CATEGORICAL DATA ANALYSIS FOR SIMPLE RANDOM SAMPLES

Avinash C. Singh
Memorial University of Newfoundland

ABSTRACT

Frequently in practice, there is uncertainty about the stability of the finite sample behavior of Pearson-Fisher's X^2 statistic with regard to the χ^2 approximation. We describe a method for checking instability in X^2 by means of an objective and a systematic way of dimensionality reduction analogous to collapsing categories. We then propose a suitable modification to X^2 which is obtained by optimally reducing the dimension of the observed count vector using principal components and subsequently constructing an optimal test based on the transformed data. Some illustrative examples are discussed.

1. INTRODUCTION

It is known from the simulation studies (see e.g. Yarnold 1970, Larntz 1978, Koehler and Larntz 1980) that even if some of the expected cell counts are very small, the Pearson-Fisher's X^2 statistic for testing fit may be quite stable in the sense of finite sample behaviour of Type I error rate in comparison to its χ^2 approximation. There are some general guidelines on the minimum expected frequency such as those suggested by Cochran (1954) and Yarnold (1970) which are helpful in practice to decide whether the χ^2 approximation may be deemed appropriate. If necessary, the number of categories is reduced by collapsing together a few cells. The interpretation of these guidelines can often be affected by subjective considerations of the user in matters such as when to collapse and with whom, especially in the case of multidimensions. An alternative approach would be to use some other approximation to X^2 for sparse contingency tables as considered by Cochran (1952), Morris (1975), McCullagh (1985), Koehler (1986) and others. However, chi-square approximation to X^2 in the presence of small counts seems to be widely used by many researchers and practitioners in various substantive fields. It would, therefore, be useful to have an objective way for checking instability in X^2 and a systematic way of collapsing if necessary. In this paper we attempt to provide an answer to this problem. A version of X^2 test suitable for the proposed method of collapsing is also given.

Consider testing fit of a model for categorical data arising from simple random samples i.e. multinomial or product multinomial sampling. We will assume multinomial for illustration purposes. With cross-classified data and for moderate samples, it often turns out that some (estimated) expected cell counts for the model under consideration are very small (e.g. 1 or so). For these situations, a solution for the problem of an objective method of collapsing is proposed. By viewing collapsing into T categories as a special case of dimensionality reduction, it is seen that the first T principal components of the observed count vector do provide a systematic and an optimal method of dimensionality reduction (or collapsing in a general sense) in the class of all linear transformations. We then propose a modified X^2 to be denoted by X_T^2 which is an optimal test statistic in the class of tests based on the given set of T principal components. The test X_T^2 coincides with the usual X^2 when there is no reduction in dimension. It may be noted that the proposed method also applies to Neyman's modified X^2 as well by using the appropriate estimated covariance matrix for computing principal components. We also present a simple test for instability by means of the difference $X^2 - X_T^2$. This would be useful in practice in order to decide whether one should collapse or not.

The test X_T^2 is expected to provide a robust alternative to the usual X^2 in the sense that only a slight loss in power (when T is close to the rank of the covariance matrix) is expected if X_T^2 were used even in the absence of the problem of instability. The proposed test X_T^2 was described earlier in a technical report (Singh 1985). In section 2, we describe a class of generalized collapsing transformations and in section 3, a statement of the problem and the proposed test are given. Section 4 contains some theoretical results used in the paper. Two illustrative examples are presented in section 5 and finally concluding remarks in section 6.

2. A GENERALIZED COLLAPSING TRANSFORMATION

Let A denote a linear transformation matrix of order $T \times k$ which would reduce the dimension k of the observed vector of cell counts to T. We shall refer to the matrix A as a generalized collapsing transformation. Notice that the usual methods of collapsing and deletion of cells can be seen as special cases of the above transformation.

It is known that principal components provide an optimal dimensionality reduction (eg. Rao 1973, p. 592). Therefore, an optimal choice of A in the class of generalized collapsing transformations is given by

$$A = M_T' = (P_1, P_2, \ldots, P_T)' \tag{2.1}$$

where P_i's are normalized eigen vectors corresponding to eigen values $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_k$ of the estimated covariance matrix

$$\hat{\Gamma} = D_{\hat{\pi}} - \hat{\pi}\hat{\pi}' . \tag{2.2}$$

Here $D_{\hat{\pi}}$ denotes a diagonal matrix with diagonal elements given by the k-vector of estimated expected proportions. It may be noted that $\hat{\lambda}_k = 0$

because the rank of $\hat{\Gamma}$ is only $k - 1$.

In order to use the optimal A, a suitable value of T is required. We propose the following method for choosing T. First set a very small non-negative ε (e.g. .01 or .005 can be used as working values). We shall refer to ε as the prescribed level of collapsing. Then we compute T from relative cumulative eigenvalues as

$$T = \max\{u: \Sigma_{i=u}^{k} \lambda_i / \Sigma_{i=1}^{k} \lambda_i \geq \varepsilon\} \tag{2.3}$$

In practice one can reasonably specify ε so that the proportion of the total variation accounted by the $k - T$ non-principal components is negligible. This would ensure a very little loss of information due to data transformation. It may be remarked that our purpose of using principal components is completely opposite of the traditional one. This is because we are interested in dropping only a few non-principal components rather than retaining a few principal ones. For a given ε, the transformation defined by (2.1) and (2.3) provides a systematic and objective method of dimensionality reduction. As ε increases, we get a hierarchy of collapsing transformations.

An intersting interpretation of the optimal transformation M_T is derived from the following observation. The observed vector (p, say) of cell proportions is in the column space of Γ (with probability one, see e.g. Moore 1978). Therefore, we can represent p as

$$p = \lambda_1 P_1 P_1' z + \lambda_2 P_2 P_2' z + \ldots \lambda_k P_k P_k' z \tag{2.4}$$

for some z in R^k . Let q denote the collapsed vector p i.e.

$$q = M_T' p . \tag{2.5}$$

It is seen from (2.4) that if the dropped eigen values are relatively very small, then we can reasonably well reconstruct p from q. The reconstructed p would be a smoothed version of p in which all entries are perturbed a bit. This shows the difference between the optimal versus usual collapsing.

The collapsing transformation based on eigen values of Γ is applicable in general for any sample design (see Singh and Kumar, 1986, for an application to complex surveys). For multinomial Γ, the eigen values are directly related to the proportion vector π in terms of the following inequality (a proof is given in section 4).

$$\pi_{(i+1)} \leq \lambda_i \leq \pi_{(i)}, \; i = 1, \ldots k, \; \pi_{k+1} = 0 \tag{2.6}$$

where $\pi_{(i)}$ denotes the i^{th} largest element of π.

The relation (2.6) implies that the instability problem caused by small proportions $\pi_{(i)}$'s is equivalent to the problem caused by small eigen values for multinomial Γ.

3. THE PROBLEM AND THE PROPOSED TEST

Consider the hypothesis of whether the variation among the observed proportions $p(k \times 1)$ fits a model specifying the probability vector π by a 'smooth' link function which is linear in $r(r < k)$ parameters θ. In other words, consider testing

$$H_0: h(\pi) = X\theta \text{ versus } K_0: \text{otherwise} \tag{3.1}$$

where h is a continuously differentiable function such that its inverse exists and X is a known $k \times r$ matrix of full rank r. The usual loglinear and logit models are special cases of the above H_0.

An asymptotically optimal test of H_0 based on p is given by rejecting for large values of the quadratic form (see Lehmann, 1959, pp. 304-313).

$$Q(\theta^*) = n(p - \pi(\theta^*))' \hat{\Gamma}^-(p - \pi(\theta^*)) \tag{3.2}$$

where θ^* is a minimum X^2 type estimator under H_0 when the metric is defined by $y'\hat{\Gamma}^- y$. Here $\hat{\Gamma}^-$ denotes a consistent estimate of a g-inverse of Γ. The asymptotic null distribution of the statistic (3.2) is χ^2_{k-r-1} which follows from the result

$$p \overset{a}{\sim} MVN(\pi, \Gamma/n) . \tag{3.3}$$

Here Γ is $D_\pi - \pi\pi'$, n is the total multinomial sample size and the symbol "$\overset{a}{\sim}$" denotes "asymptotically distributed as".

One can substitute any asymptotically equivalent estimator for θ in (3.2); for instance, the maximum likelihood estimate (mle) $\hat{\theta}$ obtained from the cell proportion p. The Pearson-Fisher's X^2 can be obtained as a special case of (3.2) when D_π^{-1} is used as a g-inverse and π is esimated by $\pi(\hat{\theta})$. That is, denoting $\pi(\hat{\theta})$ by $\hat{\pi}$, we have

$$\begin{aligned} X^2 &= n(p - \hat{\pi})' D_{\hat{\pi}}^{-1} (p - \hat{\pi}) \\ &= n \, \Sigma_{i=1}^{k} (p_i - \hat{\pi}_i)^2 / \hat{\pi}_i \end{aligned} \tag{3.4}$$

Now we can motivate the proposed test X_T^2. Suppose that it is suspected that some of the expected proportions $\hat{\pi}_i$'s are too small which might cause instability in X^2. This implies that there is a lack of sufficient observations to draw any meaningful conclusion about the discrepancy $(p_i - \pi_i)$ in those cells. We can either withhold our decision until more observations are obtained or we may wish to test for an overall pattern in π after removing heavy influence of a few cells. The latter is achieved by performing a test of H_0 based on q (collapsed p for a given ε). We assume that we are willing to sacrifice some information, the contribution of which is thought to be unreliable in the test statistic X^2 .

We thus propose the following test for H_0 in presence of the instability problem. For a given $\varepsilon > 0$, q denotes collapsed p such that $T > r$. Then the test X_T^2 rejects H_0 for large values of the quadratic form

$$\begin{aligned} X_T^2(\theta^{**}) &= n(p - \pi(\theta^{**}))' \Delta_T (p - \pi(\theta^{**})) \\ &= n(q - \psi(\theta^{**})' D_{\hat{\lambda}}^{-1} (q - \psi(\theta^{**})) \\ &= n \, \Sigma_{i=1}^{k} [P_i'(p - \pi(\theta^{**}))]^2 / \hat{\lambda}_i , \end{aligned} \tag{3.5}$$

where θ^{**} is a minimum X^2 type estimator under H_0 for the metric defined by $Y'\Delta_T Y$, the $k \times k$ matrix Δ_T is $\Sigma_{i=1}^{T} P_i P_i' / \hat{\lambda}_i$ i.e. a truncated g-inverse of $\hat{\Gamma}$ and $D_{\hat{\lambda}}$ is $\mathrm{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_T)$. The transformed parameter ψ denotes $M_T' \pi$ just as q denotes $M_T' p$.

The test statistic X_T^2 is similar to X^2 except that the matrix of the metric is modified to Δ_T. Hence θ^{**} would be different from θ^*. The asymptotic null distribution of X_T^2 is χ_{T-r}^2 which follows from the result

$$q = M_T' p \stackrel{a}{\sim} \mathrm{MVN}(\psi = M_T' \pi, \mathrm{diag}(\lambda_1, \ldots, \lambda_T)) \tag{3.6}$$

and the fact that the variable T, although random, can be regarded as fixed for our asymptotics when ε is slightly modified (see section 4).

Usually the mle $\hat{\theta}$ based on the original data p would be easily available. Then the asymptotically equivalent version of $X_T^2(\theta^{**})$ can be obtained by subtracting a correction term. We get

$$X_T^2(\hat{\theta}) = n(p-\hat{\pi})' \Delta_T (p-\hat{\pi}) - Y_T^2 \tag{3.7}$$

where B is a $k \times r$ matrix of derivatives $(\partial\pi/\partial\theta)$ and the terms Y_T^2 is $n(p-\hat{\pi})'\Delta_T B(B'\Delta_T B)^{-1}B'\Delta_T(p-\hat{\pi})$.

The correction term in (3.7) is similar to Dzhaparidze and Nikulin's (1974) modification to X^2 for Chernoff-Lehmann problem when raw data mle is used instead of grouped data mle. It may be noted that any root n-consistent estimate of θ can be used in (3.7) without violating the asymptotic equivalence.

It can be seen from section 4 that when T is maximum (i.e. $k - 1$ for $\varepsilon = 0$), then Δ_{k-1} is indeed a g-inverse of Γ and we get the usual X^2. That is

$$X_{k-1}^2 (\hat{\theta}) = X^2 \tag{3.8}$$

In fact, X_T^2 can be seen to be asymptotically optimal for testing H_0 in the class of tests based on q. Thus, with the optimal collapsing transformation for a small ε, the test X_T^2 is expected to be robust (with respect to X^2) in absence of instability problem. However, in presence of this problem, the test X_T^2 is expected to control increase in Type I error rate. This is further ensured by the observation that X_T^2 is indeed a conservative (asymptotically) test for H_0. To see this, note that in the class of tests based on q, the original hypothesis H_0 concerning k-dimensional π is reduced to testing a hypothesis H_0' concerning T-dimensional parameter ψ, i.e.

$$H_0' : \psi = M_T' h^{-1} (X\theta). \tag{3.9}$$

Since $H_0 \subset H_0'$, it follows that X_T^2 having a prescribed size α for H_0' would be conservative for H_0.

It would be useful in practice to use the difference $X^2 - X_T^2$ ($\stackrel{a}{\sim} \chi_{k-T-1}^2$) as an objective means of checking instability in X^2 for a given $\varepsilon > 0$. If for a small ε, the difference is deemed significant with respect to upper α point of χ^2 distribution with k-T-1 d.f., then the use of X_T^2 is recommended over X^2.

The test X_T^2 can be easily applied to the case of nested models. Suppose that $X = (X_1, X_2)$ and $\theta = (\theta_1, \theta_2)$ are partitioned in t and u columns $(t + u = r)$ and H_1 is specified by the condition $\theta_2 = 0$. Then X_T^2 for H_1 given that H_0 is accepted, is given by rejecting for large values of the quadratic form

$$X_T^2(H_1|H_0) = (p - \hat{\hat{\pi}})' \Delta_T (A_0 - A_1)\Delta_T(p - \hat{\hat{\pi}}) \stackrel{a}{\sim} \chi_u^2 \tag{3.10}$$

where A_0 and A_1 denote the matrices $B(B'\Delta_T B)^{-1}B'$ computed under H_0 and H_1 respectively, and $\hat{\hat{\pi}}$ is a root n-consistent estimate (e.g. mle) of π under H_1.

4. SOME THEORETICAL RESULTS

In this section we collect some results used in this article. The first proposition shows the relation between π_i's and λ_i's (eigen values of multinomial Γ i.e. $D_\pi - \pi\pi'$).

<u>Proposition 4.1.</u> Let $\pi_{(1)} \geq \pi_{(2)} \geq \ldots \geq \pi_{(k)}$ be the elements of π arranged in decreasing order. Then, for $\pi_{(k+1)} = 0$, we have

$$\pi_{(i+1)} \leq \lambda_1 \leq \pi_{(i)}, \; i = 1, \ldots, k \tag{4.1}$$

To prove (4.1) we need the following lemma from Anderson and Dasgupta (1963).

<u>Lemma 4.1.</u> Let A and B be symmetric matrices of order m. If $\lambda_i(C)$ denotes the ith rank eigen value of C, then for $j + \ell \leq i + 1$,

$$\lambda_i(A + B) \leq \lambda_j(A) + \lambda_\ell(B) \tag{4.2}$$

$$\lambda_{m-i+1}(AB) \geq \lambda_{m-j+1}(A) \; \lambda_{m-\ell+1}(B) \tag{4.3}$$

Now apply the inequality (4.2) first by setting $A = D_\pi$, $B = -\pi\pi'$ and $m = k$ to show that $\lambda_i(D_\pi - \pi\pi') \leq \lambda_i(D_\pi) = \pi_{(i)}$. Next apply (4.3) by setting $A = D_\pi$ and $B = I - \underset{\sim}{1}\, \pi'$ to show that $\lambda_{m-i+1}(D_\pi - \pi\pi') \geq \lambda_{m-i+2}(D_\pi)$ which implies that $\lambda_i(D_\pi - \pi\pi') \geq \pi_{(i+1)}$. Hence, the result

The next proposition shows that the random variable T defined by (2.3) can be regarded as fixed for our asymptotics when ε is slightly modified.

Proposition 4.2. Let ε be modified to ε^* defined by

$$\varepsilon^* = \varepsilon - \beta_n,\ 0 < \beta_n < \varepsilon,\ \beta_n \downarrow 0 \text{ but } \sqrt{n}\beta_n \to \infty. \quad (4.4)$$

Then, $T(\varepsilon^*) \to T_0(\varepsilon)$ in prob. as $n \to \infty$ (4.5)

where T_0 is T corresponding to the true Γ.

The proof is similar to the one given in Kulperger and Singh (1982) and Singh and Kumar (1986) provided that the following condition is satisifed for $i = 1,2,3, \ldots, k$,

$$\sqrt{n}|\hat{\lambda}_i - \lambda_i| = O_p(1) \quad (4.6)$$

The above condition is true in view of the previous proposition. The term β_n in ε can be chosen as $(\log n/n)^{1/2}\varepsilon$ which implies that the modification term β_n will be negligible for large n.

Henceforth we shall regard Γ as fixed asymptotically assuming β_n negligible. The next proposition shows the optimality of X_T^2.

Proposition 4.3. For the testing problem H_0 vs K_0 given by (3.1), the test X_T^2 is asymptotically UMPI (uniformly most powerful invariant) in the class of tests based on q (collapsed p as given in (3.6)). Moreover, under H_0,

$$X_T^2 \stackrel{a}{\sim} \chi^2_{T-r} \quad (4.7)$$

To prove the above proposition, first note that in the class of tests based on q, the testing problem is reduced to the modified hypothesis H_0' of (3.9). The result then follows by considering the asymptotic reduction of the problem for local alternatives to that of testing a linear hypothesis for Gaussian case (see Lehmann, 1959, p. 304). The test X_T^2 can also be obtained as a generalized score test (see Singh, 1986; also Cox and Hinkley, 1974, 321-324) and that it is optimal in the general sense of Wald (1943).

The proofs for distribution of $X^2 - X_T^2$ (used in testing instability) and $X_T^2(H_1|H_0)$ of (3.10) are similar to those given in Singh and Kumar (1986). The final proposition shows that the usual X^2 can be obtained as a special case of X_T^2 when T is k - 1, the rank of Γ.

Proposition 4.4. For T = k - 1, X_T^2 coincides with X^2.

First we show that for $B(=\partial\pi/\partial\theta)$ and $\Gamma(=D_\pi - \pi\pi')$ evaluated at the mle $\hat{\theta}$, the second term of (3.7) vanishes when T = k - 1, i.e.

$$B'\Delta_{k-1}(p - \hat{\pi}) = 0 \quad (4.8)$$

It is enough to show that $B'\Delta_{k-1}(p - \hat{\pi}) = B'D_{\hat{\pi}}^{-1}(p-\hat{\pi})$ because RHS is the score vector for θ under mulinomial. Note that from the spectral decomposition of Γ, we have

$$(D_\pi - \pi\pi')\Delta_{k-1} = P_1P_1' + \ldots + P_{k-1}P_{k-1}' \quad (4.9)$$

Next using the fact that $\underset{\sim}{1}$ is orthogonal to $D_\pi - \pi\pi'$ (which implies $\underset{\sim}{1}$ is orthogonal to $P_1, P_2, \ldots P_{k-1}$), we get

$$P_1P_1' + \ldots + P_{k-1}P_{k-1}' = I - \underset{\sim}{1}\,\underset{\sim}{1}'/k \quad (4.10)$$

Now writing $B'\Delta_{k-1}(p - \hat{\pi})$

$$= B'\Delta_{k-1}(D_\pi - \pi\pi')\Delta_{k-1}(p - \hat{\pi})$$

$$= B'\Delta_{k-1}(D_\pi - \pi\pi')D_\pi^{-1}(D_\pi - \pi\pi')\Delta_{k-1}(p - \hat{\pi})$$

$$= B'(I - \underset{\sim}{1}\,\underset{\sim}{1}'/k)'\,D_\pi^{-1}(I - \underset{\sim}{1}\,\underset{\sim}{1}'/k)(p - \hat{\pi})$$

$$= B'D_\pi^{-1}(p - \hat{\pi})$$

we can establish (4.8). Hence, the proposition.

5. EXAMPLES

We present two numerical examples for illustrating applications of X_T^2. In the first one, one would suspect the presence of instability problem due to a cell with very small expected count. It is seen that indeed X_T^2 (for $\varepsilon = .01$) accepts H_0 whereas X^2 rejects. In the second example, there does not seem any instability problem because there is no cell with very small expected count. As expected, both X^2 and X_T^2 are in agreement in rejecting H_0.

Example 5.1 Lizard Data.

We consider the problem as given in Fienberg (1978, p. 32) based on the data of Shoener (1968). Ecologists studying lizards are often interested in relationships among the variables that can be used to describe the Lizard's habitat. Table 1 contains counts for structural hatitat categories for Anolis lizards of Bimini: Sagrei Adult Males vs Angusticeps Adult Males. A total of 192 lizards were observed and for each, the perch height (variable 1 in feet, category 1 = high or '>5', 2 = Low or '≤5'), the perch diameter (variable 2 in inches, category 1 = wide or '>2.5', 2 = narrow or '≤2.5') and lizard species (variable 3, category 1 = Sagrei, 2 = Angusticeps) were recorded.

Table 1. OBSERVED AND EXPECTED COUNTS

Cell	Observed	Expected
(1,1,1)	15	12.6
(1,1,2)	21	19.566
(1,2,1)	18	20.4
(1,2,2)	1	2.444
(2,1,1)	48	50.4
(2,1,2)	3	4.444
(2,2,1)	84	81.6
(2,2,2)	2	.556

Note: cell (i,j,k) refers to ith category of variable 1, jth of variable 2, and kth of variable 3.

Consider H_0: $u_{12} = u_{123} = 0$ (see Fienberg 1978, p. 38) i.e. the hypothesis of conditional independence of variables 1 and 2 given variable 3. The mle $(n\hat{\pi})$ under H_0 are given in Table 1 and the values of X^2, G^2 and Neyman's X_N^2 are computed as 6.11, 4.88 and 4.82 respectively. When compared with $X^2_{.05,2d.f.}$ (=5.99), we see that X^2 exceeds it while G^2 and X^2_N do not. All of them exceed $X^2_{.1,2}$(=4.61). Thus there is some question as to whether the conditional independence model fits (Fienberg 1978, p. 39). Consider contributions from the eight cells to X^2 which are given below in the same order as that in Table 1.

$X^2 = .457 + .107 + .282 + .853 + .114 + .469 + .071 + 3.75$

One can see that the cell (2, 2, 2) has a rather large contribution although one would perhaps hardly notice this lack of fit in the visual inspection of table 1 for agreement between observed and expected counts. Both expected and observed counts in cell (2, 2, 2) are so small that more observations are needed in order to draw any reasonable conclusion about lack of fit in that cell. The test X^2_T on the other hand, shows strong evidence in favour of H_0. Its computation is shown below. Let $\varepsilon = .01$. For $\Gamma(=D_\pi - \pi\pi')$ evaluated at $\pi = \hat{\pi}$, the eigen values and the relative cumulative eigen values are given in Table 2. The value of T for $\varepsilon = .01$ is obtained as 6. Here the modification factor β_n of (4.4) is .0016 which does not alter the value of T.

Table 2: EIGEN VALUES ($\lambda_t \times 10^3$) IN DECREASING ORDER AND THE CORRESPONDING RELATIVE CUMULATIVE EIGEN VALUES ($\Sigma_{i=1}^t \lambda_i \times 10^2$) FOR Γ/n WHEN $\pi = \hat{\pi}$.

1	2	3	4	5	6	7	8
1.753	.860	.542	.381	.142	.075	.017	0
46.5	69.3	83.7	93.8	97.6	99.5	100	100

Now, we can write H_0 as $h(\pi) = X\theta$ where h is the log function, X is a known 8×6 matrix with entries either 1 or -1, and θ is a 6-vector of u-parameters in the standard log-linear model. However, in this setup, θ-parameters are not functionally independent due to the constraint $\pi'\underset{\sim}{1} = 1$. For computing the matrix $B(=\partial\pi/\partial\theta)$ it would be convenient to write H_0 as

$$\log \pi = u(\tilde{\theta})\underset{\sim}{1} + \tilde{X}\tilde{\theta} \tag{5.1}$$

where θ is a 5-vector of independent parameters given by $(u_{1(1)}, u_{2(1)}, u_{3(1)}, u_{13(11)}, u_{23(11)})$, $u(\theta)$ is a normalizing factor in view of the constraint $\pi'\underset{\sim}{1} = 1$ and $\tilde{X}$ is a 8 x 5 matrix of full rank 5 as follows.

$$\tilde{X} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 1 & -1 \\ 1 & -1 & -1 & -1 & 1 \\ -1 & 1 & 1 & -1 & 1 \\ -1 & 1 & -1 & 1 & -1 \\ -1 & -1 & 1 & -1 & -1 \\ -1 & -1 & -1 & 1 & 1 \end{bmatrix} \tag{5.2}$$

It is easy to see that matrix B is given by

$$(\partial\pi/\partial h)(\partial h/\partial\theta) = D_\pi(I - \underset{\sim}{1}\pi')\tilde{X} = (D_\pi - \pi\pi')\tilde{X} \tag{5.3}$$

Now using formula (3.7) with T = 6 and B as in (5.3) evaluated at $\pi = \hat{\pi} = \pi(\hat{\theta})$, we get

$$X^2_6(\hat{\theta}) = 2.001 - 1.047 = 0.954 \tag{5.4}$$

which compared with $X^2_{.3,1\ d.f.}$(=1.07) shows very strong evidence in favour of H_0. For checking instability, we also compute for T = 7

$$X^2_7(\hat{\theta}) = 6.11 = X^2 \tag{5.5}$$

Clearly, there is an unusually large amount of contribution by adding the eigen value $\hat{\lambda}_7$ by referring the difference $X^2 - X^2_6$ to a χ^2_1 distribution. This indicates presence of instability problem and hence X^2_6 should be preferred over X^2. A similar result is obtained when $\hat{\Gamma}$ is computed with $\hat{\pi} = p$. In this case at T = 6, the relative cumulative eigen value (in %) is 99.2 and at T = 5, it is 97.6, so we again use $X^2_6(\hat{\theta})$ for the new $\hat{\Gamma}$. Note that X^2_T is not defined for T < 6 because r = 5. So for Γ estimated by observed proportions p, we have

$$X^2_6(\hat{\theta}) = 2.521 - 1.633 = 0.888, \text{ and}$$

$$X^2_7(\hat{\theta}) = 4.82 = X^2_N \tag{5.6}$$

We thus conclude that the hypothesis of conditional independence of perch height and perch diameter given species is consonant with the given data.

Example 5.2 Detergent Preference Data.

Consider the data of Ries and Smith (1963) given in Fienberg (1978, p. 59) in the form of a cross-classification of 1008 consumers according to four variables. The variables (1) "softness": the softness of the laundry water used, (2) "use": the previous use of detergent brand M, (3) temperature": the temperature of the laundry water used, (4) "preference": the preference for detergent brand X over M in a consumer blind trial. The table is of dimension 3 x 2 x 2 x 2. We will illustrate computation of X^2_T for the complete independence model in the log-linear framework, i.e.,

$$H_0\text{: all u's are zero except } u, u_1, u_2, u_3, u_4. \tag{5.7}$$

For this model, it can be seen from mle of expected cell counts that there is no reason to suspect instability in X^2. The values of X^2 and G^2 are respectively 43.9 and 42.9 with 18 d.f. indicating rejection of H_0.

To compute X^2_T, we first write H_0 in the form (5.1) for convenience as in example 5.1. Here in terms of u parameters, the five independent θ parameters are $\theta_1 = u_{1(1)}$, $\theta_2 = u_{1(2)}$, $\theta_3 = u_{3(1)}$, $\theta_4 = u_{3(1)}$, $\theta_5 = u_{4(1)}$. The matrix $\tilde{X}$ is a 24 x 5 matrix with entries either -1 or 0 or 1 and can be written in a manner analogous to $\tilde{X}$ of (5.2). To compute T, the eigen values of Γ (evaluated at mle $\hat{\pi}$) are computed. It is seen that the relative cumulative eigen value for decreasing order of λ's is 97% at T = 22 and 100% at T = 23 (rank of Γ). Thus for $\varepsilon < .03$, T = 23 and X_T^2 coincides with X^2 of 43.9. The value of X_T^2 at T = 22, is found to be (using mle $\hat{\pi}$)

$$X_{22}^2(\hat{\theta}) = 43.759 - .012 = 43.747 \tag{5.8}$$

Thus $X_{22}^2(\hat{\theta})$ (at 22 - 5 = 17 d.f.) also favours rejection of H_0 as does X^2. This is expected in the absence of instability problem which is indicated by insignificance of the difference $(X^2 - X_{22}^2)$ at 1 d.f.

6. CONCLUDING REMARKS

An objective, systematic and optimal method of collasping the count vector is provided by the well-known multivariate technique of principal components. For a suitable choice of the level ε of collapsing, the proposed test X_T^2 is expected to have a stable finite sample behaviour. Moreover for small ε it is expected to be robust in view of its optimality properties. It would be noted that the test X_T^2 would have no power in detecting departures from H_0 in the direction of deleted eigen vectors. However, the rationale for using X_T^2 is that the evidence at hand for these departures is unreliable due to insufficient data and that we are willing to sacrifice this information in order to overcome the instability problem which would arise in X^2 had we used the original data. Obviously, performance of X_T^2 depends on ε. For further research, it would be desirable and it is proposed to perform a simulation study on the level and power of X_T^2 in comparison to other tests for various choices of ε analogous to Koehler and Larntz (1980).

The test X_T^2 is recommended whenever one is uncertain about the stability of X^2. The difference $X^2 - X_T^2$ can be used as a test for instability corresponding to a given level ε of collapsing. We have interpreted collapsing in the general sense of dimensionality reduction. It might be noted that for the usual method of collapsing (cf. Cochran, 1954), there does not seem available an analogous test for instability.

Finally we remark that the problem of instability as defined in this paper might also arise in using weighted least squares (WLS) test of Grizzle, Starmer, and Koch (1969). The WLS test is also subject to instability due to inversion of covariance matrix required in its computation and when the estimated covariance matrix is nearly singular. It would be useful to extend the proposed method of modifying X^2 in order to develop another version of WLS test.

ACKNOWLEDGEMENT

This research was supported by Statistics Canada and Natural Sciences and Engineering Research Council of Canada. Thanks are due to S. Kumar, Social Survey Methods Division, Statistics Canada, for his assistance in numerical computations.

REFERENCES

Anderson, T. W. and Dasgupta, S. (1963). Some inequalities on characteristic roots of matrices. Biometrika 50, 522-524..

Cochran, W. G. (1952). The χ^2 test of goodness-of-fit, Ann. Math. Statist., 23, 315-345.

Cochran, W. G. (1954). Some methods for strengthening the common χ^2 tests. Biometrics, 10, 417-451

Dzhaparidze, K. O. and Nikulin, M. S. (1974). On a modification of the standard statistic of Pearson. Theor. Prob. and Appl. 19, 851-853.

Fienberg, S. E. (1978). The Analysis of Cross-Classified Categorical Data, The MIT Press, Cambridge.

Grizzle, J. E., Starmer, C. F. and Koch, G. G. (1969). Analysis of Categorical data by linear models. Biometrics 25, 489-504.

Koehler, K. J. (1986). Goodness-of-fit test for log-linear models in sparse contingency tables. Jour. Amer. Statist. Assoc., 81, 483-493.

Koehler, K. J. and Larntz, K. (1980). An empirical investigation of goodness-of-fit statistics for sparse multinomials. Jour. Amer. Statist. Assoc. 75, 336-344.

Kulperger, R. J. and Singh, A. C. (1982). On random grouping in goodness-of-fit tests for discrete distributions. Journal of Statistical Planning and Inference 7, 109-115.

Larntz, K. (1978). Small-sample comparisons of exact levels for chi-squared goodness-of-fit statistics. Jour. Amer. Statist. Assoc. 73, 253-263.

Lehmann, E. L. (1959). Testing Statistical Hypotheses, N.Y. John Wiley.

Moore, D. S . (1978). Chi-square tests. In Studies in Statistics, Ed. R.V. Hogg, Amer. Math. Assoc. 66-106.

McCullagh, P. (1985). On the asymptotic distribution of Pearson's statistic in linear exponential-family models. Int. Statist. Rev., 53, 61-67.

Morris, C. (1975). Central limit theorem for multinomial sums. Ann. Statist. 3, 165-188.

Rao, C. R. (1973). Linear Statistical Inference and Its Applications, 2nd ed, N.Y., John Wiley.

Ries, P. N. and Smith, H. (1963). The use of chi-square for preference testing in multidimensional problems. Chemical Enginerring Progress 59, 39-43.

Schoener, T. W. (1968). The anolis lizards of Bimini: resource partitioning in a complex fauna. Ecology, 49, 704-726.

Singh, A. C. (1985). On optimal asymptotic tests for analysis of categorical data from sample surveys. Statistics Canada Methodology Branch, Working Paper No. SSMD 86-002.

Singh, A. C. (1986). On the optimality and a generalization of Rao-Robson's statistic. To appear in Commun. Statist. - Theo. Meth.

Singh, A. C. and Kumar, S. (1986). Categorical data analysis for complex surveys. Proc. Amer. Statist. Assoc., Survey Research Methods Section.

Yarnold, J. K. (1970). The minimum expectation in X^2 goodness-of-fit test and the accuracy of approximations for the null distribution. Jour. Amer. Statist. Assoc. 65, 864-886.