Spurious Correlation between Ratios with a Common Divisor

Ji-Hyun Kim [1]

ABSTRACT

One hundred years ago Karl Pearson derived an approximate formula for the correlation between ratios with a common divisor and cautioned to be wary of correlating ratios. The exact formula for the correlation between ratios is derived. It can provide a better reference point of no connection when correlating ratios with a common divisor.

KEY WORDS: Coefficient of variation; spurious correlation.

1 INTRODUCTION

Pearson (1897) pointed out the fallacy of correlation coefficient between ratios that have a common divisor. Even though X and Y are totally uncorrelated random variables, two ratios X/Z and Y/Z for another arbitrary random variable Z can have a misleadingly large value of correlation coefficient. This raised the interpretation problem of correlation coefficient. Aldrich (1995) treated the different views of Karl Pearson and G. Udny Yule on the correlation analysis including the interpretation problem.

In this paper the exact formula for the correlation coefficient between ratios is derived. The approximate formula Pearson (1897) has derived is based on the assumptions which have been shown to be faulty due to his ignoring the higher order terms in his derivation. (See the item 'Spurious Correlation' in Kotz and Johnson (1982).) The exact formula can provide a better reference point of no connection when correlating ratios with a common divisor.

[1]Ji-Hyun Kim is Associate Professor, Department of Statistics, Soong Sil University, Dongjak-Gu, Seoul 156-743, South Korea (E-mail: jhkim@stat.soongsil.ac.kr).

2 THE EXACT FORMULA FOR THE CORRELATION COEFFICIENT BETWEEN RATIOS WITH A COMMON DIVISOR

Pearson's approximate formula for the correlation between ratios with a common divisor is

$$r_{(X/Z)(Y/Z)} \simeq \frac{r_{XY}V_XV_Y - r_{XZ}V_XV_Z - r_{YZ}V_YV_Z + V_Z^2}{\sqrt{(V_X^2 + V_Z^2 - 2r_{XZ}V_XV_Z)(V_Y^2 + V_Z^2 - 2r_{YZ}V_YV_Z)}}$$

where V is the coefficient of variation, e.g. $V_X = \sqrt{\mathrm{var}(X)}/\mathrm{E}(X)$. When X, Y and Z are uncorrelated, it is simplified to

$$r_{(X/Z)(Y/Z)} \simeq \frac{V_Z^2}{\sqrt{(V_X^2 + V_Z^2)(V_Y^2 + V_Z^2)}} \tag{2.1}$$

The exact formula for the correlation can be drawn as follows.

Theorem. *If the random variables X, Y and Z are independent, the correlation between ratios X/Z and Y/Z is*

$$r_{(X/Z)(Y/Z)} = \frac{V_{1/Z}^2 \ \mathrm{sgn}(\mathrm{E}(X))\ \mathrm{sgn}(\mathrm{E}(Y))}{\sqrt{[V_X^2(1+V_{1/Z}^2) + V_{1/Z}^2][V_Y^2(1+V_{1/Z}^2) + V_{1/Z}^2]}},$$

where $\mathrm{sgn}(a) = a/|a|$.

Proof.

$$\begin{aligned}
r_{(X/Z)(Y/Z)} &= \frac{\mathrm{cov}(XW, YW)}{\sqrt{\mathrm{var}(XW)\mathrm{var}(YW)}}, \text{ where } W = 1/Z \\
&= \frac{\mathrm{E}(XYW^2) - \mathrm{E}(XW)\mathrm{E}(YW)}{\sqrt{[\mathrm{E}(X^2W^2) - \mathrm{E}^2(XW)][\mathrm{E}(Y^2W^2) - \mathrm{E}^2(YW)]}} \\
&= \frac{\mathrm{E}(X)\mathrm{E}(Y)\mathrm{E}(W^2) - \mathrm{E}(X)\mathrm{E}(Y)\mathrm{E}^2(W)}{\sqrt{[\mathrm{E}(X^2)\mathrm{E}(W^2) - \mathrm{E}^2(X)\mathrm{E}^2(W)][\mathrm{E}(Y^2)\mathrm{E}(W^2) - \mathrm{E}^2(Y)\mathrm{E}^2(W)]}} \\
&= \frac{\mathrm{E}(X)\mathrm{E}(Y)\mathrm{var}(W)}{\sqrt{[\mathrm{E}(X^2)\mathrm{var}(W) + \mathrm{E}^2(W)\mathrm{var}(X)][\mathrm{E}(Y^2)\mathrm{var}(W) + \mathrm{E}^2(W)\mathrm{var}(Y)]}} \\
&= \frac{\mathrm{E}(X)\mathrm{E}(Y)V_W^2}{\sqrt{[\mathrm{E}(X^2)V_W^2 + \mathrm{var}(X)][\mathrm{E}(Y^2)V_W^2 + \mathrm{var}(Y)]}} \\
&= \frac{V_W^2 \ \mathrm{sgn}(\mathrm{E}(X))\ \mathrm{sgn}(\mathrm{E}(Y))}{\sqrt{[(1+V_X^2)V_W^2 + V_X^2][(1+V_Y^2)V_W^2 + V_Y^2]}}. \qquad \square
\end{aligned}$$

If X and Y take positive values, or more generally $E(X)$ and $E(Y)$ take the same sign, then the formula in the theorem becomes

$$r_{(X/Z)(Y/Z)} = \frac{V_{1/Z}^2}{\sqrt{[V_X^2(1+V_{1/Z}^2)+V_{1/Z}^2][V_Y^2(1+V_{1/Z}^2)+V_{1/Z}^2]}} \tag{2.2}$$

Equation (2.2) tells that even though the random variables X, Y and Z have no relationship, the correlation $r_{(X/Z)(Y/Z)}$ takes a positive value unless Z is constant. Let us take one numerical example. Let X, Y and Z have independent and identical distribution of Uniform(1,2). Then $V_{1/Z}^2 = \mathrm{var}(1/Z)/\mathrm{E}^2(1/Z) = (1/2 - \ln^2 2)/\ln^2 2 = .0407$ and $V_X^2 = V_Y^2 = (1/12)/(1.5)^2 = .0370$, hence $r_{(X/Z)(Y/Z)} = .51$, which is close to 1/2, the value of Pearson's approximate formula.

The distinction between the approximate formula (2.1) and the exact formula (2.2) gets dilated as the difference between V_Z and $V_{1/Z}$ gets larger. For example, if X, Y and Z have independent and identical distribution of Uniform(1,11), then $V_Z = .48$, $V_{1/Z} = .76$, so that (2.1) and (2.2) are 0.50 and 0.61, respectively. More drastic difference may occur as in the real data example below.

3 SOME ARTIFICIAL AND REAL EXAMPLE

The exact formula (2.2) can be used to investigate how large the value of spurious correlation coefficient can be. We consider two special cases: Case 1; $V_X = V_Y = V_{1/Z}$, Case 2; $V_X = V_Y, V_{1/Z} = kV_X$ for some constant k.

Case 1. If $V_X^2 = V_Y^2 = V_{1/Z}^2 = c$, then $r_{(X/Z)(Y/Z)} = 1/(c+2)$, which takes values 1/2 through 1/3 while c ranges from 0 to 1.

Case 2. If $V_X^2 = V_Y^2 = c$ and $V_{1/Z}^2 = kc$, then $r_{(X/Z)(Y/Z)} = \frac{k}{k(c+1)+1}$. Note that $\lim_{k\to\infty} \frac{k}{k(c+1)+1} = \frac{1}{c+1}$, so that the correlation can be close to 1 with large k and small c. In other words, if the coefficinet of variation of $1/Z$ is large enough relative to those of X and Y, the correlation coefficient $r_{(X/Z)(Y/Z)}$ can be misleadingly large. We take one artificial example for the second case. Let X and Y be independent and uniformly distributed on (10, 11). And let Z be uniformly distributed on (5.5, 15.5). Then $V_X^2 = V_Y^2 = .000756$ and $V_{1/Z}^2 = .0927$, so that $k = 123.0$, and $r_{(X/Z)(Y/Z)} = .99$. The extreme unbalance among

variations of variables like this case ($V_Z/V_X = \sqrt{\text{var}(Z)}/\sqrt{\text{var}(X)} = 10$) may not often occur. But it suggests that the correlation between ratios can be arbitrarily close to 1 even when X, Y and Z are uncorrelated variables. The scatter plots in Figure 1 help to understand what happens. We generated 100 set of random numbers (x, y, z) from the appropriate uniform distributions mentioned above for the scatter plots. When the common divisor Z has large variation relative to X and Y as in this example, it dominantly affects the distribution of X/Z and Y/Z.

Now let us take a real data example. The data taken from the internet site with address http://www.stat.ncsu.edu/info/jse gives birth rates per 1,000 of population, death rates and other demographic features for 97 countries. (Original sources of the data are the U.N.E.S.C.O. 1990 Demographic Year Book and The Annual Register 1992) People might ask the question: Are birth rates related to death rates? Let X, Y and Z denote number of births, number of deaths and size of population, respectively. To answer the question, the correlation coefficient $r_{(X/Z)(Y/Z)}$ need to be calculated with the scatterplot. The following results were obtained: $V_X = 2.734$, $V_Y = 2.460$, $V_Z = 2.575$, $V_{1/Z} = 1.448$; $r_{(X/Z)(Y/Z)} = .486$. (We think of 97 countries as a population rather than a sample.) Assuming that X, Y and Z are independent, estimates of $r_{(X/Z)(Y/Z)}$ by (2.1) and (2.2) are .496 and .091, respectively. The approximate formula gives a value not approximate at all mainly due to the large discrepancy between the estimates of V_Z and $V_{1/Z}$. The reference point of no relation for the correlation coefficient .486 should be .091 *not* 0, even though they are not much different in this case. To avoid the proper reference point problem people might ask the question: Are the number of births X related to the number of deaths Y? The correlation coefficient r_{XY} is .978.

[Figure 1 is inserted here.]

4 DISCUSSION

The large correlation between X/Z and Y/Z when X, Y and Z are uncorrelated is spurious.

Aldrich (1995), quoting Pearson's earlier works, defined spurious correlation to be a correlation which is produced by a process of arithmetic and not by any organic relationship among the quantities dealt with. Pearson (1897) pointed out the spurious correlation problem with his famous 'bone' example and cautioned to be wary of correlating ratios. If one is forced to correlate ratios, he suggests to adopt as the point of no connection not 0, but some such value as 0.4. The equation (2.2) provides a better reference point of no connection. We can get the estimate of correlation between ratios for the case of no organic relationship among variables of interest by calculating the coefficients of variation of the variables and plugging in the equation (2.2).

REFERENCES

Aldrich, J. (1995), "Correlation Genuine and Spurious in Pearson and Yule," *Statistical Science*, 10, 364-376.

Day, A. (ed.) (1992), *The Annual Register 1992*, 234, London: Longmans.

Kotz, S. and Johnson, N. L. (1982), *Encyclopedia of Statistical Sciences*, John Wiley.

Pearson, K. (1897). "On a form of spurious correlation which may arise when indices are used in the measurement of organs," *Proc. Roy. Soc. London Ser. A*, 60, 489-498.

U.N.E.S.C.O. 1990 Demographic Year Book (1990), New York: United Nations.

Figure 1. Multiple Scatter Plots for $X, Y, Z, X/Z$ and Y/Z. The scatter plots are based on an artificial sample of 100 set of points (x, y, z) from independent uniform distributions. The coefficient of variation of Z is ten times larger than those of X and Y, so that plays a dominant role on the relationship between X/Z and Y/Z.