THEMIS SIGNAL ANALYSIS STATISTICS RESEARCH PROGRAM

GOODNESS OF FIT OF AN ASSIGNED SET OF SCORES FOR THE ANALYSIS OF ASSOCIATION IN A CONTINGENCY TABLE

by

A. M. Kshirsagar

Technical Report No. 24
Department of Statistics THEMIS Contract

February 19, 1969

Research sponsored by the Office of Naval Research
Contract N00014-68-A-0515
Project NR 042-260

DEPARTMENT OF STATISTICS
Southern Methodist University

GOODNESS OF FIT OF AN ASSIGNED SET OF SCORES FOR THE ANALYSIS OF ASSOCIATION IN A CONTINGENCY TABLE

by

A. M. Kshirsagar*
Southern Methodist University
Dallas, Texas 75222

ABSTRACT

The problem of association between two attributes in a p X q contingency table can be looked upon as the problem of relationship between two vector variables $\underline{x}$ and $\underline{y}$. If there is only one true non-zero canonical correlation between $\underline{x}$ and $\underline{y}$, the association between the two attributes is of rank 1 and in this case, one set of scores is adequate to describe the association completely; these scores are nothing but the coefficients in the canonical variates corresponding to the true non-zero canonical correlation. Given a set of hypothetical scores $\alpha_1, \alpha_2, \cdots, \alpha_p$ for the rows, one is interested in testing their goodness of fit. Tests for this are suggested in this paper. For obtaining these tests, a preliminary result about direction and collinearity factors in discriminant analysis, when S irrelevant variables are eliminated, is needed. This is derived in part one of this paper.

*This research was sponsored by the Office of Naval Research, Contract No. N00014-68-A-0515, Project No. NR042-260. Reproduction in whole or in part is permitted for any purpose of the United States Government.

Part I

1. Relationship between two vectors

The problem of association between two vectors $\underline{x}(p \times 1)$ and $\underline{y}(q \times 1)$ arises in regression analysis, multivariate analysis of variance, discriminant analysis and in contingency table analysis. This relationship has different interpretations and implications in these fields, but in each case it can be expressed in terms of canonical correlations and canonical variables. The canonical correlations $r_1, r_2, \cdots, r_p$ $(p \leq q)$ in a sample, are the roots of the equation

$$\left| \begin{array}{c|c} -r^2 C_{xx} & C_{xy} \\ \hline C_{yx} & -r^2 C_{yy} \end{array} \right| = 0 \tag{1.1}$$

and the canonical variables corresponding to r_i^2 are $\underline{\ell}'_{(i)}\underline{x}$ and $\underline{m}'_{(i)}\underline{y}$ $(i = 1, 2, \cdots, p)$, where the column vectors $\underline{\ell}_{(i)}$, $\underline{m}_{(i)}$ satisfy the equation

$$\left[\begin{array}{c|c} -r_i^2 C_{xx} & C_{xy} \\ \hline C_{yx} & -r_i^2 C_{yy} \end{array} \right] \left[\begin{array}{c} \ell_{(i)} \\ \hline m_{(i)} \end{array} \right] = 0 \tag{1.2}$$

Here

$$C = \left[\begin{array}{c|c} C_{xx} & C_{xy} \\ \hline C_{yx} & C_{yy} \end{array} \right] \tag{1.3}$$

is the matrix of the corrected sum of squares (s.s.) and sum of products (s.p.) of observations on $\underline{x}$ and $\underline{y}$ and is based on n degrees of freedom (d.f.).

The true or population canonical correlations are denoted by $\rho_1, \rho_2, \cdots, \rho_p$. If all the ρ's are null, there is no association between $\underline{x}$ and $\underline{y}$ and under the assumption of normality, this is tested by using any one of the following criteria:

$$\text{Wilks's (1932) } \Lambda \text{ criterion; } \quad \Lambda = |A|/|A + B| \tag{1.4}$$

$$\text{or} \quad \text{Pillai's (1955) criterion} \quad t_r(A + B)^{-1}B \ . \tag{1.5}$$

$$\text{where } B = C_{xy}C_{yy}^{-1}C_{yx} \ , \ A = C_{xx} - C_{xy}C_{yy}^{-1}C_{yx} \ , \ A + B = C_{xx} \tag{1.6}$$

If only $\rho_1 \neq 0$ but $\rho_2 = \cdots = \rho_p = 0$, we say that the association between $\underline{x}$ and $\underline{y}$ is of rank 1. In this case, the entire association can be adequately described by the canonical variates corresponding to ρ_1. In discriminant analysis, this means that the means of q + 1 groups to be discriminated are collinear and a single discriminant function is adequate. Testing the goodness of fit of a single discriminant function $\underline{\alpha}'\underline{x} = \alpha_1 x_1 + \cdots + \alpha_p x_p$, in this context, means that one wishes to test (1) whether $\underline{\alpha}'\underline{x}$ agrees with the true canonical variate corresponding to ρ_1 and (2) whether one linear function is adequate at all to describe completely the relationship between $\underline{x}$ and $\underline{y}$. (1) is called the 'direction' aspect and (2) is called the collinearity aspect of the goodness of fit test. Bartlett (1951) and Williams (1955) derived tests for this purpose by factorizing Wilks' Λ as

$$\Lambda = \Lambda_1 \cdot \Lambda_2 \cdot \Lambda_3 \ . \tag{1.7}$$

where (see Kshirsagar 1964)

$$\Lambda_1 = \underline{\alpha}'A\underline{\alpha}/\underline{\alpha}(A+B)\underline{\alpha} \tag{1.8}$$

$$\Lambda_2 = 1 - \frac{\underline{\alpha}'B(A+B)^{-1}B\underline{\alpha}/\underline{\alpha}'B\underline{\alpha}}{\underline{\alpha}'A\underline{\alpha}/\underline{\alpha}'(A+B)\underline{\alpha}} \tag{1.9}$$

$$\Lambda_3 = \Lambda/\Lambda_1\Lambda_2 \tag{1.10}$$

Λ_2 is the direction factor and Λ_3 is the 'partial' collinearity factor. Bartlett has given an alternative factorization also viz.

$$\Lambda = \Lambda_1\Lambda_4\Lambda_5 \ . \tag{1.11}$$

where

$$\Lambda_4 = \Lambda \left\{ 1 + \frac{\underline{\alpha}'BA^{-1}B\underline{\alpha}}{\underline{\alpha}'B\underline{\alpha}} \right\} \tag{1.12}$$

$$\Lambda_5 = \Lambda/\Lambda_1\Lambda_4 \tag{1.13}$$

Λ_4 is the collinearity factor and Λ_5 is the 'partial' direction factor. A statistic t is said to have a $\Lambda(n, p, q)$ distribution, if it is distributed as $\prod_{i=1}^{p} U_i$ where U_i's are independent and U_i has the distribution

$$\text{Const. } U_i^{\frac{n-q-i-1}{2}} (1 - U_i)^{\frac{q-2}{2}} dU_i \tag{1.14}$$

Bartlett (1951) has shown that, in this case,

$$-\left\{ n - \frac{1}{2}(p + q + 1) \right\} \log_e t \tag{1.15}$$

has a χ^2 distribution with pq d.f. in large samples. If the null-hypothesis of goodness of fit of $\underline{\alpha}'\underline{x}$ is true, he shows that Λ_2 is a $\Lambda(n-1, 1, p-1)$ and Λ_3 is an independent $\Lambda(n-2, q-1, p-1)$. Alternatively Λ_4 is $\Lambda(n-1, q-1, p-1)$ and Λ_5 is an independent $\Lambda(n-q, 1, p-1)$. Briefly, Λ_2 is based on p-1 d.f., Λ_3 on (p-1)(q-1) d.f., Λ_4 on (p-1)(q-1) d.f. and Λ_5 on (p-1) d.f.

The author has shown, in an unpublished paper (1969), that the other

criterion $t_r B(A+B)^{-1}$ can also be partitioned, analogous to this factorization of Λ, as

$$n\, t_r B(A+B)^{-1} = \gamma_1 + \gamma_2 + \gamma_3 \tag{1.16}$$

where

$$\frac{1}{n}\gamma_1 = \frac{\underline{\alpha}' B \underline{\alpha}}{\underline{\alpha}'(A+B)\underline{\alpha}} \cdot \frac{1}{n}\gamma_2 = \frac{\underline{\alpha}' B (A+B)^{-1} B \underline{\alpha}}{\underline{\alpha}' B \underline{\alpha}} - \frac{1}{n}\gamma_1$$

and

$$\frac{1}{n}\gamma_3 = t_r B(A+B)^{-1} - \frac{1}{n}\gamma_1 - \frac{1}{n}\gamma_2 \tag{1.17}$$

Here γ_2 is the 'direction' part and γ_3 is the 'collinearity' part and under the null hypothesis of goodness of fit of $\underline{\alpha}'\underline{x}$, they are distributed independently as χ^2 with $p - 1$ d.f. and χ^2 with $(p-1)(q-1)$ d.f. respectively, in large samples.

2. Elimination of Irrelevant Variables

In some situations, it so happens that one is interested in studying the relationship between --- not $\underline{x}$ and $\underline{y}$ --- but between <u>residual</u> variates $\underline{z}$ and $\underline{w}$, where the latter are obtained from $\underline{x}$ and $\underline{y}$ by eliminating the first S sample canonical variables. These first S sample canonical variables are known apriori to be irrelevant and are therefore to be excluded. Let $L_1\underline{x}$ and $M_1\underline{y}$, where

$$\underset{S \times p}{L_1} = [\underline{\ell}_{(1)} | \underline{\ell}_{(2)} | \cdots | \underline{\ell}_{(S)}]' \tag{2.1}$$

and

$$\underset{S \times q}{M_1} = [\underline{m}_{(1)} | \underline{m}_{(2)} | \cdots | \underline{m}_{(S)}]' \tag{2.2}$$

be the first S canonical variables. On account of (1.2), we find

$$C_{xx}L_1'R = C_{xy}M_1' \tag{2.3}$$

where R is the S X S diagonal matrix of r_i^2 (i = 1, 2, ···, S). One can also show from (1.2) that

$$BL_1' = (A+B)L_1'R \tag{2.4}$$

and

$$AL_1' = (A+B)L_1'(I-R) \tag{2.5}$$

Let L_2 be a (p-S) X p matrix and M_2 a (q-S) X q matrix such that

$$L_2C_{xx}L_1' = 0 \quad , \quad M_2C_{yy}M_1' = 0 \tag{2.6}$$

i.e., $L_2\underline{x}$ and $L_1\underline{x}$ are uncorrelated and so also are $M_2\underline{y}$ and $M_1\underline{y}$. From (2.4), (2.5), (2.6) it can be seen easily that

$$L_2B_1' = 0 \quad , \quad L_2AL_1' = 0 \tag{2.7}$$

We can now take

$$z = L_2\underline{x} \quad ; \quad w = M_2\underline{y} \tag{2.8}$$

as our residual variables, after eliminating $L_1\underline{x}$ and $M_1\underline{y}$. We now want to test the goodness of fit of an assigned function $\underline{\alpha}'\underline{x}$ for the relationship between $\underline{z}$ and $\underline{w}$. It is obvious that this assigned function must be so chosen that it is uncorrelated with the eliminated variables $L_1\underline{x}$; in other words, it must be a linear function, say $\underline{k}'\underline{z}$ of $\underline{z}$ alone. If so, $\underline{k}$ will satisfy

$$\underline{\alpha} = L_2'\underline{k} \tag{2.9}$$

We define C_{zz}, C_{zw}, C_{ww} in the same way as in (1.3) and then A_z, B_z and

$A_z + B_z$ as in (1.6). We can, then easily write down the new direction and collinearity factors Λ_{2z} , Λ_{3z} , Λ_{4z} , Λ_{5z} or r_{2z} , r_{3z} etc. by using $\underline{k}'\underline{z}$ instead of $\underline{\alpha}'\underline{x}$ and A_z, B_z for A and B in (1.9), (1.10), (1.12), (1.13), and (1.17). We must also replace n by n-S, p by p-S and q by q-S as S variables have been eliminated from $\underline{x}$ and from $\underline{y}$. We, however, wish to express these test statistics in terms of our old matrices A, B and the assigned vector $\underline{\alpha}$. This can be done as below:

From (2.3) and (2.6),

$$L_2 C_{xy} M_1' = 0 \tag{2.10}$$

Hence

$$\begin{aligned} B_z = C_{zw} C_{ww}^{-1} C_{wz} &= L_2 C_{xw} C_{ww}^{-1} C_{wx} L_2' \\ &= L_2 \left[C_{xy} C_{yy}^{-1} C_{yx} - C_{xy} M_1' (M_1 C_{yy} M_1')^{-1} M_1 C_{yx} \right] L_2' \\ &= L_2 B L_2' \quad , \quad \text{on account of (2.10)} \end{aligned} \tag{2.11}$$

Also

$$\begin{aligned} A_z + B_z = C_{zz} &= L_2 C_{xx} L_2' \\ &= L_2 (A+B) L_2' \end{aligned} \tag{2.12}$$

Let

$$L = \begin{array}{c} \left[\begin{array}{c} L_1 \\ \hline L_2 \end{array} \right] \begin{array}{l} S \\ p-S \end{array} \\ p \end{array} \tag{2.13}$$

Then $(A+B)^{-1} = L'(LC_{xx}L')^{-1}L$

$$= L' \left[\begin{array}{c|c} L_1 C_{xx} L_1' & 0 \\ \hline 0 & L_2 C_{xx} L_2' \end{array} \right]^{-1} L$$

on account of (2.6)

$$= \sum_{i=1}^{2} L_i (L_i C_{xx} L_i')^{-1} L_i \tag{2.14}$$

Hence

$$\underline{k}' B_z (A_z + B_z)^{-1} B_z \underline{k} = \underline{k}' L_2 B L_2' (L_2 C_{xx} L_2')^{-1} L_2 B L_2' \underline{k}$$

$$= \underline{\alpha}' B \left\{ (A + B)^{-1} - L_1' (L_1 C_{xx} L_1')^{-1} L_1 \right\} B \underline{\alpha}$$

$$= \underline{\alpha}' B (A + B)^{-1} B \underline{\alpha} \quad . \tag{2.15}$$

on account of (2.7). In exactly the same way, it can be shown that

$$\underline{k}' B_z A_z^{-1} B_z \underline{k} = \underline{\alpha}' B A^{-1} B \underline{\alpha} \tag{2.16}$$

Note also that

$$\Lambda = \frac{|A|}{|A+B|} = \frac{|LAL'|}{|L(A+B)L'|} = \frac{|L_1 A L_1'| \; |L_2 A L_2'|}{|L_1 (A+B) L_1'| \; |L_2 (A+B) L_2'|}$$

$$= \prod_{i=1}^{S} (1-r_i^2) \; \frac{|A_z|}{|A_z + B_z|}$$

$$= \Lambda_z \prod_{1}^{S} (1-r_i^2) \; , \tag{2.17}$$

on account of (2.4) and (2.5). Also

$$\underline{k}' B_z \underline{k} = \underline{k}' L_2 B L_2' \underline{k} = \underline{\alpha}' B \underline{\alpha} \tag{2.18}$$

and $k' A_z k = \underline{k}' L_2 A L_2' \underline{k} = \underline{\alpha}' A \underline{\alpha}$ (2.19)

Substituting (2.15), (2.16), (2.17), (2.18), and (2.19) in Λ_{2z} , Λ_{3z} , Λ_{4z} , Λ_{5z} , γ_{2z} and γ_{3z} , we find that these 'new' direction and

collinearity factors or parts are exactly the same as the old ones vis. Λ_2 , Λ_3 , Λ_4 , Λ_5 , γ_2 , γ_3 for $\underline{x}$ and $\underline{y}$, except that Λ must be changed to $\Lambda / \prod_1^S (1 - r_i^2)$, n to n-S , p to p-S and q to q-S.

We are now in a position to apply these results to the analysis of a contingency table, which we do in Part 2 of this paper.

Part 2

3. Association between Two Attributes

Consider a p X q contingency table with the rows corresponding to p categories $a_1, a_2, \cdots, a_p$ of an attribute 'a' and columns to q categories $b_1, b_2, \cdots, b_q$ of another attribute 'b'. Let n_{ij} $(i = 1, \cdots, p, j = 1, \cdots, q)$ be the frequency in the $(i \cdot j)^{th}$ cell. Let $n_{i\cdot}$ $(i = 1, \cdots, p)$ be the row totals and $n \cdot j$ $(j = 1, \cdots, q)$ be the column totals. Let $n = \sum_i n_{i\cdot} = \sum_j n \cdot j$ be the total frequency. We define

$$N = [n_{ij}] \quad (i = 1, \cdots, p : j = 1, \cdots, q) \tag{3.1}$$

$$D_1 = \begin{bmatrix} n_{1\cdot} & & & & & \\ & n_{2\cdot} & & & & \\ & & \cdot & & & \\ & & & \cdot & & \\ & & & & \cdot & \\ & & & & & n_{p\cdot} \end{bmatrix} \tag{3.2}$$

and

$$D_2 = \begin{bmatrix} n_{\cdot 1} & & & & & \\ & n_{\cdot 2} & & & & \\ & & \cdot & & & \\ & & & \cdot & & \\ & & & & \cdot & \\ & & & & & n_{\cdot q} \end{bmatrix} \tag{3.3}$$

The problem of assigning optimum scores to the rows and columns has received considerable attention in the literature (Yates 1948, Fisher 1940, 1950, Maung 1941, Bartlett 1951, Williams 1952). It has been shown that the vectors of optimum scores $\underline{\xi}$ and $\underline{\eta}$ corresponding to the a's and b's are obtainable from the equations

$$\left[\begin{array}{c|c} -r^2 D_1 & N \\ \hline N' & -r^2 D_2 \end{array}\right] \left[\begin{array}{c} \underline{\xi} \\ \hline \underline{\eta} \end{array}\right] = 0 \tag{3.4}$$

If we, therefore, consider two vector variables $\underline{x}(p \times 1)$ and $\underline{y}$ $(q \times 1)$, with the variance-covariance matrix,

$$\left[\begin{array}{c|c} D_1 & N \\ \hline N' & D_2 \end{array}\right] \tag{3.5}$$

it is evident from (3.4) that $\underline{\xi}'\underline{x}$ and $\underline{\eta}'\underline{y}$ are nothing but the canonical variates corresponding to the canonical correlation r^2. In other words, the association between two sets of categories in a contingency table can also be looked upon as a problem of relationship between two vector variables. In general, one set of scores will not be adequate to describe the association between 'a' and 'b' completely. We shall need as many sets of scores, as there are significant canonical correlations between $\underline{x}$ and $\underline{y}$. If, however, only one canonical correlation is significant, one set of scores will be adequate. We say, in this case, that the association is 'linear' or of rank 1.

In the notation of section 1 (of Part 1), $C_{xx} = D_1$, $C_{xy} = N$ and $C_{yy} = D_2$ and hence

$$B = ND_2^{-1}N' = \left[\sum_{j=1}^{q} \frac{n_{ij}n_{hj}}{n_{\cdot j}} \right] \qquad (i,\ h=1,\ \cdots,\ p) \tag{3.6}$$

$$A = D_1 - ND_2^{-1}N' \tag{3.7}$$

$$A + B = D_1 \tag{3.8}$$

We shall denote by A_o, B_o, and D_1^o, the matrices obtained from A, B, and D_1 respectively, by deleting the last row and the last column. It is readily observed from (1.1) that $r^2 = 1$ is a canonical correlation between $\underline{x}$ and $\underline{y}$, the corresponding canonical variates being $x_1 + \cdots + x_p$ and $y_1 + \cdots + y_q$. Obviously, these are irrelevant to our present problem of assigning scores to the a's and b's. We must therefore eliminate these variables and study the residual variates $\underline{z}$ and $\underline{w}$ as in section 2. By taking regression on $\sum_1^p x_i$ and $\sum_1^q y_j$, we can take the new variables to be

$$z_i = x_i - \frac{n_{i\cdot}}{n}(x_1 + \cdots + x_p) \quad ; \quad i = 1,\ 2,\ \cdots p-1 \tag{3.9}$$

$$\text{and} \quad w_j = y_j - \frac{n_{\cdot j}}{n}(y_1 + \cdots + y_q) \quad ; \quad j = 1,\ 2,\ \cdots,\ q-1 \tag{3.10}$$

We can easily calculate C_{zz}, C_{zw}, C_{ww} and hence A_z, B_z from these. They turn out to be

$$A_z = A_o \quad , \quad B_z = B_o - \frac{1}{n}\underline{d}_o\underline{d}_o' \quad , \tag{3.11}$$

where,

$$\underline{d}_o = \begin{bmatrix} n_{1\cdot} \\ n_{2\cdot} \\ \vdots \\ n_{p-1} \end{bmatrix} \quad , \quad \underline{d} = \begin{bmatrix} \underline{d}_o \\ \hline n_{p\cdot} \end{bmatrix} \tag{3.12}$$

Note that

$$|A_z + B_z| = |D_1^o - \frac{1}{n}\underline{d}_o\underline{d}_o'| = n_{1.}n_{2.} \cdots n_{p.}/n \tag{3.13}$$

Consider now the problem of testing the goodness of fit of a set of hypothetical scores $\alpha_1, \alpha_2, \cdots, \alpha_p$ for the rows. The null hypothesis here comprises of two aspects (i) the association between a's and b's is linear and (ii) the true scores corresponding to this linear association are $\alpha_1, \alpha_2, \cdots, \alpha_p$. (i) is the collinearity part and (ii) is the direction part of the null hypothesis.

Since we have eliminated $\sum_1^p x_i$, the assigned function $\underline{\alpha}'\underline{x}$, where $\underline{\alpha}' = [\alpha_1 , \cdots , \alpha_p]$, must - as we noticed in section 2 - be uncorrelated with Σx_i i.e.

$$\underline{d}'\underline{\alpha} = 0 \tag{3.14}$$

On account of this, $\underline{\alpha}'\underline{x}$ can be written, in terms of the residual variables $\underline{z}$ as $\underline{k}'\underline{z}$, where

$$k' = [\alpha_1-\alpha_p , \alpha_2-\alpha_p , \cdots , \alpha_{p-1} - \alpha_p] \tag{3.15}$$

We cannot obtain the 'direction' and 'collinearity' factors straightaway from section 2, in this case, because they involve $|A|$, A^{-1} and these do not exist in the present case, as

$$A\underline{e} = 0 . \tag{3.16}$$

where

$$\underset{1\times p}{\underline{e}'} = [1, 1, \cdots , 1] = [\underline{e}_o'|1] \tag{3.17}$$

and thus A is singular. We must, therefore find the direction and

collinearity factors by directly working with A_z and B_z , especially for Λ_3 , Λ_4 and Λ_5. Λ_2 does not involve A^{-1} and can be written directly.

Partition A, B and $\underline{\alpha}$ as

$$B = \begin{array}{c} \left[\begin{array}{c|c} B_o & \underline{t} \\ \hline \underline{t}' & b_{pp} \end{array}\right] \begin{array}{l} p-1 \\ 1 \end{array} \\ \begin{array}{cc} p-1 & 1 \end{array} \end{array} , \quad A = \left[\begin{array}{c|c} A_z & -\underline{t} \\ \hline -\underline{t}' & a_{pp} \end{array}\right] , \quad \alpha = \left[\begin{array}{c} \underline{\alpha}_o \\ \hline \alpha_p \end{array}\right] \begin{array}{l} p-1 \\ 1 \end{array} \tag{3.18}$$

From (3.16) and (3.18)

$$A_z \underline{e}_o = \underline{t}$$

Let

$$B\,\underline{\alpha} = \underline{f} = \begin{bmatrix} f_1 \\ \vdots \\ f_p \end{bmatrix} = \left[\begin{array}{c} \underline{f}_o \\ \hline f_p \end{array}\right] \tag{3.20}$$

so that $\quad f_i = \sum_{h=1}^{p} \sum_{j=1}^{q} n_{ij} n_{hj} \alpha_h / n_{\cdot j}$

Then $\underline{e}'\underline{f} = \underline{e}'\,B\,\underline{\alpha} = \underline{d}'\underline{\alpha} = 0$ on account of (3.14). The equations

$$A\,\underline{g} = \underline{f} \tag{3.21}$$

in the p unknowns $\underline{g}' = [g_1, \cdots, g_p] = [\underline{g}_o' | g_p]$ are soluble. A solution is

$$\underline{g} = A^{-}\,\underline{f} \tag{3.22}$$

where A^{-} is a pseudo inverse of A(see Rao 1962). But (3.21) and (3.18) yield

$$A_z \underline{g}_o - g_p \underline{t} = \underline{f}_o$$

or

$$\underline{g}_o - g_p A_z^{-1} \underline{t} = A_z^{-1} \underline{f}_o$$

or $$\underline{g}_o - g_p \underline{e}_o = A_z^{-1} \underline{f}_o \quad \text{(on account of 3.19)} \tag{3.23}$$

Also observe that

$$B_z \underline{k} = (B_o - \frac{1}{n} \underline{d}_o \underline{d}_o') \; (\underline{\alpha}_o - \alpha_p \underline{e}_o)$$

$$= B_o \underline{\alpha}_o + \alpha_p \underline{t} = \underline{f}_o \quad , \quad \text{on account of (3.20)}$$

Hence

$$\underline{k}' B_z A_z^{-1} B_z \underline{k} = \underline{f}_o' \, A_z^{-1} \underline{f}_o$$

$$= \underline{f}_o' (\underline{g}_o - g_p \underline{e}_o) \quad , \quad \text{from (3.23)}$$

$$= \underline{f}' \underline{g}$$

$$= \underline{f}' A^- \underline{f}$$

$$= \underline{\alpha}' B \, A^- \, B \, \underline{\alpha} \tag{3.24}$$

Hence Λ_{4z} and Λ_{5z} are the same as Λ_4 and Λ_5 , even if A^{-1} does not exist, provided we use A^- for A^{-1}. Hence the direction and collinearity factors are

$$\Lambda_{2z} = \Lambda_2 = 1 - \frac{\underline{\alpha}' B (A+B)^{-1} B \underline{\alpha} / \underline{\alpha}' B \underline{\alpha}}{\underline{\alpha}' A \underline{\alpha} / \underline{\alpha}' (A+B) \underline{\alpha}}$$

$$= \frac{1 - \sum_{i=1}^{p} \frac{1}{n_{i.}} f_i^2 / \sum_{1}^{p} \alpha_i f_i}{1 - \sum_{i=1}^{p} f_i \alpha_i / \sum_{1}^{p} n_{i.} \alpha_i^2} \tag{3.25}$$

But

$$\Lambda_{3z} = \Lambda_z / \Lambda_{1z} \Lambda_{2z}$$

$$= |A_z|/|A_z+B_z|\Lambda_{1z}\Lambda_{2z}$$

$$= \frac{n|A_z|}{(n_{1.}n_{2.}\cdots n_{p.})(1 - \sum_1^p \frac{1}{n_{i.}} f_i^2 / \sum_1^p \alpha_i f_i)} \tag{3.26}$$

$$\Lambda_{4z} = \frac{n|A_z|}{(n_{1.}n_{2.}\cdots n_{p.})} \quad 1 + \frac{\sum_1^p f_i g_i}{\sum_1^p f_i \alpha_i} \tag{3.27}$$

and
$$\Lambda_{5z} = \frac{\sum_1^p n_{i.}\alpha_i^2}{\Sigma n_{i.}\alpha_i^2 - \Sigma f_i\alpha_i} \cdot \frac{\Sigma\alpha_i f_i}{\Sigma n_{i.}\alpha_i^2 - \Sigma f_i\alpha_i + \Sigma f_i g_i}$$

Λ_{2z} is $\Lambda(n-2, 1, p-2)$. Λ_{3z} is $\Lambda(n-3, q-2, p-2)$, Λ_{4z} is $\Lambda(n-2, q-2, p-2)$ and Λ_{5z} is $\Lambda(n-q, 1, p-2)$. Under the null hypothesis, therefore, from (1.15)

$$-\left\{(n-2) - \frac{1}{2}(1+p-2+1)\right\}\log_e\Lambda_{2z} \text{ is } \chi^2 \text{ with } p-2 \text{ d.f.}$$

and
$$-\left\{(n-3) - \frac{1}{2}(q-2+p-2+1)\right\}\log_3\Lambda_{3z} \text{ is } \chi^2 \text{ with } (p-2)(q-2) \text{ d.f.}$$

They pertain to the direction and collinearity aspects respectively of the goodness of fit test. We can write down similar results for Λ_{4z} and Λ_{5z} of the alternative factorization.

The validity of such tests based on the assumption of normality of $\underline{x}$, for application to discrete data of contingency tables is questionable. Williams (1952) justifies this by an appeal to asymptotic normality and also by the result that elementary symmetric functions of r_i^2 have the same expected values in contingency tables, as for normally distributed $\underline{x}$. The

above tests therefore are approximate but, as pointed out by Williams (1952), adequate for practical purposes, especially when n is large.

In the above analysis, we have used Wilks' Λ as the over-all criterion for testing the association between the two attributes 'a' and 'b'. However, the usual practice, while dealing with contingency tables, is to use the χ^2 test viz., if there is no association

$$\gamma = n\left(\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}^2/(n_{i.}n_{.j}) - 1\right) \tag{3.28}$$

has a χ^2 distribution with (p-1)(q-1) d.f. But (3.28) is nothing but

$$n\, t_r\, B_z(A_z + B_z)^{-1} \tag{3.29}$$

or Pillai's criterion. This can be written, more simply as

$$n[t_r\, B(A+B)^{-1} - 1] \tag{3.30}$$

The quantity subtracted in the larger bracket of (3.30) is the eliminated root $r^2 = 1$, corresponding to $\sum_1^p x_i$.

The 'direction' and 'collinearity' parts, γ_{2z} and γ_{3z} of this over-all χ^2 of (3.29) are easily seen, from (1.16), (1.17) and (3.30), to be

$$\gamma_{2z} = n\left[\frac{\sum_1^p \frac{1}{n_{i.}} f_i^2}{\sum_1^p \alpha_i f_i} - \frac{\sum \alpha_i f_i}{\sum n_{i.}\alpha_i^2}\right], \quad \text{d.f. } (p-2) \tag{3.31}$$

and

$$\gamma_{3z} = \gamma - \frac{n \sum f_i^2/n_{i.}}{\sum \alpha_i f_i}, \quad \text{d.f. } (p-2)(q-2) \tag{3.32}$$

Under the null hypothesis, they have χ^2 distributions, for large n.

Williams (1952) has given the test of goodness of fit of a set of hypothetical scores, only for the particular cases q = 2, 3. We have here the tests for any p and q. Further, we have also given the tests, based on the alternative criterion (Pillai), which in this case is the usual χ^2 of a contingency table and is thus more in tune with the classical method of partitioning an over-all χ^2, corresponding to suspected sources of association.

REFERENCES

Bartlett, M. S. (1951). "The goodness of fit of a single hypothetical discriminant function in the case of several groups," Annals of Eugenics, 16, 199.

Fisher, R. A. (1940). "The precision of discriminant functions," Annals of Eugenics, 10, 422.

Fisher, R. A. (1950). Statistical Methods for Research Workers, 11th ed., Edinburgh, Oliver and Boyd.

Kshirsagar, A. M. (1964). "Distribution of the direction and collinearity factors in discriminant analysis," Proceedings of the Cambridge Philosophical Society, 60, 217.

Kshirsagar, A. M. (1969). "Correlation between two vector variables," unpublished.

Muang, K. (1941). "Measurement of associations in a contingency table with special reference to the pigmentation of hair and eye colour of Scottish school children," Annals of Eugenics, 11, 189.

Pillai, K. C. S. (1955). "Some new test criteria in multivariate analysis," Annals of Mathematical Statistics, 26, 117.

Rao, C. Radhakrishna (1962). "A note on a generalized universe of a matrix with applications to problems in mathematical statistics," Journal of the Royal Statistical Society, B, 24, 152-158.

Wilks, S. S. (1932). "Certain generalizations in the analysis of variance," Biometrika, 24, 471.

Williams, E. J. (1952). "Use of scores for the analysis of association in contingency tables," Biometrika, 39, 274.

Williams, E. J. (1955). "Significance tests for discriminant functions and linear functional relationships," Biometrika, 42, 360.

Williams, E. J. (1967). "The analysis of association among many variates," Journal of the Royal Statistical Society, B, 29, 199.

Yates, F. (1948). "The analysis of contingency tables with groupings based on quantitative characters," Biometrika, 35, 176.