ON THE POWER OF SOME GOODNESS-OF-FIT TESTS UNDER CLUSTER SAMPLING

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

Introduction

It is well known that the Pearson chi-squared test (X^2) and the liklihood ratio test (G^2) can yield unacceptably large significance levels under cluster sampling, and a number of alternatives have been proposed. In a recent study, Thomas and Rao (1984) compared the finite sample significance levels of these alternative test procedures for the case of a simple goodness-of-fit test, under simulated cluster sampling. From a study of a number of variants of the basic tests, they reduced the main comparison to four procedures, namely an F-based version of the Rao-Scott $\bar{\lambda}$ adjusted X^2 statistic (Rao and Scott, 1981), the Rao-Scott Satterthwaite adjusted X^2, Fay's jackknifed X^2 (Fay, 1985), and a modified Wald statistic referred to an F distribution (Fellegi, 1980; Hidiroglou et al., 1980). Except for the $\bar{\lambda}$-adjustment, these test procedures require knowledge of the estimated covariance matrix of cell estimates, while the $\bar{\lambda}$-adjustment depends only on cell design effects, or variance estimates. As expected, Thomas and Rao found that the F-based version of the $\bar{\lambda}$ adjusted X^2 statistic yielded significance levels close to the nominal level when the variation among the eigenvalues of the design effect matrix was small. In general, the Satterthwaite adjusted test and Fay's jackknifed test performed well even when variation among the eigenvalues was appreciable. They also found that for uniform probability vectors, $\underset{\sim}{\Pi}_0$, the modified Wald statistic controlled significance levels reasonably well.

Though adequate control of significance levels is essential if a statistic is to be useful, no comparison of competing statistics is complete without a comparison of their powers. This paper thus describes a Monte Carlo study of the power of the above statistics, based on the cluster sampling model used in the earlier work.

Methodology and Study Design

The study is based on simulated two-stage cluster sampling in which a k-category sample of m units is drawn independently from each of r sampled clusters, giving a total sample size n = mr. Details of the model, and of the random number generation scheme are given by Thomas and Rao (1984).

The Parameters

As in the earlier study, the degree of clustering is categorized by two parameters, $\bar{\lambda}$ and a, which represent the mean and coefficient of variation, respectively, of the eigenvalues of the 'generalized design effect matrix' (Rao and Scott, 1981). Using these two parameters, a range of cluster sampling situations can be modelled, namely: (i) multinomial sampling ($\bar{\lambda}$=1, a=0); (ii) constant design effect clustering ($\bar{\lambda}$>1, a=0); (iii) non-constant design effect clustering ($\bar{\lambda}$>1, a >0).

To keep the size of the study to a manageable level, all experiments, each consisting of 1000 independent trials, have been run at the traditional significance level α = 5%, and the reported results concentrate primarily on the case k=5, $\bar{\lambda}$ = 2, under the equiprobable null hypothesis $\underset{\sim}{\Pi}$ = (1/k,...,1/k)'. Further, the true alternative probability vector $\underset{\sim}{\Pi}$ has been restricted to the class $\underset{\sim}{\Pi}(k,q,\beta)$ defined by the vector elements

$$\pi_j(k,q,\beta) = 1/k + \beta; \quad j=1,\ldots,q,$$

$$= 1/k - q\beta/(k-q); \quad j=q+1,\ldots,k.$$

It should be noted that this class includes the set of alternatives explored by Read (1984), in his study of the 'power divergence' family of goodness-of-fit statistics.

Power Estimates and Standard Errors

The power estimates reported in this study represent, for each parameter setting, the percentage of the 1000 Monte Carlo trials in which the test statistic exceeded its nominal 5% level, leading to a correct rejection of H_o. Binomial standard errors of these point estimates of power are given in the footnote to Table 1. All test statistics were evaluated using the same set of random numbers, in order to improve the precision of power comparisons between competing statistics. Approximate standard errors of estimated power differences are given by Thomas and Rao (1985). In this paper, the discussion will be confined to power differences that are large enough to have practical importance, and which exceed their standard errors by a factor of at least two.

The Test Statistics

Only a brief summary of the four competing test statistics is given here. Details can be found in Thomas and Rao (1984).

F-based Versions of the Rao-Scott $\bar{\lambda}$ Corrections

The Rac-Scott corrected X^2 procedure refers $X^2_{c_o} = X^2/\hat{\bar{\lambda}}_o$ to χ^2_{k-1}, where $\hat{\bar{\lambda}}_o=(k-1)^{-1}\Sigma(1-\pi_{oi})\hat{d}_{oi}$. Here $\hat{d}_{oi} = \hat{v}_{ii}/[\pi_{oi}(1-\pi_{oi})]$ is the i^{th} estimated cell design effect, and $\hat{v}_{ii}$ is the i^{th} diagonal element of $\hat{\underset{\sim}{V}}$, the sample estimate of $\underset{\sim}{V}$, which is n times the covariance matrix of the estimated cell proportions $\hat{\underset{\sim}{\Pi}} = (\hat{\pi}_1,\ldots,\hat{\pi}_{k-1})$. The recommended F-based version, denoted $FX^2_{c_o}$, is obtained by referring $X^2_{c_o}/(k-1)$ to an F-distribution on (r-1) and (r-1)(k-1) degrees of freedom.

Versions of the $\bar{\lambda}$-corrected statistics can also be based on an alternative consistent estimator of $\bar{\lambda}$, denoted $\hat{\bar{\lambda}}$, obtained by replacing π_{oi} and $\hat{d}_{oi}$ in the above expression by $\hat{\pi}_i$ and $\hat{d}_i = \hat{v}_{ii}/[\hat{\pi}_i(1-\hat{\pi}_i)]$. Since Thomas and Rao (1984) found that $FX^2_{c_o}$ gives better control of significance levels than does FX^2_c, the statistic based on $\hat{\bar{\lambda}}$, primary attention will focus on the former. However, power differences between $FX^2_{c_o}$ and FX^2_c

will be summarized. Also compared will be the powers of $FX^2_{c_o}$ and the analogous statistic derived from the likelihood ratio test, namely $FG^2_{c_o}$.

The Rao-Scott Satterthwaite Correction

This test procedure refers $X^2_{S_o} = X^2_{c_o}/(1+\hat{a}^2_o)$ to a chi-squared distribution on $(k-1)/(1+\hat{a}^2_o)$ degrees of freedom. The estimate $\hat{a}_o$ depends only on the elements $\hat{v}_{ij}$ of $\hat{\underset{\sim}{V}}$, and the hypothesized probabilities $\underset{\sim}{\Pi}_o$ (see Rao and Scott, 1981). As with the $\bar{\lambda}$-corrected tests, an alternative estimator $\hat{a}$ is available which uses $\hat{\underset{\sim}{\Pi}}$ in place of $\underset{\sim}{\Pi}_o$; the corresponding statistic is denoted X^2_S. Primary attention will focus on $X^2_{S_o}$, though the differences in power between X^2_S and $X^2_{S_o}$, the form recommended by Thomas and Rao (1984), will be summarized. A Satterthwaite version of G^2, namely $G^2_{S_o}$ can be defined analogously, and the power of this variant will also be compared to that of $X^2_{S_o}$.

Fay's Jackknifed Tests

Fay's (1985) modification of the X^2 procedure is based on the statistic

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

where the normalization constants K_J and V_J are obtained by jackknifing the statistic X^2/n. X_J, and the corresponding G^2 version G_{J_2}, are referred to the critical points of $\sqrt{2}[(\chi^2_{k-1})^{\frac{1}{2}}-(k-1)^{\frac{1}{2}}]$.

The Modified Wald Statistic

The basic Wald procedure refers

$$X^2_w = n(\hat{\underset{\sim}{\Pi}} - \underset{\sim}{\Pi}_o)' \hat{\underset{\sim}{V}}^{-1} (\hat{\underset{\sim}{\Pi}} - \underset{\sim}{\Pi}_o)$$

to a chi-square distribution on (k-1) degrees of freedom. The modified procedure (Fellegi, 1980; Hidiroglou et al., 1980) refers

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

Estimated Powers of the Primary Test Statistics

This section presents and compares estimates of test power for the four statistics $FX^2_{c_o}$, $X^2_{S_o}$, X_J and F_w under an equiprobable null hypothesis. The case k=5 will first be considered in detail, and then results for k=3 and k=10 will be examined to see whether or not the identified trends persist for a wider range of values of k.

The Case of Five Categories

Two settings of the true probabilities $\underset{\sim}{\Pi}$ are examined for each value of k. For k=5, these are given by $\underset{\sim}{\Pi}(5,1,\pm 0,1)$, yielding the $\underset{\sim}{\Pi}$ vectors $(.1,.225,\ldots,.225)'$ and $(.3,.175,\ldots,.175)'$ respectively. For each setting of $\underset{\sim}{\Pi}$, four different numbers of clusters (r=50,30,20,10) and two non-multinomial clustering setups are considered, namely $\bar{\lambda}=2$, a=0 (constant design effects) and $\bar{\lambda}=2$, a=0.5 (non-constant design effects). The results are displayed in Table 1.

Table 1

Powers[1] of the Rao-Scott, Fay and Modified Wald Tests, as a function of r, a and β, for k=5

$\bar{\lambda} = 2$; m = 10

a	β	r	$FX^2_{c_o}$	$X^2_{S_o}$	X_J	F_w
0.0	-0.1	50	94.1	94.0	95.2	94.6
		30	71.5	69.9	77.0	76.8
		20	47.7	44.9	57.8	59.0
		10	22.4	19.3	32.6	31.7
	+0.1	50	85.3	84.8	84.6	76.4
		30	61.8	60.8	61.1	54.7
		20	39.7	38.4	40.4	33.6
		10	21.8	20.4	23.8	17.6
0.5	-0.1	50	89.0	88.1	84.9	75.4
		30	69.1	68.8	65.0	53.9
		20	55.1	54.3	52.0	40.3
		10	34.4	32.7	39.8	34.8
	+0.1	50	73.8	67.9	63.8	45.3
		30	50.4	45.1	41.1	28.8
		20	36.0	29.3	27.2	19.1
		10	16.9	13.6	14.7	11.7

(1) Standard errors for point estimates of powers of magnitudes 95%, 90%, 80%, 50% (and their complements) are 0.7%, 0.9%, 1.3% and 1.6% respectively.

For constant design effects, and an alternative corresponding to β =-0.1, it can be seen from the top panel of Table 1 that X_J and F_w are equally powerful. Also, X_J and F_w are both more powerful than the Rao-Scott statistics, except when r=50, in which case all powers are close to 95%. Generally speaking, there is little to choose between $FX^2_{c_o}$ and $X^2_{S_o}$ in this case. Results for the second panel of Table 1 show somewhat different trends. Again under constant design effect clustering, but with an alternative corresponding to $\beta > 0$, F_w and X_J lose their superiority. In this case, X_J and the Rao-Scott statistics exhibit similar powers, with X_J having a slight edge for small numbers of clusters. However, all three statistics are appreciably more powerful than F_w. The third and fourth panels of Table 1 display power results for the more important case of non-constant design effect clustering, in which case the relative orderings differ. The Rao-Scott statistics are now more powerful than X_J, which is in turn much more powerful than F_w. This effect is particularly noticeable when $\beta > 0$. For example, when r=50 and β = 0.1, the power of X_J is 63.8% compared to 45.3% for F_w. As in the constant design effect case, therefore, the

modified Wald statistic, F_w, is much more sensitive to the form of the alternative than are its competitors.

It is worth noting that changing the form of the alternative from $\beta < 0$ to $\beta > 0$ lowers the power of all four statistics, for both settings of a. This is true even for relatively large numbers of clusters (r=50) under constant design effect clustering, in which case the Pitman powers of $FX^2_{C_o}$, X_J and F_w are all given by the same non-central chi-squared distribution. For the case $\bar{\lambda} = 2$, a = 0 and r = 50, the actual Pitman power is 90%. From Table 1 it can be seen that for negative β, the powers of all four statistics are greater than 90%, and are all less than 90% for positive β. Since the non-centrality parameters corresponding to the equiprobable null and the two alternatives $\pi(5,1,\pm 0.1)$ are equal, the reasons for this strong dependence of power on β are not immediately apparent. As noted above, F_w is particularly sensitive to the sign of β, and an explanation of this aspect of the phenomenon, based on an idea due to Larntz (1978), is given by Thomas and Rao (1985).

A Comparison of Trends for k=3, 5 and 10

Estimates of the powers of the four primary statistics for three and ten category tests are shown in Table 2, for the case of 30 clusters. These results were made comparable to those for k=5 by a suitable choice of the parameters β and a. First β was chosen to make the Pitman powers for k=3, 5 and 10 categories equal to 90%, for the specific case $\bar{\lambda} = 2$, a = 0, r = 50. Then, for comparisons in the non-constant design effect case, a, the coefficient of variation of the eigenvalues, was chosen to make a/a_{max} the same for all three values of k, a_{max} being the maximum attainable value of a.

Results for the constant design effect case are given in the top panel of Table 2. For $\beta < 0$, it can be seen that the power advantage of F_w and X_J over the Rao-Scott statistics depends on k. For k=10, the difference is greater than it is for k=5 ; for k=3, the difference disappears, all four statistics attaining similar powers in this case.

When $\beta > 0$, and a = 0, X_J and the Rao-Scott and Fay statistics again exhibit similar power (with the exception of $X^2_{S_o}$, when k = 10), all three statistics being markedly more powerful than F_w, especially when k = 10. For the case k = 10, it is also worth noting that $X^2_{S_o}$ is less powerful than $FX^2_{C_o}$, (54.5% versus 60.9%, when $\beta > 0$), a difference that is consistent with the slight conservativeness of the Satterthwaite corrected statistic for a = 0 and k = 10 that was noted by Thomas and Rao (1984).

Results for the non-constant design effects case, a > 0, are shown in the bottom panel of Table 2. Together with the results of Table 1, these show that, for moderate to large numbers of clusters ($r \geq 30$), the effect of increasing a is to lower the power of all four tests. As for the case k=5, the Rao-Scott statistics are again more powerful than X_J, when a > 0, and considerably more powerful than F_w. When $\beta < 0$, and a = .82, for example, the powers of $X^2_{S_o}$, X_J and F_w are 66.4%, 60.9% and 47.7% respectively. Results from Table 2 clearly indicate that the power of F_w relative to its competitors drops with increasing k. Also for k = 10, under non-constant design effect sampling, $FX^2_{C_o}$ is more powerful than $X^2_{S_o}$, a trend that is to be expected since $FX^2_{C_o}$ is known to exhibit inflated significance levels in this situation.

Table 2

A Comparison of the Power Trends of the Primary Statistics for k = 3 and 10.

$\bar{\lambda} = 2$; m = 10 for k=3; m = 20 for k=10; r=30

k	a	β	$FX^2_{C_o}$	$X^2_{S_o}$	X_J	F_w
3	0.0	-.11	72.7	74.0	74.5	71.9
		+.11	65.7	67.5	66.5	60.1
10	0.0	-.06	73.0	65.3	79.7	81.3
		+.06	60.9	54.4	58.6	35.5
3	0.29	-.11	71.6	72.7	69.8	63.8
		+.11	60.3	59.4	57.4	48.8
10	0.82	-.06	70.2	66.4	60.9	47.7
		+.06	36.7	23.5	20.4	12.7

The decreased power values associated with switching alternatives from $\beta < 0$ to $\beta > 0$ can again be seen in Table 2, and the effect clearly becomes more pronounced as k increases. For k = 10, for example, the power of F_w when a = 0 goes from 81% to 36% as β changes from -.06 to +.06. This effect also interacts with the previously noted effect on power of increasing a. For the constant design effect case, with k=10, switching the sign of β lowers the powers of the Rao-Scott statistics about 10 percentage points, and the Fay statistic about 20 points. With a = .82, however, the drop in power of these statistics is drastic, being approximately 33, 43 and 40 percentage points for $FX^2_{C_o}$, $X^2_{S_o}$ and X_J respectively. Once again, however, the relative effect on power is greatest for F_w. As a result of this statistic's sensitivity both to increases in a and to the form of the alternative, its power when k=10 and a = .82 is only about one half of that of $X^2_{S_o}$ (12.7% versus 23.5%).

Estimated Powers of Variants of the Primary Statistics

This section examines the power of two variants of the Rao-Scott and Fay tests discussed above: (i) basing the Rao-Scott and Fay statistics on G^2, the likelihood ratio statistic, rather than on Pearson's X^2 ; (ii) use of $\hat{\bar{\lambda}}$ and $\hat{a}^2$ in place of $\hat{\bar{\lambda}}_o$ and $\hat{a}^2_o$ in the definition of the

Rao-Scott statistics. In the interests of brevity, results will be presented only for the case $k=5$, but similar trends were observed for $k=3$ and $k=10$.

Table 3

A Comparison of X^2 and G^2 Versions of the Rao-Scott and Fay Tests

$k=5$; $\bar{\lambda}=2$; $m=10$; $r=30$

a	β	$FX^2_{c_o}$ ($FG^2_{c_o}$)	$X^2_{S_o}$ ($G^2_{S_o}$)	X_J (G_J)
0.0	-0.1	71.5 (77.0)	69.9 (75.6)	77.0 (77.3)
	+0.1	61.8 (59.0)	60.8 (58.3)	61.1 (59.2)
0.5	-0.1	69.1 (72.5)	68.8 (71.9)	65.0 (58.7)
	+0.1	50.4 (48.3)	45.1 (41.2)	41.1 (43.4)

X^2 versus G^2 Versions of the Rao-Scott and Fay Tests

In the earlier study, Thomas and Rao (1984) found that the significance levels of both X^2 and G^2 versions of the Rao-Scott and Fay tests were very similar, though where there were differences, they tended to favour the X^2 tests. Table 3 displays the powers of the X^2 and G^2 versions of these three tests, for the case $k=5$. Results, for 30 clusters, are shown for two settings of the alternative, $\Pi(5,1,\pm 0.1)$, and for two values of a.

Some marked differences in power can immediately be seen. For the case $\beta=-0.1$, G^2 versions of both Rao-Scott tests are more powerful than X^2 versions. This relationship is reversed when $\beta=+0.1$. These results are consistent with the findings of Koehler and Larntz (1980) and Read (1984), regarding the relative powers of X^2 and G^2. These investigators showed that when one element of $\underset{\sim}{\Pi}$ is decreased towards zero, then G^2 becomes more powerful than X^2. When the single element of $\underset{\sim}{\Pi}$ is increased, and the remaining elements decreased, the power relationship is reversed, with X^2 now being more powerful than G^2. Given that both Rao-Scott statistics are simple modifications of X^2 and G^2, one would expect these trends to be manifested in the relative powers of $FX^2_{c_o}$ and $FG^2_{c_o}$, and of $X^2_{S_o}$ and $G^2_{S_o}$. No clear winner emerges from the comparison, particularly since for the more important non-constant design effect case, the gain in power from using G^2 when $\beta<0$ is offset by a similar loss due to using G^2 when $\beta>0$. Of course, if one has prior knowledge of the true form of the alternative, then the appropriate statistic can be selected.

In the constant design effect case, and for both settings of β, there are no differences of practical interest between the powers of the Fay statistics X_J and G_J. For the case $a>0$, however, there are worthwhile differences, particularly when $\beta=-0.1$, and it is interesting to note that these are in the opposite direction to the differences exhibited by X^2 and G^2 forms of the Rao-Scott statistics. Thus, when $a>0$ and $\beta=-0.1$, the power of X_J is markedly greater than that of G_J, while for $\beta=+0.1$, G_J has the edge. There is no ready explanation for this anomalous behaviour of X_J and G_J.

Table 4

The Effect of Alternative Estimates of $\bar{\lambda}$ and a on the Power of the Rao-Scott Tests

$k=5$; $\bar{\lambda}=2$; $m=10$; $r=30$

a	β	$FX^2_{c_o}$	FX^2_c	$X^2_{S_o}$	X^2_S
0.0	-0.1	71.5	70.7	69.9	70.6
	+0.1	61.8	61.0	60.8	61.1
0.5	-0.1	69.1	63.7	68.8	59.8
	+0.1	50.4	54.5	45.1	51.3

Alternative Estimates of $\bar{\lambda}$ and a for the Rao-Scott Tests

As noted earlier, the expressions given by Rao and Scott (1981) for estimating $\bar{\lambda}$ and a can, under H_o, be based either on $\underset{\sim}{\Pi}_o$ or on its consistent estimator $\underset{\sim}{\hat{\Pi}}$. Powers for both forms of the Rao-Scott statistics are shown in Table 4, for thirty clusters ($r=30$), and for the previous two values of β and a. All results shown relate to the case of five categories, i.e. $k=5$.

When $a=0$, i.e. in the constant design effect case, the powers of FX^2_c and $FX^2_{c_o}$, and of X^2_S and $X^2_{S_o}$, are virtually identical for both settings of β. However, when $a=0.5$, and $\beta=-0.1$, the power of $FX^2_{c_o}$ exceeds that of FX^2_c, and the power of $X^2_{S_o}$ exceeds that of X^2_S. When $\beta=+0.1$, i.e. when $\Pi=(.3,.175,\ldots,.175)'$, this trend is reversed, the powers of the tests based on $\underset{\sim}{\hat{\Pi}}$ now exceeding those of the test versions based on $\underset{\sim}{\Pi}_o$. An explanation of some of these trends is given in Thomas and Rao (1985).

Behaviour of F_w for Small Values of r

Thomas and Rao (1984) introduced F_w, the modified form of the standard Wald statistic X^2_w because X^2_w itself failed to provide adequate control of Type I error. X^2_w exhibited significance levels around 20%, even for 50 clusters, under multinomial sampling. For small values of r, significance levels were even higher. For goodness of fit tests on 10 categories, values as high as 95% were recorded for the case $r=10$.

Use of the modified test, F_w, which takes into account the degrees of freedom of $\hat{\underset{\sim}{V}}$, reduced these significance levels to the acceptable range in most cases. However, it is interesting to consider what happens to test power in these situations. In reducing significance levels to the acceptable range, does F_w also destroy the power of the test? Table 5 reports some results that address this question.

Table 5

The Power of F_w for Small Numbers of Clusters

k=10; $\bar{\lambda}$=2.0; a=0.0; m=20; $\underset{\sim}{\Pi}$=(0.05,0.05,0.1125,..., 0.1125)'.

r	Power(F_w)	Power($FX^2_{C_o}$)	S. Level(X^2_w)
30	90.0	91.4	29.5
20	68.0	71.3	47.3
18	59.6	63.2	54.0
16	51.5	57.7	60.3
15	49.0	56.0	69.0
14	41.0	50.0	76.0
13	34.9	42.8	80.0
12	27.8	40.2	85.0
11	17.4	36.6	93.0
10	10.2	31.0	97.0

Powers of F_w are shown for the case k=10, $\bar{\lambda}$=2.0, a=0, along with those of $FX^2_{C_o}$, which is known to control Type I error well in the constant design effect case. The alternative has the form $\underset{\sim}{\Pi}$(10,2,-0.05), i.e. $\underset{\sim}{\Pi}$ has two elements equal to 0.05, the remaining eight being equal to 0.1125. From Table 5, it can be seen that for 15 or more clusters, the power of F_w is about 90% or more of that of $FX^2_{C_o}$. However, when r=14, the ratio of the powers drops to approximately 80%, and decreases rapidly from that level to a low of 26% when r=10, the smallest number of clusters for which F_w can be evaluated. At this value of r, the significance level of X^2_w is over 90%.

Thus it appears that when the number of clusters is very close to k, the power of F_w does collapse. However, this effect is serious only when the degrees of freedom of $\hat{\underset{\sim}{V}}$, given by r-k+1, are 4 or less. It is interesting to note at this point that the variance of an F-distribution on k-1 and r-k+1 degrees of freedom does not exist when r-k+1 $\leq$ 4. From Table 5, and other results not displayed here, it is found that for k=10 and r=15, the power of F_w is never less than 60% of that of $FX^2_{C_o}$ over a range of values of $\bar{\lambda}$ and a. In these cases, the significance levels of the unmodified test are around 60%, and are all reduced to the acceptable range by F_w. Thus the power robustness of F_w for small numbers of clusters is better than might be expected. Except when degrees of freedom are very small (r-k+1 $\leq$ 4), it reduces significance levels to an acceptable range without sacrificing all of its power.

Nevertheless, this cannot be taken as a recommendation for the use of F_w instead of its competitors $FX^2_{C_o}$, $X^2_{S_o}$ and X_J. It has been noted previously that F_w is particularly susceptible to the form of the probability vector $\underset{\sim}{\Pi}$, and it has been shown that its power is inferior to its competitors in the important case of non-constant design effects.

Summary and Conclusions

Monte Carlo methods were used to examine the power performance of four basic goodness-of-fit tests, and their variants, under cluster sampling. The basic tests studied were (i) $FX^2_{C_o}$, an F-based version of the Rao-Scott $\bar{\lambda}$ adjusted X^2 statistic, (ii) $X^2_{S_o}$, the original Rao-Scott Satterthwaite adjusted X^2, (iii) X_J, Fay's jackknifed X^2 statistic, and (iv) F_w, a modified Wald statistic referred to an F distribution.

Test powers were estimated for goodness-of-fit tests involving 3,5 and 10 categories, under a number of combinations of $\bar{\lambda}$ and a, the mean and coefficient of variation, respectively, of the eigenvalues of the generalized design effect matrix. Attention focussed on the equiprobable null hypothesis, the data being generated under two basic forms of the alternative $\underset{\sim}{\Pi}$. With the exception of the previous section's analysis, both alternatives consisted of single cell deviations from the equiprobable $\underset{\sim}{\Pi}_o$, of the form $\pi_1 = \pi_{01} + \beta$, with β positive or negative, the subscript one denoting the first cell probability.

The form of the (true) alternative was found to have a major effect on power. For $\beta < 0$, i.e. for a deviation of the first true cell probability towards zero, powers of all statistics were higher than their corresponding values for $\beta > 0$, and this effect was found to be particularly marked for F_w.

For the case of constant design effects, a=0, both X_J and F_w showed similar power for $\underset{\sim}{\Pi}$ vectors having $\beta < 0$ ($\pi_1 < \pi_{01}$), both being more powerful than the Rao-Scott statistics. However for alternatives $\beta > 0$ ($\pi_1 > \pi_{01}$), the power ranking of the four statistics was different, with F_w exhibiting less power than its three competitors, all of which performed similarly. For the more important case of non-constant design effects, a > 0, both of the Rao-Scott statistics showed more power than Fay's X_J and a great deal more power than F_w. Thus F_w appears to be highly susceptible not only to changes in $\underset{\sim}{\Pi}$ but also to increases in the variability of the design effects, as measured by the parameter a.

Variants of some of the basic procedures were also examined, and though some power differences were identified, there appeared to be no consistent advantage to using any of the variants in place of the basic forms $FX^2_{C_o}$, $X^2_{S_o}$ or X_J.

In summary, this study has shown that F_w has several unattractive features. It is very sensitive to the form of $\underset{\sim}{\Pi}$, and its power is markedly less than its competitors in the important case of non-constant design effects. X_J shows much better power characteristics, and in some cases is superior to the Rao-Scott tests. It is, however, not as powerful as the latter tests when a > 0, an important case in practice. When k=3 or k=5, there is little to choose between $FX^2_{C_o}$ and $X^2_{S_o}$ in terms

of power. When k=10, the power of $FX^2_{c_o}$ is greater than that of $X^2_{S_o}$, though when $a > 0$, this power advantage can be attributed to its somewhat inflated Type I error. Thus, when significance levels, under the influence of non-constant design effects, and power, are both taken into account, the Satterthwaite corrected Rao-Scott statistic comes out ahead, with Fay's X_J a close second. Of course, if full information on $\hat{V}$ is not available, then only $FX^2_{c_o}$ can be used.

References

Fay, R.E. (1985). A jackknifed chi-squared test for complex samples. J. Amer. Statist. Assoc., 80, 148-157.

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.

Koehler, K.J. and Larntz, K. (1980). An empirical investigation of goodness-of-fit statistics for sparse multinomials. J. Amer. Statist. Assoc., 75, 336-344.

Larntz, K., (1978). Small-sample comparisons of exact levels for chi-squared goodness-of-fit statistics. J. Amer. Statist. Assoc., 73, 253-263.

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, 221-230.

Read, T.R.C. (1984). Small-sample comparisons for the power divergence goodness-of-fit statistics. J. Amer. Statist. Assoc., 79, 929-935.

Thomas, D.R. and Rao, J.N.K. (1984). A Monte Carlo study of exact levels of goodness-of-fit statistics under cluster sampling. Proc. Sec. Survey Res. Methods, Amer. Statist. Assoc., 207-211.

Thomas, D.R. and Rao, J.N.K. (1985). On the power of goodness-of-fit tests under cluster sampling. Technical Report #66, Laboratory for Research in Statistics and Probability, Carleton University/ University of Ottawa, Ottawa, Canada.