Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Chi-squared tests in dual frame surveys

by Yan Lu

Release date: December 19, 2014

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at infostats@statcan.gc.ca,

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-877-287-4369

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

To access this product

This product, Catalogue no. 12-001-X, is available free in electronic format. To obtain a single issue, visit our website, www.statcan.gc.ca, and browse by “Key resource” > “Publications.”

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under “About us” > “The agency” > “Providing services to Canadians.”

Published by authority of the Minister responsible for Statistics Canada

Cette publication est aussi disponible en français.

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard symbols

The following symbols are used in Statistics Canada publications:

.	not available for any reference period
..	not available for a specific reference period
...	not applicable
0	true zero or a value rounded to zero
0^s	value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p}	preliminary
^{r}	revised
x	suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E}	use with caution
F	too unreliable to be published
*	significantly different from reference category ($p < 0.05$)

Survey Methodology, December 2014
Vol. 40, No. 2, pp. 323-334
Statistics Canada, Catalogue No. 12-001-X

Chi-squared tests in dual frame surveys

Yan Lu[1]

Abstract

In order to obtain better coverage of the population of interest and cost less, a number of surveys employ dual frame structure, in which independent samples are taken from two overlapping sampling frames. This research considers chi-squared tests in dual frame surveys when categorical data is encountered. We extend generalized Wald's test (Wald 1943), Rao-Scott first-order and second-order corrected tests (Rao and Scott 1981) from a single survey to a dual frame survey and derive the asymptotic distributions. Simulation studies show that both Rao-Scott type corrected tests work well and thus are recommended for use in dual frame surveys. An example is given to illustrate the usage of the developed tests.

Key Words: Asymptotic properties; Chi-squared tests; Dual frame surveys; First-order corrected test; Second-order corrected test; Simulations.

1 Introduction

A general situation of a dual frame survey is depicted in Figure 1.1, where the union of frame A and frame B is denoted as the union of the three nonoverlapping domains, i.e., $A \cup B = a \cup ab \cup b$. Probability samples are selected independently from these two frames.

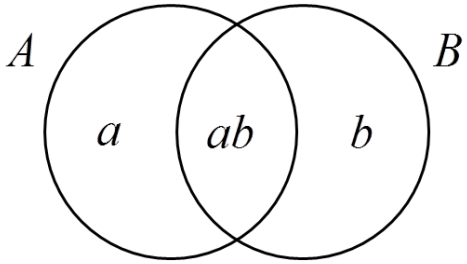

Figure 1.1: Frames A and B are both incomplete but overlapping

A dual frame survey often gives better coverage of the population, and can achieve considerable cost savings. The statistical literature has several methods for cross-sectional analyses of dual-frame survey data, see Hartley (1962, 1974), Fuller and Burmeister (1972), Skinner (1991), Skinner and Rao (1996), Lohr and Rao (2000, 2006), etc. As Rao and Thomas (1988) noted, the need to perform statistical analyses of categorical data is frequently encountered in quantitative sociological research. Pearson's chi-squared test and likelihood ratio test are both well known tests for categorical data. These methods rely on the assumption that data are obtained by simple random sampling (SRS) from one or more large population. Most current surveys have complex designs with stratification and clustering, where the SRS assumption is violated. Wald's test (Wald 1943) is one of the earliest methods proposed to assess model fit in complex designs. Fay (1979, 1985) proposed a jackknifed chi-squared test for use in complex surveys. Both Wald's (1943) and Fay's (1979) procedures require detailed survey information from which the covariance matrix can be estimated. Such detailed information is often not available in practice. Rao and Scott (1981, 1984)

1. Yan Lu, Department of Mathematics and Statistics, University of New Mexico, Albuquerque, NM, 87131-0001. E-mail: luyan@math.unm.edu.

proposed chi-squared tests for goodness of fit and independence in two-way and multi-way tables. Bedrick (1983) and Rao and Scott (1987) also studied the use of limited information on cell and marginal design effects to provide approximate tests. Thomas, Singh and Roberts (1996) described a Monte Carlo study of developed procedures for testing independence in a two-way table.

The research problem in this article arises from categorical data analysis in dual frame surveys. For example, a dual frame may consist of the online membership directories of the American Statistical Association (ASA) and the Institute for Mathematical Statistics (IMS). The overlap domain consists of the statisticians who are members from both societies. One may be interested in testing if the percentage of female in academia is the same across the three domains (domain a: members of ASA only; domain ab: members from both ASA and IMS; domain b: members of IMS only). The tests in a dual frame survey present additional challenges to those from a single frame survey because there are now two samples, each with a possibly complex sampling design and may have an unknown degree of overlap. It is possible to apply a fixed weighting constant for the overlap domain, say 1/2, and consider the union of sample A $\left(\mathcal{S}_A\right)$ and sample B $\left(\mathcal{S}_B\right)$ as a single sample. By doing so, the chi-squared tests for a single frame survey in literature such as Rao and Scott (1981) could be applied. However, this application is based on the assumption that a set of ultimate cell proportions exist for the dual frame structure, which is not necessary true. In this paper, we assume that each domain has their own set of cell proportions, under which Rao and Scott (1981) type estimator is a special case when the three sets of cell proportions in the three domains are all the same. We extend Wald's (1943) test and Rao-Scott first-order and second-order corrected tests (Rao and Scott 1981) from a single survey to a dual frame survey and derive asymptotic distributions.

This paper is organized as follows. Section 2 gives a background of the research. Section 3 proposes several chi-squared tests. Section 4 gives a small simulation study of the proposed chi-squared tests under a simple hypothesis. Section 5 gives a real example study. Finally, we give a summary in Section 6.

2 Background

2.1 Chi-squared tests in a single frame survey

Consider a one-way frequency table with k classes and associated finite population proportions $p_1, p_2, \cdots, p_k$ with $\sum_{i=1}^{i=k} p_i = 1$. Let $n_1, \cdots, n_k$ denote the observed cell frequencies in a sample falling in each of k categories with $\sum_{i=1}^{k} n_i = n$. Under SRS, the Pearson chi-squared statistic for testing simple hypothesis $H_0 : p_i = p_{0i},\ \left(i = 1, \cdots, k\right)$ is given by

$$\tilde{X}^2 = \sum_{i=1}^{k} \frac{\left(n_i - np_{0i}\right)^2}{np_{0i}}. \tag{2.1}$$

For complicated designs, $\tilde{X}^2$ involve noncentral distributions. It is natural to consider a more general statistic

$$X^2 = n\sum_{i=1}^{k} \frac{\left(\hat{p}_i - p_{0i}\right)^2}{p_{0i}}, \tag{2.2}$$

where $\hat{p}_i$ is a consistent estimator of p_i under a specified sampling design $p(s)$.

Let $\hat{\mathbf{p}} = (\hat{p}_1, \cdots, \hat{p}_{k-1})'$ represent the $k-1$ vector of estimated proportions with $\hat{p}_k = 1-(\hat{p}_1 + \cdots + \hat{p}_{k-1})$; $\mathbf{p}_0$ be the corresponding $k-1$ vector of hypothesized proportions; $\mathbf{V}$ be the $(k-1)\times(k-1)$ covariance matrix of $\hat{\mathbf{p}}$, and $\hat{\mathbf{V}}$ be the estimate of $\mathbf{V}$ obtained from the survey data. The generalized Wald statistic

$$X_W^2 = (\hat{\mathbf{p}} - \hat{\mathbf{p}}_0)' \hat{\mathbf{V}}^{-1} (\hat{\mathbf{p}} - \hat{\mathbf{p}}_0), \tag{2.3}$$

is distributed asymptotically as χ^2_{k-1} under $H_0 : p_i = p_{0i}$, $(i = 1, \cdots, k)$ for sufficiently large n.

Rao and Scott (1981) showed that under H_0, X^2 in (2.2) is distributed asymptotically as a weighted sum $\delta_1 W_1 + \cdots + \delta_{k-1} W_{k-1}$ of $k-1$ independent χ^2_1 random variables W_i, $i = 1, 2, \cdots, k-1$. The δ_is are the eigenvalues of a design effect matrix $\mathbf{P}^{-1}\mathbf{V}$, where $\mathbf{P}$ is the covariance matrix corresponding to SRS when H_0 is true, i.e. $\mathbf{P} = n^{-1}\left(diag(\mathbf{p}_0) - \mathbf{p}_0\mathbf{p}_0'\right)$. The standard result of Pearson test is recovered under SRS. Let $\hat{\delta}_i$ be an estimate of δ_i and $\hat{\delta}_\cdot = \left(\sum_{i=1}^{k-1}\hat{\delta}_i\right)\Big/(k-1)$, the Rao-Scott first order corrected test refers $X^2/\hat{\delta}_\cdot$ to χ^2_{k-1}. When the full estimated covariance matrix $\hat{\mathbf{V}}$ is known, a better approximation to the asymptotic distribution of X^2 is to match the first moment and second moment of the test statistic to a χ^2 distribution. The Rao-Scott (Rao and Scott 1981) second-order corrected test statistic considers $X_S^2 = X^2 \Big/ \left[\hat{\delta}_\cdot\left(1+\hat{a}^2\right)\right]$. This statistic is approximately a chi-squared random variable on $v = (k-1)\Big/\left(1+a^2\right)$ degrees of freedom, where $\hat{a}$ is an estimate of a with $\hat{a}^2 = \sum_{i=1}^{k-1}\hat{\delta}_i^2 \Big/ \left[(k-1)\hat{\delta}_\cdot^2\right] - 1$, and $\sum_{i=1}^{k-1}\hat{\delta}_i^2 = n^2\sum_{i=1}^{k}\sum_{j=1}^{k}\hat{\mathbf{V}}_{ij}^2 / p_{0i}p_{0j}$. If the design effects are all similar, the first and second-order corrections will behave similarly. Otherwise, the second order correction almost always performs better.

2.2 Framework of chi-squared tests and pseudo maximum likelihood estimator in dual frame surveys

The set up in this section follows from Hartley (1962) and Lu and Lohr (2010). Assume there are k categories in both surveys and the same quantities are measured. Let p_{id} be the population proportion of category i in domain d (domain d can be domain a, domain ab or domain b), with $\sum_{i=1}^{k} p_{id} = 1$. Let N_a, N_{ab} and N_b denote the population sizes of the three domains respectively, with $N_a + N_{ab} = N_A$ and $N_b + N_{ab} = N_B$. We consider the common case that N_{ab} is unknown, while N_A and N_B are constants. As a result, $\sum_{i=1}^{k} p_{ia}N_a/N_A + \sum_{i=1}^{k} p_{iab}N_{ab}/N_A = 1$ and $\sum_{i=1}^{k} p_{ib}N_b/N_B + \sum_{i=1}^{k} p_{iab}N_{ab}/N_B = 1$ (see Figure 2.1 for illustration of the proportions). The vector of proportions $\mathbf{p} = (p_1, p_2, \cdots p_{k-1})'$ for the union of the two frames is a function of the parameters p_{ia}, p_{iab}, p_{ib} and N_{ab}. For example, a natural form of p_i is

$$p_i = \frac{N_a}{N} p_{ia} + \frac{N_{ab}}{N} p_{iab} + \frac{N_b}{N} p_{ib}, \qquad \text{for} \qquad i = 1, 2, \cdots, k-1, \tag{2.4}$$

where $N = N_A + N_B - N_{ab}$.

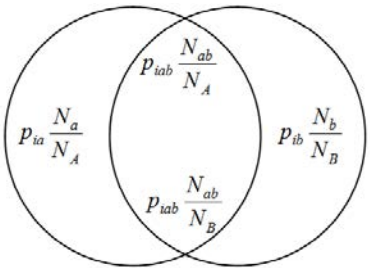

Figure 2.1: Population proportion in domains and frames

In the following, we briefly review the pseudo maximum likelihood estimator that we will use in Section 4 and Section 5. Assume independent simple random samples are taken from frames A and B respectively. The likelihood function is

$$L\left(p_{ia}, p_{iab}, p_{ib}, N_{ab}\right) \propto \prod_i \left(p_{ia}\frac{N_a}{N_A}\right)^{x_{ia}} \times \prod_i \left(p_{iab}\frac{N_{ab}}{N_A}\right)^{x_{iab}^A} \times \prod_i \left(p_{ib}\frac{N_b}{N_B}\right)^{x_{ib}} \times \prod_i \left(p_{iab}\frac{N_{ab}}{N_B}\right)^{x_{iab}^B} \quad (2.5)$$

where x_{ia}, x_{ib} represent the units falling in category i within domain a and domain b respectively; x_{iab}^A and x_{iab}^B represent the units falling in category i within the overlapping domain ab that are originally sampled from frame A and frame B respectively.

For the estimators of complex surveys, the basic idea is to use a working assumption of a multinomial distribution from a finite population to give the form of the estimators and use a design effect to adjust the cell counts to reflect the complex survey design. The pseudo likelihood function is as follows

$$\begin{aligned} L\left(p_{ia}, p_{iab}, p_{ib}, N_{ab}\right) \propto & \prod_i \left(p_{ia}\frac{N_a}{N_A}\right)^{\frac{\tilde{n}_A}{N_A}\hat{X}_{ia}} \prod_i \left(p_{iab}\frac{N_{ab}}{N_A}\right)^{\frac{\tilde{n}_A}{N_A}\hat{X}_{iab}^A} \\ & \times \prod_i \left(p_{ib}\frac{N_b}{N_B}\right)^{\frac{\tilde{n}_B}{N_B}\hat{X}_{ib}} \prod_i \left(p_{iab}\frac{N_{ab}}{N_B}\right)^{\frac{\tilde{n}_B}{N_B}\hat{X}_{iab}^B}, \end{aligned} \quad (2.6)$$

where design effect is defined as $\left\{v\left(\hat{\theta}\right) \text{ from complex survey}\right\} \Big/ \left\{v\left(\hat{\theta}\right) \text{ from SRS of same size}\right\}$, $\tilde{n}_A = n_A \Big/ \left(\text{design effect of } \hat{N}_{ab}^A\right)$, $\tilde{n}_B = n_B \Big/ \left(\text{design effect of } \hat{N}_{ab}^B\right)$, n_A and n_B are the observed sizes of S_A and S_B, and $\hat{X}_{id}$ denote the estimated counts according to the survey design. The pseudo maximum likelihood estimators (PMLEs), found by maximizing (2.6) are $\hat{p}_{ia} = \hat{X}_{ia} \big/ \hat{N}_a$, $\hat{p}_{ib} = \hat{X}_{ib} \big/ \hat{N}_b$, and

$$\hat{p}_{iab} = \frac{\frac{\tilde{n}_A}{N_A}\hat{N}_{ab}^A\hat{p}_{iab}^A + \frac{\tilde{n}_B}{N_B}\hat{N}_{ab}^B\hat{p}_{iab}^B}{\frac{\tilde{n}_A}{N_A}\hat{N}_{ab}^A + \frac{\tilde{n}_B}{N_B}\hat{N}_{ab}^B}, \quad (2.7)$$

where $\hat{p}_{iab}^A = \hat{X}_{iab}^A \big/ \hat{N}_{ab}^A$ and $\hat{p}_{iab}^B = \hat{X}_{iab}^B \big/ \hat{N}_{ab}^B$, and $\hat{N}_{ab,PML}$ is the smaller root of the quadratic function

$$\left[\tilde{n}_A + \tilde{n}_B\right]\hat{N}_{ab,PML}^2 - \left[\tilde{n}_A N_B + \tilde{n}_B N_A + \tilde{n}_A\hat{N}_{ab}^A + \tilde{n}_B\hat{N}_{ab}^B\right]\hat{N}_{ab,PML} + \left[\tilde{n}_A\hat{N}_{ab}^A N_B + \tilde{n}_B\hat{N}_{ab}^B N_A\right] = 0. \quad (2.8)$$

The estimators of the population proportions are

$$\hat{p}_{i,PML} = \frac{\left(N_A - \hat{N}_{ab,PML}\right)\hat{p}_{ia} + \hat{N}_{ab,PML}\hat{p}_{iab} + \left(N_B - \hat{N}_{ab,PML}\right)\hat{p}_{ib}}{N_A + N_B - \hat{N}_{ab,PML}}. \tag{2.9}$$

If SRSs are taken in each frame and $k = 1$, these PMLEs reduced to PMLEs in Skinner and Rao (1996).

3 Chi-squared tests in dual frame surveys

In this section, we consider the case of chi-squared tests in a dual frame survey. Some hypotheses of interest may include: a simple hypothesis $H_0: q_{ia} = p_{ia}N_a/N_A = q^A_{ia0}$, $q^A_{iab} = p_{iab}N_{ab}/N_A = q^A_{iab0}$, $q^B_{iab} = p_{iab}N_{ab}/N_B = q^B_{iab0}$, $q_{ib} = p_{ib}N_b/N_B = p_{ib0}$ (note that q_{ia}, etc., are used to simplify the notations); $H_0: p_{i,PML} = p_{i0,PML}$, in which we test whether the PMLE of proportions from the union of the two frames in (2.9) are some specific values (note that p_i can be estimated by other methods); $H_0: p_{ia} = p_{iab} = p_{ib}$, testing whether the proportions are equal in the three domains; or $H_0: p_{ij} = p_{i+}p_{+j}$, testing independence of the row classification and column classification.

Let $\boldsymbol{\eta} = \left(\mathbf{p}'_a N_a/N_A, \mathbf{p}'_{ab} N_{ab}/N_A, \mathbf{p}'_b N_b/N_B, \mathbf{p}'_{ab} N_{ab}/N_B\right)'$, $\mathbf{p}_a = \left(p_{1a}, p_{2a}, \cdots p_{ka}\right)'$, $\mathbf{p}_b = \left(p_{1b}, p_{2b}, \cdots p_{kb}\right)'$, $\mathbf{p}_{ab} = \left(p_{1ab}, p_{2ab}, \cdots p_{(k-1)ab}\right)'$, and h_i's are continuous functions. A more general hypothesis of interest may be denoted as the following:

$$H_0: h_i(\boldsymbol{\eta}) = 0, \quad i = 1, 2, \cdots r. \tag{3.1}$$

Let η_j be the j-th element of $\boldsymbol{\eta}$ and let $h(\boldsymbol{\eta}) = \left(h_1(\boldsymbol{\eta}), h_2(\boldsymbol{\eta}), \cdots h_r(\boldsymbol{\eta})\right)'$.

Assume that $\partial h_i(\boldsymbol{\eta})/\partial \eta_j$ is continuous in a neighborhood of $\boldsymbol{\eta}$ and that

$$\nabla = \frac{\partial h_i(\boldsymbol{\eta})}{\partial \eta_j} \tag{3.2}$$

has full rank. Also assume

A$_1$. There is a sequence of superpopulations $U_{A1} \subset U_{A2} \subset \ldots \subset U_{At} \subset \ldots$ as defined in Isaki and Fuller (1982).

A$_2$. Let $\tilde{n}_A$ and $\tilde{n}_B$ as defined in Section 2 and assume that $\tilde{n}_A$ and $\tilde{n}_B$ both increase such that $\tilde{n}_A/\tilde{n}_B \to \gamma$ for some $0 < \gamma < 1$.

A$_3$. Let $\pi^A_{it} = p\left(\text{psu } i \text{ is in sample from Frame } A, \text{using population } U_{At}\right)$ and $\pi^A_{ijt} = p\left(\text{psus } i \text{ and } j \text{ are in sample from Frame } A, \text{ using population } U_{At}\right)$ be the inclusion and joint inclusion probabilities for the frame-A sample from population U_{At}, and define π^B_{it}, π^B_{ijt} and U_{Bt} similarly for frame B. Assume there are constants c_1 and c_2 such that

$$0 < c_2 < \pi_{it}^F < c_1 < 1 \tag{3.3}$$

for all i and any superpopulation in the sequence, where F denotes frame A or frame B. Also assume there exists an α_t with $\alpha_t = o(1)$ such that

$$\pi_{it}^F \pi_{jt}^F - \pi_{ijt}^F \le \alpha_t \pi_{it}^F \pi_{jt}^F. \tag{3.4}$$

A$_4$. $N_{ab}/N \to \psi$ for some ψ between 0 and 1.

Theorem 1. *With assumptions $A_1 - A_4$ set out beforehand, we have the following conclusion: $\tilde{n}^{1/2}\mathbf{h}(\hat{\boldsymbol{\eta}})$ is asymptotically normal with mean $\mathbf{0}$ and asymptotic variance $\nabla\Sigma\nabla'$, where Σ is a block-diagonal matrix with blocks Σ_A and Σ_B and $\tilde{n} = \tilde{n}_A + \tilde{n}_B$. Σ_A is the asymptotic covariance matrix of $\tilde{n}^{1/2}\hat{\boldsymbol{\eta}}_A$ with $\hat{\boldsymbol{\eta}}_A = \left(\hat{\mathbf{p}}_a' \hat{N}_a/N_A, \hat{\mathbf{p}}_{ab}^{A\prime} \hat{N}_{ab}/N_A\right)'$, Σ_B is the asymptotic covariance matrix of $\tilde{n}^{1/2}\hat{\boldsymbol{\eta}}_B$ with $\hat{\boldsymbol{\eta}}_B = \left(\hat{\mathbf{p}}_b' \hat{N}_b/N_B, \hat{\mathbf{p}}_{ab}^{B\prime} \hat{N}_{ab}/N_B\right)'$ and $\hat{\boldsymbol{\eta}} = \left(\hat{\boldsymbol{\eta}}_A', \hat{\boldsymbol{\eta}}_B'\right)'$.*

Proof. The arguments given in Theorem 1 in Lu and Lohr (2010) show that $\hat{\boldsymbol{\eta}}$ is consistent for $\boldsymbol{\eta}$ and that $\hat{\boldsymbol{\eta}}$ obeys the central limit theorem, as $\tilde{n}_A$ and $\tilde{n}_B$ both increase such that $\tilde{n}_A/\tilde{n}_B \to \gamma$. Thus, since the samples S_A and S_B are selected independently, we have

$$\tilde{n}^{1/2}\left(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}\right) \xrightarrow{d} N(0, \Sigma).$$

$\mathbf{h}(\hat{\boldsymbol{\eta}})$ is consistent for $\mathbf{h}(\boldsymbol{\eta})$ because $\hat{\boldsymbol{\eta}}$ is consistent for $\boldsymbol{\eta}$. Using the delta method, $\tilde{n}^{1/2}\mathbf{h}(\hat{\boldsymbol{\eta}})$ is asymptotically normal with mean $\mathbf{0}$ and asymptotic variance $\nabla\Sigma\nabla'$.

Based on Theorem 1, the following results follow immediately.

Result 1. (Extended Wald Test) If a consistent estimator of the variance Σ is available, by Theorem 1, the generalized Wald statistic can be formed as follows:

$$X_W^2 = \tilde{n}\mathbf{h}(\hat{\boldsymbol{\eta}})'\left(\hat{\nabla}\hat{\Sigma}\hat{\nabla}'\right)^{-1}\mathbf{h}(\hat{\boldsymbol{\eta}}). \tag{3.5}$$

This test statistic is distributed asymptotically as $\chi^2(r)$ under H_0 (refer to equation 3.1), where r is the rank of ∇.

As we have noted previously, the estimate of the variance may be unstable or no closed-form estimate of Σ is available. One way we can modify the statistic in (3.5) is to first act as though the sample is a simple random sample, then modify the reference distribution used in the test to get the correct level. Equation (3.6) gives the modified statistic.

Result 2. Let

$$X_{MW}^2 = \tilde{n}\mathbf{h}(\hat{\boldsymbol{\eta}})'\left(\hat{\nabla}_0\hat{\mathbf{P}}_0\hat{\nabla}_0'\right)^{-1}\mathbf{h}(\hat{\boldsymbol{\eta}}), \tag{3.6}$$

where $\hat{\nabla}_0\hat{\mathbf{P}}_0\hat{\nabla}_0'$ can be any estimate of $\nabla\mathbf{P}\nabla$ that is consistent when H_0 is true. Matrix $\mathbf{P}_0$ is a block diagonal matrix with diagonal blocks: covariance matrix from frame A and covariance matrix from frame

B when H_0 is true and when sampling is SRS. Suppose the matrix ∇ has rank r under the null hypothesis $H_0 : \mathbf{h}(\boldsymbol{\eta}) = 0$. Then $X^2_{MW} \approx \sum_1^r \lambda_{0i} W_i$, where the λ_i's are the eigenvalues of $(\nabla \mathbf{P} \nabla')^{-1}(\nabla \Sigma \nabla')$, $W_1, \ldots, W_r$ are independent χ^2_1 random variables and λ_{0i} is the value of λ_i under H_0.

Result 3. (Extended Rao-Scott first order correction) Suppose matrix ∇ has rank r. Let X^2_{MW} be as defined in (3.6). Under the null hypothesis $H_0 : \mathbf{h}(\boldsymbol{\eta}) = 0$, the statistic $X^2_{MW} / \hat{\lambda}.$ has expectation r, where $\hat{\lambda}. = \sum \hat{\lambda}_i / r$, $\hat{\lambda}_i$ is a consistent estimate of λ_i under H_0. For example, $\hat{\lambda}_i$'s could be the eigenvalues of $\left(\hat{\nabla}_0 \hat{\mathbf{P}}_0 \hat{\nabla}_0'\right)^{-1}\left(\hat{\nabla} \hat{\Sigma} \hat{\nabla}'\right)$.

Result 4. (Extended Rao-Scott second order correction) Suppose matrix ∇ has rank r. Define

$$X^2_S = \frac{X^2_{MW}}{\hat{\lambda}.\left(1 + \hat{a}^2\right)}$$

where $\hat{a}^2 = \sum_{i=1}^{k-1} \hat{\lambda}_i^2 \Big/ \left[(k-1)\hat{\lambda}.^2\right] - 1$ is an estimate of the population value a^2. Under null hypothesis, X^2_S is distributed asymptotically as χ^2_v, a chi-square random variable with degrees of freedom $v = (k-1)/\left(1 + a^2\right)$.

4 Simulations

In this section, a small simulation has been conducted to study the proposed chi-squared tests under a simple hypothesis $H_0 : q_{ia} = p_{ia} N_a / N_A = q^A_{ia0}$, $q^A_{iab} = p_{iab} N_{ab} / N_A = q^A_{iab0}$, $q^B_{iab} = p_{iab} N_{ab} / N_B = q^B_{iab0}$, $q_{ib} = p_{ib} N_b / N_B = p_{ib0}$ to investigate the performance of chi-squared tests proposed in Section 3. We compare the percentages of samples for which the test statistics exceed the critical value to the nominal level $(\alpha = 0.05)$. R (www.r-project.org) is used to perform simulation study and other analysis.

We generated the data following Skinner and Rao (1996), with $\gamma_a = N_a / N$ and $\gamma_b = N_b / N$. A cluster sample from frame A was generated with n_p psus and m observations in each psu, and a simple random sample of n_B observations was generated for frame B. We generated the clustered binary responses for the sample from frame A by generating correlated multivariate normal random vectors and then using the probit function to convert the continuous responses to binary responses. After the sample was generated, we calculated the PML estimators of $\mathbf{p}_{id} N_d / N_A$ and $\mathbf{p}_{id} N_d / N_B$ (see Section 2.2). These estimated proportions were used to compute the chi-squared test statistics. We then compared the percentages of samples for which the test statistics exceed the critical value to the nominal level under different settings.

The simulation study was performed with factors: (1) γ_a : 0.4, (2) γ_b : 0.2, (3) clustering parameter ρ : 0.3, (4) sample sizes: n_p : 10, 30 or 50; m : 3, 5, or 10, n_B : 100, 300 or 500. (5) Simulation runs: 1,000 times for each setting and 100 times when estimating the variance covariance matrix V using bootstrapping. All runs used probability parameters $\mathbf{p}_a$: $(.3,.1,.2,.4)$, $\mathbf{p}_{ab}$: $(.3,.1,.1,.5)$, and

$\mathbf{p}_b$: $(.4,.1,.1,.4)$. Table 4.1 reported the percentages of samples for which the test statistics exceed the critical value.

Table 4.1
Comparison of the actual significance levels (%) among different tests. X^2 is the uncorrected test; X^2_{FC} is the first order corrected X^2 and X^2_{SC} is the second order corrected X^2.

$\tilde{n}_p$	m	n_B	X^2	**Wald**	X^2_{FC}	X^2_{SC}
10	3	100	12.1	17.3	5.6	4.9
30	3	300	13.6	8.4	4.8	4.8
50	3	500	15.5	10.0	6.4	3.6
10	5	100	25.7	13.5	7.5	4.9
30	5	300	29.2	9.3	7.9	5.3
50	5	500	31.5	8.5	8.1	4.9
10	10	100	46.1	21.2	6.6	5.4
30	10	300	50.2	11.5	7.5	5.6
50	10	500	58.7	8.0	9.6	5.1

Table 4.1 indicates that naively using uncorrected X^2 test for complex survey data is dangerous. With increased psu size and number of psu's, the actual significance level even reaches 62.2%. Extended Wald test doesn't perform well since the estimate of the variance may be unstable. Extended first order corrected test is acceptable with actual significance level around 7%. Extended second order corrected tests almost reach the nominal level 5%, for which is the one we recommend to use in a dual frame survey categorical data analysis.

5 Application

In this section, we give a real example to illustrate how to perform the chi-squared tests in a dual frame survey. We consider the hypothesis test $H_0: p_{ia} = p_{iab} = p_{ib}$, testing whether the proportions are equal in the three domains.

5.1 Data description and related PMLEs

Data (Lohr and Rao 2006) were originally collected for a three-frame survey of statisticians, using the online membership directories of the American Statistical Association (ASA), the Institute for Mathematical Statistics (IMS) and the Statistical Society of Canada. We treat the union of online membership directories of ASA and online membership directories of IMS as a dual frame with notation $A \cup B = a \cup ab \cup b$ (A: online membership directories of ASA; B: online membership directories of IMS; domain a: ASA member but not IMS member; domain ab: ASA member and also IMS member; domain b: IMS member but not ASA member). Note that the union of these two frames does not cover the entire population of statisticians. Many statisticians do not belong to either of the two societies, and some statisticians decline to participate in online directories. In the data set, the information of the occupation is a categorical variable with three levels: academia, industry and government. We combine

industry and government to be one level named nonacademia. Together with sex, we have a 2×2 table with four cells: female in academia, female not in academia, male in academia and male not in academia.

At the time of data collection, there were $15,500$ people in American Statistical Association (Frame A) and $4,000$ people in Institute for Mathematical Statistics (Frame B), so $N_A = 15,500$ and $N_B = 4,000$. A stratified cluster sample of size 500 was taken from frame A, of which 378 observations had information on both responses (sex and occupation). The design had 26 strata constructed by regions or states. Because of the restrictions on access to records, clusters for large states were members whose last name began with the same letter of the alphabet. There are 173 psu's in frame A. A simple random sample of size 140 was taken from frame B, in which 102 records have valid information for both responses. The weighted total of observations from frame A is $10,976$. We assume that data are missing randomly, so the nonresponse is adjusted by a fraction of $15,500/10,976$. Table 5.1 lists the number of statisticians falling in each cell within each domain.

Table 5.1
Observed data in domain a and domain ab from frame A (adjusted by a fraction of 15,500/10,976) together with observed data in domain b and domain ab from frame B.

	Domain a		Domain $ab \in A$		Domain b		Domain $ab \in B$	
	Female	Male	Female	Male	Female	Male	Female	Male
Academia	2,425	4,969	302	1,488	10	41	10	33
Nonacademia	1,959	4,091	59	209	0	3	2	3

The estimated design effect of frame A is 1.801209, so the effective sample size of n_A is $\tilde{n}_A = 378/1.8 = 210$. The effective sample size of $n_{B,eff} = n_B = 102$. The PMLEs of the estimated proportions by using (2.6) and (2.9) are listed in Table 5.2.

Table 5.2
Estimated proportions from domains and union of two frames.

	Domain a		Domain ab		Domain b		Frame $A \cup B$	
	Female	Male	Female	Male	Female	Male	Female	Male
Academia	0.180	0.370	0.186	0.701	0.185	0.759	0.182	0.452
Nonacademia	0.146	0.304	0.037	0.076	0	0.056	0.116	0.250

5.2 Test the equivalence of proportions across domains

The hypothesis of interest is whether the proportions are equal across the three domains,

$$H_0 : p_{ia} = p_{iab} \quad \text{and} \quad p_{iab} = p_{ib},\ i = 1,2,3. \tag{5.1}$$

In this example, $p_{ia},\ i = 1,2,3,4$ represent the proportion of female in academia, female not in academia, male in academia and male not in academia among ASA members respectively. Similarly define p_{iab} and p_{ib}. $\boldsymbol{\eta}$ (see Section 3) reduces to a 14×1 vector

$$\boldsymbol{\eta} = (p_{1a} N_a/N_A, p_{2a} N_a/N_A, p_{3a} N_a/N_A, p_{4a} N_a/N_A, p_{1ab} N_{ab}/N_A, p_{2ab} N_{ab}/N_A, p_{3ab} N_{ab}/N_A,$$
$$p_{1b} N_b/N_B, p_{2b} N_b/N_B, p_{3b} N_b/N_B, p_{4b} N_b/N_B, p_{1ab} N_{ab}/N_B, p_{2ab} N_{ab}/N_B, p_{3ab} N_{ab}/N_B)'.$$

Since H_0 in (5.1) only involves the simple parameters p_{ia}, p_{iab}, p_{ib} and N_{ab}, we introduce a new vector $\theta = (p_{1a}, p_{2a}, p_{3a}, p_{1ab}, p_{2ab}, p_{3ab}, N_{ab}/N_A, p_{1b}, p_{2b}, p_{3b}, p_{1ab}, p_{2ab}, p_{3ab}, N_{ab}/N_B)'$.

Let $\Omega = \left(\partial h_i(\boldsymbol{\eta})/\partial\theta_j\right)$ and $\mathbf{D}(\theta) = \left(\partial\boldsymbol{\eta}/\partial\theta_j\right)$. $\mathbf{D}(\theta)$ is found to be a block diagonal matrix with

$$\mathbf{D}_A = \begin{pmatrix}
\frac{N_a}{N_A} & 0 & 0 & 0 & 0 & 0 & -p_{1a} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & \frac{N_a}{N_A} & 0 & 0 & 0 & 0 & -p_{2a} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & \frac{N_a}{N_A} & 0 & 0 & 0 & -p_{3a} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-\frac{N_a}{N_A} & -\frac{N_a}{N_A} & -\frac{N_a}{N_A} & 0 & 0 & 0 & -p_{4a} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \frac{N_{ab}}{N_A} & 0 & 0 & p_{1ab} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \frac{N_{ab}}{N_A} & 0 & p_{2ab} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & \frac{N_{ab}}{N_A} & p_{3ab} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{N_b}{N_B} & 0 & 0 & 0 & 0 & 0 & -p_{1b} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{N_b}{N_B} & 0 & 0 & 0 & 0 & -p_{2b} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{N_b}{N_B} & 0 & 0 & 0 & -p_{3b} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{N_b}{N_B} & -\frac{N_b}{N_B} & -\frac{N_b}{N_B} & 0 & 0 & 0 & -p_{4b} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{N_{ab}}{N_B} & 0 & 0 & p_{1ab} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{N_{ab}}{N_B} & 0 & p_{2ab} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{N_{ab}}{N_B} & p_{3ab}
\end{pmatrix}.$$

Notice the relationship between $\hat{p}_{iab}$ and $\hat{p}^A_{iab}$ and $\hat{p}^B_{iab}$ from (2.7), Ω is found to be

$$\Omega = \begin{pmatrix} 1 & 0 & 0 & -\phi & 0 & 0 & 0 & 0 & 0 & 0 & -(1-\phi) & 0 & 0 & 0 \\ 0 & 0 & 0 & \phi & 0 & 0 & 0 & -1 & 0 & 0 & 1-\phi & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & -\phi & 0 & 0 & 0 & 0 & 0 & 0 & -(1-\phi) & 0 & 0 \\ 0 & 0 & 0 & 0 & \phi & 0 & 0 & 0 & -1 & 0 & 0 & (1-\phi) & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & -\phi & 0 & 0 & 0 & 0 & 0 & 0 & -(1-\phi) & 0 \\ 0 & 0 & 0 & 0 & 0 & \phi & 0 & 0 & 0 & -1 & 0 & 0 & (1-\phi) & 0 \end{pmatrix},$$

where $\phi = N_B \tilde{n}_A / \left(N_B \tilde{n}_A + N_A \tilde{n}_B \right)$. It is easy to show that $\nabla = \Omega(\mathbf{D})^{-1}$ (recall that $\nabla = \partial h_i(\boldsymbol{\eta}) / \partial \eta_j$). $\hat{\Sigma}$ is estimated by using a jackknife method by deleting one psu each time from frame A. All the results in Section 3 can be derived. The eigenvalues of $\left(\hat{\nabla}_0 \hat{\mathbf{P}}_0 \hat{\nabla}_0' \right)^{-1} \left(\nabla \Sigma \nabla' \right)$ are very close to each other, which indicates that first order corrected test perform similarly as second order corrected test. The Wald statistic, first order corrected statistic and second order corrected statistic give values of 81.48295, 72.31026 and 70.28581 respectively. Comparing to the critical value with six degrees of freedom $\chi^2(6) = 12.95$, we reject the null hypothesis that the cell proportions (female in academia, female not in academia, male in academia and male not in academia) are the same across the three domains (ASA member only, ASA and IMS member and IMS member only).

6 Conclusions

In this research, we extend Wald's (1943) test and Rao-Scott first-order and second-order corrected tests (Rao and Scott 1981) from a single survey to a dual frame survey and derive the asymptotic distributions. A limited simulation study suggests that second order corrected tests almost reach the nominal level. Although the results in this paper are for dual frame surveys, the methods are general and could be extended to more than two surveys. Our research is done in the context of survey sampling; it also applies to other settings in which data could be combined from two independent sources.

Acknowledgements

The author thanks Dr. Sharon Lohr for her valuable advisement and comments on the manuscript. The author also wants to thank the referees and the associate editor for their very helpful comments and constructive suggestions.

References

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

Fay, R.E. (1979). On adjusting the Pearson chi-square statistic for clustered sampling. In *ASA Proceedings of the Social Statistics Section*, 402-406. American Statistical Association.

Fay, R.E. (1985). A jackknifed chi-squared test for complex samples. *Journal of the American Statistical Association*, 80, 148-157.

Fuller, W.A. and Burmeister, L.F. (1972). Estimators for samples selected from two overlapping frames. In *ASA Proceedings of the Social Statistics Section*, 245-249. American Statistical Association.

Hartley, H.O. (1962). Multiple frame surveys. In *ASA Proceedings of the Social Statistics Section*, 203-206. American Statistical Association.

Hartley, H.O. (1974). Multiple frame methodology and selected applications. *Sankhyā, Series C*, 36 (3), 99-118.

Isaki, C.T. and Fuller, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

Lohr, S.L. and Rao, J.N.K. (2000). Inference from dual frame surveys. *Journal of the American Statistical Association*, 95, 271-280.

Lohr, S.L. and Rao, J.N.K. (2006). Estimation in multiple-frame surveys. *Journal of the American Statistical Association*, 101, 1019-1030.

Lu, Y. and Lohr, S. (2010). Gross flow estimation in dual frame surveys. *Survey Methodology*, vol. 36, 13-22.

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

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

Rao, J.N.K. and Scott, A.J. (1987). On simple adjustments to chi-square tests with sample survey data. *The Annals of Statistics*, 15, 385-397.

Rao, J.N.K. and Thomas, D.R. (1988). The analysis of cross-classified categorical data from complex sample surveys. *Sociological Methodology*, 18, 213-269.

Skinner, C.J. (1991). On the efficiency of raking ratio estimation for multiple frame surveys. *Journal of the American Statistical Association*, 86, 779-784.

Skinner, C.J. and Rao, J.N.K. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91, 349-356.

Thomas, D.R., Singh, A. and Roberts, G. (1996). Tests of independence on two-way tables under cluster sampling: An evaluation. *International Statistical Review*, 64(3), 295-311.

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