ON COMPARING CONTINGENCY TABLES *

By

H. FAIRFIELD SMITH **

During the early years of biometry in the first part of this century much attention was paid to inventing measures of association between two variates. When observations were qualitative, or when they could be arranged in ordered classes but with no obvious scale of measurement for the distances between classes of either one or both variates, these measures of association were called contingency coefficients. Although no longer used by many statistical workers they are still valued by some, particularly those working in sociology. Almost the only criterion used in deriving a contingency coefficient is that no association should be represented by zero, complete association by unity. They are presented as having a quantitative scale between these extremes, but no rational interpretation for the scale of any one of them seems ever to have been propounded.

A few years ago it was represented to me by an eminent mathematical statistician that discovery of the sampling distribution of biserial-η would be of great service to sociology. Let us consider the utility of the sampling distribution of any contingency coefficient in two parts: (1) when there is no association, or under the "null" hypothesis that the population value of the coefficient is zero, (2) when the population coefficient is not zero.

(1) The sampling distribution of any contingency coefficient assuming the population value to be zero, the one value having a clearly prescribed meaning, provides a test of significance for the null hypothesis that there is no association between the variates. However, when classes are not ordered and when there is no definite alternative hypothesis (and so far as is known to me no definite alternative has ever been

* Paper read at the Fifth Annual Conference of the Phil. Stat. Assoc., Manila, June 22, 1957.

** Institute of Statistics, N. Carolina State College, and FAO Agricultural Statistician, Statistical Center, University of the Philippines.

proferred when a contingency coefficient is used), the test of significance is adequately covered by the well known X^2-test for contingency tables. Excluding permutation tests which are unmanageable — except for small numbers in 2 x 2 tables for which Fisher's "exact test" has been tabulated — no other continuous approximation seems to have any chance of being more efficient against indefinite alternatives. In particular this applies to biserial-η because that coefficient is independent of ordering of the multiple classification, despite that K. Pearson (inventor of the coefficient) and some modern texts have wrongly implied that prior ordering is recognized. Application of the normal approximation to the binomial dispersion of each column of the table reduces the test for null biserial-η to the ordinary X^2-test for a 2 x n table. When classes of a contingency table can be ordered adequate tests for the null hypothesis have been worked out by Yates (1948) and by Williams (1952).

(2) Turning then to situations where association exists, debate with the propounder of the problem continued somewhat as follows:

A: What do you understand by "strength of association in a contingency table"?

B (unable to answer directly): It seems valuable to be able to test whether two populations have, or two samples indicate, the same strength of association.

Since Stuart (1953) had proposed yet another contingency coeficient (t_c) with particular reference to just that question, combined with attention to ordering of the categories, his coefficient as applied to data on grades of distance vision of males and females (See Table 1) was taken as a specific example. We could not agree on what t_c measures in these tables. Furthermore it transpired that for intuitively equal "degrees of association" t_c is not invariant under change of marginal totals, not even for the simplest case of all observations on a diagonal, when it is not necessarily unity.

A: So the expectation of t_c seems to be a ghost! What can be the use to know that ghosts in my lord's and lady's

chambers each wore a sash with the symbol .6 if we do not know how the sash or its decoration may reflect the more earthly bodies from which the ghosts have been supposed to emanate?

B: But recollect that any comparison of two such tables is equivalent to second order interactions of a three-way contingency table, and these have been notoriously difficult to interpret.

A: To substitute another mystery for the one is no help.

B: How else would you compare these tables?

This paper is an attempt to meet the challenge of the last question. It rests on the postulate (or dictum, or idea?) that a requisite for a statistical comparison is first of all to know, and to be able to define, what is to be compared. A statistician's initial maneuver must therefore be to ask his 'client': "What feature of your data is of special interest? Do you have in mind any particular way in which eye sight of two sexes may differ? What kinds of action do you consider taking and how may action be modified by possible inferences from these data?"

Before launching into contingency tables at large let us have a brief look at a much debated 2 x 2 table, say for example the frequencies of diseased and healthy individuals in two samples, one of inoculated and one of non-inoculated observees. This type of table is so simple that for description, as distinct from a test of significance, reduction to fewer statistics than the four observed frequencies is scarcely required. The only relevant simplification is to eliminate marginal frequencies which depend only on manner of sampling, size of the experiment, or are otherwise irrelevant to the population: that is to compute the pair of proportions (say p_1 and p_2) attacked by disease in each group. Apart from ancillary statistics to determine precision of the estimates, these two statistics state all of the evidence on the one and only thing of interest. They state the results succinctly in a straightforward and understandable manner. Why then seek other artificial coefficients with no interpretable meaning? For example: for one such

table Kendall (1943, 13.3-7) finds the values of three proposed contingency coefficients to be as divergent as .19, .56 and .89; and so far as I know no rational interpretation for any one of these seemingly quantitative values has ever been proposed. One of them is not even invariant for variation of sample sizes with the same p_1 and p_2. By estimating p_1 (probability of becoming diseased if inoculated) $= 3/279 \pm (p_1q_1/279)^{\frac{1}{2}}$, $p_2 = 66/473 \pm (p_2q_2/473)^{\frac{1}{2}}$, we know what exactly we are estimating and how the information can be used. Why go further merely to obscure the essential information?

Stuart's example, Table 1, presents unusually good specimens of orderly contingency tables with large numbers of observations suitable for treatment by large sample approximations. But the optometrist who made these observations is not now available for questioning. I have no clue to what he had in mind and what questions he might ask of such data. The only 'practical' problem which occurs to me is to consider whether an optical qualification for some job might be based on testing only one eye. The answer is obvious without statistical analysis beyond mere inspection of the frequency arrays: "There is considerable correlation between eye pairs, but there is also enough disagreement so that for correct assessment of individuals both eyes should be tested unless the tests are very onerous and a moderate proportion of errors can be allowed." Such conclusion from inspection may or may not suffice to determine behavior relative to some practical problems. It does not illustrate the kind of tests to be discussed here; therefore let us carry the questioning further. Whether or not the questions asked may be academic relative to this example, analogous ones may be of practical consequence with other statistically similar data.

Inspection again suffices to indicate that distribution of distance vision is distinctly different in the two sexes. Carrying the foregoing question further we might ask: what is the probability that, one eye having been tested, the other may be in the same category? The obvious answer, and the maximum likelihood estimate, is that that probability, say p_o, is estimated by the ratio of frequencies in the leading dia-

gonal to the total number of observations. For males it is $p_0(m) = .6875$, for females $p_0(f) = .7083$. The difference, $.0208 \pm .00964$ is on the border line of significance. The complements, with same standard errors, estimate the probability for incorrect diagnosis of both eyes if only one be tested.

We may proceed to further detail by asking for the relative frequencies, p_i, that a pair of eyes may differ by i grades, $i = 0, 1, 2$ or 3. Summing appropriate diagonals leads to Table 2. Treating this in the ordinary way as a 2 x 4 contingency table yields $X^2 = 23.37$ with 3 degrees of freedom, significant at $P = .00004$ against the null hypothesis that probabilities of each category are equal for both sexes. Inspection of p_i indicates that p_1 are equal for both sexes, the difference already observed for p_0 is balanced by the probability of males having eyes different by two and three grades being greater than for females. By concentrating the trend in a single degree of freedom (e.g. by Yates' test, 1948) greater significance might be indicated.

Proceeding to more general questions one might enquire, for example, whether or not each eye has similar distribution both marginally and conditionally on the other. Clearly the frequency distribution for left eyes cannot be the same when the right is, say, grade 4 or grade 2. So what we mean by similar conditional distributions is that the distributions are similar for the left eye given that the right is grade 1, and for the right given that the left is grade 1, etc. If these similarities be true then the frequency distributions of tables 1 should be symmetrical about their leading diagonals. The question can be tested by matching the frequencies in symmetrically placed cells in a 2 x 6 contingency table. These tables, one for each sex, are shown in Table 3. If the hypothesis of symmetry be true the two rows of frequencies should be equal. Each table therefore has 6 degrees of freedom which can be partitioned into one degree of freedom for the contrast of row totals, and 5 degrees of freedom for "interaction," that is for testing that the conditional distributions may be equal.

Postulating equal expectations for frequencies in each row the X^2 analyses are shown in Table 4 (cf. appendix). From it we can conclude that for males the distributions for right and left eyes appear similar, or, more carefully, that if differences exist they are smaller than can be detected with the observed sample. For females the right eye is stronger than the left more often than conversely. (Given that the two eyes differ, as measurable by the grades here reported, the probability of the right eye being stronger is estimated by $p = 1171/2181 = .5369$.) So far as present evidence goes the probability for the right eye to be stronger than the left seems fairly stable for all combinations. For some purposes it may be worth noting that the contrast 41 v. 14 contributes most to the interaction X^2: when two female eyes differ widely the estimated relative frequency for the right to be stronger increases to .65. Being an observation selected **post facto** the significance of this effect cannot be tested on the same data, it may be noted for checking on another sample.

The foregoing has not yet established that the sexes differ in frequency of right eye being stronger than left. The condition that the male ratio is not significantly different from a hypothetical .5 is not by itself evidence of difference from the estimated female ratio. The variance of the difference may be estimated using the null hypothesis $p = .5$ for both sexes (cf. appendix); this gives $\frac{1}{4}\left(\frac{1}{1013} + \frac{1}{2181}\right) = .0003614$. The difference is then $.5369 - .4847 = .0522 \pm .0190$; and a real difference between the sexes is indicated.

Had discrepancies of the upper and lower triangles of Table 1 been more complex, other partitions of the 6 degrees of freedom might be considered. One might seek to define 3 degrees of freedom to indicate discrepancies between marginal distributions of left and right eyes, and 3 for other types of contrasts. There is no particularly obvious way to define these in general. Given some hypothesis of interest the suitable tests may be indicated by respective maximum likelihood estimates as illustrated by Mather (1943, Chap. XII).

It happens that Stuart's data can be neatly described by supposing that they arise from bivariate normal distributions of quantitative variates. Analysis on these lines may be demonstrated on another occasion. It is a device which has often been used to describe contingency table distributions, for example K. Pearson's tetrachoric correlation. However, I do not consider this device relevant to true **contingency** tables which characteristically state frequencies in **qualitative** classes with no measurable underlying scale. When truly applicable it implies that observations have been grouped into grossly large class intervals, possibly because detailed data have been lost, perhaps because to record observations as greater or less than two or three fixed points may give a great saving of labour relative to individual measurements. However that may be, to fit quantitative distributions to such data is a sort of salvage operation. Even if only a few coarse categories may be required for practical application, at the research level, when variation and co-relations are to be determined, if only to evaluate optimum categories for future practice, it will usually be wise to record actual measurements.

The foregoing examples may suffice to illustrate the principle enunciated at the beginning of this paper. If two (or more) contingency tables are to be compared the first step is to ask: compared with respect to what explicitly defined characteristic? It may then usually be possible to devise a suitable test which will certainly be more meaningful than to compare values of an arbitrarily defined coefficient whose quantitative values have no tangible interpretation.

SUMMARY

Numerical values of contingency coefficients, as these have been defined in past literature, seem to be meaningless and without tangible interpretation. Comparison of two meaningless quantities seems almost, if not quite, equally abstruse. It is useless to say that one contingency table as compared to another exhibits a greater "strength of association" between two variates if strength of association be measured by an arbitrary coefficient whose scale has no apparent meaning.

This paper attempts to answer a challenge: how then should two contingency tables be compared? The answer proposed is that one should first ask: what feature of the observed bivariate frequency distribution, expressed as a contingency table, is of interest, meaningful, and relevant to future action? When that question has been answered, and only then, a statistical test may be devised to decide whether or not samples derived from two different populations indicate similar or different conditions for the characteristic explicitly defined. No single general method can be laid down for comparing any two contingency tables. Each case must be individually considered relative to meaningful characteristics of the observations. Examples are given for illustration.

TABLE 1

CLASSIFICATION OF UNAIDED DISTANCE VISION OF 3242 MEN AND 7477 WOMEN AGED 30-39: LOWEST GRADE = 1, HIGHEST GRADE = 4, DATA OF ROYAL ORDNANCE FACTORIES 1943-6; QUOTED BY STUART (1953)

Grade of right eye	*Grade of left eye*				Total
	4	3	2	1	
Men					
4	821	112	85	35	1053
3	116	494	145	27	782
2	72	151	583	87	893
1	43	34	106	331	514
Total	1052	791	919	480	3242
Women					
4	1520	266	124	66	1976
3	234	1512	432	78	2256
2	117	362	1772	205	2456
1	36	82	179	492	789
Total	1907	2222	2507	841	7477

TABLE 2

FREQUENCIES OF RIGHT AND LEFT EYES DIFFERING BY i GRADES.

	i				Total
	0	1	2	3	
Freq.					
male	2229	717	218	78	3242
female	5296	1678	401	102	7477
Totals	7525	2395	619	180	10719
p_i					
male	.6875	.2212	.0672	.0241	1
female	.7083	.2244	.0536	.0136	1
Difference	—.0208	—.0032	.0136	.0105	
St. error of diff.	.00964	.00876	.00499	.00287	

TABLE 3

COMPARISON OF SYMMETRIC CELLS OF TABLE 1.

Cells	43	32	21	42	31	41	*Total*
Males							
upper . .	112	145	87	85	27	35	491
lower . .	116	151	106	72	34	43	522
Females							
upper . .	226	432	205	124	78	66	1171
lower . .	234	362	179	117	82	36	1010

TABLE 4

ANALYSES OF X^2 FOR TABLE 3

	d.f.	*Males* X^2	*Females* X^2	*Females* P
Contrast of row totals	1	.949	11.885	<.001
Remainder (interaction)	5	3.814	7.221	.2
Total (contrasts in each of 6 pairs)	6	4.762	19.106	<.01

APPENDIX: NOTE ON THE X^2 ANALYSIS OF TABLE 4

Strictly a X^2 analysis is no longer additive after the null hypothesis has to be discarded at any one of two or more possible test points. Suppose a 2 x k contingency table with its array totals be algebraically described as follows:

$a_1 \ldots a_i \ldots a_k$	A
$b_1 \ldots b_i \ldots b_k$	B
$n_1 \ldots n_i \ldots n_k$	N

With p, q, being the hypothetical probabilities for an observation to fall in row 1 or 2, the analysis of X^2 is as follows:

Row totals $$\frac{1}{p_o q_o} \cdot \frac{(A - p_o N)^2}{N.} \qquad (1)$$

Interaction $$\frac{1}{p_1 q_1} \sum_{i=1}^{k} \frac{(a_i - n_i A/N)^2}{n_i} \qquad (2)$$

Total $$\frac{1}{p_o q_o} \sum_{i=1}^{k} \frac{(a_i - n_i p_o)^2}{n_i}$$

For any given p

$$\frac{(A - pN)^2}{N} + \Sigma \frac{(a_i - n_i A/N)^2}{n_i} \equiv \Sigma \frac{(a_i - n_i p)^2}{n_i}$$

is an algebraic identity. The variance of each deviation is pq along with N or n_i. If the a priori hypothetical p_o can be used to estimate variances in every row of the analysis the sum of X^2 for each row is equal to the total. But if row (1) indicates that p_o is not the true value, then it should not be used to evaluate the variances in row (2). Usually an estimate has to be obtained from the data, and we use $p_1 = A/N$. Row (2) is then the usual X^2 for interaction of a contingency table conditional on the marginal totals; and rows (1) and (2) no longer add to the total X^2 computed for the a priori hypothetical probability. Of course, if p_1 be substituted throughout

additivity would be regained, but the analysis would be trivial: row (1) would then be identically zero corresponding to the fact that one degree of freedom for discrepancies has been lost by fitting the hypothesis to the data at this point.

The technicality has been neglected in table 4 because when p_o is close to .5 moderate deviations therefrom have almost negligible effect on pq; for example when, as for females, $p_1 = .537$, then $p_1q_1 = .2486$ which is trivially different from $p_oq_o = .25$. The matter needs more careful attention when p or q is less than .25. Occasionally a markedly false conclusion may be inferred if the modification be then ignored. (cf. Fisher, 1925, secs. 22 and 57; Mather, 1943, chap XI.)

The X^2 test for contrast of row totals, one degree of freedom, is equivalent to the normal approximation to a simple binomial test. For example, for males, on the hypothesis $p = \frac{1}{2}$ the expectation for each row is 506.5, deviation $\pm$ 15.5 with variance $npq = 1014(.5)^2 = 253$; $t^2 = 15.5^2/253 = .949$.

LITERATURE CITED

Fisher, R. A. (1925) *Statistical methods for research workers*, Oliver & Boyd, Edinburgh.

Kendall, M. G. (1943) *Advanced theory of statistics*, vol. 1, Griffin, London.

Mather, K. (1943) *Statistical analysis in biology*, Methuen, London.

Stuart, A. (1953) Biom. 49:105-110

Williams, E. J. (1952) Biom. 39:274-289

Yates, F. (1948) Biom. 35:176-