

ANALYSIS OF CATEGORICAL DATA FROM COMPLEX SAMPLE SURVEYS USING INVERSE SAMPLING

E. Benhin¹ and J. N. K. Rao²

ABSTRACT

Analysis of complex survey categorical data using classical statistical methods without taking into account the complex nature of the data may lead to asymptotically invalid statistical inferences. Several methods have been developed that account for the survey design, but these methods require additional information such as survey weights, design effects or cluster identification for microdata. An alternate approach is to undo the complex data structures using repeated inverse sampling so that standard methods can be applied to the generated inverse-sampling data files. We propose a combined estimating equation approach to analyze such data files in the context of categorical survey data. For simplicity, we focus on goodness-of-fit test statistics under cluster sampling.

KEY WORDS: Categorical survey data; Combined estimating equation; Confidentiality; Repeated subsampling.

RÉSUMÉ

L'analyse des données catégoriques provenant d'enquêtes complexes en utilisant des méthodes statistiques classiques sans tenir compte de la nature complexe des données peut mener à des inférences statistiques asymptotiquement invalides. Plusieurs méthodes qui tiennent compte du plan du sondage ont été développées, mais celles-ci exigent de l'information additionnelle telle que les poids d'échantillonnage, les effets de plan ou l'identification de grappe pour l'ensemble des micro-données. Une approche alternative est de défaire les structures complexes des données de sorte que les méthodes standard puissent être appliquées aux fichiers de données générés par échantillonnage inverse. Nous proposons une approche d'équations d'estimation combinée pour analyser de tels fichiers de données dans le contexte des données d'enquête catégoriques. Pour des raisons de simplicité, nous concentrons sur le test d'ajustement statistique sous un échantillonnage par grappes..

MOTS CLÉS: Confidentialité; données d'enquête catégoriques; équations d'estimation combinées; sous-échantillonnage répété.

1. INTRODUCTION

Standard methods for analyzing categorical data were developed under the assumption of multinomial sampling. The use of these methods to analyze complex survey categorical data without taking into account the complex survey features may lead to asymptotically invalid statistical inferences.

In recent years many methods have been developed to analyze complex survey categorical data that take account of the design features. These methods, however, require additional information such as cluster identification or stratification, design effects or the estimated covariance matrix of ultimate cell estimates for microdata. This additional information may not be available for reasons of confidentiality.

Instead of developing complex new methods to fit the complex survey data, we propose inverse sampling methods to tailor the data to fit standard methods. Hinkins, Oh and Scheuren (1997) proposed the basic idea of inverse sampling. Rao, Scott and Benhin (2003) developed basic theory of the inverse sampling and proposed an estimating equations approach to handle complex parameters such as ratios and census regression parameters. Benhin's unpublished 2004 Ph.D. thesis studied a number of applications of the inverse sampling estimating equations method to poststratification estimation, quasi score tests and analyses of complex survey categorical data.

¹ E. Benhin (emmanuel.benhin@statcan.ca), Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6,

² J. N. K. Rao (jrao@math.carleton.ca), School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada, K1S 5B6,

The basic idea of inverse sampling is to select a subsample that has a simple random structure unconditionally or has a structure that is considerably simpler to handle than the original sample. This may lead to loss in efficiency since in general, the size of the subsample may be considerably smaller than the size of the original sample. To increase efficiency, the process is repeated independently many times and the resulting data combined for making inferences.

Section 2 summarizes the basic algorithm of inverse sampling. Some properties of the combined estimating equation method are discussed in Section 3. Design-based goodness-of-fit test statistics are discussed in Section 5. We propose inverse-sampling goodness-of-fit test statistics in Section 5. In Section 6, we present simulation results, to illustrate the performance of the proposed methods.

2. BASIC ALGORITHM OF INVERSE SAMPLING

Suppose s_0 is a sample of observations drawn from a finite population of size N according to a specified complex design. Let s^* be the subsample of size m drawn from s_0 such that the unconditional probability of s^* , $p(s^*)$, matches simple random sampling either exactly or approximately. Then

$$p(s^*) = \sum_{s_0 \supset s^*} p_0(s_0) p(s^* | s_0), \quad (2.1)$$

where $p_0(s_0)$ is the probability of selecting s_0 and $p(s^* | s_0)$ is the conditional probability of choosing s^* . Provided $p(s^* | s_0)$ does not depend on s_0 , then it follows from (2.1) that

$$p_2(s^*) = \frac{p(s^*)}{\sum_{s_0 \supset s^*} p_0(s_0)}. \quad (2.2)$$

For a given complex survey design, exact or approximate inverse sampling algorithm follows from (2.2). A number of inverse-sampling algorithms for some standard designs are discussed in Hinkins, Oh and Scheuren (1997), and these are also summarized in Rao, Scott and Benhin (2003) for ready reference.

2.1 One-stage Cluster Sampling

The three one-stage cluster sampling designs investigated by Hinkins et al. (1997) are: (1) Cluster sizes equal, M , and clusters sampled with equal design probability; (2) Cluster sizes unequal, M_i , and clusters sampled with equal probability; (3) Cluster sizes unequal, M_i , and clusters sampled with probability proportional to size M_i and with replacement.

Case 1. In the case of equal cluster sizes, M , and clusters sampled with equal probability, implementing exact simple random sampling (SRS) is difficult. Suppose s_0 contains k clusters drawn from K clusters in the population of size $N = KM$. A simple approximate method of subsampling selects one element at random from each sample cluster so that the size of s^* is k .

Case 2. For unequal cluster sizes, M_i , and clusters sampled with equal probability, it is not possible to obtain fixed size SRS either exactly or approximately. Hinkins et al. (1997) proposed an approach that artificially enlarges the population to equal cluster size case and then applies subsampling used in Case 1. We first force all clusters to have the same size by adding an appropriate number of pseudo-units to bring them up to the size of the largest sample cluster. Then select one unit at random from each sample cluster discarding any pseudo-units to obtain the final sample. This approximate method makes $p(s^* | s_0)$ depend on s_0 because the conditional probability depends on $M(s_0)$, the size of the largest sample cluster.

Case 3. For the case of probability proportional to size (PPS) sampling with replacement of unequal size clusters, Hinkins et al. (1997) proposed a simple method of subsampling which gives $p(s^*) = (1/N)^k$, where s^* now denotes an ordered simple random sample with replacement selected from the $N = \sum_{i=1}^K M_i$ units in the population. Viewing the sample clusters as ordered, we select one unit at random from each sample cluster. Note that the same cluster might appear more than once in the ordered sample. Denote the size of the cluster drawn in the i -th PPS draw by M'_i , then

$$p(s^*) = \left[\prod_{i=1}^k \frac{M'_i}{N} \right] \left[\prod_{i=1}^k \frac{1}{M'_i} \right] = \left(\frac{1}{N} \right)^k \quad (2.3)$$

where $\prod_{i=1}^k (M'_i / N)$ is the probability of drawing the ordered cluster sample. Note that s_0 is the ordered PPS sample and we have only one term in the summation in (2.1).

If the clusters are drawn with inclusion probabilities $\pi_i = kM_i / N$ and without replacement, then it is not possible to match SRS. However, we can treat the clusters as if they were drawn with replacement, as done in practice, and then apply the scheme for Case 3. This will lead to overestimation of variance, but the overestimation is not serious if the sampling fraction k / K is small (see Section 4.3 of Rao, Scott and Benhin, 2003)

2.2 Two-stage Cluster Sampling

The two two-stage cluster sampling designs investigated by Hinkins, Oh and Scheuren (1997) are: (1) Cluster sizes equal, M , and k clusters sampled with equal probability in the first stage and then simple random subsamples of equal size, m , drawn independently within each sampled cluster (PSU). (2) Unequal cluster sizes, M_i , and k clusters sampled with PPS and with replacement and then simple random subsamples of sizes, m_i , drawn independently within each cluster in the with replacement sample.

Case 1. As in the case of one-stage cluster sampling, an exact method of inverse sampling is difficult to implement. A simple approximate method of inverse sampling selects one unit at random from each of the k subsamples.

Case 2. As in Case 3 of one-stage cluster sampling, select one unit at random from each of the ordered subsamples. Hinkins, Oh and Scheuren (1997) suggested a different method: First take a simple random sample with replacement of k clusters and then with each selected cluster take one unit at random from the corresponding subsample. It appears that the first stage inverse sampling of clusters is not necessary. To see this, we note that

$$p_0(s_0) = \prod_{i=1}^k \left[\left(\frac{M'_i}{N} \right) \frac{1}{\binom{M'_i}{m'_i}} \right],$$

where m'_i is the subsample size associated with the cluster selected in the i -th draw ($i = 1, \dots, k$). We wish to draw a subsample s^* of size k such that $p(s^*) = (1/N)^k$, where $N = \sum_{i=1}^K M_i$. Also the number of terms in $\sum_{s_0 \supset s^*}$ equals

$$\prod_{i=1}^k \binom{M'_i - 1}{m'_i - 1} \text{ and}$$

$$\sum_{s_0 \supset s^*} p_0(s_0) = \prod_{i=1}^k \left[\frac{\binom{M_i'}{N}}{\binom{M_i'}{m_i'}} \frac{\binom{M_i' - 1}{m_i' - 1}}{\binom{M_i'}{m_i'}} \right] = \prod_{i=1}^k \frac{m_i'}{N}.$$

It follows from (2.2) that $p(s^* | s_0) = \prod_{i=1}^k (1/m_i')$ and hence the subsampling scheme readily follows.

3. SOME PROPERTIES OF THE COMBINED ESTIMATING EQUATION

3.1 Full-sample estimating equations

Consider the finite population parameter vector θ_N as the solution to the census estimating equations:

$$U(\theta) = \sum_{k \in U} \mathbf{u}_k(\theta) = \mathbf{0}, \quad (3.1)$$

where $\sum_{k \in U}$ denotes the summation over the finite population U of size N , and the estimating functions $\mathbf{u}_k(\theta)$ are suitably chosen (Binder, 1983, Godambe and Thompson, 1986). For example, consider the scalar case of (3.1) and let $u_k(\theta) = y_k - \theta$. This gives the population mean $\theta_N = \bar{Y}$.

The full-sample estimating equations are given by

$$\hat{U}(\theta) = \sum_{k \in s_0} w_k \mathbf{u}_k(\theta) = \mathbf{0}, \quad (3.2)$$

where w_k is the survey weight attached to the unit $k \in s_0$; if the Horvitz-Thompson estimator of $U(\theta)$ is used, $w_k = \pi_k^{-1}$ and π_k is the inclusion probability of the unit k . The solution to (3.2) gives the design-based estimator $\hat{\theta}$, which in general is nonlinear and biased. We assume that the sample size of s_0 is large enough to neglect the bias of $\hat{\theta}$.

Under regularity conditions, Binder (1983) obtained a Taylor linearization variance estimator of $\hat{\theta}$ as

$$\hat{V}_L(\hat{\theta}) = \{\hat{J}(\hat{\theta})\}^{-1} \hat{V}\{\hat{U}(\hat{\theta})\} \{\hat{J}(\hat{\theta})\}^{-1}, \quad (3.3)$$

where $\hat{V}\{\hat{U}(\hat{\theta})\}$ is the variance estimator of the estimated vector of totals, $\hat{U}(\theta)$, of the $\mathbf{u}_k(\theta)$'s evaluated at $\theta = \hat{\theta}$ and $\hat{J}(\hat{\theta})$ is the observed information matrix obtained by evaluating $\hat{J}(\theta) = -\partial \hat{U}(\theta) / \partial \theta^T$ at $\theta = \hat{\theta}$.

3.2 Combined estimating equation

Suppose s_0 is a sample of n observations drawn from a finite population according to some complex design. We assume that we have an inverse sampling algorithm (exact or approximate) that can generate a sequence of g independent subsamples s_j^* ($j = 1, \dots, g$) each of size m , where the subsamples are of simple random sampling design. Then the combined estimating equation for estimating the finite population parameter θ_N is given by

$$\hat{U}_{gc}(\theta) = \frac{1}{g} \sum_{j=1}^g \hat{U}_j^*(\theta) = \frac{1}{g} \sum_{j=1}^g \frac{N}{m} \sum_{k \in s_j^*} u_k(\theta) = \mathbf{0}. \quad (3.4)$$

The solution to (3.4), $\hat{\theta}_{gc}$ is the combined estimating equation estimator of θ_N . The linearization variance estimator of $\hat{\theta}_{gc}$ is given by

$$\hat{V}_L(\hat{\theta}_{gc}) = \left\{ \hat{J}_{gc}(\hat{\theta}_{gc}) \right\}^{-1} \hat{V} \left\{ \hat{U}_{gc}(\hat{\theta}_{gc}) \right\} \left\{ \hat{J}_{gc}(\hat{\theta}_{gc}) \right\}^{-1}, \quad (3.5)$$

where $\hat{J}_{gc}(\hat{\theta}_{gc})$ is $\hat{J}_{gc}(\theta) = -\partial \hat{U}_{gc}(\theta) / \partial \theta^T$ evaluated at $\theta = \hat{\theta}_{gc}$,

$$\begin{aligned} \hat{V} \left\{ \hat{U}_{gc}(\hat{\theta}_{gc}) \right\} &= \frac{1}{g} \sum_{j=1}^g \hat{V}_{jS}^* - \frac{1}{g} \sum_{j=1}^g \hat{U}_j^*(\hat{\theta}_{gc}) \hat{U}_j^*(\hat{\theta}_{gc})^T, \\ \hat{V}_{jS}^* &= \frac{N^2}{m} \left(1 - \frac{m}{N} \right) \frac{1}{m-1} \sum_{k \in s_j^*} \left\{ u_k(\hat{\theta}_{gc}) - \frac{1}{m} \sum_{k \in s_j^*} u_k(\hat{\theta}_{gc}) \right\} \left\{ u_k(\hat{\theta}_{gc}) - \frac{1}{m} \sum_{k \in s_j^*} u_k(\hat{\theta}_{gc}) \right\}^T. \end{aligned}$$

Benhin (2004) showed that under regularity conditions, $\hat{\theta}_{gc}$ and $\hat{V}_L(\hat{\theta}_{gc})$ are consistent, that is, $\hat{\theta}_{gc} \rightarrow_p \hat{\theta}$ and $\hat{V}_L(\hat{\theta}_{gc}) \rightarrow_p \hat{V}_L(\hat{\theta})$ as $g \rightarrow \infty$. Therefore, the combined estimating equation estimator, $\hat{\theta}_{gc}$, leads to asymptotically valid inferences regardless of the size of the subsample m . An application of the general results of the combined estimating equation method is the goodness-of-fit problem for complex survey data discussed in Section 4.

4. FULL-SAMPLE GOODNESS-OF-FIT TESTS

Consider a finite population $U = \{1, \dots, l, \dots, N\}$ of size N . Suppose U is partitioned into K non-overlapping and exhaustive categories U_i ($i = 1, \dots, K$). The size of U_i is denoted by N_i . Let p_i denote the finite population proportion for the i -th category such that $p_i \geq 0$ and $\sum_{i=1}^K p_i = 1$. Consider the goodness-of-fit problem for testing $H_0 : p_i = p_{0i}$ for specified p_{0i} . For example, in a household survey one may be interested in testing the hypothesis that the population distribution of a characteristic associated with household members is the same as found in the previous census.

A full-sample estimator of p_i based on the original sample s_0 is given by

$$\hat{p}_i = \frac{\sum_{l \in s_0} y_{il} / \pi_l}{\sum_{l \in s_0} 1 / \pi_l}, \quad i = 1, \dots, K \quad (4.1)$$

where π_l is the inclusion probability of the l -th unit, $y_{il} = 1$ if $l \in U_i$; 0 otherwise and $\sum_{i=1}^K \hat{p}_i = 1$ since $\sum_{i=1}^K y_{il} = 1$.

The generalized Wald statistic for testing H_0 is given by

$$X_w^2 = n(\hat{p} - p_0)^T \hat{V}^{-1}(\hat{p} - p_0), \quad (4.2)$$

where $\hat{\mathbf{p}} = (\hat{\mathbf{p}}_1, \dots, \hat{\mathbf{p}}_{K-1})^T$, $\mathbf{p} = (\mathbf{p}_{01}, \dots, \mathbf{p}_{0,K-1})^T$ and $\hat{\mathbf{V}}/n$ is an estimator of the covariance matrix \mathbf{V}/n , of $\hat{\mathbf{p}}$. Under regularity conditions, the statistic X_W^2 is asymptotically distributed as a χ_{K-1}^2 under H_0 as $n \rightarrow \infty$.

A Pearson chi-squared statistic for testing H_0 is given by

$$X^2 = n \sum_{i=1}^K \frac{(\hat{\mathbf{p}}_i - \mathbf{p}_{0i})^2}{\mathbf{p}_{0i}} = n(\hat{\mathbf{p}} - \mathbf{p}_0)^T \mathbf{P}_0^{-1}(\hat{\mathbf{p}} - \mathbf{p}_0), \quad (4.3)$$

and $\mathbf{P}_0 = \text{diag}(\mathbf{p}_0) - \mathbf{p}_0 \mathbf{p}_0^T$. Rao and Scott (1981) showed that under the null hypothesis, H_0 , the statistic X^2 is asymptotically distributed as $\sum_{i=1}^{K-1} \lambda_{0i} Z_i^2$, where Z_1, \dots, Z_{K-1} are independent $N(0, 1)$ random variables and λ_{0i} 's are eigenvalues of $\mathbf{A}_0 = \mathbf{P}_0^{-1} \mathbf{V}$ ($\lambda_{01} \geq \lambda_{02} \geq \dots \geq \lambda_{0,K-1}$). Rao and Scott (1981) provided simple corrections to X^2 ; (i) first-order correction; (ii) second-order correction. The first-order corrected statistic treats

$$X_c^2 = \frac{X^2}{\hat{\lambda}} \cong \frac{\sum_{i=1}^{K-1} \lambda_i Z_i^2}{\hat{\lambda}} =: Y \quad (4.4)$$

as a χ_{K-1}^2 random variable under H_0 , where $\hat{\lambda} = \sum_{i=1}^{K-1} \hat{\lambda}_i / (K-1)$ and $\hat{\lambda}_i$'s are eigenvalues of $\hat{\mathbf{A}} = \hat{\mathbf{P}}^{-1} \hat{\mathbf{V}}$ and $\hat{\mathbf{P}} = \text{diag}(\hat{\mathbf{p}}) - \hat{\mathbf{p}} \hat{\mathbf{p}}^T$. We note that $\hat{\lambda} = (K-1)^{-1} \sum_{i=1}^{K-1} (1 - \hat{p}_i) \hat{d}_i$, where $\hat{d}_i = \hat{V}_{ii} / \{\hat{p}_i(1 - \hat{p}_i)\}$ is the estimated design effect of the i -th cell and \hat{V}_{ii} is the i -th diagonal element of $\hat{\mathbf{V}}$. Thus only the estimated cell design effects are needed to implement the first-order corrected statistic. Rao and Scott (1981) also showed that treating X_c^2 as χ_{K-1}^2 under H_0 tends to give significance levels that are somewhat larger than the nominal test levels. A better approximation is the second-order correction statistic which depends on the availability of additional information on $\hat{\lambda}_i$'s. It treats

$$X_s^2 = \frac{X_c^2}{1 + \hat{a}^2} \quad (4.5)$$

as χ_ν^2 , where $\nu = (K-1)/(1 + \hat{a}^2)$ and $\hat{a}^2 = \sum_{i=1}^{K-1} (\hat{\lambda}_i - \hat{\lambda})^2 / \{(K-1)\hat{\lambda}^2\}$ is the square of the coefficient of variation of the $\hat{\lambda}_i$'s. Note that $\sum_{i=1}^{K-1} \hat{\lambda}_i^2 = \sum \sum_{i,j=1}^K \hat{V}_{ij}^2 / (\hat{p}_i \hat{p}_j)$ so that the second-order corrected statistic can be implemented if $\hat{\mathbf{V}}$ is known, the same information required for the Wald statistic, X_W^2 . Thomas and Rao (1987) showed through finite sample studies that X_s^2 performs better than X_W^2 in terms of significance levels.

It is important to note that the full-sample methods require additional design information such as cluster identifiers, stratification etc., for implementation. For reasons of confidentiality, this information may not be available on the public use microdata files.

5. INVERSE-SAMPLING GOODNESS-OF-FIT TESTS

The inverse-sampling goodness-of-fit test statistics discussed in this paper focus primarily on cluster sampling. The results, however, may be extended to other complex survey designs. Suppose s_0 is a sample from a two-stage cluster sampling in which a K -category sample of r units is drawn independently from each of m sample clusters. Suppose an inverse sample s^* of size m elements is drawn from s_0 such that the unconditional probability of s^* , $p(s^*)$, matches

simple random sampling either exactly or approximately. See Section 2 for a summary of exact or approximate inverse sampling algorithms for a number of one-stage and two-stage cluster sampling designs.

Suppose a sequence of g independent inverse samples s_j^* ($j = 1, \dots, g$), each of size m , is generated independently from s_0 . Then the inverse-sampling estimator, $\hat{\boldsymbol{p}}_g$ of $\boldsymbol{p} = (\boldsymbol{p}_1, \dots, \boldsymbol{p}_{K-1})^T$ is given by

$$\hat{\boldsymbol{p}}_g = \frac{1}{g} \sum_{j=1}^g \hat{\boldsymbol{p}}_j^* = \frac{1}{gm} \sum_{j=1}^g \sum_{l \in s_j^*} \boldsymbol{y}_l, \quad (5.1)$$

where $\boldsymbol{y}_l = (y_{1l}, \dots, y_{K-1,l})^T$, $y_{il} = 1$ if $l \in U_i$; 0 otherwise, $\hat{\boldsymbol{p}}_j^* = (\hat{p}_{1j}^*, \dots, \hat{p}_{ij}^*, \dots, \hat{p}_{K-1,j}^*)^T$, $j = 1, \dots, g$, $\hat{p}_{ij}^* = m^{-1} \sum_{l \in s_j^*} y_{il}$ and $\sum_{i=1}^K \hat{p}_{ij}^* = 1$ since for each subsample, s_j^* , $\sum_{i=1}^K y_{il} = 1$. The inverse-sampling covariance matrix of $\hat{\boldsymbol{p}}_g$ is given by

$$\hat{\boldsymbol{V}}_g = \frac{1}{g} \sum_{j=1}^g \hat{\boldsymbol{V}}_j^* - \frac{1}{g} \sum_{j=1}^g (\hat{\boldsymbol{p}}_j^* - \hat{\boldsymbol{p}}_g)(\hat{\boldsymbol{p}}_j^* - \hat{\boldsymbol{p}}_g)^T \quad (5.2)$$

and $\hat{\boldsymbol{V}}_j^* = \hat{\boldsymbol{V}}(\hat{\boldsymbol{p}}_j^*) = m^{-1} \{diag(\hat{\boldsymbol{p}}_j^*) - \hat{\boldsymbol{p}}_j^* \hat{\boldsymbol{p}}_j^{*T}\}$.

The inverse-sampling generalized Wald statistic for testing H_0 is given by

$$X_{g,W}^2 = m(\hat{\boldsymbol{p}}_g - \boldsymbol{p}_0)^T \hat{\boldsymbol{V}}_g^{-1} (\hat{\boldsymbol{p}}_g - \boldsymbol{p}_0), \quad (5.3)$$

where $\hat{\boldsymbol{p}}_g = (\hat{\boldsymbol{p}}_{g1}, \dots, \hat{\boldsymbol{p}}_{g,K-1})^T$ and $\hat{p}_{gi} = g^{-1} \sum_{j=1}^g \hat{p}_{ij}^*$, $i = 1, \dots, K$. Benhin (2004) showed that under regularity conditions, the inverse-sampling generalized Wald statistic, $X_{g,W}^2$ converges to a χ_{K-1}^2 as $g \rightarrow \infty$ and then $m \rightarrow \infty$.

The inverse-sampling Pearson chi-squared statistic for testing H_0 is given by

$$X_g^2 = m(\hat{\boldsymbol{p}}_g - \boldsymbol{p}_0)^T \boldsymbol{P}_0^{-1} (\hat{\boldsymbol{p}}_g - \boldsymbol{p}_0). \quad (5.4)$$

Benhin (2004) showed that under the null hypothesis, H_0 , the inverse-sampling Pearson chi-squared statistic, X_g^2 is asymptotically distributed as $\sum_{i=1}^{K-1} \lambda_{0i} Z_i^2$, where Z_i 's are independent $N(0, 1)$ random variables and λ_{0i} 's are the eigenvalues of $\boldsymbol{P}_0^{-1} \boldsymbol{V}$ ($\lambda_{01} \geq \lambda_{02} \geq \dots \geq \lambda_{0,K-1}$). Since in practice, λ_{0i} 's are unknown, we propose simple correction statistics to the asymptotic distribution of X_g^2 ; (i) inverse-sampling first-order corrected statistic; (ii) inverse-sampling second-order corrected statistic. The inverse-sampling first-order statistic is given by

$$X_{g,c}^2 = \frac{X_g^2}{\hat{\lambda}_g}, \quad (5.5)$$

where $\hat{\lambda}_g = (K-1)^{-1} \sum_{i=1}^{K-1} \hat{\lambda}_{gi}$, $\hat{\lambda}_{gi}$'s are the eigenvalues of $\hat{\boldsymbol{P}}_g^{-1} \hat{\boldsymbol{V}}_g$, $\hat{\boldsymbol{P}}_g = diag(\hat{\boldsymbol{p}}_g) - \hat{\boldsymbol{p}}_g \hat{\boldsymbol{p}}_g^T$ and $\hat{\boldsymbol{V}}_g$ is given by (5.2). Treating $X_{g,c}^2$ as χ_{K-1}^2 under H_0 tends to give significance levels that are somewhat larger than the nominal test levels. A better approximation is the inverse-sampling second-order corrected statistic:

$$X_{g,S}^2 = \frac{X_{g,c}^2}{1 + \hat{a}_g^2} \quad (5.6)$$

as $\chi_{\nu_g}^2$, where $\nu_g = (K-1)/(1 + \hat{a}_g^2)$ and $\hat{a}_g^2 = \sum_{i=1}^{K-1} (\hat{\lambda}_{gi} - \hat{\lambda}_g)^2 / [(K-1)\hat{\lambda}_g^2]$. We note that $\hat{\lambda}_g = (K-1)^{-1} \sum_{i=1}^K (\hat{V}_{gii} / \hat{p}_{gi})$ and $\sum_{i=1}^K \hat{\lambda}_{gi}^2 = \sum_{i=1}^K \sum_{j=1}^K \{\hat{V}_{gij}^2 / (\hat{p}_{gi}\hat{p}_{gj})\}$, where \hat{V}_{gij} is the (i,j)-th element of \hat{V}_g . We note that for the full-sample approach, the application of the second-order corrected statistic (4.5) proposed by Rao and Scott (1981) is dependent on the knowledge of $\sum_{i=1}^K \hat{\lambda}_i^2$ or \hat{V} . However, for reasons of confidentiality, public use data files may not contain design information such as cluster identifiers needed to calculate \hat{V} . On the other hand, the inverse-sampling approach has no such restrictions. In fact the inverse-sampling covariance matrix, \hat{V}_g given by (5.2) can be calculated from data files consisting of multiple inverse samples.

6. SIMULATION STUDY

The simulation study in this section compares the performance of the design-based goodness-of-fit test statistics and the inverse-sampling test statistics for testing the null hypothesis $H_0 : p = p_0$. For this study, we considered two-stage cluster sampling in which a K -category sample of r units is drawn independently from each of m sample clusters. We let $\mathbf{r}_l = (r_{l1}, \dots, r_{l,K-1})^T$ represent the vector of counts in the first $K-1$ categories for the l -th cluster ($l = 1, \dots, m$) and $\mathbf{r} = (\mathbf{r}_1, \dots, \mathbf{r}_{K-1})^T$ represent the corresponding category counts for the whole sample, that is $r_i = \sum_{l=1}^m r_{li}$ ($i = 1, \dots, K-1$). The total number of observations in the sample is thus $n = mr = \sum_{i=1}^K r_i$, where $r_K = n - \sum_{i=1}^{K-1} r_i$. Furthermore, let $\hat{\mathbf{p}} = (\hat{p}_1, \dots, \hat{p}_{K-1})^T = \mathbf{r} / n$ be the vector of cell proportions in the sample and define $\mathbf{p} = E(\hat{\mathbf{p}})$, where E denotes the expectation under a specified cluster sampling model. Let \mathbf{V} / n represent the $(K-1) \times (K-1)$ covariance of $\hat{\mathbf{p}}$ under the model.

The following parameters were controlled: (i) α , the nominal significance level for the tests; (ii) p , the model probability vector; (iii) K , the number of categories; (iv) m , the number of sample clusters; (v) r , the number of units drawn from each sample cluster; (vi) $\bar{\lambda}$, the mean of the eigenvalues λ_i ; (vii) a , the coefficient of variation of the λ_i 's. The degree of clustering is represented by $(\bar{\lambda}, a)$. For the simulation study, we considered $(\bar{\lambda} = 2, a = 0)$ which represents a constant design effect clustering. The nominal significance level used was $\alpha = 0.05$, and the model probability vector under H_0 was $\mathbf{p}_0 = (1/K, 1/K, 1/K)^T$.

Brier (1978)'s method of generating Dirichlet variates from $(K-1)$ beta random variables was used. For each of the 1000 Monte Carlo trials, independent Dirichlet vectors, \mathbf{p}_l , were generated for $m = (30, 50)$ clusters. For each cluster, a K -category conditional multinomial sample was constructed by referring each r independent $(0, 1)$ uniform variates to the appropriate interval associated with \mathbf{p}_l .

For the inverse-sampling methods we generated $g = 500, 1000, 2000$ inverse samples each of size m . The inverse samples are generated using the approximate inverse sampling algorithm for two-stage cluster sampling described in Case 1 of Section 2.2, that is, select one unit at random from each of the m full-sample clusters resulting in an inverse sample of size m . From the 1000 Monte Carlo data, we calculated the actual significance levels for the design-based methods and the inverse-sampling methods. Table 1 reports the performance of the full-sample and inverse-sampling Wald statistics with respect to the actual level of significance. The results in Table 1 show that the inverse-sampling Wald test

significance levels track the design-based Wald test significance levels even though the full-sample Wald statistic in this case is poor in tracking the nominal significance level, 5%.

Table 2 reports the performance of the full-sample and inverse-sampling second-order corrected statistics. The results in Table 2 show that both the second-order corrected inverse-sampling statistic, $X_{g,S}^2$, and the full-sample second-order corrected statistic, X_S^2 , give significance levels close to the nominal level 5%. The inverse-sampling significance levels are close to the nominal even for $g = 500$.

Table 1 – Actual significance levels (%) of the design-based and inverse-sampling Wald statistics, X_W^2 and $X_{g,W}^2$ respectively, $\mathbf{p} = (1/K, 1/K, 1/K)^T$, number of categories, $K = 3$, mean of eigenvalues, $\bar{\lambda} = 2$, the coefficient of variation of the λ_i 's, $a = 0$ and the nominal significance level, $\alpha = 5\%$

	$SL(X_W^2)$	$SL(X_{g,W}^2)$		
Sample clusters		$g = 500$	$g = 1000$	$g = 2000$
50	7.3	6.7	7.7	7.1
30	8.1	7.9	7.8	8.3

$SL(\cdot)$; Significance level, g ; number of subsamples

Table 2 – Actual significance levels (%) of the design-based and inverse-sampling second-order corrected statistics, X_S^2 and $X_{g,S}^2$ respectively, $\mathbf{p} = (1/K, 1/K, 1/K)^T$, number of categories, $K = 3$, mean of eigenvalues, $\bar{\lambda} = 2$, the coefficient of variation of the λ_i 's, $a = 0$ and the nominal significance level, $\alpha = 5\%$

	$SL(X_S^2)$	$SL(X_{g,S}^2)$		
Sample clusters		$g = 500$	$g = 1000$	$g = 2000$
50	4.7	4.6	4.4	4.8
30	4.7	4.9	4.2	5.3

$SL(\cdot)$; Significance level, g ; number of subsamples

It is important to note that the inverse-sampling methods can be implemented from a microdata file consisting of g independent subsamples each of size m . Neither the survey weights nor the cluster identifiers are needed so that confidentiality of the microdata may be preserved.

REFERENCES

- Benhin, E. (2004). *Contribution to the analysis of complex survey data and cluster-correlated biological data using inverse sampling*. Ph.D. Thesis. Carleton University.
- Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, **51**, 279-292.
- Godambe, V. P. and Thompson, M. E. (1986). Parameters of super population and survey population: their relationship and estimation. *International Statistical Review*, **54**, 127-138.
- Hinkins, S., Oh, H. L. and Scheuren, F. (1997). Inverse sampling design algorithms. *Survey Methodology*, **23**, 11-21.

- 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 on two-way tables. *Journal of American Statistical Association*, **76**, 221-230.
- Rao, J. N. K., Scott, A. J. and Benhin, E. (2003). Undoing Complex Survey Data Structures: Some Theory and Applications of Inverse Sampling. *Survey Methodology*, **29**, 107-128.
- Thomas, R. D. and Rao, J. N. K. (1987). Small-sample comparisons of level and power for simple goodness-of-fit statistics under cluster sampling. *Journal of American Statistical Association*, **82**, 630-636.