A SIMULATED COMPARISON OF CHI-SQUARED TESTS FOR COMPARING VECTORS OF PROPORTIONS FOR SEVERAL CLUSTER SAMPLES

Jeffrey R. Wilson, Arizona State University

SUMMARY

Data are simulated for several subpopulations for a two-stage cluster sampling design using mixtures of Dirichlet-Multinomial distributions. The performance of chi-squared test statistics for comparing vectors of proportions for several cluster samples are compared using the simulated data. The type I error level performance and the power of the chi-squared tests are obtained for certain combinations of parameter values. These chi-squared tests include the Wald test statistic, Wald (1943), modified chi-squared test statistics as developed by Rao and Scott (1981), Scott and Rao (1981), and tests based on a probability model as developed by Koehler and Wilson (1986) and Wilson (1986).

Key Words: Two stage cluster; type I error; power; test of homogeneity; confidence intervals;

1. Introduction

Methods for the analysis of categorical data have been extensively developed for simple random sampling with replacement. These simple samples provide vectors of observed frequencies that are closely approximated with multinomial or binomial distributions, and as such the sampling procedure is sometimes referred to as multinomial sampling. For this kind of sampling, conditions under which the Pearson goodness-of-fit statistic and log-likelihood ratio goodness-of-fit statistic have the same asymptotic chi-square distributions have been established and widely published. The regularity conditions given by Birch (1964) are frequently cited. A simple theoretical development of a similar set of conditions was recently presented by Cox (1984).

During the last decade various procedures have been proposed for obtaining test statistics for hypotheses when the data are obtained from some complex survey. These include papers by Brier (1980), Rao and Scott (1979, 1981) Holt, Scott and Ewings (1980), Wilson (1984), and Koehler and Wilson (1986).

The usual formulas for the Pearson and log likelihood ratio statistic generally do not provide reliable chi-square tests even for very large samples. Koehler and Wilson (1986) considered tests for the equality of vectors of proportions when the data are obtained from several independent two-stage cluster samples. They showed that such test statistics can be classified into three methods. One method is to construct appropriate quadratic forms, which are often referred to as Wald statistics. A second method requires probability model to describe the variation in the true vectors of proportions across all clusters for each population. Examples of this approach are given by Brier (1980), Wilson (1984) and Koehler and Wilson (1986). A third approach requires only partial information about the covariance matrix of the observed vector of frequencies. Examples of this approach are given in Bedrick (1983), Scott and Rao (1981) and Rao and Scott (1979).

In this paper test procedures from these three different methods are investigated to obtain a simulated comparison of their performance in testing the equality of vectors of proportion for two stage cluster samples. The power of the selected test statistics are computed and comparisons are made of the attained level of significance. Confidence intervals are constructed for different procedures of estimating the clustering effect.

2. Two Stage Cluster Sample Model

Consider comparing vectors of proportions from J populations. The members of each population are classified into the same set of I mutually exclusive and exhaustive categories. For the j-th population, the true proportion of members in the various categories are given by the vector $\underset{\sim}{\pi}_j = (\pi_{1j}, \ldots, \pi_{2j}, \pi_{Ij})'$, where $1 = \sum_{i=1}^{I} \pi_{ij}$ for each population. Estimates of the true vectors of proportions are obtained from independent two-stage cluster samples from each population. Each population consists of a number of clusters. A sample of K_j clusters is randomly selected with replacement and with probability proportional to size (pps) from the j-th population. Furthermore, a random sample of n_{jk} secondary units is selected with replacement from the k-th cluster selected from the j-th population and each sampled unit is classified into one of the I mutually exclusive categories. Conditionally on the cluster selected, the vector of observed frequencies for the k-th cluster selected from the j-th population, $\underset{\sim}{X}_{jk} = (X_{1jk}, X_{2jk}, \ldots, X_{IJK})'$, has a multinomial distribution with parameters $\underset{\sim}{p}_{jk} = (p_{1jk}, p_{2jk}, \ldots, p_{Ijk})'$, where $\underset{\sim}{p}_{jk}$ is the true vector of proportions for the particular cluster selected.

A two dimensional table of frequency totals can be constructed in which the rows correspond to the I categories and the columns correspond to the J populations. The j-th column of this table consists of the vector $\underset{\sim}{X}_j = \sum_{k=1}^{K_j} \underset{\sim}{X}_{jk}$ of total frequencies for the j-th population. The total number of observations obtained from the j-th population is $N_j = \sum_{k=1}^{K_j} n_{jk}$. For the sampling scheme considered here, an unbiased estimator for $\underset{\sim}{\pi}_j$ is

$$\hat{\underset{\sim}{\pi}}_j = N_j^{-1} \underset{\sim}{X}_j.$$

3. Wald Statistics

Consider testing the null hypothesis H_o: $\underset{\sim}{\pi}_j = \underset{\sim}{\pi}_o$, j=1,2,...,J, for some unknown vector

π_0, against the general alternative. When H_0 is true an unbiased estimator of π_0 is given by $\hat{\pi} = \sum_{j=1}^{J} \alpha_j \hat{\pi}_j$, where the weights are any constants such that $1 = \Sigma_{j=1}^{J} \alpha_j$ and $\alpha_j \geq 0$, for $j = 1, 2, \ldots, J$. It is easily shown that the covariance matrix for $\hat{\pi}_j - \hat{\pi}$ is

$$\Sigma_{jj} = S_j - 2\alpha_j S_j + \Sigma_{t=1}^{J} \alpha_t^2 S_t, \tag{3.1}$$

where

$$S_j = M_j + N_j^{-2}(\Sigma_{k=1}^{K_j} n_{jk}^2 - N_j)\Sigma_\ell w_{j\ell}(P_{j\ell} - \pi_j)(P_{j\ell} - \pi_j)' \tag{3.2}$$

and

$$M_j = N_j^{-1}[\mathrm{diag}(\pi_j) - \pi_j \pi_j'], \tag{3.3}$$

where $\mathrm{diag}(\pi_j)$ denotes a diagonal matrix with the elements of π_j on the diagonal, and the last sum in (3.2) is across all clusters in the j-th population with weights $w_{j\ell}$ equal to the proportion of the population in the ℓ-th cluster. The matrix of covariances between $\hat{\pi}_i - \hat{\pi}$ and $\hat{\pi}_j - \hat{\pi}$ is

$$\Sigma_{ij} = -\alpha_i S_i - \alpha_j S_j + \Sigma_{t=1}^{J} \alpha_t^2 S_t. \tag{3.4}$$

The matrix Σ with diagonal blocks given by (3.1) and off-diagonal blocks given by (3.4) is the covariance matrix for the vector of random deviations

$$\hat{d} = (\hat{\pi}_1' - \hat{\pi}', \hat{\pi}_2' - \hat{\pi}', \ldots, \hat{\pi}_J' - \hat{\pi}')'. \tag{3.5}$$

The evaluation of the test statistic necessitates the estimation of the covariance matrix Σ. A consistent and nearly unbiased estimator of Σ is obtained by replacing S_j in (3.1) with

$$\hat{S}_j = a_j N_j^{-1}(K_j - 1)^{-1} \Sigma_{i=1}^{K_j} n_{ij}(\hat{P}_{jk} - \hat{\pi}_j)(\hat{P}_{jk} - \hat{\pi}_j)' + (1 - a_j)\hat{M}_j, \tag{3.6}$$

where

$$\hat{M}_j = N_j^{-1}[\mathrm{diag}(\hat{\pi}_j) - \hat{\pi}_j \hat{\pi}_j'],$$

$$a_j = (K_j - 1)(\Sigma_{k=1}^{K_j} n_{jk}^2 - N_j)/(N_j^2 - N_j(K_j - 1) - \Sigma_{k=1}^{K_j} n_{jk}^2),$$

and $\hat{P}_{jk} = n_{jk}^{-1} X_{jk}$ is the vector of observed proportions for the k-th cluster sampled from the j-th population. The estimate of the covariance matrix is denoted by $\hat{\Sigma}$. If $n_{jk} = n_j$ for all clusters sampled from the j-th population, $a_j = 1$ and (3.6) reduces to

$$\hat{S}_j = n_j(K_j - 1)^{-1} \Sigma_{i=1}^{K_j} (\hat{P}_{jk} - \hat{\pi}_j)(\hat{P}_{jk} - \hat{\pi}_j)', \tag{3.7}$$

which is an unbiased estimator for S_j. There are other acceptable estimators for Σ which provide essentially the same value of the test statistic for large samples but may result in slightly different values for smaller samples.

Clearly, Σ and $\hat{\Sigma}$ are singular matrices. Nonsingular covariance matrices can be obtained by deleting some elements from $\hat{d}$, but it is notationally more convenient to retain redundant differences in d and use a generalized inverse of Σ in the definition of test statistics. Consequently, a Wald statistic for testing the equality of the vectors of population proportions is

$$X_w^2 = \hat{d}'\hat{\Sigma}^{-}\hat{d}. \tag{3.8}$$

Following Moore (1977), this statistic has a limiting central chi-square distribution with degrees of freedom equal to the rank of Σ when the null hypothesis is correct.

The Wald statistic in (3.8) reduces to the Pearson chi-square statistic for testing independence,

$$X_p^2 = \Sigma_{j=1}^{J} N_j \Sigma_{i=1}^{I} \hat{\pi}_i^{-1}(\hat{\pi}_{ij} - \hat{\pi}_i)^2, \tag{3.9}$$

when $\hat{\Sigma}$ is the usual estimate of the covariance matrix of d for simple random sampling. This occurs when $S_j = M_j$ in (3.7) for $j=1,2,\ldots,J$, which must occur when $n_{jk} = 1$ for all sampled clusters. For large n_{jk} values, X_p^2 may be much larger than X_w^2 when there is substantial variation among the observed vectors of proportions for the clusters sampled from a particular population. Consequently, the type I error level of the test may be greatly inflated if the Pearson chi-squared test is used for a table of frequencies obtained from two-stage cluster sampling. This fact is demonstrated in Section 7.

The accuracy of the large sample chi-square approximation for the null distribution of X_w^2 is greatly influenced by the accuracy of $\hat{\Sigma}$ as an estimator for Σ. A substantial number of sampled clusters is required to accurately estimate large covariance matrices. When a large number of clusters cannot be sampled from each population, it may be advantageous to describe variation among clusters within populations with a more parsimonious model. The Dirichlet-multinomial model described in the next section uses only one parameter to account for among cluster variation within each population.

4. Probability Model

Brier (1980) used the Dirichlet-Multinomial model to develop chi-square tests for models fits to contingency tables obtained from a single two-stage cluster sample. Under the model the vectors of category proportions for the clusters in the j-th population have a Dirichlet-distribution with mean vector π_j. The properties of this model have been studied by Mosimann (1962) and Good (1965). Koehler and Wilson (1986) extended the model to obtain chi-square tests for comparing vectors of proportions obtained from several independent two-stage cluster samples. A test for assessing fit of the model was also presented in Wilson (1986).

In this simulated study we consider among others, three chi-square tests as developed by Koehler and Wilson (1986) under the Dirichlet

Multinomial model. The test statistics considered for testing H_o: $\pi_j = \pi_o$, j=1,2,..., J; are

$$X^2_{DM} = \sum_{j=1}^{J} N_j \hat{C}_j^{-1} \sum_{i=1}^{I} \hat{\pi}_i^{-1} (\hat{\pi}_{ij} - \hat{\pi}_i)^2 \qquad (4.1)$$

where $\hat{C}_j$ measures the clustering effect in the jth population. Koehler and Wilson (1986) gave three different methods of computing C_j, thereby producing three chi-square tests. The first method estimator (C_{jB}) uses properties of the Dirichlet-multinomial model and assumes that a sufficiently large number of clusters are sampled from the j-th population so that properties of a multivariate normal distribution can be used to accurately approximate the covariance matrix. Denote the resulting chi-square test by X^2_{DMB}.

A second estimator for C_j ($\hat{C}_{jV}$) assumes that the covariance matrix is diagonal, but uses the data to estimate the variances. Denote the resulting chi-square test by X^2_{DMV}.

The third estimator for C_j ($\hat{C}_{jW}$) uses a direct estimate of the covariance matrix. It does not assume that this covariance matrix is a diagonal matrix, nor does it use the idea of approximating normality through a transformation of the vector. Denote the resulting test statistic by X^2_{DMW}.

Each method for estimating C_j results in a lack-of-fit test for the Dirichlet-Multinomial model. Denote these lack-of-fit tests by X^2_{BJ}, X^2_{VJ}, and X^2_{WJ}, respectively. A detailed discussion of these lack-of-fit test are given by Wilson (1984). An approximate distribution and test for the statistic C_j have also been considered by Wilson (1986).

5. Partial Information on Covariance Matrix

Rao and Scott (1979) and Scott and Rao (1981) have developed a first order correction for which the Pearson chi-square statistic is divided by an estimated average design effect. We examine the performance of some of these adjusted tests. Other modified versions of the usual Pearson statistic have been considered by Holt, Scott and Ewings (1980), and Bedrick (1983) to name a few. In this study we consider three modified chi-square test statistics. These three test statistics, X^2_{RSB}, X^2_{RSV} and X^2_{RSW} are similar except that the estimated average design effects are different. For X^2_{RSB} the average design effect σ is based on the estimator C_{jB} through the formula

$$\hat{\sigma} = (J-1)^{-1} \sum_{j=1}^{J} \hat{C}_{jB}(1 - N_j N^{-1}).$$

Similarly, X^2_{RSV} and X^2_{RSW} are calculated based on $\hat{C}_{jV}$ and C_{jW} respectively.

6. Design of Simulation Study

In order to study the performance of the test statistics discussed in the last three sections a simulation study was performed, since it was not very convenient to obtain the exact distribution for these statistics. The data were generated from two stage cluster sampling schemes. Brier (1978) method of generating Dirichlet variates from Beta random variables was used to simulate the data. Thomas and Rao (1984) have used this distribution to study exact levels of chi-squared goodness-of-fit statistics. Sampling schemes examined here are obtained from selected combinations of the given parameters which are defined as i) I, the number of categories which was chosen to be 5 based on a simulation study conducted by Thomas and Rao (1984), ii) K, the Dirichlet constant, iii) n, the number of units drawn per cluster, iv) π, the model probability vector, v) r, the number of independent clusters, vi) α, the nominal significance level for the test statistics and vii) J, the number of subpopulations. The values considered for the parameters are K = 5; α = .05; r = 10, 25, 50; n = 10, 25, 55; I = 5; π_1 = (.2, .2, .2, .2, .2)' π_2 = (.50, .20, .15, .10, .05)' and π_3 = (.80, .05, .05, .05, .05); J = 2 and 3. Sampling from populations with the same Dirichlet prior ensured that the hypothesis was true and all population parameters are known exactly. These tests would check the type I error performance of the various test statistics under two stage cluster sampling. Though adequate control of significance levels is essential if a test statistic is to be useful, no comparison of competing statistic is complete without a comparison of their powers. Power values were obtained by comparing vectors of proportions for data drawn from populations with different model probability vector.

The performance of tests to check the model assumptions was also investigated. Confidence intervals for the estimators of the clustering effects were also obtained.

7. Results of Simulation

All results given represent the proportion of actual rejections of a true hypothesis of a 5 percent nominal level and 1200 independent trials.

7.1 Wald Statistics

Table 7.1 gives the actual significance levels (SL) for the Wald test statistic and the Pearson statistics (Special Wald) for simulated data based on a Dirichlet constant, k = 5. The number of clusters vary from 10 to 50. The cluster sizes vary from 10 to 55. The data are simulated from three different Dirichlet Multinomials with prior probability vectors A = (.50, .20, .15, .10, .05)', B = (.20, .20, .20, .20, .20)' and C = (.80, .05, .05, .05 .05)'. The performance of X^2_W deteriorates when the number of clusters decreases, even when the cluster sizes are large. Overall X^2_w seems to work best when the data are generated from a Dirichlet distribution with model probability vector π = (.20, .20, .20, .02, .20)'; and a large number of clusters. The significance levels for the Pearson statistic, X^2_P are greatly inflated, the levels range from 12 to 28 percent. These findings are in part similar to the results of Thomas and Rao (1984) for goodness-of-fit hypothesis. The significance levels for X^2_P are about twice the corresponding values for X^2_W. Table 7.1 also shows that X^2_W

Table 7.1

Actual Significance Levels for the Wald Test Statistic and Pearson Statistic

r	n	π_1	π_2	$SL(X_w^2)$	$SL(X_p^2)$
10	10	A	A	.10	.25
	25	A	A	.13	.28
	55	A	A	.10	.28
25	10	A	A	.07	.14
	25	A	A	.08	.16
	55	A	A	.08	.17
	25	B	B	.05	.14
	10	C	C	.08	.15
	25	C	C	.08	.17
50	10	A	A	.06	.12
	55	A	A	.07	.13
	55	A	A	.07	.13
	10	B	B	.06	.14
	55	B	B	.07	.15

k = 5 α = .05
A = (.50, .20, .15, .10, .05)'
B = (.20, .20, .20, .20. .20)'
C = (.80, .05, .05, .05, .05)'

is sensitive to the non uniformity of the model probability vector especially when the number of clusters are few.

Comparison of the powers values of X_w^2 and X_p^2 are given in Table 7.2 for the case of 10, 25 and 50 clusters. The test statistic, X_w^2 has increased power values as the number of clusters increases. In all cases X_w^2 was more powerful than X_p^2. This result is expected as X_w^2 makes use of all the information on the clusters, whereas X_p^2 uses the summarized data.

Table 7.2

Comparison of Powers of the Wald Statistic and the Pearson Statistic

r	n	π_1	π_2	Power(X_w^2)	Power(X_p^2)
10	10	A	B	.75	.71
	10	A	C	.88	.67
	25	A	B	.70	.69
	25	A	C	.85	.65
	55	A	B	.85	.67
25	10	A	C	.92	.85
	25	A	B	.90	.81
50	25	A	B	.93	.86

A = (.50, .20, .15, .10, .05)'
B = (.20, .20, .20, .20, .20)'
C = (.80, .05, .05, .05, .05';

7.2 Probability Models

Table 7.3 compares the actual significance levels of X_{DMB}^2, X_{DMW}^2, and X_{DMV}^2 for a range of values of cluster sizes and selected numbers of clusters. The Dirichlet constant parameter, k is equal to 5.

The test statistic, X_{DMB}^2 performs well. The significance level is quite stable. In all the cases examined X_{DMB}^2 exhibits a lower significance level than X_{DMW}^2 or X_{DMV}^2. The statistic X_{DMW}^2 is quite insensitive to the cluster size or the number of clusters. The significance levels for X_{DMV}^2 are greatly inflated at times. It improves in performance as the number of cluster increases, but, for small cluster sizes X_{DMV}^2 had significance levels as large as 22 percent.

Table 7.3

Comparison of Actual Significance Levels for X_{DMB}^2, X_{DMW}^2, X_{DMW}^2 (Dirichlet Multinomial Model)

r	n	π_1	π_2	$SL(X_{DMB}^2)$	$SL(X_{DMW}^2)$	$SL(X_{DMV}^2)$
10	10	A	A	.06	.17	.20
	25	A	A	.06	.15	.22
	55	A	A	.05	.15	.21
25	10	A	A	.05	.15	.08
	25	A	A	.06	.23	.11
	55	A	A	.06	.25	.10
	25	B	B	.05	.20	.08
	10	C	C	.07	.23	.11
	25	C	C	.07	.29	.12
50	10	A	A	.06	.11	.07
	55	A	A	.06	.12	.08
	10	B	B	.05	.12	.07
	55	B	B	.06	.14	.08

k = 5 α = .05
A = (.50, .20, .15, .10, .05)'
B = (.20, .20, .20, .20. .20)'
C = (.80, .05, .05, .05, .05)'

Estimates of powers for X_{DMB}^2, X_{DMW}^2, and X_{DMV}^2 are compared in Table 7.4 for selected combinations of cluster sizes and number of clusters. In all cases examined, X_{DMB}^2 was the most powerful. For large numbers of clusters X_{DMV}^2 was more powerful than X_{DMW}^2. Thus the significance levels and the power values suggest that X_{DMW}^2 should be used only for samples including a large number of clusters from each population. However, in cases of large number of clusters X_w^2 would provide a more reliable test and would be just as easy to compute. Since X_w^2 does not require a model for the distribution of the true vectors of probabilities among clusters, it is sometimes recommended when a large number of clusters are sampled from each population. The ease with which X_{DMB}^2 can be computed and the power and reliability of it is an important consideration. The chi-square approximation is more accurate for X_{DMB}^2 than X_w^2 both in the case of few sampled clusters and a large number of sampled clusters. Like X_w^2 both X_{DMV}^2 and X_{DMW}^2 are both sensitive to the non uniformity of the model probability vector.

Table 7.4

Comparison of Powers of the X_{DMB}^2, X_{DMW}^2, and X_{DMV}^2

r	n	π_1	π_2	Power(X_{DMB}^2)	Power(X_{DMW}^2)	Power(X_{DMV}^2)
10	10	A	B	.93	.83	.77
	25	A	B	.90	.80	.75
	55	A	B	.93	.82	.74
	10	A	C	.92	.84	.71
	25	A	C	.93	.84	.71

Table 7.4 (Continued)

Comparison of Powers of the X^2_{DMB}, X^2_{DMW}, and X^2_{DMV}

r	n	π_1	π_2	Power(X^2_{DMB})	Power(X^2_{DMW})	Power(X^2_{DMV})
25	25	A	B	.94	.74	.87
	10	A	C	.94	.79	.91
50	25	A	B	.93	.87	.92
	10	A	C	.93	.88	.92
	55	A	B	.93	.84	.92

A = (.50, .20, .15, .10, .05)'
B = (.20, .20, .20, .20, .20)'
C = (.80, .05, .05, .05, .05)'

7.3 Partial Information

Significance levels for X^2_{RSB}, X^2_{RSV} and X^2_{RSW} are given in Table 7.5. The significance levels for X^2_{RSB} are quite stable. As the number of clusters increases, the performance of X^2_{RSV} improves. The statistic X^2_{RSW} had significance levels as high as 28 percent. The values for X^2_{RSW} are greatly inflated. In all cases X^2_{RSB} had smaller significance levels.

Table 7.5

Comparisons of Actual Significance Levels for X^2_{RSB}, X^2_{RSV}, and X^2_{RSW}

r	n	π_1	π_2	SL(X^2_{RSB})	SL(X^2_{RSV})	SL(X^2_{RSW})
10	10	A	A	.05	.18	.13
	25	A	A	.05	.19	.13
	55	A	A	.04	.18	.13
25	10	A	A	.04	.08	.14
	25	A	A	.06	.10	.22
	55	A	A	.05	.10	.24
	25	B	B	.04	.08	.20
	10	C	C	.06	.09	.21
	25	C	C	.07	.10	.28
50	10	A	A	.05	.06	.11
	55	A	A	.06	.08	.12
	10	B	B	.05	.06	.19
	55	B	B	.05	.08	.14

k = 5 α = .05
A = (.50, .20, .15, .10, .05)'
B = (.20, .20, .20, .20. .20)'
C = (.80, .05, .05, .05, .05)'

A comparison of powers for the three statistics in Table 7.6 shows that X^2_{RSB} is the most powerful and very stable for comparing proportions. The powers for X^2_{RSV} and X^2_{RSW} fluctuate greatly. For a large number of clusters X^2_{RSV} is more powerful than X^2_{RSW}.

For the comparison of vectors of proportions from several populations, the chi-square test X^2_{DMB} provided by the estimators C_{jB}, j = 1, 2, ..., J; for the Dirichlet Multinomial model is quite similar to a modification provided by Rao and Scott (1979) and Scott and Rao (1981). Thus, when the sampling scheme involves complex samples from several different populations, a simple improvement to the method of correcting chi-square tests is to compute an average design effect for each population, and average them to obtain a single correction factor only if they are reasonably similar. Otherwise, the chi-square test statistic should be modified by using a different correction factor for each population. The latter type of correction is provided by X^2_{DMB} for comparing vectors of proportions for several populations.

Table 7.6

Comparison of Powers of the X^2_{RSB}, X^2_{RSV}, and X^2_{RSW}

r	n	π_1	π_2	Power(X^2_{RSB})	Power(X^2_{RSV})	Power(X^2_{RSW})
10	10	A	B	.94	.80	.85
	25	A	B	.92	.77	.83
	55	A	B	.93	.76	.82
	10	A	C	.93	.75	.86
	25	A	C	.94	.73	.85
25	25	A	B	.93	.89	.75
	10	A	C	.94	.91	.80
50	25	A	B	.94	.93	.87
	10	A	C	.93	.93	.88
	55	A	B	.93	.91	.85

A = (.50, .20, .15, .10, .05)'
B = (.20, .20, .20, .20, .20)'
C = (.80, .05, .05, .05, .05)' k = 5

7.4 Lack-of-fit Tests

Table 7.7 gives the actual significance levels for the lack-of-fit tests, X^2_{Bj}, X^2_{Vj}, and X^2_{Wj} for selected cases of cluster sizes, and selected numbers of clusters when the data are simulated from three Dirichlet Multinomials with different model probability vectors. Table 7.7 indicates that X^2_{Wj} is unacceptable, X^2_{Bj} is very stable and X^2_{Vj} improves in its performance as the number of clusters increases. In all cases X^2_{Bj} exhibits a lower significance level without being excessively conservative. The statistic X^2_{Bj} is slightly sensitive to the non uniformity of the model probability vector.

Table 7.7

Actual Significance Levels for the Lack-of-fit Tests X^2_{Bj}, X^2_{Vj}, X^2_{Wj} (Dirichlet Multinomial Model)

r	n	π_1	SL(X^2_{Bj})	SL(X^2_{Wj})	SL(X^2_{Vj})
10	10	A	.02	.44	.15
	25	A	.03	.59	.19
	55	A	.04	.58	.22
	55	C	.04	.58	.21
25	10	A	.05	.95	.14
	25	A	.05	.93	.07
	55	A	.05	.92	.09
	25	B	.05	.94	.07
	10	C	.04	.92	.15
	25	C	.05	.97	.16
50	10	A	.03	.65	.03
	55	A	.05	.69	.05
	10	B	.05	.58	.07
	55	B	.05	.68	.06

$k = 5 \quad \alpha = .05$
A = (.50, .20, .15, .10, .05)'
B = (.20, .20, .20, .20. .20)'
C = (.08, .05, .05, .05, .05)'

7.5 Interval Estimates for Clustering Effects

Interval estimates were obtained for each of the three procedures described for the clustering effect. The procedures based on a normality assumption (C_{jB}) works well when the data are simulated from Dirichlet Multinomials with non uniform model probability vector. In all cases the procedures yield confidence intervals which are much tighter when the number of clusters are larger. For few clusters the normality assumption procedure based on a diagonal covariance matrix (C_{jV}) gave lower bounds less than zero which we set at zero. Estimator C_{jB} produced the shortest confidence intervals. These intervals always covered the true value except when the data are simulated from a uniform model probability Dirichlet vector. The non normality assumption procedure, C_{jW} gave confidence intervals which are shorter than the procedures, C_{jV}. The results of this study show that the procedures proposed by Koehler and Wilson (1986) give consistently good coverage at the nominal 95% rate with a large number of clusters. However, the normality assumption procedure C_{jB} works well for a few number of clusters.

8. Discussion

Data simulated from Dirichlet multinomial distribution with model probability vectors π_1 = (.50, .20, .15, .10, .05)' π_2 = (.20, .20, .20, .20, .20)' and π_3 = (.80, .05, .05, .05, .05)', were used to examine the type 1 error performance and the power of chi-squared test statistic for the comparison of vectors of proportion from several subpopulations under cluster sampling. A study of test statistics from three different methods, Wald statistics, Probability Models and Partial Information suggests that the methods of interest are the Probability Models and the Partial Information. For Wald statistics the procedure requires a considerable number of clusters to achieve the desired significance level. The Pearson statistic which is a special case of the Wald statistic is definitely unacceptable in both large and small samples. The statistic X^2_{DMB} of the probability model method is the most powerful and stable. It achieves the desired significance level. However, it requires knowledge of the data from each sampled cluster. The test X^2_{RSB} is the most reliable and stable of the partial information method. It achieves the desired significance level. X^2_{RSB} requires knowledge of only the estimated cell design effects and the frequency counts per subpopulation.

Thomas and Rao (1984) found that tests formed analogous to X^2_{RSB} performed well and noted less than desired performances of Wald tests. However, their work examined goodness-of-fit tests under cluster sampling. In this study we concentrated on comparing vectors of proportions from several subpopulations. The interest is in tests of homogeneity.

Wilson (1984) and Koehler and Wilson (1986) presented lack-of-fit tests for checking when the data satisfies the assumptions of the Dirichlet Multinomial models. In this study we found that X^2_{Bi} is a very reliable and powerful statistic in providing a check for Dirichlet Multinomial data, regardless of the cluster sizes, number of clusters selected or model probability vector used in generating the data. Estimating the clustering effect in a given subpopulation depends on the model probability vector in the Dirichlet distribution. The estimator C_{jB} works well when the model probability vector is uniformly distributed over the categories. The results of this study suggest that the Dirichlet Multinomial model may be a useful technique for analyzing data from two stage cluster samples. The process of estimating the clustering effect which is reflected in the design effect seems to indicate that a test of fit for the model is necessary. Care must be taken to ensure that the data are not approximated by Dirichlet Multinomials based on uniform model probability vectors. Further work is still necessary to investigate the test of independence and the performance of these statistics in non Dirichlet Multinomial situations.

ACKNOWLEDGMENTS

This research was supported in part by a Research Incentive grant from the Bureau of Research at Arizona State University.

References

Bedrick, E.J. (1983), "Adjusted goodness-of-fit tests for survey data," Biometrika, 70, 591-595.

Birch, M.W. (1964), "A new proof of the Pearson-Fisher Theorem,"Annals of Mathematical Statistics, 35, 817-824.

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

Brier, S. S. (1980), "Analysis of contingency tables under cluster sampling," Biometrika, 67, 591-596.

Cox, C. (1984), "An elementary introduction to maximum likelihood estimation for multinomial models: Birch's Theorem and the delta method," American Statistician, 38, 283-287.

Good, I. J. (1965), Estimation of Probabilities, Cambridge, Mass: Massachusetts Institute of Technology Press.

Holt, D. Scott, A.J. and Ewings, P.O. (1980). Chi-square tests with survey data. J. Roy. Statist. Soc. A, 143, 302-320.

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