THE ANALYSIS OF FREQUENCY DATA BASED ON AN ALTERNATIVE FORM OF CONSTRUCTING A CONTINGENCY TABLE

Jeffrey R. Wilson, Arizona State University
William Warde, Oklahoma State University

SUMMARY

The effects of obtaining a test statistic for the test of independence when the data are obtained from a certain common complex sampling scheme is examined in this paper. The data are summarized in a two way contingency table where the cell entries are not necessarily an integer value. This alternative method of obtaining a two way table based on ratio estimators is compared to the traditional summing procedure in the construction of contingency tables. A Wald test statistic based on Wald (1943) for the test of independence is obtained for this particular sampling scheme and compared to these two forms of table construction techniques. This paper shows that an alternative approximation to the Wald test statistic for independence is to construct a Pearson type statistic based on the alternative table presented here rather than constructing a Pearson Statistic on the usual contingency table, as is sometimes done in complex sampling schemes.

1. Introduction

Test statistics are obtained for the tests of homogeneity and independence under a stratified cluster sampling scheme. In the past, several researchers have resorted to the use of the Pearson statistic in the presence of complex sampling procedures. Rao and Scott (1981, 1984), examined the behavior of the X^2 statistic under complex designs by examining the eigenvalues of the product of the inverse of the covariance matrix under simple random sampling and the covariance matrix for the actual sampling scheme. Holt, Scott, and Ewings (1980) showed that a correction factor for X^2 based on the design effects works well for the test of homogeneity. However, they demonstrated that for tests of independence an appropriate modifying factor is more difficult to complete. Cohen (1976), Brier (1980), Wilson and Koehler (1984), Wilson (1984) have considered models as a means of reducing the sample size required for variance estimation and producing useful test statistics.

In this paper consideration is given to an alternative form of the construction of the contingency table under the specified sampling scheme. The researcher is advised in the summarizing of the data. Test statistics are constructed based on the traditional form of table construction and based on the alternative form of construction. The covariance matrix for each of these cases is constructed based on the assumption that the cell proportions are multinomially distributed. A Wald test statistic is obtained based on the actual design and ignoring the multinomial assumption. A comparison is made with those statistics obtained under the multinomial assumption and the Wald test statistic. A numerical example based on data obtained from a Wild life study Rolley and Warde (1985) is given in section 6 to demonstrate some of the results.

2. Model

In wildlife studies, it is common to attach radio transmitters to a number of animals and release them. The animals are then located repeatedly by radio telemetry and categorized as being in one of several habitats. It is apparent that repeated locations on the same animal are not independent samples. Researchers attempt to study differences in habitat usage by animals of different ages and sexes.

We therefore consider a sampling scheme consisting of J subpopulations or strata (defined by age and sex of the animal). From each subpopulation, n_j animals are sampled from an unknown population of size N_j. Each animal selected represents a cluster of observations.

Let x_{ijk} denote the number of observations in the ith category (habitat) which came from the kth sampled cluster (animal) of the jth subpopulation; $i=2, \ldots I$; $j=1, 2, \ldots J$; $k = 1,2,\ldots, n_j$.
Let

$$\underset{\sim}{x}_{jk} = (x_{1jk}, x_{2jk}, \ldots, x_{Ijk})'$$

be the observed vector of frequencies for the kth sampled cluster in the jth subpopulation. Assume that $\underset{\sim}{x}_{jk}$ is distributed as a multinomial distribution with parameters x_{+jk} and

$$\underset{\sim}{P}_{jk} = (P_{1jk}, P_{2jk}, \ldots, P_{Ijk})'.$$

Define the total sample size on the kth cluster of the jth subpopulation as

$$x_{+jk} = \sum_{i=1}^{I} x_{ijk}, \tag{2.1}$$

and let the fixed total sample size, x_{+j} for the jth subpopulation be

$$x_{+j} = \sum_{k=1}^{n_j} x_{+jk}. \tag{2.2}$$

Note that x_{+jk} represents a random sample with replacement from the X_{+jk} (unknown) observations. Since $\underset{\sim}{x}_{jk}$ has a multinomial distribution then the density function is

$$f(\underset{\sim}{x}_{jk}; \underset{\sim}{P}_{jk}, x_{+jk}) = x_{+jk}!(\prod_{i=1}^{I} x_{ijk}!)^{-1} \prod_{i=1}^{I} P_{ijk}^{x_{ijk}} \tag{2.3}$$

Define

$$\hat{P}_{jk} = x_{+jk}^{-1} x_{jk} \quad (2.4)$$

as the observed vector of proportions for the kth cluster of the jth subpopulation, which is an unbiased estimator of the true proportion of times, P_{jk}, that the kth cluster of the jth subpopulation is seen over the I categories. Let

$$\pi_j = (\pi_{1j}, \pi_{2j}, \ldots, \pi_{Ij})'$$

be the true vector of proportions for the jth subpopulation. These data can be cross classified into a two way contingency table of dimension (IXJ) where rows represent the I categories and columns represent the J subpopulations. This sort of arrangement is very familiar with categorical data.

Define

$$\pi_j = \sum_{k=1}^{N_j} X_{+j}^{-1} X_{+jk} P_{jk}. \quad (2.5)$$

The vector π_j is a weighted linear combination of the true proportion vectors for the N_j clusters within the jth subpopulation, where X_{+jk} is the total number of observations in in the kth cluster in the jth subpopulation and

$$X_{+j} = \sum_{k=1}^{N_j} X_{+jk} \quad (2.6)$$

is the total number of observations in the jth subpopulation. Let

$$x_{ij} = \sum_{k=1}^{n_j} x_{ijk} \quad (2.7)$$

be the total sample size for the jth subpopulation in category i, then the expected value (denoted by E) of the vector $x_j = (x_{1j}, x_{2j}, \ldots, x_{Ij})'$ is

$$E(x_j) = E\{\sum_{k=1}^{n_j} E(x_{jk}|n_j\}$$

$$= E\{\sum_{k=1}^{n_j} x_{+jk}P_{jk}\}$$

$$E(x_j) = \sum_{k=1}^{N_j} n_j X_{+j}^{-1} X_{+jk} x_{+jk} P_{jk} \quad (2.8)$$

Define the I dimensional observed vector of proportions for the jth subpopulation as

$$\hat{\pi}_j = x_{+j}^{-1} x_j, \quad (2.9)$$

then under the usual contingency table assumption of fixed marginals,

$$E(\hat{\pi}_j) = \sum_{k=1}^{N_j} n_j X_{+j}^{-1} X_{+jk} x_{+j}^{-1} x_{+jk} P_{jk}$$

$$= \sum_{k=1}^{N_j} n_j \alpha_{jk} \hat{\alpha}_{jk} P_{jk} \quad (2.10)$$

where

$$\alpha_{jk} = X_{+j}^{-1} X_{+jk}, \quad (2.11)$$

is an unknown constant and

$$\hat{\alpha}_{jk} = x_{+j}^{-1} x_{+jk}. \quad (2.12)$$

is considered known.

So $\hat{\pi}_j$ overestimates the true proportions, since the sum of the weightsin (2.10) is greater than one. However, if all the clusters of the same subpopulation are of equal sample size then π_j is an unbiased estimator of π_j. Such an equality condition is rather difficult to satisfy in practice, especially in the study of Wild life. Theoretically, the vector π_j is a type of combined ratio estimator and is expected to be biased (Cochran 1977).

The covariance matrix for the vector $\hat{\pi}_j$ conditional on the sample size chosen is

$$\mathrm{Var}(\hat{\pi}_j) = x_{+j}^{-2} \{\mathrm{Var}(\sum_{k=1}^{n_j} x_{+jk}P_{jk}) + E(\sum_{k=1}^{n_j} x_{+jk}(\Delta_{P_{jk}} - P_{jk} P_{jk}'))\}$$

$$= x_{+j}^{-2} \{\sum_{k=1}^{N_j} [x_{+jk}^2 n_j \alpha_{jk}(1-\alpha_{jk}) P_{jk} P_{jk}' + x_{+jk} n_j \alpha_{jk}(\Delta_{P_{jk}} - P_{jk}P_{jk}')]\} \quad (2.13)$$

$$= n_j x_{+j}^{-1}\{\sum_{k=1}^{N_j} \hat{\alpha}_{jk}\alpha_{jk}[\Delta_{P_{jk}} + (x_{+jk} - x_{+jk}\alpha_{jk} - 1)P_{jk}P_{jk}']\}.$$

$$= B_j,$$

where $\Delta_{P_{jk}}$ is a diagonal matrix with elements P_{jk}.

A consistent estimator of $\mathrm{Var}(\hat{\pi}_j)$ is given by $v(\hat{\pi}_j)$ where

$$v(\hat{\pi}_j) = n_j x_{+j}^{-1} \{\sum_{k=1}^{n_j} \hat{\alpha}_{jk}[\Delta_{\hat{P}_{jk}} + (x_{+jk} - x_{+jk} \hat{\alpha}_{jk} - 1)\hat{P}_{jk} \hat{P}_{jk}']\}. \quad (2.14)$$

and defined as $\hat{B}_j$, where $\hat{P}_{jk}$ is an unbiased estimator of P_{jk}.

3. Contingency Table

Let (x_{ij}) denote the contingency table formed using the frequencies x_{ij} (i=1, 2, ...I, j=1, 2,... J). The estimator $\hat{\pi}_j$ which is a multiple of x_j is not an unbiased estimator of π_j (2.5) unless the x_{+jk}'s are equal for all k. The estimator, $\hat{\pi}_j$ overestimates π_j except when the sample sizes of the chosen clusters of the jth subpopulation are all equal.

Consider another two way table denoted by (y_{ij}) formed using the values y_{ij} i = 1, 2, ...,I; and j = 1, 2, ..., J; where

$$y_{ij} = x_{+j} n_j^{-1} \sum_{k=1}^{n_j} x_{ijk} x_{+jk}^{-1}. \tag{3.1}$$

$$= x_{+j} n_j^{-1} \sum_{k=1}^{n_j} \hat{P}_{ijk}.$$

Define the estimator

$$\tilde{\pi}_j = y_{+j}^{-1} y_j \tag{3.2}$$

where $y_j = (y_{1j}, y_{2j}, \ldots y_{Ij})'$,

and

$$y_{+j} = \sum_{i=1}^{I} y_{ij}$$

$$= x_{+j}. \tag{3.3}$$

The estimator $\tilde{\pi}_j$ is a self weighting estimator, a desirable property in sampling. The estimator $\tilde{\pi}_j$ is a type of separate ratio estimator and is expected to perform well when the relation between x_{ijk} and x_{+jk} is constant for a given i and j. The expected value of $\tilde{\pi}_j$ conditional on the sample size chosen is

$$E(\tilde{\pi}_j) = n_j^{-1} E\{ \sum_{k=1}^{n_j} P_{jk} \}$$

$$= \pi_j. \tag{3.4}$$

Hence $\tilde{\pi}_j$ is an unbiased estimator of π_j regardless of the differences among the among the clusters' sample sizes. The covariance matrix for the vector, $\tilde{\pi}_j$ conditional on the sample size chosen is

$$\mathrm{Var}(\tilde{\pi}_j) = n_j^{-2} \{\mathrm{Var} \sum_{k=1}^{n_j} P_{jk} +$$

$$E(\sum_{k=1}^{n_j} x_{+jk}^{-1}(\Delta_{P_{jk}} - P_{jk} P'_{jk}))\}$$

$$= n_j^{-1}\{ \sum_{k=1}^{N_j} \alpha_{jk}(1-\alpha_{jk}) P_{jk} P'_{jk} +$$

$$x_{+jk}^{-1}\alpha_{jk}(\Delta_{P_{jk}} - P_{jk} P'_{jk})\}$$

$$= n_j^{-1} \sum_{k=1}^{N_j} \alpha_{jk}x_{+jk}^{-1}\{\Delta_{P_{jk}} +$$

$$(x_{+jk} - x_{jk}\alpha_{jk} -1)P_{jk} P'_{jk}\}.$$

$$= C_j. \tag{3.5}$$

A consistent estimator of $\mathrm{Var}(\tilde{\pi}_j)$ is given by $v(\tilde{\pi}_j)$ where

$$v(\tilde{\pi}_j) = n_j^{-1} \sum_{k=1}^{n_j} x_{+jk}^{-1} \{\Delta_{\hat{P}_{jk}} + (x_{+jk} -$$

$$x_{+jk}\hat{\alpha}_{jk} - 1)\hat{P}_{jk} \hat{P}'_{jk}\} \tag{3.6}$$

and defined as C_j.

The difference between the variances of $\hat{\pi}_j$ and $\tilde{\pi}_j$ is

$$\mathrm{Var}(\hat{\pi}_j) - \mathrm{Var}(\tilde{\pi}_j) = \sum_{k=1}^{N_j} \alpha_{jk}(n_j x_{+j}^{-1} \hat{\alpha}_{jk} -$$

$$n_j^{-1} x_{+jk}^{-1})V_{jk}, \tag{3.7}$$

where

$$V_{jk} = \Delta_{P_{jk}} + (x_{+jk} - x_{+jk}\alpha_{jk} - 1) P_{jk} P'_{jk} \tag{3.8}$$

Hence, the variances are the same whenever $x_{+jk} = a_j$ for all k, but it is not necessary that $a_1 = a_2 = \ldots = a_J$.

Define the coefficient of $\alpha_{jk} V_{jk}$ in the difference of variances in (3.8) as

$$R_{jk} = n_j x_{+j}^{-1} \hat{\alpha}_{jk} - n_j^{-1} x_{+jk}^{-1}$$

$$= n_j x_{+jk} x_{+j}^{-2} - n_j^{-1} x_{+jk}^{-1}. \tag{3.9}$$

Since

$$n_j^2 x_{+jk}^2 - x_{+j}^2 = (n_j x_{+jk} - x_{+j}) (n_j x_{+jk} + x_{+j}),$$

the sign of R_{jk} is unknown. R_{jk} may be negative, positive or zero for any kth cluster of the jth subpopulation. Thus a clear comparison between the variances is not possible. However, Cochran (1977) shows that unless those clusters are really alike the use of $\tilde{\pi}_j$ is likely to be more precise if the sample size in each cluster is large enough to allow estimation of the variance.

Using the table (y_{ij}) with its non integer cell values results in estimators that are unbiased but not necessarily having smaller or larger variances than estimators obtained using table (x_{ij}). The column totals of table (y_{ij}) are the same as the column totals of table (x_{ij}). The estimator based on table (y_{ij}) requires a knowledge of the separate total x_{+jk}, whereas the estimator based on table (x_{ij}) do not require a knowledge of those totals.

4. Test of Homogeneity

Consider testing the hypothesis

$$H_o: \pi_j = \pi_o \qquad j=1, 2, \ldots, J; \tag{4.1}$$

where π_o is a known vector and π_j is the true vector of proportions for the jth subpopulation. Then, a Wald type test can be formed with the biased estimator $\hat{\pi}_j - \pi_o$ and the covariance B_j, where $\hat{B}_j$ is a consistent estimator of B_j (2.13). Such a test statistic is given by

$$X^2_{1H} = \sum_{j=1}^{J} (\hat{\pi}_j - \pi_o)' \hat{B}_j^{-1} (\hat{\pi}_j - \pi_o), \tag{4.2}$$

and is asymptotically distributed as a chi-square random variable with J(I-1) degrees of freedom under H_0, (Stroud 1971). Similarly, another statistic can be computed using the unbiased estimator $\tilde{\pi}_j - \pi_o$ and the consistent estimator $\hat{C}_j$. Such a test statistic for testing (4.1) is given by

$$X^2_{2H} = \sum_{j=1}^{J} (\tilde{\pi}_j - \pi_o)' \hat{C}_j^{-1} (\tilde{\pi}_j - \pi_o), \tag{4.3}$$

where $\hat{C}_j$ is a consistent estimator of C_j (3.6). X^2_{2H} is also asymptotically distributed as a chi-square random variable with J(I-1) degrees of freedom, (Stroud 1971).

Consider the contingency table formed based on (x_{ij}) and assuming that the frequencies obtained for the jth subpopulation follows a multinomial distribution then a test statistic for testing the hypothesis in (4.1) is

$$X^2_{1Ht} = \sum_{j=1}^{J} x_{+j} \sum_{i=1}^{I} (\hat{\pi}_{ij} - \pi_{io})^2 \pi_{io}^{-1}. \tag{4.4}$$

X^2_{1Ht} is the usual Pearson Statistic for the test of homogeneity, and is used as an approximation to X^2_{2H}. Similarly for table (y_{ij}) with the same multinomial assumption, and the use of the estimated vector of proportions $\tilde{\pi}_j$, we obtain the test statistic

$$X^2_{2Ht} = \sum_{j=1}^{J} x_{+j} \sum_{i=1}^{I} (\tilde{\pi}_{ij} - \pi_{io})^2 \pi_{io}^{-1}, \tag{4.5}$$

as an approximation to X^2_{2H}. It was shown by Rao and Scott (1981) that X^2_{1Ht} is a conservative test, for testing the hypothesis in (4.1). In practice, the data are usually available in the form of table (x_{ij}), so H^2_{1Ht} is easily calculated. The statistics X^2_{1Ht} and X^2_{2Ht} are obtained from the data in the summarized tables (x_{ij}) and (y_{ij}) respectively. They do not require information on each cluster. However, the statistics X^2_{1H} and X^2_{2H} cannot be computed from the summarized data given in tables (x_{ij}) and (y_{ij}). These statistics require information on each cluster.

5. Test of Independence

Consider testing the hypothesis

$$H_o: \pi_j = \pi_o \qquad j = 1, 2, \ldots, J; \tag{5.1}$$

where π_o is an <u>unknown</u> vector and π_j is the true vector of proportions for the jth subpopulation. The unknown vector π_o can be estimated by a weighted linear combination of the J estimated vectors π_j, j =1, 2, ..., J; Thus

$$\hat{\pi}_o = \sum_{j=1}^{J} \alpha_j \hat{\pi}_j \tag{5.2}$$

for some known weights, α_j, j=1, 2 ..., J; and an estimator $\hat{\pi}_j$ given in (2.9). Similarly, one can define

$$\tilde{\pi}_o = \sum_{j=1}^{J} \alpha_j \tilde{\pi}_j \tag{5.3}$$

where $\tilde{\pi}_j$ is an unbiased estimator of π_j as given in (3.2).

The estimator $\tilde{\pi}_j - \tilde{\pi}_o$ is an unbiased estimator of $\pi_j - \pi_o$ for fixed α_j's. Let T_{jj} denote the diagonal elements of $\text{var}(\tilde{\pi}_j - \tilde{\pi}_o)$ and $T_{jj'}$ denote the off diagonal elements, $j \neq j'$, $j' = 1, 2, \ldots, J$; Then as shown in Wilson and Koehler (1984) the matrix

$$T_{jj} = C_j - 2\alpha_j C_j + \sum_{\ell=1}^{J} \alpha_\ell^2 C_\ell \tag{5.4a}$$

and

$$T_{jj'} = -\alpha_j C_j - \alpha_{j'} C_{j'} + \sum_{\ell=1}^{J} \alpha_\ell^2 C_\ell. \tag{5.4b}$$

Let $\hat{T}_{jj}$ and $\hat{T}_{jj'}$ be consistent estimators of T_{jj} and $T_{jj'}$, respectively. $\hat{T}_{jj}$ and $\hat{T}_{jj'}$ are obtained by replacing C_j with $\hat{C}_j$. A test statistic for the hypothesis in (5.1) where π_o is an unknown vector, is

$$X^2_{2I} = (\tilde{\pi}^{(J)} - \tilde{\pi}_o^{(J)})' \hat{M}^c_{H_o} (\tilde{\pi}^{(J)} - \tilde{\pi}_o^{(J)}) \tag{5.5}$$

where $\hat{M}_{H_o}$ is a consistent estimator of

$$M_{H_o} = \text{var}(\tilde{\pi}^{(J)} - \tilde{\pi}_o^{(J)}) \tag{5.6}$$

under H_o. $\hat{M}^c_{H_o}$ is the Moore Penrose inverse of $\hat{M}_{H_o}$, and the vector of vectors,

$$(\tilde{\pi}^{(J)} - \tilde{\pi}_o^{(J)}) = (\tilde{\pi}_1' - \tilde{\pi}_o', \tilde{\pi}_2' - \tilde{\pi}_o', \ldots, \tilde{\pi}_J' - \tilde{\pi}_o')'. \tag{5.7}$$

The matrix $\hat{M}_{H_o}$ has diagonal elements $\hat{T}_{jj}$ and off diagonal elements $\hat{T}_{jj'}$. Similarly, a test statistic can be constructed using the biased estimator $\hat{\pi}^{(J)} - \hat{\pi}_o^{(J)}$ and a consistent estimator of the covariance matrix

$$V_{H_o} = \text{var}(\hat{\pi}^{(J)} - \hat{\pi}_o^{(J)}). \tag{5.8}$$

A consistent estimator, $\hat{V}_{H_o}$ is similar to $\hat{M}_{H_o}$ except that $\hat{C}_j$ is replaced by $\hat{B}_j$ (2.13) in (5.3) and (5.4). Thus,

$$X^2_{1I} = (\hat{\pi}^{(J)} - \hat{\pi}_o^{(J)})' \hat{V}^c_{H_o} (\hat{\pi}^{(J)} - \hat{\pi}_o^{(J)}) \quad (5.9)$$

is a test statistic for testing H_o in (5.1). The statistics X^2_{1I} and X^2_{2I} are distributed asymptotically as a chi-square random variable with (I-1)(J-1) degrees of freedom, (Stroud 1971).

Consider using the summarized data in tables (x_{ij}) and (y_{ij}) based on the multinomial assumption. Then the test statistic based on table (x_{ij}) is

$$X^2_{1It} = \sum_{j=1}^{J} x_{+j} \sum_{i=1}^{I} (\hat{\pi}_{ij} - \hat{\pi}_{io})^2 \hat{\pi}_{io}^{-1} \quad (5.10)$$

$$\geq \sum_{i=1}^{(I-1)(J-1)} Z_i^2 \quad \text{by Rao \& Scott (1981)}$$

The Z_i's are standard normal variates. The statistic based on table (y_{ij}) is

$$X^2_{2It} = \sum_{j=1}^{J} x_{+j} \sum_{i=1}^{I} (\tilde{\pi}_{ij} - \tilde{\pi}_{io})^2 \tilde{\pi}_{io}^{-1} . \quad (5.11)$$

The statistic X^2_{1It} is the usual Pearson Statistic for the test of independence. It is normally used by researchers as an approximate statistic when the covariance matrix cannot be or is too complicated to estimate to construct of the Wald test statistic. The statistic X^2_{2It} is similar to X^2_{1It} in structure and is an approximation to the statistic X^2_{2I}.

In section 6 in the analysis of the Wild life study data the statistics X^2_{2I}, X^2_{2It} and X^2_{1It} are related by the expression

$$E\{X^2_{2I}\} \leq E\{X^2_{2It}\} \leq E\{X^2_{1It}\}. \quad (5.12)$$

Thus, having the table constructed with (y_{ij}) as the cell values and using the multinomial assumption results in a less conservative test and a better approximation to the Wald test than the usual Pearson statistic, which is obtained from the use of table (x_{ij}). Hence in the case where the sampling scheme is as described in section 2 and the estimation of the covariance matrix needed in computing the Wald test, is too complicated, a reasonable approximation is obtained by constructing the alternative contingency table (y_{ij}), and assuming multinomial sampling. These results suggest that one can obtain better results in terms of approximations in making the adjustments to the construction of the table and then using the multinomial assumption. This requires that the researcher be forewarned about the method of summarization.

6. Numerical Example

Data from the study of the diel patterns of habitat use by male and female bobcats in southeastern Oklahoma, Rolley and Warde (1985) were analyzed using the test statistics X^2_{2I}, X^2_{1It} and X^2_{2It} in (5.5), (5.10) and (5.11) respectively. The data are reproduced in Tables 6.1a and 6.1b. There are J=2 subpopulations, male bobcats and female bobcats. For the male subpopulations, there are $n_1 = 5$ clusters with vector x_{+1} = (352, 125, 74, 23, 95)'. For the female subpopulation there are $n_2 = 9$ clusters with vector, x_{+2} = (195, 19, 90, 72, 26, 74, 60, 95, 52)'. There are I=5 categories of interest; pine, deciduous, mixed pine, grassfields, and brush. These categories are assumed to be nonoverlapping and well defined.

The contingency tables (x_{ij}) and (y_{ij}) are given in Tables 6.2a and 6.2b respectively. Table 6.2a is the traditional way of constructing a contingency table while Table 6.2b is the alternative technique proposed in this paper and based on a type of separate ratio estimator. Our hypothesis of interest is H_0: $\pi_j = \pi_0$ (j=1, 2) for some unknown π_0. The idea here is to investigate whether or not the male and female bobcats have the same habitat preferences.

TABLE 6.1a

Diel Patterns of Habitat Use by Five Female Bobcats in Southeastern Oklahoma

	Bobcats				
HABITATS	1	2	3	4	5
Pine	227	80	50	9	39
Deciduous	53	10	3	4	0
Mixed Pine	53	30	20	9	11
Grass Fields	8	5	1	0	31
Brush	11	0	0	1	14
Total	352	125	74	23	95

TABLE 6.1b

Diel Patterns of Habitat Use by Nine Male Bobcats in Southeastern Oklahoma

HABITATS	1	2	3	4	5	6	7	8	9
Pine	145	11	49	38	6	33	46	39	19
Deciduous	3	1	11	11	13	3	0	5	8
Mixed Pine	26	2	21	15	6	18	9	14	13
Grass Fields	11	4	4	2	1	1	2	30	1
Brush	10	1	5	6	0	19	3	7	11
Total	195	19	90	72	26	74	60	95	52

Under Table 6.2a the estimated vectors are for female bobcats $\hat{\pi}_1$ = (.605, .105, .184, .067, .039)' and for male bobcats $\hat{\pi}_2$ = (.565, .081, .181, .082, .091)'. The usual Pearson statistic X^2_{1It} given in (5.10), is 18.289. Under Table 6.2b the estimated vector for female bobcats, $\tilde{\pi}_1$ = (.553, .089, .234, .081, .044)' and for male bobcats, $\tilde{\pi}_2$ = (.513, .121, .189, .084, .093)'. The alternative statistic X^2_{2It} given in (5.11) is 9.171. From Tables 6.1a and 6.1b the statistics X^2_{1I} in (5.9) and X^2_{2I} in (5.5) were calculated. These are Wald statistics. The diagonal elements of the covariance matrix used in calculating X^2_{1I} are (.01283, .00032, .00177, .00706, .00209, .00863, .00096, .00271, .00233, .00224)'. The diagonal elements of the covariance matrix used in calculating X^2_{2I} are (.02574, .00069, .00363, .01405, .00425, .01786, .00203, .00573, .00464, .00470)'. The statistic X^2_{2I} based on the separate type estimator has the value 6.664 and the statistic X^2_{1I} based on the combined type estimator has the value 8.830. Statistics X^2_{1It} and X^2_{2It} are the approximations to X^2_{2I}. While X^2_{1It} is an unsuitable approximation the statistic X^2_{2It} is a reasonable estimator.

When these statistics are considered to be distributed as chi-square random variables with 4 degrees of freedom, we are led to rejecting the null hypothesis at the 5% significant level, if we use X^2_{1It}, the usual Pearson Statistic. All other statistics considered in this example led to supporting the claim that the bobcats (males and females) have about the same habitat preference in Southeastern Oklahoma.

TABLE 6.2a

A Habitat by Sex Two Way Contingency Table for Bobcats in Southeastern Oklahoma

	Females	Males
Pine	405	386
Deciduous	70	55
Mixed Pine	123	124
Grassfields	45	56
Brush	26	62

TABLE 6.2b

A Habitat by Sex Two Way Alternative Contingency Table for Bobcats in Southeastern Oklahoma

	Females	Males
Pine	369.689	350.227
Deciduous	59.541	82.719
Mixed Pine	156.278	129.011
Grassfields	53.921	57.524
Brush	29.570	63.443

7. DISCUSSION

The presence of clustering in the collection of sample data can have a severe effect on certain test statistics obtained from the frequency data in a usual contingency table. Such computed statistics are usually too large in numerical value. A better approximation is to construct the table based on a separate type estimator and then to use the usual techniques of constructing Pearson statistics. This technique has its greatest gain when the clusters differ greatly.

Rao and Scott (1981, 1984), Bedrick (1983), Wilson and Koehler (1984), Brier (1980) and Holt, Scott and Ewings (1980) have considered model that leads to a correction of the usual Pearson Statistics. Their works rely on summarized data through the usual construction of a contingency table. However, in this paper no correction is considered for the usual Pearson Statistic. Here the changes are suggested prior to the summarized data. There is no need for matrix inversion or the computation of several covariances. Eigen values are not needed. The computer programs necessary are readily available. They are the same as when multinomial sampling is conducted and a Pearson Statistic computed.

REFERENCES

Bedrick, E.J. (1983). Adjusted Chi-squares tests for cross-classified tables of survey data. Biometrika 70: 591-595.

Brier, S.S. (1980). Analysis of Contingency tables under cluster Sampling. Biometrika 67: 591-596.

Cohen, J.E. (1976). The distribution of the chi-square statistic under clustered sampling from contingency tables. J. Amer. Statist. Assoc. 71: 665-670.

Cochran, W.G. (1977). Sampling Techniques. John Wiley & Sons, 3rd Edition, New York.

Holt, D., Scott, A.J., and Ewings, P.D. (1980). Chi-squared tests withsurvey data. J.R. Statist. Soc., Series A. 143 (3): 303-320.

Rao, J.N.K, and Scott, A.J. (1981). The Analysis of Categorical Data from complex surveys. J. Amer. Statist. Assoc. 76: 221-30.

Rao, J.N.K. and Scott, A.J. (1984). On Chi-squared test for Multiway Contingency Tables with all proportions estimated from survey data. The Annals of Statistics 1984, Vol. 2, No. 1, 46-60.

Rolley, R.E. and Warde, W.D. (1985). Bobcat habitat use and activity patterns in Southeastern Oklahoma. Journal of Wildlife Management (in press).

Stroud, T.W. F. (1971). On obtaining large sample tests from asymptotic normal estimations. Annals of Mathematic Statistics Association, 72, 881-885.

Wald, A. (1943). Tests of Statistical hypotheses concerning several parameters when the number of observations is large. Trans. Amer. Math. Soc. 54: 426-482.

Wilson, J.R. (1984). Statistical Methods for frequency data from complex sampling schemes. Ph.D. Dissertation, Iowa State University, Ames, Iowa 50011.

Wilson, J.R. and Koehler, K.J. (1984). Testing equality of vectors of proportions for several cluster samples. Proceedings of Joint Statistical Association Meetings 1984. Section on Survey Research Methods, 201-206.