Eleven Ways to Look at the Chi-Squared Coefficient for Contingency Tables

Boris Mirkin*
School of Computer Science and Information Systems,
Birkbeck College, London WC1E 7HX, UK

Abstract

This article has been written in recognition of the 100th aniversary of introduction of the concept of association between categorical variables by Yule and Pearson. The most popular among the contingency coefficients, Pearson's chi-squared, estimates the bias of a cross-classification from the statistical independence. Also, it measures association *per se* between the row and column variables. The purpose of this article is to present a collection of eleven definitions for the chi-squared coefficient related to either of these goals. One of the quoted definitions of the chi-squared coefficient seems especially appealing as an association measure: the averaged relative Quetelet index of category-to-category associations.

Key Words: summary association measure, category-to-category association, aggregation of contingency tables, decomposition of the variance.

Contents

*Boris Mirkin is Professor of Computer Science, School of Computer Science and Information Systems, Birkbeck College, London WC1E 7HX, UK (e-mail: mirkin@dcs.bbk.ac.uk). This material was prepared in DIMACS, Rutgers University, NJ. The author thanks the Office of Naval Research for its support under grant number N00014-96-1-0208 to Rutgers University. He is grateful to H.-H. Bock, L. Wilkinson and anonymous referees for many correcting and clarifying remarks.

1 INTRODUCTION

This paper reviews a number of interpretations of the celebrated Pearson's chi-squared coefficient for contingency tables in a manner following Rodgers and Nicewander (1988), in which a collection of interpretations of Pearson's product-moment correlation coefficient have been presented on its 100th aniversary. A similar aniversary occurs for chi-squared: it was 100 years ago when two remarkable statisticians, K. Pearson and G.U. Yule, published their papers, Pearson (1900a) and Yule (1900), devoted to evaluating of the degree of association between two qualitative variables according to their cross-classification.

K. Pearson viewed qualitative categories as intervals of an underlying continuous variate and thus wanted to introduce a measure akin to his product-moment correlation coefficient between quantitative variables. G. Yule claimed that such categories as "vaccinated/non-vaccinated" cannot be presented this way and deserve to be treated without references to the quantitative case.

The chi-squared coefficient was proposed by K. Pearson as an application of his ideas for testing observed frequencies against expected values with the chi-square distribution that he invented for this purpose (Pearson 1900b). In this context, the coefficient was to be used only for testing independence in a bivariate distribution, not for evaluating association. However, the coefficient came to be used as an association measure, notably by A.A. Tschuprow and H. Cramér.

A number of alternative summary contingency measures were suggested. A unique attempt to summarize these developments was undertaken by L. Goodman and W.H. Kruskal (in Goodman and Kruskal 1954 and several follow-up papers printed afterwards as a monograph, Goodman and Kruskal 1979). These authors insisted that any useful measure must be meaningful in an operational situation, of which perhaps the most general is the situation requiring prediction of row categories when column categories are known or vice versa. They "*have been unable to find any convincing published defence of* χ^2*-like statistics as measures of association*" (Goodman and Kruskal 1954, p. 740) and have proposed instead a number of prediction-based indices such as the coefficient of reduction of proportional prediction error, the Wallis-Goodman-Kruskal's τ (see formula (3.13)).

The current author's encounters with the chi-squared coefficient in the field of approximation clustering (Mirkin 1985, 1998) made him believe that this coefficient, actually, does not much differ from the index of proportional reduction error and has the same predictive powers, as reflected in sections 3.6, 3.7 and 5.4 below.

In our view, interpretations of the chi-squared coefficient fall in five categories: (a) related to the concept of independence in a bivariate distribution, (b) evaluating summary category-to-category associations, (c) defining the data scatter in bilinear aggregation models for contingency data as such, (d) formulated in geometric terms of the underlying indicator variables, and (e) pertaining to specifics of fourfold contingency tables. Accordingly, there are five sections in the remainder, each devoted to one of these subjects.

Specifically, in section 2, two traditional interpretations of the chi-squared as a device for testing independence or homogeneity are presented. In section 3, the coefficient is interpreted as a summary association measure - in terms of either of two category-to-category association indexes proposed by A. Quetelet (Quetelet 1832). Especially simple is the interpretation showing that the Quetelet coefficients averaged according to the contingency data lead to either the chi-squared coefficient, X^2, or the denominator of τ, Y^2, respectively. These also suggest representation of the data table in such a way that the pattern of category-to-category associations is displayed.

Section 4 describes two aggregation models for contingency tables that involve the chi-squared coefficient as a measure of the data variance in a corresponding Pythagorean decomposition of the data scatter into the explained and unexplained components. Section 5 deals with the sets of indicator variables corresponding to row or column categories. There are three interpretations here: two of them refer to the similarity between row- and column-related linear subspaces, and the third one measures that part of the scatter of the row indicators that is explained by the column indicators in a linear model relating them. It appears that, depending on the normalization scale used for the row indicator variables, either X^2 or Y^2 comes as the part of the total variance explained by the column categories and, respectively, either a normalized version of X^2 or τ comes as the proportion of the total variance explained by the column categories.

It should be probably noted that the contents of sections 4 and 5 require from the reader some knowledge in the multivariate data analysis: such topics as projection operators, latent values and singular-value-decomposition are touched there. Section 6 concludes with interpretation of the chi-squared coefficient (and τ) in the four-fold table as the determination coefficient between corresponding binary variables. This implies a most straightforward interpretation of the coefficient in terms of the prediction error.

2 A STATISTIC FOR INDEPENDENCE

2.1 Chi-Squared as an Independence Test Statistic

A contingency, or cross-classification, data table usually corresponds to two variables with disjoint categories corresponding, respectively, to rows and columns of the table; the (i,j)-th entry of such a table counts the number (or proportion) of the cases when row category i and column category j co-occur. More formally, it is assumed that the counted observations constitute a set O of N entities which is cross-classified into a partition $I = \{i_1, ..., i_m\}$ and a partition $J = \{j_1, ..., j_n\}$ according to categories of each of the two variables. One variable has m categories $i \in I$ and the other n categories $j \in J$. To correspond to the partitions, categories of each of the variables must be disjunctive and defined for all N entities in O. The table of the co-occurrence numbers, N_{ij}, or cardinalities of pair-wise intersections $i \cap j$ $(i \in I;\ j \in J)$ is referred to as the contingency, or cross-classification, table. The cardinalities of classes $i \in I$ and $j \in J$ usually are called marginals and denoted by N_{i+} and N_{+j} since, in given assumptions, they are sums of co-occurrence entries, N_{ij}, in rows, i, and columns, j, respectively.

A data table from Goodman (1986, p.249) (referring to original data from Srole, Langner, Michael, Oppler and Rennie 1962) illustrates the further discussion (see Table 1).

The proportions, $p_{ij} = N_{ij}/N$, $p_{i+} = N_{i+}/N$, and $p_{+j} = N_{+j}/N$ are also frequently used as contingency table entries. The matrix $P = (p_{ij})$ of proportions found by relating the entries in Table 1 to their total, $N = 1660$, is presented in Table 2 (along with the marginal row and column).

When the entity set, O, is a random sample from a population, the contingency table $P = (p_{ij})$ is considered as a sample-based estimate to the underlying bivariate distribution involving the two

Table 1: Cross-classification of 1660 subjects according to their mental health and their parents' socioeconomic status.

Mental health status	Parents' socioeconomic status A	B	C	D	E	F	Total
Well	64	57	57	72	36	21	307
Mild symptom formation	94	94	105	141	97	71	602
Moderate symptom formation	58	54	65	77	54	54	362
Impaired	46	40	60	94	78	71	389
Total	262	245	287	384	265	217	1660

categorical variables. To distinguish the underlying distribution parameters from the observed data entries, corresponding Greek symbols can be used: π_{ij}, π_{i+}, and π_{+j} as the underlying population characteristics for observed p_{ij}, p_{i+}, and p_{+j}, respectively.

When the underlying distribution satisfies the hypothesis of statistical independence

$$\pi_{ij} = \pi_{i+}\pi_{+j},$$

the observed values may deviate from this because of sampling bias. To evaluate how great the bias is, Pearson (1900a) proposed using his chi-squared measure of deviation which is equal, in this case, to

$$\chi^2 = N \sum_{i\in I} \sum_{j\in J} \frac{(p_{ij} - \pi_{i+}\pi_{+j})^2}{\pi_{i+}\pi_{+j}} \tag{2.1}$$

The distribution of this statistic converges to the chi-square distribution (with $mn - 1$ degrees of freedom) when sampling is done according to the standard multinomial model.

When the underlying marginal values, π_{i+} and π_{+j}, are unknown, as typically happens, they can be replaced with the empirically-driven maximum-likelihood estimates,

$$p_{i+} = \sum_{j\in J} p_{ij}, \qquad p_{+j} = \sum_{i\in I} p_{ij} \tag{2.2}$$

so that χ^2 in (2.1) becomes NX^2 where

$$X^2 = \sum_{i\in I} \sum_{j\in J} \frac{(p_{ij} - p_{i+}p_{+j})^2}{p_{i+}p_{+j}} \tag{2.3}$$

Table 2: The relative frequency table, $P = (p_{ij})$, per cent, according to the cross-classification in Table 1, along with the marginals separated by solid lines.

3.9	3.4	3.4	4.3	2.2	1.3	18.5
5.7	5.7	6.3	8.5	5.8	4.3	36.3
3.5	3.3	3.9	4.6	3.3	3.3	21.8
2.8	2.4	3.6	5.7	4.7	4.3	23.4
15.8	14.8	17.2	23.1	16.0	13.1	100

In the standard conditions of multinomial sampling, the distribution of NX^2 asymptotically converges to the χ^2 distribution, with $(m-1)(n-1)$ degrees of freedom because of equations (2.2). This convergence allows for testing in large samples the hypothesis of independence. For the data in Table 2, $X^2 = 0.0277$, which seems not that much, to signal any deviation from the independence. However, $NX^2 = 45.985$, so that, with $3 \times 5 = 15$ degrees of freedom and $N = 1660$ observations, the hypothesis of statistical independence is rejected with the level of significance 0.00005.

Such examples of the obvious discordance between a very low value of X^2 and very high level of confidence in rejecting the independence hypothesis may have contributed to the claim made by many statisticians that the magnitude of X^2 gives no indication on the degree of association. This is not quite correct, as will be seen in the next sections.

For the sake of convenience, in the remainder only X^2, not NX^2, will be considered as the Pearson's chi-squared coefficient. (Sometimes, this X^2 is referred to as the phi-squared.)

2.2 Chi-Squared as a Homogeneity Testing Index

Another aspect of (2.3) is related to the case when the contingency table represents what is called product-multinomial model, organized according to a different sampling scheme. It is assumed, in this model, that the numbers p_{+j} are not a characteristic of the sample O, but have been prespecified by the sample design. In the example of Table 1, this may happen when the numbers of representatives from different parents' statuses (column's marginals) have been a priori determined in collecting the data.

In such a situation, the table is just a set of columns, and the researcher may question how homogeneous the distributions in these columns are. The model for testing homogeneity is that there is an underlying probability distribution of the rows, π_i, and the data are homogeneous when $p(i/j) = \pi_i$ for any $i \in I$ and $j \in J$, that is, when

$$\frac{p_{ij}}{p_{+j}} = \pi_i. \tag{2.4}$$

Within each column $j \in J$, the Pearson's chi-squared measure of goodness-of-fit is

$$\sum_{i \in I} \frac{(p_{ij}/p_{+j} - \pi_i)^2}{\pi_i}$$

or, with the maximum likelihood estimates p_{i+} substituted for π_i,

$$\sum_{i \in I} \frac{(p_{ij}/p_{+j} - p_{i+})^2}{p_{i+}}$$

Now, if we take the weighted sum of these as an aggregate measure of goodness-of-fit,

$$H^2 = \sum_{j \in J} p_{+j} \sum_{i \in I} \frac{(p_{ij}/p_{+j} - p_{i+})^2}{p_{i+}} \tag{2.5}$$

it is easy to see that H^2 equals X^2, which reveals it to be a degree-of-homogeneity measure.

3 A SUMMARY ASSOCIATION INDEX

3.1 Extremum Association Patterns

Table 3: A typical contingency pattern for maximum X^2.

Category	j_1	j_2	j_3	j_4	j_5	j_6	Total
i_1	185	0	0	0	0	0	185
i_2	0	0	118	100	0	0	218
i_3	0	363	0	0	0	0	363
i_4	0	0	0	0	10	224	234
Total	185	363	118	100	10	224	1000

Before interpreting the chi-squared coefficient as an association measure, let us take a look at what patterns of contingency correspond to its minimum and maximum values as the "no-association" and "complete association" patterns, respectively.

The formula (2.3) shows that the minimum value of X^2 is zero. This occurs if and only if the variables are statistically independent, which fits into the traditional view of the "no-association" pattern.

To analyze the maximum of X^2, let us use another, also well-known, expression for X^2 that can be derived from (2.3) with a little arithmetic:

$$X^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{p_{ij}^2}{p_{i+}p_{+j}} - 1 \tag{3.1}$$

Assuming that the number of rows m is not greater than that of columns, n, let us consider a part of the sum in (3.1) corresponding to $i \in I$:

$$X_i = \sum_{j=1}^{n} \frac{p_{ij}^2}{p_{i+}p_{+j}} = \sum_{j=1}^{n} p(j/i)p(i/j)$$

where $p(j/i) = p_{ij}/p_{i+}$ and $p(i/j) = p_{ij}/p_{+j}$ are conditional probabilities. Since $\sum_{j=1}^{n} p(j/i) = 1$, X_i is the average $p(i/j)$ according to distribution $p(j/i)$ over J. Therefore, $X_i \leq \max_{j \in J} p(i/j) \leq 1$. To reach its maximum value 1, $\max_{j \in J} p(i/j) = p_{ij^*}/p_{j^*}$ must have $p_{ij^*} = p_{j^*}$ so that only one nonzero entry, (i, j^*), is present in the entire column j^*. Thus, the maximum of X^2 is $\min(m,n) - 1$, and it is reached if and only if the data table has the following pattern: (1) all row marginals are non-zero, and (2) in every column, one and only one entry is non-zero.

Such a contingency pattern is shown in Table 3 and it displays a pattern of logical association between row and column categories: i_1 is equivalent to j_1, i_2 follows from j_3 and j_4, i_3 is equivalent to j_2, and i_4 follows from j_5 and j_6. It is assumed, as usually, that a category i follows from a category j if observations falling in j are also falling in i. The categories are equivalent if they follow from each other.

Thus, the implicit assumption underlying the chi-squared coefficient is that the variables I and J are "completely" associated when and only when there is a mapping of "logical" association from J to I such that any category $i \in I$ is equivalent to a category $j \in J$ (one-to-one matching) or follows from a subset of categories $j \in J$ (many-to-one matching). This also applies when the cardinality of J

is smaller than the cardinality of I by exchanging the row and column related terms. The meaning of "complete association" caught by X^2 does conform to an intuitive meaning of the concept of logical association, though may seem over-restrictive in some contexts.

3.2 Normalized Versions

To make X^2 lie between 0 and 1, and take the extreme values under the independence and complete association, respectively, it may be scaled following H. Cramér:

$$C^2 = \frac{X^2}{\min(m-1, n-1)}. \tag{3.2}$$

Somewhat fuzzier scaling is referred to as Thschuprow's

$$T^2 = \frac{X^2}{\sqrt{(m-1)(n-1)}} \tag{3.3}$$

Rather surprisingly, both of these heuristic formulations can be supported within certain geometrical frameworks (as shown below in sections 5.2 and 5.3, respectively).

3.3 Category-to-Category Quetelet Indexes

The formulation (3.1) suggests that Pearson's chi-squared coefficient can be looked at as a summary characteristic of "local" category-to-category associations.

The association between categories $j \in J$ and $i \in I$ is reflected in the conditional probability $p(i/j) = p_{ij}/p_{+j}$ as related to the average rate p_{i+} of i. In particular, A. Quetelet (1832) recommended using either the (absolute) probability change

$$w_{ij} = p(i/j) - p_{i+} = (p_{ij} - p_{i+}p_{+j})/p_{+j}, \tag{3.4}$$

or the relative change

$$q_{ij} = w_{ij}/p_{i+} = p_{ij}/(p_{i+}p_{+j}) - 1 = (p_{ij} - p_{i+}p_{+j})/(p_{i+}p_{+j}) \tag{3.5}$$

as a measure of the "degree of influence" of j towards i (see also Yule 1900). These functions will be referred to as the absolute and relative Quetelet indexes, respectively.

Table 4: Matrix, $Q = (q_{ij})$, of the relative Quetelet coefficients for the data in Table 2, per cent.

32.1	25.8	7.4	1.4	-26.5	-47.7
-1.1	5.8	0.9	1.3	0.9	-9.8
1.5	1.1	3.9	-8.0	-6.6	14.1
-25.1	-30.3	-10.8	4.5	25.6	39.6

The matrix $Q = (q_{ij})$ for the data in Table 2 is displayed in Table 4. It graphically shows that the left two parents' statuses are positively related to mental category "Well" while the two right parents' statuses to category "Impaired". The absolute Quetelet index matrix, $W = (w_{ij})$ demonstrates a similar pattern, though with smaller entries.

3.4 Chi-Squared as the Covariance of Absolute Quetelet Indexes

Following Gilula (1981), let us consider, for a contingency table $P = (p_{ij})$, corresponding absolute Quetelet index matrices, $W = (w_{ij})$ and $W' = (w'_{ij})$, where

$$w_{ij} = (p_{ij} - p_{i+}p_{+j})/p_{+j},\ w'_{ij} = (p_{ij} - p_{i+}p_{+j})/p_{i+}$$

thus expressing absolute changes of row or coulmn probabilities when the columns (or rows) become known. Obviously,

$$w_{ij}w'_{ij} = \frac{(p_{ij} - p_{i+}p_{+j})^2}{p_{i+}p_{+j}} = p_{i+}p_{+j}q_{ij}^2 = \frac{p_{ij}^2}{p_{i+}p_{+j}} - 2p_{ij} + p_{i+}p_{+j} \tag{3.6}$$

Thus,

$$X^2 = \sum_{i\in I}\sum_{j\in J} w_{ij}w'_{ij} \tag{3.7}$$

Since the means of W and W' are zeros, (3.7) can also be interpreted as the covariance coefficient between corresponding indexes (assuming the uniform probability distribution of entries in $I \times J$, Gilula 1981).

3.5 Chi-Squared as a Variance of the Relative Quetelet Index

By using another equation in (3.6), $w_{ij}w'_{ij} = p_{i+}p_{+j}q_{ij}^2$, chi-squared can also be interpreted as the variance of the relative Quetelet index, q_{ij} in (3.5), when probability of the entry is $p_{i+}p_{+j}$ (Gilula and Krieger 1983). This distribution corresponds to the experiment in which all $i \in I$ and $j \in J$ are randomly and independently chosen with probabilities p_{i+} and p_{+j}, respectively. Thus,

$$X^2 = \sum_{i\in I}\sum_{j\in J} p_{i+}p_{+j}q_{ij}^2 \tag{3.8}$$

The average value of q_{ij} (under distribution $p_{i+}p_{+j}$) is zero, which allows us to consider formula (3.8) as the variance.

3.6 Chi-Squared as the Average Relative Quetelet Index

The chi-squared coefficient is the average relative Quetelet coefficient according to the bivariate distribution represented in the original table P (Mirkin 1985):

$$X^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} p_{ij}q_{ij} \tag{3.9}$$

It should be noted that the (i, j)-th item in (3.9),

$$p_{ij}q_{ij} = p_{ij}^2/(p_{i+}p_{+j}) - p_{ij} \tag{3.10}$$

differs from the (i, j)-th item (3.6) in (2.3), (3.8), and (3.7), by $p_{i+}p_{+j} - p_{ij}$. Curiously, the differences summed up within a row or column, disappear. For instance, $\sum_i (p_{i+}p_{+j} - p_{ij}) = p_{+j} - p_{+j} = 0$. Thus

Table 5: The entries in the left part are (3.10) and in the right part, (3.6), according to the data in Table 2. All entries are multipled by 1000.

Constituents of X^2 in (3.9)						Constituents of X^2 in (3.8)					
A	B	C	D	E	F	A	B	C	D	E	F
12.4	8.9	2.5	0.6	-5.8	-6.0	3.0	1.8	0.2	0.0	2.1	5.5
-0.6	3.3	0.6	1.1	0.5	-4.2	0.0	0.2	0.0	0.0	0.0	0.5
0.5	0.3	1.5	-3.7	-2.1	4.6	0.0	0.0	0.1	0.3	0.1	0.6
-6.9	-7.3	-3.9	2.5	12.0	16.9	2.3	3.2	0.5	0.1	2.5	4.8

the ranges of items (3.10) in (3.9) within any column and row must be greater than the ranges of items (3.6) in (3.8) because items (3.10) must have different signs while all items (3.6) are positive.

In the example of children's mental health cross-classified with the parents' statuses, the matrices of items (3.10) in (3.9) are shown in the left-hand part of Table 5 versus the items (3.6) in (2.3) (or, equivalently, in (3.7) and (3.8)), in the right-hand part of Table 5. We can see that the pattern in the left side of Table 5 is more "contrast" than the pattern in the right side because of the greater ranges of items $p_{ij}q_{ij}$ (3.10) than of $p_{i+}p_{+j}q_{ij}^2$ (3.6). Also, the pattern in the left side of Table 5 follows that of the Q matrix in Table 4 and displays the same pattern of interaction.

3.7 An Asymmetric Form of the Chi-Square Coefficient

In this section, an asymmetric analogue to the formulation (3.9) of X^2 will be explored, especially in its relation to the reduction-of-error coefficients proposed by Wallis, Goodman and Kruskal (Goodman and Kruskal 1954).

Averaging Absolute Quetelet Index. Let us average the absolute Quetelet coefficient, $w_{ij} = p_{ij}/p_{+j} - p_{i+}$ (Mirkin 1985):

$$Y^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} p_{ij}w_{ij} = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{p_{ij}^2}{p_{+j}} - \sum_{i=1}^{n} p_{i+}^2 \qquad (3.11)$$

For the data in Table 1, $Y^2 = 0.0059$.

The right-hand expression in (3.11) has been considered by Goodman and Kruskal (1954, p. 769-760) following an earlier suggestion by W.A. Wallis. They derived a chi-squared-like, though asymmetric, formula

$$Y^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{(p_{ij} - p_{i+}p_{+j})^2}{p_{+j}}, \qquad (3.12)$$

as well as the following interpretation of Y^2.

Reduction of the Proportion of Errors. The concept of proportional prediction refers to the situation when entities are chosen randomly from the a population (under the bivariate distribution $P = (p_{ij})$), and for any entity, its I-category is predicted as i either with probability p_{i+} (Case 1: when nothing is known of its J-category) or, p_{ij}/p_{+j}, (Case 2: when its J-category, j, is assumed to be known) (Goodman and Kruskal 1954). The average proportion of correct predictions in the Case 1 is $\sum_{i=1}^{n} p_{i+}^2$ and, in the Case 2, $\sum_{j=1}^{m}\sum_{i=1}^{n} \frac{p_{ij}^2}{p_{+j}}$, which leads to expression (3.11) as the increase in proportion of correct predictions, or the decrease in proportion of errors, when more exact information (Case 2 versus Case 1) becomes known.

Yet another interpretation of Y^2 has been given by Light and Margolin (1971) using the so-called Gini coefficient.

Qualitative Analogue to the Explained Variance. For a nominal variable I with m categories its Gini coefficient, or qualitative variance, can be defined as

$$G(I) = 1 - \sum_{i=1}^{m} p_{i+}^2$$

(see, for instance, Light and Margolin 1971). Obviously, $G(I) = 0$, when all the frequencies are zero except for one which is 1, and $G(I)$ reaches its maximum, $(m-1)/m$, when the distribution is uniform, that is, all $p_{i+} = 1/m$. In the case of data in Table 1, $G(I) = 0.732$.

Within a category $j \in J$, the variation is equal to $G(I/j) = 1 - \sum_{i \in I}(p_{ij}/p_{+j})^2$, which makes, on average, what can be called an analogue to the within-group variation: $G(I/J) = \sum_{j \in J} p_{+j} G(I/j) = 1 - \sum_{j \in J} \sum_{i \in I} p_{ij}^2 / p_{+j}$.

Thus, the between-group, or explained, variation is $G(I) - G(I/J)$ which equals Y^2 in (3.11). The proportion of explained variance, $(G(J) - G(J/I))/G(J) = Y^2/G(J)$, is

$$\tau = \frac{\sum_{j \in J} \sum_{i \in I} p_{ij}^2 / p_{+j} - \sum_{i=1}^{m} p_{i+}^2}{1 - \sum_{i=1}^{m} p_{i+}^2}, \tag{3.13}$$

which is referred to as Goodman and Kruskal's tau or tau-b.

In our example, $\tau = (0.732 - 0.726)/0.732 = 0.009$ since $G(J/I) = 0.726$ for the data in Table 1.

In section 5.4, this interpretation will be extended, via a linear model, to other contingency measures including the chi-squared coefficient.

4 DATA SCATTER IN CONTINGENCY DATA ANALYSIS

In this section, there will be presented two interpretations that are related to aggregate representations of the data table, one on a visual display (section 4.1), the other with sheer lumping (section 4.2).

4.1 Chi-Squared as the Data Scatter in Correspondence Factor Analysis

Correspondence Analysis (CA) is a method visually displaying both row and column categories in such a way that the distances between the presenting points reflect the pattern of co-occurrences in contingency table $P = (p_{ij})$. There have been several equivalent approaches developed for introducing the method, some of them involving the observations counted in the table P, such as the canonical correlation approach described in section 5.2, some not (see, for example, Diday, Lemaire, Pouget and Testu 1982; Goodman 1986; Benzécri 1992). Here we introduce CA in terms of the matrix P as is.

To be specific, let us concentrate on the problem of finding just two "underlying" factors, $u_1 = \{(v_1(i)), (w_1(j))\}$ and $u_2 = \{(v_2(i)), (w_2(j))\}$, with $I \cup J$ as their domain, such that each row $i \in I$ is

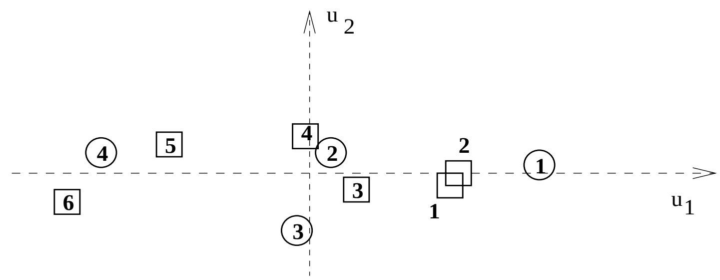

Figure 1: CA display for the rows and columns of Table 1 represented by circles and squares, respectively.

displayed as point $u(i) = (v_1(i), v_2(i))$ and each column $j \in J$ as point $u(j) = (w_1(j), w_2(j))$ on the plane as shown in Figure 1.

The coordinate row-vectors, v_l, and column-vectors, w_l, constituting u_l $(l = 1, 2)$ are calculated to approximate the relative Quetelet coefficients according to equations:

$$q_{ij} = \mu_1 v_1(i) w_1(j) + \mu_2 v_2(i) w_2(j) + \epsilon_{ij} \tag{4.1}$$

where μ_1 and μ_2 are positive reals, by minimizing the weighted least-squares criterion

$$E^2 = \sum_{i \in I} \sum_{j \in J} p_{i+} p_{+j} \epsilon_{ij}^2 \tag{4.2}$$

with regard to μ_l, v_l, w_l, subject to conditions of weighted ortho-normality:

$$\sum_{i \in I} p_{i+} v_l(i) v_{l'}(i) = \sum_{j \in J} p_{+j} w_l(j) w_{l'}(j) = \begin{cases} 1, & l = l' \\ 0, & l \neq l' \end{cases} \tag{4.3}$$

where $l, l' = 1, 2$.

The weighted criterion E^2 is equivalent to the unweighted least-squares criterion applied to the matrix with entries $a_{ij} = q_{ij}\sqrt{p_{i+}p_{+j}} = (p_{ij} - p_{i+}p_{+j})/\sqrt{p_{i+}p_{+j}}$. This implies that the factors are determined by the singular-value decomposition of matrix $A = (a_{ij})$. More explicitly, the optimal values μ_l and row-vectors $f_l = (v_l(i)\sqrt{p_{i+}})$ and column-vectors $g_l = (w_l(j)\sqrt{p_{+j}})$ are the maximal singular values and corresponding singular vectors of matrix A, defined by the equations: $Ag_l = \mu_l f_l$, $f_l A = \mu_l g_l$ (Diday et al. 1982; Goodman 1986; Benzécri 1992).

These equations, rewritten in terms of v_l and w_l, are considered to justify the joint display: the row-points appear to be averaged column-points and, vice versa, the column-points appear to be averaged versions of the row-points. The mutual location of the row-points is considered as justified by the fact that the between-row-point Euclidean distances squared, $d^2(u(i), u(i'))$, approximate the so-called chi-square distances between corresponding rows of the contingency table (Diday et al. 1982, Benzécri 1992). Here $u(i) = (v_1(i), v_2(i))$ for v_1 and v_2 rescaled in such a way that their norms are equal to μ_1 and μ_2, respectively. The chi-square distance can be defined in terms of Quetelet coefficients as $x^2(i, i') = \sum_{j \in J} p_{+j}(q_{ij} - q_{i'j})^2$. This applies also to the between-column distances defined analogously.

The values μ_l^2 are latent values of matrix $A^T A$. As is known, the sum of the latent values of a matrix is equal to its trace, defined as the sum of diagonal entries, that is, $Tr(A^T A) = \sum_{l=1}^{L} \mu_l^2$ where L is the rank of A. On the other hand, direct calculation shows that the sum of diagonal entries of $A^T A$ is

$$Tr(A^T A) = \sum_{i,j} (p_{ij} - p_{i+}p_{+j})^2/(p_{i+}p_{+j}) = X^2. \quad (4.4)$$

Thus,

$$X^2 = \mu_1^2 + \mu_2^2 + E^2 \quad (4.5)$$

which can be seen as a decomposition of the contingency data scatter, measured by the X^2, into contributions of the individual factors, μ_l^2, and unexplained residuals, E^2. (Here, $l = 1, 2$, but, actually, the number of factors sought can be raised up to the rank of matrix A). In a common situation, the first two latent values account for a major part of X^2, thus justifying the use of the plane of the first two factors for visualization of the interrelations between I and J.

In particular, for the matrix in Table 2, μ_s^2=0.0260 and μ_s^2=0.0014, accounting, respectively, for 94% and 5% of the data scatter measured by X^2=0.0277. The categories are displayed in the plane of vectors u_1 and u_2 renormalized in such a way that the norms of v_1, w_1 and v_2, w_2 are μ_1 and μ_2, respectively (Figure 1). We can see that the pattern of association observed earlier in terms of the frequency changes now is displayed in terms of the spatial arrangement of the categories.

4.2 Chi-Squared as Data Scatter in Aggregation of Contingency Tables

Another data-scatter based model arises when categories $i \in I$ and $j \in J$ are assumed to be elements of more general categories (Mirkin 1998). The hypothetical general categories are represented as classes of partitions V on I and W on J. For a pair of classes, $v \in V$ and $w \in W$, let us denote the summary frequency as $p_{vw} = \sum_{i\in v} \sum_{j\in w} p_{ij}$. The matrix of aggregate frequencies, p_{vw}, will be denoted as $P(V, W)$ while the original table is $P = P(I, J)$. The aggregate relative Quetelet coefficients, q_{vw}, are defined as $q_{vw} = (p_{vw} - p_{v+}p_{+w})/(p_{v+}p_{+w})$, where p_{v+} and p_{+w} are row- and column-marginals in $P(V, W)$.

If, for instance, we combine two intermediate symptom formations in Table 1 into a larger category, "Symptom", while maintaining the two other categories separate, and combine the parents' socioeconomic statuses pair-wise, we arrive at the aggregate version of Table 1 presented in Table 6 (along with the corresponding relative Quetelet index values on the right).

Let us score all potential aggregations according to their closeness to the original matrix in terms of the Quetelet coefficients. The scoring function is again the weighted least squares:

$$E^2(V, W) = \sum_{v\in V} \sum_{w\in W} \sum_{i\in v} \sum_{j\in w} p_{i+}p_{+j}(q_{ij} - q_{vw})^2 \quad (4.6)$$

It is not difficult to prove that

$$X^2(I, J) = X^2(V, W) + E^2(V, W) \quad (4.7)$$

which can be interpreted as decomposition of the original data scatter, $X^2(I, J)$, in two parts: the explained one, $X^2(V, W)$, to be maximized, and the unexplained one, $E^2(V, W)$, to be minimized.

Table 6: Cross-classification of 1660 subjects according to aggregate categories: N_{vw}, on the left, and q_{vw}, per cent, on the right.

	Parents' status				Parents' status		
Mental health status	A+B	C+D	E+F	Total	A+B	C+D	E+F
Well	121	129	57	307	29.0	4.0	-36.1
Symptom	300	388	276	964	1.9	-0.4	-1.4
Impaired	86	154	149	389	-27.6	-2.1	31.9
Total	507	671	482	1660			

Thus, in the model of contingency table aggregation, the chi-squared coefficient measures both the data scatter, $X^2(I,J)$, and the explained part of it, $X^2(V,W)$. This may be used for scoring the aggregate table against the original one. For the aggregate Table 6, the value of $X^2 = 0.0244$. The aggregation keeps 0.0244/0.0277=88.1% of the original value of X^2, which is not that bad for a table shrunk from 24 to 9 entries.

5 GEOMETRIC INTERPRETATIONS IN THE ENTITY SPACE

In this section, we further advance in the models of multivariate statistics to generate interpretations for X^2 by additionally involving elements of the sample O counted in the data table, $P = (p_{ij})$.

5.1 Indicator Variables and Nominal Subspaces

The following multivariate representation of a contingency table can be found in many texts on multivariate statistics (see, for instance, Kendall and Stuart 1979, Diday et al. 1982, Goodman 1986).

The set of m categories, I, partitions the set of N entities O, counted in the contingency table, into classes $i \in I$ that can be represented by a binary $N \times m$ matrix, $\mathbf{r}=(r_{oi})$, whose columns correspond to categories $i \in I$, and rows, to objects $o \in O$. The entry $r_{oi} = 1$ for $o \in i$ and $r_{oi} = 0$ for $o \notin i$. Thus, columns of $\mathbf{r}$ are binary indicators of categories $i \in I$. An analogous $N \times n$ matrix, $\mathbf{s}=(s_{oj})$, can be set for J: $s_{oj} = 1$ when $o \in J$ and $s_{oj} = 0$ otherwise.

The issue we are going to address is of evaluating similarity between I and J in terms of the linear subspaces spanning these matrices.

The linear subspace, $L(I)$, spanning $\mathbf{r}$'s columns is the set of N-dimensional vectors $z = (z_o)$ defined as follows: $z \in L(I)$ if and only if there exists an m-dimensional vector, $a = (a_i)$, such that for any $o \in O$, $z_o = \sum_{i=1}^{m} r_{oi}a_i$, or, in matrix terms, $z=\mathbf{r}a$. One can think of the vector $a = (a_i)$ as a numerical coding system for the categories $i \in I$. For the column categories, $j \in J$, the spanning subspace $L(J)$ is defined analogously.

If some components of a are equal to each other, for instance, $a_{i'} = a_{i''}$, then corresponding categories, i' and i'' in this case, get the same codes in $z = \mathbf{r}a$ and become indistinguishable in the coding a. This means that among the numerical codings are 'many-to-one', not only 'one-to-one' codings. The extremal 'all-to-one' coding assigns the same numerical value α to all categories i, which means that the bisector, the unidimensional space of all the equicomponent vectors $a = \alpha\mathbf{u}$, where

$\mathbf{u}=(1,1,...,1)^T$, is contained in $L(I)$ for any I. This bisector unidimensional space will be denoted by $L(U)$. The linear subspaces of all nominal variables on O overlap on $L(U)$ just because of the format of representation, not because of the variables' contents. We can remove this nuisance at once by orthogonally subtracting $L(U)$ from all $L(I)$, but this seems too technical in the context of this paper. For the sake of simplicity, we deal with this only when needed.

5.2 Chi-Squared as the Summary Canonical Correlation Between Nominal Subspaces

The canonical correlation coefficient is defined as the maximum correlation between vectors $z \in L(I)$ and $y \in L(J)$, that is, a solution to the problem of maximization of the scalar product (z, y) with regard to all normed and centered $z = \mathbf{r}a \in L(I)$ and $y = \mathbf{s}b \in L(J)$. The solution to this problem can be found with matrix $P_I P_J$, where P_I and P_J are orthogonal projection matrices to the spaces $L(I)$ and $L(J)$, respectively. The (orthogonal) projection matrix $P_{L(X)}$ to the space, $L(X)$, spanning columns of a matrix, X, is known to be equal to $P_{L(X)} = X(X^T X)^{-1} X^T$. It is not difficult to prove that the (o', o'')-th entry of the $N \times N$ matrix P_I is equal to zero when o' and o'' belong to different categories of I and it is equal to $1/N_{i+}$ when both o' and o'' belong to the same category $i \in I$. Thus, the only non-zero entries in P_I are those corresponding to the entities $o \in O$ covered by the same category, $i \in I$: they are all equal to the same quantity, $1/N_{i+}$, for all pairs (o', o'') with $o', o'' \in i$. Similarly P_J is defined with regard to the column set J: each its nonzero entry (o', o'') is equal to $1/N_{+j}$ when both o' and o'' belong to the same category $j \in J$.

It appears, the first-order optimim conditions for the maximum correlation problem amount to the problem of finding latent values of the product, $P_I P_J$; they are usually referred to as canonical correlations. All the latent values can be represented as scalar products, (z, y), between elements of mutually-oriented orthonormal bases $z \in L(I)$ and $y \in L(J)$, where z and y are the corresponding left and right latent vectors of $P_I P_J$ (Diday et al. 1982). The trivial latent value, 1, corresponds to "correlation" between bisector vectors in the "parasite" subspace, $L(U)$, that is a part of both $L(I)$ and $L(J)$.

The sum of all canonical correlations between $L(I)$ and $L(J)$ (except for the trivial correlation 1) is equal to the trace of $P_I P_J$ short of 1, $Tr(P_I P_J) - 1$, which thus can be considered as a "correlation-wise" measure of similarity between the spaces. It is not difficult to see that

$$Tr(P_I P_J) = \sum_{i \in R} \sum_{j \in S} \frac{N_{ij}^2}{N_{i+} N_{+j}} \tag{5.1}$$

Indeed, the $Tr(P_I P_J)$ is the sum of products, $p_{o'o''}(I) p_{o'o''}(J)$, of corresponding entries in P_I and P_J over all $o', o'' \in O$. The product is not zero if both o' and o'' belong to the intersection of some row and column categories, i and j. The number of entities in this intersection is N_{ij} and the number of entity pairs is N_{ij}^2; the corresponding entries are $1/N_{i+}$ in P_I and $1/N_{+j}$ in P_J. This leads to N_{ij}^2 products $(1/N_{i+})(1/N_{+j})$ as the contribution of the intersection of i and j to $Tr(P_I P_J)$ and proves (5.1).

Therefore, the sum of all non-trivial correlations, $Tr(P_I P_J) - 1$, is exactly the chi-squared, X^2.

Since dimensions of the spaces $L(I)$ and $L(J)$ are m and n, respectively, the number of non-trivial canonical correlations is equal to $\min(m-1, n-1)$. Thus, the Cramér coefficient (3.2) is the average non-trivial canonical correlation between spaces $L(I)$ and $L(J)$.

5.3 Chi-Squared as the Scalar Product of Centered Projection Matrices

One more interpretation of X^2 can be built upon a holistic view of the spaces $L(I)$ and $L(J)$. The projection matrix to a linear subspace can be considered an analogue to the normal vector to a hyperplane. Thus, correlation between variables can be scored according to similarities between the orthogonal projection matrices to corresponding subspaces.

Let us consider the $N \times N$ projection matrices P_I and P_J as $N \times N$ vectors and remove the effect of the "parasite" part, $L(U)$, by subtracting the projection matrix, P_U to $L(U)$, from them. All elements of P_U are equal to $1/N$ which is simultaneously the average value of the entries in either P_I or P_J. Thus, $P_I - P_U$ and $P_J - P_U$ are centered versions of P_I and P_J.

The scalar product of these matrices is $(P_I - P_U, P_J - P_U) = (P_I, P_J) - (P_U, P_I) - (P_U, P_J) + (P_U, P_U)$. It is not difficult to see that $(P_U, P_I) = (P_U, P_J) = (P_U, P_U) = 1$. On the other hand, $(P_I, P_J) = Tr(P_I P_J)$, which has been already calculated in (5.1).

Thus, X^2 is the scalar product of the centered projection matrices: $(P_I - P_U, P_J - P_U) = Tr(P_I P_J) - 1 = X^2$ (Saporta 1975).

Moreover, the Tschuprow coefficient T^2 is the product-moment correlation coefficient between P_I and P_J (Mirkin 1985, Saporta 1988):

$$Corr(P_I, P_J) = \frac{(P_I - P_U, P_J - P_U)}{\|P_J - P_U\| \|P_J - P_U\|} = \frac{X^2}{\sqrt{(m-1)(n-1)}} = T^2 \qquad (5.2)$$

Indeed, $\|P_I - P_U\|^2 = \sum_i (1/N_{i+})^2 N_{i+}^2 - (1/N)^2 N^2 = m - 1$ and, similarly, $\|P_J - P_U\|^2 = n - 1$.

5.4 Chi-Squared as the Between-Group Variance for a Nominal Variable

In the two preceding subsections we looked at the spaces generated by I and J symmetrically. However, X^2 appears also in an asymmetrical view at I and J, assuming one, say I, as presented by its category indicator matrix, $\mathbf{r}$, while the other, J, making a grouping of the entity set O according to its non-overlapping categories $j \in J$ (Mirkin 1998).

For any column-centered data matrix, $Y = (y_{ok})$, whose rows correspond to objects $o \in O$ counted in the contingency table $P = (p_{ij})$, the partition of O according to J provides for the following decomposition of the summary data variance (Diday et al. 1982):

$$\sum_{k=1}^{M} \sigma_k^2 = \sum_{j \in J} \sum_{k=1}^{M} p_{+j} c_{jk}^2 + \sum_{k=1}^{M} \sum_{j \in J} p_{+j} \sigma_{jk}^2 \qquad (5.3)$$

where $c_j = (c_{jk})$ is the gravity center of entities in category j, that is, $c_{jk} = \sum_{o \in j} y_{ok}/N_{+j}$; $\sigma_k^2 = \sum_{o \in O} y_{ok}^2/N$ is the overall variance of column k in Y, and $\sigma_{jk}^2 = \sum_{o \in S_j} (y_{ok} - c_{jk})^2/N_{+j}$ is its variance within category j.

The items in the right-hand side of (5.3) are referred to as the inter-group variance and within-group variance, respectively.

Let us take the category indicator matrix $\mathbf{r}$=(r_{oi}) of variable I, with $r_{oi} = 1$ if category $i \in I$ covers o and $r_{oi} = 0$, otherwise. Let us rescale $\mathbf{r}$ by the standard transformation, $y_{oi} = (r_{oi} - a_i)/b_i$ where a_i is the mean of the column i in $\mathbf{r}$, that is, $a_i = p_{i+}$, and b_i a scale factor that will be taken care of later. Thus Y, in this case, is a centered matrix representing categories, $i = 1, ..., m$, that substitute $k = 1, ..., M$ in the decomposition (5.3). Then the elements of (5.3) obviously are:

1. The total variance on the left, denoted as $S^2(I)$, is:

$$S^2(I) = \sum_i \sigma_i^2 = \sum_i \frac{p_{i+}(1 - p_{i+})}{b_i^2} \qquad (5.4)$$

2. The within-category means:

$$c_{ji} = \sum_{o \in j} [(s_{oi} - p_{i+})/b_i]/N_{+j} = (p_{ij} - p_{i+}p_{+j})/(p_{+j}b_i).$$

3. The intergroup variance, the left-hand term in the right part of (5.3), denoted as $S^2(I/J)$:

$$S^2(I/J) = \sum_{j=1}^{n} \sum_{i=1}^{m} p_{+j} c_{ji}^2 = \sum_{i,j} \frac{(p_{ij} - p_{i+}p_{+j})^2}{p_{+j} b_i^2} \qquad (5.5)$$

Let us specify the scaling factor by $b_i = \sqrt{p_{i+}}$, thus assuming the less frequent categories to be more important by contributing more to $S(I)$ in (5.4). With this b_i, $S^2(I) = m - 1$ and $S^2(I/J) = X^2$ in (5.5) $(i \in I)$.

Thus, X^2 is the intergroup variance of a data matrix corresponding to nominal variable I, when the grouping is done according to the other variable, J. The data matrix consists of the binary columns standardized by shifting the data entries by the column's means, p_{+j}, and rescaling by the column means' square roots.

Rather interestingly, the Wallis coefficient, Y^2, has the same meaning when no rescaling of the I categories has been done, that is, when $b_i = 1$ for all $i \in I$. Indeed, $S^2(I/J)$ in (5.5) is Y^2, in this case.

Referring to the interpretation of the intergroup variance as the explained part of the total data variance, we may consider the relative intergroup variance, $S^2(I/J)/S^2(I)$, as a normalized association measure. With respect to $b_i = \sqrt{p_{i+}}$ or $b_i = 1$, the relative intergroup variance is equal to either $M^2 = X^2/(m-1)$ or τ (3.13), respectively. Curiously, this perspective has led us to an asymmetrically normalized version of X^2, $M^2 = X^2/(m-1)$, that seems an intermediate between Cramér' and Tschuprow's normalizations. In the example of Table 1, $M^2 = 0.0277/3 = 0.0092$, thus almost coinciding with its relative intergroup variance counterpart, $\tau = 0.009$.

6 CHI-SQUARED FOR THE FOUR-FOLD CONTINGENCY TABLE

6.1 Correlation/Determination between Binary Variables

In the case when both of the variables, I and J, are binary and have two categories each, the

Table 7: Four-fold format for cross-classifications of binary variables.

Category	Y	$\overline{Y}$	Total
X	a	b	a+b
$\overline{X}$	c	d	c+d
Total	a+c	b+d	1

Table 8: A standard four-fold table for two-way prediction of rows by columns.

[!]

Category	X	$\overline{X}$	Total
Y	ϵf	$(1-\epsilon)g$	$\epsilon f + (1-\epsilon)g$
$\overline{Y}$	$(1-\epsilon)f$	ϵg	$(1-\epsilon)f + \epsilon g$
Total	f	g	1

contingency table is typically presented as the so-called four-fold table in Table 7.

In this case, the chi-squared coefficient is equal to

$$X^2 = \tau = \frac{(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} \tag{6.1}$$

which is closely related to the formula for the product-moment correlation coefficient, ρ, between category (dummy) variables corresponding to a pair of categories, each taken from either variable:

$$\rho = \frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} \tag{6.2}$$

In fact, $X^2 = \rho^2$; that is, both X^2 and τ are equal to the determination coefficient in this case. This is no coincidence: the set of all the numerical codings for the case of only two categories can be covered with just linear transformations since they are specified by two constants that can be associated to the slope and intercept. Thus, the determination coefficient, ρ^2, that is equal to the share of the overall variance explained in the class of linear transformations only, must coincide with the share of the overall variance explained in the class of all numerical codings, expressed in M^2 or τ. For the four-fold table, $M^2 = X^2$.

6.2 Standard Prediction Table

To calibrate the values of X^2 in a manner resembling the calibration of the Shannon's entropy in bits (Brillouin 1962), a standard prediction table can be utilized. Such a table is presented in Table 8 from Mirkin (1985, p. 81).

Here, f and g are given proportions of the categories X and $\overline{X}$, respectively. The parameter ϵ whose values are between 0 and 1, specifies the extent of correctness in prediction of rows when columns are known. For instance, $\epsilon = 0$ predefines exact correspondence between X and $\overline{Y}$ and between $\overline{X}$ and Y, while $\epsilon = 1$ means that X and Y, as well as $\overline{X}$ and $\overline{Y}$, exactly match each other. Values $\epsilon = 0.1$ and $\epsilon = 0.9$ support either of these conclusions with 10% error. The worst correspondence is when $\epsilon = 0.5$, the case of statistical independence.

With a little arithmetic, we can see that

$$X^2 = \frac{(1-2\epsilon)^2 f(1-f)}{\epsilon(1-\epsilon)(1-2f)^2 + f(1-f)} \tag{6.3}$$

which becomes

$$X^2 = (1-2\epsilon)^2 \tag{6.4}$$

when $f = 0.5$.

The latter expression shows that $X^2 = 0.64$ corresponds to 90% correct prediction, $X^2 = 0.36$ to 80% correct prediction and $X^2 = 0.04$ to 60% correct prediction.

For $f \neq 0.5$, when distribution of X is far from uniform, smaller values of X^2 may correspond to better predictions. If, for instance, $f = 0.1$, then $X^2 = 0.04$ corresponds to 66% correct predictions.

7 CONCLUSION

Eleven definitions of the chi-squared coefficient as either an independence/homogeneity test statistic or summary association measure have been presented.

In the context of association measurement, the average relative Quetelet index format (section 3.6) seems especially important because the items (3.10) in (3.9) graphically display the positive and negative constituents of the overall category-to-category associations in the contingency table, totalling to X^2. Also, this format shows how this index parallels that of the (asymmetric) indices of proportional error reduction, Y^2 and τ. The parallel is extended in section 5.4 leading to an asymmetric version M^2 of the chi-squared.

The normalized versions of chi-squared also have been explained in some of the association measure frameworks: the Cramér's coefficient is just the average canonical correlation and Tschuprow's coefficient is the product-moment correlation coefficient between projection matrices. The latter, actually, relates X^2 and T^2 to the Pearson product-moment correlation with its numerous interpretations (Rodgers and Nicewanders 1988).

The calibration of chi-squared in section 6.2 may be useful as a straightforward prediction-driven interpretation of the index values.

References

Benzécri, J.-P. (1992), *Correspondence Analysis Handbook*, New York: Marcel Dekker.

Brillouin, L. (1962), *Science and Information Theory*, New York: Academic Press.

Diday, E., Lemaire, J., Pouget, J., and Testu, F. (1982), *Éléments d'Analyse de Données*, Paris: Dunod.

Gilula, Z. (1981), "On Pearson's Chi-Squared Statistic for Independence in Two-Way Contingency Tables," *The American Statistician, 35*, 2, 96-97.

Gilula, Z., and Krieger, A. M. (1983), "The Decomposability and Monotonicity of Pearson's Chi-Square for Collapsed Contingency Tables with Applications," *Journal of Amer. Stat. Association, 78*, No. 381, 176-180.

Goodman, L. A. (1986), “Some Useful Extensions of the Usual Correspondence Analysis Approach and the Usual Log-Linear Models Approach in the Analysis of Contingency Tables,” *International Statistical Review, 54*, 3, 243-309.

Goodman, L. A., and Kruskal, W. H. (1954), “Measures of Association for Cross Classifications,” *Journal of American Statistical Association, 49*, 732-764.

—(1979), *Measures of Association for Cross Classifications*, New York: Springer-Verlag.

Kendall, M. G. and Stuart, A. (1979), *The Advanced Theory of Statistics, 2*, 4th ed. New York: Hafner.

Light, R. J. and Margolin, B. H. (1971), “An Analysis of Variance for Categorical Variables,” *Journal of American Statistical Association, 66*, 534-544.

Mirkin, B. (1985), *Grouping in Socioeconomic Research*, Moscow: Finansy i Statistika (in Russian).

—(1998), “Least-Squares Structuring, Clustering, and Related Computational Issues,” *The Computer Journal, 41*, 518-536.

Pearson, K. (1900a), “On the Correlation of Characters Not Quantitatively Measurable,” *Phil. Trans. Roy. Soc., A, 195*, 1.

—(1900b), “On a Criterion That a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is Such That It Can Be Reasonably Supposed to Have Arisen in Random Sampling,” *Phil. Mag., 50*, 157-175. Also, in *Karl Pearson's Early Statistical Papers* (1948), Cambridge: University Press, 339-357.

Quetelet, A. (1832), “Sur la Possibilité de Mesurer l'Influence des Causes qui Modifient les Éléments Sociaux,” *Lettre à M. Willermé de l'Institut de France*, Bruxelles.

Rodgers, J. L., and Nicewander, W. A. (1988), “Thirteen Ways to Look at the Correlation Coefficient,” *The American Statistician*, 42, no. 1, 59-66.

Saporta, G. (1975), *Liaisons Entre Plusieurs Ensembles de Variables et Codages de Données Qualitatives*, Paris: Universite Paris VI.

—(1988), “About Maximal Association Criteria in Linear Analysis and in Cluster Analysis,” *Classification and Related Methods of Data Analysis*, (ed.) H.-H. Bock, Amsterdam: Elsevier, 541-550.

Srole, L., Langner, T. S., Michael, S. T., Opler, M. K., and Rennie, T. A. (1962), *Mental Health in the Metropolis: The Midtown Manhattan Study*, New York: McGraw-Hill.

Yule, G. U. (1900), “On the Association of Attributes in Statistics: with Illustrations from the Material of the Childhood Society,” *Phil. Trans., A, 194*, 257-319. Also: A. Stuart and M. G. Kendall (1971) *Statistical Papers of George Udny Yule*, New York: Hafner Publishing Company, 7-70.