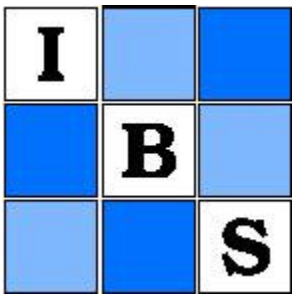

BIOMETRICS 45, 979–996
September 1989

A Reduction Factor in Goodness-of-Fit and Independence Tests for Clustered and Weighted Observations

Jai Won Choi

National Center for Health Statistics, 3700 East–West Highway, Hyattsville, Maryland 20782, U.S.A.

and

Richard B. McHugh

University of Minnesota, Minneapolis, Minnesota 55455, U.S.A.

SUMMARY

Situations often arise in a large-scale household survey where a complex probability sample of clusters rather than of individuals is drawn from a large population. Typically, the clusters of such complex samples include a number of correlated members. The responses of these members are then weighted to obtain estimates for the population. Such weighted data are commonly published by the National Center for Health Statistics and other U.S. federal agencies.

Frequently, problems arise when such data are tested by usual chi-square test statistics for goodness of fit or independence. Researchers have discovered that the usual chi-square tests provide spuriously inflated results when applied to cluster samples and that new methods are required to correct such problems.

This paper proposes a strategy for a goodness-of-fit or independence test based on correlated and weighted data arising in cluster samples, and provides a factor that validly reduces the inflation of the usual chi-square statistics.

This method is applied to the chronic condition data collected from the St Paul–Minneapolis, Minnesota, primary sampling unit (PSU) during the 1975 National Health Interview Survey (NHIS). This analysis, together with simulation studies presented elsewhere, provides evidence that the usual chi-square statistics from such data can be corrected for the impacts of clustering and weighting by use of the proposed reduction factor.

1. Introduction

Large-scale household surveys, such as the National Health Interview Survey (NHIS), use probability samples of clusters. The data from such complex samples are then weighted to yield consistent estimates of population parameters. For example, since 1957, weighted data from clustered samples have been published by the National Center for Health Statistics (NCHS).

Several authors have recognized that severe errors in analysis can arise when the data from such complex samples are subjected to the usual chi-square tests of independence or goodness of fit. Efforts to correct for spurious inflation in such tests have been based on two approaches. The design-based approach provides inferences with respect to the sampling distribution of estimates over repetitions of the same sample design. Such an approach has been used by Fellegi (1980), Holt, Scott, and Ewings (1980), Rao and Scott (1981, 1984), Bedrick (1983), Landis et al. (1984), Koch, Freeman, and Freeman (1975), and Fay (1985).

Key words: Estimation; Goodness-of-fit test; Independence test; Intracluster correlation; One- or two-stage clustering; Probability model; Reduction factor; Weighted counts.

The model-based approach—employed by Altham (1976), Cohen (1976), Brier (1980), and Fienberg (1979)—postulates a probability distribution to model the sample data. This paper extends the model-based approach to deal with complex sample data such as that obtained from the NHIS. The published data tapes from the NHIS contain the essential information required for the method presented here—individual weights and cluster sizes. With this information available, the present approach permits the computation of a factor that can be applied to the usual Pearson chi-square statistic, X^2, in order to correct the impact of sample weighting and clustering on that statistic.

Following the introduction, Section 2 presents the model and related material for single-stage clustering. In Section 3, estimation and tests of goodness of fit and independence are considered. A numerical example using the 1975 NHIS, in which the household is the cluster, is presented in Section 4. Extensions are given in Section 5 to different sizes of clusters, to two-stage clustering, and to multivariate analysis of variance estimation of intracluster correlation. General comments are presented in Section 6.

2. One-Stage Clustering

2.1 *The Available Information*

Based on the NCHS data tapes as a prototype, the information available for statistical inference consists of (i) the individual sample weights; (ii) the cluster membership; (iii) the varying sizes of the clusters; (iv) the fact that usually more than one element in each cluster is measured; and (v) the condition that measurements are categorized into cells. As an example of (iv) and (v), two or more persons in a household may be interviewed as to age and their ages then dichotomized into less than 45 years or 45 and older.

2.2 *The Weights*

Since most NCHS samples are selected by drawing sample units with equal probability from sampled clusters at the last stage, and by drawing the clusters using probability proportional to population size (pps) at each stage before the last, overall selection probabilities for every individual unit are expected to be equal or self-weighting (Kendall and Stuart, 1968, p. 195). In fact, most government surveys are based on self-weighting sample designs. In cases where the survey weights are taken to be the inverse of the inclusion probabilities, the weights of a self-weighting sample are constant.

The sample weight may reflect not only the probability of a person's inclusion in the sample, but also the composition of the current U.S. population. For example, in the NHIS sample, the original weights were adjusted to be consistent with the person's age–sex–race group of the current U.S. population.

Bryant, Baird, and Miller (1971) described such weighting procedures and summarized the results for the National Health Examination Surveys, and Bean (1973) did likewise for the NHIS data. The details of the computing method of the NHIS sample weighting have also been presented in an unpublished document by NCHS (1975). The weights are further discussed in Section 3.2.

2.3 *The Model*

The finite population U of interest is decomposed into A exclusive and exhaustive clusters $U_1, \ldots, U_i, \ldots, U_A$. The ith cluster U_i consists of B_i final units as $U_i = (U_{i1}, \ldots, U_{ij}, \ldots, U_{iB_i})$. Within the final unit, U_{ij}, the generic measurement cell will be denoted by h, with a total of r such cells, $h = 1, \ldots, r$. Let $N = \sum_{i=1}^{A} B_i$ denote the total number of population units.

The indicator variables of the model are defined as follows:

$$x_{ijh} = \begin{cases} 1 & \text{if the response of the unit } U_{ij} \text{ falls into the } h\text{th cell} \\ 0 & \text{otherwise} \end{cases} \tag{2.1}$$

for $i = 1, \ldots, A; j = 1, \ldots, B_i$; and $h = 1, \ldots, r$.

Denote the vector of population cell frequencies by $\mathbf{Y}' = (Y_1, \ldots, Y_r)$, and the corresponding proportions by the vector $\boldsymbol{\pi}' = (\pi_1, \ldots, \pi_r)$, where

$$\pi_h = \frac{Y_h}{N}, \qquad \text{where } Y_h = \sum_{i=1}^{A} \sum_{j=1}^{B_i} x_{ijh} \quad \text{with} \quad \pi_h > 0 \quad \text{and} \quad \sum_{h=1}^{r} \pi_h = 1. \tag{2.1a}$$

The model-based approach assumes that the x_{ijh} are the realized random variables for the unit U_{ij} that are related by the following mathematical model (using the symbol E for expectation operator):

$$\begin{aligned} \mathrm{E}(x_{ijh}) &= \pi_h; \\ \mathrm{E}(x_{ijh}, x_{i'j'h'}) &= p_{hh'} \qquad \text{if} \quad i = i', j \neq j'; \end{aligned} \tag{2.2}$$

where

$$p_{hh'} = \begin{cases} \theta\pi_h + (1-\theta)\pi_h^2 & \text{for} \quad h = h' \\ (1-\theta)\pi_h\pi_{h'} & \text{for} \quad h \neq h'. \end{cases} \tag{2.3}$$

Let θ denote the common intraclass correlation ($0 \leqslant \theta \leqslant 1$) regardless of the size of cluster and cells, corresponding to all possible pairs of units in a cluster. Assume that the probability of a randomly selected unit falling into the hth cell is π_h. Then, in (2.3), p_{hh} is the probability that the first member of a pair falls in cell h with probability π_h and the second member also falls in cell h with probability $[\theta \cdot 1 + (1-\theta)\pi_h]$. Similarly, in (2.3), $p_{hh'}$ is the probability that one member of a pair falls in cell h with probability π_h and the other in cell h' with probability $(1-\theta)\pi_{h'}$ for $h \neq h'$. Note that the θ in the model does not depend on the labels of the cluster and cell.

The joint expectation (2.2) is for $i = i'$. For $i \neq i'$, we have $\mathrm{E}(x_{ijh}, x_{i'j'h'}) = \pi_h\pi_{h'}$ when the clusters are independent, and one member came from cluster i and the other from cluster i'. However, because one person cannot be classified into two cells, we have $\mathrm{E}(x_{ijh}, x_{i'j'h'}) = 0$ for $i = i', j = j'$, and $h \neq h'$.

The model (2.3) may not be suitable if the clusters are obviously different. Cohen (1976) used the model (2.3) for clusters of two members. The model for clusters of varying sizes is discussed in Section 3.5 as an extension for more general applications.

2.4 *Sample*

Statistical inference is to be made on the basis of a sample S of negligible sampling fraction denoted by $S = \{(i, j): i \in S^*, j \in S_i\}$, where S^* is a sample of a clusters, and S_i is a sample of b_i units in the ith cluster. Here, the sample clusters are indexed by i as $S^* = (u_1, \ldots, u_i, \ldots, u_a)$, and the sampled units in the ith cluster are indexed by j as $S_i = (u_{i1}, \ldots, u_{ij}, \ldots, u_{ib_i})$. Note that the indexes i and j for the population units U_{ij} are not the same as those used for the sample units u_{ij}.

The definition (2.1) of random variables x_{ijh} for the population remains the same for the sample units u_{ij}, where $i = 1, \ldots, a; j = 1, \ldots, b_i$; and $h = 1, \ldots, r$. We do not specify the sampling design except that it is a probability measure on the set of all possible samples.

The cluster sizes are assumed known, and the sampling and weighting are assumed to be independent of the realization of the cell counts. The (i, j)th sample response is multiplied by the weight w_{ij} in the estimation process.

3. Inference in One-Stage Clustering

3.1 *Estimation of the* π_h

The vector $\boldsymbol{\pi}$ of population cell proportions, $\pi_1, \ldots, \pi_{r-1}$, deleting one redundant cell, can be estimated unbiasedly using the sample weights w_{ij} of Section 2.2 and the indicator variables x_{ijh} of (2.1), as follows.

The vector of weighted cell counts for the ith cluster S_i will be denoted by $\mathbf{y}_i' = (y_{i1}, \ldots, y_{ir-1})$, where y_{ih} is defined by

$$y_{ih} = \sum_{j=1}^{b_i} w_{ij} x_{ijh}. \tag{3.1}$$

Hence, for the hth cell, the weighted cell count is

$$y_h = \sum_{i}^{a} \sum_{j}^{b_i} w_{ij} x_{ijh}.$$

The total of the weights in the sample will be denoted by $y = \sum_i^a \sum_j^{b_i} w_{ij}$. Then the estimator of $\boldsymbol{\pi}' = (\pi_1, \ldots, \pi_h, \ldots, \pi_{r-1})$ is denoted by $\hat{\boldsymbol{\pi}}' = (\hat{\pi}_1, \ldots, \hat{\pi}_h, \ldots, \hat{\pi}_{r-1})$, where the unbiased estimator, $\hat{\pi}_h$, is defined as

$$\hat{\pi}_h = \frac{y_h}{y}. \tag{3.2}$$

As noted in Section 2.2, the weights w_{ij} are the inverses of the corresponding selection probabilities. Hence, the weights in such cases are fixed since the selection probabilities of individual units are known. As the w_{ij}'s and b_i's are fixed numbers, y is also fixed. Actually, y is the population total N for a pps sample.

In some surveys, the weights are further adjusted by post-stratified ratios or some other methods involving sample values. Although this adjustment may cause variation in the weights from person to person, even for self-weighting samples, as discussed in Section 2.2, such variation is small.

One source of verification is provided by a simulation study (Choi, unpublished Ph.D. thesis, University of Minnesota, 1980) in which four sets of weights were randomly generated with different variation for a set of unweighted data. Regardless of this variation in the weights, the study produced approximately the same variance (and chi-square test statistics of Sections 3.4 and 3.5) when the coefficient of variation of the weights remained less than 30%.

Moreover, the adjusted weights remained nearly the same as the original weights when the post-stratified ratios deviated slightly from unity. In the adjustments of the original NHIS weights, the post-stratified ratios were used and differed little from unity. The assumption of fixed weights thus appears robust for statistical inferences in the case of NHIS data. In the derivations following, we assume that the weights are fixed. The case of ratios substantially different from unity is discussed further in Section 6.

3.2 *Precision of the* $\hat{\pi}_h$

The covariance matrix of the $\hat{\pi}_h$'s of (3.2) can now be derived as follows. A direct calculation will show that the covariance matrix $\mathbf{V}_i$ of the y_{ih} of (3.1) is given by

$$\operatorname{var}(y_{ih}) = \sum_{j \neq j'}^{b_i} w_{ij} w_{ij'} (p_{hh'} - \pi_h^2) + \pi_h(1 - \pi_h) \sum_{j=1}^{b_i} w_{ij}^2 \tag{3.3}$$

on the main diagonal, and

$$\operatorname{cov}(y_{ih}, y_{ih'}) = \sum_{j \neq j'}^{b_i} w_{ij} w_{ij'} (p_{hh'} - \pi_h \pi_{h'}) - \pi_h \pi_{h'} \sum_{j=1}^{b_i} w_{ij}^2 \tag{3.4}$$

on the off-diagonal, where $p_{hh'}$ is given in (2.3).

Since the clusters $S_1, S_2, \ldots, S_a$ in the sample S are assumed to be independent, we can replace $p_{hh'}$ by the model (2.3) in equations (3.3) and (3.4), and sum over all clusters to obtain the covariance matrix $\mathbf{V}$ of $\hat{\boldsymbol{\pi}}$:

$$\operatorname{var}(\hat{\pi}_h) = \frac{G}{y} \frac{\pi_h(1 - \pi_h)}{y} \tag{3.5}$$

on the main diagonal and

$$\operatorname{cov}(\hat{\pi}_h, \hat{\pi}_{h'}) = \frac{G}{y} \frac{(-\pi_h \pi_{h'})}{y} \tag{3.6}$$

on the off-diagonal, where

$$G = \theta \sum_{i}^{a} \sum_{j \neq j'}^{b_i} w_{ij} w_{ij'} + \sum_{i}^{a} \sum_{j}^{b_i} w_{ij}^2. \tag{3.7}$$

(3.5) and (3.6) are the regular multinomial variance and covariance, respectively, multiplied by the known factor G/y. G shows the combination of the weighting and correlation effects. Note that the multiplier G/y^2 is bounded as $0 < G/y^2 \leqslant 1$ where the equality holds if $\theta = 1$ and $a = 1$.

3.3 *Chi-Square Test for Goodness of Fit*

Consider the null hypothesis H_0: $\boldsymbol{\pi} = \boldsymbol{\pi}_0$, where $\boldsymbol{\pi}_0$ is specified. We assume that, under certain restrictive conditions on weights, the vector $\hat{\boldsymbol{\pi}}$ of cell proportions from the a independently distributed random clusters follows

$$\sqrt{y}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0) \rightarrow \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}) \quad \text{for large } a, \tag{3.8}$$

under the null hypothesis H_0. Here, $\boldsymbol{\Sigma}$ is the covariance matrix $\mathbf{V}$, given in (3.5) and (3.6), multiplied by y, and a nondegenerate matrix; "$\rightarrow \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$" means asymptotically distributed as multivariate normal with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$. Here the asymptotic normality holds for large a as large a implies large y.

The varying sizes of clusters would not disturb this asymptotic property under the model (2.3) as long as the dimension of the response vectors from these clusters remains the same. In fact, regardless of the distribution of the independent random variables, the sum of these independent random variables is asymptotically normal (Cramér, 1946, p. 214).

The convergence in distribution (3.8) is the only necessary condition required to derive the asymptotic distribution of Q shown below, and assures that the quadratic form Q is asymptotically distributed as chi-square with $r - 1$ degrees of freedom. The general Wald statistic testing H_0 (Wald, 1943) is

$$Q = y(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0)' \, \boldsymbol{\Sigma}^{-1} (\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0) \rightarrow \chi^2_{r-1} \tag{3.9}$$

for large a. Equation (3.9) can be rewritten as

$$Q = \frac{y}{G} X^2 \qquad \text{where} \quad X^2 = \sum_{h}^{r} y \frac{(\hat{\pi}_h - \pi_{h0})^2}{\pi_{h0}}, \tag{3.10}$$

where X^2 is the usual Pearson goodness-of-fit test statistic, and y/G is

$$\frac{y}{G} = \frac{\sum_{i=1}^{a} \sum_{j=1}^{b_i} w_{ij}}{\theta \sum_{i=1}^{a} \sum_{j \neq j'}^{b_i} w_{ij} w_{ij'} + \sum_{i=1}^{a} \sum_{j=1}^{b_i} w_{ij}^2}. \tag{3.11}$$

We call y/G a reduction factor because it reduces a usual chi-square statistic X^2 to a smaller one since $1/y \leqslant y/G \leqslant 1$. The equality in the upper bound holds for $\theta = 0$ and $w_{ij} = 1$. The equality in the lower bound holds for $\theta = 1$ and $a = 1$.

This suggests that the chi-square statistic based on correlated and/or weighted data should be corrected by multiplying the statistic by the reduction factor of y/G. The factor y/G may be simplified as shown in Section 3.5 when the average of weights is used.

The quadratic form Q takes into account both the squared random deviations and the impact of clustering and weighting. The adjustment factor y/G corrects those clustering and weighting impacts on the test statistic X^2, when X^2 is based on correlated and weighted observations.

The asymptotic distribution similar to (3.10) for unweighted data has been given previously (Cohen, 1976).

In (3.8), we note that the covariance matrix is $\boldsymbol{\Sigma} = \mathbf{P}(G/y)$, where $\mathbf{P} = (\mathbf{D}_{\pi} - \boldsymbol{\pi}\boldsymbol{\pi}')$ is the covariance matrix that would be obtained if the cell counts came from a simple random sample and $\mathbf{D}_{\pi}$ is the diagonal matrix with the vector $\boldsymbol{\pi}$. Note that the usual chi-square statistic X^2 shown in equation (3.10) can be expressed as

$$X^2 = y(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0)' \mathbf{P}_0^{-1} (\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0), \tag{3.12}$$

where $\mathbf{P}_0$ is the value of $\mathbf{P}$ for $\boldsymbol{\pi} = \boldsymbol{\pi}_0$. The quadratic form (3.12) is distributed as

$$\sum_{h}^{r-1} \lambda_{0h} Z_h^2, \tag{3.13}$$

where $Z_1^2, \ldots, Z_{r-1}^2$ are asymptotically independent χ_1^2 random variables, and the λ_{0h} are the eigenvalues of the matrix product $\mathbf{P}_0^{-1} \boldsymbol{\Sigma}_0$, where $\mathbf{P}_0$ and $\boldsymbol{\Sigma}_0$ are the covariance matrices of $\mathbf{P}$ and $\boldsymbol{\Sigma}$, respectively, under H_0 (Rao and Scott, 1981, Theorem 1).

However, this matrix product $\mathbf{P}_0^{-1} \boldsymbol{\Sigma}_0$ can be written as $\mathbf{P}_0^{-1}(\mathbf{P}_0 G/y) = \mathbf{I}(G/y)$, since $\boldsymbol{\Sigma}_0 = \mathbf{P}_0(G/y)$. Hence, $\lambda_{0h} = (G/y)$ for all nonzero eigenvalues of $\mathbf{I}(G/y)$, and $\mathbf{I}$ is the identity matrix of dimension $r - 1$. Thus, the usual chi-square statistic X^2 is the χ_{r-1}^2 random variable when X^2 is divided by the reduction factor G/y, that is, $Q = X^2/(G/y) = X^2(y/G)$. Bishop, Fienberg, and Holland (1975, p. 473) prove that when $G/y = 1$, the quadratic form Q has the χ_{r-1}^2 distribution asymptotically.

It will be recalled that the derivation of the above goodness-of-fit test assumed fixed weights. It is important to note that in certain applications, departure from this assumption can lead to substantial test inaccuracy. In particular, the goodness-of-fit test will be nonrobust if the π_{h0} depend on variables also employed in post-stratification. For example, if the hypothesized distribution included separate cells for males and females according to some other characteristics, and if adjustments to the distribution by gender were included in the post-stratification, the result might be severely in error.

However, the test of independence presented below is much less affected by such violations of the assumption of fixed weights. In addition, the test of independence is used much more frequently than the goodness-of-fit test.

3.4 *Chi-Square Test of Independence*

Consider the problem of an independence test for a general null hypothesis $\mathbf{f}(\boldsymbol{\pi}) = \mathbf{0}$. In particular, the hypothesis of independence between the R rows and C columns in a two-way table can be written as H_0: $f_{rc}(\boldsymbol{\pi}) = \pi_{rc} - \pi_{r+}\pi_{+c} = 0$ for $r = 1, \ldots, (R - 1)$ and

$c = 1, \ldots, (C-1)$, where population cell vector $\boldsymbol{\pi}' = (\pi_{11}, \pi_{12}, \ldots, \pi_{RC-1})$, π_{rc} is the population proportion for the (r, c)th cell,

$$\pi_{r+} = \sum_{c=1}^{C} \pi_{rc}, \qquad \pi_{+c} = \sum_{r=1}^{R} \pi_{rc}, \qquad \sum_{r=1}^{R} \sum_{c=1}^{C} \pi_{rc} = 1.$$

We assume that the $F_{th}(\boldsymbol{\pi}) = \partial f_t(\boldsymbol{\pi})/\partial \pi_h$ are continuous in the neighborhood of the true $\boldsymbol{\pi}$ for $t = 1, \ldots, T = (R-1)(C-1)$, and $h = 1, \ldots, (RC-1)$. Denote the matrix of the partial derivatives $F_{th}(\boldsymbol{\pi})$ evaluated at $\boldsymbol{\pi}$ by the $T \times (RC-1)$ matrix $\mathbf{F}(\boldsymbol{\pi}) = \mathbf{F}$ of rank T.

Then from the linear approximation $\mathbf{f}(\hat{\boldsymbol{\pi}}) \approx \mathbf{f}(\boldsymbol{\pi}) + \mathbf{F}(\boldsymbol{\pi})(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi})$, and the usual regularity conditions (for example, Bishop et al., 1975, p. 509), we assume that the random vector

$$\sqrt{y}[\mathbf{f}(\hat{\boldsymbol{\pi}}) - \mathbf{f}(\boldsymbol{\pi})] \to \mathrm{N}(\mathbf{0}, \mathbf{FMF}') \quad \text{under } H_0, \tag{3.14}$$

where $\mathbf{f}(\boldsymbol{\pi})' = (f_1(\pi), \ldots, f_T(\pi))$ is a $(1 \times T)$ vector, $\mathbf{f}(\hat{\boldsymbol{\pi}})$ is a consistent estimator of the vector $\mathbf{f}(\boldsymbol{\pi})$, $\mathbf{FMF}'$ is the $T \times T$ covariance matrix of the $\sqrt{y}\mathbf{f}(\hat{\boldsymbol{\pi}})$, and $\mathbf{M}$ is the $(RC-1) \times (RC-1)$ covariance matrix of $\hat{\boldsymbol{\pi}}$.

If we have a consistent estimator $\hat{\mathbf{F}}\hat{\mathbf{M}}\hat{\mathbf{F}}'$ available, the Wald statistic to test H_0 is written as

$$Q_T = y\mathbf{f}(\hat{\boldsymbol{\pi}})'(\hat{\mathbf{F}}\hat{\mathbf{M}}\hat{\mathbf{F}}')^{-1}\mathbf{f}(\hat{\boldsymbol{\pi}}). \tag{3.15}$$

If a model or sampling design permits estimation of the covarance matrix $\mathbf{FMF}'$, we may use such an estimator and the model in Section 2.3 to obtain the test statistic Q_T.

A proper model often provides a closed form of the covariance matrix, as is the case in this section. Upon application of the model (2.3), we can obtain the covariance matrix

$$\hat{\mathbf{F}}\hat{\mathbf{M}}\hat{\mathbf{F}}' = \frac{G}{y}\hat{\mathbf{F}}\hat{\mathbf{P}}\hat{\mathbf{F}}', \tag{3.16}$$

where $\hat{\mathbf{F}}\hat{\mathbf{P}}\hat{\mathbf{F}}'$ is a consistent estimator of the covariance matrix $\mathbf{FPF}'$ of $\sqrt{y}\mathbf{f}(\hat{\boldsymbol{\pi}})$ obtained as if the data arose from a multinomial sample. We assume that $\hat{\mathbf{F}}\hat{\mathbf{M}}\hat{\mathbf{F}}'$ is invertible; otherwise, the generalized inverse can be used.

For the two-way table, we can invert the covariance matrix as

$$(\hat{\mathbf{F}}\hat{\mathbf{M}}\hat{\mathbf{F}}')^{-1} = \frac{y}{G}(\hat{\mathbf{F}}\hat{\mathbf{P}}\hat{\mathbf{F}}')^{-1}, \quad \text{where} \quad (\hat{\mathbf{F}}\hat{\mathbf{P}}\hat{\mathbf{F}}')^{-1} = \hat{\mathbf{P}}_r^{-1} \otimes \hat{\mathbf{P}}_c^{-1}$$

and $\otimes$ denotes the usual direct matrix product. $\hat{\mathbf{P}}_r$ and $\hat{\mathbf{P}}_c$ are the estimators of $\mathbf{P}_r = \mathrm{diag}(\boldsymbol{\pi}_r) - \boldsymbol{\pi}_r\boldsymbol{\pi}_r'$ and $\mathbf{P}_c = \mathrm{diag}(\boldsymbol{\pi}_c) - \boldsymbol{\pi}_c\boldsymbol{\pi}_c'$, respectively, for $\boldsymbol{\pi} = \hat{\boldsymbol{\pi}}$, and $\boldsymbol{\pi}_r' = (\pi_{1+}, \ldots, \pi_{R-1,+})$ and $\boldsymbol{\pi}_c' = (\pi_{+1}, \ldots, \pi_{+,C-1})$.

We can rewrite (3.15) as

$$Q_T = \frac{y}{G}X_T^2, \qquad \text{where} \quad X_T^2 = \sum_{r}^{R}\sum_{c}^{C} \frac{y(\hat{\pi}_{rc} - \hat{\pi}_{r+}\hat{\pi}_{+c})^2}{\hat{\pi}_{r+}\hat{\pi}_{+c}}, \tag{3.17}$$

and X_T^2 is the usual chi-square statistic for tests of independence as if the sample were a simple random sample, and can be written as

$$X_T^2 = y\mathbf{f}(\hat{\boldsymbol{\pi}})'(\hat{\mathbf{F}}\hat{\mathbf{P}}\hat{\mathbf{F}}')^{-1}\mathbf{f}(\hat{\boldsymbol{\pi}}). \tag{3.18}$$

Following directly from standard results on quadratic forms (Rao and Scott, 1981, Theorem 2), X_T^2 is distributed as

$$\sum_{t=1}^{T} \delta_{t0}\omega_t^2, \tag{3.19}$$

where $\omega_1^2, \ldots, \omega_T^2$ are independent χ_1^2 random variables, and the δ_{t0} are the nonzero eigenvalues of $(\mathbf{FPF}')^{-1}(\mathbf{FMF}')$ under H_0.

Substituting $\mathbf{FPF}'(G/y)$ for $\mathbf{FMF}'$, we can express $(\mathbf{FPF}')^{-1}(\mathbf{FMF}')$ as

$$(\mathbf{FPF}')^{-1}(\mathbf{FPF}')(G/y) = \mathbf{I}(G/y),$$

where $\mathbf{I}$ is the identity matrix of dimension T and, therefore, the nonzero eigenvalues $\delta_{t0} = (G/y)$ for all $t = 1, \ldots, T$.

Thus, Q_T is distributed as $\chi^2_{(R-1)(C-1)}$ when the usual chi-square statistic X_T^2 is divided by G/y, that is, $Q_T = X_T^2(y/G)$, as shown in (3.17).

If a model different from that given by equation (2.3) is used, the eigenvalues may vary rather than have a common value G/y. Adjustment of Q_T may compensate for such variation (Satterthwaite, 1946).

In Section 3.3, the cell proportions were specified by H_0 in the goodness-of-fit test, but they were not specified in H_0 for the test of independence. Thus, they must be estimated. If $y/G = 1$ and maximum likelihood estimation is used, Bishop et al. (1975, Theorem 14.9.4) have shown that Q_T is asymptotically distributed as χ_T^2. The reduction factor y/G is the same as the one we have presented previously in equation (3.11). The reduction factor is further discussed in Sections 3.5 and 5.1 for more general situations.

If it is possible to find a consistent covariance estimator of $(\mathbf{FMF}')$, we can always find a reduction factor to correct the usual X_T^2, regardless of the method used for the estimation of $\boldsymbol{\pi}$.

It is often difficult to find a closed form of a covariance matrix from a complex sample and, therefore, it is often impossible to find an exact reduction factor without some model assumptions. Some researchers apply resampling methods (Efron, 1982) to obtain a covariance matrix for complex sample survey data. For example, Fellegi (1980) used the balanced half sample replication method. Landis et al. (1984), and Shah (unpublished document, Research Triangle Institute, Research Triangle Park, North Carolina, 1981), and Hidiroglou and Rao (paper presented at International Statistical Institute meetings, Buenos Aires, 1981) employed Taylor approximations to estimate such covariance matrices, when the weighting was not assumed to be fixed.

3.5 *Reduction Factor for Different Sizes of Clusters*

If the intracluster correlation remains constant only for clusters of the same size, these clusters may be grouped into the same-sized clusters, forming, say, K groups, with K intracluster correlations to be indexed by the subscript k, $k = 1, \ldots, K$.

Suppose that a clusters in the sample consist of a_1 clusters of b_1 members, a_2 clusters of b_2 members, ..., and a_K clusters of b_K members. The whole sample thus can be reclassified into K groups according to their sizes. The subscripts i and j now indicate the numbers of clusters and elements, $i = 1, \ldots, a_k$ and $j = 1, \ldots, b_{ki}$, within the same-sized clusters of the kth group. For the k-sized clusters, all clusters have k units and $b_{ki} = b_k$. Then the number of units in the k-sized clusters is $n_k = b_k a_k$, and total sample units $n = \sum_k b_k a_k$ for unweighted data. For the weighted counts, these are $y_k = \sum_{i=1}^{a_k} \sum_{j=1}^{b_{ki}} w_{ijk}$ and $y = \sum_{k=1}^{K} y_k$, where w_{ijk} is the weight of the jth member in the ith cluster that belongs to the kth group. Now the inference is conditional on the distribution of cluster sizes in the realized sample, but this is again reflected only in the model (3.20) below.

We can modify the model (2.3) by replacing θ by θ_k, which now depends on the cluster size k, that is,

$$p_{khh'} = \begin{cases} \theta_k \pi_h + (1 - \theta_k)\pi_h^2 & \text{for } h = h', \\ (1 - \theta_k)\pi_h \pi_{h'} & \text{for } h \neq h'. \end{cases} \tag{3.20}$$

Note that the parameters π_h are independent of the cluster size k, and θ_k is now the intracluster correlation common to the clusters of size k.

The covariance matrix of $\hat{\boldsymbol{\pi}}$ for the goodness-of-fit test and that of $\mathbf{f}(\hat{\boldsymbol{\pi}})$ for the test of independence are the same as $\boldsymbol{\Sigma}$ in Section 3.3, and $\mathbf{M}$ in Section 3.4, respectively, except for the factor G/y. Using the previous assumptions on the sampling, weighting, and known cluster sizes along with the new model (3.20), one can show that the new covariance matrices for the K groups of clusters are the same as $\boldsymbol{\Sigma}$ and $\mathbf{M}$, respectively, with the new constant factor G/y. Hence, the respective quadratic forms for the model (3.20) are also the same as Q and Q_T with the new reduction factor of

$$\frac{y}{G} = \frac{y}{\sum_{k=1}^{K} (\theta_k \sum_{i=1}^{a_k} \sum_{j \neq j'}^{b_k} w_{ijk} w_{ij'k} + \sum_{i=1}^{a_k} \sum_{j=1}^{b_k} w_{ijk}^2)}, \tag{3.21}$$

which now depends on cluster size. For clusters of size 1, the first term of G disappears for $\theta_k = 0$.

Under the conditions discussed in Sections 2.2 and 3.1, the average w of fixed weights w_{ijk} can be used in the reduction factor. Here, $\bar{w} = y/n$, where y is the total of the weighted counts and n is the total of the unweighted counts from the sample. Replacing w_{ijk} by $\bar{w}$, the reduction factor (3.21) reduces to

$$\frac{y'}{G'} = \frac{1}{\bar{w}[1 + (1/n) \sum_k n_k \theta_k (b_k - 1)]}. \tag{3.22}$$

Thus, the usual chi-square value can be adjusted when it is multiplied by one of the reduction factors shown in (3.21) or (3.22) for the clusters of different sizes. The latter expression is especially useful when individual weights are relatively homogeneous and $\bar{w}$ is known approximately.

For clusters with a common intracluster correlation, the reduction factor is given by (3.11) for varying weights in general. In addition, for the special case of a common weight, the reduction factor becomes

$$\frac{y''}{G''} = \frac{1}{w[1 + \theta(b - 1)]}. \tag{3.23}$$

This is also readily obtained from (3.21) for w_{ijk}, θ_k, and b_k replaced by w, θ, and b, respectively. Thus, in (3.23), G'' is the product of the weighting effect w and the design effects $[1 + \theta(b - 1)]$.

We can easily obtain reduction factors for unweighted data from (3.11), (3.21), (3.22), or (3.23), replacing $w_{ij} = 1$ for all i and j (and k). For instance, when the w_{ij} are replaced by 1, (3.22) reduces to

$$\frac{y_{\text{un}}}{G_{\text{un}}} = \frac{1}{1 + (1/n) \sum_k n_k \theta_k (b_k - 1)}, \tag{3.24}$$

and (3.23) to

$$\frac{y'_{\text{un}}}{G'_{\text{un}}} = \frac{1}{1 + \theta(b - 1)}, \tag{3.25}$$

which is the reduction factor shown by Altham (1976).

Rao and Scott (1981) called the factor G/y the generalized design effect. We note that $(G/y)^{-1}$ is the constant multiplier to correct the usual chi-square test statistic. The factor G/y was factored out from the covariance matrix, leaving the regular form of the multinomial covariance matrix.

As noted above, the reduction factor also depends on the estimation of the intracluster correlation θ. The estimation of θ is discussed in Sections 3.6 and 5.

3.6 *Estimation of the Intracluster Correlation* θ

The intracluster correlation has been discussed in the recent statistical literature by a number of statisticians, who often estimated the correlation independently from the models used for the development of test statistics. Such independent methods may often be the only alternative when there is no direct method for obtaining an accurate estimate from the original model.

Brier (1980) and Kleinman (1973) used the method of moments to estimate intracluster correlations. Donner and Koval (1980) and Stanish and Taylor (1983) estimated intracluster correlations of one-stage clustering for continuous data by linear model assumptions. Cohen (1976) estimated the intracluster correlation by maximum likelihood (ML) for clusters of two members. Landis and Koch (1977) used multivariate analysis of variance (MANOVA) to estimate such a correlation.

We extend the ML method here, and briefly present the MANOVA method for more general applications in Section 5.2.

Maximum likelihood estimation Suppose the clusters are grouped according to their size. Let K be the number of different sizes for clusters in a given sample. The intracluster correlation θ_k is then to be estimated for each group of the same size clusters ($k = 2, \ldots, K$). Here we estimate θ_k by ML, using an independent model (3.26), when clusters include more than two members.

The following probability model for the relationship among k members in the same cluster was introduced by Altham (1976),

$$p_{h_1,\ldots,h_k} = \begin{cases} \theta_k \pi_h + (1 - \theta_k)\pi_h^k & \text{for } h_1 = \cdots = h_k = h, \\ (1 - \theta_k)\pi_{h_1}\pi_{h_2} \cdots \pi_{h_k} & \text{otherwise.} \end{cases} \tag{3.26}$$

The upper part of (3.26) is the probability that all other members fall in the hth cell, given that the first member falls in that cell. The lower part of (3.26) is the probability that the first member falls in the cell h_1, the second member in cell h_2, ..., and kth member in cell h_k. The common intracluster correlation for the clusters of size b_k is θ_k, where $0 \leqslant \theta_k \leqslant 1$ and $\theta_k = 0$ for the clusters of one member. Note that the model (3.26) does not depend on the subscripts i, j, and h; only θ_k depends on the size of the cluster in the model.

θ_k may be estimated by ML assuming the following distribution for the k-sized clusters. We denote the k-sized clusters by the symbol $x_{h_1 h_2 \cdots h_k}$, with the distribution function of

$$f_k = b_k! \prod_{h_1 h_2 \cdots h_k}^{r} \frac{(p_{h_1 h_2 \cdots h_k})^{x_{h_1 h_2 \cdots h_k}}}{x_{h_1 h_2 \cdots h_k}!}. \tag{3.27}$$

The subscript k symbolizes the clusters including k members and r is the number of categories. We use the above distribution of dimension k not for the entire sample, but only for the clusters of k members, and estimate the intracluster correlation coefficient θ_k only for the k-member clusters. We may similarly define an independent distribution for the clusters of another size to estimate its intracluster correlation. The only difference among these functions is the dimension of x.

We use ML for the estimation of the intracluster correlation θ_k for the k-member clusters, assuming the model (3.36) and distribution (3.27). The ML estimator $\hat{\theta}_k$ of θ_k and $\text{var}(\hat{\theta}_k)$ for $k > 2$ are given in the Appendix.

The above method of estimating θ_k is applied to clusters of two members and of three members in the numerical example of Section 4.

4. A Numerical Example

During the 1975 National Health Interview Survey in the St. Paul–Minneapolis primary sampling unit (PSU), the 259 households in the sample were identified as including one, two, or three members. The sample consisted of 81 single-member households, 112 two-member households, and 66 three-member households. There were no age restrictions on the respondents.

All 503 members of the 259 households interviewed were classified into one of four categories: (1) less than 45 years of age with no chronic conditions, (2) less than 45 years of age with one or more chronic conditions, (3) 45 years of age or older with no chronic conditions, and (4) 45 years of age or older with one or more chronic conditions. In Table 1, the 503 persons and their weighted counts of 877,325 were classified by age and chronic condition status. Table 1 also shows the estimates of population proportions from the weighted and unweighted counts.

We want to test whether the proportions estimated by unweighted or weighted data fit the specified null hypothesis of independence between age and chronic condition. The basic parameters of interest are shown in Table 2.

Table 3 shows the estimates of intracluster correlations $\hat{\theta}_k$ by ML, and their standard deviations for each group of equal-sized clusters of one, two, and three members from the unweighted data (see the Appendix). The last two columns include the unweighted and weighted counts, respectively.

The individual weights in the data tape ranged from 1,500 to 1,750, except for a few extreme cases. The mean weight $\bar{w} = 1{,}744$ is obtained from the known weighted population estimate of 877,325 persons in the St. Paul–Minneapolis PSU, divided by the known sample count of 503 persons in this PSU. These two numbers—877,325 and 503—are

Table 1
The counts and proportions from unweighted and weighted data, National Health Interview Survey, 1975

Category	Y−	YC	O−	OC	Total
Population proportions	π_{11}	π_{12}	π_{21}	π_{22}	
Unweighted counts	208	85	112	98	503
Estimated prop.	.4135	.1690	.2227	.1948	1
Weighted counts	364,665	150,483	194,806	167,371	877,325
Estimated prop.	.4157	.1715	.2220	.1908	1

Y: Less than 45 years of age, O: 45 years or over;
−: No chronic condition, C: One or more chronic conditions.

Table 2
The test of independence between chronic conditions and age

		Chronic condition		
		−	C	
Age	Y	π_{11}	π_{12}	π_{1+}
	O	π_{21}	π_{22}	π_{2+}
		π_{+1}	π_{+2}	

Y: Less than 45 years of age, O: 45 years or over;
−: No chronic condition, C: One or more chronic conditions.
+: Marginal sum of respective subscripts.

Table 3
ML *estimates of intracluster correlations and their standard errors according to the three different sizes of the clusters, and the numbers of unweighted and weighted counts*

Cluster size b_k	Number of clusters a_k	Intracluster correlation $\hat{\theta}_k$	Standard deviation $\sqrt{\text{var}(\hat{\theta}_k)}$	Unweighted counts n_k	Weighted counts y_k
1	81	0	0	81	143,999
2	112	.5189	.0617	224	387,801
3	66	.1285	.0566	198	345,525

Table 4
Test of independence between chronic conditions and age, based on unweighted and weighted data given in Table 1

Types of data	Pearson chi-square statistic	Reduction factor	Adjusted chi-square statistic
Unweighted counts	16.476	.75061	12.37*
Weighted counts	26,604.258	.00043	11.38*

* Significant at $\alpha = .01$ (1 d.f.).

provided in the NCHS tapes, so that the average weight can easily be obtained from the tapes.

The independence test was performed on both types of data, one based on the unweighted counts and the other on the weighted data. The chi-square test scores before and after adjustment are presented in Table 4.

Using the intracluster correlations of one-, two-, and three-member clusters given in Table 3, the simplified formula (3.17) provided the reduction score of .00043 for weighted counts, much smaller than that of .75061 from the simplified formula (3.19) for the unweighted data.

The data in Table 1 are used for the test of independence by the formula (3.13). As seen in Table 4, the test results, if not adjusted, were usually too big to be reliable, especially for the weighted data. For instance, when the scores are not adjusted, the ordinary chi-square values are 26,604.258 for the weighted data and 16.476 for the unweighted data. They reduce to 11.38 and 12.37, respectively, when they are multiplied by the respective reduction factors.

As expected, the null hypothesis of independence between the chronic condition and age was rejected by the adjusted scores for both original and weighted data. If the null hypothesis is not rejected using the nonadjusted score, then it will not be rejected for the adjusted score. However, if the former is rejected, the original test score must be adjusted to see whether the reduced score should also be rejected.

5. Extensions

5.1 *Reduction Factor for Two-Stage Clustering*

In a two-stage design of the NHI survey, the first-stage cluster is a segment (a group of four households) and the second-stage cluster is the household. We may ignore the largest cluster (primary sampling unit, PSU) since the intracluster correlation between the members in the PSU is very weak in general.

Suppose that a population consists of two-stage clustering, and that the last-stage cluster includes a number of final units. This two-stage clustering gives a three-stage population. We extend the notations used for the two-stage population (one-stage clustering) to three-stage population (two-stage clustering).

Indexing the first-stage clusters, second-stage clusters, and the final units by i, j, and l, respectively, let the finite population U include A independent first-stage clusters, $U_1, \ldots, U_i, \ldots, U_A$, with the ith cluster decomposed into B_i second-stage clusters as $U_i = (U_{i1}, \ldots, U_{ij}, \ldots, U_{iB_i})$, and, finally, the (i, j)th second-stage cluster U_{ij} is decomposed into $U_{ij} = (U_{ij1}, \ldots, U_{ijl}, \ldots, U_{ijM_{ij}})$ final units. Each final unit belongs to one of the r categories indexed by $h = 1, \ldots, r$.

We define the probability for the relationship among elementary units in the first-stage cluster separately from the relationship of the elementary units in the second-stage clusters. Each stage clustering will be defined by different intracluster correlations. Denote the first-stage intracluster correlation by θ_1, and the second-stage intracluster correlation by θ_2.

We assume the following relationship holds for the members in the respective clusters:

$$\mathrm{E}(x_{ijlh}) = \pi_h \qquad \text{for all} \quad i, j, \text{ and } l;$$

$$\mathrm{E}(x_{ijlh}, x_{i'j'l'h'}) = \begin{cases} \pi_h\pi_{h'} & \text{if} \quad i \neq i' \\ u_{1hh'} & \text{if} \quad i = i', \ j \neq j' \\ u_{2hh'} & \text{if} \quad i = i', \ j = j', \ \text{and } l \neq l' \\ \pi_h & \text{if} \quad i = i', \ j = j', \ l = l', \text{and } h = h' \\ 0 & \text{if} \quad i = i', \ j = j', \ l = l', \text{and } h \neq h', \end{cases} \tag{5.1}$$

where $u_{1hh'}$ reflects the pairwise relationship of final units in the first-stage cluster (or segment) as defined by

$$u_{1hh'} = \begin{cases} \theta_1\pi_h + (1 - \theta_1)\pi_h^2 & \text{for} \quad h = h' \\ (1 - \theta_1)\pi_h\pi_{h'} & \text{for} \quad h \neq h', \end{cases} \tag{5.2}$$

and $u_{2hh'}$ reflects the pairwise relationship of the final units in the second-stage cluster (or household) as defined by

$$u_{2hh'} = \begin{cases} \theta_2\pi_h + (1 - \theta_2)\pi_h^2 & \text{for} \quad h = h' \\ (1 - \theta_2)\pi_h\pi_{h'} & \text{for} \quad h \neq h'. \end{cases} \tag{5.3}$$

Any one pair in the last two-stage cluster can be defined by either (5.2) or (5.3), but not by both, in order to avoid duplications. For instance, a pair defined by (5.3) cannot be included in the pairs defined by (5.2). Thus, (5.2) defines the pairwise relationship for the first-stage clusters, while (5.3) defines the pairwise relationship for the second-stage clusters. Note that neither model depends on subscripts i, j, and l.

The models require that the intracluster correlations are restricted as $0 \leqslant \theta_1 \leqslant 1$ and $0 \leqslant \theta_2 \leqslant 1$. Sometimes, the estimates $\hat{\theta}_1$ and $\hat{\theta}_2$ may be negative and the systems (5.2) or (5.3) break down. The estimates may be replaced by zero (see the Appendix).

Consider a sample S from the population U, with slight variation in the weights. We assume that these weights are known positive integers, and that the realization of cell frequencies is not dependent on the sampling and weighting, but is regulated by the model (5.1)–(5.3). Denote the sample S by

$$S = \{(i, j, l): \ i \in S^*, \ j \in S_i, \ l \in S_{ij}\},$$

where S^* is now a sample of a independent clusters, S_i is the sample of b_i second-stage clusters, and S_{ij} is a sample of m_{ij} units in the (i, j)th second-stage cluster. The units in S_i are regulated by the rules defined in (5.2), and the units in S_{ij} follow the definition (5.3).

These three stages are indexed by $i = 1, \ldots, a$ for the first-stage clusters, $j = 1, \ldots, b_i$ for the second-stage clusters, and $l = 1, \ldots, m_{ij}$ for the units in the second-stage cluster. We also assume that the b_i and the m_{ij} are fixed numbers.

The weight of the (i, j, l)th element is now denoted by w_{ijl} for the two-stage clustering. Denote the total of the weights in the sample S by $y = \sum_{i=1}^{a} \sum_{j=1}^{b_i} \sum_{l=1}^{m_{il}} w_{ijl}$, the cell counts by $y_h = \sum_{i=1}^{a} \sum_{j=1}^{b_i} \sum_{l=1}^{m_{il}} x_{ijlh} w_{ijl}$, and the cell proportions by $\hat{\pi}_h = y_h/y$, where the random variables $x_{ijlh} = 1$ if the (i, j, l)th unit belongs to cell h, and $x_{ijlh} = 0$, otherwise. All the assumptions on the $\boldsymbol{\pi}$, $\mathbf{f}(\boldsymbol{\pi})$, and weights seen in the one-stage clustering also hold here.

The covariance matrix of $\hat{\boldsymbol{\pi}}$ for a goodness-of-fit test and the covariance matrix of $\mathbf{f}(\hat{\boldsymbol{\pi}})$ for an independence test in two-stage clustering can be derived by the same procedures previously used for one-stage clustering.

It can be shown that the quadratic form for the goodness-of-fit test takes the form of (3.5)–(3.7) and that of the independence test has the form of (3.16) with the new constant factor G/y (Choi, 1981). Consequently, the goodness-of-fit test takes the form of (3.10) and the independence test takes the form of (3.17) with the new constant reduction factor of

$$\frac{y}{G} = \frac{y}{\sum_i^a \sum_j^{b_i} \sum_l^{m_{ij}} w_{ijl}^2 + \theta_1 \sum_i^a \sum_j^{b_i} \sum_{l \neq l'}^{m_{ij}} w_{ijl} w_{ijl'} + \theta_2 \sum_i^a \sum_{j \neq j'}^{b_i} \sum_l^{m_{ij}} \sum_{l'}^{m_{ij'}} w_{ijl} w_{ij'l'}}, \tag{5.4}$$

where G now includes three terms. Note $0 < y/G \leq 1$. All the previous statements for asymptotic distributions for the one-stage clustering hold for the two-stage clustering case. If the sample is self-weighting or if the weights do not vary much, the average of weights may be used to simplify the reduction factor (5.4) as

$$\frac{y}{G} = \frac{1}{\bar{w}\{1 + \theta_1[(1/n) \sum_i^a \sum_j^{b_i} m_{ij}^2 - 1] + \theta_2 (1/n)[\sum_i^a m_{i+}^2 - \sum_i^a \sum_j^{b_i} m_{ij}^2]\}}. \tag{5.5}$$

If there is no correlation in the cluster or $\theta_1 = \theta_2 = 0$, the last two terms of G will disappear and G is a minimum, and hence y/G is a maximum. On the other hand, if $\theta_1 = \theta_2 = 1$, G is a maximum; therefore, y/G is a minimum.

If $w_{ijl} = w$, $m_{ij} = m$, and $b_i = b$, then (5.4) reduces to

$$\frac{y'}{G'} = \frac{1}{w[1 + \theta_2(m - 1) + \theta_1 m(b - 1)]}, \tag{5.6}$$

where G' also depends on the weight, and the first- and the second-stage intracluster correlations. If $b = 1$, the last two terms of G' are zeros since $b = 1$ implies $m = 1$. The reduction factor for unweighted data is obtained simply by setting $w = 1$ from (5.5) or (5.6).

We can rewrite y/G for K groups of the same-sized clusters by redefining the models (5.2) and (5.3) with new parameters θ_{1k} and θ_{2k}, which then depend on the cluster size as already discussed in Section 3.5 for one-stage clustering. The core problem in the reduction factor is that of estimating the intracluster correlation.

When the data arise from two-stage clustering, it may be possible to extend ML estimation to the estimation of the intracluster correlations θ_1 defined in the model (5.2), and θ_2 defined in the model (5.3).

It has been shown that MANOVA can be extended to two-stage estimation of θ_1 for first-stage clustering and θ_2 for second-stage clustering (Choi and Casady, 1982). Further, the numerical example in Section 5.2 based on one-stage clustering suggests that MANOVA may be preferable to ML.

5.2 *Multivariate Analysis of Variance* (MANOVA) *Estimation*

We briefly review the MANOVA estimation of intracluster correlations for one-stage clustering.

Landis and Koch (1977) developed a MANOVA method to estimate intracluster correlation for one-stage clustering. They estimated θ with the usual ANOVA terms as

$$\hat{\theta} = \frac{\sum_{h=1}^{r} \{SS_{chh}/[d(a-1)] - SS_{ehh}/[d(n-a)]\}}{\sum_{h=1}^{r} \{SS_{chh}/[d(a-1)] + (d-1)SS_{ehh}/[d(n-a)]\}},$$

where $d = (n^2 - \sum b_i^2)/[n(a-1)]$, a is the number of clusters, b_i the number of elements in the ith cluster, and n the sum of b_i. SS_{chh} and SS_{ehh} are the hth category MANOVA sums of squares due to clusters and residual errors. Here the estimator $\hat{\theta}$ is biased; however, the bias becomes small in large samples.

Numerical example The data of 30 patients who were examined for psychiatric diagnoses, involving five determinations per subject by six physicians (Landis and Koch, 1977), were used to estimate θ. Here the patients are considered as clusters, and the five diagnoses as the members in the cluster in a one-stage clustering.

Choi (1981) compared ML estimation with MANOVA estimation, using these data. The results are presented in Table 5. It appears that the ML estimate is close to the MANOVA estimate for clusters of size 2, but that the ML estimate becomes small faster than the MANOVA estimate when the cluster size increases.

Table 5
Comparison between ML *and* MANOVA *estimation*

Method of estimation	Number of diagnoses per person (cluster size)				
	2	3	4	5	6
ML	.644	.433	.329	.285	.166
MANOVA	.653	.545	.500	.496	.440

This empirical evidence suggests that the ML method may not provide appropriate estimates, when the cluster size increases sharply, and many cells remain empty or small, perhaps because in this case the data do not fit the distribution. It appears that MANOVA estimation works better, especially for larger clusters. However, further simulation comparisons are required to obtain a better picture of this trend.

6. Comments

The proposed reduction factor provides the minimum and maximum value of the chi-square test statistic if $\hat{\theta}$ is set equal to 1 and 0, respectively, in the reduction factor. In this way, a preliminary test can be done even without information on $\hat{\theta}$.

If a maximum value of $\hat{\theta}$ was obtained from past experience [for example, some NCHS data suggest $\max(\hat{\theta}) = \frac{1}{2}$ or $\frac{1}{4}$], we may establish a better conservative estimate of the chi-square test statistic by the application of such $\max(\hat{\theta})$ in the reduction factor without actual information about the covariance matrix.

When the corresponding log-linear model (Bishop et al., 1975) has closed-form estimates of parameters and the model fits the data, the log-likelihood-ratio test G^2 provides a test asymptotically equivalent to the Pearson test statistic X^2 (Bedrick, 1983; Fay, 1985) when both G^2 and X^2 are corrected for design effects. However, it is not clear how these chi-square statistics would behave when a log-linear model fits the data and the model used in

this paper does not. A comparison of the results of the model approach used here with other tests from design-based methods is also of interest. Additional study is needed to answer these questions.

We assumed that the weights are fixed in the test when the post-stratified ratios deviate little from 1. If this is not the case, other kinds of weights might be considered. For example, sample-based weights could be considered as random variables arising from clustering and post-stratified ratio estimations. The initial research presented in this paper may be extended to such complex situations by treating design effects separately from weighting effects. For instance, one may use the same approach for the estimation of design effects, but with repeated Taylor approximations of weights, so that the product of the covariance matrix of weighted cells and the sampling effects may give the covariance matrix of weighted and correlated observations.

Extension might also be made to the case of cluster sizes considered as random variables under a given sample survey design.

No stratification is discussed in this paper. But if the sample is selected by stratified random sampling with proportional (self-weighting) allocation, such stratification always results in a smaller variance than is given by a comparable simple random sample. Therefore, the usual chi-square statistic based on a stratified random sample is always smaller than the chi-square statistic based on a simple random sample if the stratification is done skillfully. If the model can be used to define the pairwise relationship for the members in the independent strata, it may be possible to derive a closed form of the covariance matrix for the entire sample and consequently to obtain a correction factor that will increase the size of the usual chi-square test statistic.

It may also be possible to estimate the covariance matrix for the data from other types of sample designs and obtain the correction factor for the respective chi-square statistics.

There are other examples of clusters for which this reduction factor may be applicable. Animals in the same litter or plants in the same plot are often correlated. The patients examined by the same physician, responses taken by the same interviewer, or manufactured goods produced by the same machine, may also be correlated. Therefore, the method presented in this paper may be applicable to analyses arising in animal or plant studies, and to the study of the biases of examiners, interviewers, or machines.

Acknowledgements

The authors thank Dr Robert J. Casady, Department of Labor, Dr Gad Nathan, Hebrew University, the editor, and two anonymous readers whose comments helped to improve this paper.

Résumé

Au cours des enquêtes à domicile à grande échelle, on s'intéresse fréquemment non pas à des individus, mais à des groupes issus de la population par des méthodes d'échantillonage complexes. Typiquement, les groupes provenant de tels échantillons complexes contiennent des individus corrélés. Les réponses de ces individus sont alors pondérées, afin d'obtenir des estimations pour la population entière. Ces données pondérées sont d'habitude publiées par le National Center for Health Statistics ou d'autres agences fédérales Américaines.

Les problèmes sont fréquents quand on teste l'adéquation de ces données ou leur indépendance par les statistiques habituelles du Chi-2: celles ci surestiment gravement les résultats dans le cas de ces échantillons de données groupées. De nouvelles méthodes sont nécessaires pour résoudre de tels problèmes.

Cet article propose une stratégie pour les tests d'adéquation et d'indépendance, dans le cas de données corrélées et pondérées provenant de tels groupes. Il fournit un facteur qui réduit notablement la surestimation de la statistique usuelle du Chi-2.

Cette méthode est appliquée sur des données recueillies à St. Paul–Minneapolis au cours de l'Enquête Nationale de Santé (NHIS) réalisée en 1975 (enquête par entretiens). Cette analyse, associée à des études par simulation présentées ailleurs, démontrent que les statistiques habituelles du Chi-2 sur de telles données peuvent être corrigées des effets du regroupement et de la pondération grace au facteur de réduction proposé.

References

Altham, P. M. E. (1976). Discrete variable analysis for individuals grouped into families. *Biometrika* **63,** 263–269.

Bean, J. (1973). *Estimate and Sample Variance.* Series 2, No. 38. Hyattsville, Maryland: The National Center for Health Statistics.

Bedrick, E. J. (1983). Adjusted chi-squared tests for cross-classified tables of survey data. *Biometrika* **70,** 591–595.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice.* Cambridge, Massachusetts: The Massachusetts Institute of Technology Press.

Brier, S. S. (1980). Analysis of contingency tables under cluster sampling. *Biometrika* **67,** 591–596.

Bryant, E. E., Baird, J. B., and Miller, H. W. (1971). *Sample Design and Estimation Procedures.* Series 2, No. 43, Hyattsville, Maryland: The National Center for Health Statistics.

Choi, J. W. (1981). A further study on the analysis of categorical data from weighted cluster sample surveys. In 1981 *ASA Proceedings of the Section on Social Science Statistics*, 217–222.

Choi, J. W. and Casady, R. J. (1982). χ^2-testing of categorical data from nested design using the correction factor estimated from analysis of variance components. In 1982 *ASA Proceedings of the Section on Social Science Statistics*, 531–536.

Cohen, J. E. (1976). The distribution of chi-squared statistic under clustered sampling from contingency tables. *Journal of the American Statistical Association* **71,** 665–670.

Cramér, H. (1946). *Mathematical Methods of Statistics.* Princeton, New Jersey: Princeton University Press.

Donner, A. and Koval, J. J. (1980). The estimation of intraclass correlation in the analysis of family data. *Biometrics* **36,** 19–25.

Efron, B. (1982). *The Jackknife, the Bootstrap, and Other Resampling Plans.* Philadelphia, Pennsylvania: Society for Industrial and Applied Mathematics.

Fay, R. E. (1985). Complex samples. *Journal of the American Statistical Association* **80,** 148–157.

Fellegi, I. P. (1980). Approximate tests of independence and goodness of fit based upon stratified multistage samples. *Journal of the American Statistical Association* **75,** 261–268.

Fienberg, S. E. (1979). The use of chi-square statistics for categorical data problems. *Journal of the Royal Statistical Society, Series B* **41,** 54–64.

Holt, D., Scott, A. J., and Ewings, P. O. (1980). Chi-squared tests with survey data. *Journal of the Royal Statistical Society, Series A* **143,** 302–320.

Kendall, M. G. and Stuart, A. (1968). *The Advanced Theory of Statistics, Vol.* 3. New York: Hafner.

Kleinman, J. C. (1973). Proportions with extraneous variance: Single and independent samples. *Journal of the American Statistical Association* **68,** 46–54.

Koch, G. G., Freeman, D. H., and Freeman, J. L. (1975). Strategies in the multivariate analysis of data from complex surveys. *International Statistical Review* **43,** 59–78.

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

Landis, J. R., Lepkowski, J. M., Eklund, S. A., and Stehouwer, S. A. (1984). *A Statistical Methodology for Analyzing Data from a Complex Survey: The First National Health and Nutrition Examination Survey.* Series 2, No. 92. Hyattsville, Maryland: The National Center for Health Statistics.

Rao, J. N. K. and Scott, A. J. (1981). The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association* **76,** 221–230.

Rao, J. N. K. and Scott, A. J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *The Annals of Statistics* **12,** 46–60.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics* **2,** 110–114.

Stanish, W. M. and Taylor, N. (1983). Estimation of the intraclass correlation coefficient for the analysis of covariance model. *The American Statistician* **37,** 221–224.

Wald, A. (1943). Test of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society* **54,** 426–482.

Received August 1984; *revised January* 1986, *June* 1987, *and October* 1988.

APPENDIX

Maximum Likelihood Estimation of Intracluster Correlation

Using (3.26) and the $\pi_r = 1 - \sum_{h=1}^{r-1} \pi_h$ in the function (3.27), we obtain the two maximum likelihood equations for the unweighted data of k-size clusters, one for θ_k and the other for $\boldsymbol{\pi}$. The first equation, $\partial \ln f_k/\partial\theta_k = 0$, reduces to

$$\sum_{h}^{r} \frac{x_{hh\ldots h}(1 - \pi_h^{b_k-1})}{[\theta_k/(1 - \theta_k) - \pi_h^{b_k-1}]} = a_k - \sum_{h}^{r} x_{hh\ldots h}, \tag{A.1}$$

and the second one, $\partial \ln f_k/\partial\pi_h = 0$, $h = 1, \ldots, r-1$, reduces to

$$\pi_h = \frac{n_{kh} - (b_k - 1)x_{hh\ldots h}/[1 + (1/\theta_k - 1)\pi_h^{b_k-1}]}{n_k - \sum_h [(b_k - 1)x_{hh\ldots h}]/[1 + (1/\theta_k - 1)\pi_h^{b_k-1}]}. \tag{A.2}$$

The random variable $x_{hh\ldots h}$ is the sum of all clusters of size k whose members fall in the hth cell, and the parameter π_h is the hth cell proportion in the population. The equations (A.1) and (A.2) are consistent with the results known to be true for $\theta_k = 0$ and $\theta_k = 1$.

These solutions depend only on the diagonal elements of $[x_{h_1h_2\cdots h_k}]$ and $[n_{kh}]$ when $h_1 = h_2 = \cdots = h_k = h$. Sometimes $\hat{\theta}_k$ is negative; then it may be equated to zero for practical applications as suggested by Fienberg (1979). Details of the derivation are presented in Choi's unpublished Ph.D. thesis (University of Minnesota, 1980). If the results shown in (A.1) and (A.2) exist between the interval of 0 and 1, (A.1) and (A.2) provide the solutions for the estimation of θ_k and π_h. Replacing the parameters with known estimates, one may also estimate the variance of $\hat{\theta}_k$ by

$$\begin{aligned}\operatorname{var}(\hat{\theta}_k) &= \frac{1}{-\mathrm{E}[\partial^2 \ln f_k/\partial\theta_k^2]} \\ &= \frac{1}{a_k\{\sum_h^r \pi_h(1 - \pi_h^{b_k-1})^2/[\theta_k + (1 - \theta_k)\pi_h^{b_k-1}] + (1 - \sum_h^r \pi_h^{b_k})/(1 - \theta_k)\}}.\end{aligned}$$

As seen in (A.1) and (A.2), these estimators are not given in explicit form. The final solutions can be obtained by the Newton–Raphson iterative method as described below. To start the first iteration, we may use the sample estimate of θ_k from equation (3.26).

Consider only $h_1 = h_2 = \cdots = h_k = h$ in (3.26) and replace $p_{h_1h_2\cdots h_k}$ by $\hat{p}_{h_1h_2\cdots h_k} = x_{h_1\cdots h_k}/a_k$ and π_h by $\hat{\pi}_h^0 = n_{kh}/n_k$. Summing over h and solving for θ_k, we obtain an initial value $\hat{\theta}_k^0$ (say) for θ_k as

$$\hat{\theta}_k^0 = \frac{\sum_{h=1}^r \hat{p}_{h_1h_2\cdots h_k} - \sum_{h=1}^r \hat{\pi}_h^0}{1 - \sum_h^r \hat{\pi}_h^2}.$$

The solutions of $(\hat{\theta}_k, \hat{\pi}_1, \ldots, \hat{\pi}_{(r-1)})$ from equations (A.1) and (A.2) exist. We start the iteration with the initial values $\hat{\theta}_k^0$ and $\hat{\pi}_h^0$. Let t denote the generic iteration. For $t = 0$, we:

1. Find $\hat{\theta}_k^{t+1}$, solving (A.1), using $\hat{\pi}_h^0$ for π_h.
2. Find $\hat{\pi}_h^{t+1}$ from (A.2), using $\hat{\theta}_k^{t+1}$ for θ_k.
3. Increment t by 1 and repeat steps 1 and 2, stopping when the solution converges to a specified level.

We can repeat the above procedures for each of K groups for the estimation of θ_k for $k = 2, \ldots, K$. When $k = 1$, $\theta_k = 0$. It appears that the ML method is an attractive alternative for small-sized clusters.