STATISTICAL ANALYSIS OF VARIABLES WITH SHARED TERMS

Havens C. Tipps U.S. Commission on Civil Rights

A potential problem with Pearsonian correlation coefficients has been recognized almost since the development of the coefficient itself. This paper describes the results of experimental tests of this statistical problem which has been causing concern and distress among a sizeable number of researchers. The problem arises most often when variables are created through making ratios, proportions and various kinds of indexes. It has been maintained that when certain types of pairs of numbers with common terms are correlated, statistics are produced which are spurious, misleading, or in other ways inappropriate. If the problem is as severe as some have asserted, a significant part of our empirical findings are suspect. This paper examines the nature of the problem and provides evidence that the problem is indeed worth stressing under certain limited conditions.

The pairs of variables below illustrate some of the types of combinations which have been described as problematic Pearsonian correlation coefficients:

1. Organizational size with administrative intensity:
(administrators plus non-administrators)
correlated with
(administrators/admin plus non-admin.) or:
(organization size)
correlated with
(administrators/organizational size)
2. Population size with suicide rate:
(population)
correlated with
(number of suicides/population) or:
(suicides plus non-suicides)
correlated with
(suicides/suicides plus non-suicides)
3. Percent black with population density:
(blacks/population)
correlated with
(population/area)
4. Population size with growth:
(population at time 1)
correlated with
(population at time 2) or:
(population at time 1)
correlated with
(population at time 1 plus change)
5. Proportion nonwhite with education index:
(nonwhites/population)
correlated with
(those with some college/population)
6. Social origin with social mobility:
(father's status)
correlated with
(son's status - father's status)

Needless to say, we are talking about types of variables and types of relationships which are very important to social scientists. In each of those pairs of variables numerical terms from one variable are duplicated in the other. In some situations this factor has been described as having the variables "definitionally dependent" on each other so that it is virtually assured that a non-zero correlation will be found between the two variables. The issues for this study are the extent to which the shared terms influence the value of the correlation coefficients and the meaning of this influence (if any) for actual researchers correlating variables with shared terms.

In the past few years interest in this problem has spread as with the publication of several important and often misinterpreted articles. The problem has also been extended to path coefficients by Karl Schuessler.[1] This project actually started with an article by Freeman and Kronenfeld (2) that bothered me and renewed a debate on the issue between myself and some colleagues. The article was titled "Problems of Definitional Dependency: The Case of Administrative Intensity." Freeman and Kronenfeld focused on a particular research problem in the study of organizations which involved the problematic condition of correlating variables with shared terms.

The interesting and theoretically important variable at the center of the Freeman and Kronenfeld analysis is the composition of organizations in terms of the administrative and productive components. Some organizations have a relatively large administrative component and some have a relatively small administrative component. A key theoretical issue involving this variable is the relationship between the organizations size and the administrative component. There are reasons to believe that as organizations get larger the proportion of the workers who hold administrative positions systematically changes. With the advancement of work in this area different patterns will probably be discovered for different types of organizations.

Freeman and Kronenfeld noted that many researchers have found a negative relationship indicating that the larger organizations have a smaller proportion of the work force holding administrative positions. The statistical concern of the authors is over the fact that the correlation of the size of an organization with the proportion of those workers who are in administrative positions involves the issue of "definitional dependency" as the authors call it. The correlation is between (x+y), with "x" being the administrative and "y" being the non-administrative workers, and either (x/y) or (x/x+y) depending on the method of measuring the administrative variable as a ratio or a proportion.

Through a mathematical demonstration they conclude that the reason why almost all data from research on this question exhibit an "exponentially shaped decreasing pattern," is that "this pattern is primarily the result of the coordinate transformation, and not because of any inherent relationship between x and y." (p. 112)

I had a difficult time accepting the implication of a correlation such as this being due to definitional dependence. There are just too many studies in the literature and especially in unpublished work where the correlations are zero and near zero under similar conditions to be easily convinced that there is a definitional

dependence when a unit's size is correlated with a proportion or rate based on the unit's size. The field of urban sociology seems filled with near zero correlations of that type. My two primary concerns were to provide experimental evidence concerning the extent and nature of the problem and to emphasize the limited applicability of the problem.

To create experimental conditions approximating the problematic situation data sets were randomly generated so variables could be created fitting the required specifications. The IBM Fortran Scientific Subroutine Package, using the GAUSS subroutine, was used to generate 100 data sets of 50 observations for each of three variables (X,Y, and Z). The generation process produces a normally distributed random number for each request with a given mean and standard deviation. In this case the mean used was 500 and the standard deviation was 125. The ratios, sums, and differences were then calculated and correlated for each of the 100 data sets.

Table 1 provides some of the statistics calculated from the 100 correlations produced for each type of relationship. Although 100 statistics (correlation coefficients in this case) is clearly not enough to generate a sampling distribution it seems to be a large enough sample of samples to reach some important conclusions about some of the problematic correlations. It is ample, it seems, to demonstrate that there is indeed a "problem" with some of the coefficients and the problem is numerically substantial. Equally important is that it seems that some of the coefficients which Freeman and Kronenfeld (1973) have described as spurious from the analysis of the formulas seems not to be "definitially dependent" in the samples generated for this study.

The table contains twenty-six types of relationships which involve shared terms. Freeman and Kronenfeld refer mainly to relationship number 4 and number 8.

The mean of both of these sets of correlations is close to zero and the number of significant correlations (12) and 9 respectively) is not especially far from the expected five at the .05 level. But most of the other sets of correlations are clearly far above the expected five significant correlations and seven have all 100 correlations being significant. In additional experiments performed by some colleagues at the University oc Cincinnati the means and the standard deviations of X,Y, and Z were modified in various ways for some of the types of relationships.3 Under conditions of varying means and standard deviations of the components, even relationships like number 4 and number 8 exhibit correlations that are vastly different from zero. Clearly there is room for concern over this issue.

Table 1

Statistics from the distribution of 100 correlation coefficients (r) with X,Y, and Z being random numbers*

	Form of the Relationship		Mean of r's	Std. Dev.	Range: Min.	Max.	Number of r's P .05
1.	X	Y	-.003	.13	-.33	.32	6
2.	X	$\frac{X}{Y}$	.623	.12	.13	.83	98
3.	$\frac{X}{Y}$	Y	-.694	.07	-.84	-.49	100
4.	X+Y	$\frac{X}{Y}$	-.040	.16	-.38	.32	12
5.	X	X+Y	.709	.06	.51	.84	100
6.	X	X-Y	.711	.07	.42	.84	100
7.	X	$\frac{X}{X+Y}$	.702	.08	.40	.86	100
8.	X+Y	$\frac{X}{X+Y}$	.022	.15	.36	.41	9
9.	$\frac{X}{X+Y}$	X-Y	.983	.01	.96	.99	100
10.	X+Y	X-Y	.017	.14	-.39	.30	4
11.	$\frac{X}{Y}$	X-Y	.935	.06	.51	.98	100
12.	X	$\frac{Y}{Z}$	-.018	.13	-.43	.27	2
13.	X+Y	$\frac{Y}{Z}$	.419	.11	.15	.66	87
14.	X+Y	$\frac{Z}{Y}$	-.484	.10	-.77	-.18	96
15.	$\frac{X}{Y}$	$\frac{Z}{Y}$	.569	.13	.20	.99	98
16.	$\frac{X}{Y}$	$\frac{Y}{Z}$	-.433	.10	-.63	-.13	92
17.	$\frac{X}{X+Y}$	$\frac{Z}{Y}$	.521	.11	.17	.77	98
18.	$\frac{X}{X+Y}$	$\frac{Y}{Z}$	-.433	.11	.69	-.10	91
19.	X-Y	$\frac{Z}{Y}$	.499	.09	.21	.74	98
20.	X-Y	$\frac{Y}{Z}$	-.444	.11	-.68	-.12	92
21.	X+Y	Y-Z	.486	.10	.20	.74	95
22.	X-Y	Y-Z	-.501	.10	-.71	-.19	98
23.	$\frac{X}{X+Y}$	Y-Z	-.489	.10	-.71	-.14	99
24.	$\frac{Z}{X+Y+Z}$	$\frac{X}{X+Y+Z}$	-.490	.10	-.73	-.19	96
25.	$\frac{Z}{X+Y}$	$\frac{X}{X+Y}$	-.007	.15	-.38	.38	7
26.	$\frac{Z}{X+Y}$	X+Y	-.578	.10	-.81	-.34	100

+Each of the 100 data sets has 50 observations for X,Y and Z, with each value selected as a normally distributed random number with a mean of 500 and a standard deviation of 125.

Practical Implications for Researchers

Given that there is ample evidence that correlations between composite variables with shared terms often exhibit high correlations when the components are not correlated, we must ask what the impact of this should be for researchers. What seems to need stressing most is the limited applicability of the problem for the theoretical and empirical issues which researchers typically face. The problem of "definitional dependence" is only "a problem" when the researcher is really interested in the variables X,Y and/or Z rather than the ratios, proportions or differences, which the researchers constructs. This is very rarely the case. Generally, it seems, when we use percentages, ratios or differences, we are actually interested in the percentages, ratios and differences as meaningful varioables in themselves. As Fuguitt and Lieberson reported:

> Discussions of this problem have centered on the purpose and assumptions of the analysis to be undertaken. Several writers have regretted Pearson's choice of the word spurious to refer to this phenomenon. A number have pointed out that there is nothing intrinsically spurious about the correlation, through interpretations may indeed be spurious, as in inferring from a ration correlation the size or direction of a component correlation or vice versa. A basic distinction here is whether the ratio or difference score is taken to be the basic variable describing the population under study or whether one's major interest really focuses on the component measures. If the former is the case, some authors argue that spurious correlation is not a problem (Yule, 1910;Kuh and Meyer, 1955; Rangarajan and Chatterjee, 1969). Logan (1972,p.67) gives as an example the association of speed (miles per hour) with gasoline consumption (miles per gallon). The basic interest here is in whether cars that go faster burn gasoline at a greater rate, and not in the associations between component variables _ miles traveled, time elapsed, or gasoline consumed. Just as this example utilizes a common numerator in the two ratios (miles), sociologists may likewise try to claim an inherent interest in ratios with a common denominator; for example, the correlation between per capita energy consumption of nations and their per capita gross national product.4

Freeman and Kronenfeld, among many other statistical analysists, fail to realized or appreciate the importance of the composite variables. Administrative intensity is an important variable itself (as a ratio or a proportion) and as such the relationship between the components of the composite variables are irrelevant when the variable is correlated with organizational size. In fact, I can not think of a single major research finding that should be discredited because of the definitional dependence issue. The isolated case where researchers construct composite measures, correlate them with variables sharing some terms, and then try to infer to the components of the composite variables is clearly inappropriate. In such cases the construction of composite variables is usually to control for an additional variable. Although the construction of composite measures is not a very appropriate method for multivariate controls, the standard methods of control still apply and should be used. Thus standardization, partial correlations and multiple regression can offer solutions to the problem of "definitional dependency" if the primary interest is in the component variables.

In the normal situation rates, proportions, ratios and differences are used as variables with theoretical importance. Correlations between proportions, rates, ratios, or differences and other variables which may or may not share terms are subject to the same sources of misinterpretation as other correlations. We must always be aware of possible effects of other variables, sampling error, etc. But is has not been established that special problems exist for this type of correlation. In fact, since Yule clarified the issue in 1910 by advising that one simply state in advance whether one was interested in the components of the ratios or the ratios themselves,5 there has not been a dispute over the statistical issue. The experimental evidence provided in this paper further documents the existence of cause for concern. The key issue remaining is the validity of studying variables which are in the form of ratios, proportions and differences under the types of conditions discussed above. This question of the validity of the variables is a theoretical question that can only be answered on the grounds of the particular substantive areas. In the more common areas of investigation such variables are definitely established as legitimate and often they are the most important variables. Often they seem to be the purest form of measurement we have since they are based on counting, although they are not an artifact of the size of the populations subject to the count.

References

1. Karl Schuessler, "Ratio Variables and Path Models," in Arthur S. Goldberger and O.D. Duncan (eds), Structural Equation Models in Social Science, New York: Seminar Press, 1973.
2. J.H. Freeman and J.E. Kronenfeld, "Problems of Definitional Dependency: The Case of Administrative Intensity," Social Forces, Vol. 52, Sept. 1973.
3. W.E. Feinberg, J. Trotta and H.C. Tipps, "Correlations of Ratio Variables Involving two components: Investigation of an Alternative Significance Test," unpublished paper, 1976.
4. Glenn V. Fuguitt and Stanley Lieberson, "Correlation of Ratios or Difference Scores Having Common Terms," in H.L. Costner (ed.), Sociological Methodology 1973-1974, San Francisco Jossey-Bass, 1974. Yule, G.U., "On the interpretation of correlations between indicies or ratios," Journal of the Royal Statisical Society 73 (June):644-647, 1910. Kuh, E. and Meyer, J.R., "Correlation and regression estimates when data are ratios," Econometrica 23:400-416, 1955. Rangarajan, C. and Chatterjee, S., "A note on comparison between correlation coefficients of original and transformed variables," American Statistician 23 (October): 28-28, 1969.
5. G.U. Yule: see above.