Statistics Canada
Statistique Canada

Methodology Branch

Business Survey Methods Division

Direction de la méthodologie

Division des méthodes d'enquêtes entreprises

Canada

CHI-SQUARE TESTS FOR THE ANALYSIS OF THREE-WAY CONTINGENCY TABLES FROM THE CANADA HEALTH SURVEY

M.A. Hidiroglou and J.N.K. Rao

Working Paper No. BSMD 85-006E

Statistics Canada Statistique Canada

Ottawa

CHI-SQUARE TESTS FOR THE ANALYSIS OF THREE-WAY CONTINGENCY TABLES FROM THE CANADA HEALTH SURVEY

M.A. HIDIROGLOU and J.N.K. RAO

JANUARY 1983

TABLE OF CONTENTS

PAGE

1. INTRODUCTION 1

2. DESCRIPTION OF THE CANADA HEALTH SURVEY (CHS) 2

3. ESTIMATION OF TOTAL COUNTS AND PROPORTIONS 4

4. NOTATION FOR THREE-DIMENSIONAL TABLES 6

5. THREE-DIMENSIONAL TABLES AND CHI-SQUARE TESTS 7

6. COMPUTATIONS FOR ESTIMATED NOMINAL LEVELS 16

7. EMPIRICAL RESULTS 18

8. SUMMARY 31

BIBLIOGRAPHY 32

TABLES

Table 1: Estimated population counts (in thousands), c.v.'s and DEFF's in a three-way table (sex × drug use × age group) -- Adjusted for Age-Sex counts 21

Table 2: (Unadjusted) estimated population counts (in thousands), c.v.'s and DEFF's in a three-way table (sex × drug use × age group) 22

Table 3: Standardized Multinomial Residuals e_{ijk} and Design Based Residuals $\hat{e}_{ijk}$ for the Age-Sex Adjusted Case 23

Table 4: Summary of Statistics for Independence Hypotheses in a Three-Way Table (Age-Sex Adjusted Case) - Example 1 24

Example 2 25

Table 5: Female Population 15 Years and Over by Frequency of Breast Self-Examination × age × education (in thousands) "SMALL" 2×3×3 27

TABLE 6: Female Population 15 Years and Over by Frequency of Breast Self-Examination × age × education (in thousands) "LARGE" 3×4×4 28

TABLE OF CONTENTS

PAGE

TABLE 7: Female Population 15 Years and Over by Frequency of Breast Self-Examination by Age and Education "SMALL" 2×3×3 29

TABLE 8: Female Population 15 Years and Over by Frequency of Breast Self-Examination by Age and Education "LARGE" 3×4×4 30

1. INTRODUCTION

The results of the Canada Health Survey (1981) have been summarized in the form of cross-classified tables of counts, commonly referred to as contingency tables. The units of the sampled population have been cross-classified according to sets of categories such as sex (male, female), age (young, middle-age, old) and biomedical characteristics giving rise to multidimensional tables. Such tables present special problems of analysis and interpretation. Until recently, statistical and computational techniques limited researchers to analyze multidimensional tables by examining the categorical variables two at a time (tests of independence or homogeneity). This type of analysis was limited because it did not allow the simultaneous examination of pairwise relationships and it ignored the possibility of three-factor and higher order interactions among the variables.

Recently, the analysis of multidimensional tables has been greatly aided through the use of loglinear models. Loglinear models express the logarithm of the cross-classified expected cell frequencies as a linear combination of main effects and interaction terms. These models permit the simultaneous examination of pairwise relationships and the effect of the higher order interaction terms. The analysis of multidimensional contingency tables using the loglinear approach is quite sophisticated and quite a few analytical tools are now available (see Fienberg 1980, Haberman 1978). However, these methods cannot be applied straightforwardly to categorical data arising from complex sample surveys without taking into account the effect of stratification and clustering. The effect of sample design has been studied by Rao and Scott (1981), Holt, Scott and Ewings (1980), Fellegi (1980), and Hidiroglou and Rao (1981) for two-way tables. The design may be taken into account by either using a Wald statistic (Koch et al 1975) which incorporates the sampling procedure into the covariance matrix or by suitably modifying the customary Pearson chi-square statistic. The correction to the Pearson chi-square statistic is a deflation factor which depends on design effects of estimated cell proportions and certain estimated marginal proportions (see Section 5).

In this report, three hypotheses postulated for three-way tables will be studied. These hypotheses are as follows:

(a) Complete independence (mutual independence): this type of independence is a natural extension to the two-way independence with the three categorical variables not depending on each other in all possible ways;

(b) Multiple independence of one category with the remaining two categories;

(c) Conditional independence of two categories given the third category.

The tests corresponding to the above hypotheses will be studied by taking the sample design into account and applying them to some three-way tables from the Canada Health Survey (1978-79).

2. DESCRIPTION OF THE CANADA HEALTH SURVEY (CHS)

The broad objectives of the Canada Health Survey (1978-79) were to provide reliable information on the health status of Canadians. The antecedents to health status, such as risk factors, current health status and consequences of the health status were measured on rotating monthly samples of households. The data had two intended uses: (a) to monitor health status on a broad basis and identify problem priorities, and (b) to help develop, implement and evaluate preventive and remedial efforts.

The information collected was made up of two main components. The first, known as the Interview Component, used two types of questionnaires. The first questionnaire covered items which in general required probing by an interviewer and could be obtained for the entire household from a suitable member, such as questions relating to accidents and injuries, chronic conditions, hearing, and disability days.

The second questionnaire covered data which could be sensitive and could only be reliably answered by the person concerned. Due to its content and the need for respondent completion, each household member 15 years old and above was asked to complete it. This questionnaire included queries on alcohol use, tobacco use and health related activities. The field organization of Statistics Canada collected data for the Interview Component using part-time interviewers.

The second component known as the Physical Measures Component was divided into two parts. The first part included physical measurements of blood pressure, cardiorespiratory fitness,height,weight and skinfolds on persons in various age groups. The second part involved the taking of blood samples from persons three years and over in order to determine immune status as well as biochemical and trace metal levels. These physical measurements were performed by part-time nurses employed by the Victorian Order of Nurses under contract to Health and Welfare Canada.

The required annual sample size was 12,000 households (40,000 persons) in the Interview Component from 100 sample geographical clusters in monthly samples of 10 households per cluster. A sub-sample of 4,200 households was selected for the Physical Measures Component in 50 of the 100 Interview sample clusters at the rate of 7 out of the 10 interview households per cluster per month. These 100 clusters were allocated initially to the provinces proportional to the square root of their 1971 Census populations. Three major strata were formed within each province, these strata being major cities, other urban areas and rural areas. Quebec and Ontario were further subdivided into three health regions. The allocation of the provincial culsters to the major strata was done proportional to their respective 1971 Census populations with the requirement that the minimum allocation to a stratum be two clusters.

In each of the major cities, a minimum of two clusters were selected. In the other urban areas a systematic sample of one to eight cities was selected within each province with probability proportional to their 1971 Census populations. Each selected city was allocated one cluster; for major cities and selected urban areas, each cluster was a group of geographically spread blocks. A circular systematic sample of 10 dwellings was then selected within each cluster for the Interview Component. Collapsing of strata was adopted to ensure at least two sample cities in the combined stratum. For rural strata, within each province, three-stage design employing systematic sampling at each stage was adopted. For all major strata, the clusters were chosen in groups of three to allow for rotation between clusters within a three month period with a return to the sample cluster every quarter. New households were selected every month.

3. ESTIMATION OF TOTAL COUNTS AND PROPORTIONS

The Canada Health Survey may be described as a multi-stage stratified cluster sample design. In order to simplify the development of variance formulae, we assume that the primary sampling units (clusters) have been sampled with replacement. This is a reasonable assumption to make because of low sampling fractions at the first stage sampling. The estimates of total are adjusted for post-stratification,using the projected Census age-sex distribution at the provincial level. The resulting estimates should be substantially more efficient than the unadjusted ones for health characteristics closely related to age and sex.

In order to provide estimators for total counts of categorical variables at the provincial level, some preliminary notation is required. To this end, define the variable ${}_a y_{i(hct)}$ for the t-th unit in the c-th sample primary of the h-th stratum as one if the unit belongs to the i-th category (one-way notation) and the a-th age-sex group and zero otherwise $(i = 1, 2, \ldots, I+1)$. In the rural and other urban areas strata, we have three-stage sampling so that the t-th unit corresponds to a sampled dwelling in a sampled second-stage unit. The basic sampling weight for the (hct)-th unit will be denoted by w_{hct}.

The ranges for the pre-script a and subscripts (hct) are as follows: $a = 1, 2, \ldots, A$; $h = 1, 2, \ldots, L$; $c = 1, 2, \ldots, n_h$; $t = 1, \ldots, m_{hc}$. Correspondingly, the indicator variable ${}_ax_{hct}$ will be associated with the (hct)-th unit. This variable takes the value one if the (hct)-th unit belongs to the a-th age-sex group and zero otherwise.

The adjusted estimator of total count in the i-th category at the provincial level is

$$\hat{N}_i = \Sigma_a ({}_a\hat{N}_i / {}_a\hat{N})\ {}_aN, \quad i = 1, \ldots, I+1 \tag{3.1}$$

where

$${}_a\hat{N}_i = \Sigma_h \Sigma_c \{\Sigma_t\ w_{hct}\ {}_ay_{i(hct)}\} = \Sigma_h \Sigma_c\ {}_aB_{i(hc)}, \text{ say}$$

$${}_a\hat{N} = \Sigma_h \Sigma_c \{\Sigma_t\ w_{hct}\ {}_ax_{hct}\} = \Sigma_h \Sigma_c\ {}_aB_{hc}, \text{ say}$$

and ${}_aN$ is the projected census population of the province in the a-th age-sex group at the time of survey.

An estimator of the covariance between two estimated cell counts $\hat{N}_i$ and $\hat{N}_\ell$ is

$$\text{cov}(\hat{N}_i, \hat{N}_\ell) = \Sigma_h \frac{n_h}{n_h - 1} \Sigma_c (z_{ihc} - \bar{z}_{ih})(z_{\ell hc} - \bar{z}_{\ell h}) \tag{3.2}$$

where

$$z_{ihc} = B_{i(hc)} - \Sigma_a ({}_a\hat{N}_i / {}_a\hat{N})\ {}_aB_{hc}$$

and

$$B_{i(hc)} = \Sigma_a\ {}_aB_{i(hc)}, \ \bar{z}_{ih} = \Sigma_c\ z_{ihc}/n_h \ .$$

If $\hat{N}_i(1), \ldots, \hat{N}_i(m)$ denote the estimates (3.1) for the provinces in a region (or Canada) the aggregate estimate will be $\hat{N}_i(+) = \hat{N}_i(1) + \ldots + \hat{N}_i(m)$ and the estimator for the proportion, p_i, in the i-th category is

$$\hat{p}_i = \hat{N}_i(+)/\hat{N}_+(+) \tag{3.3}$$

where $\hat{N}_+(+) = \Sigma_i \hat{N}_i(+)$. Hence, noting that $\hat{p}_i$ is a combined ratio estimator, the covariance of $\hat{p}_i$ and $\hat{p}_\ell$ is estimated as

$$\text{cov}(\hat{p}_i,\hat{p}_\ell) = [\hat{N}_+(+)]^{-2} \Sigma_f \hat{\sigma}_{i\ell}(f) . \tag{3.4}$$

Here, $\hat{\sigma}_{i\ell}(f)$, for the f-th province, is given by (3.2) with z_{iht} replaced by

$$[B_{i(hc)} - \hat{p}_i B_{+(hc)}] - \Sigma_a ({}_aB_{hc}/{}_a\hat{N}) [{}_a\hat{N}_i - \hat{p}_i \, {}_a\hat{N}_+] \tag{3.5}$$

where

$$B_{+(hc)} = \Sigma_i B_{i(hc)} \text{ and } {}_a\hat{N}_+ = \Sigma_i \, {}_a\hat{N}_i .$$

The unadjusted estimator of p_i is $p_i^* = N_+^*(+)/N_i^*(+)$ where $N_i^*(+) = \Sigma_f N_i^*(f)$, $N_{+(+)}^* = \Sigma_i N_i^*(+)$ and $N_i^*(f) = \Sigma_a \, {}_a\hat{N}_i(f)$. Noting that p_i^* is a combined ratio estimator, we have

$$\text{cov}(p_i^*, p_\ell^*) = [N_+^*(+)]^{-2} \Sigma_f \sigma_{i\ell}^*(f) \tag{3.6}$$

where $\sigma_{i\ell}^*(f)$, for the f-th province, is given by (3.2) with z_{ihc} replaced by $B_{i(hc)} - p_i^* B_{hc}$, where $B_{hc} = \Sigma_i B_{i(hc)}$.

4. NOTATION FOR THREE-DIMENSIONAL TABLES

A three-way table consists of frequencies classified by the categories of three variables (e.g. age, sex and drug use). The three variables will be labelled as A (rows), B (columns) and C (layers) and suppose that correspondingly they have I, J, and K categories respectively. The theoretical probability of a randomly chosen observation in the population falling into cell (i,j,k), (i = 1, ..., I;

$j=1, \ldots, J$; $k=1, 2, \ldots, K$) will be denoted as q_{ijk} where $q_{ijk} = N_{ijk}/N_{+++}$, $N_{+++} = \Sigma_i \Sigma_j \Sigma_k N_{ijk}$ and N_{ijk} is the population count for the (i,j,k)-th cell at the Canada level (for brevity, the (+) notation to represent summing over provinces has been dropped). Corresponding estimates obtained from blown-up sample counts will be denoted as $\hat{q}_{ijk}$ for age-sex adjusted counts where $\hat{q}_{ijk} = \hat{N}_{ijk}/\hat{N}_{+++}$ and $\hat{N}_{ijk}$ is an estimator for N_{ijk} with $\hat{N}_{+++} = \Sigma_i \Sigma_j \Sigma_k \hat{N}_{ijk}$. Similarly q^*_{ijk} will denote proportions based on unadjusted counts. The symbol "+" used as a subscript means that the summation is over the symbol that used to be in that position.

5. THREE-DIMENSIONAL TABLES AND CHI-SQUARE TESTS

Loglinear models may be used to express the log probabilities, $\ell n\ q_{ijk}$, in a three-dimensional table as a linear combination of row, column, depth effects and the associated two-factor and three-factor interactions between them. The saturated model for three-dimensional tables is given by

$$\begin{aligned} \ell n\ q_{ijk} = u &+ u_{1(i)} + u_{2(j)} + u_{3(k)} \\ &+ u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)} \\ &i=1, 2, \ldots, I;\ j=1, 2, \ldots, J;\ k=1, 2, \ldots, K; \end{aligned}$$

where the u-terms sum to zero when summed over any subscript i,j,k. For instance,

$$\sum_i u_{123(ijk)} = \sum_j u_{123(ijk)} = \sum_k u_{123(ijk)} = 0.$$

Deletion of the three-factor interaction u_{123} or specified two-factor interaction terms u_{12}, u_{13} or u_{23} lead to different hierarchical models with respect to the cell proportions q_{ijk}. In this report, three types of hierarchical models for three-way tables will be studied.

Footnote: In a hierarchical model, if a u-term is included, all its lower order terms should also be included; conversely, if a u-term is deleted all its higher order terms should also be deleted. Hierarchical models facilitate meaningful interpretation of u-terms and also simplify the analysis. For instance, the interpretation of a main effect u_1 in the presence of the interaction u_{12} is not very meaningful.

a. Complete Independence: A*B*C

The definition of complete independence of the three variables A, B and C is given by the null hypothesis

$$H_o: \; u_{123} = u_{12} = u_{13} = u_{23} = 0$$

which is equivalent to (<=>)

$$H_o: \; f_{ijk}(\underset{\sim}{\theta}) = q_{i++} \, q_{+j+} \, q_{++k} \tag{5.1}$$

$$i = 1, 2, \ldots, I; \; j = 1, 2, \ldots, J; \; k = 1, 2, \ldots, K,$$

where $u_{12} = 0$ means $u_{12(ij)} = 0$ for all i,j and similarly for other interactions terms and $\underset{\sim}{\theta}$ denotes the vector of independent u-terms included in the model. The estimated proportions under (5.1) are given by $f_{ijk}(\underset{\sim}{\hat{\theta}}) = \hat{q}_{i++} \, \hat{q}_{+j+} \, \hat{q}_{++k}$ where $\hat{q}_{i++} = \sum_j \sum_k \hat{q}_{ijk}$, etc.

b. Multiple Independence: A*(BC)

If one factor A is independent of the remaining factors B and C, the corresponding null hypothesis is given by

$$H_o: \; u_{12} = u_{13} = u_{123} = 0 \iff f_{ijk}(\underset{\sim}{\theta}) = q_{i++} \, q_{+jk} \tag{5.2}$$

$$i = 1, 2, \ldots, I; \; j = 1, 2, \ldots, J; \; k = 1, 2, \ldots, K.$$

The hypotheses B*(AC) or C*(AB) are analogous to (5.2). The estimated proportions under the null hypothesis (5.2) are given by $f_{ijk}(\underset{\sim}{\hat{\theta}}) = \hat{q}_{i++} \, \hat{q}_{+jk}$.

c. Conditional Independence: (A*B)|C

If we consider a two-variable wedge taken from a three-variable table, then if the two-variable independence definition is applied to such a wedge, we say that two variables (A and B) are conditionally independent of one another given the third variable (C). The null hypothesis for (A*B)|C is

$$H_o:\ u_{12} = u_{123} = 0 \iff f_{ijk}(\underset{\sim}{\theta}) = q_{i+k}\, q_{+jk}/q_{++k} \tag{5.3}$$

$$i = 1, 2, \ldots, I;\quad j = 1, 2, \ldots, J;\quad k = 1, 2, \ldots, K.$$

Once more, the null hypotheses (B*C)|A and (C*A)|B are analogous to (5.3). The estimates proportions under the model (5.3) are given by $f_{ijk}(\underset{\sim}{\hat{\theta}}) = (\hat{q}_{i+k}\ \hat{q}_{+jk})/\hat{q}_{++k}$.

d. No three-factor interaction

In the loglinear model framework, this hypothesis is given by H_o: $u_{123} = 0$, meaning that the association in the two-way table corresponding to the level of third factor is constant for all levels. This hypothesis cannot be expressed in the form of independence or conditional independence, unlike the hypotheses (a) - (c). The estimates $f_{ijk}(\underset{\sim}{\hat{\theta}})$ are obtained by solving the equations $q_{ij+} = \hat{q}_{ij+}$, $q_{+jk} = \hat{q}_{+jk}$ and $q_{i+k} = \hat{q}_{i+k}$ iteratively for q_{ijk} using the well-known iterative proportional fitting procedure (IPFP). The resulting estimates $f_{ijk}(\underset{\sim}{\hat{\theta}})$ are consistent, as also the estimates for hypotheses (a) - (c).

A statistic used for testing the aforementioned hypotheses is the customary Pearson chi-square given by

$$X_P^2 = n \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \{\hat{q}_{ijk} - f_{ijk}(\underset{\sim}{\hat{\theta}})\}^2 / f_{ijk}(\underset{\sim}{\hat{\theta}}). \tag{5.4}$$

In the case of multinomial sampling (simple random sampling) and large samples, X_P^2 is approximately distributed as a chi-square (χ^2) random variable with degrees of freedom t (say) under the null hypothesis H_o, where t = IJK-I-J-K+2, (I-1)(JK-1), (I-1)(J-1)K and (I-1)(J-1)K-1 for the hypotheses a-d respectively. However, when clustering and/or stratification is involved, as in the case of the CHS design, X_P^2 would no longer be distributed approximately as a χ^2 random variable under H_o. Rao and Scott (1982) have shown, for general loglinear models, that X_P^2 is approximately distributed as a weighted sum $\sum_1^t \delta_i W_i$ of independent χ^2 random variables W_i, each with one degree of freedom. The weights δ_i (eigenvalues of a "design effect" matrix) can be interpreted as generalized design effects (DEFF's) (see Rao and Scott, 1981). With a significant cluster effect, the Statistic X_P^2 could lead to a much larger type I error α (probability of rejection H_o when it is true) than the nominal (say α 0.05 or 0.10) level. Hence, it is necessary to modify X_P^2 to take the survey design into account. Such a statistic is given by

$$X_P^2(\hat{\delta}_.) = X_P^2/\hat{\delta}_. \tag{5.7}$$

where $\hat{\delta}_. = \sum_1^t \hat{\delta}_i/t$ and $\hat{\delta}_.$ is its sample estimate (Rao and Scott 1982).

Rao and Scott have shown that $\hat{\delta}_.$ for the hypotheses (a) - (c), may be obtained knowing only the estimated cell design effects and the estimated design effects of two-way marginals $\hat{q}_{ij+}$, $\hat{q}_{i+k}$, $\hat{q}_{+jk}$ and the one-way marginals $\hat{q}_{i++}$, $\hat{q}_{+j+}$ and $\hat{q}_{++k}$. The hypothesis (d), however does not permit a representation in terms of cell DEFF's and marginal DEFF's.

Denote by $\sigma_{ijk,i'j'k'}$, the (ijk, i'j'k')th element of the covariance matrix $\hat{\underset{\sim}{h}} = (\hat{h}_{111}, \ldots, \hat{h}_{11K}; \ldots; \hat{h}_{IJ1}, \ldots, \hat{h}_{IJK})^T$ where $\hat{h}_{ijk} = \hat{q}_{ijk} - f_{ijk}(\hat{\underset{\sim}{\theta}})$. Then

$$\sum_1^t \delta_i = \bar{\delta} \doteq E(X_P^2) = n \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K \frac{\sigma_{ijk,ijk}}{q_{ijk}(\underset{\sim}{\theta})} \tag{5.5}$$

and

$$2 \sum_1^t \delta_i^2 \doteq V(X_P^2) = 2n^2 \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K \sum_{i'=1}^I \sum_{j'=1}^J \sum_{k'=1}^K \frac{\sigma^2_{ijk,\ i'j'k'}}{q_{ijk}(\underset{\sim}{\theta})\ q_{i'j'k'}(\underset{\sim}{\theta})} \tag{5.6}$$

For complete independence (A*B*C), $t = IJK-I-J-K+2$ and $\hat{\delta}_{.}$ may be obtained from

$$t\,\hat{\delta}_{.} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} \hat{d}_{o,ijk}\,[1 - f_{ijk}(\hat{\underset{\sim}{\theta}})]$$

$$- \sum_{i=1}^{I} (1-\hat{q}_{i++})\,\hat{d}_{(A)i} - \sum_{j=1}^{J} (1-\hat{q}_{+j+})\,\hat{d}_{(B)j}$$

$$- \sum_{k=1}^{K} (1-\hat{q}_{++k})\,\hat{d}_{(C)k}, \qquad (5.7)$$

where $f_{ijk}(\hat{\underset{\sim}{\theta}}) = \hat{q}_{i++}\,\hat{q}_{+j+}\,\hat{q}_{++k}$ and $\hat{d}_{o,ijk}$ is the (i,j,k)-th cell DEFF under H_o; that is

$$\hat{d}_{o,ijk} = \frac{var(\hat{q}_{ijk})}{[\hat{q}_{i++}\,\hat{q}_{+j+}\,\hat{q}_{++k}\,(1-\hat{q}_{i++}\,\hat{q}_{+j+}\,\hat{q}_{++k})/n]}$$

with $var(\hat{q}_{ijk})$ being the estimated variance, under the design, of the estimated proportion $\hat{q}_{ijk}$. Correspondingly, $\hat{d}_{(A)i}$, $\hat{d}_{(B)j}$ and $\hat{d}_{(C)k}$ are the DEFF's of the i-th row, j-th column and k-th layer marginals, i.e.,

$$\hat{d}_{(A)i} = \frac{var(\hat{q}_{i++})}{[\hat{q}_{i++}\,(1-\hat{q}_{i++})/n]} ,$$

with $\hat{d}_{(B)j}$ and $\hat{d}_{(C)k}$ similarly defined. The estimated variance of $\hat{q}_{i++}$ is obtained from (3.4).

The covariance of $\hat{q}_{ijk}$ and $\hat{q}_{i'j'k'}$, denoted as $\sigma_{ijk,i'j'k'}$, is estimated as

$$\hat{\sigma}_{ijk,i'j'k'} = \hat{N}_{+++}^{-2} \sum_{f}\Bigg[\sum_{h} \frac{n_h(f)}{n_h(f)-1} \sum_{c}\left(z_{ijk(hc)}(f) - \bar{z}_{ijk(h)}(f)\right)$$

$$\left(z_{i'j'k'(hc)}(f) - \bar{z}_{i'j'k'(h)}(f)\right)\Bigg] \qquad (5.8)$$

with obvious extension to the notation given in Section 2, where the accumulation of the covariance is performed by summing over provinces (f). The z-values that enter into (5.8) are

$$\begin{aligned} z_{ijk(hc)}(f) = {} & B_{ijk(hc)}(f) - \hat{q}_{i++}\,\hat{q}_{+j+}\,B_{++k(hc)}(f) \\ & - \hat{q}_{i++}\,\hat{q}_{++k}\,B_{+j+(hc)}(f) - \hat{q}_{+j+}\,\hat{q}_{++k}\,B_{i++(hc)}(f) \\ & + 2\,\hat{q}_{i++}\,\hat{q}_{+j+}\,\hat{q}_{++k}\,B_{+++(hc)}(f) \\ & - \Sigma_a\left({}_aB_{hc}(f)/{}_a\hat{N}(f)\right)\left\{{}_a\hat{N}_{ijk}(f) - \hat{q}_{i++}\,{}_a\hat{N}_{+j+}(f)\,\hat{q}_{++k}\right. \\ & - \hat{q}_{i++}\,\hat{q}_{+j+}\,{}_a\hat{N}_{++k}(f) - \hat{q}_{+j+}\,\hat{q}_{++k}\,{}_a\hat{N}_{i++}(f) \\ & \left. + 2\,\hat{q}_{i++}\,\hat{q}_{+j+}\,\hat{q}_{++k}\,{}_a\hat{N}_{+++}(f)\right\}. \end{aligned} \qquad (5.9)$$

For multiple independence A*(BC), $t = (I-1)(JK-1)$ and $\hat{\delta}_{.}$ may be obtained from

$$t\,\hat{\delta}_{.} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} \hat{d}_{o,ijk}\left(1 - q_{ijk}(\hat{\underset{\sim}{\theta}})\right)$$

$$- \sum_{i=1}^{I}(1-\hat{q}_{i++})\,\hat{d}_{(A)i} - \sum_{j=1}^{J}\sum_{k=1}^{K}(1-\hat{q}_{+jk})\,\hat{d}_{(BC)jk} \qquad (5.10)$$

with $\hat{d}_{o,ijk}$ and $\hat{d}_{(A)i}$ defined as before and

$$\hat{d}_{(BC)jk} = \frac{\mathrm{var}(\hat{q}_{+jk})}{[\hat{q}_{+jk}(1-\hat{q}_{+jk})/n]}\,.$$

The estimated variance of $\hat{q}_{+jk}$ is obtained from (3.4).

The covariance of $\sigma_{ijk,i'j'k'}$ is estimated from (5.9) using

$$z_{ijk(hc)}(f) = B_{ijk(hc)}(f) - \hat{q}_{i++}\, B_{+jk(hc)}(f) - \hat{q}_{+jk}\, B_{i++(hc)}(f)$$

$$+ \hat{q}_{i++}\, \hat{q}_{+jk}\, B_{+++(hc)}(f)$$

$$-\Sigma_a \left({}_aB_{hc}(f)/{}_a\hat{N}(f)\right) \left\{{}_a\hat{N}_{ijk}(f) - \hat{q}_{i++}\, {}_a\hat{N}_{jk}(f)\right.$$

$$\left. - \hat{q}_{+jk}\, {}_a\hat{N}_{i++}(f) + \hat{q}_{i++}\, \hat{q}_{+jk}\, {}_a\hat{N}_{+++}(f)\right\} . \qquad (5.11)$$

For conditional independence (A*B)|C, $t = (I-1)(J-1)K$ and $\hat{\delta}_.$ may be obtained from

$$t\,\hat{\delta}_. = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \hat{d}_{o,ijk}(1 - \hat{q}_{i+k}\, \hat{q}_{+jk}/\hat{q}_{++k})$$

$$+ \sum_{k=1}^{K} \hat{d}_{(C)k}(1 - \hat{q}_{++k}) - \sum_{i=1}^{I} \sum_{k=1}^{K} d_{(AC)ik}\, (1 - \hat{q}_{i+k})$$

$$- \sum_{j=1}^{J} \sum_{k=1}^{K} \hat{d}_{(BC)jk}\, (1 - \hat{q}_{+jk}) \qquad (5.12)$$

with obvious definitions of $\hat{d}$'s in the above expression. The covariance $\hat{\sigma}_{ijk,k'j'k'}$ is estimated from (5.12) using

$$z_{ijk(hc)}(f) = B_{ijk(hc)}(f) - \frac{\hat{q}_{+jk}}{\hat{q}_{++k}} B_{i+k(hc)}(f)$$

$$- \frac{\hat{q}_{i+k}}{\hat{q}_{++k}} B_{+jk(hc)}(f) + \frac{\hat{q}_{i+k}\, \hat{q}_{+jk}}{\hat{q}^2_{++k}} B_{++k(hc)}(f)$$

$$- \Sigma_a \left({}_aB_{hc}(f)/{}_a\hat{N}(f)\right) \left\{{}_a\hat{N}_{ijk}(f) - \frac{\hat{q}_{+jk}}{\hat{q}_{++k}}\, {}_a\hat{N}_{i+k}(f)\right.$$

$$\left. - \frac{\hat{q}_{i+k}}{\hat{q}_{++k}}\, {}_a\hat{N}_{+jk}(f) + \frac{\hat{q}_{i+k}\, \hat{q}_{+jk}}{\hat{q}^2_{++k}}\, {}_a\hat{N}_{+++}(f)\right\}. \qquad (5.13)$$

For each of the aforementioned hypotheses, the Wald statistic is given by

$$X_W^2 = \hat{\underset{\sim}{h}}^T \hat{\underset{\sim}{\Sigma}}^- \hat{\underset{\sim}{h}} \tag{5.14}$$

where $\hat{\underset{\sim}{\Sigma}}^-$ is the generalized Moore-Penrose inverse of the estimated covariance matrix of $\hat{h}$. The Wald statistic is asymptotically χ^2 with t degrees of freedom under the particular null hypothesis of interest H_o, where t is as given previously. The statistic (5.14) is unique for any choice of generalized inverse. It requires the knowledge of the full estimated covariance matrix, unlike $X_P^2(\hat{\delta}_.)$. Moreover, as t increases, $\hat{\underset{\sim}{\Sigma}}$ and hence X_W^2 are likely to become unstable: empirical evidence is provided in Section 7.

The Pearson statistic X_P^2 could have also been modified using the mean of the cell design effects (DEFF) or the mean of the eigenvalues of the design effect matrix for the individual cell proportions. These modifications are given by $X_P^2(\hat{d}_{1,.}) = X_P^2/\hat{d}_{1,.}$ and $X_P^2(\hat{\lambda}_{1,.}) = X_P^2/\hat{\lambda}_{1,.}$ where

$$\hat{d}_{1,.} = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \hat{d}_{1,ijk}/(IJK) \tag{5.15}$$

and

$$\hat{\lambda}_{1,.} = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} (1-\hat{q}_{ijk})\, \hat{d}_{1,ijk}/(IJK-1) \tag{5.16}$$

where $\hat{d}_{1,ijk} = \mathrm{var}(\hat{q}_{ijk})/[\hat{q}_{ijk}(1-\hat{q}_{ijk})/n]$.

These modified statistics, although depend only on the cell DEFF's are not likely to be satisfactory in practice since they do not depend on the hypothesis under consideration, unlike $X_P^2(\hat{\delta}_.)$.

Standardized residuals, taking the design into account, are given by

$$\hat{e}_{ijk} = \hat{h}_{ijk}/(\hat{\sigma}_{ijk,ijk})^{\frac{1}{2}} \;. \tag{5.17}$$

The $\hat{e}_{ijk}$'s are approximately $N(0,1)$ under H_o. The corresponding standardized residuals under the assumption of multinomial sampling (Haberman, 1973) are given by

a) $$e_{ijk} = \frac{\hat{q}_{ijk} - \hat{q}_{i++}\,\hat{q}_{+j+}\,\hat{q}_{++k}}{[\hat{q}_{i++}\,\hat{q}_{+j+}\,\hat{q}_{++k}\,(1-\hat{q}_{i++}\,\hat{q}_{+j+} - \hat{q}_{i++}\,\hat{q}_{++k} - \hat{q}_{+j+}\,\hat{q}_{++k} + 2\,\hat{q}_{i++}\,\hat{q}_{+j+}\,\hat{q}_{++k})/n]^{\frac{1}{2}}}$$

in the case of complete independence A*B*C,

b) $$e_{ijk} = \frac{\hat{q}_{ijk} - \hat{q}_{i++}\,\hat{q}_{ijk}}{[\hat{q}_{i++}\,\hat{q}_{+jk}\,(1-\hat{q}_{i++})(1-\hat{q}_{+jk})/n]^{\frac{1}{2}}}$$

in the case of multiple independence A*(BC),

c) $$e_{ijk} = \frac{\hat{q}_{ijk} - \hat{q}_{i+k}\,\hat{q}_{+jk}/\hat{q}_{++k}}{\left[\frac{\hat{q}_{i+k}\,\hat{q}_{+jk}}{\hat{q}_{++k}}\left(1-\frac{\hat{q}_{i+k}}{\hat{q}_{++k}}\right)\left(1-\frac{\hat{q}_{+jk}}{\hat{q}_{++k}}\right)/n\right]^{\frac{1}{2}}}$$

in the case of conditional independence (A*B)|C.

Standardized residuals are helpful in detecting particular deviations from H_o.

6. COMPUTATIONS FOR ESTIMATED NOMINAL LEVELS

As noted in Section 5, the asymptotic distribution of the Pearson statistic X_P^2 under H_o is of the form $\sum_{i=1}^{t} \delta_i W_i$ where the W_i's are independent χ_1^2 random variables. Estimated nominal levels for X_P^2 and its modifications $X_P^2(b_.) = X_P^2/b_.$ for suitable $b_.$, are obtained using Satterthwaite's (1946) approximation. This approximation requires the knowledge of $\delta_.$ and the coefficient of variation, c, of the δ_i's. This approximation treats

$$X_P^2(S) = X_P^2/[\delta_.(1+c^2)] \tag{6.1}$$

as χ_ν^2, a χ^2 random variables with $\nu = t/(1+c^2)$ degrees of freedom. The coefficient of variation may be obtained from (5.6) by noting that

$$\begin{aligned} V(X_P^2) &\doteq 2 \sum_{i=1}^{t} \delta_i^2 \\ &= 2[t + \sum_{i=1}^{t} (\delta_i - \delta_.)^2/\delta_.^2]\, \delta_.^2 \\ &= 2\, t[1+c^2]\, \delta_.^2 \end{aligned} \tag{6.2}$$

or

$$c^2 = V(X_P^2)/[2\, t\, \delta_.^2] - 1, \tag{6.3}$$

where $t\, \delta_. = E(X_P^2)$ is given by (5.5).

The asymptotic significance level or type I error of X_P^2 and $X_P^2(b_.)$ are obtained as

$$\begin{aligned} SL(X_P^2) &= Pr[X_P^2 \geq \chi_t^2(\alpha) | H_o] \\ &\doteq P[\chi_\nu^2 \geq \chi_t^2(\alpha)/\{\delta_.(1+c^2)\}] \end{aligned} \tag{6.4}$$

and

$$\hat{e}_{ijk} = \hat{h}_{ijk}/(\hat{\sigma}_{ijk,ijk})^{\frac{1}{2}} \ . \qquad (5.17)$$

The $\hat{e}_{ijk}$'s are approximately $N(0,1)$ under H_o. The corresponding standardized residuals under the assumption of multinomial sampling (Haberman, 1973) are given by

a) $$e_{ijk} = \frac{\hat{q}_{ijk} - \hat{q}_{i++}\,\hat{q}_{+j+}\,\hat{q}_{++k}}{[\hat{q}_{i++}\,\hat{q}_{+j+}\,\hat{q}_{++k}\,(1-\hat{q}_{i++}\,\hat{q}_{+j+} - \hat{q}_{i++}\,\hat{q}_{++k} - \hat{q}_{+j+}\,\hat{q}_{++k} + 2\,\hat{q}_{i++}\,\hat{q}_{+j+}\,\hat{q}_{++k})/n]^{\frac{1}{2}}}$$

in the case of complete independence A*B*C,

b) $$e_{ijk} = \frac{\hat{q}_{ijk} - \hat{q}_{i++}\,\hat{q}_{ijk}}{[\hat{q}_{i++}\,\hat{q}_{+jk}\,(1-\hat{q}_{i++})(1-\hat{q}_{+jk})/n]^{\frac{1}{2}}}$$

in the case of multiple independence A*(BC),

c) $$e_{ijk} = \frac{\hat{q}_{ijk} - \hat{q}_{i+k}\,\hat{q}_{+jk}/\hat{q}_{++k}}{\left[\frac{\hat{q}_{i+k}\,\hat{q}_{+jk}}{\hat{q}_{++k}}\left(1-\frac{\hat{q}_{i+k}}{\hat{q}_{++k}}\right)\left(1-\frac{\hat{q}_{+jk}}{\hat{q}_{++k}}\right)/n\right]^{\frac{1}{2}}}$$

in the case of conditional independence (A*B)|C.

Standardized residuals are helpful in detecting particular deviations from H_o.

6. COMPUTATIONS FOR ESTIMATED NOMINAL LEVELS

As noted in Section 5, the asymptotic distribution of the Pearson statistic X_P^2 under H_o is of the form $\sum_{i=1}^{t} \delta_i W_i$ where the W_i's are independent χ_1^2 random variables. Estimated nominal levels for X_P^2 and its modifications $X_P^2(b_.) = X_P^2/b_.$ for suitable $b_.$, are obtained using Satterthwaite's (1946) approximation. This approximation requires the knowledge of $\delta_.$ and the coefficient of variation, c, of the δ_i's. This approximation treats

$$X_P^2(S) = X_P^2/[\delta_.(1+c^2)] \qquad (6.1)$$

as χ_ν^2, a χ^2 random variables with $\nu = t/(1+c^2)$ degrees of freedom. The coefficient of variation may be obtained from (5.6) by noting that

$$V(X_P^2) \doteq 2 \sum_{i=1}^{t} \delta_i^2$$

$$= 2[t + \sum_{i=1}^{t} (\delta_i - \delta_.)^2/\delta_.^2]\, \delta_.^2$$

$$= 2\, t[1+c^2]\, \delta_.^2 \qquad (6.2)$$

or

$$c^2 = V(X_P^2)/[2\, t\, \delta_.^2] - 1, \qquad (6.3)$$

where $t\,\delta_. = E(X_P^2)$ is given by (5.5).

The asymptotic significance level or type I error of X_P^2 and $X_P^2(b_.)$ are obtained as

$$SL(X_P^2) = Pr[X_P^2 \geq \chi_t^2(\alpha) | H_o]$$

$$\doteq P[\chi_\nu^2 \geq \chi_t^2(\alpha)/\{\delta_.(1+c^2)\}] \qquad (6.4)$$

and

$$\begin{aligned} SL[X_P^2(b_{\cdot})] &= Pr[X_P^2(b_{\cdot}) \geq \chi_t^2(\alpha)|H_o] \\ &\doteq P[\chi_\nu^2 \geq \{b_{\cdot}\ \chi_t^2(\alpha)\}/\{\delta_{\cdot}(1+c^2)\}] \qquad (6.5) \end{aligned}$$

where $\chi_t^2(\alpha)$ is the upper α percentage point of a χ^2 random variable with t degrees of freedom. The test statistic

$$X_P^2(S;\ \alpha) = X_P^2(S)\ \chi_t^2(\alpha)/\chi_\nu^2(\alpha)$$

gives nominal significance level under Satterthwaite's approximation:

$$\begin{aligned} SL[X_P^2(S;\ \alpha)] &= P[X_P^2(S;\ \alpha) \geq \chi_t^2(\alpha)] \\ &= P[\chi_\nu^2 \geq \chi_\nu^2(\alpha)] \\ &= \alpha\ . \end{aligned}$$

As noted before, the Wald statistic X_W^2 gives the nominal significance level for large samples: i.e. $SL(X_W^2) = P[X_W^2 \geq \chi_t^2(\alpha)] \doteq P[\chi_t^2 \geq \chi_t^2(\alpha)] = \alpha$. Estimated significance levels are obtained by replacing the respective parameters $b_{\cdot}$, $\delta_{\cdot}$ and c by their estimates $\hat{b}_{\cdot}$, $\hat{\delta}_{\cdot}$ and $\hat{c}$.

7. EMPIRICAL RESULTS

Two examples using data from the Canada Health Survey will be presented. The first example (sex × drug use × age group) will contrast the estimated design effects using age-sex adjusted cell counts and unadjusted cell counts. The second example (frequency of breast self-examination × education × age group) will contrast the analysis on two different sized tables using the same data set and adjusted age-sex counts. These two examples are mainly intended as illustrations of the effect of survey design on customary chi-square tests applied to three-way tables and the performance of modified statistics and the Wald statistic.

. Example 1

Consider the estimated population counts cross-classified by sex male, female), drug use (five categories: 0, 1, 2, 3, 4+ drug classes in a 2-day period) and age group (four categories: 0-14, 15-44, 45-64, 65+). Here, $n = 31,668$; $I = 2$, $J = 5$, and $K = 4$. The sex, drug use and age group categories will be denoted by A, B and C respectively. The estimated counts (in thousands) of the population reporting in each drug × sex category × age group category are given in Tables 1 and 2.

Comparing Tables 1 and 2, it is clear that poststratification adjustment for age-sex has led to a substantial reduction in cell DEFF's, the average DEFF $d_{1,.}$ being 1.614 compared to $d^*_{1,.} = 3.024$ for the unadjusted cell. The d_{ijk}'s vary considerably across all cross-classifications, ranging from 0.468 to 3.906 (similarly for d^*_{ijk}'s).

The proposed measures of the design effects matrix along with the values of X^2_P, its modifications, the Wald statistic X^2_W and associated significance levels (for nominal size 0.05) are given below for the test of complete independence hypothesis A*B*C:

Unadjusted for Age-Sex Case

$d^*_{1,\cdot} = 3.024$, $\delta^*_{\cdot} = 2.922$, $\lambda^*_{1,\cdot} = 3.011$, c.v.$(\delta^*_{\cdot}) = 1.531$;

$X_P^2 = 3630$, $X_P^2(\delta^*_{\cdot}) = 1242$, $X_P^2(\lambda^*_{1,\cdot}) = 1205$, $X_P^2(d^*_{1,\cdot}) = 1200$,

$X_P^2(S,0.05) = 966$, $X_W^2 = 3743$;

$SL(X_P^2) \doteq 0.882$, $SL[X_P^2(\delta^*_{\cdot})] \doteq 0.157$, $SL[X_P^2(\lambda^*_{1,\cdot})] \doteq 0.140$,

$SL[X_P^2(d^*_{1,\cdot})] \doteq 0.138$.

Adjusted for Age-Sex Case

$\hat{d}_{1,\cdot} = 1.614$, $\hat{\delta}_{\cdot} = 2.092$, $\hat{\lambda}_{1,\cdot} = 1.615$, c.v.$(\hat{\delta}) = 1.538$;

$X_P^2 = 3634$, $X_P^2(\hat{\delta}_{\cdot}) = 1737$, $X_P^2(\hat{\lambda}_{1,\cdot}) = 2250$, $X_P^2(\hat{d}_{1,\cdot}) = 2251$,

$X_P^2(S,0.05) = 1348$, $X_W^2 = 9.0248 \times 10^7$;

$SL(X_P^2) \doteq 0.720$, $SL[X_P^2(\hat{\delta}_{\cdot})] \doteq 0.158$, $SL[X_P^2(\hat{\lambda}_{1,\cdot})] \doteq 0.343$,

$SL[X_P^2(\hat{d}_{1,\cdot})] \doteq 0.344$.

For both the adjusted case and unadjusted case, the estimated significance level (SL) is unacceptably high, being 0.72 and 0.88 respectively compared to the nominal level of 0.05. In the unadjusted case, the modification $X_P^2(\delta^*_{\cdot})$, reduces SL to around 0.16 from 0.88, but it is not totally satisfactory (0.16 compared to the nominal level 0.05) due to the large c.v.'s of the δ^*_i's (c.v.$(\delta^*_{\cdot}) = 1.531$). The modifications $X_P^2(\lambda^*_{1,\cdot})$ and $X_P(d^*_{1,\cdot})$ have essentially the same SL (Fellegi 1980 proposed the latter modification: X_P^2/(average cell DEFF)). For the adjusted case, the $\hat{\delta}_{\cdot}$ modification has brought down the SL from 0.72 to 0.16; however, the $\hat{\lambda}_{1,\cdot}$ and $\hat{d}_{1,\cdot}$ modifications provide SL around 0.34. For both the adjusted and unadjusted cases, the Wald statistic X_W^2 is surprisingly much higher than all the other statistics, including X_P^2 (especially the value for the adjusted case).

The poststratification adjustment clearly provides a more powerful test here (compare $X_p^2(S,0.05) = 966$ in the unadjusted case to $X_p^2(S,0.05) = 1348$ in the adjusted case). Since the Satterthwaite correction $X_p^2(S, 0.05)$ for both the adjusted and unadjusted cases is much higher than the upper percentage of χ_{31}^2, viz, $\chi_{31}^2(0.05) = 19.2$, the null hypothesis of complete independence is not tenable.

The standardized residuals to detect deviations from H_o for the case of complete independence are provided in Table 3. The multinomial residuals e_i are much larger than the design-based residuals $\hat{e}_i$, especially for females over 45 whose drug consumption exceeds two or more drugs in a two-day period. Examining the residuals (adjusted case) in Table 3, one may conclude that in the age group 45-64 significantly more females than males use drugs and both males and females in the age group 65+ use two or more drugs.

Table 4 summarizes the statistics associated with the three hypotheses a, b and c permitting a representation of $\hat{\delta}_.$ in terms of cell DEFF's and marginal DEFF's. Note that $\hat{\delta}_.$ differs from hypothesis, whereas $\hat{d}_{1,.}$ and $\hat{\lambda}_{1,.}$ are the same regardless of the hypothesis. The value of the Wald statistic, X_W^2, in each case is much higher compared to the other chi-squared statistics, including X_p^2. The volatility of the Wald statistic is quite surprising: this may be due to the instability of the covariance matrix. Alternative Wald statistics (e.g. those based on weighted least squares for the loglinear model), however, might have better stability than X_W^2. The effect of survey design on SL of X_p^2 is quite severe here (SL ranging from 0.300 to 0.762) compared to the nominal level 0.05, for the adjusted age-sex case. The modification $X_p^2(\hat{\delta}_.)$ brought the SL down to about 0.14 on the average. The modifications $X_p^2(\hat{d}_{1,.})$ and $X_p^2(\hat{\lambda}_{1,.})$ are not as stable as $X_p^2(\hat{\delta})$ eventhough the SL is quite close to 0.05 in two cases, viz A*(BC) and (A*C)|B. It should be noted that we have computed significance levels using the Satterthwaite approximation and it is not clear if this approximation remains accurate when the c.v. of $\hat{\delta}_i$'s is very large.

TABLE 1: Estimated population counts (in thousands), c.v.'s and DEFF's in a three-way table (sex × drug use × age group) -- Adjusted for Age-Sex counts

SEX	DRUGS				
	0	1	2	3	4+
	AGE: 0-14				
MALE	1576a 0.025b 1.397c	881 0.037 1.741	275 0.055 1.166	79 0.130 1.826	22 0.248 1.896
FEMALE	1477 0.020 0.900	880 0.048 2.869	253 0.069 1.655	67 0.149 2.064	21 0.259 1.919
	AGE: 15-44				
MALE	3766 0.011 0.753	1277 0.041 3.096	364 0.063 1.996	99 0.143 2.796	17 0.221 1.124
FEMALE	2753 0.023 2.313	1694 0.026 1.691	736 0.043 1.925	252 0.105 3.906	60 0.113 1.058
	AGE: 44-64				
MALE	1117 0.022 0.786	641 0.034 1.051	272 0.063 1.559	106 0.121 2.153	37 0.181 1.693
FEMALE	753 0.036 1.388	750 0.031 1.033	428 0.047 1.335	227 0.056 0.981	120 0.092 1.397
	AGE: 65+				
MALE	297 0.049 0.981	286 0.042 0.687	187 0.042 0.468	77 0.106 1.202	39 0.127 0.864
FEMALE	257 0.051 0.927	336 0.071 2.342	255 0.069 1.689	166 0.088 1.792	117 0.114 2.126

NOTE: "a" stands for estimated population count

"b" stands for coefficient of variation of the cell

"c" stands for the cell DEFF

TABLE 2: (Unadjusted) estimated population counts (in thousands), c.v.'s and DEFF's in a three-way table (sex × drug use × age group)

SEX	DRUGS				
	0	1	2	3	4+
	AGE: 0-14				
MALE	1524[a] 0.053[b] 2.705[c]	860 0.084 4.785	272 0.114 3.018	73 0.171 2.504	21 0.288 2.365
FEMALE	1439 0.060 2.788	874 0.090 6.178	246 0.098 2.335	68 0.157 2.601	22 0.342 2.966
	AGE: 15-24				
MALE	3481 0.046 4.137	1176 0.052 3.198	344 0.100 3.472	92 0.162 3.048	16 0.251 1.522
FEMALE	2565 0.042 3.326	1598 0.053 3.273	685 0.068 2.721	240 0.127 4.912	56 0.117 1.481
	AGE: 45-64				
MALE	1050 0.040 2.896	599 0.064 4.228	255 0.088 3.530	96 0.124 2.354	35 0.188 1.925
FEMALE	711 0.048 5.276	703 0.059 3.567	395 0.064 2.747	213 0.064 1.139	112 0.099 1.654
	AGE: 65+				
MALE	262 0.074 3.809	253 0.085 3.574	168 0.082 1.829	69 0.133 1.892	34 0.151 1.084
FEMALE	238 0.072 3.699	310 0.103 5.134	240 0.068 1.792	156 0.097 2.220	109 0.143 3.291

NOTE: "a" stands for estimated population count

"b" stands for coefficient of variation of the cell

"c" stands for the cell DEFF

TABLE 3: Standardized Multinomial Residuals e_{ijk} and Design Based Residuals $\hat{e}_{ijk}$ for the Age-Sex Adjusted Case

SEX	DRUGS				
	0	1	2	3	4+
			AGE: 0-14		
MALE	9.15a 4.51b	5.79 2.89	-6.03 -3.80	-8.38 -4.85	-7.75 -5.55
FEMALE	1.50 1.02	4.71 1.96	-8.96 -5.57	-10.68 -6.13	-8.44 -5.52
			AGE: 15-44		
MALE	40.39 30.84	-18.79 -10.30	-26.29 -12.04	-22.41 - 9.55	-19.47 -18.68
FEMALE	-6.17 -3.81	3.82 1.94	6.02 2.62	-0.97 -0.35	-9.99 -5.71
			AGE: 45-64		
MALE	-2.33 -1.27	-0.50 -0.26	0.80 0.35	0.49 0.24	-1.18 -0.66
FEMALE	-29.01 -19.59	7.58 4.23	18.67 8.60	22.64 10.26	22.29 7.52
			AGE: 65+		
MALE	-23.36 -17.76	-0.88 -0.79	11.14 8.01	7.91 3.56	7.94 4.02
FEMALE	-28.40 -21.94	4.51 1.72	22.06 8.07	30.71 8.28	39.48 7.54

NOTE: "a" Standardized residuals for multinomial sampling: e_{ijk}

"b" Standardized residuals for the given design : $\hat{e}_{ijk}$

TABLE 4: Summary of Statistics for Independence Hypotheses in a Three-Way Table (Age-Sex Adjusted Case) - Example 1

STATISTIC	INDEPENDENCE HYPOTHESIS						
	COMPLETE	MULTIPLE			CONDITIONAL		
	A*B*C	A*BC	B*AC	C*AB	A*B|C	A*C|B	B*C|A
$\hat{d}_{1,\cdot}$	1.614	1.614	1.614	1.614	1.614	1.614	1.614
$\hat{\delta}_{\cdot}$	2.092	1.400	2.251	2.087	1.626	1.394	2.311
$\hat{\lambda}_{1,\cdot}$	1.615	1.615	1.615	1.615	1.615	1.615	1.615
c.v.$(\hat{d}_{1,i})$	0.449	0.449	0.449	0.449	0.449	0.449	0.449
c.v.$(\hat{\delta}_i)$	1.539	1.024	1.369	1.270	0.857	1.048	1.115
X_P^2	3634	1000	3446	2558	95	23	2493
$X_P^2(S,0.05)$	1348	608	1226	1003	52	14	1257
X_W^2	9.025×10^7	4.923×10^7	1.286×10^4	6.825×10^7	1199	1692	2.984×10^4
SL(X_P^2)	0.720	0.334	0.762	0.719	0.429	0.300	0.776
SL$[X_P^2(\hat{d}_{1,\cdot})]$	0.344	0.056	0.391	0.319	0.098	0.059	0.390
SL$[X_P^2(\hat{\delta}_{\cdot})]$	0.160	0.111	0.143	0.134	0.095	0.110	0.119
SL$[X_P^2(\hat{\lambda}_{1,\cdot})]$	0.343	0.054	0.391	0.318	0.097	0.059	0.389

. Example 2

Selected female health practices and associated risks (such as the use of birth control pills or hormones and smoking), frequency of the Pap smear test and of breast self-examination were queried in the self-completed questionnaire. This questionnaire applies to the population aged 15 and over. The results of the analysis on breast self-examination will be discussed in this example. The categories of interest are education, frequency of breast self-examination and age. The counts arising from these categories will be summarized in two tables of different dimensions. The first table referred to as "SMALL", is cross-classified by education (secondary or less, some post-secondary and post-secondary), frequency of breast-examination (monthly and quarterly, less often, never) and age (15-24, 25-44, and 45+). Here $n = 9918$, $I = 2$, $J = 3$, and $K = 3$. The education, frequency and age variables will be denoted by A, B and C respectively. The second table referred to as "LARGE" is cross classified using the same variables of the first table in a finer cross-classification: education (secondary or less, some post-secondary, post-secondary), frequency of breast examination (monthly, quarterly, less often, never) and age (15-19, 20-24, 25-44, 45-64). Here $n = 8657$, $I = 3$, $J = 4$ and $K = 4$. The estimated age-sex adjusted counts (in thousands) of the population in "SMALL" are given in Table 5 and correspondingly the counts for "LARGE" are given in Table 6. The corresponding categorical analyses on "SMALL" and "LARGE" are given in Tables 7 and 8 respectively.

Comparing the results of Table 5 to Table 6, several points may be noted. Firstly, in both tables the coefficients of variation and design effects vary considerably between cells. Secondly, the design effects do not seem in any way to be related to the count size.

Tables 7 and 8 summarize the statistics associated with the three hypotheses a, b and c permitting a representation of $\hat{\delta}_.$ in terms of cell DEFF's and marginal DEFFs. As previously noted in Table 4, $\hat{\delta}_.$ differs from hypothesis to hypothesis, whereas $\hat{d}_{1,.}$ and $\hat{\lambda}_{1,.}$ are the same regardless of the hypothesis. As expected, the coefficients of variation for $\hat{d}_{1,i}$'s and $\hat{\delta}_i$'s are larger with the "LARGE" contingency table as opposed to those associated with the "SMALL" contingency table. It is clear from both Tables 7 and 8 that $SL(X_P^2)$ is unacceptably high, ranging from 0.494 to 0.621 for the "SMALL" contingency table and from 0.721 to 0.826 for the "LARGE" contingency table whereas $\alpha = 0.05$. Hence the effect of survey design on X_P^2 is severe. The corrected statistics $X_P^2(\hat{\lambda}_{1,.})$ and $X_P^2(\hat{d}_{1,.})$, where $\hat{\lambda}_{1,.}$ and $\hat{d}_{1,.}$ do not depend on the hypothesis, have essentially the same performance for the "SMALL" or "LARGE" contingency tables. In the "SMALL" table, all three corrected statistics perform reasonably well although $X_P^2(\hat{\delta}_.)$ is slightly more stable across the hypotheses. However in the case of the "LARGE" contingency table, $X_P^2(\hat{\delta}_.)$ performs consistently better than $X_P^2(\hat{\lambda}_{1,}.)$ or $X_P^2(\hat{\lambda}_{1,.})$ (SL ranging from 0.13 to 0.17 compared to 0.13 to 0.25), but not entirely satisfactory due to large c.v. of the $\hat{\delta}_i$'s. It may also be noted that, unlike the empirical results for two-way tables previously reported (Hidiroglou and Rao 1981), $X_P^2(\hat{\lambda}_{1,.})$ or $X_P^2(\hat{d}_{1,.})$ may not be conservative for three-way tables; in fact, as shown in Tables 7 and 8 their SL could be quite high.

The values associated with the Wald statistic are very much larger for the "LARGE" table compared to the "SMALL" table. This phenomenon could be attributed to the instability of the estimated covariance matrix used in the computation of the Wald statistic.

TABLE 5: Female Population 15 Years and Over by Frequency of Breast Self-Examination × age × education (in thousands) "SMALL" 2×3×3

EDUCATION	MONTHLY & QUARTERLY	LESS OFTEN	NEVER
	AGE: 15-24		
Secondary or less ...	462 a 0.05 b 1.35 c	214 0.08 2.02	817 0.04 1.84
Some Post-Secondary and Post-Secondary .	203 0.10 2.43	147 0.12 3.30	162 0.10 3.17
	AGE: 25-44		
Secondary or less ...	975 0.04 1.92	446 0.08 3.79	539 0.05 2.10
Some Post-Secondary and Post-Secondary .	582 0.07 2.78	253 0.07 1.70	200 0.11 2.06
	AGE: 45+		
Secondary or less ...	1169 0.03 1.56	454 0.06 1.58	905 0.05 2.84
Some Post-Secondary and Post-Secondary .	305 0.09 1.95	117 0.12 2.90	101 0.12 1.97

NOTE: "a" stands for estimated population count

"b" stands for the coefficient of variation of the cell

"c" stands for the cell DEFF

TABLE 6: Female Population 15 Years and Over
by Frequency of Breast Self-Examination
× age × education (in thousands)
"LARGE" 3×4×4

EDUCATION	MONTHLY	QUARTERLY	LESS OFTEN	NEVER
	AGE: 15-19			
Secondary or less .	92a 0.14b 2.45c	79 0.10 1.07	108 0.13 2.52	615 0.04 1.51
Some Post-Secondary	11 0.36 2.03	10 0.28 1.14	23 0.29 2.72	59 0.17 2.39
Post-Secondary	2 0.56 1.07	2 0.53 0.95	- 1.30 0.90	5 0.51 1.69
	AGE: 20-24			
Secondary or less .	147 0.12 3.11	144 0.10 1.88	106 0.12 2.06	202 0.09 2.31
Some Post-Secondary	41 0.24 3.38	27 0.17 1.08	54 0.19 2.88	44 0.19 2.10
Post-Secondary	53 0.18 2.40	56 0.12 1.09	70 0.20 4.05	54 0.14 1.49
	AGE: 25-44			
Secondary or less .	486 0.06 2.94	488 0.05 2.18	446 0.08 4.41	539 0.05 1.81
Some Post-Secondary	60 0.20 3.41	64 0.13 1.59	56 0.15 1.85	43 0.18 1.96
Post-Secondary	213 0.11 3.54	244 0.08 2.00	197 0.09 2.05	157 0.13 3.91
	AGE: 45-64			
Secondary or less .	469 0.03 0.77	408 0.06 2.31	312 0.07 2.08	520 0.07 3.37
Some Post-Secondary	26 0.24 2.07	40 0.18 1.85	26 0.20 1.54	13 0.32 1.91
Post-Secondary	72 0.12 1.48	69 0.17 2.71	71 0.15 2.14	38 0.16 1.37

NOTE: "a" stands for estimated population count
"b" stands for the coefficient of variation of the cell
"c" stands for the cell DEFF

TABLE 7: Female Population 15 Years and Over
by Frequency of Breast Self-Examination
by Age and Education
"SMALL" 2×3×3

STATISTIC	INDEPENDENCE HYPOTHESIS						
	COMPLETE	MULTIPLE			CONDITIONAL		
	A*B*C	A*BC	B*AC	C*AB	A*B|C	A*C|B	B*C|A
$\hat{d}_{1,\cdot}$	2.291	2.291	2.291	2.291	2.291	2.291	2.291
$\hat{\delta}_{\cdot}$	2.331	2.136	2.642	2.402	2.422	2.170	2.880
$\hat{\lambda}_{1,\cdot}$	2.294	2.294	2.294	2.294	2.294	2.294	2.294
c.v.$(\hat{d}_{1,i})$	0.299	0.299	0.299	0.299	0.299	0.299	0.299
c.v.$(\hat{\delta}_i)$	0.931	0.750	0.697	0.792	0.566	0.623	0.549
X_P^2	938	625	496	693	212	409	299
$X_P^2(S,0.05)$	343	260	168	253	80	172	96
X_W^2	853	810	153	624	113	50	34
$SL(X_P^2)$	0.621	0.531	0.608	0.602	0.494	0.495	0.591
$SL[X_P^2(\hat{d}_{1,\cdot})]$	0.100	0.062	0.129	0.105	0.085	0.059	0.145
$SL[X_P^2(\hat{\delta}_{\cdot})]$	0.098	0.084	0.079	0.087	0.069	0.074	0.068
$SL[X_P^2(\hat{\lambda}_{1,\cdot})]$	0.105	0.062	0.129	0.104	0.084	0.059	0.144

TABLE 8: Female Population 15 Years and Over
by Frequency of Breast Self-Examination
by Age and Education
"LARGE" 3×4×4

STATISTIC	INDEPENDENCE HYPOTHESIS						
	COMPLETE	MULTIPLE			CONDITIONAL		
	A*B*C	A*BC	B*AC	C*AB	A*B|C	A*C|B	B*C|A
$\hat{d}_{1,\cdot}$	2.158	2.154	2.155	2.159	2.153	2.154	2.153
$\hat{\delta}_{\cdot}$	2.391	2.467	2.133	2.417	2.174	2.549	2.169
$\hat{\lambda}_{1,\cdot}$	2.153	2.149	2.150	2.154	2.149	2.149	2.148
c.v.$(\hat{d}_{1,i})$	0.402	0.402	0.402	0.402	0.402	0.402	0.402
c.v.$(\hat{\delta}_i)$	1.710	1.580	1.408	1.600	1.221	1.455	1.357
X_P^2	2125	958	1124	1789	169	772	873
$X_P^2(S,0.05)$	678	297	424	570	64	235	322
X_W^2	14091	2774	5239	10412	471	792	70160
SL(X_P^2)	0.826	0.797	0.766	0.808	0.721	0.777	0.731
SL$[X_P^2(\hat{d}_{1,\cdot})]$	0.245	0.248	0.142	0.238	0.134	0.249	0.146
SL$[X_P^2(\hat{\delta}_{\cdot})]$	0.174	0.161	0.148	0.163	0.129	0.148	0.142
SL$[X_P^2(\hat{\lambda}_{1,\cdot})]$	0.247	0.249	0.143	0.239	0.135	0.251	0.147

8. SUMMARY

The theory developed for testing hypotheses in a three-way table has been applied to two data sets derived from the Canada Health Survey summary tapes. The main conclusions from this study may be summarized as follows:

(i) The cell design effects (DEFFs) vary considerably across cells and the coefficient of variation of these DEFFs is quite large,

(ii) Postratification adjustment leads to a substantial reduction of cell DEFFs. However, the distortion in significance level (SL) of customary chi-square statistics, X^2, remains severe even after poststratification adjustment,

(iii) The corrected statistic $X_p^2(\hat{\delta}_.)$ performs better than $X_p^2(\hat{\lambda}_{1,.})$ or $X_p^2(\hat{d}_{1,.})$ in controlling SL, especially when the coefficient of variation of the $\hat{\delta}_i$'s is small,

(iv) The behaviour of $X_p^2(\hat{\lambda}_{1,.})$ and $X_p^2(\hat{d}_{1,.})$ is essentially the same,

(v) The Satterthwaite correction $X_p^2(S;\alpha)$ is recommended as the statistics to use to correct for the effect of design,

(vi) The Wald statistic X_W^2 behaves erratically both in the adjusted and unadjusted cases, especially as the dimension of the three-way table increases. This may be due to the instability of the estimated covariance matrix used in the computation of the Wald statistic.

ACKNOWLEDGEMENT

The authors would like to thank Philip Giles for preparing the necessary tapes for the three-way analyses contained in this report, A.V. Winkworth, Director of Business Survey Methods Division and G. Hole, Director of Institutions and Agriculture Survey Methods Division for supporting this reasearch.

BIBLIOGRAPHY

FELLEGI, I.P. (1980), Approximate Tests of Independence and Goodness of Fit Based on Stratified Multistage Samples, Journal of the American Statistical Association, 75, 261-268.

FIENBERG, S.E. (1980), The Analysis of Cross-Classified Categorical Data. Second Edition, MIT Press.

HABERMAN, S.J. (1973), The Analysis of Residuals in Cross-Classified Tables, Biometrics, 29, 205-220.

HABERMAN, S.J. (1978). Analysis of Qualitative Data. Volume 1: Introductory Topics. New York, Academic Press.

HIDIROGLOU, M.A. and RAO, J.N.K. (1981), Chisquare Tests for the Analysis of Categorical Data from the Canada Health Survey, presented at the 43rd ISI Conference in Buenos Aires.

HOLT, D., SCOTT, A.J., and EWINGS, P.O. (1980), Chi-Squared Tests with Survey Data, Journal of the Royal Statistical Society, Sec. A, 143, 302-320.

KOCH, G.G., FREEMAN, D.H. Jr., and FREEMAN, J.L. (1975), Strategies in the Multivariate Analysis of Data From Complex Surveys, International Statistical Review, 43, 59-78.

RAO, J.N.K. and SCOTT, A.J. (1981), The Analysis of Categorical Data from Complex 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. (1982), On Chisquare Tests for Multiway Contingency Tables with cell Proportions Estimated from Survey Data, in preparation.

SATTERTHWAITE, F.E. (1946), An Approximate Distribution of Estimates of Variance Components, Biometrics, 2, 110-114.

The Health of Canadians, (1981), Report of the Canada Health Survey, Health and Welfare and Statistics Canada. Statistics Canada Catalogue 82-538E.