A MONTE CARLO STUDY OF EXACT LEVELS OF GOODNESS-OF-FIT STATISTICS UNDER CLUSTER SAMPLING

D. Roland Thomas and J.N.K. Rao, Carleton University

Introduction

The consequences of ignoring the effects of the design, and analysing survey data as if they arise from a simple random sample are now well known. For categorical data analysis the standard Pearson chi-squared tests (X^2) and likelihood ratio tests (G^2) can yield unacceptably large significance levels under cluster sampling. A number of alternatives that take account of the design have been developed. Weighted least squares methods based on the Wald Statistic (Koch, Freeman and Freeman, 1975) have been extensively used by several survey organizations, and computer software based on this approach is available. Fay (1979) proposed jackknifed chi-squared test statistics based on a replication strategy, and has applied this approach to hierarchical log-linear models via a computer program CPLX (Fay, 1983a).

Two alternative test statistics for categorical data have been proposed by Rao and Scott (1981, 1984), based on an asymptotic analysis of the distribution of X^2 and G^2. The first statistic, X_C^2 or G_C^2, was designed for use with published tables, for which neither the full covariance matrix required for the Wald Statistic, nor the detailed replicate level data required by Fay's method, are generally available. In the k-category goodness of fit case, for example, this method requires knowledge only of the variance (or design effects) of the k cell estimates, instead of the full covariance matrix. The second statistic, X_S^2 or G_S^2, uses a Satterthwaite approximation to the asymptotic distribution of X^2 and G^2, but requires knowledge of the full covariance matrix.

In this paper, the finite sample relative performance of the above test statistics is assessed, under simulated cluster sampling. The Monte Carlo study is confined to the simple goodness-of-fit problem.

The Cluster Sampling Model

We will consider two-stage cluster sampling in which a k-category sample of m units is drawn independently from each of r sample clusters. Let $\underset{\sim}{m}_\ell = (m_{\ell 1},\ldots,m_{\ell,k-1})'$ represent the vector of category counts for the ℓ^{th} cluster, $\ell = 1,\ldots,r$, and let $\underset{\sim}{m} = (m_1,\ldots,m_{k-1})'$ represent the category counts for the whole sample. The total number of observations in the sample is thus $n = mr = \sum_{\ell=1}^{r} m_\ell$. Further, let $\hat{\underset{\sim}{\pi}} = (\hat{\pi}_1,\ldots,\hat{\pi}_{k-1})' = \underset{\sim}{m}/n$ be the vector of cell proportions in the sample, and define $\underset{\sim}{\pi} = E(\hat{\underset{\sim}{\pi}})$, where E denotes expectation under the model, yet to be defined. Similarly, let $\underset{\sim}{V}/n$ represent the $(k-1)\times(k-1)$ covariance matrix of $\hat{\underset{\sim}{\pi}}$.

For a goodness of fit hypothesis $H_o: \underset{\sim}{\pi} = \underset{\sim}{\pi}_0$ on the model vector $\underset{\sim}{\pi}$, Rao and Scott (1981) showed that the Pearson X^2 test (hence also G^2) has the asymptotic null distribution $\lambda_1 Z_1^2 + \ldots + \lambda_{k-1} Z_{k-1}^2$, where the Z_i^2 are independent, one degree of freedom chi-squared random variables, and the λ_i's are eigenvalues of the "generalized design effect" matrix $\underset{\sim}{\Delta}^{-1}\underset{\sim}{V}$. Here $\underset{\sim}{\Delta} = \mathrm{diag}(\underset{\sim}{\pi}) - \underset{\sim}{\pi}\underset{\sim}{\pi}'$. When all the λ_i are equal to unity, X^2 recovers the traditional chi-squared distribution on k-1 degrees of freedom. Clearly then, our model must be capable of generating alternative patterns of eigenvalues.

Brier (1978) proposed a model for cluster sampling in which second stage sampling within each cluster was conditionally multinomial, based on probability vectors $\underset{\sim}{p}_\ell$, with the first stage $\underset{\sim}{p}_\ell$'s being sampled independently from a Dirichlet distribution having parameters ν (> 0) and $\underset{\sim}{\pi}$. Under this model all the eigenvalues λ_i are equal to $(m+\nu)/(1+\nu)$. This model mimics only one aspect of cluster sampling, namely $\bar{\lambda} > 1$, so we need to extend Brier's model to generate non-constant design effects. An appropriate extension is obtained by drawing $\underset{\sim}{p}_\ell$, the multinomial probability vector for the ℓ^{th} cluster, from a mixture of L Dirichlet distributions, having parameters $(\nu, \underset{\sim}{\pi}_j)$, $j=1,\ldots,L$. Under this model, we have $\underset{\sim}{\pi} = \Sigma\beta_j\underset{\sim}{\pi}_j$, and

$$\underset{\sim}{\Delta}^{-1}\underset{\sim}{V} = \left(\frac{m+\nu}{1+\nu}\right)I + \frac{(m-1)\nu}{(1+\nu)} \cdot \sum_{j=1}^{L} \beta_j \underset{\sim}{\Delta}^{-1}(\underset{\sim}{\pi}_j - \underset{\sim}{\pi})(\underset{\sim}{\pi}_j - \underset{\sim}{\pi})' \qquad (1)$$

where the mixture weights β_j and Dirichlet parameter vectors $\underset{\sim}{\pi}_j$ satisfy

$$\Sigma\,\beta_j = 1, \quad \text{and} \quad \Sigma\,\beta_j\underset{\sim}{\pi}_j = \underset{\sim}{\pi}. \qquad (2)$$

When $L = 2$, this model yields one distinct and k-2 equal eigenvalues, which can be explicitly evaluated, giving (Rao and Scott, 1979):

$$\bar{\lambda} = \sum_{i=1}^{k-1} \lambda_i/(k-1) = \frac{m+\nu}{1+\nu} + \frac{(m-1)\nu\delta}{(k-1)(1+\nu)} \qquad (3)$$

$$a = \frac{\left(\sum_{i=1}^{k-1}(\lambda_i - \bar{\lambda})^2\right)^{\frac{1}{2}}}{\bar{\lambda}(k-1)^{\frac{1}{2}}} = \frac{(k-2)^{\frac{1}{2}}(m-1)\nu\delta}{[(k-1)(m+\nu)+(m-1)\nu\delta]}, \qquad (4)$$

where $\delta = \beta_1\beta_2(\underset{\sim}{\pi}_1 - \underset{\sim}{\pi}_2)'\Delta^{-1}(\underset{\sim}{\pi}_1 - \underset{\sim}{\pi}_2)$, $0 \le \delta \le 1$, and a is the coefficient of variation of the λ's. Thus non-zero values of a can be modelled, and this mixture model with $L = 2$ has been adopted for the study. Though other, possibly more realistic distributions of eigenvalues, could be modelled using $L > 2$, the adopted model has the advantage of simplicity. It can generate suitably large values of a for fixed values of $\bar{\lambda}$, and can hence be used to simulate the behaviour of the various test statistics over a wide range of conditions, from multinomial through highly non-homogeneous clustering.

Design of the Monte Carlo Study

The Parameters

The parameters to be controlled are: (1) α, the nominal significance level for the tests; (2) $\underset{\sim}{\pi}$, the model probability vector; (3) k, the number of categories; (4) r, the number of independent clusters; (5) m, the (constant) number of units drawn per cluster; (6) $\bar{\lambda}$, the mean of the eigenvalues of $\underset{\sim}{\Delta}^{-1}\underset{\sim}{V}$, the generalized design effect matrix; (7) a, the coefficient of variation of the eigenvalues.

From equations (2), (3) and (4), it can be seen that, for fixed values of k and m, the parameters $\underset{\sim}{\pi}$, $\bar{\lambda}$ and a are functions only of ν, β_1 and $\underset{\sim}{\pi}_1$. The latter parameters are not controlled in the study, but are varied to provide the desired combinations of values of the controlled parameters. Given the large number of these, it is not feasible to examine a complete factorial set of combinations. Thus the bulk of the Monte Carlo simulation has been carried out for one value of $\bar{\lambda}$ ($\bar{\lambda} = 2$) under the equiprobable case $\underset{\sim}{\pi} = (1/k,\ldots,1/k)'$.

Generation of Random Numbers

Brier's (1978) method of generating Dirichlet variates from k-1 beta random variables was used. The betas were generated using subroutine GGBTR (IMSL, 1980), while the required source of uniforms was supplied by the generator GOFCAF (NAG, 1983), a multiplicative congruential generator of modulus 2^{59}. For each of the 1000 Monte-Carlo trials, independent Dirichlet k-vectors were generated for fifty clusters. Then, for each cluster, a k-category conditional multinomial sample was constructed by referring each of m independent (0,1) uniforms to the appropriate interval associated with p_ℓ. For given values of m, k, $\underset{\sim}{\pi}$, $\bar{\lambda}$ and a, all test statistics under consideration were then applied to the same subset of r independent sampled clusters, thus increasing the precision of comparisons between different test procedures at the same parameter settings (Schruben and Margolin, 1978; Olson, 1974).

The precision of comparisons between the same test procedures at different settings was also increased by a synchronized reuse of the basic set of uniform random numbers. For each of the 1000 sets of 50 clusters, all test procedures were applied in turn to the $1000 \times r$ array of clusters, for r = 5, 10, 15, 20, 30 and 50. Thus, test statistics for two different numbers of clusters r_1 and r_2 were correlated by having the $\min(r_1,r_2)$ clusters in common. For different values of k and m, correlations were induced by re-using distinct streams of uniforms for each Dirichlet vector and for each of the sets of uniforms used to generate the conditional multinomial distributions, for each of the 1000 Monte Carlo trials.

Test Statistics

(1) X^2 and G^2 Statistics. The test procedure refers

$$X^2 = n\sum_{i=1}^{k}(\hat{\pi}_i-\pi_{0i})^2/\pi_{0i}, \text{ or } G^2 = 2n\sum_{i=1}^{k}\hat{\pi}_i\log(\hat{\pi}_i/\pi_{0i}),$$

to χ^2_{k-1}, a chi-square variable with k-1 d.f.

(2) The Wald Statistic: The test procedure refers

$$X^2_W = n(\hat{\underset{\sim}{\pi}}-\underset{\sim}{\pi}_0)'\,\hat{\underset{\sim}{V}}^{-1}(\hat{\underset{\sim}{\pi}}-\underset{\sim}{\pi}_0)$$

to χ^2_{k-1}, where $\hat{\underset{\sim}{V}}$ is given by

$$\hat{\underset{\sim}{V}} = \frac{1}{m(r-1)}\sum_{\ell=1}^{r}(\underset{\sim}{m}_\ell - \frac{1}{r}\underset{\sim}{m})(\underset{\sim}{m}_\ell - \frac{1}{r}\underset{\sim}{m})'\ .$$

An alternative test is obtained by referring

$$F_W = \frac{(r-k+1)}{(k-1)(r-1)}X^2_W \quad \text{to} \quad F_{(k-1),(r-k+1)}\ ,$$

(Fellegi, 1980; Hidiroglou et al., 1980).

(3) Fay's X_J and G_j Statistics. Fay's (1979) version of X^2 is defined in terms of the following quantities: $\hat{\underset{\sim}{\pi}}(-\ell) = r(r-1)^{-1}(\hat{\underset{\sim}{\pi}} - n^{-1}\underset{\sim}{m}_\ell)$, $Q^2(-\ell) = \Sigma(\hat{\pi}_i(-\ell)-\pi_{0i})^2/\pi_{0i}$, $P(\ell) = n(Q^2(-\ell)-Q^2)$, where $Q^2 = X^2/n$. Then, the jackknife statistic X_J is given by

$$X_J = \frac{(X^2)^{\frac{1}{2}} - (K_J)^{\frac{1}{2}}}{(V_J/8X^2)^{\frac{1}{2}}}\ ,$$

where $K_J = r^{-1}(r-1)\Sigma P(\ell)$ and $V_J = r^{-1}(r-1)\cdot\Sigma P^2(\ell)$. The jackknife version of G^2, denoted G_J, is defined in an entirely analogous way. Both X_J and G_J are referred to the critical points of $\sqrt{2}[(\chi^2_{k-1})^{\frac{1}{2}} - (k-1)^{\frac{1}{2}}]$.

(4) Rao and Scott's $\bar{\lambda}$ Corrections. The method refers

$$X^2_c = X^2/\bar{\hat{\lambda}}\ , \text{ or } G^2_c = G^2/\bar{\hat{\lambda}}$$

to χ^2_{k-1}, where $\bar{\hat{\lambda}} = (k-1)^{-1}\Sigma(1-\hat{\pi}_i)\hat{d}_i$, $\hat{d}_i = \hat{v}_{ii}/\hat{\pi}_i(1-\hat{\pi}_i)$ is the i^{th} estimated cell design effect, and $\hat{v}_{ii}$ is the i^{th} diagonal element of $\hat{\underset{\sim}{V}}$. An alternative test is obtained by referring

$$FX^2_c = X^2_c/(k-1)\ , \text{ or } FG^2_c = G^2_c/(k-1)$$

to $F_{(k-1),(r-1)(k-1)}$, an F distribution on k-1 and $(r-1)(k-1)$ degrees of freedom. See Thomas and Rao (1984) for details. When H_o is true, the modified estimator $\bar{\hat{\lambda}}_o$, with $\hat{\pi}_i$ replaced by π_{0i}, is also a consistent estimator of $\bar{\lambda}$. Modified X^2 and G^2 statistics based on $\bar{\hat{\lambda}}_o$ will be denoted by $X^2_{c_o}$, $G^2_{c_o}$ and $FX^2_{c_o}$, $FG^2_{c_o}$ respectively.

(5) Rao and Scott's Satterthwaite Corrections. The procedure consists of referring

$$X^2_S = X^2_c/(1+\hat{a}^2)$$

to $\chi^2_{k^*}$, where $k^* = (k-1)/(1+\hat{a}^2)$. The estimate $\hat{a}^2$ can be obtained via the expression

$$\Sigma\hat{\lambda}_i^2 = \sum_{i,j=1}^{k}\hat{v}^2_{ij}/(\hat{\pi}_i\hat{\pi}_j)\ .$$

As before, a version of X^2_S can be obtained by replacing $\hat{\pi}$ by π_o, and this version is denoted by $X^2_{S_o}$. Satterthwaite versions of G^2, namely G^2_S and $G^2_{S_o}$, can be defined analogously. F-based versions, FX^2_S abd $FX^2_{S_o}$, are obtained by referring $X^2_c/(k-1)$ and $X^2_{c_o}/(k-1)$ to $F_{k^*,(r-1)k^*}$. Similarly, FG^2_S and $FG^2_{S_o}$ are obtained.

(6) Fellegi's Correction. The procedure refers $X^2_F = X^2/\bar{\hat{d}}$, or $G^2_F = G^2/\bar{\hat{d}}$ to χ^2_{k-1}, where $\bar{\hat{d}}$ is the mean of the cell design effects $\hat{d}_i$ (Fellegi, 1980).

Results

All results are given in terms of realized significance levels, i.e. the proportion of actual rejections of a correct hypothesis, at a nominal level of $\alpha = 5\%$ in 1000 independent trials.

X^2 and G^2 tests.

Table 1 gives the actual significance levels (SL) for the uncorrected X^2 and G^2 tests, for the case of $r = 50$ clusters. The results are in fact quite insensitive to r, the number of clusters, even for values of r as low as 5. Clearly, these uncorrected tests are unacceptable unless $\bar{\lambda}$ is close to unity, i.e. unless the effect of the clustering is very small. The distortion in significance levels is primarily related to $\bar{\lambda}$; for constant $\bar{\lambda}$, increasing the coefficient of variation of the λ_i's, namely a, appears to decrease the significance levels, though the relative effect of changes in a is minor. It can also be seen that for constant $\bar{\lambda}$, the performance of X^2 and G^2 deteriorates rapidly as k, the number of categories, increases. For example, $SL(X^2)$ for $\bar{\lambda} = 2.0$ and $a = 0$ increases from 20.8 to 50.3 as k increases from 3 to 10.

Table 1

Actual Significance Levels (%) for the Unadjusted Tests X^2 and G^2

$r = 50$, $\alpha = 5\%$, $\underset{\sim}{\pi} = (1/k,\ldots,1/k)'$

k	$\bar{\lambda}$	a	m	$SL(X^2)$	$SL(G^2)$
3	1.5	0.0	10	13.4	13.7
3	1.5	0.5	10	13.6	13.9
3	2.0	0.0	10	23.3	23.0
3	2.0	0.5	10	20.8	20.7
5	2.0	0.0	10	31.7	32.1
5	2.0	0.5	10	28.3	28.7
10	2.0	0.0	20	50.3	50.0
10	2.0	0.5	20	48.4	46.6
10	2.0	1.0	20	44.5	44.2
10	1.05	0.0	20	6.0	6.4

Wald X^2 versus Wald F tests.

Table 2 compares the actual significance levels (SL) of the X^2_W and Hotelling's F_W versions of the Wald statistic, for a range of values of r. Several important conclusions can be drawn. First, X^2_W performs poorly even for $r = 50$ clusters when $k = 10$, yielding an actual significance level close to 20%. As k decreases, its performance improves, the actual level for $k = 3$, $r = 50$ being 6.0%. For a given combination (k,m,a), the performance of X^2_W deteriorates rapidly as r decreases; for $k = 5$, $m = 10$, $a = 0.5$, its significance level goes from 12.6% at $r = 30$ to 37.4% at $r = 10$. Clearly, unless k is small (≤ 5), the chi-squared version of the Wald test must be used with caution. Even for small k, it should not be used unless the number of clusters is 50 or more. It should be noted that this poor behaviour of X^2_W is not merely a function of large $\bar{\lambda}$ and a. Even for a $\bar{\lambda}$ of 1.05 and $a = 0$ (i.e., approximately the multinomial case), X^2_W has an actual significance level of 17.4% for $k = 10$ and $r = 50$. These findings confirm the warnings given by Fay (1983b) regarding the use of the Wald procedure.

Table 2

Comparison of the Actual Significance Levels (%) of X^2_W and F_W

$\alpha = 50\%$, $\bar{\lambda} = 2.0$, $\underset{\sim}{\pi} = (1/k,\ldots,1/k)'$

k	m	a	r	$SL(X^2_W)$	$SL(F_W)$
3	10	.5	50	6.0	4.7
			30	7.1	5.5
			10	15.5	7.4
5	10	.5	50	9.1	5.9
			30	12.6	7.7
			10	37.4	10.8
10	20	1.0	50	20.5	7.9
			30	32.5	10.4
			20	49.4	12.5
			10	95.4	5.5

From the point of view of significance level, the F_W version of the Wald test is more stable, though it too attains an excessive significance level of over 12% for $k = 10$ and $r = 20$. Though F_W gives better control of test size, its power for small to moderate numbers of clusters, however, is likely to be small.

Rao-Scott procedures based on $\hat{\pi}$ and π_o.

In general, the $\bar{\lambda}$ and Satterthwaite adjusted tests are not very sensitive to the choice of π_o or $\hat{\pi}$ in the calculation of $\hat{\bar{\lambda}}$ and $\Sigma\hat{\lambda}^2_i$. For moderate to large r (30-50), differences attributable to the use of $\hat{\pi}$ or π_o are minor. For small numbers of clusters ($r = 10$), use of π_o results in lower attained significance levels, which is beneficial when the coefficient of variation of λ_i is not small. Henceforth, the results are given for procedures based on π_o.

Table 3

Variants of the Rao-Scott $\bar{\lambda}$ Adjusted Test: Significance Levels

$\alpha = 5\%$, $\bar{\lambda} = 2.0$, $\underset{\sim}{\pi} = (1/k,\ldots,1/k)'$

k	m	a	r	$SL(X^2_{C_o})$	$SL(FX^2_{C_o})$	$SL(X^2_F)$
3	10	0.0	50	5.5	5.1	5.4
			30	4.7	4.0	4.4
			10	6.7	4.7	6.3
3	10	0.5	50	5.9	5.5	5.7
			30	7.2	6.7	6.7
			10	10.7	7.1	9.8
10	20	0.0	50	5.5	5.3	5.9
			30	4.7	4.1	5.1
			10	5.0	3.5	5.4
10	20	1.0	50	11.5	11.1	11.4
			30	12.7	12.1	13.1
			10	14.1	12.6	15.9

Variants of the Rao-Scott $\bar{\lambda}$ Test.

Significance levels for $X^2_{C_o}$ and $FX^2_{C_o}$ are shown in Table 3 for a selection of k and a combinations that exhibit both liberal and conservative behaviour. Also shown is Fellegi's heuristic adjustment to X^2. It can be seen that $X^2_{C_o}$ can become overly liberal for large values of a , as expected. In all cases, $FX^2_{C_o}$ exhibits a lower significance level than $X^2_{C_o}$, without being excessively conservative, even for the case k = 10, a = 0. Thus $FX^2_{C_o}$ will be used from now on in preference to $X^2_{C_o}$. Fellegi's procedure yields significance levels that are similar to those produced by $X^2_{C_o}$, a conclusion that holds true for a wide variety of test conditions. For this reason, Fellegi's procedure will not be discussed separately in what follows.

$X^2_{S_o}$ versus $FX^2_{S_o}$.

The modification $FX^2_{S_o}$ always yields lower SL than does $X^2_{S_o}$, but tends to be unnecessarily conservative for $k \geq 5$. However, for k = 3 the lower SL of $FX^2_{S_o}$ are advantageous. Thus, in the comparisons of Table 4, $X^2_{S_o}$ is used for $k \geq 5$, while $FX^2_{S_o}$ is used for k = 3.

Overall Comparisons of the Rao-Scott, Fay and Wald Tests.

It can be seen from Table 4 that, for the equiprobable case, the significance levels of tests based on X^2 and G^2 are quite similar. When there are noticeable differences, they usually favour X^2 , e.g. for k = 10, r = 10, a = 1.0 , the $\bar{\lambda}$, Satterthwaite and Fay procedures are two to three percentage points more liberal for G^2 than for X^2. Remaining comparisons will therefore focus on X^2 based tests.

As previously noted, $FX^2_{C_o}$, the F version of the Rao-Scott $\bar{\lambda}$ adjusted test, can be liberal for large values of a , particularly so for k = 10. For k = 3 and k = 5, $FX^2_{C_o}$ behaves well over a wide range of values of r. It should be noted that application of $FX^2_{C_o}$ requires knowledge of only the estimated cell design effects, whereas the other tests require knowledge of $\hat{\underset{\sim}{V}}$, or the replicate level data.

Table 4

Significance Levels of Rao-Scott and Fay Tests: X^2 and G^2 versions

$\alpha = 5\%$, $\bar{\lambda} = 2.0$, $\underset{\sim}{\pi} = (1/k,\ldots,1/k)'$

k	m	a	r	$SL(X^2_{S_o})$	$SL(G^2_{S_o})$	$SL(FX^2_{C_o})$	$SL(FG^2_{C_o})$	$SL(X_J)$	$SL(G_J)$
3	10	.5	50	5.1*	5.0*	5.5	5.7	4.9	5.0
			30	5.1*	5.3*	6.7	7.0	5.4	5.2
			10	5.6*	5.9*	7.1	7.9	9.2	8.3
5	10	.5	50	5.1	5.1	6.2	6.5	4.8	5.0
			30	5.1	4.9	6.2	6.7	5.8	6.0
			10	7.8	7.6	9.0	9.3	10.4	10.9
10	20	1.0	50	7.3	8.2	11.1	11.2	5.6	4.4
			30	8.1	8.8	12.1	11.8	8.0	7.1
			10	6.5	9.5	12.6	14.9	13.8	16.5

* These values correspond to $FX^2_{S_o}$ and $FG^2_{S_o}$.

Fay's X_J procedure exhibits some interesting characteristics. Though only an approximate test for $a > 0$, it displays the best 'asymptotic' behaviour of the four competing tests studied. (See Table 2 for results for F_W.) For r = 50, its actual significance levels are very close to 5%. It does not exhibit any tendency to conservativeness, and for $r \geq 30$, it limits the actual significance levels to 8% or less. However, for r = 10, X_J can become quite liberal, exhibiting significance levels well over 10%.

Over the complete range of k , r and a studied, the original Satterthwaite approximation $X^2_{S_o}$, supplemented by $FX^2_{S_o}$ for k = 3 , seems to provide the best compromise. Though a little liberal for the extreme case k = 10 , a = 1.0 , significance levels are for the most part within $\pm$ 3 points of the nominal 5% level.

The Effect of Skewness in $\underset{\sim}{\pi}$.

The effect of varying degrees of skewness in $\underset{\sim}{\pi}$, the population probability vector, has been examined for the case k = 5 (see Thomas and Rao, 1984). The results are summarized below.

If the minimum cell expected count is ≥ 1 , then skewness has little effect on significance levels for the statistics $FX^2_{C_o}$, $X^2_{S_o}$ and X_J . The actual significance levels of F_W , however, become more liberal with increased skewness, especially for small numbers of clusters.

If the minimum expected cell count per cluster is 0.5, then for the case of moderate r , there is again no evidence of a skewness effect on the significance levels of $FX^2_{C_o}$, $X^2_{S_o}$ and X_J .

For small numbers of clusters, however, there is clear evidence of increasing liberality with increasing skewness for these three procedures. Results for F_W follow its previous pattern, but the effects are more pronounced. Even for moderate r, the significance level for the least skewed case studied (.3, .3, .3, .05, .05) is at least twice as great as in the uniform case (.2, .2, .2, .2, .2), which puts F_W into the unacceptable category. For small r, significance levels of F_W are even higher, reaching over 26% for the highly skewed case (.8, .05, .05, .05, .05).

Summary and Conclusions

Monte Carlo techniques were used to examine the type I error performance of a number of chi-squared goodness-of-fit test statistics under cluster sampling. A study of a number of variants of the basic statistics under consideration has reduced the comparison to four procedures, namely an F-based version of the Rao-Scott $\bar{\lambda}$ adjusted X^2 statistic, the original Rao-Scott Satterthwaite adjusted X^2, Fay's jackknifed X^2 and a modified Wald statistic referred to an F distribution. The $\bar{\lambda}$ adjusted X^2 statistic depends only on the cell design effects unlike the others. This statistic performs well provided that the coefficient of variation a of the λ_i's, the eigenvalues of the design effect matrix, is small. In general, the Satterthwaite adjusted test and Fay's jackknifed test perform well even when a is not small. The modified Wald statistic behaves reasonably well for goodness-of-fit tests of uniform probability vectors $\underset{\sim}{\pi}$, but it is sensitive to skewness in $\underset{\sim}{\pi}$, particularly when the expected cell count per cluster is less than one.

References

Brier, S.S. (1978). *Categorical Data Models for Complex Data Structures*. Unpublished Ph.D. Dissertation, School of Statistics, University of Minnesota.

Fay, R.E. (1979). On adjusting the Pearson chi-square statistic for cluster sampling. *Proc. Soc. Statist. Sec.*, Amer. Statist. Assoc., 71, 665-670.

Fay, R.E. (1983a). CPLX-Contingency table analysis for complex sample designs, program documentation. Unpublished report.

Fay, R.E. (1983b). Replication approaches to the log-linear analysis of data from complex samples. Unpublished report.

Fellegi, I.P. (1980). Approximate tests of independence and goodness of fit based on stratified multistage samples. *J. Amer. Statist. Assoc.*, 71, 665-670.

Hidiroglou, M.A., Fuller, W.A., and Hickman, R.D. (1980), MINI CARP: A program for stratified and cluster samples. Survey Section, Iowa State University, Ames, Iowa.

IMSL (1950). IMSL Library, Edition 8, International Mathematical and Statistical Libraries Inc., Houston, Texas.

Koch, G.G., Freeman, D.H. Jr. and Freeman, J.L. (1975). Strategies in the multivariate analysis of data from complex surveys. *Intl. Statist. Rev.*, 43, 59-78.

NAG (1983). The NAG FORTRAN Mark 10 Library, Numerical Algorithms Group, Oxford, England.

Olson, C.L. (1974). Comparative robustness of six tests in multivariate analysis of variance. *J. Amer. Statist. Assoc.*, 69, 894-908.

Rao, J.N.K. and Scott, A.J. (1979). Chisquared tests for analysis of categorical data from complex surveys. *Proc. Sec. Survey Res. Methods*, Amer. Statist. Assoc., 58-66.

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. *J. Amer. Statist. Assoc.*, 76, 46-60.

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

Schruben, L.W. and Margolin, B.H. (1978). Pseudorandom number assignment in Statistically designed simulation and distribution sampling experiments. *J. Amer. Statist. Assoc.*, 73, 504-525.

Thomas, D.R. and Rao, J.N.K. (1984). A Monte Carlo study of exact levels for chi-squared goodness-of-fit statistics under cluster sampling. Technical Report #35, Laboratory for Research in Statistics ane Probability, Carleton University/University of Ottawa, Ottawa, Canada.