The Spurious Correlation of Ratios That Have Common Variables: A Monte Carlo Examination of Pearson's Formula

WILLIAM P. DUNLAP
JÖRG DIETZ
JOSE M. CORTINA
Department of Psychology
Tulane University

ABSTRACT. Pearson (1897) investigated correlations of ratios of bone measurements and found that although the correlations among the original measures were low, the correlations among ratios with common measures were about .5. To understand this result, he developed an approximate equation for the correlations of ratios. In the present study, Monte Carlo methods were used to show that Pearson's equation is fairly accurate and that correlations among ratios with common elements (e.g., X/C, Y/C) are indeed at least partly spurious, as Pearson concluded. This finding should serve as a two-fold warning to those who might correlate ratios that have common elements: (a) Interpretation of the observed relationship between such ratios may at best be tenuous, and (b) the relationships among the elements themselves may call into question the usefulness of addressing a hypothesis that can be tested only with a correlation between ratios that share elements.

RATIO VARIABLES (sometimes called index variables) are formed by dividing one variable into another (Cohen & Cohen, 1983, p. 73). Neyman (1952) provided a fictitious example of the problems that can occur when correlating ratio variables, by presenting a table of counts of births, storks, and women in 54 counties. Unfortunately, larger counties had more storks, births, and women. To control for county size, Neyman computed rates per county, dividing the numbers of storks by women and dividing births by women. He then correlated the two ratios, finding a correlation of .62, which supported the notion that storks served as a mediator variable between women and offspring. In Neyman's example, the partial correlation between storks and births, with women controlled, is in fact 0. Neyman went on to argue that it is not the positive correlation between ratios that has a common denominator that is spurious; instead, the idea that dividing by the common element would somehow control that variable is what is spurious.

Address correspondence to William P. Dunlap, Department of Psychology, Tulane University, New Orleans, LA 70118.
A century ago, Pearson (1897) showed that the correlation between two ratio variables that have the same denominator may equal .5, even when the original variables from which the ratios were formed are uncorrelated. He concluded that the "part [of the correlation between ratio variables that] is solely due to the nature of [the] arithmetic, . . . is spurious" (p. 491). However, the accuracy of Pearson's formula, introduced below, may be open to question because Pearson assumed that terms of the third and fourth powers of the coefficients of variation were unnecessary.

Our purposes in the present study were (a) to examine empirically the accuracy of Pearson's formula when the ratio variables share denominators or numerators or when the numerator of one ratio is also the denominator of the other and (b) to examine the implications for research of our findings. For those purposes, data were generated through Monte Carlo simulation for uncorrelated, moderately correlated, and highly correlated ratio components that represented "raw variables" (Chayes, 1971). Other forms of correlations between ratio variables, such as correlations between composite percentages, will not be specifically addressed in this article (see Aitchison, 1986, for a thorough review on correlations of compositional data).

Although the issue of spuriousness of correlations between ratio variables that have a common element has been raised by numerous authors (Anderson & Paine, 1978; Bettis & Mahajan, 1985; Cohen & Cohen, 1983; Firebaugh & Gibbs, 1985; Fuguit & Lieberson, 1974; Logan, 1982; Long, 1979; Lynn & Bond, 1992; Kuh & Meyer, 1955; MacMillan, Hambrick, & Day, 1982; McNemar, 1962), those correlations are still used in industrial and organizational psychology (Ostroff, 1992; Ryan, Schmit, & Johnson, 1996), social psychology (McClintock & Hunt, 1975; Levine, Martinez, Brase, & Sorenson, 1994), educational psychology (Benton, Corkill, Sharp, Downey, & Khramtsova, 1995), research in organizational behavior (Boeker, 1992; Smith, Grimm, Gannon, & Chen, 1991), and management science (Amit & Wernerfelt, 1990; Buzzell & Gale, 1987; Hambrick, MacMillan, & Day, 1982; Haunschild, 1994; Kelm, Narayanan, & Pinches, 1995; Miller & Bromiley, 1990). For example, Smith et al. (1991), in a study involving domestic airlines, examined the relationship between profitability, which was operationalized as the ratio of net profit to revenue, and absorbed slack, which was operationalized as the ratio of expenditures to revenue. Smith et al. reported a correlation of -.94 between those index variables, which share the common denominator. Although Smith et al. may have expected a positive relationship between those constructs, there is no theoretical reason for believing that they are interchangeable, as suggested by their observed correlation. Nevertheless, they failed to recognize common ratio elements as a contributor to the strength of the relationship.

Researchers who look for advice in the methodological literature on how to deal with correlations among ratio variables that have a common element find a range of recommendations, from seeing no problem to a recognition that the problem might be insurmountable (Pearson, 1897). For example, Yule (1910) was the first to raise the argument that if the theory that drives the research refers to ratios rather than to the original variables, the correlations among ratios are proper and correlations.
among the original variables may be deemed spurious. Fuguitt and Lieberson (1974) stated that “spurious correlation is not an issue in correlating ratios or differences having common terms, provided that one’s interest is exclusively grounded in the composite variables rather than in the components” (p. 141). Long (1979) went a step further, stating that “the use of ratio variables with common components . . . , regardless of the position of the shared component, does not constrain or make more likely one sign or direction of association over another (p. 38).”

Others, however, have recommended caution in using correlations with ratio variables that have a common element. For example, Tittle (1969) suggested that the observed correlation should be compared with the maximum possible correlation computed from Pearson’s formula, instead of a null correlation. Logan (1982) advocated the use of semipartial correlations, partialling the common element, to inspect for artifactual inflation of zero-order correlations. The problem with that approach is that it requires an independent measure of the common element, which often is not available. Firebaugh and Gibbs (1985) recommended the use of regression analysis rather than correlation analysis, but Kronmal (1993) showed that the use of ratios in regression analysis, as in correlation analysis, often leads to misleading results. Cohen and Cohen (1983) gave specific advice that one should not correlate ratio variables that have a common element unless the correlation between numerator and denominator is high. The obvious lack of agreement and consistency in recommendations about correlating ratio variables that share components was our primary motivation for the present article. One way to enlarge the agreement in the discussion on correlations between ratio variables is to ground the discussion in empirical evidence, which can be achieved through Monte Carlo computer simulation. There have been several Monte Carlo studies of correlations among ratios that share common variables. Chayes (1971), for example, investigated the effects of different coefficients of variation for cases in which all original variables were uncorrelated. Chayes found, as expected, that the accuracy of Pearson’s (1897) equation was best when the coefficients of variation were smallest. The coefficient of variation for the present study was 0.14, about in the middle of the range examined by Chayes. We added to Chayes’s work by examining cases of moderately and highly correlated original variables.

MacMillan and Daft (1979, 1980) also performed an empirical investigation of correlations among ratios with common components; in their model, the numerator changed systematically with the size of the denominator. They reported little evidence of spuriousness. However, Logan (1982) criticized MacMillan and Daft because they overconstrained their data by assuming specific linear relationships between the ratio components. Pendleton, Newman, and Marshall (1983) conducted a Monte Carlo study of the impact of using ratio variables in multiple-regression equations in which all predictors had the same common denominator. Under most conditions studied, the resulting regression equations were dramatically inaccurate, indicating severe spuriousness. This spuriousness is well captured in Pearson’s ratio correlation formula, which we shall discuss next.
Mathematical Development

Approximate Correlations of Ratio Variables

Pearson (1897) gave a formula for approximate correlations among ratio variables; it involved the simple correlations among all relevant variables and the coefficients of variation of all of the variables. This formula has been repeated by others (e.g., Cohen & Cohen, 1983) in their treatment of the problem of correlations among ratio or index variables. The coefficient of variation of a variable is its population standard deviation divided by its mean. Because Pearson dropped all terms when the coefficient of variation had a power higher than 2, he was assuming that the coefficients of variation are all rather small, as was the case in the example he presented. The formula (Pearson, 1897, Equation 4) is

\[ r_{(Y/Z)(X/W)} = \frac{r_{xy}v_xv_y - r_{yw}v_yv_w - r_{xz}v_xv_z + r_{zw}v_zv_w}{(v_y^2 + v_z^2 - 2r_{yz}v_yv_z)^{1/2}(v_x^2 + v_w^2 - 2r_{xz}v_xv_w)^{1/2}}, \]  

(1)

where the symbol \( V \) represents the coefficient of variation.

To further simplify this formula to get indications as to how intercorrelations of ratio variables might behave, we assumed that the coefficients of variation of all variables were equal, in which case the coefficients of variation all canceled out, giving the following equation:

\[ r_{(Y/Z)(X/W)} = \frac{(r_{xy} - r_{yw} - r_{xz} + r_{zw})}{[2(1 - r_{yz})^{1/2}(1 - r_{xz})^{1/2}].} \]  

(2)

Common Denominators

To examine the case in which two ratio variables have a common denominator, we left \( X \) and \( Y \) in Equation 1 as they were, but we set the common element \( C = Z = W \) (see also, e.g., Schuessler, 1974). After these substitutions were made and the terms were rearranged, the resulting equation was

\[ r_{(Y/Z)(X/C)} = \frac{(1 - r_{yc} - r_{xc} + r_{xy})}{[2(1 - r_{yc})^{1/2}(1 - r_{xc})^{1/2}].} \]  

(3)

Even if the correlations among the original variables all equaled 0, the correlation among ratios with a common denominator would not equal 0; instead, Equation 3 simplified to .5. Thus, according to Pearson's (1897) equation, the spurious correlation produced by the common denominator is .5 when all variables are uncorrelated.

Common Numerators

To examine the case of correlations among ratios that have a common numerator, we set the common element \( C = Y = X \) in Equation 1 and set the denominators \( Z \) and \( W \) equal to \( Y \) and \( X \), respectively. Those substitutions resulted in the following equation (which was identical to Equation 3) for correlations among index variables with a common numerator:
\[
\rho_{(CYXC/X)} = \frac{(1 - r_{YC} - r_{XC} + r_{XY})}{2(1 - r_{YC})^{1/2}(1 - r_{XC})^{1/2}}.
\] (4)

Again, it is clear that when all three original variables are independent, the correlation among ratios sharing a common numerator will not be 0 but will equal .5. Thus, ratios sharing a numerator will be spuriously correlated as badly as those sharing denominators.

**Numerator Common to Other Denominator**

Finally, if the numerator of one variable is the denominator of the other variable, substituting \( Y, X, \) and \( C \) appropriately yields the following:

\[
\rho_{(CYXY/CX)} = -\frac{(1 - r_{YC} - r_{XC} + r_{XY})}{2(1 - r_{YC})^{1/2}(1 - r_{XC})^{1/2}}.
\] (5)

The only change from Equations 3 and 4 above was a change in the sign of the numerator. Again, with uncorrelated original variables, the correlation between ratios will not be 0 but instead will equal -.5. The fact that the correlation was negative makes sense, in that a large value for \( C \) will simultaneously make the first ratio larger and the second ratio smaller. (See Table 1 for the results of these theoretical relationships for Equations 3 and 4. For Equation 5, the values in Table 1 must be multiplied by -1.) Because the limits for \( r_{XY} \) are \( r_{XC} \times r_{YC} + \frac{1}{- [(1-r_{XC}^2)(1-r_{YC}^2)]^{1/2}}, \) certain combinations of the three correlations shown in the tables cannot exist. For example, if \( r_{XC} \) and \( r_{YC} \) both equal .8, the smallest \( r_{XY} \) can be is .28. Consequently, the 10 combinations of correlation conditions in the bottom left corners of the tables have been left blank.

<table>
<thead>
<tr>
<th>( r_{XC} )</th>
<th>( .0 )</th>
<th>( .1 )</th>
<th>( .2 )</th>
<th>( .3 )</th>
<th>( .4 )</th>
<th>( .5 )</th>
<th>( .6 )</th>
<th>( .7 )</th>
<th>( .8 )</th>
<th>( .9 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( r_{YC} )</td>
<td>( .50 )</td>
<td>( .55 )</td>
<td>( .60 )</td>
<td>( .65 )</td>
<td>( .70 )</td>
<td>( .75 )</td>
<td>( .80 )</td>
<td>( .85 )</td>
<td>( .90 )</td>
<td>( .95 )</td>
</tr>
<tr>
<td>( .1 )</td>
<td>( .44 )</td>
<td>( .50 )</td>
<td>( .56 )</td>
<td>( .61 )</td>
<td>( .67 )</td>
<td>( .72 )</td>
<td>( .78 )</td>
<td>( .83 )</td>
<td>( .89 )</td>
<td>( .94 )</td>
</tr>
<tr>
<td>( .2 )</td>
<td>( .38 )</td>
<td>( .44 )</td>
<td>( .50 )</td>
<td>( .56 )</td>
<td>( .62 )</td>
<td>( .69 )</td>
<td>( .75 )</td>
<td>( .81 )</td>
<td>( .88 )</td>
<td>( .94 )</td>
</tr>
<tr>
<td>( .3 )</td>
<td>( .29 )</td>
<td>( .36 )</td>
<td>( .43 )</td>
<td>( .50 )</td>
<td>( .57 )</td>
<td>( .64 )</td>
<td>( .71 )</td>
<td>( .78 )</td>
<td>( .86 )</td>
<td>( .93 )</td>
</tr>
<tr>
<td>( .4 )</td>
<td>( .17 )</td>
<td>( .25 )</td>
<td>( .33 )</td>
<td>( .42 )</td>
<td>( .50 )</td>
<td>( .58 )</td>
<td>( .67 )</td>
<td>( .75 )</td>
<td>( .83 )</td>
<td>( .92 )</td>
</tr>
<tr>
<td>( .5 )</td>
<td>( .00 )</td>
<td>( .10 )</td>
<td>( .20 )</td>
<td>( .30 )</td>
<td>( .40 )</td>
<td>( .50 )</td>
<td>( .60 )</td>
<td>( .70 )</td>
<td>( .80 )</td>
<td>( .90 )</td>
</tr>
<tr>
<td>( .6 )</td>
<td>( -.25 )</td>
<td>( -.12 )</td>
<td>( .00 )</td>
<td>( .12 )</td>
<td>( .25 )</td>
<td>( .38 )</td>
<td>( .50 )</td>
<td>( .62 )</td>
<td>( .75 )</td>
<td>( .87 )</td>
</tr>
<tr>
<td>( .7 )</td>
<td>( -.67 )</td>
<td>( -.50 )</td>
<td>( -.33 )</td>
<td>( -.17 )</td>
<td>( .00 )</td>
<td>( .17 )</td>
<td>( .33 )</td>
<td>( .50 )</td>
<td>( .67 )</td>
<td>( .83 )</td>
</tr>
<tr>
<td>( .8 )</td>
<td>( -.75 )</td>
<td>( -.50 )</td>
<td>( -.25 )</td>
<td>( .00 )</td>
<td>( .25 )</td>
<td>( .50 )</td>
<td>( .75 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( .9 )</td>
<td>( -.50 )</td>
<td>( .00 )</td>
<td>( .50 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note.* Empty cells represent impossible conditions.
The above equations give approximations of the correlation to expect among ratios sharing common elements, given that the coefficients of variation are small and all roughly equal. If the correlation between ratios is taken as an estimate of the correlation between $X$ and $Y$, the differences between the entries in the body of Table 1 and the column headings estimate the extent of spuriousness induced by the common element. We determined the accuracy of the approximations by examining the simulations described below.

Monte Carlo Simulations With Ratio Variables

Method

For the simulation, we wrote a program in FORTRAN, compiled with an XL FORTRAN Compiler/6000 Version 3 and executed on an IBM AIX RISC System/6000 computer. Subroutines from the IMSL statistical library (IMSL, Inc., 1987) were used to generate random standard multivariate-normal deviates. First, the population matrix of correlations among three original variables (labeled $Y$, $X$, and $C$ in the text of this article) was entered, and this matrix was decomposed into Cholesky factors by subroutine CHFAC from the IMSL package. Then, we used subroutine RNMVN to generate samples of size 50 of the three variables, which were multivariate-normal with means equal to 0 and variances of 1. Those variables were next multiplied by 10, and 70 was added to each; the variables then had population means of 70, population standard deviations of 10, and population coefficients of variation equal to 0.143. Finally, we formed the various ratio variables and then computed and averaged the correlations between pairs of ratios across 10,000 iterations (see Tables 2 to 4).

Results

We calculated empirical correlations between $Y/C$ and $X/C$, ratios sharing a denominator, where $r_{XC} = r_{YC}$ and $r_{XY}$ were set equal to .0 through .9, in increments of .1 (see Table 2). Comparing the empirical values (see Table 2) with the estimates from Pearson’s (1897) formula (see Table 1), one can see that although the agreement is fairly close, the empirical values exceed the Pearson estimates by as much as 2 to 3 points in the second decimal place when ratios have a common denominator. Accuracy appears to be greatest at the right side of Table 1 but to decrease as one moves toward conditions at the left side of Table 1. Finally, as $r_{XC}$ and $r_{YC}$ increase, approaching the point where correlation conditions can no longer exist (blanks in the table), the empirical correlations turn negative, as predicted by Pearson’s equation, and behave most erratically, in that differences in magnitude between adjacent table entries are largest.

The correlations between $C/Y$ and $C/X$, ratios with a common numerator, are shown in Table 3. Comparing Table 3 with Table 1 (the estimates from Pearson’s
TABLE 2

Empirical Correlations Between X/C and Y/C, Ratios With Common Denominations, as Functions of $r_{XY}$ and $r_{XC} = r_{XY}$

<table>
<thead>
<tr>
<th>$r_{XC}$</th>
<th>.0</th>
<th>.1</th>
<th>.2</th>
<th>.3</th>
<th>.4</th>
<th>.5</th>
<th>.6</th>
<th>.7</th>
<th>.8</th>
<th>.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r_{YC}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.0</td>
<td>.52</td>
<td>.57</td>
<td>.61</td>
<td>.66</td>
<td>.71</td>
<td>.76</td>
<td>.81</td>
<td>.85</td>
<td>.90</td>
<td>.95</td>
</tr>
<tr>
<td>.1</td>
<td>.46</td>
<td>.52</td>
<td>.57</td>
<td>.62</td>
<td>.68</td>
<td>.73</td>
<td>.78</td>
<td>.84</td>
<td>.89</td>
<td>.95</td>
</tr>
<tr>
<td>.2</td>
<td>.40</td>
<td>.46</td>
<td>.51</td>
<td>.58</td>
<td>.63</td>
<td>.70</td>
<td>.76</td>
<td>.82</td>
<td>.88</td>
<td>.94</td>
</tr>
<tr>
<td>.3</td>
<td>.31</td>
<td>.38</td>
<td>.44</td>
<td>.51</td>
<td>.58</td>
<td>.65</td>
<td>.72</td>
<td>.79</td>
<td>.86</td>
<td>.93</td>
</tr>
<tr>
<td>.4</td>
<td>.19</td>
<td>.27</td>
<td>.35</td>
<td>.43</td>
<td>.51</td>
<td>.59</td>
<td>.67</td>
<td>.75</td>
<td>.84</td>
<td>.92</td>
</tr>
<tr>
<td>.5</td>
<td>.02</td>
<td>.12</td>
<td>.22</td>
<td>.31</td>
<td>.41</td>
<td>.51</td>
<td>.61</td>
<td>.70</td>
<td>.80</td>
<td>.90</td>
</tr>
<tr>
<td>.6</td>
<td>-.22</td>
<td>-.10</td>
<td>.02</td>
<td>.14</td>
<td>.26</td>
<td>.38</td>
<td>.51</td>
<td>.63</td>
<td>.75</td>
<td>.87</td>
</tr>
<tr>
<td>.7</td>
<td>-.64</td>
<td>-.47</td>
<td>-.31</td>
<td>-.15</td>
<td>.02</td>
<td>.18</td>
<td>.34</td>
<td>.50</td>
<td>.67</td>
<td>.83</td>
</tr>
<tr>
<td>.8</td>
<td>-.73</td>
<td>-.48</td>
<td>-.23</td>
<td>.01</td>
<td>.25</td>
<td>.50</td>
<td>.75</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.9</td>
<td></td>
<td>-.49</td>
<td>.01</td>
<td></td>
<td>.50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note.* Means of original variables equaled 70, standard deviations were 10, and $N = 50$. Empty cells represent impossible conditions.

1897 formula), one can see that whereas the agreement is again fairly close, the empirical values are less than the Pearson estimates by as much as 3 to 4 points in the second decimal place when ratios have a common numerator. The other characteristics noted in Table 2 are also true of Table 3.

The empirical correlations between $Y/C$ and $C/X$, ratios in which the denominator of the first is the numerator of the second, are shown in Table 4. Comparing Table 4 with Table 1, one can see that the signs have been reversed, as predicted by Equation 5. In terms of absolute values, the agreement again is fairly close, but the empirical values are less than the Pearson estimates by as much as 3 to 4 points in the second decimal place when the denominator of one ratio is the numerator of the other. In contrast to the previous empirical correlations, the most accurate Pearson predictions are seen in the upper right corner of Table 4. Again, the most dramatic changes in correlations occur when the conditions at the bottom and left of the table are approached, where it seems that correlations change rapidly, from negative to positive.

**Discussion**

The present findings show that Pearson's (1897) equation provides fairly accurate estimates of the values of correlations of ratios that we studied, at least under the following conditions: (a) the coefficients of variation are small; (b) the population coefficients of variation are equal; (c) the correlations of the unique variables with the common variable are equal; and (d) the original variables are sampled from a multivariate normal distribution. Having a small coefficient of
TABLE 3
Empirical Correlations Between C/Y and C/X, Ratios With Common Numerators, as Functions of \(r_{XY}\) and \(r_{XC} = r_{YC}\).

<table>
<thead>
<tr>
<th>(r_{YC})</th>
<th>.0</th>
<th>.1</th>
<th>.2</th>
<th>.3</th>
<th>.4</th>
<th>.5</th>
<th>.6</th>
<th>.7</th>
<th>.8</th>
<th>.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>.0</td>
<td>.46</td>
<td>.51</td>
<td>.56</td>
<td>.62</td>
<td>.67</td>
<td>.72</td>
<td>.78</td>
<td>.83</td>
<td>.89</td>
<td>.94</td>
</tr>
<tr>
<td>.1</td>
<td>.41</td>
<td>.47</td>
<td>.52</td>
<td>.58</td>
<td>.64</td>
<td>.69</td>
<td>.75</td>
<td>.82</td>
<td>.88</td>
<td>.94</td>
</tr>
<tr>
<td>.2</td>
<td>.34</td>
<td>.41</td>
<td>.47</td>
<td>.53</td>
<td>.60</td>
<td>.66</td>
<td>.73</td>
<td>.80</td>
<td>.86</td>
<td>.93</td>
</tr>
<tr>
<td>.3</td>
<td>.26</td>
<td>.33</td>
<td>.40</td>
<td>.47</td>
<td>.55</td>
<td>.62</td>
<td>.69</td>
<td>.77</td>
<td>.84</td>
<td>.92</td>
</tr>
<tr>
<td>.4</td>
<td>.14</td>
<td>.23</td>
<td>.31</td>
<td>.39</td>
<td>.47</td>
<td>.56</td>
<td>.64</td>
<td>.73</td>
<td>.82</td>
<td>.91</td>
</tr>
<tr>
<td>.5</td>
<td>-.01</td>
<td>.08</td>
<td>.18</td>
<td>.28</td>
<td>.38</td>
<td>.48</td>
<td>.58</td>
<td>.68</td>
<td>.79</td>
<td>.89</td>
</tr>
<tr>
<td>.6</td>
<td>-.25</td>
<td>-.13</td>
<td>-.01</td>
<td>.11</td>
<td>.23</td>
<td>.36</td>
<td>.48</td>
<td>.61</td>
<td>.74</td>
<td>.87</td>
</tr>
<tr>
<td>.7</td>
<td>-.65</td>
<td>-.49</td>
<td>-.33</td>
<td>-.17</td>
<td>-.01</td>
<td>.15</td>
<td>.32</td>
<td>.49</td>
<td>.65</td>
<td>.83</td>
</tr>
</tbody>
</table>
| .8         | -.73| -.49| -.25| -.01| .24 | .49 | .74 | .
| .9         |     |     |     |     |     |     |     | .49 | .00 | .49 |

*Note.* Means of original variables equaled 70, standard deviations were 10, and \(N = 50\). Empty cells represent impossible conditions.

Variation is important for several reasons. First, it justifies Pearson's dropping terms that involved powers higher than the square of the coefficients of variation. Second, if the mean is large relative to the standard deviation, the probability of having a denominator near 0 becomes vanishingly small, such that in a simulation, the erratic ratios involving very small denominators are avoided. Third, Chayes (1971) verified by simulation that the larger the coefficient of variation, the less accurate Pearson's equation becomes.

To summarize the inaccuracy of Pearson's (1897) equation, we can state that it slightly underestimates correlations among ratios with common denominators and slightly overestimates the correlations of ratios with common numerators or correlations in which the denominator of one ratio equals the numerator of the other. Overall, however, the inaccuracy is limited to a maximum of 4 points in the second decimal place, under the conditions investigated. Because the inaccuracy of Pearson's equation is relatively small, the values in Table 1 can be used safely to reach some general conclusions about correlations between ratio or index variables with common components when the coefficients of variation are small. When the coefficients of variation are large, greater inaccuracy will result.

Pearson's (1897) conclusion that if the original variables were independent, the correlation among ratio variables with a common denominator could quite easily be equal to .5 is not only true but generalizes to other cases such as common numerators (Table 1). Also, as seen along the major diagonal of Table 1, the .5 correlation generalizes to all cases in which the correlations among the original variables are equal. It is also clear that if one is interested in the relationship between X and
Y, the contamination by any common element in forming ratios will obscure that relationship (Table 1). There is no way of even estimating the X–Y relationship from correlated ratios without knowing a great deal about the original variables. The correct X–Y correlation will be replicated with ratios only if their correlation with the common variable is .5, and then only under conditions in which other factors such as coefficients of variation are the same (Table 1).

Furthermore, as Pearson (1897) pointed out, the correlation of ratios is affected not only by the correlations among the original variables but also by the coefficients of variation, which, in the interest of simplicity, we set equal for the present investigation. The lack of ability to reproduce correlations associated with components of ratios is particularly problematic when using large organizational databases, such as the PIMS (Profit Impact of Market Strategy; e.g., Buzzell & Gale, 1987; Hambrick, MacMillan, & Day, 1982). The PIMS database contains data for many ratio variables but does not contain meaningful absolute data for the components of those ratios, because of the proprietary information they would reveal. Reviewers of the PIMS database (Anderson & Paine, 1978; Lubatkin & Pitts, 1983) have acknowledged the multicollinearity of PIMS variables, but they have not specifically addressed the spuriousness of correlations among PIMS variables because of mathematical artifacts.

On the other hand, what if one were interested in the ratio variable per se and not its components, as suggested by Yule (1910), Fuguitt and Lieberson (1974), and Kasarda and Nolan (1979)? The correlation of a ratio with other variables might be meaningful and the interpretation may be relatively unambiguous as

### TABLE 4
Empirical Correlations Between Y/C and C/X, Denominator of First Ratio Common to Numerator of Second Ratio, as Functions of $r_{XY}$ and $r_{XC} = r_{YC}$

<table>
<thead>
<tr>
<th>$r_{XC} = r_{YC}$</th>
<th>.0</th>
<th>.1</th>
<th>.2</th>
<th>.3</th>
<th>.4</th>
<th>.5</th>
<th>.6</th>
<th>.7</th>
<th>.8</th>
<th>.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note.* Means of original variables equaled 70, standard deviations were 10, and $N = 50$. Empty cells represent impossible conditions.
long as the other variables are not ratios constructed with components common to the first ratio. In the latter case, however, any interpretation must be tempered by the knowledge that no good method exists for estimating what proportion of the correlation of ratios sharing elements is spurious and what part is not. Examining the extreme outcomes seen in Table 1 should cause the researcher to be conservative in such instances and not to base an important decision, either theoretical or applied, on such tenuous relationships.

Consider the position of Yule (1910), Fuguitt and Lieberson (1974), and Kasarda and Nolan (1979) from another perspective. A question about ratios is answered perfectly well by ratios. When those ratios have common elements, however, questions about the relationships among such ratios may be of little value. As Neyman (1952) argued earlier regarding stork rates and birth rates, the correlation may not be spurious, but its interpretation may be quite misleading.

Finally, an important point is that when the correlations with the common element become large, then the changes in the resulting correlations are dramatic, switching from positive to negative and showing the greatest rates of change. This result is important, because Cohen and Cohen (1983) said that one should not correlate ratio variables unless the correlation between numerator and denominator is high—but that is just the point at which such correlations are most unstable.

REFERENCES


Pearson, K. (1897). Mathematical contributions to the theory of evolution—On a form of spurious correlation which may arise when indices are used in the measurement of organs. Proceedings of the Royal Society of London, 60, 489–497.


Journal of Sociology, 80, 379–396.

Manuscript received December 2, 1996
Revision accepted for publication March 20, 1997