

Simultaneous Confidence Intervals for Proportions Under Cluster Sampling

D. ROLAND THOMAS¹

ABSTRACT

The paper describes a Monte Carlo study of simultaneous confidence interval procedures for $k > 2$ proportions, under a model of two-stage cluster sampling. The procedures investigated include: (i) standard multinomial intervals; (ii) Scheffé intervals based on sample estimates of the variances of cell proportions; (iii) Quesenberry-Hurst intervals adapted for clustered data using Rao and Scott's first and second order adjustments to X^2 ; (iv) simple Bonferroni intervals; (v) Bonferroni intervals based on transformations of the estimated proportions; (vi) Bonferroni intervals computed using the critical points of Student's t . In several realistic situations, actual coverage rates of the multinomial procedures were found to be seriously depressed compared to the nominal rate. The best performing intervals, from the point of view of coverage rates and coverage symmetry (an extension of an idea due to Jennings), were the t -based Bonferroni intervals derived using log and logit transformations. Of the Scheffé-like procedures, the best performance was provided by Quesenberry-Hurst intervals in combination with first-order Rao-Scott adjustments.

KEY WORDS: Simultaneous inference; Complex surveys; Monte Carlo.

1. INTRODUCTION

Survey results are often presented as estimated proportions (or percentages) of population units belonging to two or more distinct categories. Examples include many sociological studies (see for example Black and Myles 1986), marketing studies and opinion polls. As noted by Fitzpatrick and Scott (1987), inference on category proportions is often based on single binomial confidence intervals, even when more than two category proportions are being examined. This paper describes a study of several procedures for constructing simultaneous confidence intervals for the proportions π_i , $i = 1, \dots, k$, of population units belonging to each of k distinct categories, using data from a two-stage cluster sample. Standard simultaneous confidence interval procedures for categorical data problems, reviewed by Hochberg and Tamane (1987), are based on the assumption of multinomially distributed sample counts, and are thus appropriate for data from simple random samples. When the data have been collected using sample survey designs that involve clustering, standard procedures are likely to perform poorly, as is the case when standard multinomial based tests are applied to data from complex sample surveys. In the latter case, it has been shown by many workers that clustering can lead to unacceptably high Type I error rates (see, for example, Fellegi 1980; Rao and Scott 1979, 1981; Holt, Scott and Ewing 1980). For simultaneous confidence intervals, therefore, it is natural to expect that clustering will lead to coverage probabilities that are lower than multinomial theory indicates.

Estimation of simultaneous confidence intervals (SCI's) is an important adjunct to hypothesis testing. The present study thus represents a natural follow-up to Thomas and Rao's (1987) investigation of test statistics for the simple goodness of fit problem, under

¹ D. Roland Thomas, School of Business, Carleton University, Ottawa, Ontario, K1S 5B6.

simulated cluster sampling. In this paper, adaptations of the standard SCI procedures are proposed, and their performance in small samples is evaluated using Monte Carlo techniques.

The cluster sampling model that is used in the Monte Carlo study is described in Section 2, and the SCI procedures to be examined are presented in Section 3. In Section 4, the design of the Monte Carlo experiment is described, together with procedures for evaluating confidence interval performance. The main results of the study are presented in Sections 5 through 7, followed in Section 8 by some final conclusions and recommendations.

2. THE CLUSTER SAMPLING MODEL

This investigation will focus on two-stage sampling in which a k -category sample of m units is drawn independently from each of r sampled clusters.

For a sample of size $n = mr$, let $\mathbf{m} = (m_1, \dots, m_{k-1})'$ represent the category counts for the whole sample, where $m_k = n - \sum_{i=1}^{k-1} m_i$. In terms of proportions, let $\hat{\boldsymbol{\pi}} = (\hat{\pi}_1, \dots, \hat{\pi}_{k-1})' = \mathbf{m}/n$ be the vector of category proportions for the full sample. Further, define $\boldsymbol{\pi} = \mathbf{E}(\hat{\boldsymbol{\pi}})$, where \mathbf{E} denotes expectation under a suitable model of cluster sampling, and let \mathbf{V}/n represent the $(k-1) \times (k-1)$ covariance matrix of $\hat{\boldsymbol{\pi}}$. Following Rao and Scott (1981), the ordinary design effect for the linear combination $\mathbf{c}'\hat{\boldsymbol{\pi}}$ of category proportions is $\mathbf{c}'\mathbf{V}\mathbf{c}/\mathbf{c}'\mathbf{P}\mathbf{c}$, where \mathbf{P} is n times the covariance matrix of $\hat{\boldsymbol{\pi}}$ under multinomial sampling, *i.e.*, $\mathbf{P} = \text{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'$, and \mathbf{c} is a vector of dimension $k-1$. The largest design effect taken over all possible linear combinations is given by the largest eigenvalue of the design effect matrix $\mathbf{D} = \mathbf{P}^{-1}\mathbf{V}$. The eigenvalues of \mathbf{D} , denoted in decreasing order by $\lambda_1, \lambda_2, \dots, \lambda_{k-1}$, were termed generalized design effects by Rao and Scott (1981), and provide a quantitative summary of the variance inflation associated with a particular design, relative to simple random sampling. Under the multinomial distribution, corresponding to simple random sampling from large populations, $\lambda_j = 1 \forall j$. Designs involving clustering usually yield generalized design effects greater than one on the average, *i.e.*, $\bar{\lambda} = \sum_{j=1}^{k-1} \lambda_j / (k-1) > 1$. Furthermore, studies of real survey data (Hidioglou and Rao 1987; Rao and Thomas 1988) reveal significant variation among the λ_j 's. This is conveniently represented by their coefficient of variation, given by

$$a = \left(\sum_{j=1}^{k-1} \lambda_j^2 / [(k-1)\bar{\lambda}^2] - 1 \right)^{1/2}. \quad (1)$$

A suitable model of cluster sampling must therefore be capable of generating generalized design effects such that $\bar{\lambda} > 1$ and $a > 0$.

Brier (1981) proposed a model of two-stage cluster sampling in which individual clusters are represented by vectors of category probabilities, $\mathbf{p}_\ell = (p_{\ell 1}, p_{\ell 2}, \dots, p_{\ell, k-1})'$, $\ell = 1, \dots, r$, where for each cluster, $p_{\ell k} = 1 - \sum_{i=1}^{k-1} p_{\ell i}$. Each \mathbf{p}_ℓ was independently drawn from a Dirichlet distribution with mean $\boldsymbol{\pi}$, *i.e.* $\mathbf{E}(\mathbf{p}_\ell) = \boldsymbol{\pi}$, and second stage sampling of the m units per cluster was multinomial, conditional on the realized value of \mathbf{p}_ℓ for that cluster. Let the vector of counts for each cluster be $\mathbf{m}_\ell = (m_{\ell 1}, \dots, m_{\ell, k-1})$, with $m_{\ell k} = m - \sum_{i=1}^{k-1} m_{\ell i}$. Thus for the full sample, $\mathbf{m} = \sum_{\ell=1}^r \mathbf{m}_\ell$, and in terms of proportions, $\hat{\boldsymbol{\pi}} = \sum_{\ell=1}^r \hat{\boldsymbol{\pi}}_\ell$, where $\hat{\boldsymbol{\pi}}_\ell = \mathbf{m}_\ell/m$. Brier (1981) showed that under this model, $\mathbf{E}(\hat{\boldsymbol{\pi}}) = \boldsymbol{\pi}$ and $\mathbf{V}(\hat{\boldsymbol{\pi}}) = d\mathbf{P}/n$, *i.e.*, the covariance matrix of $\hat{\boldsymbol{\pi}}$ is proportional to the multinomial covariance matrix, with the constant of proportionality $d > 1$. Under this model, the design effect matrix is given by $\mathbf{D} = d\mathbf{I}_{k-1}$, where \mathbf{I}_{k-1} is the identity matrix of order $k-1$. Thus $\lambda_i = d \forall i$, so that $\bar{\lambda} = d$ and $a = 0$. Brier's model can therefore represent variance inflation ($\bar{\lambda} > 1$), but cannot

represent the unequal generalized design effects encountered in practice. Thomas and Rao (1987) used an extension of Brier's model in which the first stage p_i 's are sampled independently from a mixture of two Dirichlet distributions, representing a population composed of two distinct classes of clusters. This model, which is a special case of that proposed by Rao and Scott (1979), generates one distinct and $k - 2$ equal eigenvalues, with $\bar{\lambda}$ and a being explicit functions of the Dirichlet parameters. This greatly facilitates the design of the Monte Carlo study by allowing for convenient control of the values of the clustering measures $\bar{\lambda}$ and a . Since it satisfies the basic requirements outlined above ($\bar{\lambda} > 1, a > 0$), Thomas and Rao's (1987) model will be used in this study.

3. SIMULTANEOUS CONFIDENCE INTERVAL PROCEDURES

3.1 Scheffé Intervals

A standard Scheffé argument, based on the asymptotically exact probability statement

$$P\left(n(\hat{\pi} - \pi)' \hat{V}^{-1} (\hat{\pi} - \pi) \leq \chi_{k-1}^2(\alpha)\right) = 1 - \alpha \tag{2}$$

leads to simultaneous confidence intervals for linear combinations, $\ell' \pi$, of the category probabilities, where ℓ is a vector of dimension $(k - 1)$. Appropriate choices of ℓ then yield SCI's on the individual cell probabilities given by

$$\pi_i \in \left\{ \hat{\pi}_i \pm (\hat{v}_{ii})^{1/2} (A/n)^{1/2} \right\}, i = 1, \dots, k, \tag{3}$$

where $A = \chi_{k-1}^2(\alpha)$ is the upper α percent point of a chi-squared distribution on $k - 1$ degrees of freedom, and \hat{v}_{ii} is the i^{th} diagonal element of a consistent estimator of V (as $r \rightarrow \infty$) given by

$$\hat{V} = \frac{n}{r(r - 1)} \sum_{l=1}^r (\hat{\pi}_l - \hat{\pi}) (\hat{\pi}_l - \hat{\pi})'. \tag{4}$$

Note that when the endpoint of an interval lies outside $[0, 1]$, definition (3) must be modified by truncating the endpoint to 0 or 1 as appropriate. For multinomial sampling, \hat{v}_{ii} can be replaced by $\hat{\pi}_i (1 - \hat{\pi}_i)$, in which case the Scheffé intervals reduce to those proposed by Gold (1963). The latter will be referred to as Scheffé-Gold intervals. The Scheffé intervals of equation (3) will be conservative, *i.e.*, will have coverage exceeding $(1 - \alpha)$ asymptotically since they make use of only a finite number of the available ℓ directions (see Miller 1981, page 63). In fact, they will become very conservative as k increases, as can be shown using the following argument due to Goodman (1965). The coverage of the Scheffé intervals is equal to one minus the probability of occurrence of at least one of the events $\{(\hat{\pi}_i - \pi_i)^2 / (\hat{v}_{ii}/n) > \chi_{(k-1)}^2(\alpha)\}$, $i = 1, \dots, k$; since the random variables $(\hat{\pi}_i - \pi_i)^2 / (\hat{v}_{ii}/n)$ each have chi-squared distributions on one degree of freedom asymptotically, the probability of each individual event can be evaluated. Using the Bonferroni inequality, lower bounds for the coverage can then be obtained; for a nominal coverage of 95% with $k = 3, 5, 8$ and 12 , these bounds are .9571, .9896, .9986 and .9999 respectively.

3.2 Modified Quesenberry-Hurst Intervals

Under the assumption of multinomial sampling, Quesenberry and Hurst (1964) solved the large sample probability statement

$$P\left\{X^2 = n \sum_{i=1}^k \frac{(\hat{\pi}_i - \pi_i)^2}{\pi_i} \leq A\right\} = 1 - \alpha \quad (5)$$

for the cell probabilities π_i , to get the SCI's

$$\pi_i \in \left\{ \frac{\hat{\pi}_i + A/2n \pm (A/n)^{1/2} [\hat{\pi}_i(1 - \hat{\pi}_i) + A/4n]^{1/2}}{1 + A/2n} \right\}. \quad (6)$$

Under multinomial sampling, these intervals are asymptotically equivalent to Scheffé and Scheffé-Gold intervals, and will therefore exhibit similar asymptotic conservativeness.

Quesenberry-Hurst (Q-H) intervals can be adapted for use with clustered survey data using the first and second order corrections to the distribution of X^2 proposed by Rao and Scott (1981). Corresponding first and second order SCI's are obtained by replacing A in equation (3) by

$$A^{(1)} = \hat{\lambda}A \text{ and } A^{(2)} = \hat{\lambda}(1 + \hat{a}^2) \chi_v^2(\alpha) \quad (7)$$

respectively, where $v = (k - 1)/(1 + \hat{a}^2)$ and $\hat{\lambda}$, an estimate of the mean of the generalized design effects, is given by (Rao and Scott, 1981)

$$\hat{\lambda} = (k - 1)^{-1} \sum_{i=1}^k (1 - \hat{\pi}_i) \hat{d}_i, \quad (8)$$

where \hat{d}_i , $i = \dots, k$ is an estimated cell design effect given by $\hat{d}_i = \hat{v}_{ii}/\hat{\pi}_i(1 - \hat{\pi}_i)$. The coefficient of variation, a , is estimated by replacing $\bar{\lambda}$ in equation (1) by $\hat{\lambda}$, and $\sum \lambda_i^2$ by the estimate $\sum \hat{\lambda}_i^2 = \sum \sum \hat{v}_{ij}^2/\hat{\pi}_i\hat{\pi}_j$. It turns out (see Thomas 1989) that the second order modified intervals are unnecessarily conservative, so that only the first-order modified Q-H intervals will be discussed in the remainder of the paper.

3.3 Simple Bonferroni Intervals

Since (loosely speaking) each $\hat{\pi}_i$ is asymptotically $N(\pi_i, v_{ii}/n)$, the intervals

$$\pi_i \in \left\{ \hat{\pi}_i \pm (v_{ii}/n)^{1/2} z_{\alpha'/2} \right\}, \quad (9)$$

will have large sample coverage at least $(1 - \alpha)$ by the Bonferroni inequality, where $\alpha' = \alpha/k$ and $z_{\alpha'/2}$ is the upper $\alpha'/2$ percent point of the standard normal distribution. Intervals (9) are equivalent to Scheffé intervals with A in equation (3) replaced by $A^{(3)} = \chi_1^2(\alpha')$. As noted

by Goodman (1965), they will be shorter than Scheffé intervals for the usual values of α and k ; e.g., $\alpha = 1\%$, 5% , or 10% . Goodman's (1965) multinomial Bonferroni intervals are given by equation (9) with \hat{v}_{ii} replaced by $\hat{\pi}_i (1 - \hat{\pi}_i)$. All endpoints of simple Bonferroni intervals that lie outside $[0, 1]$ will be truncated to 0 or 1 as appropriate.

3.4 Transformed Bonferroni Intervals

For suitably smooth g , $g(\hat{\pi}_i)$ will be asymptotically $N(g(\pi_i), [g'_i(\pi_i)]^2 v_{ii}/n)$, where $g'_i(\pi_i)$ denotes the partial derivative $\partial g(\pi_i)/\partial \pi_i$ evaluated at π_i . Bonferroni intervals can then be obtained by inverting corresponding intervals on the $g(\pi_i)$'s, giving

$$\pi_i \in \left\{ g^{-1}(g(\hat{\pi}_i) \pm g'_i(\hat{\pi}_i) (\hat{v}_{ii}/n)^{1/2} z_{\alpha'/2}) \right\}. \tag{10}$$

Three g functions will be investigated: the square root $g_1(\pi_i) = \pi_i^{1/2}$ (previously investigated by Bailey 1980, for the case of multinomial sampling); the natural logarithm $g_2(\pi_i) = \ln(\pi_i)$; and the logit $g_3(\pi_i) = \ln(\pi_i/(1 - \pi_i))$. Interval endpoints that lie outside $[0, 1]$ will again be truncated to 0 or 1 as necessary.

Transformed Bonferroni intervals based on a jackknifed estimator of the variance of $g(\hat{\pi})$ have also been examined (see Thomas 1989). It was found that there is little advantage to using jackknifed variance estimates; Taylor series variance estimates are therefore recommended for their simplicity. Intervals based on jackknife variance estimates will not be considered further in this paper.

3.5 Variants of the Above Intervals

Scheffé Intervals: Following Thomas and Rao (1987), Scheffé intervals can be modified by replacing the critical constant A in equation (3) by $A^{(4)} = (k - 1)(r - 1)(r - k + 1)^{-1} F_{(k-1), (r-k+1)}(\alpha)$, where $F_{(k-1), (r-k+1)}(\alpha)$ is the upper α percent point of an F distribution on $(k - 1)$ and $(r - k + 1)$ degrees of freedom.

Quesenberry-Hurst Intervals: Variants of the modified Quesenberry-Hurst (Q-H) intervals can also be defined, corresponding to the F forms of the first and second order corrected test statistic proposed by Thomas and Rao (1987). These again turn out to be conservative, and will not be considered further.

Bonferroni Intervals: Heuristic arguments (see the appendix to Thomas and Rao 1987) suggest that the simple Bonferroni intervals can be improved by replacing $z_{\alpha'/2}$ in (9) by $t_{r-1}(\alpha'/2)$, the upper $\alpha'/2$ percentage point of Student's t distribution on $r - 1$ degrees of freedom. This strategy will also be applied to the transformed Bonferroni intervals.

4. THE DESIGN OF THE MONTE CARLO STUDY

4.1 Parameters and Random Numbers

The parameters to be controlled are: (i) the nominal coverage level $(1 - \alpha)$ of the SCI; (ii) π , the model probability vector; (iii) k , the number of categories; (iv) r , the number of sample clusters; (v) m , the number of units drawn from each sampled cluster; (vi) $\bar{\lambda}$, the mean of the generalized design effects (eigenvalues); (vii) a , the coefficient of variation of the generalized design effects. The nature and degree of clustering is represented by the pair $(\bar{\lambda}, a)$ as follows: (a) multinomial sampling ($\bar{\lambda} = 1, a = 0$); (b) constant design effect clustering ($\bar{\lambda} > 1, a = 0$); (c) non-constant design effect clustering ($\bar{\lambda} > 1, a > 0$).

Individual Monte Carlo experiments were run for particular combinations of k , $\bar{\lambda}$, a and r_{max} , the latter being the maximum number of clusters generated in one computer run. Most experiments were run at two values of $\bar{\lambda}$, namely 1.5 and 2.0, two values of a , namely $a = 0$ (constant design effects) and $a > 0$ (one level of non-constant design effects), for equiprobable categories ($\pi_i = 1/k, i = 1, \dots, k$). Three values of k ($k = 3, 5, 8$) were initially selected to cover the range of numbers of categories commonly encountered in goodness-of-fit tests. An additional run was subsequently done for the case $k = 12, \bar{\lambda} = 2$ and $a > 0$ to check on the range of applicability of the results. The number of units per cluster was set at $m = 10$ for $k = 3, 5$ and 8 , and at $m = 20$ for $k = 12$. Preliminary investigations showed coverage rates to be insensitive to the value of this parameter. For comparability of results over k , the non-zero settings of a were selected to make a/a_{max} the same for each selected value of k , where $a_{max} = (k - 2)^{1/2}$ is the maximum possible value of a . For $k = 5$, the non-zero value of a was set at 0.5, which is typical of the values encountered in practice, e.g., $\hat{a} = 0.43$ for $k = 5$, as reported by Rao and Thomas (1988).

The initial focus on equiprobable categories allowed for a cost effective assessment of the influence of k , $\bar{\lambda}$ and a on coverage rates, and eliminated many of the possible SCI variants from further consideration. Additional experiments reported in Section 7 show that the procedures that passed this initial screening can in fact be applied when the cell probabilities are markedly unequal. Vectors of unequal probabilities were confined to the class $\pi(k, q, \phi)$, defined by the elements $\pi_i = \phi, i = 1, \dots, q$ and $\pi_i = (1 - q\phi)/(k - q), i = q + 1, \dots, k$.

For details of the generation of the random clusters from the mixture Dirichlet multinomial distribution, the reader is referred to Thomas and Rao (1987). Each Monte Carlo experiment consisted of 1000 sets of up to 100 independent clusters, grouped into nested subsets. All SCI procedures were applied in turn to each subset, using two nominal coverage levels (95% and 90%), thus improving the precision of comparisons between procedures at the same parameter settings, and between the same SCI procedures for different numbers of clusters. Most of the results presented will be for 95% nominal coverage; trends for 90% coverage were found to be qualitatively similar.

4.2 Evaluation Procedures

The percentage of Monte Carlo trials for which at least one of the k confidence intervals fails to cover the true parameter value is reported, and used for a preliminary screening of the main SCI procedures. This is a measure of the family error rate, which is equivalent to the actual significance level of the SCI when the latter is viewed as a test of goodness-of-fit. The family error rate, which will be referred to in this paper as the total error rate ER_T , is used in place of the more commonly reported actual coverage rate (equal to one hundred percent minus the total error rate) because it can be conveniently split into two one-sided rates which will provide information on the symmetry or 'unbiasedness' of each SCI procedure. Jennings (1987) argued that coverage rates alone can provide a misleading assessment of single parameter confidence interval procedures, and recommended that the number of times that an interval falls above and below the true parameter value should be separately reported. In this paper, Jennings' suggestion has been adapted to simultaneous confidence intervals on $\pi_i, i \in I$, where I is the index set $\{1, \dots, k\}$, by counting the number of Monte Carlo trials for which:

- (a) more intervals fall above their corresponding $\pi_i, i \in I$, than fall below;
- (b) more intervals fall below their corresponding $\pi_i, i \in I$, than fall above;
- (c) the same number (> 0) of intervals fall above their corresponding $\pi_i, i \in I$, as fall below.

Upper and lower error rates are then defined as $ER_U = [n_a + (n_c/2)]/N_t$ and $ER_L = [n_b + (n_c/2)]/N_t$, respectively, where N_t represents the number of Monte Carlo trials, and n_a , n_b and n_c denote the counts (a) through (c), respectively. The sum of ER_U and ER_L is clearly equal to the total error rate, ER_T . These one-sided error rates will be used to compare SCI procedures whose overall error rates are acceptably close to the nominal rate α , over a range of parameter settings and cluster strengths. Average interval lengths and corresponding standard errors have also been computed, and will be used as final discriminators in the selection of the recommended procedures.

5. A SUMMARY OF RESULTS FOR TOTAL ERROR RATES

All results in this section are given in terms of the total error rate ER_T , defined in Section 4. For lack of space, tables are presented only for the case of unequal design effects, ($a > 0$), with $\bar{\lambda} = 2$. More detailed results are given in Thomas (1989). In interpreting the tabulated results, it should be noted that for 1000 Monte Carlo trials, binomial standard errors of point estimates of true ER_T 's having magnitudes 5%, 10% and 20% are 0.7%, 0.9% and 1.3% respectively. As a general rule deviations from nominal rates, and differences between the error rates of different SCI procedures will be noted only when they are large enough to have practical significance, and exceed their Monte Carlo standard errors by a factor of at least two.

5.1 Multinomial Procedures

Results for multinomial intervals will only be summarized here; for details see Thomas (1989). Under cluster sampling, error rates for Goodman's Bonferroni intervals (see equation (9) with \hat{v}_{ii} replaced by $\hat{\pi}_i(1 - \hat{\pi}_i)$) are unacceptably high except for values of $\bar{\lambda}$ close to 1, *i.e.*, unless the effect of clustering is small. The Scheffé-Gold and multinomial Quesenberry-Hurst intervals, on the other hand, can yield error rates that are close to the nominal value in certain cases, whenever their inherent conservativeness balances the error inflating effects of clustering (see also Andrews and Birdsall 1988). Unfortunately, this is not always the case; both procedures can display inflated error rates ($ER_T \geq 2\alpha$) for realistic combinations of category numbers and clustering strengths.

Multinomial procedures should therefore not be used with complex survey data. Procedures are clearly required that directly account for the clustering, and provide good coverage for the required number of categories, over a wide range of clustering conditions.

5.2 The Scheffé Procedures

Total error rates for the χ^2 -based Scheffé procedure of equation (3) and its F -based variant are summarized in Table 1 as functions of r , for the case $\alpha = 5\%$, $\bar{\lambda} = 2$ and $a > 0$. More detailed graphs are given in Thomas (1989).

For the values of k studied, ER_T for the χ^2 -based Scheffé procedure of equation (3) increases rapidly as the number of clusters decreases, so that it should never be used for small numbers of clusters. The F -based variant, on the other hand, keeps ER_T reasonably close to or below $\alpha = 5\%$ for all r . As r increases, ER_T for F -based Scheffé remains fairly constant for the case $k = 3$, but becomes increasingly conservative for $k \geq 5$, as does the χ^2 version. These empirical trends with varying r can be explained in terms of two competing effects. As r increases, error rates for both procedures approach their asymptotic levels which are bounded above by 4.29%, 1.04% and 0.14%, for $k = 3, 5$ and 8 respectively (see Section 3.1).

Table 1
 Total Error Rates for Scheffé and Modified Q-H Intervals;
 $\alpha = 5\%$, $\bar{\lambda} = 2$, $m = 10$

<i>k</i>	<i>a</i>	<i>r</i>	Total Error Rate (ER_T)		
			Scheffé (χ^2 based)	Scheffé (<i>F</i> based)	Modified Q-H (first order)
3	.29	15	9.2	5.9	5.0
3	.29	30	5.7	4.7	5.1
3	.29	50	5.4	5.0	5.4
5	.5	15	8.8	5.2	4.3
5	.5	30	4.0	3.0	2.7
5	.5	50	2.5	2.0	2.0
8	.71	15	12.7	7.4	2.4
8	.71	30	4.2	3.0	2.7
8	.71	50	2.7	1.6	2.5
8	.71	100	0.8	0.7	2.3

As *r* decreases, however, the conservativeness of the Scheffé procedures (for $k \geq 5$) will be increasingly swamped by the effects of increasing non-normality of the estimated proportions, $\hat{\pi}$. For the *F*-based version, the inflation in error rate due to non-normality is less than for the chi-squared version of equation (3), with the result that ER_T for the *F*-based version never seriously exceeds the nominal 5% rate. For moderate levels of clustering ($\bar{\lambda} = 1.5$), the behaviour of the *F*-based procedure is qualitatively similar to that described above for the case $\bar{\lambda} = 2$. From the point of view of total error rate, therefore, the *F*-based Scheffé procedure is useable over a wide range of clustering situations, though its possible conservativeness is a disadvantage.

5.3 Modified Quesenberry-Hurst Intervals

Total error rates for the first order modified Quesenberry-Hurst (Q-H) procedure of Section 3.2 are also shown in Table 1 for $\alpha = 5\%$, $\bar{\lambda} = 2$ and $a > 0$.

Total error rates are close to or below the nominal 5% for all combinations of *r* and *k* shown. For moderate to large numbers of clusters ($r \geq 30$), error rates for $k = 5$, and 8 are very similar, being approximately one half of the nominal rate (true also when $k = 12$). For the case of constant design effects (see Thomas 1989), error rates for first order modified Q-H intervals are conservative for $k \geq 5$, particularly for large *r*. The absence of this Scheffé-like conservativeness for the more realistic case of unequal design effects shown in Table 1 can again be explained using the argument of Section 3.1. From equation (6), it is easily seen that the asymptotic coverage of the first-order modified Q-H intervals is given by one minus the probability that at least one of the random variables $(\hat{\pi}_i - \pi_i)^2 / (\bar{\lambda}\pi_i(1 - \pi_i)/n)$, $i = 1, \dots, k$, will exceed the critical value $\chi_{k-1}^2(\alpha)$ asymptotically. When $a > 0$, these individual random variables will not all be asymptotically distributed as chi-squared on one degree of freedom, so that the bound of Section 3.1 does not apply. The true bound on the error rate will be inflated since at least one of the random variables $(\hat{\pi}_i - \pi_i)^2 / (\bar{\lambda}\pi_i(1 - \pi_i)/n)$ will be stochastically larger than $(\hat{\pi}_i - \pi_i)^2 / (v_{ii}/n)$, whenever $a > 0$.

Trends for the case $\bar{\lambda} = 1.5$ are similar (Thomas 1989). Overall, the results show that from the point of view of total error rates, first-order modified Q-H intervals provide a safe but somewhat conservative SCI procedure under realistic clustering conditions.

5.4 Simple Bonferroni Intervals

Total error rates for the simple Bonferroni intervals given by equation (9) are summarized in Table 2 for the case $\alpha = 5\%$, $\bar{\lambda} = 2$, $a > 0$, and $k = 3, 5$ and 8 . Also shown are corresponding error rates for the t -based variants described in Section 3.5.

From Table 2, it is evident that the error performance of both sets of SCI's is poor, both showing a strong tendency to high error rates for small to medium numbers of clusters when k , the number of categories, is five or more. Using critical values of Student's t distribution to compensate for the variability in the estimated variances of the category proportions clearly has the effect of generally lowering error rates. As can be seen from Table 2, however, this strategy is unable to prevent significant error rate inflation in the t -based intervals as the number of clusters decreases, except when $k = 3$. The trend to inflated error rates for small numbers of clusters (for both z and t -based intervals), is due to the increasing non-normality of the $\hat{\pi}_i$'s with decreasing r . This trend gets progressively more severe as k increases, which is to be expected since non-normality will become more pronounced, for a given value of r , as the values of the π_i 's get smaller. This is precisely what happens with increasing k in the case under study, for which $\hat{\pi}_i = 1/k \forall i$.

When $k = 3$, error rates for the t -based procedure are essentially constant, and close to the nominal level. For $k = 8$, on the other hand, ER_T varies from close to 20% at $r = 15$ to approximately 8% at $r = 100$. From Table 2, and other results not shown, it appears that for $k \geq 8$, simple t -based intervals approach their Bonferroni limits very slowly as $r \rightarrow \infty$. Also, for $k \leq 5$, error rates are close to the nominal level for moderate to large numbers of clusters ($r \geq 40$). Results for constant design effects, and for the case $\bar{\lambda} = 1.5$ are consistent with the above. From the point of view of total error rates (or equivalently of coverage rates), it is clear that simple t -based Bonferroni intervals are useable in practice over a range of realistic clustering situations only if $k \leq 5$ and $r \geq 40$.

Table 2
Total Error Rates for z and t -Based Simple Bonferroni Intervals;
 $\alpha = 5\%$, $\bar{\lambda} = 2$, $m = 10$

k	a	r	Total Error Rate (ER_T)	
			z -based	t -based
3	.29	15	10.0	5.6
3	.29	30	6.3	4.9
3	.29	50	6.5	5.5
5	.50	15	15.0	9.7
5	.50	30	8.8	7.2
5	.50	50	7.2	5.5
8	.71	15	29.6	19.1
8	.71	30	15.0	11.0
8	.71	50	11.5	9.8
8	.71	100	8.1	7.8

5.5 Transformed Bonferroni Intervals

The more detailed results given in Thomas (1989) demonstrate that the problem of error rate inflation exhibited by simple z -based Bonferroni intervals is not solved by the use of transformations alone. All three transformed z -based intervals again display severely inflated error rates for small to medium numbers of clusters. Fortunately, the effect of transformations on the t -based Bonferroni intervals is very different, as can be seen from the results summarized in Table 3.

For $k = 3, 5$ and 8 , error rates for the log and logit intervals are close to the nominal 5% for all r values shown, with the logit intervals yielding slightly lower rates than the log intervals (see the footnote to Table 3). The t -based square root intervals, on the other hand, exhibit the undesirable characteristic of error rate inflation for small r , when $k \geq 8$; they will not be considered further. For large numbers of categories ($k = 12$), both log and logit intervals do exhibit some error rate inflation for intermediate numbers of clusters ($r = 30$). This is not a serious drawback, however, as this number of categories is rarely encountered in practice. Results for constant design effects, and for the case $\bar{\lambda} = 1.5$ are generally similar to those described above.

It thus appears that for the ranges of k , r , $\bar{\lambda}$ and α that are likely to be encountered in practice, log and logit transformations (which reduce the non-normality in $\hat{\pi}$) used in combination with t -based critical values (which compensate for the variability in the estimated variances) do yield intervals that provide the desired degree of control. These intervals will be explored further in Section 6 in terms of the symmetry of their error rates.

Table 3
Total Error Rates¹ for t -based Transformed Bonferroni Intervals;
 $\alpha = 5\%$, $\bar{\lambda} = 2$, $m = 10$ for $k \leq 8$, $m = 20$ for $k = 12$

k	α	r	Total Error Rate (ER_T)		
			t -based Transformed Bonferroni		
			Square Root	Log	Logit
3	.29	15	4.5	4.6	3.3
3	.29	30	3.6	4.0	3.5
3	.29	50	4.6	5.6	4.1
5	.5	15	6.4	4.7	4.6
5	.5	30	4.6	4.2	3.5
5	.5	50	4.3	4.5	4.0
8	.71	15	12.0	5.9	5.2
8	.71	30	6.2	6.6	5.2
8	.71	50	5.9	5.4	5.2
8	.71	100	4.9	3.9	4.2
12	.91	15	17.0	6.7	6.5
12	.91	30	12.9	10.1	10.2
12	.91	50	8.2	6.5	6.3

¹ For $k = 8$ and $r = 50$, the correlation between ER_T estimates for log and logit intervals is 0.92. Assuming this is typical for all r and k , the Monte Carlo standard error of the difference in log and logit error rates is approximately 0.3%.

Table 4

Percentage Asymmetry (PER_U)¹ in the Total Error Rate for the Viable Procedures;
 $a > 0^2$, $r = 50$, $m = 10$ for $k \leq 8$, $m = 20$ for $k = 12$

α	k	$\bar{\lambda}$	$PER_U = (ER_U/ER_T) \times 100\%$			
			Scheffé (F -based)	Modified Q-H (first order)	t -based Bonferroni (log)	logit
5%	5	1.5	19.2	58.7	61.0	48.9
5%	5	2.0	0.0	45.0	61.1	48.8
5%	8	1.5	0.0	63.2	67.5	56.8
5%	8	2.0	0.0	65.2	64.9	49.0
5%	12	2.0	0.0	46.9	53.8	51.6
10%	5	1.5	16.3	49.4	59.2	48.4
10%	5	2.0	6.1	50.0	61.8	48.6
10%	8	1.5	0.0	60.7	67.3	55.8
10%	8	2.0	0.0	65.6	60.7	50.0
10%	12	2.0	0.0	47.5	56.0	51.4

¹ For $k = 8$, $\bar{\lambda} = 2$ and $\alpha = 5\%$, the correlation between PER_U estimates for log and logit intervals is 0.82. Assuming this is typical, Monte Carlo standard errors for differences in log and logit PER_U 's are approximately 4% and 3% for $\alpha = 5\%$ and 10%, respectively.

² For values of a for specific k , see Table 3.

6. ERROR RATE SYMMETRIES FOR THE VIABLE PROCEDURES

This section presents results on error rate symmetry based on the decomposition of the total error rate ER_T into its two additive components ER_U and ER_L , as described in Section 4. The measure used in the tables is $(ER_U/ER_T) \times 100\%$, *i.e.*, the upper error rate expressed as a percentage of the total error rate. It will be denoted PER_U . A symmetric SCI will have an empirical PER_U that is close to 50%; a PER_U that is greater (less) than 50% will indicate an increased probability of non-coverage due to intervals lying above (below) their respective π_i 's. For values of percentage symmetry between 50% and 80%, 95% confidence intervals on the true PER_U are approximately $(PER_U \pm 14)\%$ and $(PER_U \pm 10)\%$ for total error rates of 5% and 10% respectively.

6.1 Modified Scheffé and Quesenberry-Hurst Intervals

Percentage symmetry results for the F -based Scheffé and the first order Quesenberry-Hurst (Q-H) intervals are given in Table 4 for a selection of parameter values. It can be seen that the Scheffé procedure displays extreme asymmetry, making it an unattractive SCI. The first order modified Q-H procedure displays only moderate asymmetry, and is therefore the better of the two in practice.

The source of the asymmetry in the Scheffé intervals is again the non-normality of the untransformed $\hat{\pi}_i$'s. In particular, the fact that "small" $\hat{\pi}_i$'s generate "small" estimates of the variances $v_{\hat{\pi}_i}$ and hence shorter intervals (*cf.* the multinomial case where $\hat{v}_{\hat{\pi}_i} = \hat{\pi}_i(1 - \hat{\pi}_i)/n$, $i = 1, \dots, k$) increases the probability that non-covering intervals will lie below their respective π_i 's. This tendency to asymmetry will increase as the total error rate decreases, making the F -based Scheffé procedure particularly vulnerable to this effect. Since Scheffé intervals differ from simple Bonferroni intervals only through the critical constant used, asymmetry is also to be expected in the latter though it should not be as severe given that error rates for simple Bonferroni intervals are liberal. This is confirmed by study results, *e.g.*, $PER_U = 4.9\%$ for simple t -based Bonferroni intervals when $r = 50$, $k = 8$ and $a = 0.71$.

6.2 *t*-Based Transformed Bonferroni Intervals

Table 4 also gives percentage symmetry results for *t*-based Bonferroni intervals based on the log and logit transformations. The results of the table suggest that logit intervals do provide more symmetric coverage than the log intervals, when *k* is in the range 5 to 8. Thus logit intervals might be considered preferable in practice to log intervals from the point of view of error rate symmetry.

7. UNEQUAL CELL PROBABILITIES

Table 5 presents results on total error rates and error rate symmetry under unequal cell probabilities for the *t*-based log and logit transformed Bonferroni procedures, together with results for the first order modified Q-H procedure. Results are tabulated for six sets of unequal probabilities, three for the case $k = 5, \bar{\lambda} = 2, a = 0.5$, namely $\pi(5, 3, .3)$, $\pi(5, 2, .425)$ and $\pi(5, 1, .8)$, (see Section 4.1), and three for the case $k = 8, \bar{\lambda} = 2, a = 0.71$, namely $\pi(8, 3, .25)$, $\pi(8, 2, .35)$ and $\pi(8, 1, .65)$. For each π vector, the remaining $k - q$ elements all equal 0.05. Results for equiprobable cells are also displayed in Table 5 for comparison.

It can be seen that deviations from equiprobability do affect total error rates for the modified Q-H procedure, particularly when $k = 8$. With the first element $\pi_1 = 0.65$ the total error rate of modified Q-H is close to its error rate under equiprobability. For the other two cases studied ($\pi_1 = \pi_2 = .35$, and $\pi_1 = \pi_2 = \pi_3 = 0.25$), total error rates are considerably lower, closer in fact to the modified Q-H results obtained for the constant design effect case (see Thomas 1989). This difference in total error rates occurs because the pattern of cell design effects is different for each set of unequal probabilities, though the pattern of generalized design effects (the λ 's) remains the same ($\lambda_1 = 2 + 2\sqrt{3}, \lambda_j = 2 - \sqrt{3}/3, j = 2, \dots, 7$ for $\bar{\lambda} = 2, a = \sqrt{2}/2 = .707$). When $\pi_1 = 0.65$, the cell design effects are $d_1 = 5.7, d_i = 1.82, i = 2, \dots, 8$.

Table 5
 The Effect of Unequal Cell Probabilities on the Total Error Rates (ER_T) and Percentage Asymmetries (PER_U) of the Modified Q-H and Transformed Bonferroni Procedures;
 $r = 50, \bar{\lambda} = 2, a = 5\%, m = 10$

<i>k</i>	$\pi(k, q, \phi)$	Procedures					
		Modified Q-H (first order)		<i>t</i> -based Bonferroni			
		ER_T	PER_U	(log)		(logit)	
		ER_T	PER_U	ER_T	PER_U	ER_T	PER_U
5	$\pi(5, 1, 0.8)$	3.2	7.3	5.6	75.9	4.4	62.5
5	$\pi(5, 2, 0.425)$	1.4	82.1	4.8	57.2	4.6	47.8
5	$\pi(5, 3, 0.3)$	1.5	76.7	4.2	51.2	3.9	38.5
5	equi-prob.	2.0	45.0	4.5	61.1	4.0	48.8
8	$\pi(8, 1, 0.65)$	2.7	63.0	6.3	68.3	5.4	55.6
8	$\pi(8, 2, 0.35)$	0.6	83.3	4.9	58.2	4.4	51.2
8	$\pi(8, 3, 0.25)$	0.7	100	5.2	68.2	4.6	63.1
8	equi-prob.	2.5	66.5	6.0	64.0	5.2	49.0

Use of a uniform adjustment factor ($\hat{\lambda}$) will thus seriously underestimate the variance of the first estimated cell probability, leading to inflation of the error rate of the modified Q-H procedure. That the nominal error rate $\alpha = 5\%$ is not exceeded is due to the inherent conservativeness of modified Q-H intervals in the constant design effect case (see Section 5.3). When $\pi_1 = \pi_2 = 0.35$, corresponding design effects are $d_1 = d_2 = 2.36$, $d_i = 1.97$, $i = 3, \dots, 8$. These are much closer to constant design effects ($d_i = 2.0$, $i = 1, \dots, 8$) hence the conservative behaviour of the intervals in this case. It can also be seen from Table 5 that conservative ER_T 's are associated with highly asymmetric error rates.

Despite the variation in cell design effects implied by the different probability vectors of Table 5, it can be seen that the transformed Bonferroni procedures exhibit very stable performance. Total error rates (for 50 clusters) are close to the nominal rate ($\alpha = 5\%$) for both log and logit intervals, and neither exhibits serious asymmetry. Total error rates corresponding to unequal probabilities do decrease with decreasing r over the range $r = 50$ to $r = 15$ when $k = 8$ (results not shown). Variations in ER_T are not severe, however; when $r = 15$ clusters the minimum rate for the cases examined is approximately 2%.

8. SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

In the search for procedures that take direct account of the survey design and that provide adequate control of error rates and error rate symmetry over a wide range of problem and clustering situations, Scheffé intervals based on estimated cell variances must be rejected: the chi-squared version of equation (3) on the grounds of poor error control, and the F -based version on the grounds of extreme asymmetry. Modifications to Quesenberry-Hurst intervals are somewhat conservative, though the version based on the first order Rao-Scott correction does provide a viable procedure. For Bonferroni intervals, the benefits of using critical points of the t -distribution instead of the standard normal are substantial. Even so, intervals based on $\hat{\pi}$ and its square root provide inadequate control of total error rates, particularly for small numbers of clusters when the distribution of $\hat{\pi}$ becomes increasingly non-normal. On the other hand, t -based Bonferroni intervals using both the log and logit transformations provide good control of total error rates and error rate symmetry, and are clearly superior to all other competing intervals. Both log and logit transformed intervals (t -based) also appear to provide good control of error rates and error rate symmetry when the cell probabilities are unequal, differing in the cases studied by a ratio (maximum to minimum) of up to sixteen. From the point of view of total error rates there is little to choose between the log and logit intervals, though error rates for the latter are consistently a little lower. Logit intervals are superior from the point of view of symmetry, however. Estimates of confidence interval lengths (detailed results not shown) also favour the logit intervals, despite their slightly lower error rates. For example, for the equiprobable case with $\alpha = 5\%$, $k = 5$, $\bar{\lambda} = 2$, $a = 0.5$ and $r = 50$, the average length of the confidence interval on π_1 (expressed as a 95% confidence interval) was $.1915 \pm .0014$ for the log-based interval, and $.1850 \pm .0014$ for the logit-based interval. For the case of unequal probabilities, with $\alpha = 5\%$, $k = 8$, $\bar{\lambda} = 2$, $a = 0.71$, $r = 50$, $\pi_1 = 0.65$ and $\pi_2 = 0.05$ (see Table 5), 95% confidence intervals for the average lengths of log and logit intervals were: for π_1 , $.2865 \pm .0012$ and $.2776 \pm .0011$, respectively; for π_2 , $.0806 \pm .0010$ and $.0789 \pm .0011$, respectively.

Before final recommendations are made, it is necessary to consider possible limitations imposed by the design of the Monte Carlo study. A potentially limiting feature is the use of a single specific sampling design, namely two-stage cluster sampling with SRS at the second

stage, given that practitioners will encounter data collected using a range of survey designs that might include stratification and multiple levels of unit selection. For large samples, the relevant distribution theory requires knowledge only of first and second moments, assuming that a suitable central limit theorem applies (see for example Rao and Scott 1981). This study will therefore yield valid recommendations for large numbers of clusters, or more generally for large numbers of degrees of freedom for variance estimation (Rao and Thomas 1988), as long as the covariance matrix V/n and hence the generalized design effects can be appropriately modelled. Since the Dirichlet mixture model used in this study yields generalized design effects having means and coefficients of variation that are typical of those found in practice, recommendations based on a large number of clusters or degrees of freedom (fifty or more) can be made with confidence. For small to moderate numbers of clusters, quantitative results may differ from design to design. Since the basic mechanisms underlying the results exhibited in this study, namely increasing non-normality of $\hat{\pi}$ for decreasing r plus the inherent conservativeness of Scheffé-like procedures, will apply in general, it is expected that the qualitative trends for the different statistics examined will be generalizable across a wide variety of designs, even when the number of clusters is not large. The basic aim of the study has been to identify procedures whose control of error rates is robust to variations in the study parameters, namely the number of categories, the number of clusters, the strength of clustering, and the skewness of the vector of category probabilities. The combination of parameters examined has covered much of the range likely to be encountered in practice, so it is reasonable to suggest that the robustness exhibited by the log and logit transformed Bonferroni intervals might extend to variations in survey design, for moderate numbers of clusters (or degrees of freedom). Further research on this question is clearly required.

Subject to these caveats, t -based Bonferroni simultaneous confidence intervals based on the logit transformation are recommended for assessing up to $k = 12$ proportions of varying magnitude, under realistic clustering conditions. If conservativeness is deemed to be an asset, the first-order modified Quesenberry-Hurst procedure can be safely used. Both procedures require only a knowledge of the variances (or design effects) of the estimated cell proportions.

ACKNOWLEDGMENTS

I wish to thank J.N.K. Rao for many valuable discussions and for his comments on a draft of this paper. Thanks are also due to Steve Brockwell both for useful suggestions and for excellent programming assistance, and to Paul Bertelman for programming assistance during the latter part of the project. Finally I wish to thank two reviewers and an associate editor for suggestions that significantly improved the presentation of the results. This research was supported by a grant from the Natural Science and Engineering Research Council of Canada.

REFERENCES

- ANDREWS, R.W., and BIRDSALL, W.C. (1988). Simultaneous confidence intervals: a comparison under complex sampling. Paper presented at the 1988 American Statistical Association Annual Meeting, Chicago.
- BAILEY, B.J.R. (1980). Large sample simultaneous confidence intervals for the multinomial probabilities based on transformation of the cell frequencies. *Technometrics*, 22, 583-589.

- BLACK, D., and MYLES, J. (1986). Dependent industrialization and the Canadian class structure: a comparative analysis of Canada, the United States, and Sweden. *Canadian Review of Sociology and Anthropology*, 23, 157-181.
- BRIER, S.S. (1980). Analysis of contingency tables under cluster sampling. *Biometrika*, 67, 591-596.
- FELLEGI, I.P. (1980). Approximate tests of independence and goodness of fit based on stratified multistage samples. *Journal of the American Statistical Association*, 75, 261-268.
- FITZPATRICK, S., and SCOTT, A.J. (1987). Quick simultaneous confidence intervals for multinomial proportions. *Journal of the American Statistical Association*, 82, 875-878.
- GOLD, R.Z. (1963). Tests auxiliary to χ^2 tests in a Markov chain. *Annals of Mathematical Statistics*, 34, 56-74.
- GOODMAN, L.A. (1965). On simultaneous confidence intervals for multinomial proportions. *Technometrics*, 7, 247-254.
- HIDIROGLOU, M.A., and RAO, J.N.K. (1987). Chi-squared tests with categorical data from complex surveys: Part I – simple goodness-of-fit, homogeneity and independence in a two-way table with applications to the Canada Health Survey (1978-1979). *Journal of Official Statistics*, 3, 117-132.
- HOCHBERG, Y., and TAMANE, A.C. (1987). *Multiple Comparison Procedures*. New York: Wiley.
- HOLT, D., SCOTT, A.J., and EWING, P.D. (1980). Chi-squared tests with survey data. *Journal of the Royal Statistical Society, Ser. A*, 143, 303-320.
- JENNINGS, D.E. (1987). How do we judge confidence-interval adequacy? *The American Statistician*, 41, 335-337.
- MILLER, R.G. (1981). *Simultaneous Statistical Inference*, 2nd Edition. New York: Springer-Verlag.
- QUESENBERRY, C.P., and HURST, D.C. (1964). Large sample simultaneous confidence intervals for multinomial proportions. *Technometrics*, 6, 191-195.
- RAO, J.N.K., and SCOTT, A.J. (1979). Chi-squared tests for analysis of categorical data from complex surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 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. *Journal of the American Statistical Association*, 76, 261-230.
- 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.
- THOMAS, D.R. (1989). An investigation of simultaneous confidence interval procedures for proportions, under cluster sampling. Working Paper WPS 89-02, School of Business, Carleton University.
- 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.