Estimation of Product Moment Correlation Coefficients Through the Use of the Ratio of Contingency Coefficient to the Maximal Contingency Coefficient[1]

LeRoy A. Stone and Marlo A. Skurdal
University of North Dakota

Over sixty years ago, Pearson (1904) in his fundamental paper on the theory of contingency clearly indicated some of the difficulties of comparing the coefficients of relationship, correlation and contingency. Pearson did show that, with certain reservations concerning fineness of subdivision in classification, the coefficient of contingency is essentially identical with the product moment correlation coefficient as deduced from a normal correlation surface.

In a practical sense, contingency coefficients are not directly comparable unless derived from the same size contingency tables and they are not directly comparable to product moment correlation coefficients because of a limitation regarding upper limits and because of a measurement restriction problem. The upper limits for contingency coefficients are a function of the number of categories. The upper limit for a 2 X 2 table is .707; for a 3 X 3 table, .816; for a 4 X 4 table, .866; for a k X k table, $\sqrt{(k - 1)/k}$.

Over four decades ago, Kelley (1924) presented corrections which may be applied to make contingency coefficients estimates of product moment correlations.[2] The corrections are most tedious and time consuming to make. One correction is for number of categories. The other correction requires the assumptions that the underlying traits are continuous and normal in distribution. McNemar (1962, p. 201) suggests that if the assumptions of normally distributed continuous variables are tenable and if one is justified in reducing a more than four-cell contingency table to a 2 X 2 table, one can instead determine the value of tetrachoric r.

The purpose of the present paper is to suggest another and more simplified approach to use when one desires to compare a contingency coefficient to a product moment correlation coefficient. This approach is not dissimilar

---

[1]Based on a paper read at the Psychometric Society meeting, September 2, 1966, New York.

[2]The need for correcting contingency coefficients has also been shown by Harris and Treloar (1927) and by Harris and Chi Tu (1929).

to what some investigators, namely those using coefficients of correlation in factor analysis, have done to make the phi coefficient supposedly comparable to r by computing the ratio, ϕ / ϕ_{max}, when the ϕ_{max} value has been determined by an equation developed by either Ferguson (1941) or Guilford (1965) involving the marginal means, p_i and p_j

We will attempt to empirically demonstrate that the ratio, contingency coefficient/maximal contingency coefficient (C/C_{max}), is also directly comparable to the product moment correlation coefficient. The bivariate data used in this investigation were obtained from 45 statistics textbooks. Product moment correlation coefficients were computed using 74 sets of bivariate data (Ns ranged from 20 to 6835 and rs ranged from 0.00 to 1.00). When data were cast into 2 X 2 contingency tables, an attempt was always made so as to have dichotomies as near to .50--.50 proportions as possible. However, the achievement of such .50--.50 proportions was seldom possible. The dichotomization was also done so that no contingency table cell would have an expected value of less than five.

Inequality of means in correlated, dichotomized variables has an effect upon the size of a contingency coefficient computed from such bivariate data. The data from 38 of the 74 bivariate data sets were recast into 2 X 2 tables so that the marginal proportions, p_i, q_i, p_j, and q_j would vary widely. However, adherence to the restriction that expected values for cells must not be less than five was followed. Some of these bivariate data sets were cast into as many as 11 different 2 X 2 contingency tables. With each data set, the C/C_{max} ratio which best approximated the computed correlation coefficient was selected. These selected C/C_{max} ratios were then statistically compared to the product moment correlation coefficients. The product moment correlation coefficient between the selected C/C_{max} ratio values and the correlation coefficients was high (r = .934, N = 74, p <.001). As should be expected there was a very high linear relationship between values from these two relationship indices. The intraclass correlation coefficient between these two sets of relationship estimation values was only slightly lower (R = .924, p <.001) and represented an estimate of agreement between the two sets of relationship estimations when they had been classified in 20 groups in which the interval size was .05, e.g., .00 - .04, .05 - .09, .10 - .14, etc.

The test for the difference between the product moment correlation coefficient (mean r = .592, S.D. = .257) and the C/C_{max} ratio (mean C/C_{max} = .558, S. D. = .247) was significant (C.R.= 3.97, p <.001). It appeared that the C/C_{max} ratio model provided a conservative estimate of the correlation coefficient.

Inspection of all of the computed C/C_{max} ratios, from 2 X 2 tables, (see Table 1) showed that the ratios which corresponded most closely to the product moment correlation coefficients were not always the ones which were associated with fourfold tables having dichotomies nearer to .50 - .50 proportions. However, we're lead to believe that the C/C_{max} ratios which best approximated the product moment correlation coefficients generally were from the fourfold tables where $p_i \cong p_j \cong .50$.

Twenty of the 74 bivariate data sets were also cast into 3 X 3 tables. Contingency coefficients and C/C_{max} ratios were computed and were compared to the product moment correlation coefficients. With 10 of these bivariate data sets, the C/C_{max} ratios when compared to the product moment correlation coefficients were less adequate than when the C/C_{max} ratios were computed from 2 X 2 tables. Two of the bivariate data sets were also cast into 4 X 4 tables, contingency coefficients and C/C_{max} ratios were computed, and were compared to the product moment correlation coefficients. One of these two C/C_{max} ratios represented a more accurate estimate of the correlation coefficient than did the C/C_{max} ratios computed from 2 X 2 and 3 X 3 tables. From this limited evidence it cannot be said that the C/C_{max} ratio computed from 3 X 3 or 4 X 4 tables provide more accurate estimates of the product moment correlation coefficients than those C/C_{max} ratios computed from 2 X 2 tables.

The implications of these conclusions for the use of the C/C_{max} ratio are not clear. However, it would appear, based on this empirical demonstration, that the C/C_{max} ratio may be used as a "quick and dirty" estimate of the relationship measure provided by the product moment correlation model. No mathematical justification is offered for this contingency coefficient ratio, C/C_{max}. However, it has been pointed out by Guilford (1965) that he has not seen any mathematical justification regarding the ratio, ϕ/ϕ_{max}, as an index of relationship and it has received wide use as a statistical device.

Table 1

Relationship Statistics, r and C/C_{max}, Computed with Differing Marginal Values (Arranged According to N Size)

N	r	C/C_{max}	p_i	q_i	p_j	q_j	N	r	C/C_{max}	p_i	q_i	p_j	q_j
20	.60	.69	.40	.60	.40	.60	56	.72	.71	.59	.41	.52	.48
		.51	.40	.60	.65	.35			.63	.36	.64	.38	.62
32	.53	.48	.69	.31	.56	.44			.53	.29	.71	.23	.77
		.44	.47	.53	.56	.44	64	.00	.02	.69	.31	.69	.31
		.43	.47	.53	.50	.50			.05	.69	.31	.94	.06
		.69	.69	.31	.69	.31	64	.25	.24	.94	.06	.31	.69
		.76	.78	.22	.78	.22			.23	.69	.31	.31	.69
		.20	.12	.88	.12	.88			.18	.69	.31	.69	.31
35	.43	.36	.37	.63	.37	.63			.10	.94	.06	.69	.31
		.35	.49	.51	.51	.49	64	.50	.46	.69	.31	.69	.31
		.52	.60	.40	.66	.34			.42	.31	.69	.69	.31
		.30	.37	.63	.40	.60			.33	.94	.06	.69	.31
40	.68	.56	.42	.58	.30	.70			.28	.94	.06	.94	.06
		.54	.68	.32	.72	.28			.24	.06	.94	.69	.31
49	.97	.93	.83	.17	.80	.20			.09	.06	.94	.94	.06
		.88	.55	.45	.61	.39	64	.75	.69	.69	.31	.69	.31
		.84	.41	.59	.41	.59			.60	.94	.06	.94	.06
		.78	.41	.59	.61	.39			.59	.31	.69	.69	.31
		.44	.69	.31	.94	.06			.43	.34	.66	.69	.31
		.17	.06	.94	.69	.31			.41	.34	.66	.48	.52
64	1.00	1.00	.69	.31	.69	.31	92	.67	.64	.50	.50	.55	.45
		1.00	.94	.06	.94	.06			.63	.28	.72	.29	.71
65	.76	.53	.37	.63	.35	.65			.79	.38	.62	.38	.62
		.47	.37	.63	.65	.35	99	.24	.12	.54	.46	.56	.44
69	.93	.89	.36	.64	.36	.64			.04	.34	.66	.29	.71
		.88	.49	.51	.54	.46			.03	.19	.81	.14	.86
		.78	.23	.77	.20	.80			.46	.54	.46	.43	.57
		.76	.09	.91	.09	.91	100	.38	.38	.50	.50	.54	.54

Table 1

(continued)

N	r	C/C_{max}	p_i	q_i	p_j	q_j
72	.75	.75	.44	.56	.53	.47
		.76	.44	.56	.46	.54
75	.07	.08	.40	.60	.67	.33
		.11	.61	.39	.67	.33
		.03	.61	.39	.33	.67
		.00	.40	.60	.33	.67
85	.54	.52	.69	.31	.69	.31
		.51	.69	.31	.48	.52
		.60	.92	.08	.85	.15
106	.78	.79	.27	.73	.27	.73
		.79	.40	.60	.27	.73
		.85	.40	.60	.44	.56
		.69	.27	.73	.44	.56
		.67	.65	.35	.44	.56
		.56	.93	.07	.91	.09
		.56	.65	.35	.58	.42
		.54	.40	.60	.58	.42
110	.13	.09	.62	.38	.56	.44
		.20	.45	.55	.56	.44
		.06	.26	.74	.32	.68
		.03	.36	.64	.41	.59
113	.37	.36	.73	.27	.71	.29
		.44	.62	.38	.60	.40
		.46	.50	.50	.45	.55
120	.60	.52	.65	.35	.66	.34
		.50	.38	.62	.66	.34
		.49	.38	.62	.35	.65
		.47	.65	.35	.35	.65
193	.79	.86	.51	.49	.38	.62
		.70	.51	.49	.76	.24
		.33	.17	.83	.12	.88
		.22	.33	.67	.32	.68
100	.82	.83	.50	.50	.56	.44
		.80	.81	.19	.76	.24
		.75	.72	.28	.63	.37
		.71	.50	.50	.45	.55
100	1.00	1.00	.46	.54	.46	.54
		1.00	.64	.36	.64	.36
		1.00	.79	.21	.79	.21
140	.59	.61	.16	.84	.62	.38
		.56	.33	.67	.46	.54
		.69	.33	.67	.62	.38
		.44	.16	.84	.46	.54
141	.82	.76	.59	.41	.50	.50
		.73	.75	.25	.60	.40
		.69	.39	.61	.42	.58
		.65	.39	.61	.50	.50
149	.68	.65	.42	.58	.51	.49
		.58	.58	.42	.42	.58
		.57	.71	.29	.29	.71
		.49	.81	.19	.20	.80
188	.03	.06	.68	.32	.61	.39
		.07	.90	.10	.71	.29
		.10	.44	.56	.52	.48
192	.08	.07	.43	.57	.64	.36
		.04	.43	.57	.36	.64
192	.48	.45	.39	.61	.58	.42
		.54	.48	.52	.41	.59
281	.55	.56	.31	.69	.30	.70
		.57	.41	.59	.40	.60

Table 1

(continued)

N	r	C/C_{max}	p_i	q_i	p_j	q_j
		.70	.23	.77	.38	.62
		.89	.51	.49	.61	.39
		.90	.69	.31	.61	.39
		.91	.69	.31	.76	.24
		.60	.23	.77	.19	.81
		.60	.51	.49	.19	.81
		.57	.23	.77	.61	.39
		.46	.51	.49	.89	.11
		.40	.51	.49	.09	.91
193	.86	.82	.45	.55	.48	.52
		.82	.43	.57	.47	.53
202	.57	.57	.70	.30	.27	.73
		.58	.44	.56	.46	.54
		.62	.58	.42	.46	.54
225	.80	.77	.79	.21	.20	.80
		.76	.56	.44	.40	.60
		.85	.45	.55	.50	.50
		.52	.21	.79	.20	.80
		.52	.50	.50	.50	.50
310	.69	.59	.47	.53	.45	.55
		.56	.71	.29	.67	.33
		.53	.86	.14	.77	.23

References

Ferguson, G. A. "The Factorial Interpretation of Test Difficulty." Psychometrika, VI (1941), 323 - 333.

Guilford, J. P. "The Minimal Phi Coefficient and the Maximal Phi." Educational and Psychological Measurement, XXV (1965), 3 - 8.

Harris, J. A. and Chi Tu "A Second Category of Limitations in the Applicability of the Contingency Coefficient." Journal of the American Statistical Association, XXIV (1929), 367 - 375.

Harris, J. A. and Treloar, A. E. "On a Limitation in the Applicability of the Contingency Coefficient." Journal of the American Statistical Association, XXII (1927), 460-472.

Kelley, T. L. Statistical Method. New York: Macmillan, 1924.

McNemar, Q. Psychological Statistics (3d ed.). New York: Wiley, 1962.

Pearson, K. "On the Theory of Contingency and Its Relation to Association and Normal Correlation." Drapers' Company Research Memoirs, Biometric Series, London, I (1904), 1 - 35.