is 653. From this we get the sum of the squares of the deviations from the averages of all the subject values in the table by subtracting the product of the sum of the subject series and its average or $(136 \times 4.53 = 616)$ which gives 37.

Probably the best way for the reader to convince himself that this method is valid and to find out why it is valid is to experiment with simple illustrative cases, carrying the work through by the standard method as described by Rugg and by this method.

We now have two numbers, 62 and 37, which are the sums of squares of deviations and from them we must find the two standard deviations and the ratio between them which will be the correlation ratio. The work is as follows:

$$\eta = \frac{\sqrt{\frac{37}{30}}}{\sqrt{\frac{62}{30}}} = \frac{\sqrt{37}}{\sqrt{62}} = \frac{6.083}{7.874} = 0.77$$

Instead of dividing 37 and 62 by 30 before finding the square roots it is better to eliminate the two 30's. This shortens the work and does not change the relationship between the numerator and denominator.

The correlation ratio is 0.77 whereas the coefficient of correlation is 0.73. The difference is frequently much greater than it is here. The added work necessary to find the correlation ratio is slight. It consists for the most part in the two added lines of figures at the bottom of the table which are easy to get because they are derived from the other computations of the rows above them which have already been used in computing the coefficient of correlation.

There are two correlation ratios for each table of distribution, just as there are two coefficients of regression for each coefficient of correlation. In the case here given the other correlation ratio would be found by working with the relative series and the rows of numbers in the original table instead of with the subject numbers and the columns. The added columns of figures would be at the right of the main table instead of at the bottom. Generally the two correlation ratios are nearly identical so that it is not necessary to find both of them.

Following the method described, the formula for the correlation ratio becomes

$$\eta = \frac{\sigma_{AS}}{\sigma_S}$$

in which the subscript *AS* indicates that the numerator of the fraction is the standard deviation of the averages of the subject values in each column from the average of all the subject values, and the subscript *S* indicates that the denominator is the standard deviation of all the subject values in the table.