Analysis of Singly Ordered Two-way Contingency Tables

J.C.W. RAYNER john_rayner@uow.edu.au
School of Mathematics and Applied Statistics, University of Wollongong, Northfields Avenue, NSW 2522, Australia

D.J. BEST john.best@cmis.csiro.au
CSIRO Mathematical and Information Sciences, PO Box 52, North Ryde, NSW 1670, Australia

Abstract. Pearson's statistic is investigated for nominal-ordinal two-way contingency tables in which we wish to test for identical rows. The statistic is expressed as a sum, the first summand of which is a statistic given by Yates (1948), and examines location effects for the nominal category. The second and subsequent summands reflect the corresponding moments: for example, dispersion, skewness, kurtosis etc. The summands are shown to be weakly optimal in that they are score statistics.

Keywords: Categorical data, logistic model, log-linear model, orthonormal polynomial, two-way data.

1. Introduction

In the social sciences and in more specialized areas such as sensory evaluation, it is common to obtain categorized ratings for a number of items. For example, if eyesight is under consideration, one might have five categories of eyesight ranging from poor to excellent, with observations on both males and females, giving a two by five contingency table of responses. As one of the categorizations is ordered, it is possible to do a more thorough analysis than that given by the usual X_P^2 test for a two-way contingency table. Sometimes, although the X_P^2 test may not be significant, an effect may be suggested by one of a number of analyses suggested in the literature, including

(i) giving the categories equi-spaced scores and using a regression analysis as in Yates (1948);

(ii) using nonequi-spaced scores based on mid-rank values as in Bross (1958), Conover (1998, p.281) and Nair (1986);

(iii) linear logistic models as in McCullagh (1980);

(iv) log-linear models and user defined assigned scores as in Agresti (1984, p.84);

(v) the cumulative chi-square method of Taguchi (1966); see also Nair (1986) and Hamada and Wu (1990) for a discussion of this method and reasons for not using it; and

(vi) analysis of variance and user defined assigned scores as in Box and Jones (1986) or Nair (1990).

We demonstrate here that a method using the summands or components of X_P^2 compares well with more recent methods and has the appeal of being weakly optimal and simple both conceptually and arithmetically. The method generalizes readily to other models, and to multi-way tables; see Beh and Davy (1998) and Beh and Davy (1999).

Our approach involves a family of simple parametric distributions. With sufficiently many parameters our models will fit the data exactly, but in practice highly parametrised models are not needed. Our parameters are related to moments, and we usually find it is sufficient to include only location and dispersion parameters for each row; skewness and kurtosis parameters could be included, but rarely would it be necessary to include more parameters. The test statistics we recommend for testing are asymptotically independent, assessing if the rows agree with regard to their location, dispersion, skewness etc. We suggest simultaneously assessing location and dispersion, and combining the remainder into a residual unless there are a *priori* reasons for doing otherwise.

2. One-Way Tables

We now describe some results for one-way tables because our two-way results are strongly motivated by the corresponding one-way results.

Best and Rayner (1987) gave formulae for obtaining components of the usual X_P^2 goodness of fit statistic for the multinomial, where there are n observations categorized into c classes with known class probabilities $p_1, p_2, \ldots, p_c$. These components may be correlated in small samples but are asymptotically uncorrelated and have the sum of their squares equal to X_P^2. If the categories are ordered and the components are based on orthogonal polynomials, then, for example, the first two components identify linear and quadratic effects, i.e. loosely location and dispersion effects. Subsequently we suppose the numbers of observations in the c classes are $N_1, N_2, \ldots, N_c$, where $n = N_1 + N_2 + \ldots + N_c$. 'Asymptotic' results mean $n \to \infty$.

Both the linear and the quadratic components are asymptotically distributed as standard normal variables, and power studies in Best and Rayner (1987) indicated these components compete well with a variety of other statistics when alternatives involve location and dispersion effects. For completeness, we give the relevant formulae. The orthogonal polynomials have, for $j = 1, \ldots, c$,

$$g_0(x_j) = 1, g_1(x_j) = (x_j - \mu)/\sqrt{\mu_2} \text{ and}$$

$$g_2(x_j) = a\{(x_j - \mu)^2 - \mu_3(x_j - \mu)/\mu_2 - \mu_2\},$$

in which

$$\mu = \sum_{j=1}^{c} x_j p_j, \mu_r = \sum_{j=1}^{c} (x_j - \mu)^r p_j \text{ and a } = \left(\mu_4 + \mu_3^2/\mu_2 - \mu_2^2\right)^{-0.5}.$$

The components are given explicitly by

$$\widehat{V}_u = \sum_{j=1}^{c} N_j g_u(x_j)/\sqrt{n}, u = 1, \ldots, c-1.$$

These components depend on the orthogonal polynomials, which are most conveniently given by using the explicit formulae for the g_1 and g_2, and then the recurrence relations of Emerson (1968). The $\hat{V}_u^2$ are score statistics in their own right, and hence provide weakly optimal directional tests (each seeks to detect alternatives in a one dimensional parameter space), so supplementing the omnibus nature of the X_P^2 test (that seeks to detect alternatives in a $c-1$ dimensional parameter space). The latter is based on the statistic

$$X_P^2 = \sum_{j=1}^{c} (N_j - np_j)^2/(np_j) = \hat{V}_1^2 + \ldots + \hat{V}_{c-1}^2.$$

Lancaster (1969, p.134) demonstrated such a partition of X_P^2 into components for the particular case $p_1 = \ldots = p_c$.

3. Two Way Tables

Consider the following product multinomial model. We have an r by c contingency table with cell probabilities $p_{ij}, i = 1, \ldots, r$, and $j = 1, \ldots, c$, such that $p_{i1} + \ldots + p_{ic} = 1$ for $i = 1, \ldots, r$. Observations $N_{i1}, \ldots, N_{ic}$ are taken on the cells of each of the i rows, yielding row totals $n_{i.}, i = 1, \ldots, r$ that were known before the data were collected. Column totals are random variables and are denoted by $N_{.j}, j = 1, \ldots, c$; the total count is $n_{..}$. Suppose that the columns are ordered categories, and it is of interest to compare rows for similarity of location and dispersion effects. The null hypothesis is equality of the corresponding row probabilities. If $p_{.j} = (p_{1j} + \ldots + p_{rj})/r$ for $j = 1, \ldots, c$, we test the null hypothesis $p_{ij} = p_{.j}$ for $i = 1, ..., r$, and $j = 1, ..., c$, against the alternative hypothesis, not the null. As the conditional probability $p_{j|i} = p_{ij}/p_{i.} = p_{ij}$ since $p_{i.} = 1$ for all i, the null hypothesis could also be written as $p_{j|i} = p_{.j}$ for $i = 1, \ldots, r$ and $j = 1, \ldots, c$: the conditional row probabilities are the same as the marginal probabilities. The usual X_P^2 statistic is derived in, for example, Conover (1998, section 4.2) to be

$$X_P^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} (N_{ij} - n_{..}\hat{p}_{ij})^2/(n_{..}\hat{p}_{ij}),$$

where $\hat{p}_{ij} = (n_{i.}/n_{..})(N_{.j}/n_{..})$. This X_P^2 statistic examines all deviations from what is expected under a homogeneity model. As in the one-way table, it is appropriate to decompose X_P^2 to obtain more informative directional tests. The summands of

the decomposition ($\hat{V}_{ui}$ subsequently) are not quite components as we usually use the term, as they are not independent, not even asymptotically.

To decompose X_P^2 statistics for the two-way table, $\hat{V}_1$ and $\hat{V}_2$ defined earlier can be calculated for each row (provided $c \geq 3$), yielding $\hat{V}_{1i}$ and $\hat{V}_{2i}$, for $1 \leq i \leq r$. In these $\hat{V}_{ui}$ the p_j are now taken to be $(N_{\cdot j}/n_{\cdot\cdot})$ for $1 \leq j \leq c$, and n is taken as $n_{i\cdot}, 1 \leq i \leq r$, and N_j becomes N_{ij}. Using the subsequent $g_u(x_j)$, further statistics $\hat{V}_{ui}$ can be defined by

$$\hat{V}_{ui} = \sum_{j=1}^{c} N_{ij}\hat{g}_u(x_j)/\sqrt{n_{i\cdot}}, u = 1, \ldots, c-1 \text{ and } i = 1, \ldots, r$$

where the hats indicate that the maximum likelihood estimators $\hat{p}_{ij} = N_{\cdot j}/n_{\cdot\cdot}, j = 1, \ldots, c$ have been used in the construction of the orthonormal polynomials.

We will show that

$$\sum_{u=1}^{c-1}\sum_{i=1}^{r} \hat{V}_{ui}^2 = X_P^2,$$

and that this is an alternative decomposition of X_P^2 to that given by Lancaster (1969, Theorem 6.2). We also show that the $\hat{V}_{ui}$ are score statistics for an appropriate model; this implies weak optimality. See Rayner and Best (1989, section 3.4) for a discussion of this optimality.

A measure of the location effect for the whole table is $\hat{V}_{11}^2 + \ldots + \hat{V}_{1r}^2$, which, when $x_j = j$ for $j = 1, \ldots, c$, is just the statistic Q of Yates (1948). Similarly the overall dispersion effect can be assessed by $\hat{V}_{21}^2 + \ldots + \hat{V}_{2r}^2$, provided $c \geq 3$. Provided $c \geq u + 1$, a measure of the uth moment departure from the null hypothesis is $\hat{V}_{u1}^2 + \ldots + \hat{V}_{ur}^2$. This interpretation could be called *diagnostic;* see Rayner and Best (1999) for a discussion and references. It should be noted though that the departure from the null could be due to moments between the $u+1$ th to the $2u$ th. However, if the model is correct then in large samples significance will be due to moments up to the u th, and we believe that is where most attention should focus.

If $\hat{V}_{u1}^2 + \ldots + \hat{V}_{ur}^2$, is significant then 2 by c tables could be examined in a multiple comparisons fashion. These statistics are asymptotically independent and each is well approximated by the χ_{r-1}^2 distribution. "Asymptotic" means $n_{\cdot\cdot} \to \infty$. Effectively we have a decomposition of X_P^2 into components that assesses, under the hypothesis of independence, the agreement of the rows of the table in regard to specific moment effects, up to the $c-1$ th moment. The statistics are asymptotically independent, and hence so are the assessments. By analogy with our goodness of fit work, we expect that the most significant effects will be in the first two to four moments.

4. Partitioning X_P^2 Using Score Statistics

We test for equality of the corresponding row probabilities by first setting, for $j = 1, \ldots,$ c and $i = 1, \ldots, r$,

$$p_{ij} = \left\{ 1 + \sum_{u=1}^{k} \theta_{ui} g_{uj} / \sqrt{n_{i.}} \right\} p_{.j}. \tag{4.1}$$

In (4.1) we note the following.

- The θ_{ui} are real valued parameters.
- The $\{g_{uj}\}$ is taken to be orthonormal:

$$\sum_{j=1}^{c} g_{uj} g_{vj} g_{.p} = \delta_{uv} \text{ for } u, v = 1, \ldots, c-1,$$

  where δ_{uv} is the Kronecker delta, $\delta_{uv} = 1$ for $u = v$, and $= 0$ for $u \neq v$. Typically the g_{uj} depend on the $p_{.j}$; we require that they do not depend on the row, since otherwise row comparison would be virtually impossible. A number of choices for g_{uj} are possible, but for ordered categories a ready interpretation is available if we use the $g_u(x_j)$ of section 2. Then each θ_{ui} reflects the u th moment shift of the distribution defined by the i th row from that defined on the $\{p_{.j}\}$.
- In the goodness of fit context we call k the *order* of the model. It can be at most $c - 1$, when the model becomes saturated, and an identity similar to Fisher's identity (given, for example, by Lancaster 1969, Theorem 2.1, Corollary 2) would result. Normally k would be chosen to be at most four, and more usually two.

Again in the goodness of fit context, we note that in Rayner and Best (1989) we say that a statistic is *partitioned* into components if the sum (or sum of squares) of the components gives that statistic, and the components are at least asymptotically independent. Our partition of X_P^2 using the $\hat{V}_{ui}^2$ does not have even asymptotic independence, and could be thought of as an arithmetic rather than a statistical partition. On the other hand, the sums $\hat{V}_{u1}^2 + \ldots + \hat{V}_{ur}^2$ are asymptotically independent and provide a partition in our usual sense.

In using the $\{g_{uj}\}$, we are effectively assigning scores $\{j\}$ to the ordered categories. The derivations generalise to user-assigned scores $\{x_j\}$ as anticipated by the discussion in section 3. It should, however, be emphasised that our approach assumes user-assigned and not estimated scores.

Colleagues have commented that although models of the form (4.1) are well-known to sometimes give excellent results, they can also produce negative probabilities. However (4.1) is asymptotically equivalent to

$$p_{ij} = C(\theta) \left\{ \exp \left[\sum_{u=1}^{k} \theta_{ui} g_{uj} / \sqrt{n_{i.}} \right] \right\} p_{.j}$$

where $\theta = (\theta_{11}, \ldots, \theta_{1r}, \ldots, \theta_{k1}, \ldots, \theta_{kr})^T$. Of course, this model cannot produce estimates of probabilities that are negative. The score tests from the two different models are asymptotically equivalent, but the derivations for (4.1) given here, messy as they are, are simpler than for the exponential model above. Whatever the asymptotic optimality probabilities of the test statistics from both models, they will ultimately be judged on their practicality, convenience and small sample properties. By these standards, the statistics derived here are not inferior to *any* of the available possibilities!

To test for equality of the corresponding row probabilities, take θ_{ui} to be the $(u-1)r + i$ th element of a vector θ. We test H_0: $\theta = 0$ against K: $\theta \neq 0$ with $p_{.1}, \ldots, p_{.(c-1)}$ as nuisance parameters; $p_{.c}$ is omitted from the set of nuisance parameters because the constraints $p_{i1} + \ldots + p_{ic} = 1,\ i = 1, \ldots, r$ imply $p_{.1} + \ldots + p_{.c} = 1$.

Subsequently we write $f = (\sqrt{n_{1.}}, \ldots, \sqrt{n_{r.}})^T$, $f^* = (\sqrt{n_{1.}}, \ldots, \sqrt{n_{(r-1).}})^T$ and I_n for the n by n identity matrix. Proofs of the three results that follow are in the Appendix. These derivations are anticipated in Best, Rayner and Stephens (1998).

Theorem 1: For the model (4.1), the information matrix evaluated at $\hat{p}_{.j} = N_{.j}/n_{..}, j = 1, \ldots, c$, is given by the direct sum of k matrices $I_r - ff^T/n_{..}$. This information matrix is singular.

The score statistic involves the inverse of the information matrix. One way to overcome the information matrix being singular is to omit $\theta_{1r}, \ldots, \theta_{kr}$ from the model. In modelling terms, the reason for doing this is that the θ's model differences between the row distributions and the average (fitted) distribution $\{N_{.j}/n_{..}\}$. If there are no differences for the first $r-1$ rows, then there will be no difference for the r th row.

Theorem 2: For $u = 1, \ldots, k$ and $i = 1, \ldots, r$ define $\hat{V}_{ui} = \Sigma_j N_{ij} \hat{g}_{uj} / \sqrt{n_{i.}}$. The score statistic for the model

$$p_{ij} = \{1 + \sum_{u=1}^{k} \theta_{ui} g_{uj} / \sqrt{n_{i.}}\} p_{.j},$$

for $i = 1, \ldots, r-1$ (**not** r as in (4.1)) and $j = 1, \ldots, c-1$, with $p_{rj} = p_{.j} - p_{1j} - \ldots - p_{(r-1)j}$ for $j = 1, \ldots, c-1$, and $p_{ic} = 1 - p_{i1} - \ldots - p_{i(c-1)}, i = 1, \ldots, r$, is

$$\hat{S}_k = \hat{V}_1^T \hat{V}_1 + \ldots \hat{V}_k^T \hat{V}_k$$

in which, $\hat{V}_u = (\hat{V}_{u1}, \ldots, \hat{V}_{ur}^T)$. The $\hat{V}_u$ are asymptotically independent.

Note that the hats indicate that the maximum likelihood estimators $\hat{p}_{.j} = N_{.j}/n_{..}$,

$j = 1, \ldots, c$, have been used in the construction of the orthonormal functions. Also $\hat{V}_u$ gives information about the deviations of order u from the fitted distribution $\{N_{\cdot j}/n_{\cdot\cdot}\}$; there are contributions to this information from all r rows. The asymptotic covariance matrix of $(\hat{V}_{u1}, \ldots, \hat{V}_{u(r-1)})$ is derived incidentally in the proof of Theorem 2 to be $I_{r-1} + f^* f^{*T}/n_{r\cdot}$. It follows that the $\hat{V}_{ui}$ are correlated and hence are not components. Also, since $\hat{V}_u$ is asymptotically $r-1$ variate normal with mean zero and covariance matrix $I_{r-1} + f^* f^{*T}/n_{r\cdot}$, the contribution to $\hat{S}_k$ from the u th order terms, $\hat{V}_{u1}^2 + \ldots + \hat{V}_{ur}^2$, asymptotically has the χ^2_{r-1} distribution. This uses a result for multivariate normal random variables, given for example, in Stuart and Ord (1987, 15.10). Given the asymptotic independence of the $\hat{V}_u, \hat{S}_k$ has asymptotic distribution $\chi^2_{k(r-1)}$.

Theorem 3: $\hat{S}_{(c-1)} = X_P^2$. Although dependent, the $\hat{V}_{ui}$ partition X_P^2 arithmetically in that the sum of the squares of all $r(c-1)\hat{V}_{ui}$ add to give X_P^2.

As remarked before the statement of Theorem 2, there are only $(r-1)(c-1)$ functionally independent θ's, for one row is determined from the average multinomial distribution by the other $r-1$ row distributions. Each $\hat{V}_{ua}$ corresponds to a θ_{ua}, and assesses the deviation of the u th moment of the i th row distribution $\{p_{i1}, \ldots, p_{ic}\}$ from the u th moment of the distribution defined by the $\{N_{\cdot 1}/n_{\cdot\cdot}, \ldots, N_{\cdot c}/n_{\cdot\cdot}\}$. In fact $\hat{V}_{ua}^2$ could be derived as the score statistic for the model: $p_{aj} = \{1 + \theta_{ua} g_{uj}/\sqrt{n_{a\cdot}}\}p_{\cdot j}$, and $p_{ij} = p_{\cdot j}$ for all $(r-2)(c-1)$ other p_{ij} in the first $r-1$ rows. Therefore each $\hat{V}_{ua}$ is the basis of a strongly directional test, with one dimensional parameter space $\{\theta_{ua}\}$. In the same vein we confirm that $\hat{V}_{u1}^2 + \ldots + \hat{V}_{ur}^2$ has the χ^2_{r-1} distribution by observing that it is the score statistic for the model $p_{ij} = \{1 + \theta_{ui} g_{uj}/\sqrt{n_{i\cdot}}\}p_{\cdot j}, i = 1, \ldots, r-1, j = 1, \ldots, c-1$. It has $r-1$ dimensional parameter space $\{\theta_{u1}, \ldots, \theta_{u(r-1)}\}$. It thus plays a useful intermediate role, being "more directional" than the $(r-1)(c-1)$ dimensional X_P^2, and "more omnibus" than each of the $\hat{V}_{u1}^2, \ldots, \hat{V}_{ur}^2$ singly.

Although we have not done a thorough analysis, we suspect that orthogonal polynomials can also be used as part of a log-linear model approach. In that case, the log-likelihood ratio statistic would be partitioned rather than Pearson's X_P^2. We would expect such test statistics to perform very similarly to those we have just introduced. Everitt (1992, section 7.3) for example, indicates how to proceed.

5. Other Methods for Ordered Data

5.1. Nair's Method

In simulation studies in Best, Rayner and Stephens (1998), we found that the statistics $\hat{V}_{11}^2 + \ldots + \hat{V}_{1r}^2$ and $\hat{V}_{21}^2 + \ldots + \hat{V}_{2r}^2$ are identical in numerical value to the location and dispersion statistics described by Nair (1986). Nair's location statistic is just the Kruskal-Wallis statistic adjusted for ties which is often applied to ranked one-way analysis of variance data. Eubank et al. (1987) showed that a number of commonly used statistics like the one on which the Kruskal-Wallis test is based, are components of Pearson's X_P^2.

Nair (1986) defined location scores

$$\ell_k = (t_k - 0.5)/\sqrt{\{\sum_{j=1}^{c} N_{.j}(t_j - 0.5)^2/n_{..}\}}$$

in which $t_k = (N_{.1} + \ldots + N_{.(k-1)} + N_{.k}/2)/n_{..}$, for $1 \leq k \leq c$, and also dispersion scores

$$d_k = e_k/\sqrt{\{\sum_{j=1}^{c} N_{.j}e_j^2/n_{..}\}}$$

in which $e_k = \ell_k\{\ell_k - (N_{.1}\ell_1^3 + \ldots + N_{.c}\ell_c^3)/n_{..}\} - 1$, for $1 \leq k \leq c$. If we define location and dispersion effects

$$\tau_i = N_{i1}\ell_1 + \ldots + N_{ic}\ell_c \text{ and } \omega_i = N_{i1}d_1 + \ldots + N_{ic}d_c,$$

then $\tau_i/\sqrt{n_{i.}}$ and $\omega_i/\sqrt{n_{i.}}$ are analogous to $\hat{V}_{1i}$ and $\hat{V}_{2i}$.

Nair's location scores are proportional to the midrank for category k. Graubard and Korn (1987) criticized the use of rank scores for contingency table analysis on the grounds that they may not give enough weight to extreme categories. The same sort of criticism may also apply to other data-dependent or estimated scores such as those given in the next section. Nair's statistics can also be derived as in section 3 if we use mid-rank scores rather than the "natural" scores $1, 2, \ldots, c$. So his statistics are special cases of our partition of X_P^2 statistics.

5.2. Logistic Models

The partition of X_P^2 given in Theorem 3 is relevant when either the rows (or columns) have ordered categories and where columns (or rows) have nominal categories. Another model suggested for use in this situation is the logistic model. McCullagh (1980) suggested the model:

$$\log\{(N_{i(j+1)} + \ldots + N_{ic})/(N_{i1} + \ldots + N_{ij})\} = (\alpha_j + \tau_i')/\omega_i'$$

in which $1 \leq i \leq r, 1 \leq j \leq c-1$ and $\tau_1' + \ldots + \tau_r' = 0$.

Iterative methods are needed for maximum likelihood estimation of the parameters. Agresti (1984, Appendix B.3) gave details. Notice that the parameters $\alpha_j, 1 \leq j \leq c-1$, estimate the scores which are not arbitrarily assigned while the τ_i' and ω_i' are location and dispersion parameters with $1 \leq i \leq r$. However it should be noted that there is some evidence, for example Agresti (1984, p.225), Newell (1986) and Box and Jones (1986), that in some cases there is little difference in both location and dispersion effects for assigned scores and estimated scores.

The logistic method we have just described requires iteration. However our X_P^2 method, which includes Nair's midrank scores method, does not. Further, if we take $\tau_i^* = \hat{V}_{1i}\sqrt{n_{i.}}$ with $1 \leq i \leq r$, we have

$$\sum_{i=1}^{r} \tau_i^* = \sum_{i=1}^{r} = \sqrt{n_{i.}}\{\sum_{j=1}^{c} N_{ij} g_1(j)/\sqrt{n_{i.}}\} = \sum_{j=1}^{c} N_{.j} g_1(j) = 0.$$

We now have

$$\sum_{i-1}^{r} \tau_i = \sum_{i-1}^{r} \tau_i' = \sum_{i-1}^{r} \tau_i^* = 0.$$

In this sense, the τ_i, τ_i', and τ_i^* are all contrasts. For completeness we also define $\omega_i^* = \hat{V}_{2i}\sqrt{n_{i.}}, i = 1, \ldots, r$.

5.3. ANOVA Analysis

Box and Jones (1986) and Nair (1990) suggested the use of analysis of variance (ANOVA) methods, and user defined assigned scores, to analyse ordered categorical data. However such an analysis relies on more assumptions to justify its use, and these additional assumptions may be difficult to justify. Sometimes the ANOVA method can give a non-orthogonal analysis, which is less convenient from many standpoints. Further, Brown (1988) has done a small simulation study which indicates that, when compared to X_P^2 tests, the ANOVA tests have actual sizes further from the nominal sizes. For these reasons we do not consider ANOVA methods further, although we have often used them in the past for the analysis of ordered categorical data. It may be worth extending the simulation study given by Brown (1988).

5.4. Comparison

The X_P^2 method, that includes Nair's method, and McCullagh's method both partition the total X_P^2 value into location, dispersion and residual effects. The location

and dispersion test statistics are each associated with $r-1$ degrees of freedom, and have asymptotic χ^2_{r-1} distributions. A simulation study reported in Best, Rayner and Stephens (1998) observed that, for the limited range of alternatives considered, the location and location plus dispersion tests for all methods considered have power almost exactly the same. Example 1 below demonstrates a difficulty with the use of the logistic model/McCullagh analysis. Our preference is for the X^2_P method. The $\hat{V}_{ui}$ are easily interpreted, give a complete and detailed scrutiny of the data, do not involve iterative calculations, and give an orthogonal partition of X^2_P.

6. Examples

Example 1. A taste-test experiment form Bradley et al. (1962) gave the response frequencies shown in Table 1.

Table 1. Response frequencies from a taste test experiment

	Response Category				
Product	- -	-	ϕ	+	++
1	9	5	9	13	4
2	7	3	10	20	4
3	14	13	6	7	0
4	11	15	3	5	8
5	0	2	10	30	2

Table 2. Taste test data summarized by location and dispersion parameters and three methods of analysis

	Analysis					
	X^2_P components		McCullagh		Nair	
Product	τ_i^*	ω_i^*	τ_i'	ω_i'	τ_i	ω_i
1	-0.22	2.40	0.07	1.25	-0.35	2.11
2	10.00	-0.87	0.56	1.05	9.85	-2.14
3	-25.06	-1.90	1.11	0.90	-24.97	-0.04
4	-11.02	14.60	-0.50	1.66	-10.41	16.74
5	26.30	-14.23	0.98	0.51	25.89	-16.68

This is a somewhat unusual taste-test as it appears the judges did not taste each product and so cannot be eliminated as blocks. However, it could be claimed that presenting one product per judge gives a more realistic consumer assessment of the products (McBride, 1986). For this contingency table, recalling that we have

defined $\tau_i^* = \hat{V}_{1i}\sqrt{n_{i\cdot}}$ and $\omega_i^* = \hat{V}_{2i}\sqrt{n_{i\cdot}}$, we obtain the summaries shown in Table 2.

The analyses are very similar. All indicate that treatment 5 is most liked, as τ_5, τ_5' and τ_5^* are the highest τ values, and that the judges agree on this, as ω_5, ω_5' and ω_5^* are the smallest ω values.

The high ω_4, ω_4' and ω_4^* values indicate the judges responses were most spread for treatment 4. Reference to the data indicates this spread is real, and not a spurious dispersion effect due to a large location effect, as discussed in Hamada and Wu (1990). Treatment 3 was least liked, as τ_3, τ_3' and τ_3^* are the smallest τ values, and the judges were in fair agreement about this, as ω_3, ω_3' and ω_3^* are the second or third smallest of the ω values. Notice the agreement between the analyses and, in particular, the closeness of the X_P^2 and Nair results.

We can also partition the value of the usual X_P^2 statistic as in the analysis of variance. For the X_P^2 analysis we get the analysis shown in Table 3.

Table 3. Partition of χ_P^2 for taste test data

Effect	df	SS	χ^2 p-value
location	4	36.58	0.000
dispersion	4	9.93	0.042
residual	8	27.33	0.001
total	16	73.84	

Table 4. Alternative partitions of the log-likelihood statistic

Effect	df	SS_1	χ^2 p-value	SS_2	χ^2 p-value
location	4	36.11	0.000	40.89	0.000
dispersion	4	27.76	0.000	22.98	0.000
residual	8	21.34	0.006	21.34	0.006
total	16	85.21		85.21	

However, because the logistic analysis is not orthogonal we get different analyses depending on whether location or dispersion effects are removed first. In fact we have either of the partitions shown in Table 4. In this case the conclusions from either logistic analysis are the same, but it is not clear that this would always be so.

Example 2. The overall X_P^2 value of 73.84 for the taste test data of Example 1 was highly significant. However, it can be the case that the overall X_P^2 is not significant but that the examination of location and/or dispersion statistics will indicate a significant effect. To illustrate this point consider the data in Table 5 which are taken

from Armitage (1955) and which are concerned with two treatments for ulcers and subsequent categorization of ulcer severity.

Table 5. Responses for two ulcer treatments

Treatment	# larger	# slightly healed	# most healed	# healed
A	12	10	4	6
B	5	8	8	11

For this ulcer data $X_P^2 = 5.91$ on three degrees of freedom implying a p-value in excess of 10% if the usual χ^2 approximation for X_P^2 is assumed.

However the location statistic $\hat{V}_{11}^2 + \hat{V}_{21}^2 = 5.26$, with a p-value between 1% and 5%. Treatment B is superior to treatment A, tending to have more responses in the number healed/mostly healed categories. The residual is 0.65 on two degrees of freedom, indicating there are no dispersion and no skewness effects. This example emphasizes the need to look at the $\hat{V}_{ui}$ and not just X_P^2, as in X_P^2 the insignificant dispersion and skewness effects have masked a significant location effect.

Appendix: Proof of Theorems

Note that sample values of the N_{ij} are written n_{ij} etc. The logarithm of the likelihood for our model is

$$\ell = \text{ constant } + \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} \log p_{ij}.$$

If θ_a and p_b are typical elements of θ and p respectively, then the efficient score is defined by $V = (\partial \ell / \partial \theta_a)$, and the information matrix by $M = I_{\theta\theta} - I_{\theta p} I_{pp}^{-1} I_{p\theta}$, in which $I_{\theta\theta} = (-E[\partial \ell / \partial \theta_a \partial \theta_b]), I_{\theta p} = (-E[\partial^2 \ell / \partial \theta_a \partial p_b]), I_{p\theta} = I_{\theta p}^T$, and $I_{pp} = (-E[\partial^2 \ell / \partial p_a \partial p_b])$. The score statistic is of the form $V_0^T M_0^{-1} V_0$, in which the subscript zero indicates evaluation under the null hypothesis. Asymptotically V_0 has a multivariate normal distribution with mean zero and covariance matrix M_0 (see, for example, Cox and Hinkley, 1974, Chapter 9). Subsequently we will need to find the inverse of I_{pp} for our model. For this purpose the following lemma will be needed; it may be easily verified.

Lemma: Let a be a constant, D an n by n diagonal matrix, and w an n by 1 vector. Provided $1 + a w^T D^{-1} w \neq 0$, put $b = -a/(1 + a w^T D^{-1} w)$. Then

$$(D + a\ ww^T)^{-1} = D^{-1} + b D^{-1} w w^T D^{-1}.$$

In our model $p = (p_{.1}, \ldots, p_{.(c-1)})$ is a $(c-1)$ by 1 vector and $\theta = (\theta_{11}, \ldots, \theta_{1r}, \ \ldots, \ \theta_{k1}, \ldots, \theta_{kr})$ is kr by 1. Note that we write θ_{wa} for a typical element of θ, where θ_{wa} is the $(w-1)r + a$ th element of θ. The same convention is used for the efficient score and elsewhere. Note that for $u = 1, \ldots, k$

$$\sum_{j=1}^{c} g_{uj} p_{\cdot j} = 0.$$

We call these the zero mean conditions, and they follow from our choice of orthonormal functions.

Subsequently we write 1_n for the n by 1 vector with every element 1. In the derivations that follow, the $p_{ic}, i = 1, \ldots, r$, will be treated as dependent variables.

Theorem 1 Proof: Using

$$\ell = \text{ constant } + \Sigma_i \Sigma_j n_{ij} \{\log(1 + \Sigma_u \theta_{ui} g_{uj} / \sqrt{n_{i\cdot}}) + \log p_{\cdot j}\},$$

we find

$$\partial \ell / \partial \theta_{wa} = \Sigma_j n_{aj} g_{wj} / \{\sqrt{n_{a\cdot}}(1 + \Sigma_u \theta_{uo} g_{uj} / \sqrt{n_{a\cdot}})\},$$

$$\partial \ell / \partial p_{\cdot s} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{n_{ij} \Sigma_u \theta_{ui} (\partial g_{uj} / \partial p_{\cdot s})}{\sqrt{n_{i\cdot}}(1 + \Sigma_u \theta_{ui} g_{uj} / \sqrt{n_{i\cdot}})} + \{n_{\cdot s} / p_{\cdot s} - n_{\cdot c} / p_{\cdot c}\}$$

using $\Sigma_i n_{ij} = N_{\cdot j}$, and

$$\partial^2 \ell / \partial \theta_{wa} \partial \theta_{zb} = -\delta_{ab} \Sigma_j n_{aj} g_{wj} g_{zj} (\sqrt{n_{a\cdot}} + \Sigma_u \theta_{ua} g_{uj})^{-2}.$$

After further differentiation and some manipulation we find that

$$\partial^2 \ell / \partial p_{\cdot s} \partial \theta_{wa} = \Sigma_j (n_{aj} / \sqrt{n_{a\cdot}})(\partial g_{wj} / \partial p_{\cdot s}) + \text{ terms zero when } \theta = 0 \text{ and}$$

$$\partial^2 \ell / \partial p_{\cdot s} \partial p_{\cdot t} = -\delta_{st} n_{\cdot t} / p_{\cdot t}^2 - n_{\cdot c} / p_{\cdot c}^2 + \text{ terms zero when } \theta = 0.$$

Taking E_0 of the second order derivatives and evaluating at $\hat{p}_{\cdot j} = N_{\cdot j} / n_{\cdot\cdot}, j = 1, \ldots, c$, gives $I_{\theta\theta} = I_{kr}$ using the orthonormality conditions. Also $I_{pp} = (n_{\cdot\cdot} / p_{\cdot c} + \delta_{st} n_{\cdot\cdot} / p_{\cdot t}) = \text{ diag}(n_{\cdot\cdot} / p_{\cdot s}) + (n_{\cdot\cdot} / p_{\cdot c}) 1_{(c-1)} 1_{(c-1)}^T$. The matrix $I_{\theta p}$ is of dimension kr by $(c-1)$ and has typical element $(\sqrt{n_{a\cdot}} \Sigma j p_{\cdot j} [\partial g_{uj} / \partial p_{\cdot s}])$. To simplify this, differentiate the zero mean conditions $\Sigma_j g_{uj} p_{\cdot j} = 0$ with respect to $p_{\cdot s}$, to give $0 = g_{us} - g_{uc} + \Sigma_j p_{\cdot j} [\partial g_{uj} / \partial p_{\cdot s}]$, so that $\Sigma_j p_{\cdot j} [\partial g_{uj} / \partial p_{\cdot s}] = g_{uc} - g_{us}$. It follows that $I_{\theta p} = (\sqrt{n_{a\cdot}} [g_{us} - g_{uc}])$.

To evaluate $I_{\theta p} I_{pp}^{-1} I_{p\theta}$ we first need I_{pp}^{-1}, and by the lemma of this appendix we obtain $n_{\cdot\cdot} I_{pp}^{-1} = \text{ diag}(p_{\cdot s}) - (p_{\cdot s} p_{\cdot t})$. Now on using

$$\sum_{j=1}^{c-1} (g_{uc} - g_{uj}) p_{\cdot j} = \sum_{j=1}^{c} (g_{uc} - g_{uj}) p_{\cdot j} = g_{uc} - \sum_{j=1}^{c} g_{uj} p_{\cdot j} = g_{uc} \text{ and}$$

$$\begin{aligned} \textstyle\sum_{j=1}^{c-1} (g_{uj} - g_{uc}) p_{\cdot j} (g_{ws} - g_{wc}) &= \textstyle\sum_{j=1}^{c} \{g_{uj} g_{ws} - g_{uc} g_{ws} - g_{wc} g_{uj} + g_{uc} g_{wc}\} p_{\cdot s} \\ &= \textstyle\sum_{j=1}^{c} g_{uj} g_{ws} p_{\cdot j} + g_{uc} g_{wc} = \delta_{uw} + g_{uc} g_{wc}, \end{aligned}$$

we find that $I_{\theta p}I_{pp}^{-1}I_{p\theta}$ is the direct sum of k equal matrices $I_r - ff^T/n_{..}$. The stated information matrix now follows. It is singular because f is a latent vector with zero latent root.

Theorem 2 Proof: After omitting $\theta_{1r}, \ldots, \theta_{kr}$ from the model, the information matrix becomes M^*, the direct sum of k matrices $I_{r-1} - f^*f^{*T}/n_{..}$, where f^* is the $(r-1)$ by 1 vector formed from f by omitting $\sqrt{n_{r.}}$. From the lemma this matrix has inverse $I_{r-1} + f^*f^{*T}/n_{r.}$, and the inverse of the information matrix is the direct sum of k such matrices. The efficient score is

$$\hat{V}^* = (\hat{V}_{11}, \ldots, \hat{V}_{1(r-1)}, \ldots, \hat{V}_{k1}, \ldots, \hat{V}_{k(r-1)})^T,$$

where the hats indicate that the maximum likelihood estimators of the nuisance parameters are required, namely $\hat{p}_{.j} = N_{.j}/n_{..}, j = 1, \ldots, c$. If we put

$$\hat{V}_w^* = (\hat{V}_{w1}, \ldots, \hat{V}_{w(r-1)})^T, w = 1, \ldots, k,$$

then $\hat{V}^*$ can also be expressed as

$$\hat{V}^* = (\hat{V}_1^{*T}, \ldots, \hat{V}_k^{*T})^T.$$

Substitution gives the score statistic as

$$\hat{S}_k = \hat{V}^{*T}\hat{M}^{*-1}\hat{V}^* = \sum_{w=1}^{k} \hat{V}_w^{*T}\{I_{r-1} + f^*f^{*T}/n_{r.}\}\hat{V}_w^*.$$

The contribution of the w th order terms is

$$\hat{V}_w^{*T}\{I_{r-1} + f^*f^{*T}/n_{r.}\}\hat{V}_w^{*T} = \sum_{i=1}^{r-1} \hat{V}_{wi}^2 + \{\sum_{i=1}^{r-1}\sum_{j=1}^{c} N_{ij}\hat{g}_{wj}\}^2/n_{r.}.$$

This simplifies if we first notice that

$$\sum_{i=1}^{r}\sum_{j=1}^{c} N_{ij}\widehat{g_{wj}} = \sum_{j=1}^{c} N_{.j}\hat{g}_{wj} = n_{..}\sum_{j=1}^{c} \hat{g}_{wj}\hat{p}_{.j} = 0,$$

using $N_{.j} = n_{..}\hat{p}_{.j}$ and the zero mean conditions. In terms of the $\hat{V}_w^*$, this is

$$\sqrt{n_{1.}}\hat{V}_{w1} + \ldots + \sqrt{n_{r.}}\hat{V}_{wr} = 0, w = 1, \ldots, k.$$

So

$$\sum_{i=1}^{r-1}\sum_{j=1}^{c} N_{ij}\hat{g}_{wj} = \sum_{i=1}^{r-1} \sqrt{n_{i.}}\hat{V}_{wi} = -\sqrt{n_{r.}}\hat{V}_{wr} \text{ and}$$

$$\hat{V}_w^{*T}\{I_{r-1} + f^* f^{*T}/n_{r.}\}\hat{V}*_w = \hat{V}_{w1}^2 + \ldots + \hat{V}_{wr}^2, \text{ for } w = 1, \ldots, k.$$

Theorem 3 Proof: The proof is similar to that of Rayner and Best (1989, Theorem 5.1.2). Write $H = (\hat{g}_{wj})$, and for $i = 1, \ldots, r$,

$$\hat{U}_i = (\hat{V}_{1i}, \ldots, \hat{V}_{(c-1)i})^T \text{ and } N_i = (N_{i1}, \ldots, N_{ic})^T.$$

Then by definition $\hat{U}_i = HN_i/\sqrt{n_{i.}}$. If we now put

$$\hat{V}_{1i}^2 + \ldots + \hat{V}_{(c-1)i}^2 = X_i^2,$$

the sum of the squares of the $\hat{V}_{ui}$ corresponding to each row, then

$$X_i^2 = \hat{U}_i^T \hat{U}_i = N_i^T H^T H N_i / n_{i.}.$$

Putting $\hat{p} = (\hat{p}_{.1}, \ldots, \hat{p}_{.c})^T$, the zero mean condition implies $H\hat{p} = 0$. The orthonormality condition may be expressed as

$$H^* \text{ diag}(\hat{p}_{.s}) H^{*T} = I_c,$$

where H^* is H augmented by a c th row of ones. This implies that

$$\text{diag}(\hat{p}_{.s}^{-1}) = H^{*T} H^* = H^T H + 11^T,$$

where 1 is c by 1 vector of ones. This gives

$$X_i^2 = \hat{U}_i^T \hat{U}_i = (N_i - n_{i.}\hat{p})^T H^T H (N_i - n_{i.}\hat{p})/n_{i.}$$

$$= (N_i - n_{i.}\hat{p})^T \{ \text{ diag}(\widehat{p_{.s}^{-1}}) - 11^T\}(N_i - n_{i.}\hat{p})/n_{i.}$$

$$= \Sigma_j (N_{ij} - n_{i.}\hat{p}_{.j})^2/(n_{i.}\hat{p}_{ij}).$$

This is of the form of Pearson's X_P^2, and is clearly the contribution to X_P^2 from the i th row. Summing over rows gives

$$X_P^2 = \Sigma_i \Sigma_j (N_{ij} - n_{i.}\hat{p}_{.j})^2/(n_{i.}\hat{p}_{.j}) = \Sigma_i X_i^2.$$

The non-zero covariance matrix establishes the dependence.

References

1. Agresti, A. *Analysis of Ordinal Categorical Data.* New York: Wiley, 1984.
2. Armitage, P. Tests for linear trends in proportions and frequencies. *Biometrics,* 11, 375-386, 1955.
3. Beh, E.J. and Davy, P.J. Partitioning Pearson's chisquared statistic for a completely ordered three-way contingency table. *Australia & New Zealand Journal of Statistics,* 40(4), 465-477, 1998.
4. Beh, E.J. and Davy, P.J. Partitioning Pearson's chisquared statistic for an ordered three-way contingency table: Part 2. *Australia & New Zealand Journal of Statistics*, 41(2), 233-246, 1999.
5. Best, D.J. and Rayner, J.C.W. Goodness-of-fit for grouped data using components of Pearson's X^2. *Computational Statistics and Data Analysis,* 5, 53-57, 1987.
6. Best, D.J.; Rayner, J.C.W. and Stephens, L.G. Small sample comparison of McCullagh and Nair analyses for nominal-ordinal categorical data. *Computational Statistics and Data Analysis,* 28, 217-223, 1998.
7. Box, G. and Jones, S. Discussion of Nair, V., Testing in industrial experiments with ordered categorical data, *Technometrics* 28, 283-294, *Technometrics* 28, 295-301, 1986.
8. Bradley, R.A.; Katti, S.K. and Coons, I.J. Optimal scaling for ordered categories. *Psychometrika,* 27, 355-374, 1962.
9. Bross, J. How to use ridit analysis. *Biometrics,* 14, 18-38, 1958.
10. Brown, G.H. The statistical comparison of reproduction rates for groups of sheep. *Australian Journal of Agricultural Research,* 39, 899-905, 1988.
11. Conover, W.J. *Practical Nonparametric Statistics.* (3rd ed.), New York: Wiley, 1998.
12. Cox, D.R. and Hinkley, D.V. *Theoretical Statistics.* London: Chapman and Hall, 1974.
13. Emerson, P.L. Numerical construction of orthogonal polynomials from a general recurrence formula. *Biometrics,* 24, 695-701, 1968.
14. Eubank, R.L.; La Riccia, V.N. and Rosenstein, R.B. Test statistics derived as components of Pearson's phi-squared distance measure. *Journal of American Statistics Association,* 82, 816-825, 1987.
15. Everitt, B.S. *The Analysis of Contingency Tables.* (2nd ed.), London: Chapman and Hall, 1992.
16. Graubard, B.I. and Korn, E.L. Choice of column scores for testing independence in ordered 2xK contingency tables. *Biometrics,* 43, 471-476, 1987.
17. Hamada, M. and Wu, C.F.J. A critical look at accumulation analysis and related methods. *Technometrics,* 32, 119-130, 1990.
18. Lancaster, H.O. *The Chi-Squared Distribution.* New York: Wiley, 1969.
19. McBride, R.L. Hedonic rating of food: single or side-by-side sample presentation. *Journal of Food Technology,* 21, 355-363, 1986.
20. McCullagh, P. Regression models for ordinal data. *Journal of the Royal Statistical Society* B, 109-142, 1980.
21. Nair, V. Testing in industrial experiments with ordered categorical data. *Technometrics,* 28, 283-311, 1986.
22. Nair, V. Discussion of Hamada, M. and Wu, C.F.J. A critical look at accumulation analysis and related methods. *Technometrics,* 32, 119-130, *Technometrics,* 32, 151-152, 1990.
23. Newell, G.J. Are rating scales linear? *Australian Marketing Researcher,* 10, 53-63, 1986.
24. Rayner, J.C.W. and Best, D.J. *Smooth Tests of Goodness of Fit.* New York: Oxford University Press, 1989.
25. Rayner, J.C.W. and Best, D.J. Contingency tables, ordered. Editors: Kotz, S.; Read, C.B. and Banks, D.L., in *Encyclopedia of Statistical Sciences.* New York: Wiley, 1999
26. Stuart, A.S. and Ord, J.K. *Kendall's Advanced Theory of Statistics,* Vol. 1, London: Griffin, 1987.
27. Taguchi, G. *Statistical Analysis* (in Japanese). Tokyo: Maruzen, 1966.
28. Yates, F. The analysis of contingency tables with groupings based on quantitative characters. *Biometrika,* 35, 176-181, 1948.