Size and Power of Independence Tests for $R \times C$ Tables from Complex Surveys

D.R. Thomas, Carleton University; A.C. Singh and G. Roberts, Statistics Canada
D.R. Thomas, School of Business, Carleton University, Ottawa, Canada, K1S 5B6

I. Introduction

A variety of procedures currently exist for testing hypotheses on categorical data drawn from complex sample surveys, all of which are designed to avoid the well-documented problem of inflation of test significance levels that is a characteristic of the classical multinomial-based procedures.These procedures fall into three main classes: (i) methods based on the Wald statistic, see for example Koch, Freeman and Freeman(1975); (ii) simple first and second order corrections to the classical X^2 and G^2 tests, see Brier (1980), Fellegi (1980), Rao and Scott (1981, 1984, 1987) and Bedrick (1983); (iii) methods based on jackknifing the classical X^2 and G^2 tests as proposed by Fay (1979, 1985). All three classes represent general solutions which have been applied to a wide variety of problems. Variants of these techniques have also been proposed. Thomas and Rao (1987) used F-based (rather than chi-squared) variants of both Wald and adjusted X^2 procedures. Singh (1985), developed a general approach for reducing instabilities in the covariance matrix of parameter estimates and applied this to the Wald procedure to attain better control of Type I error; see also Singh and Kumar (1986). Several reviews of the Wald and adjusted X^2 procedures have been published, see for example Hidiroglou and Rao (1987) and Rao and Thomas (1988). For details of the application of the jackknifed X^2 methodology, the reader is referred to Fay (1982).

Since the rationale for all complex survey procedures is based on asymptotic theory, it is important that the finite sample characteristics of the above methods be thoroughly explored. Monte Carlo studies have been reported by Brier (1980), Fellegi (1980), Wilson (1986), Fay (1983, 1987), Thomas and Rao (1987) and Thomas (1989). In particular, Thomas and Rao (1987) conducted a Monte Carlo study of all competing procedures (excepting that of Singh, 1985) for the case of a goodness-of-fit test under two-stage cluster sampling. Besides corroborating Fay's (1985) criticism of the Wald test, these authors established that both the second order corrections to X^2 (and G^2) and the Fay jackknife procedures performed well in practice, the former having a slight edge in terms of control of Type I error, particularly, when the "degrees of freedom" for variance estimation was small. Despite this research, much remains to be done. In fact, do not at present know the extent to which the results of Thomas and Rao (1987) can be generalized to a test of independence on a two-way table, perhaps the most heavily used test of all in the context of categorical data. The remainder of this paper considers in detail the design and execution of a Monte Carlo study of procedures for testing independence in a two-way table under a model of two-stage cluster sampling.

II. Design Requirements

1. Some Notation

For an $r \times c$ contingency table having a fixed total of n observations, let $\pi_{ij}, i = 1,\ldots,r,\ j = 1,\ldots,\ c$, represent the individual cell probabilities, and $\pi_{i.}, i = 1,\ldots,\ r$, and $\pi_{.j},\ j = 1,...,c$, represent the marginal probabilities. In vector form these will be denoted $\boldsymbol{\pi} = (\pi_{11},\ldots,\pi_{rc})'$, $\boldsymbol{\pi}_R = (\pi_{1.},\cdots,\pi_{r})'$, and $\boldsymbol{\pi}_C = (\pi_{.1},\ldots,\pi_{.c})'$. Let $\hat{\pi}_{ij}$, $\hat{\pi}_{i.}$ and $\hat{\pi}_{.j}$ represent consistent estimates of the corresponding probabilities under some suitable model of cluster sampling, and let $\mathbf{V}/n$ represent the $rc \times rc$ (singular) variance-covariance matrix of $\hat{\boldsymbol{\pi}}$, the $(rc \times 1)$ vector of estimates of cell probabilities arranged in lexicographic order. In this report, $\mathbf{V}$, and the corresponding multinomial quantities defined below will be refered to as scaled covariance matrices. Let $\widehat{\mathbf{V}}$ be a consistent estimate of $\mathbf{V}$ under the model. Also, let $\mathbf{P} = \text{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'$ represent the scaled multinomial covariance matrix corresponding to the vector $\boldsymbol{\pi}$, which is consistently estimated by $\widehat{\mathbf{P}} = \text{diag}(\hat{\boldsymbol{\pi}}) - \hat{\boldsymbol{\pi}}\hat{\boldsymbol{\pi}}'$. Finally, denote by $\mathbf{P}_R$ and $\mathbf{P}_C$ the corresponding scaled multinomial covariance matrices for the marginals, with consistent estimates $\widehat{\mathbf{P}}_R$ and $\widehat{\mathbf{P}}_C$. We will consider models of two-stage cluster sampling in which units are drawn independently from each of L independent clusters, each cluster characterized by a vector of probabilities $\mathbf{p}_\ell, \ell = 1,\ldots,L$. Following Rao and Scott (1981, 1984), we can define several different sets of generalized design effect matrices and their corresponding generalized deffs (eigenvalues), as follows:

(i) $\lambda_k,\ k = 1,\ldots,\ (rc-1)$. These are eigenvalues (or generalized deffs) of the matrix $\mathbf{D}_{RC} = \mathbf{P}^{(t)-1}\mathbf{V}^{(t)}$, which give a measure of variance inflation for the full vector of estimates $\hat{\boldsymbol{\pi}}$. The superscript (t) denotes a trimmed matrix obtained by deletion of the last row and column of the full matrix in question. These λ_k's are the relevant deffs for a goodness-of-fit hypothesis on $\boldsymbol{\pi}$. For multinomial samples, $\lambda_k = 1\ \forall\ k$; under cluster sampling, we expect that the mean $\bar{\lambda} > 1$.

(ii) $\lambda_{R(k)},\ k = 1,\ldots,r-1$ and $\lambda_{C(k)},\ k = 1,\ldots,c-1$. These are eigenvalues of the deff matrices $\mathbf{D}_R = \mathbf{P}_R^{(t)-1}\mathbf{V}_R^{(t)}$ and $\mathbf{D}_C = \mathbf{P}_C^{(t)-1}\mathbf{V}_C$, corresponding to marginal estimates $\hat{\boldsymbol{\pi}}_R$ and $\hat{\boldsymbol{\pi}}_c$. Again, the superscript (t) denotes a trimmed matrix. These are the relevant deffs for a goodness of fit test on the table marginals.

(iii) $\delta_k,\ k = 1,\ldots,(r-1)(c-1)$. These are the eigenvalues of the generalized deff matrix $\mathbf{D}_I$ corresponding to the test of independence. $\mathbf{D}_I$ can be expressed in the form

$$\mathbf{D}_I = \left(\mathbf{C}'\mathbf{D}_{\boldsymbol{\pi}}^{-1}\mathbf{C}\right)^{-1}\left(\mathbf{C}'\mathbf{D}_{\boldsymbol{\pi}}^{-1}\mathbf{V}\mathbf{D}_{\boldsymbol{\pi}}^{-1}\mathbf{C}\right) \qquad 2.1.1$$

where $\mathbf{C}$ is the completion of the design matrix $\mathbf{X}$ for the independence form of the loglinear model

$$\ln(\boldsymbol{\pi}) = \mathbf{X}\boldsymbol{\beta} \qquad 2.1.2$$

In other words, $\mathbf{X}'\mathbf{C} = 0$, where $\mathbf{C}$ is of maximum column rank. Also, $\mathbf{D}_\pi$ is the diagonal matrix $\text{diag}(\boldsymbol{\pi})$, having the elements $\pi_{ij} = \pi_{i.}\pi_{.j}$ on its diagonal. Under multinomial sampling, $\delta_k = 1\ \forall\ k$; for cluster sampling (and complex designs generally), we expect $\bar{\delta}$, the mean of the δ_k's, to be greater than one. Estimates of these generalized deffs yield the first order corrected tests of Rao and Scott (1981), which are asymptotically exact for constant design effects, i.e. $\bar{\delta} = \delta_k,\ \forall k$. The second order Rao-Scott tests account for variations among the δ_k's.

2. Model Requirements

The model should be a plausible representation of two stage cluster sampling, and should be capable of:

(i) Modelling different row and column generalized design effects, i.e., $\bar{\lambda}_R \neq \bar{\lambda}_C$. A real example of such a case is given by Rao and Thomas (1988).

(ii) Modelling a range of values of $\bar{\delta}$ for given values of $\bar{\lambda}_R$ and $\bar{\lambda}_C$ so that the direct effect of changes in $\bar{\delta}$ can be properly assessed.

(iii) Modelling unequal design effects, i.e., some $\lambda_{R(k)} \neq \bar{\lambda}_R$, some $\lambda_{C(k)} \neq \bar{\lambda}_C$, some $\delta_k \neq \bar{\delta}$.

(iv) Providing independent control of $CV(\delta)$, the coefficient of variation of the generalized design effects δ_k, $k = 1, \ldots, (r-1)(c-1)$, over a range of values of $\bar{\lambda}_R$, $\bar{\lambda}_C$, and $\bar{\delta}$. Preliminary Monte Carlo results discussed in Section 4 show that large $CV(\delta)$'s tend to produce liberal test significance levels for all but one of the procedures under study.

(v) Modelling patterns of marginal probabilities other than the equiprobable case $\pi_{i\cdot} = 1/r$, $\pi_{\cdot j} = 1/c$, $\forall\, i, j$.

(vi) Modelling deviations from $H_0 : \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}$, so that the powers of the competing procedures can be assessed.

An additional constraint is imposed by cost, both in terms of computer cycles and programming time. Thomas, Singh and Roberts (1989), hereafter referred to as TSR, discuss several models of two-stage cluster sampling that can be relatively easily implemented. These include Brier's (1980) Dirichlet multinomial (DM) model and its extension to DM mixtures Thomas and Rao, 1987). They also include an extension of Brier's model to logistic normal distributions made possible by Scott and Rao's (1981) generalization of the method. Unfortunately, these various models all fail to satisfy one or more of the above conditions. A new model, based on a "modified logistic normal" (MLN) distribution, has therefore been developed for use in this study. The MLN model can, in principle, satisfy all the design conditions; the preliminary implementation described in this paper satisfies every condition except condition (iv).

III. The Modified Logistic Normal Model.

1. Introduction

Given cell probabilities π_{ij}, $i = 1, \ldots, r$, $j = 1, \ldots, c$, the MLN model generates $(rc \times 1)$ vectors of non integer pseudo-counts $\mathbf{m}_\ell$ that satisfy $E(\mathbf{m}_\ell) = m\boldsymbol{\pi}$. L independent draws of pseudo-counts will be used to represent L clusters, and an estimator $\hat{\boldsymbol{\pi}}$ of the vector of probabilities $\boldsymbol{\pi}$ will be found that:

(i) is consistent (and asymptotically unbiased) as $L \to \infty$;

(ii) has a covariance matrix that exhibits the variance inflation characteristic of two-stage clustering.

The cell probabilities can be generated, under the independence hypothesis, from preset marginals $\boldsymbol{\pi}_R$ and $\boldsymbol{\pi}_C$, or, if non-independence is to be simulated, they can be generated using the Bahadur representation

$$\pi_{ij} = \pi_{i\cdot}\pi_{\cdot j} + \rho_{ij}\{\pi_{\cdot j}(1-\pi_{\cdot j})\}^{1/2}\{\pi_{i\cdot}(1-\pi_{i\cdot})\}^{1/2} \qquad 3.1.1$$

The data generation scheme will be discussed first, followed by a description of the sample estimators and their properties. For proofs, see TSR (1989).

2. Data Generation

Stage I: Let $\mu_{ij} = \ln(\pi_{ij})$, $i = 1, \ldots, r$, $j = 1, \ldots, c$. Then draw a sample of L cluster proportions $\mathbf{p}_\ell = (p_{11\ell}, \ldots p_{rc\ell})'$, $\ell = 1, \ldots L$ according to the following scheme.

(a) Draw $X_{ij\ell}$ from $N(\mu_{ij}, \sigma_{ij}^2)$, where σ_{ij}^2 are constants which will be chosen to yield desired values of $\bar{\lambda}_R$, $\bar{\lambda}_C$, and $\bar{\delta}$.

(b) Set $y_{ij\ell} = \exp\{X_{ij\ell} - \sigma_{ij}^2/2\}$. 3.2.1

(c) Form $p_{ij\ell} = y_{ij\ell} \Big/ \left\{\sum_i \sum_j y_{ij\ell}\right\} = y_{ij\ell}/T_\ell$. 3.2.2

Then we have:

Result 1

$$(a) \quad \sum_i \sum_j p_{ij\ell} = 1, \quad \ell = 1, \ldots, L. \qquad 3.2.3$$

$$(b) \quad E(y_{ij\ell}) = \pi_{ij}; \quad E(T_\ell) = 1; \qquad 3.2.4$$

$$(c) \quad V(y_{ij\ell}) = \pi_{ij}^2\gamma_{ij}^2; \quad V(T_\ell) = \sum_i \sum_j \pi_{ij}^2\gamma_{ij}^2; \qquad 3.2.5$$

$$\text{where} \quad \gamma_{ij}^2 = \exp(\sigma_{ij}^2) - 1. \qquad 3.2.6$$

Stage II: For each $\mathbf{p}_\ell$, select a "modified multinomial sample" of non-integer "size" mT_ℓ as follows:

(a) Draw a multinomial sample of size m conditional on $\mathbf{p}_\ell$ and denote it $\mathbf{m}_\ell^*$, where $\sum_i \sum_j m_{ij\ell}^* = m$, and m is integer.

(b) Form $\mathbf{m}_\ell = T_\ell \mathbf{m}_\ell^*$, so that $\sum_i \sum_j m_{ij\ell} = mT_\ell$.

The total sample "size" is now $\sum_{\ell=1}^{L} mT_\ell = m \sum_{\ell=1}^{L} T_\ell = n$, which is random. However, $E(n) = mL$.

Result 2

$$(a) \quad E(\mathbf{m}_\ell) = m\boldsymbol{\pi}; \qquad 3.2.7$$

$$(b) \quad V(\mathbf{m}_\ell) = m\mathbf{P} + m^2 \text{ diag } (\pi_{ij}^2\gamma_{ij}^2), \qquad 3.2.8$$

where diag$(\cdot)$ denotes a diagonal matrix of order $rc \times rc$, and $\mathbf{P} = \text{ diag } (\pi_{ij}) - \boldsymbol{\pi}\boldsymbol{\pi}'$.

3. Estimation

A natural estimator of π_{ij} is $\tilde{\pi}_{ij} = \left(\sum_{\ell=1}^{L} m_{ij\ell}\right)/mL$, $i = 1, \ldots, r$, $j = 1, \ldots, c$, for which $E(\tilde{\boldsymbol{\pi}}) = \boldsymbol{\pi}$, and $\mathbf{V}(\tilde{\boldsymbol{\pi}}) = (m^2L)^{-1}\mathbf{V}(\mathbf{m}_\ell)$. Unfortunately, $\sum_i \sum_j \tilde{\pi}_{ij} = \left(\sum_{\ell=1}^{L} T_\ell\right)/L = \overline{T} \neq 1$. Such an estimator is not acceptable in practice, even though $E(\overline{T}) = 1$. An alternative estimator is given by

$$\hat{\pi}_{ij} = \Sigma m_{ij\ell}/m\Sigma T_\ell = \tilde{\pi}_{ij}/\overline{T} \qquad 3.3.1$$

From the earlier results, $\hat{\pi}_{ij}$ is a consistent estimator of π_{ij} as $L \to \infty$. Also, it can be shown that

Result 3

$$\mathbf{V}(\hat{\boldsymbol{\pi}}) \doteq (mL)^{-1}(\mathbf{P} + m\mathbf{P}\mathbf{D}_{\boldsymbol{\gamma}^2}\mathbf{P}) \qquad 3.3.2$$

where $\mathbf{D}_{\boldsymbol{\gamma}^2} = \text{ diag}(\gamma_{ij}^2)$, and the error in the Taylor series approximation to $V(\hat{\pi})$ is $o(L^{-1})$.

The scaled covariance matrix of $\hat{\boldsymbol{\pi}}$ again demonstrates the extra-multinomial variation typical of two-stage cluster sampling. As $\sigma_{ij}^2 \to 0$, $\gamma_{ij}^2 \to 0$ and the scaled form of $\mathbf{V}(\hat{\boldsymbol{\pi}})$ reduces to the multinomial covariance matrix $\mathbf{P}$.

The next task is to develop an estimator of $\mathbf{V}(\hat{\boldsymbol{\pi}})$, which can be used in the Monte Carlo study. The estimator $\hat{\boldsymbol{\pi}}$ can be written in the form

$$\hat{\boldsymbol{\pi}} = L^{-1}\sum_{\ell=1}^{L}\left(\frac{T_\ell}{\overline{T}}\right)\hat{\mathbf{p}}_\ell, \qquad 3.3.3$$

where

$$\hat{\mathbf{p}}_\ell = \mathbf{m}_\ell^*/m = \mathbf{m}_\ell/mT_\ell. \qquad 3.3.4$$

The cluster probabilities $\hat{\mathbf{p}}_\ell$ are *i.i.d.*, but the weights $T_\ell/\overline{T}$ are not. The weights are correlated as a result of their common

denominator $\overline{T}$. The $\hat{\mathbf{p}}_\ell$ themselves do not, therefore, lead to a natural variance estimator. A consistent estimator of $\mathbf{V}(\hat{\pi})$ given in Result 3 can, however, be obtained (TSR, 1989).

The elements of the model are now in place: (i) a random number generating scheme; (ii) a consistent estimator $\hat{\pi}$, together with a consistent estimator of its variance; (iii) a convenient expression for the asymptotic variance of $\hat{\pi}$ which can be used to design the experiment and define the required parameter settings. A drawback of this scheme is that $\hat{\pi}$ is not model unbiased for π. TSR (1980) show, however, that for the cases considered in this study the bias will have a negligible effect.

4. Controlling the Average Generalized Deffs

As noted in Section 2, the model must allow for experimental control of the average generalized design effects $\bar{\lambda}_R, \bar{\lambda}_C$, and $\bar{\delta}$, which amounts to finding values for the elements γ_{ij}^2 of the vector $\boldsymbol{\gamma}^2$ that yield specific values of $\bar{\lambda}_R, \bar{\lambda}_C$, and $\bar{\delta}$. One method of achieving this has been developed by TSR (1989), based on the asymptotic form of $\mathbf{V}(\pi)$.
After some algebra, closed form expressions for $\bar{\lambda}_R$, $\bar{\lambda}_C$ and $\bar{\delta}$ can be derived that are linear in the γ_{ij}^2's. For given values of $\bar{\lambda}_R$ and $\bar{\lambda}_C$, a set of values of the γ_{ij}^2's corresponding to the maximum attainable value of $\bar{\delta}$ can then be found by linear programming (LP). Values of the γ_{ij}^2's corresponding to intermediate design settings of $\bar{\delta}(< \bar{\delta}_{\max})$ can be obtained by interpolating between sub-optimal "feasible basic" solutions to the LP. Details are given in TSR (1989). One drawback of this LP method is its failure to control variation in the individual design effects, as quantified by their coefficients of variation. Of particular concern is the value of $CV(\delta)$, which has a major effect on the performance of the test statistics. A method of selecting values of the γ_{ij}^2's corresponding to specific values of $\bar{\lambda}_R$, $\bar{\lambda}_C$, $\bar{\delta}$ and $CV(\delta)$ is being developed.

IV The Monte Carlo Study

1. The Statistics Examined

The following independence test statistics were included in the preliminary study. For details see Rao and Thomas (1988).

(1) The multinomial-based Pearson X^2 test;

(2) The multinomial-based log-likelihood G^2 test;

(3) A Wald test $X_W^2 = n\hat{\mathbf{h}}'\mathrm{Var}(\hat{\mathbf{h}})^{-1}\hat{\mathbf{h}}$, where $\hat{\mathbf{h}}$ has elements $(\hat{\pi}_{ij} - \hat{\pi}_{i\cdot}\hat{\pi}_{\cdot j})$.

(4) An F-based version of the above, defined as $F_w = [L-(r-1)(c-1)+1]\,/[(L-1)(r-1)(c-1)]X_w^2$, referred to an F distribution on $(r-1)\,(c-1)$ and $(L-(r-1)(c-1)+1)$ degrees of freedom.

(5) The first order Rao-Scott correction to X^2; denoted X_c^2 .

(6) An F-based version of the first order Rao-Scott correction, referred to an F distribution on $(r-1)(c-1)$ and $(r-1)(c-1)(L-1)$ degrees of freedom; denoted FX_c^2.

(7) A conservative version of the above referred to an F on $(L-1)$ denominator degrees of freedom (see Rao and Thomas, 1988); denoted $F^*X_c^2$.

(8) The first order Rao-Scott correction applied to G^2; denoted G_c^2.

(9) The second order Rao-Scott (Satterthwaite) correction applied to X^2; denoted X_S^2.

(10) The Singh (1985) $Q^{(T)}$ statistic, a stabilized version of the chi-square based Wald test, evaluated for several different values of ϵ, the eigenvalue cut off parameter; denoted $Q^{(T)}$.

(11) An F-based version of the above, defined as $FQ^{(T)} = [L-T+1]/[(L-1)T]Q^{(T)}$, referred to an F distribution on T and $L-T+1$ degrees of freedom.

(12) The Fay jackknifed procedure applied to X^2; denoted X_J^2.

(13) The Fay jackknifed procedure applied to G^2; denoted G_J^2.

2. Parameter Settings

This preliminary study concentrated on a single 3×3 table with cell probabilities,

$$\pi_R = (1/2, 1/3,\ 1/6)' \text{ and } \pi_C = (1/6,\ 1/3,\ 1/2)'.$$

A single experiment generated results for $L = 15$, 30, 50, 70 and 100 clusters, with $m = 20$ conditional multinomial draws per cluster. Empirical test significance levels were estimated based on 1000 Monte Carlo trials, for three nominal settings of 1%, 5% and 10%. The overall strategy followed that used by Thomas and Rao (1987); thus 100 clusters were first generated for each Monte Carlo iteration, each succeeding sample of L clusters then being a subset of the previous one. All test statistics were applied to the same data. Estimates of significance levels for different statistics for the same number of clusters will thus be highly positively correlated; estimates for the same statistic for different values of L should also be positively correlated, increasing the precision of the comparisons. For value of L, TSR (1989) give detailed results for four combinations of $\bar{\lambda}_R, \bar{\lambda}_C$, $\bar{\delta}$, for the case $\alpha = 5\%$, each combination exhibiting different degrees of variation among the δ_i's namely, $CV(\delta) = 0.28$, 0.40, 0.58, and 0.80. For lack of space, detailed results are given in Table 1 of this paper only for three values of $VC(\delta)$ and for two values of L, namely $L = 15$ and 50. Conclusions based on the full set of results will, however, be incorporated into the discussion of Table 1. Selected results for the different settings of α (1%, 5% and 10%) are shown in Table 2. Preliminary results on power are presented in Table 3.

3. Discussion of Table 1 and the Results in TSR (1989).

The three panels of Table 1 have $CV(\delta)$'s of .28, .40 and .81 respectively. For the first panel, $\bar{\delta} = 1.77$; $\bar{\delta} = 2$ for the remaining two. Values of $\bar{\delta}$ and $CV(\delta)$ in this range have been reported in the literature. For example, for a 2×4 table taken from the Canada Health Survey, Hidiroglou and Rao (1987) estimated $\bar{\delta} = 1.77$ and $CV(\bar{\delta}) = 0.47$. Several conclusions can be drawn from Table 1, and the results in TSR (1989).

(i) X^2 and G^2: Significance levels (SL) are severely inflated as expected. These procedures were included for reference only.

(ii) X_W^2: Even for large numbers of clusters ($L \geq 50$), significance levels are inflated. For small numbers of clusters, significance levels are approximately equal to $SL(X^2)$ and $SL(G^2)$. Again, this result was expected, based on empirical evidence from earlier studies.

(iii) F_W: The results confirm Thomas and Rao's (1987) findings for the goodness-of-fit test. Significance levels are much lower than for X_W^2, though still somewhat inflated for small L, and also for large $CV(\delta)$.

(iv) X_c^2, G_c^2, FX_c^2 and $F^*X_c^2$: Again these results are consistent with the findings of Thomas and Rao (1987). For small $CV(\delta)$, X_c^2 performs adequately; ($5.9 \leq$ SL $(X_c^2) \leq$ 6.8 for $CV(\delta) = .28$). There is little to choose between X^2 and G_c^2 for the sample sizes studied here. As $CV(\delta)$ increases, X_c^2 tends to become liberal, particularly for small numbers of clusters ($L = 15$) as can be seen from Table 1. Since $SL(FX_c^2) < SL(X_c^2)$ everywhere, the F-based version is the better of the two. $F^*X_c^2$ is excessively conservative for small L, when $CV(\delta)$ is small to moderate. (SL$(F^*X_c^2) = 2.7$ for $L = 15$, and $CV(\delta) = .28$). It is not recommended for use with tests of independence unless $L \geq 30$, except when $CV(\delta)$ is large ($\geq .50$).

(v) X_S^2: For small to moderate values of $CV(\delta)$, the second order Rao-Scott statistic yields SL's that are within two Monte Carlo standard errors ($\pm$ 1.4%) of the nominal 5% level (Table 1, and TSR, 1989). For large values of $CV(\delta), SL(X_S^2)$ tends to be somewhat liberal ($8.5 \leq$ SL $(X_s^2) \leq 9.8$ for $CV(\delta) = .81$). For $CV(\delta)$ as large as 0.58, X_S^2 performs adequately (TSR, 1989).

(vi) X_J^2 and G_J^2: There is little to choose between these two versions of Fay's procedure; the discussion will refer to X_J^2. For small to moderate values of $CV(\delta)$, X_J^2 provides good control for large numbers of clusters. As L decreases, $SL(X_J^2)$ increases, and for $L = 15$ it is definitely liberal as can be seen from Table 1. This characteristic of X_J^2 was previously noted by Thomas and Rao (1987) for the goodness-of-fit test.

(vii) $Q^{(T)}$: Significance levels are presented for a range of values of ϵ, the eigenvalue cutoff parameter. The largest values of ϵ yield the least inflated SL. For $\epsilon = .1$ and .05, SL's are generally inflated compared to the nominal $\alpha = 5\%$, and compared to the competing procedures FX_c^2, X_c^2 and X_J. For $\epsilon = .025$ and .01, the degree of SL inflation is unacceptable.

(viii) $FQ^{(T)}$: It is to be expected that better control of significance levels will be exhibited when eigenvalue trimming is applied to the F-based Wald statistic. This is confirmed by the results in Table 1 and TSR (1989). Within the range $.05 \leq \epsilon \leq .1$, $FQ^{(T)}$ yields good control of significance levels across all conditions studied. For $\epsilon = .025$ and .01, SL's are close to the nominal 5% level for $CV(\delta) \leq .4$, but exhibit some inflation for $CV(\delta) \geq .58$ (TSR, 1989, and Table 1). Conclusions regarding the merits of the FQ family of statistics relative to FX_c^2, X_s^2 and X_J must await a discussion of their relative powers.

4. Discussion of Table 2

Before considering power, it is interesting to see how the statistics that provide good SL control at $\alpha = 5\%$ perform at $\alpha = 1\%$ and $\alpha = 10\%$ (X^2 is included for comparison only). The two panels of Table 2 display this information for the minimum and maximum values of $CV(\delta)$ studied.

It can immediately be seen that as nominal α rates decrease, the relative level of control worsens rapidly. For example, $SL(X^2)$ is about ten times the nominal value when $\alpha = 1\%$, and about three times its nominal value when $\alpha = 10\%$. For $CV(\delta) = .28$, the various Rao-Scott procedures do somewhat better than both FX_W^2 and X_J^2, and about as well as $FQ^{(T)}$. However, for $CV(\delta) = .81$, all procedures except $FQ^{(T)}$ do relatively poorly (SL's from 2% - 4% when α =1%). There are two main conclusions to be drawn, in addition to those previously discussed in detail for the case α =5%.

(i) For α =10%, the procedures X_S^2 and X_J^2 which have been strongly recommended in other studies provide reasonable, though slightly liberal, control across a wide range of values of $CV(\delta)$. The $FQ^{(T)}$ statistic for $\epsilon = .05$ and does better, and appears to keep SL's very close to the nominal 10%.

(ii For α =1%, the only test statistic that maintains control at the nominal level for both values of $CV(\delta)$ is $FQ^{(T)}$. The best of the competition, other than $Q^{(T)}$, is $F^*X_C^2$, which exhibits SL's of 0.7 and 2.5% for $CV(\delta) = .25$ and .81, respectively.

For smaller values of ϵ, the degree of SL control exhibited by $FQ^{(T)}$ diminishes (results not shown). Unfortunately, as will be seen below, the power of $FQ^{(T)}$ is relatively low when $\epsilon \geq .05$. For this reason, modified versions of $FQ^{(T)}$ were included in which T was controlled at a minimum value T_0. In general, T_0 should be a high fraction (say 4/5) of the length of the vector $\hat{\mathbf{h}}$ in the definition of X_W^2. In the present study, $\hat{\mathbf{h}}$ has four elements and T_0 was set at 3; thus no more than one eigenvalue was deleted. The corresponding test was denoted $FQ^{(T)}$. Also, in order that ϵ might be kept as small as possible (in the interests of power), a combined test $F_W * FQ^{(T_0)}$ was considered which rejects H_0 only when both individual tests reject. Clearly, the combination test is more conservative than either test taken individually.

5. Discussion of Table 3.

A preliminary comparison of the powers of the potentially viable procedures (from the point of view of SL control) is displayed in Table 3. It involves only one setting of $\bar{\delta}$ and $CV(\delta)$, so that conclusions are tentative. SL values are given, together with empirical powers (percentage rejections) at a nominal α-level of 5%, for deviations from H_0 of $\rho = .02$ (with $L = 50$) and $\rho = .07$ (with $L = 15$). For a definition of ρ, see equation 3.1.1. Also included in Table 3 (in square parentheses) are estimated powers corrected for differences in SL's. These were obtained by graphing SL against power, for the three values of nominal α, and interopolating to estimate power at SL = 5%. These provide a rough but useful guide; final results will be based on a more accurate determination of empirical cut-offs. In the table, all SL's and powers are based on 1000 Monte Carlo intervals, except for F_W. In some cases, particularly when $L = 15$ and $\rho = .07$, F_W is not defined for all iterations due to the presence of zero $\hat{\pi}_{ij}$'s; for F_W, therfore, SL and power estimates refer only to cases where F_W was defined. For computing SL's and power for the combination test $F_W * FQ^{(T_0)}$, only $FQ^{(T_0)}$ was applied wherever F_W was not defined. For ease of interpretation, the discussion will focus on two broad classes of test statistic: (i) those based on Pearson's X^2, namely FX_c^2, $F^*X_c^2$, X_S^2 and X_J; (ii) those based on the Wald test X_w^2, namely F_W, $FQ^{(T_0)}$ and $F_W * FQ^{(T_0)}$. In the former class, all statistics show comparable power. However, X_S has the best performance overall in view of its better control of SL. In the latter class, F_w does not in general yield good SL control, while $FQ^{(T_0)}$ and the combination statistic (data not shown) do better, particularly as ϵ increases. Thus it is of interest to compare X_S^2 with $FQ^{(T_0)}$ (and with the combination test). For $\epsilon = .01$, $FQ^{T_0)}$ and X_S^2 has similar power on the average; on the other hand, SL's are somewhat inflated for $FQ^{(T_0)}$, while x_S^2 displays excellent control. For $\epsilon = .025$, $FQ^{(T_0)}$ displays good SL control and reasonable power, though in this case its power is lower than that of X_S^2 for both values of ρ studied. It is worth noting that the power of $FQ^{(T_0)}$ decreases rapidly as ϵ increases. In practice, therefore, $FQ^{(T_0)}$

should be used with the smaller values of ϵ. The combination test also appears to be potentially useful in controlling SL when ϵ is small (data not shown), and it is currently being investigated further.

6. Conclusion

Most of the conclusions drawn by Thomas and Rao (1987) on the basis of goodness-of-fit tests appear to hold for tests of independence in two way tables. Of course, only one 3×3 table has been examined, so the results are preliminary. The simulation results suggest that for testing independence in $R \times C$ tables, X_S^2, the second-order Rao-Scott connected X^2 test has the best performance of all test procedures examined. When $CV(\delta)$ is large, X_S^2 (and most of its competitors its competitors) is liberal and this situation is currently under further investigation. It should be noted that X_S^2 requires full information on $\hat{\mathbf{V}}$, the covariance matrix of proportions. Given only partial variance information, FX_c^2 can be used instead.

The X_S^2 procedure uses the full $\hat{V}$ to correct for the use of a working (multinominal) covariance matrix in X^2. The Wald test procedure X_w^2, on the other hand, uses the full covariance matrix directly, resulting in an asymptotically optimal test. All Wald-based procedures studied here attempt to stabilize this asymptotically optimal test, i.e. to combine its high asymptotic power with adequate SL control. Interestingly, the test X_S^2 appears to give the best-finite sample perfomance, even though it is not based on an optimal test. Nevertheless, the Wald procedure provides a flexible and widely applicable approach to the construction of optimal tests whenever a consistent estimate of the covariance matrix is available, and it is natural to seek modifications to Wald tests whenever their stability is in doubt. For the problem studied here, the modification F_w corrects to some extent; its power is good but its SL control is not fully satisfactory. The test $FQ^{(T_0)}$ is an improvement on F_w; if a Wald-based test is desired, it offers reasonable control of significance levels and power. Working rules for closing T_0 and ϵ are being investigated further for other two-dimensional tables, as well as for three-dimensional tables.

V. Acknowledgements

The work of D.R. Thomas was partially supported by a grant from the Natural Sciences and Engineering Research Council of Canada (NSERC) and by a contract from Statistics Canada. A.C. Singh was also partially supported by an NSERC grant held at Carleton University as an Adjunct Research Professor. The authors wish to thank Peter Found, a student at the University of Waterloo, for his important contribution to this project. While on a work-study assignment at the Social Survey Methods Division of Statistics Canada, he developed, in GAUSS, a prototype of the simulation package used in this study.

VI. References

Bedrick, E.J. (1983), Adjusted Chi-squared Tests for Cross-Classified Tables of Survey Data. *Biometrika, 70*, 591-596.

Brier, S.E. (1980). Analysis of Contingency Tables under Cluster Sampling. *Biometrika, 67*, 591-595.

Fay, R.E. (1979). On Adjusting the Pearson Chi-Square Statistic for Cluster Sampling. *Proceedings of the American Statistical Association*, Social Statistics Section, 402-405.

Fay, R.E. (1982). Contingency Table Analysis for Complex Sample Designs: CPLX. *Proceedings of the American Statistical Association, Survey Research Methods Section*, 44-52.

Fay, R.E. (1985). A Jackknifed Chi-squared Test for Complex Samples. *Journal of the American Statistical Association, 80*, 148-157.

Fay, R.E. (1987). Additional Evaluation of Chi-Square Methods for Complex Samples. Paper presented at the Annual Meeting of the American Statistical Association, San Francisco CA, August, 1987.

Fellegi, I.P. (1980). Approximate Tests of Independence and Goodness-of-Fit Based on Stratified Multistage Samples. *Journal of the American Statistical Association*, 71, 665-670.

Hidiroglou, M.A. and Rao, J.N.K. (1987). Chi-squared Tests with Categorical Data from Complex Surveys: Part I – II *Journal of Official Statistics, 3* 117-132; 133-140.

Koch, G.G., Freeman, D.H. and Freeman, J.L. (1975). Strategies in the Multivariate Analysis of Data from Complex Surveys. *International Statistical Review*, 93, 59-78.

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 Multi-way Tables with Cell Proportions Estimated from Survey Data. *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. *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.

Scott, A.J. and Rao, J.N.K. (1981). Chi-Squared Tests for Contingency Tables with Proportions Estimated from Survey Data. In *Current Topics in Survey Sampling* (D. Krewski, R. Platek and J.N.K. Rao Eds.), Academic Press, New York, 247-266.

Singh, A.C. (1985). On Optimal Asymptotic Tests for Analysis of Categorical Data From Sample Surveys. Methodology Branch. Working paper, No. SSMD 86-002, Statistics Canada.

Singh, A.C. and Kumar, S. (1986). Categorical Data Analysis for Complex Surveys. *Proceedings of the American Statistical Association, Survey Research Methods Section*, 252-257.

Thomas, D.R. and Rao, J.N.K. (1987). Small-Sample Comparisons of Level and Power for Simple Goodness-of-fit Statistics Under Cluster Sampling. *Journal of the American Statistical Association, 82*, 630-636.

Thomas, D.R., Singh, A.C. and Roberts, G. Size and Power of Independence Tests for $R \times C$ Tables from Complex Surveys: Design Issues and a Monte Carlo Study. Methodology Branch Working Paper, Statistics Canada, forthcoming.

Thomas, D.R. (1989). Simultaneous Confidence Interval Procedures for Proportions, Under Cluster Sampling. *Survey Methodology*, in press.

Wilson, J.R. (1986). A Simulated Comparison of Chi-Squared Tests for Comparing Vectors of Proportions for Several Cluster Samples. *Proceedings of the American Statistical Association, Survey Research Methods Section*, 246-251.

Table 1
Test Significance Levels for Nominal $\alpha = 5\%$

	$CV(\delta)$					
$\bar{\lambda}_R$	2.0		1.5		3.0	
$\bar{\lambda}_C$	1.5		2.13		1.5	
$\bar{\delta}$	1.77		2.0		2.0	
No. of Clusters	15	50	15	15	15	
Procedures						
X^2	23.2	23.9	26.0	30.4	19.8	24.1
X^2_W	22.7	8.6	22.1	10.4	26.0	12.6
F_W	7.5	6.0	9.0	7.4	10.9	10.0
X^2_c	6.8	6.2	8.1	6.9	11.5	10.0
FX^2_c	5.6	5.9	6.3	6.1	10.3	9.8
$F^*X^2_c$	2.7	4.7	2.2	4.7	6.6	9.2
X^2_S	4.6	4.9	4.8	4.6	9.2	9.3
X^2_J	7.5	6.5	9.5	7.4	11.9	8.8
$Q^{(T),\epsilon}$ $= .1$	9.5	7.1	9.1	7.3	9.1	6.7
$.05$	12.0	7.5	10.2	7.7	12.9	8.7
$.025$	15.4	8.6	13.0	8.9	18.3	12.4
$.01$	21.1	8.5	18.8	10.3	25.0	12.6
$FQ^{(T),\epsilon}$ $=.1$	3.8	5.5	5.2	5.4	4.9	5.3
$.05$	4.5	5.4	4.9	6.0	6.3	6.2
$.025$	4.7	6.0	5.1	6.5	7.9	9.8
$.01$	6.6	6.0	7.1	7.4	10.4	10.0

Table 2
Control of Test Significance Levels[1] as a Function of Nominal α-Level: $L = 30$, $\bar{\delta} = 2$.

	$CV(\delta)$					
	.40			.81		
	α-level			α-level		
Procedure	1%	5%	10%	1%	5%	10%
X^2	14.1	27.9	37.7	11.4	22.9	32.2
F_W	2.1	8.0	13.3	3.1	9.9	16.5
FX^2_c	1.5	5.8	10.9	3.8	9.6	15.6
$F^*X^2_c$	0.7	4.4	8.8	2.3	7.9	12.9
X^2_S	0.9	4.8	9.1	3.3	8.8	14.2
X_J	2.7	7.7	13.1	3.3	8.8	14.7
$Q^{(T),\epsilon}$ $=.1$	2.5	8.2	13.9	2.8	8.0	13.8
$.05$	3.3	9.3	15.3	3.8	9.9	16.5
$FQ^{(T),\epsilon}$ $=.1$	1.3	5.6	10.8	1.2	5.2	10.0
$.05$	1.3	5.8	11.5	1.3	6.0	11.2

[1] 4000 trials; 95% C.I. on 1%, ± 0.3%; on 5%, ± 0.7%

Table 3
Comparison of Power of the Viable Procedures;
Nominal $\alpha = 5\%$, $\bar{\delta} = 2.0$, $CV(\delta) = 0.40$

	Number of Clusters			
	L=15		L=50	
Procedure	SL	Power [1] ($\rho = .07$)	SL	Power[1] ($\rho = .02$)
F_w	9.0	99.8 [99][2]	7.4	50.6 [41]
FX^2_c	6.3	96.1 [95][1]	6.1	36.0 [32]
$F^*X^2_c$	2.2	91.5 [95]	4.7	32.6 [33]
X^2_s	4.8	92.7 [93]	4.6	30.9 [33]
X^2_J	9.5	99.4 [97]	7.4	40.1 [33]
$FQ^{(T_0)}$,				
$\epsilon = .05$	5.3	59.4 [59]	6.0	9.0 [8]
$\epsilon = .025$	5.1	62.7 [62]	6.5	29.2 [24]
$\epsilon = .01$	7.1	80.7 [77]	7.4	50.2 [41]
$F_W * FQ^{(T_0)}$				
$\epsilon = .05$	3.5	54.3	4.7	59.3
$\epsilon = .025$	4.2	62.6	6.1	62.6
$\epsilon = .01$	7.0	80.6	7.4	80.6

[1] First power estimate is the percentage of rejections for a nominal $\alpha = 5\%$; estimate in square parentheses is interpolated power corresponding to an empirical test level of $\alpha = 5\%$.

[2] In this case, F_W was computable only for 497 out of 1000 M.C. trials.