χ^2-TESTING OF CATEGORICAL DATA FROM NESTED DESIGN USING THE CORRECTION FACTOR ESTIMATED FROM ANALYSIS OF VARIANCE COMPONENTS

Jai W. Choi and Robert J. Casady
National Center for Health Statistics

INTRODUCTION

In recent literature some statisticians discussed the effects of complex survey design on test statistic. Rao and Scott (1979) derived such effect from examining the eigenvalues of the matrix $\mathbf{A} = \mathbf{P}^{-1}\mathbf{V}$ where $\mathbf{P}$ is the covariance matrix under simple random sampling situation and $\mathbf{V}$ is that arising from actual complex sampling. Fellegi (1978) used an effective sample size in a test statistic, when the balanced repeated replication method is used to obtain the variance of data based on nonsimple random sampling. Cohen (1976), Altham (1976), and Choi (1980) used a model approach to find effects of cluster sampling on test statistics.

In this paper, we propose a device for the estimation of intracluster correlations in nested design for categorical data when usual analysis of variance terms are utilized. Using these results, we also propose a method to find a correction method for goodness of fit test statistic based on nested survey design.

In the first section, a nested random effect model is introduced. Sections 2 introduces the definitions and notations used in the succeeding sections of this paper and section 3 is allocated for the estimation of the intracluster correlation coefficient in the first stage cluster and that in the second stage cluster. In section 4, the adjustment of chi-squared goodness of fit test statistic is discussed. A simple numerical example is given in the last section.

1. NESTED MODEL

The purpose of this section is to discuss the nested random effect model for multivariate data from unbalanced design, which is analogous to multiway analysis of variance (ANOVA) model discussed in Schette (1959, p 248).

If y_{ijk} denotes the kth measurement of the j th secondary unit in the i th primary unit, we may write

$$y_{ijk} = \mu + c_i + t_{ij} + e_{ijk} \tag{1.1}$$

for $i=1,\dots,r$, $j=1,\dots,d_i$, and $k=1,\dots,m_{ij}$.

The usual assumptions for estimation are that (c_i), (t_{ij}), and (e_{ijk}) are independently identically distributed with zero mean and variances σ_T^2, σ_C^2, and σ_e^2 respectively and that (c_i), (t_{ij}) and (e_{ijk}) are uncorrelated.

The unbiased estimates of σ_C^2, σ_T^2, and σ_e^2 may be obtained from the linear combination of the mean square errors in usual ANOVA table so that

$$\sigma_Y^2 = \sigma_C^2 + \sigma_T^2 + \sigma_e^2 \tag{1.2}$$

which is the measure of reliability of quality of the given measurement of c_i and t_{ij}. Denote the intracluster correlation coefficient among member in the first stage cluster (or PSU) by

$$\rho_C = \frac{\sigma_C^2}{\sigma_Y^2} \tag{1.2a}$$

and the intracluster correlation coefficient among the members in the second stage cluster (segment) by

$$\rho_T = \frac{\sigma_T^2}{\sigma_Y^2} \tag{1.2b}$$

In single stage situation, Cohen (1976) and Altham (1976) presented maximum likelihood estimate of intracluster correlation using a probability model defining the relationship between members. Landis and Koch (1977) applied random effect model to one way analysis of variance in order to estimate intracluster correlation coefficient.

2. DEFINITIONS AND NOTATIONS.

This section outlines definitions and notations for categorical data which permits the estimation of variance components in a nested design situation. The variables that are considered for the multivariate case of q response categories are assumed to have multinomial distribution.

Suppose that a sample of n elementary units is selected from a three stage nested design with replacement, i.e. First, r PSU's are taken from R PSU's by PPS design and secondly, d_i segments randomly selected from D_i segments in the ith PSU for $i=1,\dots,r$. Thirdly, m_{ij} elements are randomly taken from M_{ij} units in segment j for $j=1,\dots,d_i$

Denote each element by y_{ijk} $(k=1,\dots,m_{ij})$, which have multinomial distribution with parameter $\boldsymbol{\pi} = (\pi_1,\dots,\pi_q)$ where π_h $(h=1,\dots,q)$ is the probability that randomly selected unit y_{ijk} falls in category h.

Define $y_{ijkh} = 1$ if the (ijk)th persons falls in category h and = 0 otherwise for $i=1,\dots,r$, $j=1,\dots,d_i$, $k=1,\dots,m_{ij}$, and $h=1,\dots,q$.

Let $y_{ijk} = (y_{ijk1}, y_{ijk2},\dots, y_{ijkq})$ be the vector of q indicator variables for the (ijk)th person, then $\sum_h y_{ijkh} = 1$ for all i, j, and k.

Under this situation, the standard assumptions for (y_{ijkh}) to have multinomial ANOVA model (1.1) is

$$E(y_{ijkh}) = \pi_h \tag{2.1}$$

$$\sigma_h^2 = \mathrm{var}(y_{ijkh}) = \pi_h(1-\pi_h) \tag{2.2}$$

Let $E(y_{ijkh}y_{ijk'h})$

$$= p(y_{ijkh} = y_{ijk'h} = 1) = \delta_{Thh} \text{ for } k\neq k' \quad (2.3)$$

$E(y_{ijkh}y_{ij'k'h})$

$$= p(y_{ijkh} = y_{ij'k'h} = 1) = \delta_{Chh} \text{ for } j\neq j' \quad (2.4)$$

Then it follows that for $h=1,\ldots,q$,

$$\delta_{Thh} = \rho_{Thh}\pi_h(1-\pi_h) + \pi_h^2 \quad \text{for } k\neq k' \quad (2.5)$$

$$\delta_{Chh} = \rho_{Chh}\pi_h(1-\pi_h) + \pi_h^2 \text{ for } j\neq j' \text{ and } k\neq k' \quad (2.6)$$

where ρ_{Chh} and ρ_{Thh} are the intracluster correlation coefficient for the members in the first stage cluster and that in the second stage cluster respectively for the hth response category, $h=1,\ldots,q$ As a result

$$\rho_{Thh} = \frac{\delta_{Thh} - \pi_h^2}{\pi_h(1-\pi_h)} \quad (2.7)$$

$$\rho_{Chh} = \frac{\delta_{Chh} - \pi_h^2}{\pi_h(1-\pi_h)} \quad (2.8)$$

Further $\sigma^2_{Ch} = \rho_{Chh}\,\pi_h(1-\pi_h)$, (2.9)

$$\sigma^2_{Th} = \rho_{Thh}\,\pi_h(1-\pi_h) \quad (2.10)$$

$$\sigma^2_{eh} = (1-\rho_{Chh}-\rho_{Thh})\,\pi_h(1-\pi_h), \quad (2.11)$$

and $\sigma^2_{Yh} = \sigma^2_{Ch} + \sigma^2_{Th} + \sigma^2_{eh}$, for $h=1,\ldots,q$. (2.12)

Thus, intracluster correlation coefficient within the PSU in (2.7) is

$$\rho_{Chh} = \frac{\sigma^2_{Ch}}{\sigma^2_{Ch} + \sigma^2_{Th} + \sigma^2_{eh}} \quad (2.13)$$

and intracluster correlation coefficient within the segment in (2.8) is

$$\rho_{Thh} = \frac{\sigma^2_{Th}}{\sigma^2_{Ch} + \sigma^2_{Th} + \sigma^2_{eh}} \quad (2.14)$$

The correlation structure from response categories can be developed by letting the pairwise probability of agreement on the classification of given persons between the h th and h' th response categories in multiple determinations be denoted by

$$\delta_{Chh'} = p(y_{ijkh} = y_{ij'k'h'} = 1) \quad (2.15)$$

$$= E(y_{ijkh}, y_{ij'k'h'}) \quad \text{for } i\neq i',\ j\neq j'$$

$$\delta_{Thh'} = p(y_{ijkh} = y_{ij\,k'h'} = 1) \quad (2.16)$$

$$= E(y_{ijkh}, y_{ij'k'h'}) \text{ for } i=i',\ j\neq j'$$

for $i=1,\ldots,r$, $j,j'=1,\ldots,d_i$, $k,k'=1,\ldots,m_{ij}$, and $h,h'=1,\ldots,q$, then it follows that for $h\neq h'$

$$\delta_{Chh'} = \rho_{Chh'}(\pi_h\pi_{h'}(1-\pi_h)(1-\pi_{h'}))^{\frac{1}{2}} + \pi_h\pi_{h'} \quad (2.17)$$

$$\delta_{Thh'} = \rho_{Thh'}(\pi_h\pi_{h'}(1-\pi_h)(1-\pi_{h'}))^{\frac{1}{2}} + \pi_h\pi_{h'} \quad (2.18)$$

where $\rho_{Chh'}$ is the within-PSU intracluster correlation coefficient for (h,h') th response categories and $\rho_{Thh'}$ is the within-segment intra-cluster correlation coefficient for the (h,h') response categories. As a result, for $h,h'=1,\ldots,q$, $(h\neq h')$

$$\rho_{Chh'} = \frac{\delta_{Chh'} - \pi_h\pi_{h'}}{(\pi_h(1-\pi_h)\,\pi_{h'}(1-\pi_{h'}))^{\frac{1}{2}}} \quad (2.19)$$

$$\rho_{Thh'} = \frac{\delta_{Thh'} - \pi_h\pi_{h'}}{(\pi_h(1-\pi_h)\,\pi_{h'}(1-\pi_{h'}))^{\frac{1}{2}}} \quad (2.20)$$

The assumptions (2.6) and (2.17) are true under a certain limited constraint and similarly for (2.5) and (2.18).

Those entire structures for $h,h'=1,\ldots,q$ can be summarized in the matrix notation written in boldface.

$$\mathbf{\Delta}_C = (\delta_{Chh'}) \quad (2.21)$$

$$\mathbf{\Delta}_T = (\delta_{Thh'}) \quad (2.22)$$

which denotes the qxq symmetric matrix of pairwise agreement and disagreement probabilities defined previously in (2.5) and (2.6) for $h=h'$ and (2.17) and (2.18) for $h\neq h'$. Write q x q matrices

$$\mathbf{\Phi}_C = (\rho_{Chh'}) \quad (2.22a)$$

$$\mathbf{\Phi}_T = (\rho_{Thh'}) \quad (2.22b)$$

$\mathbf{\Phi}_C$ the symmetric matrix with ρ_{Chh} $(h=h')$ intraclass correlation) on diagonal elements and $\rho_{Chh'}$ $(h\neq h')$ (interclass correlation) on the off-diagonal elements in PSU. Similarly $\mathbf{\Phi}_T$ is the symmetric matrix of correlations among members in the segment. Denote q x q diagonal matrix

$$\mathbf{\Lambda} \text{ with } \left(\sqrt{\pi_1(1-\pi_1)}\ \ldots\ \sqrt{\pi_q(1-\pi_q)}\right) \quad (2.22c)$$

on the main diagonal and 1 x q row vector $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_q)$. We have

$$\mathbf{\Delta}_C = \mathbf{\Lambda}\mathbf{\Phi}_C\mathbf{\Lambda} + \boldsymbol{\pi}'\boldsymbol{\pi} \quad (2.22d)$$

$$\mathbf{\Delta}_T = \mathbf{\Lambda}\mathbf{\Phi}_T\mathbf{\Lambda} + \boldsymbol{\pi}'\boldsymbol{\pi} \quad (2.22e)$$

It follows that

$$\mathbf{\Phi}_C = \mathbf{\Lambda}^{-1}(\mathbf{\Delta}_C - \boldsymbol{\pi}'\boldsymbol{\pi})\mathbf{\Lambda}^{-1} \quad (2.23)$$

$$\mathbf{\Phi}_T = \mathbf{\Lambda}^{-1}(\mathbf{\Delta}_T - \boldsymbol{\pi}'\boldsymbol{\pi})\mathbf{\Lambda}^{-1} \quad (2.24)$$

Thus the paramaters σ^2 are the main diagonal elements of q x q matrix

$$\mathbf{\Phi}_C = (\mathbf{\Delta}_C - \boldsymbol{\pi}'\boldsymbol{\pi}), \quad (2.24a)$$

σ^2_{Th} are the main diagonal elements of

$$\mathbf{\Phi}_T = (\mathbf{\Delta}_T - \boldsymbol{\pi}'\boldsymbol{\pi}), \quad (2.24b)$$

and σ^2_{eh} are the main diagonal elements of

$$\Phi_e = P - \Phi_C - \Phi_T \tag{2.24c}$$

$$P = D_\pi - \pi\pi' \tag{2.24d}$$

where D_π is the q x q diagonal matrix with elements of the vector π on the main diagonal. The measure of overall average intracluster correlation coefficient can be estimated by

$$\rho_C = \frac{tr(\Phi_C)}{tr(P)} \tag{2.25}$$

$$\rho_T = \frac{tr(\Phi_T)}{tr(P)} \tag{2.26}$$

tr (H) is the trace of H . One may observe that the only information required for the estimation is the diagonal elements of these matrices. Other types of ratio may also be considered such as determinants or the largest eigenvalues. A better picture may emerge when the relatioship between these values becomes known.

3. ESTIMATION OF INTRACLUSTER CORRELATION FROM ANOVA TERMS.

This section is concerned with multivariate analysis of variance calculation involving the sum of squares and their expected values that are used to estimate the variance components and hence the corresponding intracluster correlation coefficient of respective stages of clustering discussed in section 2. The notations are summarized in the table below.

Table 1 ANOVA table

Source	Mean Squares (MS)
Total (Y)	$\frac{1}{n-1}\sum_{i=1}^{r}\sum_{j=1}^{d_i}\sum_{k=1}^{m_{ij}}(\mathbf{y}_{ijk}-\bar{\mathbf{y}})(\mathbf{y}_{ijk}-\bar{\mathbf{y}})'$
1st stage cluster: PSU(C)	$\frac{1}{r-1}\sum_{i=1}^{r} m_i(\bar{\mathbf{y}}_i-\bar{\mathbf{y}})(\bar{\mathbf{y}}_i-\bar{\mathbf{y}})'$
2nd stage cluster: segment(T)	$\frac{1}{d-r}\sum_{i=1}^{r}\sum_{j=1}^{d_i} m_{ij}(\bar{\mathbf{y}}_{ij}-\bar{\mathbf{y}}_i)(\bar{\mathbf{y}}_{ij}-\bar{\mathbf{y}}_i)'$
Error(e):	$\frac{1}{n-d}\sum_{i=1}^{m}\sum_{j=1}^{d_i}\sum_{k=1}^{m_{ij}}(\mathbf{y}_{ijk}-\bar{\mathbf{y}}_{ij})(\mathbf{y}_{ijk}-\bar{\mathbf{y}}_{ij})'$

$$\mathbf{y}_{ijk} = (y_{ijk1}; \cdots , y_{ijkq})$$

$$d = \sum_{i=1}^{r} d_i \qquad m_i = \sum_{j=1}^{d_i} m_{ij} \qquad n = \sum_{i}^{r}\sum_{j=1}^{d_i} m_{ij}$$

$$\bar{\mathbf{y}} = \frac{\sum_{i=1}^{r}\sum_{j=1}^{d_i}\sum_{k=1}^{m_{ij}} \mathbf{y}_{ijk}}{n} \tag{3.1}$$

$$\bar{\mathbf{y}}_i = \frac{\sum_{j=1}^{d_i}\sum_{k=1}^{m_{ij}} \mathbf{y}_{ijk}}{m_i} \tag{3.1a}$$

$$\bar{\mathbf{y}}_{ij} = \frac{\sum_{k}^{m_{ij}} \mathbf{y}_{ijk}}{m_{ij}} \tag{3.1b}$$

m_{ij} is the number of persons in the (ij)th segment or ultimate sampling cluster. The expected values of the mean square errors are shown below

$$E(MS_Y) = (P - \theta\,\Phi_C - \zeta\,\Phi_T) \tag{3.1c}$$

$$E(MS_C) = (P + \alpha\,\Phi_C + \beta\,\Phi_T) \tag{3.2}$$

$$E(MS_T) = (P + \gamma\,\Phi_C + \delta\,\Phi_T) , \tag{3.3}$$

$$E(MS_e) = (P - \Phi_C) \tag{3.4}$$

where $\alpha = (E-r-(A/n))/(r-1)$
$\beta = (n-E-((B-A)/n))/(r-1)$
$\gamma = ((n-E)/\ (d-r))-1$ (3.5)
$\delta = (n-E)/(d-r)$
$\theta = A/n(n-1)$
$\zeta = (B-A)/\ n(n-1)$
P is given in (2.24d) (3.6)

$$A = \sum_{i=1}^{r}\sum_{j=1}^{d_i} m_{ij}(m_{ij}-1) \tag{3.7}$$

$$B = \sum_{i=1}^{r} m_i(m_i-1) \tag{3.8}$$

$$E = \sum_{i=1}^{r}\frac{1}{m_i}\sum_{j=1}^{d_i} m_{ij}^2 \tag{3.9}$$

More than one unbiased estimates for Φ_C, Φ_T, and Φ_e can be possible. A unique solution may be possible through multivariate least square method (suggested by Dr. Ron Forthofer, Univ. of Tx). A set of unbiased estimates are

$$\hat{\Phi}_T = \frac{\zeta\ MS_C - (\zeta+\beta)MS_e + \beta\ MS_Y}{\zeta(1+\alpha) + \beta(1-\theta)} \tag{3.11}$$

$$\hat{\Phi}_C = \frac{(1-\theta)MS_C + (\alpha+\theta)MS_e - (1+\alpha)MS_Y}{\zeta(1+\alpha) + \beta(1-\theta)} \tag{3.12}$$

$$\hat{\Phi}_e = MS_e - \hat{\Phi}_T \tag{3.13}$$

Thus, we can use the ANOVA mean square matrices to construct the unbiased estimates of these parameters. In particular, the unbiased estimates of the variance components due to PSUs as shown in (2.9) can be obtained by

$$\hat{\sigma}_C^2 = \mathrm{Diag}\,(\hat{\Phi}_C) \tag{3.14}$$

where Diag (H) **denotes the vector** determined by the main diagonal elements of the matrix H, the unbiased estimate of variance components due to segments as shown in (2.10) can be obtained by

$$\hat{\sigma}^2_T = \text{Diag}(\hat{\phi}_T) \quad (3.15)$$

and the unbiased estimate of residual error as shown in (2.11) can be obtained by

$$\hat{\sigma}^2_e = \text{Diag}(MS_e - \hat{\phi}_T) \quad (3.16)$$

The corresponding unbiased estimates of the total variance of each of the q responses as shown (2.12) can be obtained by

$$\hat{\sigma}^2_Y = \hat{\sigma}^2_C + \hat{\sigma}^2_T + \hat{\sigma}^2_e \quad (3.17)$$

The correlation matrix Φ_C and Φ_T shown in (2.23) and (2.24) may be estimated by

$$\hat{\Phi}_C = \Lambda^{-1}\hat{\phi}_C\Lambda^{-1} \quad (3.18)$$

$$\hat{\Phi}_T = \Lambda^{-1}\hat{\phi}_T\Lambda^{-1} \quad (3.19)$$

where Λ is the q x q matrix given in (2.22c).

Since the main diagonal elements in (3.18) and (3.19) reflect the intracluster correlation coefficients of the respective q response categories. An overall average level of them is of great interest, which is used in adjusting a test statistic for such clustering impacts in section 4.

Denote (2.25) and (2.26) by (3.20) and (3.21) below:

$$\tilde{\rho}_C = \frac{\sum_{h=1}^{q} \rho_{Chh}\sigma^2_{Yh}}{\sum_{h=1}^{q} \sigma^2_{Yh}} \quad (3.20)$$

$\tilde{\rho}_C$ is a weighted average of the cell correlation coefficients due to the clustering of PSU's with weights being the corresponding total variances of each category. Similarly denote a weighted average due to the clustering of segments in the PSU by

$$\tilde{\rho}_T = \frac{\sum_{h=1}^{q} \rho_{Thh}\sigma^2_{Yh}}{\sum_{h=1}^{q} \sigma^2_{Yh}} \quad (3.21)$$

An unbiased estimate of the numerator of $\tilde{\rho}_C$ can be obtained by $\mathbf{1}\,\hat{\sigma}^2_C{}' = \text{tr}(\hat{\phi}_C)$ and that of $\tilde{\rho}_T$ by $\mathbf{1}\,\hat{\sigma}^2_T{}' = \text{tr}(\hat{\phi}_T)$, where $\mathbf{1}$ is 1xq vector of 1's.

An unbiased estimate of the denominator of $\tilde{\rho}_C$ and $\tilde{\rho}_T$ can be obtained by $\mathbf{1}\,\sigma^2_Y{}' = \text{tr}(\hat{\mathbf{P}})$ where $\hat{\mathbf{P}} = \hat{\phi}_C + \hat{\phi}_T + \hat{\phi}_e$. Thus, consistent estimates of $\tilde{\rho}_C$ and $\tilde{\rho}_T$ are:

$$\hat{\rho}_C = \frac{\mathbf{1}\,\hat{\sigma}^2_C{}'}{\mathbf{1}\,\hat{\sigma}^2_Y{}'} \quad (3.22)$$

and

$$\hat{\rho}_T = \frac{\mathbf{1}\,\hat{\sigma}^2_T{}'}{\mathbf{1}\,\hat{\sigma}^2_Y{}'} \quad (3.23)$$

These results are applied to adjustment of test statistics in section 4. Cohen (1976) found the maximum likelihood estimate of intracluster correlation coefficient when the multinomial distribution is assumed. Landis and Koch (1977) used the average intracluster correlation in one stage clustering and applied to measure overall reliability for response categories.

4 IMPACTS OF NESTED DESIGN ON χ^2 TEST STATISTIC

The impacts of sample survey design on the test statistics are generally called design effect (Rao and Scott, 1979, Fellegi, 1978, Kish and Frankel, 1974). The design effect can be identified by investigating variance covariance structure of a statistic based on complex survey data.

The adjustment of goodness of fit test statistic based on the nested design is shown below.

Let $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_r$ be a set of r independent vectors of dimension q-1:

$$\mathbf{y}_i = (y_{i1}, \ldots, y_{iq-1}) \quad (4.1)$$

$$y_{ih} = \sum_{j=1}^{d_i} \sum_{k=1}^{m_{ij}} y_{ijkh} \quad (4.2)$$

i=1,..,r, and h=1,...,q-1, and y_{ijkh} is indicator variable defined in section 2.

$\mathbf{y} = (y_1, y_2, \ldots, y_{q-1})$, where $y_h = \sum_{i=1}^{r} y_{hi}$. $n = \sum_{h=1}^{q} y_h$.

$\hat{\boldsymbol{\pi}} = (\hat{\pi}_1, \ldots, \hat{\pi}_{q-1})$ where $\hat{\pi}_h = y_h / n$.

$$E(y_{ijkh}) = \pi_h \quad (4.2a)$$

$$E(y_{ijkh}, y_{i'j'k'h'}) = \begin{cases} \delta_{Chh'} & \text{for } j \neq j' \\ \delta_{Thh'} & \text{for } k \neq k' \\ \pi_h \pi_{h'} & \text{for } i \neq i' \end{cases} \quad (4.2b)$$

where $\delta_{Chh'}$ and $\delta_{Thh'}$ are given in (2.5) and (2.16) for h=h' and (2.17) and (2.18) for h≠h'.

The variance covariance matrix of $\mathbf{y}$ can be written as

$$\boldsymbol{\Sigma} = n\,P + A\,\boldsymbol{\Lambda}\Phi_T\boldsymbol{\Lambda} + (B-A)\,\Lambda\Phi_C\boldsymbol{\Lambda} \quad (4.2e)$$

of which the elements are: for h=h'

$$\sigma^2_h = n\pi_h(1-\pi_h) + \rho_{Thh}A\,\pi_h(1-\pi_h) + \rho_{Chh}(B-A)\pi_h(1-\pi_h),$$

$$\sigma_{hh'} = -n\pi_h\pi_{h'} + \rho_{Thh'}A(\pi_h(1-\pi_h)\pi_{h'}(1-\pi_{h'}))^{\frac{1}{2}}$$

$$+ \rho_{Chh'}(B-A)(\pi_h(1-\pi_h)\pi_{h'}(1-\pi_{h'}))^{\frac{1}{2}} \quad \text{for } h \neq h'$$

A and B are defined in (3.7) and (3.8), Λ in (2.22c), Φ_C and Φ_T in (2.19) and (2.20), and **P** in (2.24d) with only q-1 columns and rows. If $\rho_{Chh'} = \rho_C$ and $\rho_{Thh'} = \rho_T$ for h,h'= 1,...,q-1, Σ in (4.2e) reduces to

$$\tilde{\Sigma} = n\,\mathbf{P} + (g-n)\,\mathbf{b'b} \qquad (4.2f)$$

of which the elements are:

$$\sigma_h^2 = n\pi_h(1-\pi_h) + (g-n)\pi_h(1-\pi_h) \quad (h=h') \qquad (4.3)$$

$$\sigma_{hh'} = -n\pi_h\pi_{h'} + (g-n)\left(\pi_h(1-\pi_h)\pi_{h'}(1-\pi_{h'})\right)^{\frac{1}{2}} \quad (h\neq h')$$

$$\mathbf{b} = \left(\sqrt{\pi_1(1-\pi_1)}, \ldots, \sqrt{\pi_{q-1}(1-\pi_{q-1})}\right) \qquad (4.2g)$$

$$g = n\left(1 + \rho_T\frac{A}{n} + \rho_C\frac{(B-A)}{n}\right) \qquad (4.4)$$

ρ_C and ρ_T are the average intracluster correlation coefficients defined in (2.25) and (2.26) and estimated by (3.22) and (3.23).

If $m_{ij}=m$, and $d_i = d$, than g reduces to

$$g' = n\left(1+\rho_T(m-1) + \rho_C m(d-1)\right) \qquad (4.5)$$

When the sample included different sizes of clusters, the weighted average of m_{ij}'s and d_i's often gives a better result than using the largest values of m_{ij} and d_i if r is reasonably large. The weighting may be made according to the size of clusters (Choi, 1980).

If d =1, g' further reduces to $n(1+\rho_T(\bar{m}-1))$, it becomes one stage clustering situation. And furthermore, when $\bar{m}=1$ and $\bar{d}=1$, g' reduces to n, this is also true if $\rho_C=\rho_T=0$ regardless of the nature of nested design.

Using the results in (4.2f), the covariance matrix **V**, say, for the vector $\sqrt{n}\,(\hat{\pi}-\pi)$, can be written by

$$\mathbf{V} = \mathbf{P} + \frac{(g-n)}{n}\,\mathbf{b'b} \qquad (4.7)$$

where **b** is 1x(q-1) vector defined in (4.2g). The inverse of matrix **V** (see Donald Morrison p69) is given by

$$\mathbf{V}^{-1} = \mathbf{P}^{-1} - \frac{(g-n)/n}{1+\frac{(g-n)}{n}\,\mathbf{b}\,\mathbf{P}^{-1}\mathbf{b'}}\,\mathbf{P}^{-1}\mathbf{b'b}\,\mathbf{P}^{-1}$$

$$= \mathbf{P}^{-1} - \mathbf{Z}\ (\text{say}) \qquad (4.8)$$

where **Z** is so defined. In binomial situation, $\mathbf{V}^{-1}$ reduces to $n/(\pi(1-\pi)g)$. For an unbiased estimate $\hat{\pi}$ and for the hypothesis $\pi=\pi_o$ (specified), where $\pi_o = (\pi_{o1},\ldots,\pi_{oq-1})$, one can write a quadratic form

$$Q_{cc} = n\,(\hat{\pi}-\pi_o)\,\mathbf{V}^{-1}\,(\hat{\pi}-\pi_o)' \qquad (4.9)$$

Q_{cc} can be written into two terms: $Q_{cc} = Q_1 - Q_2$, where

$$Q_1 = n\,(\hat{\pi}-\pi_o)\,\mathbf{P}^{-1}\,(\hat{\pi}-\pi_o)' \qquad (4.9a)$$

$$Q_2 = n\,(\hat{\pi}-\pi_o)\,\mathbf{Z}\,(\hat{\pi}-\pi_o)' \qquad (4.9b)$$

Since $\mathbf{V}^{-1}$, $\mathbf{P}^{-1}$, and **Z** are positive semidefinite,

$$Q_{cc} \leq Q_1 \qquad (4.9c)$$

Q_1 is the usual form of goodness of fit test statistic, i.e.

$$Q_1 = \sum_{i=1}^{q} \frac{n\,(\hat{\pi}_i - \pi_{oi})^2}{\pi_{oi}} \qquad (4.10)$$

One may observe that $Q_2 > 0$ under the situation $(\rho_T A + \rho_C(B-A)) > 0$. The equality in (4.9c) holds if $\rho_C = \rho_T = 0$. In most practical cases, $\hat{\rho}_C > 0$ and and $\hat{\rho}_T > 0$. If $\hat{\rho}_C$ and/or $\hat{\rho}_T$ are negative, these estimates could be replaced by zero for practical application.

Q_1 is the maximum value of chi-square statistic obtainable regardless of the nature of dependence between members in the cluster. If Q_1 is not significant when referred to $\chi^2(q-1)$, the hypothesis $\pi=\pi_o$ should be accepted whatever the Q_{cc} value is. But when Q_1 is significant, Q_{cc} should be adjusted for design effect in order to find the actual significance of Q_{cc}.

If the full covariance matrix **V** is known, one can always construct an asymptotically correct Wald statistics. Rao and Scott (1979) introduced a simple approximation to the distribution of Q that required only very limited information about **V**, that is, $tr(P^{-1}V)/(q-1) = \bar{\lambda}$. $\bar{\lambda} \geq 1$ for the clustered data. $\bar{\lambda}$ can be written as

$$\bar{\lambda} = \frac{tr(\mathbf{P}^{-1}\mathbf{V})}{q-1} = 1 + \frac{(g-n)}{n(q-1)}\,tr(\mathbf{P}^{-1}\mathbf{bb'})$$

$$= 1 + \frac{(g-n)}{n(q-1)}\sum_{h=1}^{q-1}\left(\frac{b_h^2}{\pi_h} + \frac{b_h}{\pi}\sum_{i=1}^{q-1} b_i\right) \qquad (4.11)$$

where g is given in (4.4), **b** in (4.2g), and $\pi = 1 - \sum_{i=1}^{q-1}\pi_i$.

Thus, the modified statistic is

$$Q_{rao} = \frac{Q_1}{\bar{\lambda}} \qquad (4.12)$$

Both Q_{cc} and Q_{rao} can be considered as a $\chi^2(q-1)$ random variable. The expected values of Q's are same as $\chi^2(q-1)$. The variance is also same under certain conditions (Rao and Scott, 1979).

A consistent estimate $\hat{g}$ of g can be obtained by substituting $\hat{\rho}_T$ and $\hat{\rho}_C$ in g. Thus, it gives $\hat{\lambda}$.

In order to find Q_2, the matrix $\mathbf{Z}$ should be known:

$\mathbf{Z} = G\,\mathbf{P}^{-1}\mathbf{b}'\mathbf{b}\mathbf{P}^{-1}$ where $G = (g-n)/(n+(g-n)f)$, and $f = \mathbf{b}\,\mathbf{P}^{-1}\,\mathbf{b}'$. The elements of $\mathbf{Z}$ are:

$$z_{hh'} = G\left(\frac{b_h}{\pi_h} + \frac{1}{\pi}\sum_{h=1}^{q-1} b_h\right)^2 \quad \text{for } h=h' \tag{4.13}$$

$$z_{hh'} = G\left(\frac{b_h}{\pi_h} - \frac{1}{\pi}\sum_{h=1}^{q-1} b_h\right)\left(\frac{b_{h'}}{\pi_{h'}} - \frac{b}{\pi}\sum_{h'=1}^{q-1} b_{h'}\right) \quad \text{for } h\neq h'$$

It is generally true that $f > 1$ and consequently $1 > G > 0$. $G = 0$ if $\rho_C = \rho_T = 0$.

Using these scalar forms, one can avoid matrix operations in order to obtain Q_2.

For Q_{cc}, the design effect can be adjusted by subtracting Q_2 from a conventional chi-square test statistic Q_1. Only information reguired is Q_2 for such adjustment. On the other hand, for Q_{rao}, the knowledge of a full variance covariance matrix is required to correct the test statistic Q_1.

5 EXAMPLE

A simple example is presented here for an illustrative purpose. Suppose that the sampling is done with replacement. Three PSU's are selected by PPS design. Here design feature does not matter as far as the models fit for pairwise relationship in the cluster. The sample segments in the PSU are randomly selected. Thirdly the elementary units (e.u.) in the segment are also randomly selected. These steps are illustrated in the table below.

PSU No.	No. Seg	No.of Elem.	$\mathbf{y}_{ijk}$	$\mathbf{y}_{ij}$	$\mathbf{y}_{i}$
1	$d_1=2$	$m_{11}=2$	(0,1,0),(0,1,0)	(0,1,0)	$(\frac{1}{2},\frac{1}{2},0)$
		$m_{12}=2$	(1,0,0),(1,0,0)	(1,0,0)	
2	$d_2=2$	$m_{21}=1$	(1,0,0)	(1,0,0)	(1,0,0)
		$m_{22}=2$	(1,0,0),(1,0,0)	(1,0,0)	
3	$d_3=1$	$m_{31}=2$	(0,1,0),(0,1,0)	(0,1,0)	(0,1,0)

$\bar{\mathbf{y}} = (5/9,\ 4/9,\ 0)$, $n=9$, $n(n-1)=72$, $A=8$, $B=20$, $(B-A)=12$, $E=17/3$

$$MS_y = \frac{5}{18}\begin{vmatrix}1 & & \\ & 1 & \\ & & 0\end{vmatrix} \quad MS_C = \frac{11}{18}\begin{vmatrix}1 & & \\ & 1 & \\ & & 0\end{vmatrix} \quad MS_T = \frac{1}{2}\begin{vmatrix}1 & & \\ & 1 & \\ & & 0\end{vmatrix}$$

Here MS(error)= **0** in order to simplify the calculation although it is not realistic.

$\alpha = 8/9$, $\beta = 1$, $\gamma = 2/3$, $\delta = 5/3$, $\theta = 1/9$, $\zeta = 1/6$, $\beta(1-\theta)+\zeta(1+\alpha) = 65/54$

$$\hat{\boldsymbol{\phi}}_C = \frac{41}{130}\begin{vmatrix}1 & & \\ & 1 & \\ & & 0\end{vmatrix} \quad \hat{\boldsymbol{\phi}}_T = \frac{1}{65}\begin{vmatrix}1 & & \\ & 1 & \\ & & 0\end{vmatrix} \quad \hat{\boldsymbol{\phi}}_e = \frac{-1}{65}\begin{vmatrix}1 & & \\ & 1 & \\ & & 0\end{vmatrix}$$

$\hat{\rho}_C = 2/41 \quad \hat{\rho}_T = 1 \quad \pi_o = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$

$Q_1 = 14/3 = 4.666$,

$Q_2 = 3.5656$,

$Q_{cc} = Q_1 - Q_2 = 1.1010$ (2 d.f.)

$\bar{\lambda} = 3.7534$,

$Q_{rao} = 4.666/3.7534 = 1.2433$ (2 d.f.)

Thus, the data fit to the specified value π_o for both procedures. In this case, Q_{cc} and Q_{rao} gives approximately same results. However, the Q_2 is generally effected by the redundant cell deleted and thus may have to be adjusted for other situations.

REFERENCES

Altham, P.M.E. (1976). Discrete variable analysis for individuals grouped into families. Biometrika 63, 263-69.

Choi, J.W. (1980). Unpublished Thesis, University of Minnesota.

Cohen, J. E.(1976). The distribution of the chi-squared statistic under clustered sampling from contingency tables. Journal of the American Statistical Association 71, 665-70.

Morrison, D. F. (1976). Multivariate Statistical Method. McGraw-Hill Book Co.

Fellegi, I. P. (1978). Approximate test of independence and goodness of fit based on stratified multistage samples. Survey Methodology, Statistics Canada

Fienberg, S.E. (1979). The use of chi-squared statistics of categorical data problems. Journal of the Royal Statistical Society B. 41, 1, 54-64.

Kish, L and Frankel, M.R. (1974). Inference from complex samples. Journal of the Royal Statistical Society B, 36, 1-37.

Landis, J. R. and Koch. G.G. (1977). A one-way components of variance model for categorical data. Biometrics 33, 671-679.

Rao, J.N.K. and Scott, A.J. (1979). Chi-square tests for analysis of categorical data from complex surveys. Proceedings in Social Statistics, ASA meeting, Washington, D.C.

Scheffe, H. (1959). The analysis of variance, John Wiley and Sons.