# ADDITIONAL EVALUATION OF CHI-SQUARE METHODS FOR COMPLEX SAMPLES

Robert E. Fay, U.S. Bureau of the Census[1]
Statistical Methods Div., U.S. Bureau of the Census, Washington DC 20233

## 1. INTRODUCTION

The effects of complex sample designs on the analysis of categorical data have received considerable attention. Complex designs typically seriously affect the Pearson or likelihood-ratio chi-square tests for categorical models. A number of alternative tests have been proposed under various sets of assumptions about the nature of the complex design. Three approaches represent general solutions: the Wald test (Koch, Freeman, and Freeman 1975), adjustments to the original chi-square tests (Rao and Scott 1981, 1984), and the jackknifed chi-square test (Fay 1985).

The Wald test incorporates an estimate of the covariance matrix of the estimated cell frequencies into both estimation under the model and testing. Although this approach was the first of the three general solutions to be introduced, the specific manner in which the estimated covariance matrix of the estimated cell probabilities is incorporated leads to appreciable instability in many applications (Thomas and Rao 1984, 1987). Recently, Singh and Kumar (1986) proposed a modification to the Wald test to lessen the effect of this source of variability.

The procedures proposed by Rao and Scott (1981, 1984) employ the standard estimation methods -- maximum-likelihood estimation applied to the weighted cell estimates as if they were cell counts from a multinomial distribution -- and adjustments to the usual chi-square tests. Two principal forms of these procedures are available. In the first, relationships between the variance under the multinomial distribution and the sampling variance under the complex design are examined for both the cells and margins. These relationships are then incorporated into an adjustment factor by which the chi-square tests are divided. The procedure compares the resulting adjusted chi-square test to the chi-square distribution on the same number of degrees of freedom as appropriate under multinomial sampling. The method is particularly useful in applications to published tables if the required information about the variances under the complex design for the cells and marginal tables is also available.

The second method proposed by Rao and Scott incorporates an estimate of the covariance matrix under the complex design for the cells of the estimated cross-classification. In practice, this method requires returning to the original data to compute the estimated covariance matrix, since such matrices are rarely published. The method is again based on an adjustment to the original chi-square tests, but in this case both the test statistic and the original degrees of freedom are altered and interpreted according to an approximation due to Satterthwaite (1946). The second method requires more extensive calculation than the first, but its performance is sufficiently superior (Thomas and Rao 1984, 1987) that it clearly represents the preferred method over the first, when such calculations are possible.

The jackknifed chi-square test (Fay 1985) employs replication to determine the effect of the complex sample design on the original Pearson or likelihood-ratio tests. This method also employs standard maximum-likelihood estimators applied to the observed cell estimates, as do the methods of Rao and Scott. The behavior of the chi-square tests recomputed according to a replication method reflecting the complex sample design is used to derive a new test of significance. In earlier comparisons (Thomas and Rao 1984, 1985) the jackknifed test performed approximately as well overall as the method proposed by Rao and Scott employing the Satterthwaite approximation. In practice, application of the jackknifed test requires access to the original data in order to form the necessary replicate samples.

This paper will provide further evidence on the relative performance of the last two methods: the use of the Satterthwaite approximation as described by Rao and Scott, and the jackknifed chi-square tests. These methods have thus far appeared to give the best results (Thomas and Rao 1987). Both require access to the original data or derivation of summaries from the data that would not usually be produced otherwise.

Section 2 of the paper describes the form of the test statistics evaluated. The description in this section is intended to specify clearly what statistics are being assessed, but the reader is referred to the original references on these procedures for their derivation and details of their application.

Section 3 describes the Monte Carlo design and analytic approach. A choice exists on how the test proposed by Rao and Scott should actually be implemented: Section 4 presents a strategy not explicitly suggested by Rao and Scott that appears to give the best performance. Section 5 summarizes the results from the Monte Carlo evaluations and compares these findings to earlier work. Section 6 summarizes the conclusions. An appendix, excluded from the Proceedings for the sake of space but available from the author, presents more extensive comparisons derived from the Monte Carlo study.

## 2. TESTS EVALUATED IN THE STUDY

Both methods to be evaluated here are applicable to a variety of parametric models for cross-classified data. The scope of this paper, however, will be restricted to log-linear models, since these models represent most of the probable application of these methods.

Let the vector $\mathbf{Y} = \{Y_i\}$ denote an estimated cross-classification based on a complex sample design. $\mathbf{Y}$ may represent unweighted counts from a clustered sample or other complex design, or weighted estimates of totals for a population, as is frequently the case for federal and other national sample surveys. Let $N = \mathbf{e}'\mathbf{Y}$, where $\mathbf{e} = (1,1,\ldots,1)'$, denote the total estimated count for the table, and $\hat{\mathbf{p}} = N^{-1}\ \mathbf{Y}$ be the usual estimated cell proportions from the sample.

We will consider two log-linear models for $\mathbf{p}$, the vector of true proportions in the population.

The first model will be of the form:

$$\ln \mathbf{p} = u_1(\theta_1)\, \mathbf{e} + \mathbf{X}_1\, \theta_1, \qquad (2.1)$$

where $\ln\ \mathbf{p} = \{\ln p_i\}$ is the vector of log proportions, $\mathbf{X}_1$ represents a given design matrix, and $\theta_1$ denotes a vector of unknown parameters. The function $u_1(\theta_1)$ takes a value depending on the parameters $\theta_1$ such that the sum of probabilities over the cells of the table is one. The second model is

$$\ln \mathbf{p} = u(\theta)\, \mathbf{e} + \mathbf{X}\theta, \qquad (2.2)$$

where $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$, $\theta = (\theta_1', \theta_2')'$. Model (2.2) thus implies model (2.1) as a special case. The purpose is to test the improvement of model (2.2) over model (2.1) when model (2.2) is assumed to hold, i.e. to evaluate the hypothesis $\theta_2 = 0$. As a special case, (2.2) may be taken to be the saturated model, that is, the fully parameterized model that fits any set of positive probabilities exactly. In this case the test represents an evaluation of the overall fit of model (2.1).

Let $\hat{\mathbf{p}}_1 = \{p_{1i}\}$ denote the maximum-likelihood estimates of the cell proportions under model (2.1) based on the multinomial likelihood, and $\hat{\mathbf{p}}_2 = \{p_{2i}\}$ denote the corresponding estimates under model (2.2). The Pearson chi-square test for this comparison is given by

$$X^2 = N \sum_i (p_{1i} - p_{2i})^2/p_{1i}, \qquad (2.3)$$

and the likelihood-ratio chi-square by

$$G^2 = 2N \sum_i p_i \ln(p_{2i}/p_{1i}), \qquad (2.4)$$

The method employing the Satterthwaite approximation as proposed by Rao and Scott requires the matrix $\mathbf{P} = \{P_{ij}\}$, where

$$P_{ii} = p_i - p_i^2 \qquad (2.5)$$

$$P_{ii'} = -p_i p_{i'} \quad \text{for } i \neq i' \qquad (2.6)$$

and the estimated covariance matrix $\mathbf{V}$ equal to N times the estimated covariance matrix of $\hat{\mathbf{p}}$ under the complex sample design. Then, let

$$\tilde{\mathbf{X}}_2 = (\mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{P}\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{P})\mathbf{X}_2 \qquad (2.7)$$

$$\mathbf{M}^* = (\tilde{\mathbf{X}}_2'\mathbf{P}\tilde{\mathbf{X}}_2)^{-1}\, (\tilde{\mathbf{X}}_2'\mathbf{V}\tilde{\mathbf{X}}_2). \qquad (2.8)$$

The sum of the eigenvalues of $\mathbf{M}^*$ may be found as the trace of $\mathbf{M}^*$, and the trace of $\mathbf{M}^*\mathbf{M}^*$ gives the sum of squares of the eigenvalues of $\mathbf{M}^*$. The Satterthwaite approximation is to compute the integer $k'$ nearest to $(\mathrm{tr}(\mathbf{M}^*))^2/\mathrm{tr}(\mathbf{M}^*\mathbf{M}^*)$, and to compare $X^2_S = (k'/\mathrm{tr}(\mathbf{M}^*))X^2$ to the chi-square distribution on $k'$ degrees of freedom. (A further variation could be based on the incomplete gamma distribution without requiring rounding of $k'$ to an integer, but this refinement is unlikely to yield much additional improvement, except perhaps for values between 1 and 2, and is not considered here.)

The jackknifed chi-square test is based upon recomputing the chi-square tests, (2.3) and (2.4), or differences of (2.3) of two nested models, each compared to the saturated model (i.e., differences of the usual Pearson tests for two different models, where $\hat{\mathbf{p}}_2 = \hat{\mathbf{p}}$ in each calculation) for a series of replicate samples based on the sample data. Each replicate is of the form $\mathbf{Y} + \mathbf{W}^{(h,j)}$, $h=1,\ldots,H$, $j=1,\ldots,J_h$, where H represents a total number of strata, and $J_h$ the number of replicates in stratum h, and where

$$\sum_j \mathbf{W}^{(h,j)} = 0 \qquad (2.9)$$

for each h, such that the usual replication-based estimator of $\mathbf{V}^*$, the sampling covariance matrix of $\mathbf{Y}$ under the complex design, is given by

$$\mathbf{V}^* = \sum_h b_h \sum_j \mathbf{W}^{(h,j)} \otimes \mathbf{W}^{(h,j)}. \qquad (2.10)$$

In (2.10), $\otimes$ denotes the standard outer product, i.e., the usual cross-product matrix. If the same replication method is used to compute the matrix $\mathbf{V}$ in (2.8), the relationship is

$$N\,\mathbf{V} = \mathbf{V}^* + (\mathbf{e}'\mathbf{V}^*\mathbf{e})\hat{\mathbf{p}} \otimes \hat{\mathbf{p}} - \mathbf{V}^*\mathbf{e} \otimes \hat{\mathbf{p}} - \hat{\mathbf{p}} \otimes \mathbf{V}^*\mathbf{e} \qquad (2.11)$$

Details on how familiar replication methods, such as the jackknife and half-sample replication, may be represented in this form appear in Fay (1985).

Let $X^2_{(1)}(\mathbf{Y})$ denote the value of the Pearson chi-square test for evaluating the fit of $\hat{\mathbf{p}}_1$ and $X^2_{(2)}(\mathbf{Y})$ the test for $\hat{\mathbf{p}}_2$. Define

$$R_{hj} = \{X^2_{(1)}(\mathbf{Y}+\mathbf{W}^{(h,j)}) - X^2_{(2)}(\mathbf{Y}+\mathbf{W}^{(h,j)})\} - \{X^2_{(1)}(\mathbf{Y}) - X^2_{(2)}(\mathbf{Y})\} \qquad (2.12)$$

$$k^* = \sum_h b_h \sum_j R_{hj} \qquad (2.13)$$

$$v^* = \sum_h b_h \sum_j R_{hj}^2 \qquad (2.14)$$

$$X_J = \frac{\{X^2_{(1)}(\mathbf{Y}) - X^2_{(2)}(\mathbf{Y})\}^{1/2} - \{k^+\}^{1/2}}{\{v^*/\{8\{X^2_{(1)}(\mathbf{Y}) - X^2_{(2)}(\mathbf{Y})\}\}\}^{1/2}} \qquad (2.15)$$

where $k^+$ is $k^*$ when the latter is positive, 0 otherwise. A similar statistic, $G_J$, is obtained by replacing $X^2$ by $G^2$ throughout. The test procedure is to compare $X_J$ or $G_J$ to critical values tabulated in Fay (1983 or 1985).

There are close connections between the jackknife tests and procedures developed by Rao and Scott. In particular, under the asymptotic conditions considered by Rao and Scott (1984) and those in Fay (1985), $k^*$ and $\mathrm{tr}(\mathbf{M}^*)$ are consistent estimators of the same quantity under the null hypothesis. Furthermore, under some conditions (Fay 1985) the two test procedures are asymptotically equivalent.

## 3. MONTE CARLO DESIGN

The study examines the behavior of the test statistics for a 27-cell table representing a cross-classification of three variables, each with three levels. Ten test hypotheses are considered, using both $X^2$ and $G^2$:

1. The independence model, [1], fitting each margin separately.
2. The model, [2], specifying a two-way interaction between variables 1 and 2. This model is equivalent to the hypothesis of independence of variable 3 and the joint distribution of variables 1 and 2.
3. The model, [3], specifying the interactions of variables 1 and 2 and 2 and 3. This model is equivalent to the conditional independence of variables 1 and 3 given the value of variable 2.
4. The no-three factor interaction model, [4], which includes parameters corresponding to each pair of variables. This model requires iterative computation.
5. The comparison of [1] and [2].
6. The comparison of [1] and [3].
7. The comparison of [1] and [4].
8. The comparison of [2] and [3].
9. The comparison of [2] and [4].
10. The comparison of [3] and [4].

In parts of the study, a subset of these comparisons are used. Although the results for both $X^2$ and $G^2$ have been computed and saved, only the results for $X^2$ and the complex sample analogues are presented in this paper for tests of overall fit (hypotheses 1 through 4) and only the results for $G_J$ for model comparisons, since these choices represent the best general strategy (e.g., Fay 1983). Both $X^2_S$, based on (2.3), and $G^2_S$, the Satterthwaite correction to $G^2$, will be considered for model comparisons.

The Monte Carlo samples were generated on a PC (IBM-compatible at the AT level, with a math coprocessor) with a multiplicative congruent random number generator with modulus $2^{31}-1$. The multiplier was the first overall choice of Fishman and Moore (1984). Possible later work will examine the effect of substituting the method of Fushimi (1983), but the method chosen appeared to give satisfactory performance demonstrated by the behavior of the Pearson and likelihood-ratio statistics.

The study uses samples that are multinomial samples but that can also be treated as complex samples. For example, a multinomial sample of size 200 can be treated as a clustered sample of 20 clusters of 10 observations each; a clustered sample of 50 clusters of 4 observations each; a stratified clustered sample for 20 strata, each with 2 sample clusters of 5 observations each; a stratified clustered sample for 50 strata, each with 2 sample clusters of 2 observations each; or a stratified clustered sample for 5 strata, each with 5 sample clusters of 8 observations each. These options are used in this study, although not all five options are employed in each part of the study.

The principal interest is in methods designed for complex samples, but the study focuses on their performance for simple random samples. The intention here is to provide some notion of the performance of the methods for complex samples relative to the usual chi-square tests in a situation in which the latter are also suitable. Further comments on this choice are included in the concluding section.

## 4. REFINEMENT OF THE METHODS OF RAO AND SCOTT

The matrix **P**, defined by (2.5) and (2.6) and used in (2.7) and (2.8), plays an important part in the test defined by Rao and Scott. Some latitude is possible in its definition, since $p_i$ and $p_i'$ may be replaced in (2.5) and (2.6) by any consistent estimator of **p**, for example, $\hat{p}_1$ or $\hat{p}_2$ (in the case that $\hat{p}_2$ differs from $\hat{p}$). Rao and Scott (1984) were aware of this choice (e.g., Thomas and Rao 1987).

Table 1 offers some insight into the consequences in the estimation of **P** for moderate size samples. Two measures are considered; ideally, both should average approximately 1 for this problem. The first measure, $R_1$, represents the sum of the design effects compared to the degrees of freedom in the test. Because the Monte Carlo samples yield a multinomial distribution, each design effect is in fact 1. For measures of overall fit, use of $\hat{p}$ to define **P** can lead to systematic underestimates of the average design effect, although no discernable effect is seen for comparisons of models. The second ratio, $R_2$, represents a measure of the spread of

Table 1 Relative Performance of the Components of the Satterthwaite Approximation for Choices of **P**, Average Values of Ratios

| | For **P** based on: | | | |
|---|---|---|---|---|
| | $\hat{p}$ | | $\hat{p}_1$ | |
| | $R_1$[a] | $R_2$ | $R_1$ | $R_2$ |
| | 50 Clusters of 4 | | | |
| Model [1] | .98 | 1.31 | 1.00 | 1.46 |
| Model [4] | .93 | 1.12 | 1.00 | 1.20 |
| [1] - [2] | 1.01 | 1.07 | 1.00 | 1.08 |
| [1] - [4] | 1.01 | 1.20 | 1.00 | 1.26 |
| [3] - [4] | 1.01 | 1.07 | 1.00 | 1.08 |
| | 20 Clusters of 10 | | | |
| Model [1] | .98 | 1.95 | 1.00 | 2.13 |
| Model [4] | .93 | 1.38 | 1.00 | 1.47 |
| [1] - [2] | 1.01 | 1.21 | 1.01 | 1.22 |
| [1] - [4] | 1.00 | 1.60 | 1.00 | 1.68 |
| [3] - [4] | 1.00 | 1.21 | 1.00 | 1.22 |
| | 50 Clusters of 2 | | | |
| Model [1] | .88 | 1.18 | 1.00 | 1.53 |
| Model [4] | .69 | 1.06 | .99 | 1.32 |
| [1] - [2] | 1.01 | 1.04 | 1.00 | 1.08 |
| [1] - [4] | 1.01 | 1.13 | 1.00 | 1.28 |
| [3] - [4] | 1.00 | 1.04 | 1.00 | 1.08 |
| | 20 Clusters of 5 | | | |
| Model [1] | .88 | 1.75 | 1.00 | 2.20 |
| Model [4] | .70 | 1.24 | .99 | 1.57 |
| [1] - [2] | 1.01 | 1.18 | 1.00 | 1.22 |
| [1] - [4] | 1.01 | 1.53 | 1.00 | 1.68 |
| [3] - [4] | 1.00 | 1.18 | 1.00 | 1.22 |

Note: Based on Monte Carlo sample sizes of 1000. The standard errors for the averages are approximately .01 or less.

[a] $R_1 = tr(M^*)/k$, where k is the degrees of freedom of the test, and $R_2 = tr(M^*M^*)k'/tr(M^*)^2$, where k′ is the degrees of freedom of $\hat{M}$, which may be different from k for **P** based on $\hat{p}$.

the eigenvalues, equal to 1 plus the relative variance of the eigenvalues. With $k'$ equal to the rank of $\mathbf{M}$, $\hat{\mathbf{p}}$ yields better ratios than $\hat{\mathbf{p}}_1$, although in many cases the improvement is not dramatic.

If a choice was required between $\hat{\mathbf{p}}$ or $\hat{\mathbf{p}}_1$ for all calculations, the relatively more important role of $R_1$ would favor $\hat{\mathbf{p}}_1$. In this paper, however, $\mathbf{p}_1$ will be employed to compute $R_1$ and $\hat{\mathbf{p}}$ to compute $R_2$. In other words, $k'$ will be defined as the integer closest to $(tr(\mathbf{M}^*_2))^2/tr(\mathbf{M}^*_2\mathbf{M}^*_2)$, where $\mathbf{M}^*_2$ is based on $\hat{\mathbf{p}}$. The adjusted test is $X^2_S = (k'/tr(\mathbf{M}^*_1))\ X^2$, where $\mathbf{M}_1$ is based on $\hat{\mathbf{p}}_1$. This test represents almost twice the computation of the test based only on $\hat{\mathbf{p}}_1$, but it appears sensible to use the best available version of the Rao and Scott tests in this study.

## 5. RESULTS

### 5.1 Type I Error Rates

Tables 2 and 3 show the type I error rates for the chi-square tests, the adjusted tests based on the Satterthwaite correction, and the jackknifed tests, that is, their actual rejection rates when the null hypothesis is true. The tests were evaluated at the nominal .05 level. The same random seed formed the basis for the random samples for each given number of observations, so that comparisons across different sets for the same sample size are far more accurate than comparisons based on independent samples of the same size would be. The appendix contains further information on the rates of agreements among the tests.

The presumed probabilities satisfied model [1], i.e., the independence model, and consequently all other models as well. The marginal proportions were taken to be .2, .3, and .5 for each of the three variables. At 100 observations, consequently, most samples included at least one sample zero.

In Table 2, the actual rejection rates by $X^2$ for test of overall fit is consistently closer to the nominal level than the complex sample alternatives for these multinomial samples, but not by a great deal. A conservative tendency may be noted for $X^2_S$, especially for calculations based on 20 clusters instead of 50. The number of clusters has less effect on the performance of $X_J$ over this range, but $X_J$ rejects too often at sample sizes of 100. Fay (1983) presented further empirical evidence that cells with sample zeros have a greater effect on $X_J$ than $X^2$.

The complex sample alternatives perform essentially as well as the original chi-square tests for model comparisons, over the range considered in Table 3. Again, some relative improvement in the performance of $X^2_S$ and $G^2_S$ may be noted for 50 clusters relative to 20.

Table 2 Rejection Rates for the Pearson Chi-Square Tests of Overall Fit, for Their Jackknifed Versions, and for the Versions Based on the Satterthwaite Correction, at the Nominal 5 Percent Level, as Percentages, Under the Null Hypothesis

| | | Simple 20 | | Simple 50 | |
|---|---|---|---|---|---|
| | $X^2$ | $X^2_S$ | $X_J$ | $X^2_S$ | $X_J$ |
| | | 200 Observations | | | |
| Model [1] | 5.5 | 1.0 | 5.6 | 2.8 | 6.1 |
| Model [2] | 4.9 | 1.6 | 6.0 | 2.7 | 6.4 |
| Model [3] | 4.0 | 1.7 | 6.2 | 2.7 | 5.5 |
| Model [4] | 6.4 | 3.1 | 6.9 | 3.7 | 7.0 |
| | | 100 Observations | | | |
| Model [1] | 5.2 | 1.4 | 5.7 | 2.9 | 6.0 |
| Model [2] | 5.1 | 1.0 | 8.5 | 2.3 | 9.1 |
| Model [3] | 5.4 | .8 | 10.8 | 2.1 | 10.7 |
| Model [4] | 6.8 | 1.6 | 10.3 | 1.9 | 9.6 |

Note: Based on Monte Carlo sample sizes of 1000 and a table of 27 cells, for 20 and 50 clusters under the simple jackknife. See text for an explanation of the models.

Table 3 Rejection Rates for the Likelihood-Ratio and Pearson Chi-Square Tests of Model Comparisons, for Their Jackknifed Versions, and for the Versions Based on the Satterthwaite Correction, at the Nominal 5 Percent Level, as Percentages, Under the Null Hypothesis

| | $G^2$ | $G^2_S$ | $G_J$ | $X^2$ | $X^2_S$ |
|---|---|---|---|---|---|
| | 200 Observations, 20 Clusters | | | | |
| [1] - [2] | 5.0 | 4.2 | 5.8 | 4.5 | 3.7 |
| [1] - [3] | 5.3 | 3.4 | 4.6 | 5.0 | 3.7 |
| [1] - [4] | 5.3 | 2.4 | 4.4 | 5.7 | 2.6 |
| [2] - [3] | 4.7 | 3.5 | 4.8 | 4.2 | 2.7 |
| [2] - [4] | 5.0 | 3.8 | 4.4 | 5.0 | 3.5 |
| [3] - [4] | 4.5 | 4.7 | 5.0 | 4.3 | 4.2 |
| | 200 Observations, 50 Clusters | | | | |
| [1] - [2] | 5.0 | 4.7 | 5.2 | 4.5 | 4.2 |
| [1] - [3] | 5.3 | 3.2 | 4.3 | 5.0 | 3.7 |
| [1] - [4] | 5.3 | 4.5 | 4.6 | 5.7 | 4.5 |
| [2] - [3] | 4.7 | 4.0 | 4.1 | 4.2 | 3.5 |
| [2] - [4] | 5.0 | 3.7 | 3.9 | 5.0 | 3.5 |
| [3] - [4] | 4.5 | 4.3 | 4.2 | 4.3 | 3.7 |
| | 100 Observations, 20 Clusters | | | | |
| [1] - [2] | 6.1 | 5.4 | 4.9 | 4.9 | 4.1 |
| [1] - [3] | 5.9 | 4.3 | 4.0 | 5.3 | 3.8 |
| [1] - [4] | 7.2 | 4.1 | 4.5 | 7.1 | 3.8 |
| [2] - [3] | 5.3 | 4.1 | 4.3 | 4.2 | 3.3 |
| [2] - [4] | 7.2 | 5.5 | 5.0 | 5.6 | 4.0 |
| [3] - [4] | 7.4 | 6.6 | 6.0 | 6.4 | 4.7 |
| | 100 Observations, 50 Clusters | | | | |
| [1] - [2] | 6.1 | 5.7 | 4.4 | 4.9 | 4.5 |
| [1] - [3] | 5.9 | 4.1 | 4.1 | 5.3 | 3.6 |
| [1] - [4] | 7.2 | 6.1 | 4.6 | 7.1 | 5.7 |
| [2] - [3] | 5.3 | 4.7 | 3.6 | 4.2 | 3.7 |
| [2] - [4] | 7.2 | 5.6 | 4.3 | 5.6 | 3.6 |
| [3] - [4] | 7.4 | 6.8 | 5.4 | 6.4 | 5.3 |

Note: Based on Monte Carlo sample sizes of 1000 and a table of 27 cells, for 20 and 50 clusters under the simple jackknife. See text for an explanation of the models.

## 5.2 Type II Error Rates

Tables 4 and 5 examine the power of the tests, i.e., the tests' abilities to reject when the null hypothesis is false. In both tables, the actual probabilities satisfy model [3]. In Table 4, the interaction patterns favor more sample zeros than the independence model; in Table 5, the distribution of probabilities leads to fewer sample zeros than under independence.

In Table 4, the jackknifed tests of overall fit are the most powerful, even compared to the Pearson test. In contrast, the jackknifed tests of overall fit are not as powerful as the Pearson in Table 5. In other words, the jackknifed tests of overall fit appear to be particularly sensitive to sample zeros, at some cost in power against alternative hypotheses seldom yielding sample zeros. This behavior appears to originate in the jackknifed test's use of replication to estimate variability: a cell with a sample estimate of zero contributes to the overall chi-square but is estimated to have no variance under replication. This pattern was analyzed earlier (Fay 1983) in more detail. The performance of $X^2_S$ lags just behind $X_J$ in Tables 4 and 5. Again, calculations based on 50 clusters show some advantage over 20.

Table 4 Rejection Rates for the Pearson and Likelihood-Ratio Chi-Square Tests, for Their Jackknifed Versions, and for the Versions Based on the Satterthwaite Correction, at the Nominal 5 Percent Level, as Percentages, Under a Fixed Alternative

| | Sample Size | | | |
|---|---|---|---|---|
| | 100 | 200 | 400 | 800 |
| Model [1] | | | | |
| $X^2$ | 11.2 | 24.6 | 59.7 | 96.1 |
| $X^2_S$ Sm 20 | 4.7 | 12.5 | 43.1 | 89.2 |
| $X^2_S$ Sm 50 | 8.9 | 20.7 | 55.4 | 94.7 |
| $X_J$ Sim 20 | 15.9 | 30.0 | 64.4 | 95.7 |
| $X_J$ Sim 50 | 17.5 | 32.1 | 67.4 | 97.5 |
| $X_J$ Str 50 | 25.0 | 36.7 | 68.8 | 97.4 |
| Model [2] | | | | |
| $X^2$ | 6.4 | 11.5 | 28.6 | 65.8 |
| $X^2_S$ Sm 20 | 3.1 | 4.6 | 17.2 | 49.2 |
| $X^2_S$ Sm 50 | 4.6 | 8.0 | 23.4 | 58.8 |
| $X_J$ Sim 20 | 16.2 | 17.3 | 36.9 | 70.0 |
| $X_J$ Sim 50 | 16.8 | 19.1 | 38.0 | 72.0 |
| $X_J$ Str 50 | 25.5 | 25.9 | 40.7 | 74.9 |
| [1] - [3] | | | | |
| $G^2$ | 30.5 | 52.1 | 85.8 | 99.6 |
| $G^2_S$ Sm 20 | 26.9 | 47.2 | 81.5 | 99.3 |
| $G^2_S$ Sm 50 | 28.6 | 48.8 | 81.7 | 99.4 |
| $G_J$ Sim 20 | 25.7 | 45.6 | 83.7 | 99.4 |
| $G_J$ Sim 50 | 26.6 | 45.7 | 83.7 | 99.6 |
| $G_J$ Str 50 | 34.1 | 51.7 | 85.7 | 99.7 |
| $X^2$ | 22.9 | 46.4 | 83.1 | 99.6 |
| $X^2_S$ Sm 20 | 20.3 | 42.6 | 78.5 | 99.0 |
| $X^2_S$ Sm 50 | 21.8 | 43.8 | 78.8 | 99.2 |
| [2] - [3] | | | | |
| $G^2$ | 21.5 | 32.8 | 60.3 | 93.9 |
| $G^2_S$ Sm 20 | 20.2 | 30.5 | 55.7 | 91.8 |
| $G^2_S$ Sm 50 | 22.5 | 34.4 | 59.8 | 93.3 |
| $G_J$ Sim 20 | 20.1 | 30.7 | 59.6 | 92.8 |
| $G_J$ Sim 50 | 19.7 | 30.3 | 59.5 | 93.1 |
| $G_J$ Str 50 | 23.6 | 35.4 | 62.0 | 94.0 |
| $X^2$ | 14.1 | 28.0 | 55.0 | 93.3 |
| $X^2_S$ Sm 20 | 13.0 | 24.6 | 50.6 | 90.1 |
| $X^2_S$ Sm 50 | 14.2 | 27.9 | 55.5 | 92.2 |

Note: Based on Monte Carlo sample sizes of 1000 and a table of 27 cells. See text for an explanation of the models. Model [3] fits the data. "Sm" and "Sim" denote the simple jackknife, with the stated number of clusters; "Str" 50 indicates the stratified jackknife with two clusters in each of 50 strata.

Table 5 Rejection Rates for the Pearson and Likelihood-Ratio Chi-Square Tests, for Their Jackknifed Versions, and for the Versions Based on the Satterthwaite Correction, at the Nominal 5 Percent Level, as Percentages, Under a Fixed Alternative

| | Sample Size | | | |
|---|---|---|---|---|
| | 100 | 200 | 400 | 800 |
| Model [1] | | | | |
| $X^2$ | 19.8 | 35.3 | 70.2 | 94.7 |
| $X^2_S$ Sm 20 | 6.0 | 14.2 | 40.5 | 82.6 |
| $X^2_S$ Sm 50 | 9.2 | 21.1 | 52.2 | 89.4 |
| $X_J$ Sim 20 | 12.5 | 24.1 | 57.8 | 91.0 |
| $X_J$ Sim 50 | 13.3 | 25.6 | 60.4 | 92.4 |
| Model [2] | | | | |
| $X^2$ | 11.5 | 20.1 | 37.9 | 69.0 |
| $X^2_S$ Sm 20 | 2.8 | 7.6 | 18.3 | 46.0 |
| $X^2_S$ Sm 50 | 5.2 | 11.8 | 24.8 | 55.5 |
| $X_J$ Sim 20 | 13.3 | 17.0 | 32.9 | 61.9 |
| $X_J$ Sim 50 | 12.9 | 17.7 | 34.0 | 63.8 |
| [1] - [3] | | | | |
| $G^2$ | 25.4 | 44.4 | 81.5 | 99.0 |
| $G^2_S$ Sm 20 | 16.6 | 30.5 | 70.8 | 96.2 |
| $G^2_S$ Sm 50 | 16.6 | 30.9 | 73.0 | 97.1 |
| $G_J$ Sim 20 | 19.1 | 40.4 | 78.8 | 97.7 |
| $G_J$ Sim 50 | 19.8 | 39.6 | 80.4 | 98.4 |
| $X^2$ | 29.4 | 49.1 | 83.1 | 99.2 |
| $X^2_S$ Sm 20 | 20.0 | 36.9 | 74.1 | 96.9 |
| $X^2_S$ Sm 50 | 20.5 | 36.6 | 76.5 | 97.9 |
| [2] - [3] | | | | |
| $G^2$ | 18.4 | 30.7 | 57.2 | 89.1 |
| $G^2_S$ Sm 20 | 12.4 | 20.1 | 46.7 | 83.4 |
| $G^2_S$ Sm 50 | 14.3 | 22.6 | 50.2 | 85.8 |
| $G_J$ Sim 20 | 14.6 | 27.3 | 55.6 | 88.4 |
| $G_J$ Sim 50 | 14.4 | 27.0 | 55.2 | 89.3 |
| $X^2$ | 18.3 | 31.4 | 58.9 | 89.5 |
| $X^2_S$ Sm 20 | 12.9 | 21.0 | 48.7 | 84.2 |
| $X^2_S$ Sm 50 | 14.8 | 24.4 | 52.2 | 86.5 |

Note: Based on Monte Carlo sample sizes of 1000 and a table of 27 cells. See text for an explanation of the models. Model [3] fits the data. See the notes to Table 4 for abbreviations.

Tables 4 and 5 show $G_J$ falling just behind $G^2$. The performance of $G^2_S$ is slightly less good, but still acceptable. The relative advantage of $X^2$ and $G^2$ varies with the problem, as do the relative virtues of $X^2_S$ and $G^2_S$.

Appendix Table A.1 presents additional type I evaluation for the jackknifed tests. Tests of overall fit with $X_J$ calculated under a jackknife for the two-per-stratum case appear to be adversely affected by small sample sizes to a greater degree than those for simple clustering. No comparable results are yet available for $X^2_S$.

## 6. CONCLUSIONS

Thomas and Rao (1987) note the closeness in performance between the jackknifed test and the Satterthwaite approximation, but imply a slight edge to the latter. If the results of this paper are used instead, the comparison might lean slightly the other direction. The use of the multinomial distribution in this paper may have thrown an advantage to the jackknifed tests. The overall situation appears close to a draw.

In fact, the major conclusion from this study is that both tests perform almost as well as the familiar $X^2$ and $G^2$ that have been the recognized standards for multinomial samples. Both the jackknifed and Satterthwaite versions are designed for complex samples under very few assumptions, yet compare favorably to tests that rest heavily on the multinomial assumption. Consequently, there appears to be relatively little room left for improvement on these two complex sample methods, and their use should be strongly encouraged for any complex sampling situation.

Acknowledgements: The author would like to thank Lynn Weidman and Nathaniel Schenker for helpful comments on an earlier draft.

Footnote:

[1] This paper reports research undertaken by a member of the Census Bureau's staff. The views expressed are attributable to the author and do not necessarily reflect those of the Census Bureau.

## References

Fay, R.E. (1983), "CPLX - Contingency Table Analysis for Complex Samples, Program Documentation," unpublished report, U.S. Bureau of the Census.

______ (1985), "A Jackknifed Chi-Square Test for Complex Samples," Journal of the American Statistical Association, 80, 148-157.

Fishman, G.S., and Moore, L.R. (1984), "An Exhaustive Analysis of Multiplicative Congruent Random Number Generators with Modulus $2^{31}$-1," Technical Report UNC/ORSA/TR-84/5, University of North Carolina at Chapel Hill.

Fushimi, M. (1983), "Increasing the Orders of Equidistribution of the Leading Bits of the Tausworthe Sequence," Information Processing Letters, 16, 189-192.

Koch, G.G., Freeman, D.H., and Freeman, J.L. (1975), "Strategies in the Multivariate Analysis of Data from Complex Surveys," International Statistical Review, 43-78.

Rao, J.N.K., and Scott, A.J. (1981), "The Analysis of Categorical Data from Complex Samples: Chi-Squared Tests for Goodness-of-Fit and Independence in Two-Way Tables," Journal of the American Statistical Association, 76, 221-230.

______ (1984), "On Chi-Squared Tests for Multiway Contingency Tables With Cell Proportions Estimated from Survey Data," Annals of Statistics, 12, 46-60.

Satterthwaite, F.E. (1946), "An Approximate Distribution of Estimates of Variance Components," Biometrics, 2, 110-114.

Singh, A.V. and Kumar, S. (1986), "Categorical Data Analysis for Complex Samples," Proceedings of the Section on Survey Research Methods, American Statistical Association, Washington, DC.

Thomas, D.R., and Rao, J.N.K. (1984), "A Monte Carlo Study of Exact Levels of Goodness-of-Fit Statistics Under Cluster Sampling," Proceedings of the Section on Survey Research Methods, Washington DC: American Statistical Association, 207-211.

______ (1985), "On the Power of Some Goodness-of-Fit Tests Under Cluster Sampling," Proceedings of the Section on Survey Research Methods, Washington, DC: American Statistical Association, 291-296.

______ (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.