BIASED CORRELATION COEFFICIENTS AND CHI-SQUARE STATISTICS WITH REGIONAL DATA

William R. Latham*

1. Introduction

This paper demonstrates that standard computational procedures used for calculating coefficients of correlation and chi-square statistics can produce biased results unless observations containing zero values for all variables are omitted. Furthermore it is shown that such observations are likely to be present when data such as those in input-output studies and the various censuses are used in a disaggregated form.

An increasing ability on the part of investigators to utilize large data arrays that can be manipulated by faster, larger digital computers has been accompanied by an increasing availability of highly-disaggregated data on a large variety of magnitudes. For example, the 1963 Input-Output Study divides the economy into about 367 sectors(the 1958 study utilized only 86) and the 1963 Census of Manufactures provides data on plant locations at the county level for industries disaggregated to nonzero observation for only a fraction of the possible values, selection of pairs of industries or counties from them is likely to yield many pairs of zero observations. The fact that such arrays are often available in computer-readable form, such as magnetic tapes, may permit investigators to manipulate the data sets without actually examining the nature of the individual observations and without realizing the extent or consequences of zero observations.[1]

In the following sections the theoretical effects of zero observations on coefficients of correlation and on chi-square statistics for contingency tables are stated and illustrative examples using data from the Census of Manufactures are presented.

2. Correlation Coefficients

The sample correlation coefficient, r, is often calculated as

$$\frac{n\Sigma xy - \Sigma x\Sigma y}{\{[n\Sigma x^2 - (\Sigma x)^2]\,[n\Sigma y^2 - (\Sigma y)^2]\}^{\frac{1}{2}}}$$

where the summations are over all values of the two variables, x and y, and n is the sample size. It is clear that the value of r varies with n provided all other magnitudes remain unchanges. As n increases without bound the value of r^2 approaches a limit of $\frac{(\Sigma xy)^2}{\Sigma x^2\Sigma y^2}$, and thus the value of r will approach the (positive) square root of this limit from below.[2]

In addition to increasing the value of r, larger values of n will directly influence hypothesis tests regarding r. Provided the joint distribution of x and y is bivariate normal, the hypothesis that the population correlation coefficient, ρ, is equal to zero may be tested by constructing a t statistic as $\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ and comparing with tabulated values of t with n-2 degrees of freedom for any desired level of significance. This value is seen to vary directly with n as does r. Test of the hypothesis that ρ equals some value other than zero are usually performed using Fisher's z - transformation and constructing the standard normal deviate, $Z = \frac{z - \zeta H}{\sigma_z}$, which is compared with tabulated values for any desired level of significance. Because

$$z = \ln\frac{1+r}{1-r},$$

*Instructor of economics, University of Delaware.

$\zeta_H = \ln \frac{1+\rho_H}{1-\rho_H}$ (where ρ_H is the hypothesized value of ρ), and

$$\sigma_z = \frac{1}{\sqrt{n-3}} ,$$

Z also will vary directly with n.[3]

One use of the correlation coefficient is as a measure of the strength of the geographic association between two industries where the magnitudes correlated are measures of employment for each industry in a number of geographic regions as has been done by Florence [3], McCarty, et al. [5], Richter [6], and Streit [7]. Recent investigations by the author have also employed the correlation coefficient as a measure of association, but much less aggregated data have been used than in the previous studies. The level of aggregation has been reduced both spatially and industrially as described below and in greater detail in [4].

In the following example a set of regions is initially defined as 20 of the Office of Business Economics' 173 economic regions of the United States. The industries used here are Office, Computing and Accounting Machines (1957 SIC code 357) and Motor Vehicles and Parts(1957 SIC code 371). The employment estimates for each of these industries in the 20 regions are shown in Columns (1) and (2) of Table 1.[4] The coefficient of correlation for the two distributions is .1330 and the value of t is .5692.

The data were then disaggregated spatially by dividing each of the OBE regions into a segment consisting of a Standard Metropolitan Statistical Area within the OBE region and a second segment consisting of the area within the OBE region but outside of the SMSA. The employment estimates for the two industries in the 40 new regions are shown in columns (1) and (2) of Table 2. Using the same calculating procedure as above, the correlation coefficient between the distributions of employment estimates in the two industries using the spatially disaggregated data is .0817 and the value of t is .5054.[5] Note, however, that the disaggregation resulted in five observations for which the estimated employment in both industries is 0. Further disaggregation down to the county level and below would produce still more such paired-zero observations. The effect of paired-zero observations in this case is to increase the value of r and t in a spurious fashion. The fact that neither industry has located in some geographic areas should not lead one to conclude that the two industries tend to associate themselves with each other spatially. Such a conclusion would permit the inflation of the coefficient of correlation to the limit described above simply by finer and finer geographic subdivisions contributing pairs of zero observations which increase n but do not affect any other magnitudes. Omitting the five paired zero observations yields a value for r of .0661, which is 19% lower than the value with the paired zeroes included, and reduces the value of t to .3804.

The original data were next disaggregated from 3-digit SIC code industries to 4-digit SIC code industries. Columns (3) and (4) of Table 1 show the employment estimates for Computing and Related machines (SIC code 3571) and Truck Trailers (SIC code 3715). The correlation coefficient and value of t for these two distributions are, respectively, -.0942 and -.4016 with the 8 paired zero observations included and -.1963 and -.6330 without them (a reduction in the value of r of 108%).

Thus, the theoretical effect on r and t of increasing n without changing any other magnitudes has been found to occur when Census of Manufactures data are disaggregated in either their spatial or industrial dimensions. In some situations paired zeroes might be quite meaningful and, if so, then increases in r because of such observation would not be considered spurious. For example, correlations between measures of incomes and education levels should be influenced by pairs of zero observations. In the example above, however, the increase in the value of r can be considered spurious and paired zeroes should be omitted from the calculations of the correlation coefficients in such cases.[6]

The following are two examples of studies in which bias of the type identified above may have influenced the results. Others could have been

TABLE 1. ESTIMATED NUMBER OF EMPLOYEES IN FOUR S.I.C. INDUSTRIES IN 20 O.B.E. REGIONS[a]

Region					
OBE Number	Name	SIC 357	SIC 371	SIC 3571	SIC 3715
		(1)	(2)	(3)	(4)
009	Buffalo, N.Y.	383	16236	346	0
012	Binghamton, N.Y.	5689	3906	3138	0
017	Baltimore, Md.	174	8167	131	12
026	Charlotte, N.C.	415	826	69	0
029	Columbia, S.C.	408	37	377	31
032	*Augusta, Ga.	6	37	0	0
043	*Columbus, Ga.	6	37	0	0
044	*Atlanta, Ga.	100	11816	0	0
048	*Chattanooga, Tenn.	0	24	0	0
058	Champaign, Ill.	6	383	6	0
064	Columbus, Ohio	236	2171	0	31
079	Davenport, Ia.	155	342	0	69
080	Cedar Rapids, Ia.	0	100	0	69
106	*Des Moines, Ia.	6	380	0	0
122	*Amarillo, Tex.	6	93	0	0
129	*Austin, Tex.	0	167	0	0
140	*Beaumont, Tex.	0	12	0	0
146	Albuquerque, N.M.	31	118	31	100
148	Denver, Colo.	67	1088	55	377
159	Boise City, Id.	0	61	0	6

*Indicates a region in which neither SIC 3571 nor SIC 3715 is located.

[a] Estimates are based on data contained in the *Census of Manufactures, Location of Manufacturing Plants by County, Industry and Employment Size*. (1963)

TABLE 2. ESTIMATED NUMBER OF EMPLOYEES IN FOUR SIC INDUSTRIES IN 20 SMSA's AND 20 SURROUNDING AREAS[a]

Region	SIC 357 (1)	SIC 371 (2)	SIC 3571 (3)	SIC 3715 (4)
# Buffalo, N.Y. SMSA	37	15717	0	0
Surrounding Area	346	519	346	0
Binghamton, N.Y. SMSA	2792	6	2786	0
Surrounding Area	2897	3900	352	0
Baltimore, Md. SMSA	168	5523	131	6
Surrounding Area	6	2644	0	6
# Charlotte, N.C. SMSA	0	93	0	0
Surrounding Area	415	733	69	0
# Columbia, S.C. SMSA	31	6	0	0
Surrounding Area	377	31	377	31
# Augusta, Ga. SMSA	31	0	0	0
* Surrounding Area	0	0	0	0
# Columbus, Ga. SMSA	6	37	0	0
#* Surrounding Area	0	0	0	0
# Atlanta, Ga. SMSA	100	10928	0	0
# Surrounding Area	0	888	0	0
# Chattanooga, Tenn. SMSA	0	12	0	0
# Surrounding Area	0	12	0	0
Champaign, Ill. SMSA	6	37	6	0
# Surrounding Area	0	346	0	0
# Columbus, Ohio SMSA	236	2065	0	0
Surrounding Area	0	106	0	31
Davenport, Ia. SMSA	155	273	0	69
# Surrounding Area	0	69	0	0
Cedar Rapids, Ia. SMSA	0	100	0	69
#* Surrounding Area	0	0	0	0
# Des Moine, Ia. SMSA	6	143	0	0
# Surrounding Area	0	237	0	0
# Amarillo, Tex. SMSA	6	87	0	0
# Surrounding Area	0	6	0	0
# Austin, Tex. SMSA	0	167	0	0
#* Surrounding Area	0	0	0	0
Beaumont, Tex. SMSA	0	12	0	0
#* Surrounding Area	0	0	0	0
Albuquerque, N.M. SMSA	31	112	31	100
Surrounding Area	0	6	0	0
Denver, Colo. SMSA	67	1051	55	377
# Surrounding Area	0	37	0	0
# Boise City, Id. SMSA	0	55	0	0
Surrounding Area	0	6	0	0

*Indicates a region in which neither SIC 357 nor SIC 371 is located.
#Indicates a region in which neither SIC 3571 nor SIC 3715 is located.
[a] Estimates are based on data contained in the Census of Manufactures, Location of Manufacturing Plants by County, Industry and Employment Size. (1963)

cited but these serve as indicators of one class of study likely to be subject to this bias.

(1) Richter [6] used the 1958 Census of Manufactures to construct a matrix of employment estimates for 2-and 3-digit SIC industries in a group of large SMSA's. Because this matrix contained a significant number of zero observations, measures of geographic association for some pairs of industries, calculated using this data, may have been biased. Utilization of more disaggregated data would certainly increase the potential for bias.

(2) Czamanski [2] computes four correlation coefficients for every pair of industries in an expanded form of the 1958 Input-Output Study using the interindustry transactions of his data. Even at his level of aggregation (89 industries) many paired zero observations were included in his calculations, possibly biasing his final results. Any attempt to apply his methods to a more disaggregated set of data would certainly encounter significant bias unless paired zeroes were discarded.

3. Chi-square Statistics

Rather than computing a correlation coefficient as in the above example, one might compare the spatial distributions of the two industries simply by counting the number of geographic regions in which are found (a)both industries, (b) neither industry, (c) SIC 357 but not SIC 371, and (d) 371 but not 357 and performing a χ^2 test to see whether the sample provides evidence indicating that the two industries are distributed differently. Such a procedure has been used by Czamanski [1]. χ^2 is computed $\Sigma[\frac{(f_o - f_c)^2}{f_c}]$ for large samples and as $\Sigma[\frac{(|f_o - |f_c - \frac{1}{2})^2}{f_c}]$ for small samples where f_o and f_c represent, respectively, the observed frequencies and the expected frequencies if both industries are distributed as their marginal frequency distributions.

The variation of $\chi^2$with sample size is more complex than the variation of the correlation coefficient described in the preceding section. This is because the observed frequencies in one cell of the contingency table also vary with n,even when disaggregation results only in the addition of paired zero observations. χ^2 may be calculated for the two-by-two case as $\frac{(ad-bc)^2n}{(a+b)(c+d)(a+c)(b+d)}$ where a, b, c and d refer to the cells of the contingency table and n = a+b+c+d. If disaggregation results only in paired-zero observations being added to d, then the value in cell d at any time will be equal to its original value, d', plus the number of observations can derive n'-d' = a+b+c and thus d = n-(a+b+c). Substituting this into the expression above for χ^2permits it to be expressed as a function of only the constants a, b, and c and the single variable n.

$$\chi^2 = \frac{\{a[n-(a+b+c)]-bc\}^2 n}{(a+b)(a+c)\{c+[n-(a+b+c)]\}\{b+[n-(a+b+c)]\}}$$

Attempting to find the limit as n becomes large and applying l'Hospital's rule gives:

$$\lim_{n\to\infty} = \lim_{n\to\infty} \frac{3a^2n-2a^2(a+b+c)-2abc}{(a+b)\ (a+c)}$$

In this expression it can be seen that χ^2 does not have a finite limit in cases such as the one considered above and that its value can be made arbitrarily large by adding paired-zero observations.

Using the data for SIC industries 357 and 371 in columns (1) and (2)of Table 1, the observed and expected frequencies are:

Observed

	$f_{371}>0$	$f_{371}=0$	Total f
$f_{357}>0$	14	1	15
$f_{357}=0$	5	0	5
Total f	19	1	20

Expected

	$f_{371}>0$	$f_{371}=0$	Total f
$f_{357}>0$	14.25	.75	15
$f_{357}=0$	4.75	.25	5
Total f	19	1	20

χ^2 calculated using Yates's continuity correction for small samples is .351.

Disaggregating spatially into SMSA and non-SMSA regions the frequencies from columns (1) and (2) of Table 2 are:

Observed

	$f_{371}>0$	$f_{371}=0$	Total f
$f_{357}>0$	18	1	19
$f_{357}=0$	16	5	21
Total f	34	6	40

Expected

	$f_{371}>0$	$f_{371}=0$	Total f
$f_{357}>0$	16.15	2.85	19
$f_{357}=0$	17.85	3.15	21
Total f	34	6	40

χ^2 calculated as above is now 1.433 but there are 5 paired-zero observations. χ^2 without the paired zero observations is reduced to .007. Because the number of degress of freedom in each case is identical, the inclusion of paired-zero observations serves to raise the value of χ^2 with no offsetting change in critical values for testing χ^2. Additional disaggregation would raise it sufficiently to permit rejection of an hypothesis (for some levels of significance) that could not have been rejected without these observations.

Disaggregating the spatially disaggregated industries above yields the following frequencies from columns (3) and (4) of Table 2:

Observed

	$f_{3715}>0$	$f_{3715}=0$	Total f
$f_{3571}>0$	5	3	8
$f_{3571}=0$	7	25	32
Total f	12	28	40

Expected

	$f_{3715}>0$	$f_{3715}=0$	Total f
$f_{3571}>0$	2.4	5.6	8
$f_{3571}=0$	9.6	22.4	32
Total f	12	28	40

χ^2 is calculated as 1.875. Omitting the paired-zero observations, the value of χ^2 increases to 2.093 contrary to the preceding results. The discrepancy occurs because χ^2 is not a monotonically increasing function of n and, for some distributions, the predicted increase in χ^2 occurs only for sufficiently large increases in n. In this case another 10 zero observations would produce a value of $\chi^2 = 3.042$.

4. Conclusion

The preceding sections have shown how straight-forward application of computational formulas for the coefficient of correlation and χ^2 for contingency tables to data containing pairs of zero observations can lead to biased estimates of these measures. The occurrence of pairs of zero observations was shown to be likely when spatial industrial data are disaggregated spatially or industrially. It was suggested that computational algorithms should provide for the elimination of paired zero observations when such observations are not meaningful. Investigators employing other measures and other data set should consider the applicability of generalizations of the results presented.

FOOTNOTES

[1] For example, the 1963 *Census of Manufactures, Location of Manufacturing Plants by County, Industry and Employment Size* is available on magnetic tape from the Bureau of the Census. A description of the tape states that only 100,769 records are on the tape. An appreciation of the fact that the potential number of records, considering both the number of counties in the U. S. (more than 3100) and the number of manufacturing industries (more than 290), is nearly a million should alert the investigator to the existence of zero values for approximately 90% of the potential records. A description of the *Input-Output Structure of the U.S. Economy: 1963*, which is available on magnetic tape, does not alert one to the pervasiveness of zero entries in it, however, because every potential record is present, even when its value is zero.

[2]

$$\lim_{n\to\infty} r = \lim_{n\to\infty} \frac{[\ \Sigma xy - \frac{\Sigma x\Sigma y}{n}\ ]^2}{[\ \Sigma x^2 - \frac{(\Sigma x)^2}{n}\ ]\ [\ \Sigma y^2 - \frac{(\Sigma y)^2}{n}\ ]} = \frac{(\Sigma xy)^2}{\Sigma x^2 \Sigma y^2}$$

which must be ≥ 0 because all quantities are squared, and must also be ≤ 1 by the Schwarz inequality.

[3] One cannot make comparisons of t and z values with and without the zero observations without also allowing r to vary with n.

[4] The employment estimates used here are based on data from the 1963 *Census of Manufactures, Location of Manufacturing Plants by County, Industry and Employment Size* [8]. The method of estimation is described in Richter [6].

[5] The variation in the coefficient of correlation between less and more highly disaggregated industries is not of interest in this paper. It has been discussed in [3] and [5].

[6] In the example used the addition of paired zero observations has obviously caused the joint distributions of the variables to depart from normality and this invalidates any tests of r. In some applications the number of paired zero observations added may be small enough, however, not to significantly alter the joint distribution, especially if the distributions of nonzero observations have means close to zero.

REFERENCES

1. Czamanski, S., "A Model of Urban Growth," *Papers, Regional Science Association*, 12 (1965), 177-200.

2. _____., "Some Empirical Evidence of the Strengths of Linkages Between Groups of Related Industries in Urban-Regional Complexes," *Papers, Regional Science Association*, 27 (1971), 137-150.

3. Florence, P. Sargent, *Investment, Location and Size of Plant*, Cambridge: University Press, 1948.

4. Latham, W. R., *The Impact of Agglomerative Economies on Industrial Location*. Unpublished Ph.D. disseration: University of Illinois, 1973.

5. McCarty, H. H., J. C. Hook and G. S. Knos, *The Measurement of Association in Industrial Geography*. Iowa City: State University of Iowa Press, 1956.

6. Richter, C. E., "The Impact of Industrial Linkages on Geographic Association," Journal of Regional Science, 9 (1969), 19-28.

7. Streit, M., "Spatial Associations and Economic Linkages Between Industries," Journal of Regional Science, 9 (1969), 177-188.

8. U.S., Bureau of the Census, 1963 Census of Manufactures, Location of Manufacturing Plants by County, Industry and Employment Size.