INDEPENDENCE TESTS FOR TWO-WAY TABLES UNDER CLUSTER SAMPLING

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

Key Words: Monte Carlo Study, Complex Surveys

1. INTRODUCTION

There is a well-documented problem of inflated significance levels if classical multinomial-based procedures are used when testing hypotheses on categorical data from complex surveys. Because of this problem, a variety of procedures have been developed specifically for survey data. However, the rationale for these procedures is generally based on asymptotic theory. It is thus important that there be thorough investigation of the finite population characteristics of these methods.

A few Monte Carlo studies have been reported in the literature. One of these, by Thomas and Rao (1985, 1987), looked at available procedures for a goodness of fit test under two-stage cluster sampling. Many people assumed that their results can be generalized to the test of independence in two-way tables. The first objective of this current study is to validate this assumption.

As well, a number of different procedures are now available for testing independence that were not considered in the goodness of fit study. The second objective of this study is to look at their finite sample properties.

A summary of the study is presented below. Full details are given in Thomas, Singh and Roberts (1995).

2. DESIGN REQUIREMENTS

2.1 Notation required

Consider an r × c contingency table, with π_{ij}, i=1,2,...r, j=1,2,...c being the individual cell probabilities, and $\pi_{i.}$, i=1,2,...r, and $\pi_{.j}$, j=1,2...c being the row and column marginal probabilities respectively. These probabilities may be represented in vector form as $\pi=(\pi_{11},...,\pi_{rc})'$, $\pi_R=(\pi_{1.},...\pi_{r.})'$, and $\pi_C=(\pi_{.1},...,\pi_{.c})'$.

The independence hypothesis can be expressed in two equivalent forms:
1) the residual form - H_0: $h_{ij}=\pi_{ij}-\pi_{i.}\pi_{.j}=0$
and
2) the loglinear form - H_0: $\ln(\pi_{ij})=\mu^*+\mu_{1(i)}+\mu_{2(j)}$.
The different formulations give rise to different statistics for testing the independence hypotheses.

The following three different sets of design effects, and functions of them, are relevant when examining the characteristics of the different test statistics:
1) $\lambda_{R(k)}$, $k=1,2,...(r-1)$, the eigenvalues of the design effect matrix $\boldsymbol{D}_R=\boldsymbol{P}_R^{(t)-1}\boldsymbol{V}_R^{(t)}$ arising from the test of goodness of fit on the row marginals π_R, where $\boldsymbol{V}_R$ denotes the covariance matrix of consistent estimates of the row marginals, $\boldsymbol{P}_R$ denotes the corresponding multinomial covariance matrix, and the superscript (t) denotes a trimmed matrix, obtained by deleting the last row and column of the matrix in question. The mean of the $\lambda_{R(k)}$ will be denoted by $\bar{\lambda}_R$.
2) $\lambda_{C(k)}$, $k=1,2,...(c-1)$, the eigenvalues of the design effect matrix arising from the test of goodness of fit on the column marginals π_C. The mean of the $\lambda_{C(k)}$ will be denoted by $\bar{\lambda}_C$.
3) δ_k, $k=1,2,...(r-1)(c-1)$, the eigenvalues of the generalized design effect matrix $\boldsymbol{D}_I$ corresponding to the test of independence. $\boldsymbol{D}_I$ can be expressed in the form $\boldsymbol{D}_I=n(Z'D_\pi^{-1}Z)^{-1}(Z'D_\pi^{-1}V(\hat{\pi})D_\pi^{-1}Z)$, where n is the size of sample taken, Z is the completion of the design matrix for the independence form of the loglinear model, $V(\hat{\pi})$ is the covariance matrix of a consistent estimate of π, and D_π is a diagonal matrix with the elements $\pi_{ij}=\pi_{i.}\pi_{.j}$ on its diagonal.
The mean of the δ_k will be denoted by $\bar{\delta}$. A measure of the variation among the δ_k will be denoted by $a(\delta)$, which is defined as $a(\delta)=[\sum_{1}^{\nu}\delta_i^2/\nu\bar{\delta}^2 - 1]^{1/2}$, with $\nu=(r-1)(c-1)$.

2.2 Model requirements

The model must be a plausible representation of two-stage sampling, and be capable of:
(I) modelling different row and column design effects, ie. $\bar{\lambda}_R\neq\bar{\lambda}_C$.
(ii) modelling a range of values of $\bar{\delta}$ for given values of $\bar{\lambda}_R$ and $\bar{\lambda}_C$.
(iii) modelling unequal design effects, ie. 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 $a(\delta)$ over a range of values of $\bar{\lambda}_R, \bar{\lambda}_C$, and $\bar{\delta}$.
(v) modelling patterns of marginal probabilities other than the equiprobable case, $\pi_{i.}=1/r$, $\pi_{.j}=1/c$, $\forall i,j$.
(vi) modelling deviations from $H_0:\pi_{ij}=\pi_{i.}\pi_{.j}$, so that the powers of the competing procedures can be assessed.

3. GENERATING CLUSTERED DATA

Several models of two-stage cluster sampling were considered for this study, but those found in the literature seemed intractable to use for satisfying the modelling constraints described in 2.2 above. A new model, based on a "modified logistic normal" distribution, was therefore developed for use in this study.

Given cell probabilities π_{ij}, $i=1,...r$, $j=1,...c$, the modified logistic normal (MLN) model generates (rc × 1) vectors of non-integer pseudo-counts m_k that satisfy $E(m_k)=m\pi$. The cell probabilities can be generated either under the independence hypothesis from preset marginals π_R *and* π_C, or, if deviations from independence are to be simulated, using the Bahadur representation
$\pi_{ij}=\pi_{i.}\pi_{.j}+\rho_{ij}[\pi_{i.}(1-\pi_{i.})]^{1/2}[\pi_{.j}(1-\pi_{.j})]^{1/2}$, with $-1\le\rho_{ij}\le 1$.

An estimator $\hat{\pi}$ of π was found that is consistent as $L\to\infty$ and that exhibits the variance inflation that is characteristic of two-stage clustering. As well, a convenient expression for the asymptotic variance **V** of $\hat{\pi}$ was developed which was used to design the experiment and define the required parameter settings. A consistent estimator for **V** was also found.

As stated in 2.2 above, the model must allow for experimental control of the various parameters. It was found that these parameters cannot be freely selected. Thus, a method based on the ideas of linear programming was used to determine an envelope of admissible values of $\bar{\delta}$ *and* $a(\delta)$ for selected values of the marginal probabilities, $\bar{\lambda}_R$ *and* $\bar{\lambda}_C$.

4. THE MONTE CARLO STUDY

4.1 The Test Statistics Examined

The following statistics for testing row-column independence were examined in the study. More complete descriptions are given in the full paper.

1) The Pearson X^2 and the loglikelihood G^2 tests

2) First-order Rao-Scott corrections to X^2 *and* G^2, denoted by X_c^2 *and* G_c^2

3) An F-based version and conservative F-based version of X_c^2 and an F-based version of G_c^2, denoted respectively FX_c^2, $F^*X_c^2$, *and* FG_c^2

4) The second-order Rao-Scott correction to X^2 *and* G^2, denoted by X_S^2 *and* G_S^2 respectively

5) F-based versions of X_S^2 *and* G_S^2, denoted FX_S^2 *and* FG_S^2 respectively

6) A corrected X^2 test due to Fellegi, denoted X_F^2

7) Fay jackknifed procedures applied to X^2 *and* G^2, denoted X_J^2 *and* G_J^2 (Fay, 1985)

8)Two Bonferroni procedures: Bf(R) consisting of (r-1)(c-1) simultaneous tests on cell residuals and Bf(LL) consisting of the same number of simultaneous tests constructed from a loglinear representation of the independence hypothesis

9) Two Wald procedures: $X_W^2(R)$ being based on cell residuals and $X_W^2(LL)$ being based on a loglinear representation of the independence hypothesis

10) F-based versions of the two Wald procedures described in 9), denoted $F_W(R)$ *and* $F_W(LL)$ respectively

11) A modified Wald test based on an heuristic suggested by Morel

12) Singh's $Q^{(T)}$ procedure applied to the Wald procedures described in 9), denoted $Q^{(T)}(R)$ *and* $Q^{(T)}(LL)$ respectively (Singh, 1985)

13) F-based versions of the 2 procedures described in 12), denoted $FQ^{(T)}(R)$ *and* $FQ^{(T)}(LL)$ respectively

14) An alternative to Singh's stabilization of the Wald test, called EV1, that uses a ridge-type adjustment of the eigenvalues. Four versions of the procedure are explored, using both chi-squared and F-based forms of both the residual and loglinear Wald statistics, and are denoted $X_W^2(R, EV1)$, $X_W^2(LL, EV1)$, $F_W^2(R, EV1)$, and $F_W^2(LL, EV1)$

15) A second eigenvalue adjustment procedure, denoted EV2, that is based on a logarithmic transformation of the eigenvalues. It also results in four distinct versions, denoted in the same way as in 14), except with "EV1" replaced by "EV2".

4.2 Parameter Settings for 3 × 3 Table

The major part of the study assessed the relationship of the parameters $\bar{\lambda}_R$, $\bar{\lambda}_C$, $\bar{\delta}$, *and* $a(\delta)$ to the performance of the selected test statistics using a 3 × 3 table with marginal probabilities

$$\pi_R = \pi_C = (1/2,1/3,1/6)'.$$

For this table, separate Monte Carlo experiments were carried out for each of the 52 admissible parameter combinations shown in Table 1. Each experiment consisted of 4000 Monte Carlo trials, and generated results for L=100, 70, 50, 30 and 15 clusters, each with 20 conditional multinomial draws per cluster. The cluster data were generated using a strategy similar to that reported by Thomas and Rao (1987); in particular, 100 independent clusters were first generated for each Monte Carlo trial and then each succeeding sample of L clusters was selected as a subset of the previous one.

For the study of Type I error control, data were generated to conform to the null hypothesis of independence. For the study of power, data were generated under an alternative hypothesis, using cell probabilities obtained from the Bahadur representation given in 3 above, with $\rho_{ij}=\rho=.03$.

4.3 Other Parameter Settings Examined

Those test statistics exhibiting good Type I error control and acceptable power for the 3 × 3 table were tested on a 3 × 3, a 3 × 4, and a 4 × 4 contingency table using a single combination of study parameters given by $\bar{\lambda}_R=\bar{\lambda}_C=\bar{\delta}=2.0$, $a(\delta)=0.75$, *and* $m=30$. Probabilities for three-category margins were the same as for the main experiment. For four-category margins, probabilities were set equal to

$\pi_R=\pi_C=(1/3,5/18,2/9,1/6)'$.

To compare test powers across the 3 tables, cell probabilities under the alternative hypothesis were obtained using $\rho=0.02$, 0.0125, *and* 0.01.

4.4 Empirical Measures Used

The Type I error control and power of each test statistic were assessed using empirical significance levels (ESL's) and empirical powers (EP's) respectively. These are the proportions of Monte Carlo trials leading to rejection of the independence hypothesis, when the data are generated under the null hypothesis and under the alternative, respectively. Both ESL's and EP's were recorded for nominal test levels ranging from $\alpha=0.1\%$ to $\alpha=10\%$. Because all test statistics were calculated on the same data and because of the subsetting to obtain different cluster counts, ESL's and EP's for the same statistic across different values of L are therefore positively correlated, as are ESL's and EP's for different statistics measured using the same number of clusters.

4.5 Approaches Used to Summarize Results of Type I Error Control for 3 × 3 Tables

(I) Regression summary:

As an aid to summarizing and understanding the results of the Monte Carlo experiment, the regression equation

$$ESL=\beta_0 + \beta_1\left(\frac{(\bar{\lambda}_R+\bar{\lambda}_C)}{2} - 1\right) + \beta_2\left|\frac{\bar{\lambda}_R-\bar{\lambda}_C}{2}\right| + \beta_3(\bar{\delta} - 1) + \beta_4 a(\delta) + \beta_5 a^2(\delta) + \beta_6 L + \beta_7/L + error$$

was fitted by ordinary least squares to the ESL's corresponding to a nominal test level of 5%. A separate equation was fit to the 260 observations for each test statistic obtained from the 52 parameter combinations and the 5 values of L. As a result of the study design, ESL's for a given test statistic are independent across different combinations of the parameters $\bar{\lambda}_R$, $\bar{\lambda}_C$, $\bar{\delta}$ *and* $a(\delta)$, but are correlated across the 5 values of L used with each parameter combination. Thus, observations of ESL for each parameter combination form independent clusters of 5 correlated observations, with the result that OLS estimates of the standard errors of the regression parameters are biased. Appropriate standard error estimates were obtained using PC-CARP under the assumption of equally weighted independent clusters.

The following different techniques were among those then used to interpret the regression equations:

1) The standard expression for R^2 was used as a measure of fit of the equations.

2) Different test procedures were compared by examining fitted significance levels for particular values of the independent variables.

3) The joint effect of subsets of the parameters was studied by evaluating the minimum and maximum of their joint contribution to the regression function over all combinations of the parameters

II) Overall comparison of ESL's

While the fitted regression equations provide an overview of the extent to which a given test procedure controls Type I error and the extent to which the ESL is affected by the various study parameters, and also reveal important differences among major classes of test procedures, a detailed comparative assessment of individual procedures requires a direct examination of the ESL's themselves. This was done by calculating and examining average ESL's over all combinations of

$\bar{\lambda}_R$, $\bar{\lambda}_C$ *and* $\bar{\delta}$ for different values of L and $a(\delta)$.

III) Comparison of ESL's over α Values
For those test statistics that appear to provide adequate control of significance levels at α=5%, based on the methods described above, a comparison of ESL's at α=1% and α=10% was done, for a selection of values of the study parameters.

4.6 Approaches Used to do Power Comparisons

For a statistic to be considered viable, it must provide good control of Type I error over a wide range of conditions and also must provide adequate power under the conditions likely to be encountered in practice. To assess the power of those test statistics appearing to have good Type I error control, EP's averaged over the 52 parameter combinations for each cluster count were compared. As well, for selected parameter settings, EP's adjusted for Type I error were compared, in recognition of the fact that statistics that are susceptible to Type I error inflation will, in the non-null case, have an apparent power advantage.

5. STUDY RESULTS: CONTROL OF TYPE I ERROR AND POWER

A summary of the Type I error control of the better procedures is shown in Table 2, for the three sizes of contingency tables explored. The following general comments may be made on the basis of Table 2 and of the other results given in the full report.

1) The regression equation described in 4.4(I) is a useful summary of the relation between the ESL and the parameters under study, with R^2 being well above 70% for most procedures. The one procedure with a low R^2 - Bf(LL) - has this characteristic because it maintains excellent control of Type I error at the 5% level, so that there is relatively little variation around the mean to explain.

2) The ESL's of most test procedures are not highly sensitive to different values of $\bar{\lambda}_R$, $\bar{\lambda}_C$, *and* $\bar{\delta}$.

3) The ESL's for most test procedures increase with increased variability in the cell deffs, as measured by $a(\delta)$.

4) Test procedures vary in their degree of sensitivity to $a(\delta)$ and number of clusters.

5) For those test procedures with both a residual version and a loglinear version, the loglinear versions generally have better Type I error control.

6) Do not use X^2, G^2, $X^2_W(R)$, *or* $X^2_W(LL)$, since all have wildly inflated Type I error rates over a wide range of conditions.

7) Do not perform a test of independence, using any of the test procedures studied , at nominal levels lower than 5%, since the control of Type I error seems very poor in that range.

8) Only a limited exploration of the power properties of the test statistics has been done to date. However, it is already obvious that, for those procedures with both a residual and loglinear version, the loglinear version has much better power.

9) The loglinear Bonferroni procedure and the eigenvalue adjusted procedures outperform their competitors with respect to Type I error for all three table sizes considered.

10)Similarly, the loglinear Bonferroni procedure is the most powerful procedure for all 3 contingency tables and all values of L.

6. CONCLUSIONS AND RECOMMENDATIONS

One major aim of this study was to determine whether or not Thomas and Rao's (1987) conclusions regarding goodness-of-fit tests can be extended to tests of independence in two-way tables, under cluster sampling. Their main conclusions do apply, although there are some minor differences, such as FX^2_S outperforming X^2_S with respect to Type I error control - the opposite to which was found for the goodness-of-fit test.

The second objective of this study was to examine in detail several families of test procedures not considered by Thomas and Rao. The most striking conclusion of this phase of the study was the success of the Bonferroni procedure Bf(LL), which provided the best control of Type I error and the highest power across the full range of parameters under study. The second best from the point of view of Type I error control was the procedure $F_W(LL,EV1)$, while the runner-up in terms of power was the Singh procedure $FQ^{(T)}(LL)$; however, both procedures have the disadvantage of depending on external parameters that cannot be estimated from the data.

In recommending a test procedure to a practitioner, several issues must be considered. Purely on the basis of

Type I error control and power, the Bonferroni procedure Bf(LL) would be the method of choice. However, some practitioners might prefer a more familiar test procedure, and in addition, might be reluctant to use other test statistics that depend on preselected constants. They would therefore choose among the procedures previously considered by Thomas and Rao (1987). The results of the current study show that the second order Rao-Scott procedure FX_S^2 provides reasonable control of Type I error and adequate power, and is thus a viable choice. There is some further evidence that its control of Type I error improves with table size, which is an attractive characteristic. It should also be noted that when complete survey information is not available, as is frequently the case in practice, practitioners would be forced to use one of the two F-based first order Rao-Scott procedures which require information only on cell and marginal design effects. It has been shown in this study that the conservative variant $F^*X_c^2$ can be useful whenever large variations in design effects are to be expected.

Provided the number of clusters is greater than thirty, Fay's jackknifed test G_J remains a competitor whenever full survey design information is available, and might be regarded as a natural procedure to be used when survey inference is tied to a replication strategy. Practitioners who are accustomed to using Wald-based procedures can improve on the standard Wald test by using a loglinear F-based procedure, or if they are comfortable with tests that depend on preset parameters they can select either $FQ^{(T)}(LL)$ or $F_W(LL; EVI)$. The former is the more powerful of the two, whereas the latter provides excellent control of Type I error, and good power. It should also be noted that an attractive feature of $F_W(LL; EVI)$ is that it can be implemented with only a minor modification to existing software, since it is a simple multiple of the parent Wald statistic.

7. REFERENCES

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

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

Thomas, D.R. and Rao, J.N.K. (1985). On the Power of Some Goodness-of-fit Tests Under Cluster Sampling. In *Analysis of Categorical Data From Sample Surveys: A Collection of Five Papers*, (Technical Report 66), Ottawa, Canada: Carleton University / University of Ottawa, Laboratory for Research in Probability and Statistics, 57-82.

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.R. (1995) Tests of Independence on Two-Way Tables under Cluster Sampling: An Evaluation. Carleton University School of Business Working Paper Series No. 95-04; Ottawa, Canada.

Table 1

Test Values of $\bar{\lambda}_R$, $\bar{\lambda}_C$, $\bar{\delta}$ and $a(\delta)$ for the Monte Carlo Experiment on 3 × 3 Table

$\bar{\lambda}_R$	$\bar{\lambda}_C$	$\bar{\delta}$	$a(\delta)$
1.5	1.5 (0.25)[1] 2.25	1.75	0.5
1.5	3.0	2.0	0.8
2.0	2.25	1.75 (0.25) 2.25	0.5
2.0	2.25 (0.25) 3.0	2.5	0.3 (0.1) 0.7
2.0	3.0	3.0	0.7 (0.1) 1.0
2.5	3.0	2.5 (0.25) 3.25	0.3 (0.1) 0.7

(1) To be read as: 1.5 through 2.25, in increments of 0.25.

Table 2

Empirical Significance Levels[1] as a Function of Table Dimension, for $\alpha = 5\%$; $\bar{\lambda}_R = \bar{\lambda}_C = \bar{\delta} = 2.0, a(\delta) = 0.75$.

		Table Dimension									
		3 × 3			3 × 4			4 × 4			
		L			L			L			
Test		15	30	70	15	30	70	15	30	70	RMS (5%)[2]
X_c^2		12.4	10.3	9.5	10.6	9.2	9.3	11.1	9.3	8.3	5.1
FX_c^2		10.9	9.6	9.3	8.9	8.5	8.9	9.8	8.7	8.0	4.3
$F^*X_c^2$		6.5	7.4	8.4	3.5	5.8	7.9	2.6	5.2	6.4	2.1
X_S^2		9.5	8.1	7.6	6.4	6.3	6.9	5.6	5.8	5.5	2.2
FX_S^2		7.8	7.4	7.3	4.8	5.8	6.6	4.3	5.3	5.4	1.6
G_J		9.5	7.3	6.0	9.1	7.1	6.5	8.8	7.4	5.8	2.8
$Bf(LL)$		5.8	5.2	5.0	4.9	4.6	4.9	4.5	5.5	4.9	0.4
$F_W(R)$		9.1	8.2	6.7	8.8	9.0	8.0	8.6	10.8	9.1	3.8
$F_W(LL)$		6.5	6.9	6.0	5.6	6.9	6.6	6.3	7.5	7.1	1.7
$X_W^2(LL;M)$		7.9	8.3	6.7	3.1	7.1	7.5	0.1	4.1	6.6	2.7
$FQ^{(T)}(LL)$;	$\epsilon = .1$	3.6	4.3	5.1	2.2	4.0	5.5	1.1	3.4	3.9	1.8
	$\epsilon = .05$	3.9	5.1	5.5	3.2	4.5	5.7	1.3	4.4	5.5	1.5
	$\epsilon = .02$	5.8	6.8	6.0	3.1	5.5	5.9	1.4	5.2	6.1	1.6
$F_W(R;EV1)$;	$k = 1$	4.0	4.8	4.4	3.9	4.6	4.9	4.2	5.2	4.7	0.6
	$k = .5$	5.8	5.9	5.3	5.5	6.2	6.4	6.1	7.2	6.4	1.2
$F_W(LL;EV1)$;	$k = 1$	3.1	3.4	3.6	2.5	3.2	3.9	2.7	3.2	3.2	1.8
	$k = .5$	4.3	5.0	4.4	3.8	4.6	5.3	4.1	4.7	4.6	0.6
	$k = .25$	5.0	5.8	5.1	4.7	5.5	5.9	5.1	6.0	5.7	0.6
$F_W(LL;EV2)$;	$k = 1$	2.6	3.2	3.5	2.1	3.0	3.7	2.4	2.8	3.6	2.1
	$k = .5$	4.0	4.9	4.4	3.5	4.4	5.2	3.8	4.6	4.5	0.8
	$k = .25$	5.0	5.7	5.0	4.6	5.5	5.9	5.0	6.0	5.7	0.6

(1) Monte Carlo standard error for ESL's of magnitudes 2.5%, 5%, and 10% are 0.25%, 0.35% and 0.47%, respectively.

(2) Pooled root mean square deviation from the nominal 5% level.