

THE UNIVERSITY OF TEXAS AT AUSTIN

SCHOOL OF INFORMATION

MATHEMATICAL NOTES FOR LIS 397.1

INTRODUCTION TO RESEARCH IN LIBRARY AND INFORMATION SCIENCE

Ronald E. Wyllys
Last revised: 2003 Jan 15

CORRELATION AND ANALOGOUS MEASURES

Distinctions among Correlation, Association, Contingency, and Concordance

The technique known as correlation examines the question of whether variables appear to behave in a similar fashion in the way they take on their respective values. Specifically, correlation examines the question of whether there is a tendency for the values of the variables to co-occur (i.e., to occur together) in a patterned way. For example, if two variables, X and Y, are such that large values of X tend to occur together with large values of Y, and small values of X with small values of Y, and so on, then we can say that X and Y appear to be correlated. Because of the co-occurrence of large with large and small with small, we call this *positive* correlation; if two other variables, W and Z, take on their values in such a way that large values of W tend to occur together with small values of Z, and small values of W with large values of Z, we can say that W and Z appear to be *negatively* correlated. If there is no clear tendency of certain values of a pair of variables to co-occur in a patterned way, we call the variables "uncorrelated."

Other words used similarly to "correlation" include "association," "contingency," and "concordance." Choices among these various names are linked to the types of variables involved; "correlation" is usually used to refer to pairs of interval variables and pairs of ordinal variables, and the other terms are usually used to refer to pairs of categorical variables (see Endnote 1). When one wants to talk about a pair of variables that are of different types, the matter of terminology becomes more complicated, as does the matter of assessing whether there is any tendency for their values to co-occur in a patterned way.

An Important Restriction Herein: Only Pairs of Variables of the Same Type

In this discussion, we shall confine ourselves to a highly restricted situation: that of pairs of variables both of which are of the same type. We begin by considering the case of a pair of variables both of which are *interval* variables. Properly speaking, there is even a further restriction--both variables ought to be Gaussianly distributed--, but this further restriction is frequently ignored.

The Pearson Product-Moment Correlation Coefficient

The situation of a pair of interval, Gaussianly distributed variables is the situation to which the Pearson product-moment correlation coefficient applies. First derived in 1896 by Karl Pearson, a British mathematician, this coefficient is the most widely known of the many correlation (association, concordance, contingency) coefficients. In fact, when you find a less than ideally careful author writing about simply "the correlation coefficient" without specifying which of the many coefficients he or she has in mind, you can probably safely assume that the author means the Pearson coefficient.

The essence of the Pearson coefficient is that it is a cleverly constructed arithmetic formula that has the property of measuring the degree to which two variables exhibit a tendency for their values to *co-occur in a pattern* (or to fail to do so). Here is how it works. Let us suppose that we have two variables, X and Y, and that we are curious about whether they are correlated in the sense described above. One thing is absolutely necessary for any attempt to satisfy our curiosity: it must be possible for us to observe *pairs of values* of the variables, i.e., it must be possible for us to observe how the values of X and of Y occur together.

In other words, we must be able to take a random sample such that for *each element* in the sample we can observe *both* a value of X and a value of Y. If we can do this, then we shall be able to gather a sample of pairs of values, which we can label as follows: X_1 and Y_1 will be the values observed on the first element in the sample; X_2 and Y_2 , the values observed on the second element; and in general, X_i and Y_i will be the values observed on the *i*-th (4th, 5th, etc.) element in the sample. For example, from a sample of people we might be able to obtain pairs of values for the

variables of height, X, and weight, Y; we would call the height of the sixth person in the sample X_6 and his or her weight Y_6 .

Given such a sample of n pairs of values of X and Y, we can find the mean and the (sample) standard deviation of the values of X and Y. If we employ the usual notation, these will be

$$\bar{X}, s_X, \bar{Y}, \text{ and } s_Y$$

Then we can construct the following peculiar sum

Formula 1

$$C_{X_1 - \bar{X}} C_{Y_1 - \bar{Y}} + C_{X_2 - \bar{X}} C_{Y_2 - \bar{Y}} + \dots + C_{X_n - \bar{X}} C_{Y_n - \bar{Y}}$$

The sum (see Endnote 2) in Formula 1 has some important properties. To see what they are, you should first consider what happens if value X_1 is large (i.e., greater than the mean value of the X values) and if value Y_1 is also large (greater than the mean of the Y values). In that case

$$C_{X_1 - \bar{X}} \text{ and } C_{Y_1 - \bar{Y}}$$

will be positive, and so their product will be positive. Next suppose that X_2 and Y_2 are both small (less than their respective means). Then

$$C_{X_1 - \bar{X}} \text{ and } C_{Y_1 - \bar{Y}}$$

will both be negative, which will make their product positive. In other words, the co-occurrence of similar values of X and Y (either both large or both small) leads to positive products.

Finally, suppose that X_1 is small (less than the mean of the X values) and that Y_1 is large (greater than the mean of the Y values). Then the first term in

$$C_{X_1 - \bar{X}} \text{ and } C_{Y_1 - \bar{Y}}$$

will be negative, and the second term will be positive; so that their product will be negative. In similar fashion a negative product will result when an X value is large and its paired Y value is small.

The nature of the sum in Formula 1 will depend on the kinds of pairs of values of X and Y that occur in the sample. One possibility is that most of the pairs consist of a large value of X together with a large value of Y, or of a small value of X together with a small value of Y, and only a few of the pairs consist of a large X and a small Y or vice versa. If this is the case, then in Formula 1 most of the products will be positive, and only a few will be negative. The sum will thus be a relatively large positive number. (This is what happens with positively correlated variables.)

Another possibility is that most of the pairs consist of a large value of X together with a small value of Y or vice versa, and only a few of the pairs consist of a large X together with a large Y or a small X together with a small Y. If this is the case, then most of the products will be negative, and only a few will be positive. The sum will thus be a relatively large negative number. (This is what happens with a negatively correlated variables.)

The remaining possibility is that the pairs in the sample will consist of a fairly even mix of (1) large X with large Y, (2) large X with small Y, (3) small X with large Y, and (4) small X with small Y. In this case, the products will be a fairly even mixture of positive and negative; and, since they will tend to cancel each other out, their sum will be close to zero. (This is what happens with uncorrelated variables.)

In short, by turning out to be positive, negative, or close to zero, the sum in Formula 1 captures and expresses the tendency of the values of X and Y to co-occur in a patterned fashion, or to fail to do so. This sum is, in fact, the essential part of the Pearson product-moment correlation coefficient; everything else is just adjustments. The sum has a name of its own, *covariance*, and is often denoted $Cov(X, Y)$. That is,

$$\begin{aligned} Cov(X, Y) &= C_{X_1 - \bar{X}} C_{Y_1 - \bar{Y}} + C_{X_2 - \bar{X}} C_{Y_2 - \bar{Y}} + \dots + C_{X_n - \bar{X}} C_{Y_n - \bar{Y}} \\ &= \sum C_{X_i - \bar{X}} C_{Y_i - \bar{Y}} \end{aligned}$$

The adjustments just mentioned are needed to provide a standard range of values for the measure by which we are going to compare various pairs of variables with respect to the tendency of their values to co-occur in a patterned way. There are two kinds of adjustments. The first is needed because, as a moment's reflection will indicate to you, the larger the sample size n , the larger (either positively or negatively) will be the covariance of a pair of correlated variables, simply because there will be more terms in the sum. Naturally, we want a measure of correlation that is not affected by the happenstance of sample size. To adjust for sample size, we shall divide the covariance by $n-1$, to yield the kind of special average that you are accustomed to seeing used in sample statistics.

When we divide the covariance by $n-1$, we obtain

$$\frac{1}{n-1} \sum (C_{X_i} - \bar{X})(C_{Y_i} - \bar{Y})$$

The second kind of adjustment is needed because the units of measurement used for X and Y affect the size of the covariance. For example, if you calculated the covariance of a sample of height and weight measurements with height in inches and weight in pounds, and then re-calculated it with the weights changed to ounces, the second covariance would be 16 times as large as the first. Naturally, we want a measure of correlation that is not affected by the happenstance of what the original units of measurement were. To adjust for units of measurement, we divide the terms that involve X by s_X and the terms that involve Y by s_Y , as follows:

$$\sum \frac{C_{X_i} - \bar{X}}{s_X} \frac{C_{Y_i} - \bar{Y}}{s_Y}$$

The result of combining these three adjustive divisions is the sample Pearson correlation coefficient, r_{XY} .

$$r_{XY} = \frac{\sum (C_{X_i} - \bar{X})(C_{Y_i} - \bar{Y})}{(n-1)s_X s_Y}$$

This coefficient ranges in value from +1 through 0 to -1. A value of +1 indicates perfect positive correlation (hardly ever realizable in the real world), meaning that not only do large values of X occur only with large values of Y, and small values of X only with small values of Y, but also that each

$$C_{X_i} - \bar{X}$$

deviation as measured in units of s_X is *exactly* matched by the corresponding

$$C_{Y_i} - \bar{Y}$$

deviation as measured in units of s_Y . A value of -1 indicates perfect negative correlation, which is also hardly ever realizable in the real world.

An alternative formula for the sample Pearson correlation coefficient is suited for use with calculators that have only mean and (sample) standard deviation calculations built in as keyboard functions:

$$r_{XY} = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{(n-1)s_X s_Y}$$

Note: You will sometimes see the formulas for r_{XY} written with population standard deviations, σ_X and σ_Y , instead of sample standard deviations. This is incorrect and evidences the writer's confusion of the sample Pearson correlation coefficient r_{XY} with the corresponding population Pearson correlation coefficient, ρ_{XY} , whose formula is

$$\rho_{XY} = \frac{\sum (X_i - \mu_X)(Y_i - \mu_Y)}{N\sigma_X\sigma_Y}$$

where μ_X and μ_Y are the population means of X and Y, respectively, and N is the size of the population.

The proper use of the sample Pearson correlation coefficient is in connection with a test of the hypothesis that the corresponding population Pearson correlation coefficient is 0. That is, the null hypothesis that is built into the standardized statistical procedure for the Pearson correlation coefficient is that the pair of variables under consideration, X and Y, are not correlated. Symbolically, we have $H_0: \rho_{XY} = 0$.

The procedure involves calculating the sample Pearson correlation coefficient, r_{XY} , which is the test statistic for this procedure, and then comparing the observed value of r_{XY} with the appropriate threshold value. (Strictly speaking, we compare the absolute value $|r_{XY}|$ with the threshold if, as in LIS 397.1, we are using a two-tailed test.)

The threshold value will be found in a table of thresholds for the Pearson correlation coefficient. It will be in the column corresponding to the chosen level of significance, α , of the test, and it will ordinarily be in the row corresponding to the number of degrees of freedom, df. For a pair of variables, the number of degrees of freedom will be 2 less than the sample size, or $n - 2$. (If you are using the Pearson partial correlation coefficient, a technique applicable to the study of the possible correlation of more than 2 variables, then the degrees of freedom will be $n - v$, where v is the number of variables.)

As usual, only if the (absolute) observed value of the test statistic r_{XY} exceeds the tabled threshold value should you reject the null hypothesis. If you can reject it, then you can consider the value of r_{XY} as your best estimate of the true, population Pearson correlation coefficient ρ_{XY} . Otherwise, you must conclude that $\rho_{XY} = 0$, i.e., that X and Y are uncorrelated.

You will recognize that this amounts to accepting a point estimate of a population parameter; and you will remember that in talking about the population mean, we stressed that it is preferable to work out a confidence interval for it rather than settling for a less informative point estimate. As you probably suspect, there are procedures for calculating confidence intervals for the population Pearson correlation coefficient, but we shall not discuss them in LIS 397.1.

Here is a table of thresholds for the Pearson product-moment correlation coefficient, prepared from a variety of sources and tailored to the needs of students in LIS 397.1.

Degrees of freedom	$\alpha = .05$	$\alpha = .01$	Degrees of freedom	$\alpha = .05$	$\alpha = .01$
1	.997	1.000	31	.344	.442
2	.950	.990	32	.339	.436
3	.878	.959	33	.334	.430
4	.811	.917	34	.329	.424
5	.754	.875	35	.325	.418
6	.707	.834	36	.320	.413
7	.666	.798	37	.316	.408
8	.632	.765	38	.312	.403
9	.602	.735	39	.308	.398
10	.576	.708	40	.304	.393
11	.553	.684	45	.288	.372
12	.532	.661	50	.273	.354
13	.514	.641	60	.250	.325
14	.497	.623	70	.232	.302
15	.482	.606	80	.217	.283

16	.468	.590	90	.205	.267
17	.456	.575	100	.195	.254
18	.444	.561	125	.174	.228
19	.433	.549	150	.159	.208
20	.423	.537	175	.149	.195
21	.413	.526	200	.138	.181
22	.404	.515	300	.113	.148
23	.396	.505	400	.098	.128
24	.388	.496	500	.088	.115
25	.381	.487	600	.083	.108
26	.374	.478	800	.072	.095
27	.367	.470	1000	.062	.081
28	.361	.463			
29	.355	.456			
30	.349	.449			

The Spearman Rank-Order Correlation Coefficient

Next, we consider another restricted situation, that of a pair of ordinal (ranking, rank-order) variables. This situation can be handled by a rank-order correlation coefficient, usually called "rho," which was derived by a British psychologist, Charles Spearman, in 1904. It is difficult to find practical examples of the use of this coefficient outside the area of psychology, for the essence of the coefficient is that it assesses the correlation between pairs of rankings, assigned to some set of objects or concepts, by two different persons or from two different sources of evaluations. But let us proceed anyway.

We assume that we have a set of n things (objects or concepts), to each of which two judges, X and Y, have assigned rankings. That is, each judge has considered the n things and assigned to each of them a number from 1 to n that represents his or her judgment of the thing, with 1 representing the highest rank. To give the simplest kind of useful example, we may suppose that we have five things, named I, II, III, IV, and V, and that the judges' rankings are as follows:

	<u>I</u>	<u>II</u>	<u>III</u>	<u>IV</u>	<u>V</u>
Judge X:	1	3	5	2	4
Judge Y:	3	2	4	1	5

These data must be viewed as a sample, realized on this particular set of five things, of the extent to which X and Y tend to agree in their judgments. The appropriate null hypothesis is that their overall degree of agreement is zero. In general, we would have n pairs of judgments as a sample with which to test the null hypothesis of zero overall degree of agreement.

What the Spearman rank-order correlation coefficient does for us is to let us use a sample of pairs of judgments to calculate a sample correlation coefficient, i.e., a test statistic, which we shall then compare with an appropriate threshold value from a table in order to decide whether the null hypothesis of zero overall degree of agreement should be accepted or rejected. The sample Spearman coefficient, ρ or $\rho(X, Y)$, is calculated just like the sample Pearson coefficient:

$$\rho_{X, Y} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{(n-1)s_X s_Y}}$$

In practical terms this means that calculating the Spearman coefficient is very easy if you have a computer program designed to calculate the Pearson coefficient, r_{XY} , or if you have a calculator with a built-in Pearson-coefficient function. If so, then to calculate the Spearman coefficient you can simply enter the pairs of values into the program or the calculator exactly as you would for a calculation of the Pearson coefficient. For the data above concerning the rankings by Judges X and Y, we find: $\rho(X, Y) = 0.60$.

Unfortunately for consistency, the usual pattern of a Roman letter for the sample value and a Greek letter for the corresponding population parameter has not become standard in the case of the Spearman coefficient. We are free to adopt a notation such as $\rho_{\text{pop}}(X, Y)$ for the population Spearman coefficient, and to state the null hypothesis as

$$H_0: \rho_{\text{pop}}(X,Y) = 0$$

The procedure involved in testing this null hypothesis is to calculate the test statistic, the sample Spearman coefficient, $\rho(X,Y)$, and to compare this observed value with a threshold value from the appropriate table. Although the Spearman coefficient is arithmetically equivalent to the Pearson coefficient, we cannot use the Pearson table of thresholds, for the obvious reason that the Spearman coefficient is applicable to a situation quite different from that of the Pearson coefficient. We must use a table constructed for the purpose of dealing the Spearman coefficient. Here is such a table (prepared for LIS 397.1). In the table, n denotes the number of paired ranks, and the entries are for levels of significance $\alpha = .05$ and $\alpha = .01$:

Threshold values of rho, the Spearman rank-order correlation coefficient					
<u>n</u>	<u>.05</u>	<u>.01</u>	<u>n</u>	<u>.05</u>	<u>.01</u>
5	1.000				
6	.886	1.000	21	.438	.576
7	.786	.929	22	.428	.562
8	.738	.881	23	.418	.549
9	.683	.833	24	.409	.537
10	.648	.794	25	.400	.526
11	.620	.774	26	.392	.515
12	.591	.755	27	.384	.505
13	.566	.735	28	.377	.496
14	.544	.715	29	.370	.487
15	.524	.688	30	.364	.478
16	.506	.665			
17	.490	.644			
18	.475	.625			
19	.462	.607			
20	.450	.591			

As usual, only if the (absolute) observed value of the test statistic exceeds the tabled threshold value should you reject the null hypothesis. If you can reject it, then you can consider the value of $\rho(X,Y)$ as your best estimate of the overall tendency of the judges (or more generally, the sources of the rankings) to agree. Otherwise, you must conclude that the judges exhibit no overall tendency to agree. In the example above of rankings by Judges X and Y, the observed value of ρ , .60, is considerably less than the threshold value of 1.0000 for a sample of 5 pairs of rankings and a level of significance $\alpha = .05$. Thus we decide to accept the null hypothesis that these judges exhibit no tendency to agree in their rankings.

The Phi Association Coefficient

One more restricted situation remains to be considered: that of a pair of categorical, or nominal, variables. Here is quite difficult to talk meaningfully about the degree of association, even though the chi-square test of association (derived by the British mathematician, Karl Pearson, in 1900) does enable us to accept or reject the null hypothesis of no association between a pair of categorical variables. That is, we can say that there is or is not some association, but to assign a figure to the strength of the association (if it exists) is difficult.

There is not much to be said about this situation except to mention that you occasionally see people using the coefficient denoted by ϕ (phi) and defined as

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

where χ^2 is the chi-square statistic and n is the total number of observations.

Obviously, ϕ can be meaningful only if the observed value of χ^2 in a chi-square test of association (as discussed below) is large enough so that this test's null hypothesis of no association may be rejected; however, even when that null hypothesis is rejected, it is still not clear just what ϕ means. About as far as you can go is to say, if you have two pairs of categorical variables, that one of the pairs might be shown by the ϕ coefficient to be more strongly associated than the other pair.

Endnotes

1. You may recall that interval variables are those that result from physical measurements or counts, ordinal variables are those that simply express order (or rank or sequence), and categorical variables are those that simply identify classes or categories. For example, a height of 28 cm for a book is an example of an interval variable; a ranking of 3 ("You had the 3rd highest grade in the course") is an example of an ordinal variable; and classifications such as "history major" and "English major" are examples of a categorical variable.

Ratio variables are interval variables that also possess the property of having a true zero; e.g., height is a ratio variable, but Fahrenheit temperature is not, since it has merely an arbitrary zero rather than a true zero. (The only regularly used temperature scale that has a true zero is the Kelvin scale, though the almost forgotten Rankine temperature scale also has a true zero. For both the Kelvin and Rankine scales, the zero is the "absolute zero" of physics, the temperature at which molecular motion ceases completely.) Most ordinary interval variables encountered in library and information science are also ratio variables.

2. A similar sum occurs when one calculates the torque, or moment of force, exerted by forces applied to a lever such as a wrench. The total moment of force is found by adding the products of each of the individual forces times its distance from the axis of rotation (e.g., the center of a nut that is being tightened). The similarity between the formulas for torque and for the sum displayed above gave rise to the name, "product-moment correlation."
3. There is an alternative method of calculating the sample Spearman coefficient:

$$r_{s_{X,Y}} = 1 - \frac{6 \sum D_i^2}{n(n^2 - 1)}$$

where the D_i terms represent the various pairwise differences between the judges' rankings; i.e., $D_i = X_i - Y_i$, for $i = 1, 2, \dots, n$. (For example, the differences between the rankings above by Judges X and Y are: -2, 1, 1, 1, -1.)

Useful only for simple four-function calculators, this formula is an arithmetic shortcut made possible because of the severely restricted nature of rank numbers, which must be integers and which must begin with 1 and not exceed n . If you have a calculator with a correlation function (or a trend function) built into it, you will find it much easier to enter the data directly than to use this shortcut formula. Similarly, it is much easier to enter the data directly into a statistical program package's correlation procedure than to write a program to carry out the shortcut formula.