APPROXIMATE TEST STATISTICS FOR THE TEST OF INDEPENDENCE FOR SMALL SAMPLES FROM A CLUSTER SAMPLING SCHEME

Jeffrey R. Wilson, Arizona State University

Summary

A small sample test is constructed for the test of proportions for data obtained from cluster sampling schemes. The model assumes that the covariance matrix for the specified design is a function of the covariance matrix under multinomial sampling. A wald test statistic, Wald (1943) is constructed using the assumed covariance matrix. Brier (1980), Wilson (1984), and Wilson and Koehler (1984) made use of the covariance structure obtained under a Dirichlet model for cluster sampling. However, the models considered there are somewhat restrictive, in that they assume equal sample sizes for the clusters and constant design effects. This paper considers a less restrictive model for cluster sampling schemes through identification of a patterned covariance matrix and obtains test statistics based on these schemes. The results obtained are compared to those found in Wilson (1984) and Wilson and Koehler (1984). A comparison is also made to Bedrick (1983) and Rao and Scott, (1981, 1984). Tests of hypotheses are considered and limiting chi-squared distributions are obtained for the various test statistics. A numerical example is given based on data analyzed in Wilson and Koehler (1984) and test statistics are computed for each of the above mentioned procedures. The small sample test constructed here is related to other techniques and performs well as far as the numerical values obtained.

1. Introduction

A small sample test statistic is constructed for the test of proportions for data obtained from a non-multinomial sampling scheme. The model assumes that the covariance matrix of the proportions under the design has a special form. This special patterned covariance matrix is assumed to be a function of the corresponding covariance matrix for the proportions under multinomial sampling. In the construction of test statistics for proportions under cluster sampling scheme, Brier (1980), Wilson (1984), Wilson and Koehler (1984), made use of the nice covariance structure obtained under Dirichlet Multinomial Sampling. In this paper a less restrictive model than that used in Wilson (1984) is applied to the same set of data analyzed in that paper. Comparisons are made with the results obtained here and in Wilson (1984) and Wilson and Koehler (1984). The methods proposed by Rao and Scott (1981, 1984) require some information regarding the covariance matrix, such as the design effects or generalized design effect. Here the factors used to adjust the Pearson statistic for non-multinomial sampling are obtained from the summarized data. These factors are related in some ways to Brier's (1980) method of obtaining a single factor, C, for the Dirichlet Multinomial model. In that model approach, the covariance terms must be the same, a rather stringent condition in practice. In this paper each covariance term is allowed to have different design effects.

2. Model

Consider obtaining a vector of observed proportions $\hat{\underset{\sim}{\pi}} = (\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_I)'$ of dimension I for a certain subpopulation under some cluster sampling scheme of sample size n. Let the matrix of variances and covariances for the I dimensional probability vector $\hat{\underset{\sim}{\pi}}$ (an estimate of the vector of true proportions $\underset{\sim}{\pi}$) under the specified cluster sampling scheme be of the form

$$\Omega = n^{-1}(A - \underset{\sim}{F}\underset{\sim}{F}'), \tag{2.1}$$

where A is a diagonal matrix consisting of the elements of the vector

$$(\pi_1 b_{11}, \pi_2 b_{22}, \ldots, \pi_I b_{II}),$$

and

$$\underset{\sim}{F} = (\pi_1^{\frac{1}{2}} b_{11}, \pi_2^{\frac{1}{2}} b_{22}, \ldots, \pi_I^{\frac{1}{2}} b_{II})', \tag{2.2}$$

and (unknown) $b_{ii} > 0$ (i = 1 to I). Hence Ω can be expressed as

$$\Omega = n^{-1}\{B^{\frac{1}{2}}(\Delta_{\underset{\sim}{\pi}} - \underset{\sim}{\pi}\underset{\sim}{\pi}')B^{\frac{1}{2}}\}, \tag{2.3}$$

where

$$B^{\frac{1}{2}} = \text{diag}\{b_{11}^{\frac{1}{2}}, b_{22}^{\frac{1}{2}}, \ldots, b_{II}^{\frac{1}{2}}\} \tag{2.4}$$

and

$$\Delta_{\underset{\sim}{\pi}} = \text{diag}(\pi_1, \pi_2, \ldots, \pi_I). \tag{2.5}$$

When the b_{ii}'s are all equal this model has the same covariance structure as the Dirichlet Multinomial model as considered for cluster sampling in Brier (1980) and Wilson and Koehler (1984). If the b_{ii}'s are all equal to the value one then the covariance matrix is equivalent to the covariance matrix of $\hat{\underset{\sim}{\pi}}$ under multinomial sampling. Since $(\Delta_{\underset{\sim}{\pi}} - \underset{\sim}{\pi}\underset{\sim}{\pi}')$ is singular then the covariance matrix, Ω is singular. Ω has rank I-1.

Let Σ_m denote the covariance matrix of $\hat{\underset{\sim}{\pi}}$ under multinomial sampling then,

$$\begin{aligned}\text{Cov}(\hat{\underset{\sim}{\pi}}) &= \Sigma_m \\ &= n^{-1}(\Delta_{\underset{\sim}{\pi}} - \underset{\sim}{\pi}\underset{\sim}{\pi}').\end{aligned} \tag{2.6}$$

Thus the design's covariance matrix,

$$\Omega = n^{-1}\{B^{\frac{1}{2}}(\Delta_{\pi} - \underset{\sim}{\pi}\underset{\sim}{\pi}')B^{\frac{1}{2}}\} = B^{\frac{1}{2}}\,\Sigma_m\, B^{\frac{1}{2}}. \tag{2.7}$$

Let σ_{ij} denote the ijth element of Ω, then

$$\sigma_{ii} = n^{-1}b_{ii}(\pi_i - \pi_i^2) \qquad i=j \tag{2.8}$$

and

$$\sigma_{ij} = -n^{-1}b_{ii}^{\frac{1}{2}}\, b_{jj}^{\frac{1}{2}}\, \pi_i\, \pi_j. \qquad i\neq j \tag{2.9}$$

From equation 2.8 set

$$b_{ii} = \sigma_{ii}/n^{-1}\pi_i(1-\pi_i), \tag{2.10}$$

then b_{ii} is a ratio of the variance under the specified design to the variance under the multinomial sampling. When I=2, Rao and Scott (1981) refer to such a ratio as the design effects. The matrix B is unknown but its elements can be estimated from (2.10) by estimating σ_{ii} and π_i, i=1, 2,..., I. Here we choose to use the diagonal elements as the off diagonal elements would produce an equation with the extra unknown.

3. Testing of Hypothesis

3.1 Hypothesis $\underset{\sim}{\pi} = \underset{\sim}{\pi}_o$

It is well known that test statistics for goodness of fit, independence and homogeneity are different when the sampling scheme is not multinomial. We consider these hypotheses now.

Suppose the data from a particular cluster sampling scheme with a relatively large number of clusters and sample size, n are obtained and the interest is in the hypothesis

$$H_o\colon \underset{\sim}{\pi} = \underset{\sim}{\pi}_o.\ (\underset{\sim}{\pi}_o \text{is known}) \tag{3.1}$$

Denote the observed proportions by

$$\hat{\underset{\sim}{\pi}} = (\hat{\pi}_1, \hat{\pi}_2, \ldots\ldots, \hat{\pi}_I)'. \tag{3.2}$$

Then the covariance matrix, Σ_o of the observed proportions is $\Sigma_o = n^{-1}(\Delta_{\underset{\sim}{\pi}_o} - \underset{\sim}{\pi}_o\underset{\sim}{\pi}_o')$ under H_o.

3.1.1 Test Statistic for $H_o\colon \underset{\sim}{\pi} = \underset{\sim}{\pi}_o$.

Consider constructing a Wald type statistic, Wald (1943) to test the hypothesis in (3.1), using the covariance matrix Ω in (2.7). Such a Wald test statistic is given by

$$X^2_{(ii)} = (\hat{\underset{\sim}{\pi}} - \underset{\sim}{\pi}_o)'\ \hat{\Omega}_{ii}^{-1}\ (\hat{\underset{\sim}{\pi}} - \underset{\sim}{\pi}_o) \tag{3.3}$$

where $\hat{\Omega}_{ii}$ is a consistent estimator of Ω_{ii}. Ω_{ii} is equivalent to the covariance matrix Ω where the ith row and ith column are left off. The ith row of $(\hat{\underset{\sim}{\pi}} - \underset{\sim}{\pi}_o)'$ is also left off.

Then,

$$X^2_{(ii)} = n(\hat{\underset{\sim}{\pi}} - \underset{\sim}{\pi}_o)'\ [A - F\,F']^{-1}(\hat{\underset{\sim}{\pi}} - \underset{\sim}{\pi}_o) \tag{3.4}$$

$$= n(\hat{\underset{\sim}{\pi}} - \underset{\sim}{\pi}_o)'\ [B^{-\frac{1}{2}}(\Delta^{-1}_{\underset{\sim}{\pi}_o} + \underset{\sim}{\pi}_o\ \underset{\sim}{\pi}_o')B^{-\frac{1}{2}}]\ (\hat{\underset{\sim}{\pi}} - \underset{\sim}{\pi}_o).$$

So $X^2_{(ii)}$ can be expressed as

$$X^2_{(ii)} = n \sum_{\substack{j=1\\ j\neq i}}^{I} b_{jj}^{-1}\, \pi_{oj}^{-1}\, (\hat{\pi}_j - \pi_{oj})^2 + \pi_{oi}^{-1}\, n\, [\sum_{\substack{j=1\\ j\neq i}}^{I} b_{jj}^{-\frac{1}{2}}\, (\hat{\pi}_j - \pi_{oj})]^2. \tag{3.5}$$

Note that when b_{ii}'s are all equal then $X^2_{(ii)}$ for i = 1 to I are all equal.

The choice of the ith category to be omitted makes a difference to the value of $X^2_{(ii)}$ unless the b_{ii}'s are all equal. An average of the $X^2_{(ii)}$ i=1,2...,I can be used as an approximate test. Another procedure involves using the Moore-Penrose Inverse to construct a Wald test, X^2_w given by

$$X^2_w = (\hat{\underset{\sim}{\pi}} - \underset{\sim}{\pi}_o)'\ \Omega^-(\hat{\underset{\sim}{\pi}} - \underset{\sim}{\pi}_o)$$

where Ω^- is the Moore Penrose inverse of Ω, Moore (1977). One aspect is that the mean of $X^2_{(ii)}$ i=1, 2, ...I consists of

$$X^2_{\Omega A} = \frac{I-1}{I}\, n \sum_{i=1}^{I} b_{ii}^{-1}\, \pi_{oi}^{-1}(\hat{\pi}_i - \pi_{oi})^2,$$

$$\text{for } i=1,\ 2,\ \ldots,\ I; \tag{3.6}$$

under H_o, and

$$X^2_{\Omega B} = nI^{-1} \sum_{i=1}^{I} \pi_{oi}^{-1}\ [b_{ii}^{-\frac{1}{2}}\ (\hat{\pi}_i - \pi_{oi}) - \sum_{j=1}^{I} b_{jj}^{-\frac{1}{2}}\ (\hat{\pi}_j - \pi_{oj})]^2 \tag{3.7}$$

$$nI^{-1} \sum_{i=1}^{I} \pi_{oi}^{-1}\ b_{ii}^{-1}\ (\hat{\pi}_i - \pi_{oi})^2 =$$

$$nI^{-1}\ b \sum_{i-1}^{I} \pi_{oi}^{-1}\ (\hat{\pi}_i - \pi_{oi})^2 \text{ as}$$

as all b_{ii} → a constant b. Then the conjecture is that $X^2_{\Omega A} + X^2_{\Omega B} \to n \sum_{i=1}^{I} b_{ii}^{-1}\, \pi_{oi}^{-1}(\hat{\pi}_i - \pi_{oi})^2$.

When the b_{ii}'s are less than one then the test statistics value, $I^{-1}\sum_{i=1}^{I} X^2_{(ii)} = X^2_{\Omega A} + X^2_{\Omega B}$ under the specified design is larger than the test statistic value obtained assuming that multinomial sampling is present. Such a condition will result in a loss of power. When the b_{ii}'s are greater than one then the test sta- the test statistic value, $X^2_{\Omega A}$ under the specified design is less than the test statistic value obtained assuming that multinomial sampling is present. Such a condition results in a conservative test. When all the b_{ii}'s are equal to one then the test statistic $X^2_{\Omega A}$ is equal to the usual Pearson goodness of fit statistic. This means that the clustering is negligible. When some b_{ii}'s are significantly less than one and some b_{ii}'s greater than one the relationship between $X^2_{\Omega A}$ and the usual Pearson statistics are not clearly determined. The b_{ii}'s can be estimated by taking a ratio of the variance under the specified design to the variance under multinomial sampling. This requires some knowledge of cell variances. These results agree with the findings of Rao & Scott (1981) for different sampling schemes.

The statistic $I^{-1} \sum_{i=1}^{I} X^2_{(ii)}$ can be partitioned as A + B where

$$A = n \sum_{i=1}^{I} b_{ii}^{-1} \pi_{io}^{-1} (\hat{\pi}_i - \pi_{io})^2 \qquad (3.8)$$

and

$$B = nI^{-1} [\sum^{I} \pi_{oi}^{-1} - b_{ii}^{-1} (\hat{\pi}_i - \pi_{oi})^2] +$$

$$nI^{-1} \sum_{i=1}^{I} \pi_{oi}^{-1} [b_{ii}^{-\frac{1}{2}} (\hat{\pi}_i - \pi_{oi})$$

$$- \sum_{j} b_{jj}^{-\frac{1}{2}} (\pi_j - \pi_o)]^2$$

$$= NI^{-1} \sum_{i=1}^{I} \pi_{oi}^{-1} [-2b_{ii}^{-\frac{1}{2}} (\hat{\pi}_i - \pi_{oi})$$

$$\sum_{j} b_{jj}^{-\frac{1}{2}} (\hat{\pi}_j - \pi_{oj}) + \{\sum_{j} b_{jj}^{-\frac{1}{2}} (\pi_j - \pi_{oj})\}^2]$$

$$= -2ndI^{-1} (\sum_{i=1}^{I} \pi_{oi}^{-1} b_{ii}^{-\frac{1}{2}} (\hat{\pi}_i - \pi_{oi})) + nd^2 I^{-1},$$

where

$$d = \sum_{j} b_{jj}^{-\frac{1}{2}} (\hat{\pi}_j - \pi_{oj})$$

Then d and B are close to zero if all b_{ii}'s are nearly equal, but may possibly be close to zero in other cases as well.

Since $\sqrt{n}\ (\hat{\underset{\sim}{\pi}} - \underset{\sim}{\pi}_o)$ has a limiting normal distribution with mean vector $\underset{\sim}{0}$ and covariance matrix Ω for sufficiently large n, then under H_o, the limiting distribution of $(\hat{\underset{\sim}{\pi}} - \underset{\sim}{\pi}_o)' \hat{\Omega}^- (\hat{\underset{\sim}{\pi}} - \underset{\sim}{\pi}_o)$ is an approximate chi-square random variable distributed with I-1 degrees of freedom (Moore 1977). The approximate statistic $X^2_{\Omega A}$ may be considered as a chi-square random variable with (I-1) degrees of freedom.

3.2 Test of Independence

We now consider procedures for testing the hypothesis of independence. Assume that there dence. Assume that there are J probability vectors $\underset{\sim}{\pi}_j$ for j = 1, 2, ... J; obtained from each of J subpopulations where cluster sampling is used within each subpopulation of sample size n_j such that

$$\hat{\underset{\sim}{\pi}}_j = (\hat{\pi}_{1j}, \hat{\pi}_{2j}, \ldots, \hat{\pi}_{Ij})', \qquad (3.10)$$

$$\sum_{j=1}^{J} n_j = n,$$

and the hypothesis of interest is

$$H_o:\ \underset{\sim}{\pi}_j = \underset{\sim}{\pi}_o \ (j = 1, 2, \ldots J) \qquad (3.11)$$

where $\underset{\sim}{\pi}_o$ is an unknown probability vector. Suppose that there exists an estimated covariance matrix for each $\underset{\sim}{\pi}_j$, defined as Σ_j with pqth element denoted by $\hat{\sigma}_{jpq}$. Consider a ratio estimator of the elements of the vector

$$\underset{\sim}{\pi}_0 = (\pi_{01}, \pi_{02}, \ldots, \pi_{0I})' \qquad (3.12)$$

as

$$\hat{\pi}_{oi} = \sum_{j=1}^{J} n_j b_{jii}^{-1} \hat{\pi}_{ij} [\sum_{j=1}^{I} \sum_{i=1}^{I} n_j b_{jii}^{-1}]^{-1} \qquad (3.13)$$

a linear combination of the estimated probability vectors. Let $b_{j\ell\ell}$ be estimated by

$$\hat{b}_{j\ell\ell} = \hat{\sigma}_{j\ell\ell} / n_j^{-1} \ \hat{\pi}_{o\ell j} (1 - \hat{\pi}_{o\ell j})$$

$$j = 1, 2, \ldots J; \qquad (3.14)$$

following equation (2.8). Similar to equation (2.1), let

$$\Sigma_{(\ell\ell)j} = n_j^{-1} \ [A_{(\ell\ell)j} - \underset{\sim}{F}_{(\ell\ell)j} \underset{\sim}{F}'_{(\ell\ell)j}], \qquad (3.15)$$

where $\Sigma_{(\ell\ell)j}$ is (I-1) x (I-1) matrix obtained by omitting the ℓth row and ℓth column of the covariance matrix of

$$\Sigma_j = n_j^{-1} [A_j - \underset{\sim}{F}_j \underset{\sim}{F}'_j], \qquad (3.16)$$

with

$$A_j = \text{diag}[b_{j11} \pi_{01j}, b_{j22} \pi_{02j}, \ldots, b_{jII} \pi_{0Ij}], \qquad (3.17)$$

$$F_j = (b_{j11}^{\frac{1}{2}} \pi_{01j}, b_{j22}^{\frac{1}{2}} \pi_{02j}, \ldots, b_{jII}^{\frac{1}{2}} \pi_{0Ij})'.$$

From Graybill (1969), an inverse of $\Sigma_{(\ell\ell)j}$ is

$$\Sigma^{-1}_{(\ell\ell)j} = n_j [A^{-1}_{(\ell\ell)j} + (1 - \sum_{\substack{i=1 \\ i \neq \ell}}^{I} b_{jii}^{-1} \pi_{oij})^{-1} C_{(\ell\ell)j} C'_{(\ell\ell)j}] \qquad (3.18)$$

where

$$\underset{\sim}{C}_{(\ell\ell)j} = (b_{j11}^{\frac{1}{2}}, b_{j22}^{\frac{1}{2}}, \ldots, b_{jII}^{\frac{1}{2}})', \qquad (3.19)$$

which is C_j with the ℓth category left off.

Then, a chi-square test for the hypothesis in (3.9) is

$$X^2_{\Omega D \ell} = (\underset{\sim}{\hat{\pi}}^{(J)} - \underset{\sim}{\hat{\pi}}_o^{(J)})'_{(\ell\ell)} \; [\hat{\Sigma}_{(\ell\ell)}^{(J)}]^{-1}$$

$$(\underset{\sim}{\hat{\pi}}^{(J)} - \underset{\sim}{\hat{\pi}}^{(J)})_{(\ell\ell)}, \qquad (3.20)$$

where

$$(\underset{\sim}{\hat{\pi}}^{(J)} - \underset{\sim}{\hat{\pi}}_0^{(J)})_{\ell\ell} = (\underset{\sim}{\hat{\pi}}_1'-\underset{\sim}{\hat{\pi}}_0', \underset{\sim}{\hat{\pi}}_2'-\underset{\sim}{\hat{\pi}}_0', \ldots,$$

$$\underset{\sim}{\hat{\pi}}_{J-1}' - \underset{\sim}{\hat{\pi}}_o')', \qquad (3.21)$$

with the ℓth category of each $\underset{\sim}{\hat{\pi}}_j' - \underset{\sim}{\hat{\pi}}_o'$ left off, and $\hat{\Sigma}^J_{(\ell\ell)}$ is an estimate of

$$\Sigma^{(J)} = \text{cov}\ (\underset{\sim}{\hat{\pi}}^{(J)} - \underset{\sim}{\hat{\pi}}_0^{(J)}), \qquad (3.22)$$

with the ℓth row and the ℓth column of each block diagonal left off. The covariance matrix $\Sigma^{(J)}$ is a block diagonal matrix with elements Σ_j. The covariance matrix $\Sigma^{(J)}_{\ell\ell}$ is $\Sigma^{(J)}$ when the ℓth row and ℓth column of each block are omitted. $X^2_{\Omega D \ell}$ can be expressed, by use of equation (3.18) as

$$X^2_{\Omega D \ell} = \sum_{j=1}^{I} n_j \{ \sum_{\substack{i=1\\ i\neq \ell}}^{I} (\hat{\pi}_{ij} - \hat{\pi}_{oi})^2 b^{-1}_{jii} \hat{\pi}^{-1}_{oi} +$$

$$\hat{\pi}^{-1}_{o\ell}[\sum_{\substack{t=1\\ t\neq \ell}}^{I} b^{-\frac{1}{2}}_{jtt}(\hat{\pi}_{tj} - \hat{\pi}_{ot})]^2 \} . \qquad (3.22)$$

As suggested for the X^2_{ii}'s (3.5) the conjecture is that the average of $X^2_{\Omega D \ell}$ for ℓ=1, 2, ..., I; is numerically close to

$$X^2_{\Omega DA} = \sum_{j=1}^{J} \sum_{i=1}^{I} n_j b^{-1}_{jii} \hat{\pi}^{-1}_{oi} (\hat{\pi}_{ij} - \hat{\pi}_{oi})^2 . \qquad (3.24)$$

The statistic $X^2_{\Omega DA}$ is similar in form to statistics developed in Wilson (1984) and Wilson and Koehler (1984) except that the weights used here apply to the contribution from each cell individually instead of affecting an average of those cell contributions for each cluster. It also differs from Rao and Scott (1981) and Bedrick (1983) in the way the weights are used. In their work the weights are obtained as design effects and are used as an overall divisor in the construction of the test statistic. The chi-squared test, $X^2_{\Omega D \ell}$ can be averaged over the I positive deletions to obtain an approximate statistic symmetric in the observations. An average of $X^2_{\Omega D \ell}$, ℓ = 1, . . . I; is

$$X^2_{\Omega D} = \sum_{j=1}^{J} \sum_{i=1}^{I} n_j b^{-1}_{jii} \hat{\pi}^{-1}_{oi} (\hat{\pi}_{ij} - \hat{\pi}_{oi})^2 +$$

$$(I-1)I^{-1} \sum_{t=1}^{I} \{ \sum_{t=1}^{I} \sum_{j=1}^{J} \hat{\pi}^{-1}_{0\ell} n_j [b^{-\frac{1}{2}}_{j\ell\ell} (\hat{\pi}_{\ell j} - \hat{\pi}_{o\ell})$$

$$- \sum_{t=1}^{I} b^{-\frac{1}{2}}_{jtt}(\hat{\pi}_{tj} - \hat{\pi}_{ot})]^2 \} .$$

$$(3.25)$$

Another form of a chi-square test statistic based on the Moore-Penrose inverse for the covariance matrix, $\Sigma^{(J)}$ is

$$X^2_D = (\underset{\sim}{\hat{\pi}}^{(J)} - \underset{\sim}{\hat{\pi}}_o^{(J)})'[\hat{\Sigma}^{(J)}]^- (\underset{\sim}{\hat{\pi}}^{(J)} - \underset{\sim}{\hat{\pi}}_o^{(J)}),$$

$$(3.26)$$

where $[\hat{\Sigma}^{(J)}]^-$ is the Moore-Penrose inverse of $\Sigma^{(J)}$.

3.3 Limiting Distribution of $X^2_{\Omega D}$

By the Multivariate Central Limit Theorem, $\sqrt{n_j}\ (\underset{\sim}{\hat{\pi}}_j - \underset{\sim}{\pi}_j)$ has a limiting noraml distribution for sufficiently large n_j. Since $\sqrt{n_j}\ (\underset{\sim}{\hat{\pi}}_j - \underset{\sim}{\hat{\pi}}_0)$ is a linear combination (for b_{jii} fixed) of the $\underset{\sim}{\hat{\pi}}_j$'s, then by Cramer (1946) $\sqrt{n_j}\ (\underset{\sim}{\hat{\pi}}_j - \underset{\sim}{\hat{\pi}}_0)$ also has a limiting normal distribution. Under H_o, $\sqrt{n_j}\ (\underset{\sim}{\hat{\pi}}_j - \underset{\sim}{\hat{\pi}}_0)$ has a mean vector $\underset{\sim}{0}$ and covariance matrix given by Σ_j. But $\underset{\sim}{\hat{\pi}}^{(J)} - \underset{\sim}{\hat{\pi}}_0^{(J)}$ is a linear combination of $\underset{\sim}{\hat{\pi}}^{(J)}$. Therefore, the statistic X^2_D is distributed asymptotically as a chi-square random variable with (I-1)(J-1) degrees of freedom (Moore 1977). If b_{jii} is unknown and a consistent estimator is available then the asymptotic distribution of $X^2_{\Omega D}$ is also asymptotically chi-square.

3.4 Estimating b_{ii}'s.

Consider sampling the same number of observations from each cluster then from equation 2.8 a consistent estimator of b_{jii} is

$$\hat{b}_{jii} = \hat{\sigma}_{jii}/n_j^{-1} \hat{\pi}_{ij}(1 - \hat{\pi}_{ij}) \qquad (3.27)$$

where $\hat{\sigma}_{jii}$ is the ith diagonal element of

$$\hat{\Sigma}_j = (s_j-1)^{-1} \sum_{t=1}^{s_j} (\underset{\sim}{\hat{\pi}}_{tj} - \underset{\sim}{\hat{\pi}}_j)\ (\underset{\sim}{\hat{\pi}}_{tj} - \underset{\sim}{\hat{\pi}}_j)' ,$$

$$(3.28)$$

where $\underset{\sim}{\hat{\pi}}_{tj}$, is the estimated probability vector of the tth cluster of the jth subpopulation and s_j is the number of clusters in the jth subpopulation.

A second possible estimate of the vector of b_{jii}'s following the procedure of Wilson (1984) and equivalent to Brier (1980) in estimating the cluster effect for each subpopulation category is

$$\tilde{b}_{jii} = (s_j-1)^{-1} \sum_{t=1}^{s_j} n_j\ (\hat{\pi}_{itj} - \hat{\pi}_{ij})^2 \hat{\pi}^{-1}_{ij}.$$

$$(3.29)$$

An average of $\tilde{b}_{iii}$ is equivalent to the factor C if the Dirichlet Multinomial model with its constant design effects is assumed for clustering (Wilson and Koehler 1984).

4. Numerical Example

Brier (1980) considered data pertaining to the manner in which people in Minnesota perceive the quality of their housing and their community housing. The variable of interest in this survey is the opinions of families about their homes (personal satisfaction). There were 85 families questioned in the metropolitan area and 90 questioned in the outlying area.

In each community, five homes were randomly selected and the families were questioned about two items: satisfaction with the housing in the neighborhood as a whole (unsatisfied, satisfied, very satisfied) and satisfication with their own home. The groups of five homes are the clusters. There are a total of 35 clusters, 17 in the metropolitan Minneapolis-St. Paul area 18 in the outlying region (non metropolitan area).

The hypothesis of interest is in the distribution of the responses for the two areas given as

$$H_{op}: \underset{\sim}{\pi}_j = \underset{\sim}{\pi}_o \text{ (unknown) } j = 1,\ 2;$$

for personal satisfaction categories.

The two subpopulations correspond to the non-metropolitan (non-metro) area and the metropolitan (metro) area, so J = 2. Let subscript 1 denote the non-metropolitan subpopulation area and subscript 2 denote the metropolitan subpopulation area. Then, the observed vectors of proportions are

$$\hat{\underset{\sim}{\pi}}_1 = (.5222,\ .4222,\ .0556)'$$

for the non-metropolitan area and

$$\hat{\underset{\sim}{\pi}}_2 = (.3529,\ .5059,\ .1412)\ '$$

for the metropolitan area. A test statistic, for testing H_{op}, was computed in Wilson and Koehler (1984), based on the Dirichlet Multinomial model. They constructed the statistic

$$X^2_{DMI} = \sum_{j=1}^{2} N_j \hat{C}_j^{-1} \sum_{i=1}^{3} (\hat{\pi}_{ij} - \hat{\pi}_{io})^2\ \hat{\pi}_{io}^{-1}$$

where N_j is the total sample for the j^{th} subpopulation $\hat{\underset{\sim}{\pi}}_j = (\hat{\pi}_{1j}, \hat{\pi}_{2j}, \ldots, \hat{\pi}_{Ij})'$ is the observed vector of proportions for the j^{th} subpopulation, $\hat{C}_j$ is a consistent estimator for the clustering effect, C_j, in the covariance matrix for the Dirichlet-Multinomial distribution,

$$\hat{\pi}_{io} = \sum_{j=1}^{2} N_j\ \hat{C}_j^{-1}\ \hat{\pi}_{ij}\ [\sum_{\ell=1}^{2} N_\ell \hat{C}_\ell^{-1}]^{-1},$$

$$i = 1,2,3;$$

One possible estimator for C_j following the method of Brier (1980) is

$$\hat{C}_{jB} = (I-1)^{-1}\ (s_j-1)^{-1} \sum_{t=1}^{s_j} \sum_{i=1}^{I} m_j$$

$$(\hat{\pi}_{itj} - \hat{\pi}_{ij})^2\ \hat{\pi}_{ij}^{-1},$$

where

$$\hat{\underset{\sim}{\pi}}_{jt} = (\hat{\underset{\sim}{\pi}}_{1jt}, \hat{\underset{\sim}{\pi}}_{2jt}, \ldots, \hat{\underset{\sim}{\pi}}_{Ijt})',$$

is the vector of proportions for the tth cluster of the jth subpopulation. I is the number of categories, s_j is the number of clusters, and m_j is the size of the clusters in the jtj subpopulation. The values for $\hat{C}_{jB}$, j = 1,2 for personal satisfaction are $\hat{C}_{1B}$ = 1.619 (nonmetro area) and C_{2B} = 1.632 (metro area).

The estimated proportion vector from (3.13) is

$$\hat{\underset{\sim}{\pi}}_o = (.4427,\ .4642,\ .103)'.$$

and the estimated vector of b's for the metropolitan area from (3.27) is

$$\hat{\underset{\sim}{b}}_1 = (0.9975,\ 1.0526,\ 1.1882)'.$$

A second estimate of the vector of b's from (3.29) is

$$\tilde{\underset{\sim}{b}}_1 = (2.0878,\ 1.8219,\ 1.2581)'.$$

An average of the elements of $\tilde{b}_1$ is equal to C_{1B}. Similarly for the non-metropolitan area the vector of b's are

$$\hat{\underset{\sim}{b}}_2 = (1.5250,\ .8953,\ .8438)' \text{ from } (3.27)$$

and

$$\tilde{\underset{\sim}{b}}_2 = (2.3568,\ 1.8120,\ 0.9824)', \text{ from } (3.29),$$

which also has elements with average equal to C_{2B}.

The covariance matrices $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ for the vectors $\underset{\sim}{\pi}_1$ and $\underset{\sim}{\pi}_2$ under the specified design can be estimated using $\hat{b}$ or $\tilde{b}$ in expression (3.4). Here we use $\tilde{b}$ to construct the various statistics. A constant reminder of this is given by attaching the ∿ sign to the computed statistic. The vectors F_1 and F_2 (3.17) which are used to compute Σ_1 and Σ_2 are estimated respectively by

$$\tilde{F}_1 = (0.3560,\ 0.3400,\ 0.3237)'$$

and

$$\tilde{F}_2 = (0.2879,\ 0.3687,\ 0.3841)'.$$

The statistics constructed in section 3 are used in the analysis of these data. The average chi-square test statistic $\tilde{X}^2_{\Omega D}$ is 3.1837 as given in (3.25). The conjecture statistic $\tilde{X}^2_{\Omega D}$ in (3.24) has the value 3.1871. Using the Moore-Penrose inverse, $\tilde{X}^2_D$ as used in SAS in (3.26) results in a value of 3.1521. The usual

Pearson statistic has the value of 6.807. Rao and Scott (1981) made use of the eigen values of the product of the inverse of covariance matrix under multinomial sampling and the covariance matrix under the specified design. They presented the statistic X^2/λ where $\lambda = \text{trace}\ (\Sigma_m^{-1}\hat{\Sigma})/I-1.$ This statistic has a value of 4.3806 with $\lambda = 2.3628$. Tables in Johnston and Kotz (1968) allows us to find the distribution of the Pearson Statistic, since the weights can be obtained from the calculated eigen values. The weights are 0.39, 0.21, 0.21 and 0.19. The Pearson statistic, X using these weights has a p-value less then .025. Assuming that X is distributed as a chi-square random variable with 2 degrees of freedom gives a p-value of approximately .007. The Wald test statistic Table 5.1 has a p-value less than .05.

5. DISCUSSION

A numerical comparison can be made between the results here and those obtained in Wilson and Koehler (1984). These data were analyzed in Wilson (1984) and Wilson and Koehler (1984). A summary of the test statistics values obtained are given in the following table.

Table 5.1

	Method	Test Statistic	Source
1.	Dirichlet multinomial model with normality assumptions and an assumed diagonal covariance matrix.	4.1881	Wilson (1984)
2.	Dirichlet Multinomial model with normality assumptions.	4.2079	Wilson (1984)
3.	A Wald test with no assumptions on the covariance structure.	4.16	Wilson & Koehler (1984)
4.	Model given in section 2.	3.1521	Section 3 (3.26)
5.	Approximate statistic to the model given in section 2.	3.1871	Section 3 (3.22)
6.	Pearson Statistic	6.807	Wilson & Koehler (1984)
7.	Eigen values (Rao & Scott)	4.3806	Rao & Scott

Similar results were obtained with other data sets when examined by these different methods. The indication is that the model given in this paper works fairly well in comparison to the Wald method. The statistic $\hat{X}^2_{\Omega DA}$ (conjecture) which well approximates the Wald techniques of method 4 in Table 5.1 requires only a knowledge of the variances under the particular design. It is not as restrictive as methods 1 and 2 and not as complicated in constructing as method 3. The advantage of the approximate statistics developed in this paper over the Dirichlet Multinomial technique of Wilson and Koehler (1984) and Brier (1980) is the fact that the statistics do not require constant design effects on the clusters within a particular subpopulation. Also these statistics can easily accommodate unequal cluster sample sizes. Further, this statistic does not require large amount of data as in the case with method 3.

The statistics constructed in section 3 differ from the techniques of Rao and Scott (1981) in the way in which the weights are obtained in their contribution to the formation of the test statistics. Also the techniques used to obtain these weights do not require any matrix inversion as in the case with the statistics obtained in obtaining the eigen values. Bedrick (1983) makes use of marginal design effects and cell design effects but the contribution of these design effect on the test statistics differs from the procedure adapted in this paper.

Work is being done by this author to get some guidance as to when some of these techniques should be employed. Rao and Thomas (1984) conducted a Monte Carlo study for the case of the goodness of fit problem, which has contributed to such guidance. However, this author is conducting a study to investigate the test of independence and to have some way of testing the fit of the model.

6. REFERENCES

Bedrick, E. J. 1983. Adjusted Chi-squared tests for cross-classified tables of survey data. Biometrika 70:591-595.

Bishop, M. M., Fienberg, S. E. and Holland, P. W. (1975), Discrete Multivariate Analysis: Theory and Practice, Cambridge, Massachusetts: MIT Press.

Brier, S. S. 1980. Analysis of contingency tables under cluster sampling.Biometrika 67:591-596.

Cramer, H. 1946. Mathematical Methods of Statistics. Princeton University Press, Princeton.

Rao, J. N. K. and Scott, A. J. 1981. The analysis of categorical datafrom complex surveys. J. Amer. Statist. Assoc. 76:221-230.

Rao, J. N. K. and Scott, A. J. 1984. On Chi-squared test for Multiway Contingency Tables with Cell proportions estimated from survey data. The Annals of Statistics 1984, Vol. 12, No. 1, 46-60.

Wald, A. 1943. Tests of Statistical hypotheses concerning severalparameters when the number of observations is large. Trans. Amer.Math. Soc. 54:426-482.

Wilson, J. R. 1984. Statistical methods for frequency data from complex sampling schemes. Ph.D. Dissertation, Ames Iowa State University, Ames, Iowa 50011.