ANALYSIS OF CATEGORICAL DATA IN WEIGHTED CLUSTER SAMPLE SURVEYS

Jai W. Choi, National Center for Health Statistics and Richard B. McHugh, University of Minnesota

1. Introduction

A weighted cluster sample survey design is frequently used in large demographic sample surveys. In the National Health Interview Survey conducted by the National Center for Health Statistics, households are often selected in clusters of four. In this survey, sociodemographic and health characteristics of all members of sample households are recorded. Such characteristics for each person interviewed are multiplied by a known weight that is approximately the inverse of the probability of being included in the sample on the basis of the post-stratified geographic and demographic domain of each individual. This type of weighting is necessary to estimate certain characteristics of the target population at reasonable cost in large sample survey situations.

Cohen (1976) discussed the distribution of the chi-squared statistic from contingency tables in cluster sampling when clusters consist of two members. Altham (1976) generalized Cohen's results for clusters of M members. In the present research, these results are further extended to the weighted cluster sample survey. A new chi-squared statistic is used to analyze data from cluster sampling and weighted cluster sampling, and these two results are compared. This statistic is useful in the analysis of complex survey data for investigating the effect of weighting in cluster sample survey situations. Illustrative data from the 1975 National Health Interview Survey are analyzed by these new methods.

2. Models of Association Between Cluster Members

Suppose that a sample of b clusters is randomly drawn from the population of B clusters and that the ν the cluster contains M_ν members (ν=1,...,b). We observe that b clusters include n members ($n=\sum^b_\nu M_\nu$) in the sample and B clusters include N members ($N=\sum^B_\nu M_\nu$) in the population. Suppose that each member is doubly classified in a two-way contingency table: once by a population characteristic represented by R rows and once by a different population characteristic represented by C columns. Let n_i denote the sample count of members that fall in category i and N_i the count of category i in the population, i=1, ...,q where q=RxC so that $\sum^q_i n_i= n$ and $\sum^q_i N_i= N$. Further, let π_i represent the probability that the λth member of the νth cluster falls in category i with $\Sigma\pi_i=1$ and $\pi_i> 0$ where $\lambda=1,\ldots,M_\nu$, $\nu=1,\ldots,b$, and i=1,...,q.

Denote by $P_{ij\ldots t}$ the probability that the first member of the cluster falls in category i, the second member in category j...and the last member in category t (i,j,...,t=1,...,q). Specific forms for this parameter have been recently proposed to define the probability of positive association between the members of the cluster in terms of the π_i and a ($0\le a\le 1$) as shown below in (2.1) and (2.2).

Cohen (1976) considered a random sample of b clusters, each containing two siblings. Each of the 2b sibilings is classified in one of q categories. Let P_{ij} be the probability of a family in which the first sibling falls into category i and the second sibling into category j. Cohen suggested the model relationship:

$$P_{ij} = \begin{cases} a\pi_i + (1-a)\pi_i^2 & (i=j) \\ (1-a)\pi_i\pi_j & (i\neq j) \end{cases} \tag{2.1}$$

where $P_{ij}>0, \Sigma P_{ij} = 1$, and $0\leq a\leq 1$.

The probability table of P_{ij} is symmetrical and marginally homogeneous. If a=0, $P_{ij}=\pi_i\pi_j$, the two members are totally independent and if a=1, $P_{ij}=\pi_i$(i=j) and 0 ($i\neq i$), the two members are totally dependent. Altham (1976) extended Cohen's model to the case of families of size three, employing the relationship:

$$P_{ijk}= \begin{cases} a\pi_i+(1-a)\pi_i^3 & (i=j=k) \\ (1-a)\pi_i\pi_j\pi_k & \text{otherwise} \end{cases} \tag{2.2}$$

where $P_{ijk}>0$, $\Sigma P_{ijk} = 1$, and $0\leq a \leq 1$. The extension to families of any size is straightforward.

Other models appeared in recent literature are another model by Altham (1976), Dirichlet model (Plackett and Paul 1978), and Dirichlet-Multinomial model (Fienberg 1979). The variance form of the cell counts depends totally or partially on the choice of a model and this form arising from using such a model in cluster sampling is generally different from the variance form of multinomial distribution.

3. Chi-Square Tests of a Simple Hypothesis

Pearson's statistic (1900) is used conventionally to test the simple hypothesis $H_0:\pi_i^o=\pi_i$, i=1,...,q. The statistic for testing goodness of fit is

$$Q = \sum^q_i \frac{(n_i-n\pi_i)^2}{n\pi_i} \tag{3.1}$$

However, in a complex sample survey situation when the elementary counts are dependent variables and weighted, the statistic Q is not appropriate. In this section, we discuss the problems arising from such a nonstandard situation, viz. that in which the data come from a weighted cluster sample survey.

When the data consist of a random sample of b clusters, each including two members, Cohen (1976) showed that a valid goodness of fit test statistic is

$$Q_C = \frac{Q}{1 + a}, \quad 0 \leq a \leq 1 \qquad (3.2)$$

where Q_C has a limiting χ^2_{q-1} distribution as $b \to \infty$. Altham (1976) extended Cohen's result to clusters of size M, that is,

$$Q_A = \frac{Q}{1+a(M-1)}, \quad 0 \leq a \leq 1. \qquad (3.3)$$

Q_A also approaches the distribution of χ^2_{q-1} as $b \to \infty$. When the members in the cluster are totally independent (a=0), $Q_A = Q$ and the conventional test statistic Q in (3.1) can be used as if there were no clusterings. When they are totally dependent (a=1), then $Q_A = Q/M$ because the observed sample size of bM is M times the actual number of independent observations b. Here the effective sample size is b.

Before introducing a new statistic for the weighted cluster sample survey, certain basic aspects of this survey design need emphasis.

1. Denote by $w_{\nu\lambda}$ the statistical weight for the λth member of the νth cluster ($\lambda=1,\ldots,M_\nu=1,\ldots,b$). We assume that the weights of individual members are known from previous data collection procedures regardless of the category into which each may fall later. In fact, $w_{\nu\lambda}$ is the reciprocal of the probability that the (ν,λ)th member is selected from a post-stratified geographic and demographic domain to which it belongs. Such weighting is aimed at bringing the contribution of each member into closer alignment with the known population figure of its domain. Bryant, Baird, and Miller (1973) illustrated such a weighting scheme for the National Health Examination Survey, and Bean (1974) used a similar weighting system in the National Health Interview Survey. Observe that $1 \leq w_{\nu\lambda} \leq N$ and unweighted data are obtained as a special case by setting $w_{\nu\lambda} = 1$. Further $w_{\nu\lambda} = 1$ for b=B. If only one person is sampled from the population, $w_{\nu\lambda} = N$. In practice, most sample surveys are so designed that the weights assigned to individual members are not much different.
2. We assume that a random sample of b clusters is to be drawn with replacement from the population and all the members in the cluster are included in the sample without second stage sampling within the cluster.
3. Using the known weights defined previously, we can obtain the overall weighted counts $\hat{N}$ where $\hat{N}=\sum_\nu^b \sum_\lambda^{M_\nu} w_{\nu\lambda}$, M_ν and $w_{\nu\lambda}$ are known positive integers. Following the theorem (Chung 1968 p196) the limiting distribution of b independent random variables, each containing M dependent variables, the standardized sum of the b variables coverges to normality as b becomes large. When each of the members is multiplied by a known weight, the central limit theorem still holds so long as the sum indexed by b has the finite mean and variance.
4. Define the indicator function

$$\delta_{i\nu\lambda} = \begin{cases} 1 & \text{if the } (\nu\ \lambda)\text{th member falls in the ith category} \\ 0 & \text{otherwise} \end{cases}$$

The sample count of members falling in category i is n_i where $n_i = \sum_\nu^b \sum_\lambda^{M_\nu} \delta_{i\nu\lambda}$. Denote by $\hat{N}_i$ the weighted count of members that fall in the ith category so that $\hat{N}_i = \sum_\nu^b \sum_\lambda^{M_\nu} w_{\nu\lambda}\delta_{i\nu\lambda}$ and $\hat{N}=\sum_i^q \hat{N}_i$. In both unweighted and weighted cases, the cell count is the sum of independent clusters, each including M_ν dependent ($\nu=1,\ldots,b$) variables that may or may not fall in the ith category.

The problem of interest is to investigate the goodness of fit of π_i to the weighted cell counts $\hat{N}_i$ arising from the weighted cluster sampling. The conventional chi-squared test statistic for the investigation is

$$Q_W = \sum_i^q \frac{(\hat{N}_i - \hat{N}\pi_i)^2}{\hat{N}\pi_i} \qquad (3.5)$$

Because of the possible dependence between the members in the same cluster with weighting of individual member, the joint distribution of $\hat{N}_i$ is not a multinomial distribution, therefore, the conventional statistic Q_W in (3.5) will not provide an appropriate test statistic. Under model (2.1), $\hat{Y} = (\hat{N}_1,\ldots,\hat{N}_{q-1})$ has the finite mean $\hat{N}\pi$ and covariance matrix $G(D_\pi - \pi'\pi)$ where $\pi=(\pi_1,\ldots,\pi_{q-1})$, D_π is diagonal matrix based on π and G is a known positive number defined in (3.6) below. If the simple null hypothesis is true, the correct test statistic under model (2.1) is

$$Q_T = (\hat{N}/G)\, Q_W \qquad (3.6)$$

where

$$G = a \sum_\nu^b \sum_{\lambda\neq\lambda'}^{M_\nu} w_{\nu\lambda} w_{\nu\lambda'} + \sum_\nu^b \sum_\lambda^{M_\nu} w^2_{\nu\lambda} \text{ and } 0 \leq a \leq 1.$$

G measures the combination of clustering and weighting effects in the weighted cluster sampling situation. The largest value of G is obtained at a=1 for fixed $w_{\nu\lambda}$. $G \to \infty$ if the weights become large and $a \geq 0$. Observe that $\frac{\hat{N}}{G}$ is the proportion of the effective size of the weighted sample counts $\hat{N}$ and $0 \leq \frac{\hat{N}}{G} \leq 1$, where $\frac{\hat{N}}{G} = 1$ if $w_{\nu\lambda}=1$ and a=0, and $\frac{\hat{N}}{G} \to 0$ as the weights become large for $a \geq 0$.
Thus if the model (2.1) is used and if the weights and parameter a are known, the conventional test statistic Q_W in (3.5) can be corrected by multiplying the scale factor $\hat{N}/G$ and this final result can be used to test the goodness of fit of the weighted cluster sample data for π, that is, if

$Q_W \to \frac{G}{\hat{N}} \chi^2_{q-1}$ where $G/\hat{N}$ is known constant that factored out, then $\frac{\hat{N}}{G} Q_W \to \chi^2_{q-1}$ as $b \to \infty$.

The statistic Q_T in (3.6) is true only if $0 \leq a \leq 1$. If $a < 0$, it indicates negative dependence between members in the cluster; this situation is seldom encountered in practice.

4. Estimation of Parameters

When likelihood functions do not have closed form solutions in the maximum likelihood estimations, Bishop, Fienberg, and Holland (1976 p84) suggested the use of Newton-Raphson method Dodes (1978 p76) for a solution by iteration. Cohen (1976) considered a particular case arising from a cluster sampling situation for which each cluster included two members and obtained the open form solutions for the parameters in model (2.1). Starting from the initial sample estimators, he derived the final estimators applying Newton-Raphson method to the open systems.

Cohen's method is easily extended to the clusters of three members. The general case follows readily. Using the unweighted data, we find the initial sample estimators π_i^o of π_i, that is,

$$\pi_i^o = \frac{n_i}{n}, \quad i=1,\ldots,q \tag{4.1}$$

that are unbiased and consistent under the model (2.2). From model (2.2), by considering i=j=k (only the diagonal elements) and replacing $P_{iii}=X_{iii}/b$ and $\pi_i=\pi_i^o$, we can find the initial sample estimator a^o of a by summing over the subscript i and solving for a, that is,

$$a^o = \frac{\sum_i^q \frac{X_{iii}}{b} - \sum_i^q \pi_i^{o\,3}}{1- \sum_i^q \pi_i^{o\,3}} \tag{4.2}$$

where X_{iii} is the observed number of clusters that all three members of a three member cluster fall in category i, i=1,...,q. The estimator a^o depends only on the sufficient configurations n_i and X_{iii}, which also remain true of the final estimators as will be seen below in (4.3) and (4.4).

Suppose that the joint distribution function of X_{ijk} is multinomial with parameters b and p_{ijk} which is defined in (2.2). We can obtain the open form solutions for the parameters a and π_i, that is, for a,

$$\sum_i^q \frac{X_{iii}(1-\pi_i^2)}{\frac{a}{1-a} + \pi_i^2} = b - \sum_i^q X_{iii} \tag{4.3}$$

and for π_i

$$\pi_i = \frac{n_i - \frac{2X_{iii}}{1+(\frac{1}{a}-1)\pi_i^2}}{3b - \sum_j^q \frac{2X_{jjj}}{1+(\frac{1}{a}-1)\pi_j^2}} \tag{4.4}$$

It can be easily checked that the solutions of (4.3) and (4.4) are consistent with the known results in the limiting values of a=1 amd a=0. If maximum likelihood estimator $\hat{a}$ exists in the interval between 1 and 0 and if π_i are known, the variance of the estimator $\hat{a}$ can be estimated by substituting a by $\hat{a}$ in Fisher's information. If a solution $(\hat{a},\hat{\pi}_1,\ldots,\hat{\pi}_q)$ in (4.3) and 4.4) exists, starting with the initial estimates a^o and π_i^o, we can obtain the estimators $\hat{a}$ and $\hat{\pi}_i$ by the Newton-Raphson method. If the estimates exist, this method always converges to the required set of maximum likelihood estimates; a stopping rule may be used that ensures accuracy to any desired degree; any set of starting values may be used that conforms to the model being fitted; and if direct estimates exist, the procedure yields the exact estimates in one cycle.

5. Chi-Square Test of a Complex Hypothesis

The problems of asymptotic distribution of the chi-squared statistic in testing independence arise when the parameters are unknown and the data come from weighted cluster sampling.

Consider a two-way contingency table of n_{gh} of R rows C colums based on a random sample of size n. We use usual plus "+" summation convention so that the row and column margins are expressed as n_{g+} and n_{+h}. Let π_{gh} be the unknown true cell probabilities, and π_{g+} and π_{+h} are the row and column margins. If the columns and rows are independent, then the cell probability π_{gh} satisfies the relationship

$$H_o: \pi_{gh}=\pi_{g+}\pi_{+h}, \quad g=1,\ldots,R,\ h=1,\ldots,C \tag{5.1}$$

where $\Sigma\,\pi_{gh} = 1$ and $\pi_{gh}>0$. To test the hypothesis H_o, conventionally we use the test statistic

$$X^2 = \sum_g^R \sum_h^C \frac{\left(n_{gh} - \frac{n_{g+}n_{+h}}{n}\right)^2}{\frac{n_{g+}n_{+h}}{n}} \tag{5.2}$$

When data come from a random sample of clusters, X^2 in (5.2) may not generally have a limiting chi-square distribution because of possible dependence between the members in the same cluster. If the cell counts are based on the b clusters, each including two members, Cohen (1976) showed how to correct for the inappropriateness in the use of the conventional statistic X^2 in (5.2). He proved that under H_o

$$X_C^2 = \frac{1}{1+\hat{a}} X^2 \to \chi^2_{(R-1)(C-1)} \text{ as } b \to \infty \tag{5.3}$$

where $\hat{a}$ is a consistent estimator of a $(0 \leq \hat{a} \leq 1)$.

Altham (1976) generalized Cohen's results for clusters of size M, showing that

$$X_A^2 = \frac{1}{1+\hat{a}(M-1)} X^2 \rightarrow \chi^2_{(R-1)(C-1)} \text{ as } b \rightarrow \infty \quad (5.4)$$

Brier (1979) assumed that the independent clusters have Dirichlet-Multinomial distribution, and he corrected the shortcomings of the usual chi-squared test statistic by a scale factor, derived from the covariance matrix of the Dirichlet-Multinomial distribution, which is somewhat similar with the scale factor of (5.4).

When data arise from a weighted cluster sample survey, we again want to know the degree of inappropriateness in using the conventional chi-squared statistic X_w^2 defined below in (5.5).

Using the same notations and the results shown in section 3, let $\hat{N}_{11},\ldots,\hat{N}_{RC}$ be the weighted cell counts of RxC contingency table from a weighted cluster sample survey. Let $\hat{N}_{g+}$ and $\hat{N}_{+h}$ be the row and column margins, respectively. Then the conventional test statistic is

$$X_w^2 = \sum_g^R \sum_h^C \frac{\left(N_{gh} - \frac{\hat{N}_{g+}\hat{N}_{+h}}{\hat{N}}\right)^2}{\frac{\hat{N}_{g+}\hat{N}_{+h}}{\hat{N}}} \quad (5.5)$$

Due to the effects of weighting and clustering in the process of sampling, the statistic X_w^2 in (5.5) does not have a limiting chi-square distribution. It can be shown that under H_o

$$X_T^2 = \frac{\hat{N}}{\hat{G}} X_w^2 \rightarrow \chi^2_{(R-1)(C-1)} \text{ as } b \rightarrow \infty \quad (5.6)$$

where X_w^2 is given in (5.5) and $\hat{G}$ is given in (5.7).

X_T^2 is the product of the scale factor $\frac{\hat{N}}{\hat{G}}$ and the conventional test statistic X_w^2 defined in (5.5). The effects of clustering and weighting are measured simultaneously by $\hat{G}$. If $w_{\nu\lambda} = w$ and $M_\nu = M$, $\hat{G} = bM(1+\hat{a}(M-1))w^2$ and $\hat{N}=bMw$ so that the scale factor $\frac{\hat{N}}{\hat{G}} = \frac{1}{(1+\hat{a}(M-1))w}$, which is exactly the same scale factor of Altham's result in (5.4) when w = 1.

If a is replaced by a consistent estimator $\hat{a}$ in G, we have

$$\hat{G} = \hat{a} \sum_\nu^b \sum_{\lambda\neq\lambda'}^M w_{\nu\lambda} w_{\nu\lambda'} + \sum_\nu^b \sum_\lambda^{M_\nu} w_{\nu\lambda}^2 \quad (5.7)$$

which converges to G as $b \rightarrow \infty$.

When $\hat{G} \rightarrow G$ and $X_w^2 \rightarrow \frac{G}{\hat{N}} \chi^2_{(R-1)(C-1)}$, the product of $\frac{\hat{N}}{\hat{G}}$ and X_w^2 converges to $\chi^2_{(R-1)(C-1)}$.

Cohen (1976) discusses that the result in (5.4) applies if M is the size of the largest cluster which occurs in the sample when clusters are of unequal size within a single sample. This is useful when $w_{\nu\lambda} = 1$; otherwise we may use the generalized form $C_w(a_k)$ illustrated in the numerical example for practical purposes.

6. Numerical Example

The 1975 Health Interview Survey included data from the St. Paul-Minneapolis primary sampling unit (PSU), a simple random sample of clusters of four households. The grouping of households induces no clustering effects for the variables of interest, which are age and presence of chronic conditions. Therefore, we assume that these households come from random sampling. Bean (1974) described the details of the sample survey design and procedures for weighting.

All persons residing in each sample household are interviewed for sociodemographic and health characteristics. The data are weighted to reflect the population size of the poststratified geographic and demographic domain to which the interviewed individual belongs. This weighting of elementary units is necessary to estimate certain characteristics of the target population at reasonable cost in actual sample survey situations.

However, clustering of household members and weighting of data pertaining to individuals invalidates the simple random sample assumption that is used for conventional statistical methods. The example presented in this section illustrates the use of the new test statistic in correcting the shortcomings of conventional test statistics for a valid statistical inference. This statistic corrects the combined defects of the conventional technique; that is, the effect of clustering and the differential weighting of elementary units.

Among 1,009 persons interviewed in the St. Paul-Minneapolis PSU, 503 persons are from families of one, two, or three members; 506 persons are from families of four persons or more and are excluded from this sample. Our interest is in testing of the hypothesis concerning the independence of age and the prevalence of chronic conditions among the sample persons, using the following 2X2 contingency tables.

Table 1. Unweighted Data

	O	Y	
C	98	85	
N	112	208	Total
			503

O = 45 years of age and over
Y = Under 45 years of age
C = One chronic condition or more
N = No chronic condition

In a report published by the National Center for Health Statistics (NCHS), the weighted data of 503 persons are distributed in a 2x2 table as follows:

Table 2. Weighted Data

	O	Y	
C	167,371	150,483	
N	194,806	364,665	Total
			877,325

O = 45 years of age and over
Y = Under 45 years of age
C = One chronic condition or more
N = No chronic condition

Individual weight ranges from 1,750 to 1,500, excluding a few extreme values. The average weight is about 1,744. When the conventional statistical method is used to analyze the weighted data published by NCHS, the results of chi-square testing of independence are significant in most cases as will be seen later.

The method developed in section 4 is used for the estimation of parameters. Since the data consist of the three different sizes of households, the parameters π_1 and a are separately estimated for the same size of households except one member households. For two person households, Newton-Raphson method using the initial value of $a_2^o = 0.5196$ produces the final estimator $\hat{a}_2 = 0.5189$. The initial and final estimators for the three member family are $a_3^o = 0.1647$ and $\hat{a}_3 = 0.1285$ respectively. The standard deviations of the final estimators are 0.0617 for $\hat{a}_2$ and 0.05666 for $\hat{a}_3$. For both cases, the final estimator is not much different from the initial estimator.

Now we use estimators $\hat{a}_2$ and $\hat{a}_3$ to adjust the conventional χ^2 statistic for testing of independence of columns (age) and rows (presence of chronic condition). We introduce the analyses of unweighted data in Table 3 which is compared with the analysis of weighted data in Table 4 below.

Table 3 Chi-square Test Results of Unweighted Data

Usual χ^2 Value	Scale Factor	Corrected χ^2	Remark
16.476 (1 D.F.)	$C(\hat{a}_\nu=0)=1$	16.476 (1 D.F.)**	Maximum
16.476 (1 D.F.)	$C(\hat{a}_\nu)=0.7506$	12.3673 (1 D.F.)**	
16.476 (1 D.F.)	$C(\hat{a}_\nu=1)=0.4479$	7.3797 (1 D.F.)**	Minimum
16.476 (1 D.F.)	$C(a',M')=0.4908$	8.0865 (1 D.F.)**	$a'=\max(\hat{a}_\nu)$ $M'=\max(M_\nu)$

Table 4 Chi-square Test Results of Weighted Data

Usual χ^2 Value	Scale Factor	Corrected χ^2 Value	Remark
26,604.258 (1 D.F.)	$C_w(\hat{a}_k=0)=0.00056764$	15.1016 (1 D.F.)**	Maximum
26,604.258 (1 D.F.)	$C_w(\hat{a}_k)=0.00042774$	11.3795 (1 D.F.)**	
26,604.258 (1 D.F.)	$C_w(\hat{a}_k=1)=0.00025564$	6.8011 (1 D.F.)**	Minimum

** significant at the 1% level (i.e. $\chi^2 \geq 6.635$ with 1 D.F.)

Table 3 shows the results of the analysis of Table 1. The first and third rows give the maximum and minimum chi-square values. The minimum value is obtained by usual chi-square value multiplied by minimum scale factor derived by setting $\hat{a}_\nu = 1$ in (6.1). The maximum value is attained by adjusting the usual chi-square value by the maximum scale factor attained by setting $\hat{a}_\nu = 0$ in (6.1). The second row gives an actual test score when the conventional test score was multiplied by the scale factor

$$C(a_\nu) = \frac{n_1 + n_2 + n_3}{\sum_{\nu=1}^{3} n_\nu (1+\hat{a}_\nu(M_\nu - 1))} = 0.7506 \quad (6.1)$$

where $\hat{a}_1 = 0$, $\hat{a}_2 = 0.51887$, $\hat{a}_3 = 0.12847$

$n_1 = 81$, $n_2 = 224$, $n_3 = 198$, (n_1 is the number of persons from one member households, similarly for n_2 and n_3) $M_1 = 1$, $M_2 = 2$, $M_3 = 3$.

$C(\hat{a}_\nu)$ can be interpreted as an effective sample proportion out of overall sample size n ($n = n_1 + n_2 + n_3$). The last row gives the chi-square value when usual test score is adjusted by the scale factor (6.1) by setting $a_\nu = a'$ where $a' = \max(a_1 a_2 a_3)$ and $M_\nu = M'$ where $M' = \max(M_1 M_2 M_3)$. Since the minimum chi-square value is 7.3797 with 1 D.F. and hence significant at the 1 percent level, the hypothesis of independence between age and the presence of chronic condition is rejected regardless of the real chi-square value.

Table 4 includes the conventional chi-square test scores and the scores corrected for the combination of clustering and weighting effects arising from weighted cluster sample survey. The first and third rows show the obtainable maximum and minimum chi-square values. The minimum value is obtained from the usual chi-square test score multiplied by the minimum scale factor derived by setting $\hat{a}_k = 1$ in (6.2). The maximum value is similarly attained by the product of the conventional test score and maximum scale factor obtained by setting $\hat{a}_k = 0$ in (6.2). The second row shows that the usual chi-square value is corrected by a more realistic scale factor

$$C_w(a_k) = \frac{N_1 + N_2 + N_3}{\sum_{k=1}^{3} \sum_{\nu=1}^{b_k} (\hat{a}_k \sum_{\lambda \neq \lambda'}^{M_k} w_{\nu\lambda} w_{\nu\lambda'} + \sum_{\lambda}^{M_k} w_{\nu\lambda}^2)}$$

$$= \frac{877{,}325}{2{,}051{,}103{,}490} = 0.0004277 \quad 6.2)$$

where $b_1 = 81$, $b_2 = 112$, $b_3 = 66$ (b_1 is the number of households of one member, similarly for b_2 and b_3)

$N_1 = 143{,}999$, $N_2 = 387{,}801$, and $N_3 = 345{,}525$ (N_1 is the weighted person counts from one member households, similarly for N_2 and N_3)

$\hat{a}_1 = 0$, $\hat{a}_2 = 0.51887$, $\hat{a}_3 = 0.12849$,

$M_1 = 1$, $M_2 = 2$, and $M_3 = 3$.

$C_w(\hat{a}_k)$ can be interpreted as an effective weighted sample proportion out of total weighted counts of 877,325. $C(\hat{a}_\nu)$ in (6.1) is merely a special case of $C_w(\hat{a}_k)$ when $w_{\nu\lambda} = 1$. Since the minimum chi-square test score is 6.8011 and is significant at the one percent level, the null hypothesis of independence is rejected.

When the test statistics are corrected by the methods presented in this paper, the results based on the unweighted but clustered data analyses are remarkably close to those results based on the weighted and clustered data. If the hypothesis of independence is rejected, most likely it will be rejected regardless of the type of data used for the analysis, that is, the conclustion drawn from the analysis of unweighted data will not be altered from the analysis of weighted data. For instance, the independence of age and the presence of one chronic condition or more is rejected by both types of data analysis in the present example.

References

Altham, P.M.E. (1976). Discrete Variable Analysis for Individuals Grouped into Families. Biometrika, 63, 263-369.

Bean, J. (1974). Estimate and Sampling Variance in the Health Interview Survey. Vital and Health Statistics, Series 2-No. 38, National Center for Health Statistics.

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1976). Discrete Multivariate Analysis, the MIT Press, Cambridge, Mass.

Brier, S.S. (1979). Categorical Data Models for Complex Data Structures. Ph.D. Dissertation, School of Statistics, University of Minnesota.

Bryant, E.E., Baird, J.T., and Miller, H.W. (1973). Sample Design and Estimation Procedures for a National Health Examination Survey of Children. Vital and Health Statistics, Series 2-No. 43, National Center for Health Statistics.

Chung, K.L. (1968). A Course in Probability Theory, Harcout, Brace, and World, Inc. New York.

Cohen, J. (1976). The Distribution of Chi-Squared Statistic Under Clustered Sampling from Contingency Tables. Journal of the American Statistical Association 71, 665-670.

Dodes, I.A. (1978). Numerical Analysis for Computer Science, North Holland Scientific Publishers, Ltd., New York.

Fienberg, S.E. (1979). The Use of Chi-Squared Statistics for Categorical Data Problems. Journal of the Royal Statistical Society B,41,1, 54-64.

Pearson, K (1900). On the criterion that a given system of deviation from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philosophical Magazine,5, 157-175.

Plackett, R.L and Paul, S.R. (1978). Dirichlet Models for Square Contingency Tables. Communications in Statistics: Theory and Methods, A7(10), 939-52.

Acknowledgements

The authors would like to thank Mr. E. E. Bryant, the National Center for Health Statistics, for drawing this problem to their attention and supporting the first author for this research at work.