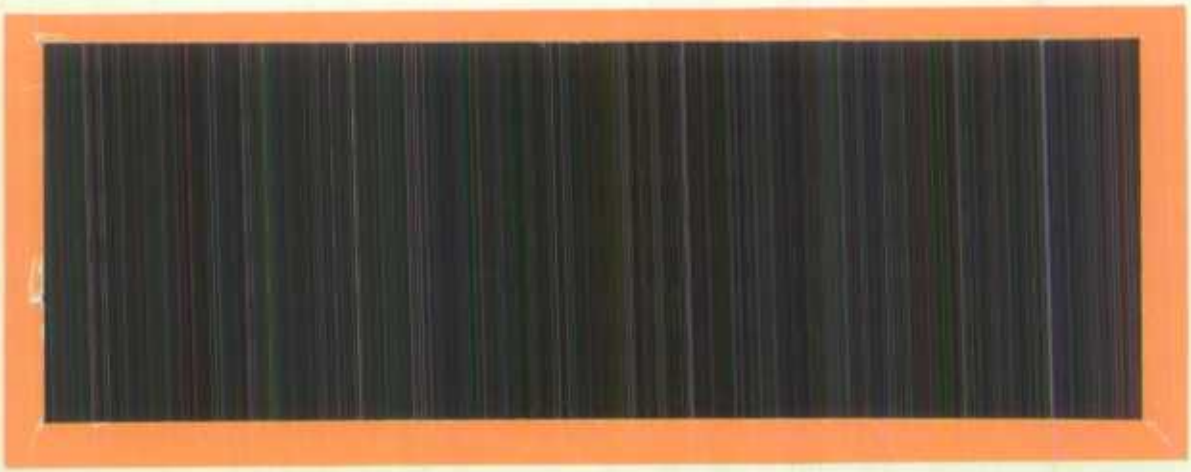
Statistics Canada | Statistique Canada



Methodology Branch

Census & Household Survey
Methods Division

Direction de la méthodologie

Division des méthodes de recensement
et d'enquêtes ménages

anadä

ANALYSIS OF CATEGORICAL DATA FROM SURVEYS WITH COMPLEX DESIGNS: SOME CANADIAN EXPERIENCES

D.A. Binder, M. Gratton, M.A. Hidiroglou,
S. Kumar and J.N.K. Rao

Number: METH 85-083E

Analysis of Categorical Data from Surveys with Complex Designs: Some Canadian Experiences[1]

D.A. Binder, M. Gratton, M.A. Hidiroglou,
S. Kumar and J.N.K. Rao[2]

ABSTRACT

Goodness of fit tests, tests for independence in a two-way contingency table, log-linear models and logistic regression models are investigated in the context of samples which are obtained from complex survey designs. Suggested approximations to the null distributions are reviewed and some examples from the Canada Health Survey and Canadian Labour Force Survey are given. Software implementation for using these methods is briefly discussed.

KEYWORDS: χ^2 statistic; Wald Statistics; Goodness of fit; Independence in two-way tables; Log-linear and logistic regression model.

1. INTRODUCTION

A sketch of the historical development of modern categorical data analysis has been given in the excellent review paper by Imrey, Koch and Stokes (1981). These techniques, applied in the context of random samples derived as independent selections from a common distribution function, are not directly applicable to survey samples collected using complex survey designs.

Koch *et al* (1975), Shuster and Downing (1976), developed asymptotically valid methods, based on the Wald statistic that take the survey design into account, but requiring access to the micro-data file or at least the full estimated covariance matrix of cell estimates. Cohen (1976) and Altham (1976) proposed a simple model for clustering and showed that the generalized Wald statistic for goodness of fit is a multiple of χ^2, when the model holds. Brier (1978) considered a similar model, but studied general hypotheses on cell probabilities, and proved that a multiple of the corresponding Pearson statistic is asymptotically distributed as a χ^2 random variable, when the model holds. Fellegi (1980) deflated the χ^2 using a correction factor based on the mean of the estimated design effects. Fay (1985) developed jackknife χ^2 and G^2 statistics, also taking the design into account, but requiring the cell estimated at the primary sampling unit level. Rao and Scott (1981) developed a correction to χ^2 (or G^2) based on the Satterthwaite to approximation to the asymptotic distribution of χ^2, requiring the full estimated covariance matrix.

In this paper, we discuss the problems of fitting models and testing hypotheses with categorical data resulting from complex designs. For data collected using complex designs, some adjustments to the classical methods described by Imrey, Koch and Stokes (1981) are necessary in order to make valid inferences. If the published tables are provided along with the cell and marginal design

[1] This paper is a revised and expanded version of that presented at the Seminar on Recent Developments in the Analysis of Large Scale Data Sets sponsored by Statistical Office of European Communities, November 16-18, 1983, Luxemburg.

[2] D.A. Binder, Institutional & Agriculture Survey Methods Division, M. Gratton, EDP Planning and Support Division, M.A Hidiroglou, Business Survey Methods Division, S. Kumar, Census & Household Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6, and J.N.K. Rao, Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario, CANADA.

effects, some of the approximations to the null distributions of our test statistics can be obtained, without having access to the complete micro-data file. On the other hand, for applications where the complete micro-data file is available, alternative approaches will be described.

For illustrative purposes, Section 2 begins with the standard goodness of fit problem. This discussion is then extended in Section 3 to tests of independence in a two-way contingency table. This leads to a general discussion of log-linear models in Section 4. Logistic regression models are described in Section 5. In Section 6 we summarize the existing situation with respect to software development for these methods at Statistics Canada. In Section 7, we discuss the appropriateness of these methods. Numerical examples are taken primarily from the Canada Health Survey. An application from the Canadian Labour Survey is given in Section 5.

2. GOODNESS OF FIT

2.1 Multinomial Sampling

Suppose we select n independent and identically distributed observations $Y_1, \ldots, Y_n$ from a discrete distribution with k categories, where $\Pr(Y = i) = \pi_i$; $\Sigma_{i=1}^{k} \pi_i = 1$. We observe the random vector $\underset{\sim}{n} = (n_1, \ldots, n_{k-1})^T$, which has a multinomial distribution. Our estimate of $\underset{\sim}{\pi} = (\pi_1, \ldots, \pi_{k-1})^T$ is given by $\underset{\sim}{p} = \underset{\sim}{n}/n$. This estimate is unbiased and has covariance matrix given by $V\{\underset{\sim}{p}\} = (\underset{\sim}{D}_{\pi} - \underset{\sim}{\pi}\underset{\sim}{\pi}^T)/n = \underset{\sim}{P}/n$, where $\underset{\sim}{D}_{\pi} = \text{diag}\{\pi_1, \ldots, \pi_{k-1}\}$. Note that $\underset{\sim}{P}^{-1} = \underset{\sim}{D}_{\pi}^{-1} + (\underset{\sim}{1}\underset{\sim}{1}^T/\pi_k)$. Asymptotically, $n^{1/2}(\underset{\sim}{p} - \underset{\sim}{\pi}) \rightarrow N(\underset{\sim}{0}, \underset{\sim}{P})$. For a given $\underset{\sim}{\pi}_o$, the goodness of fit problem is to test the hypothesis.

$$H_o: \underset{\sim}{\pi} = \underset{\sim}{\pi}_o,$$

against the alternative

$$H_1: \underset{\sim}{\pi} \neq \underset{\sim}{\pi}_o. \tag{2.1}$$

Letting $\underset{\sim}{P}_o$ represent $\underset{\sim}{P}$ evaluated at $\underset{\sim}{\pi}_o$, the Wald statistic for this test is

$$W_1 = n(\underset{\sim}{p} - \underset{\sim}{\pi}_o)^T \underset{\sim}{P}_o^{-1} (\underset{\sim}{p} - \underset{\sim}{\pi}_o)$$

$$= n\sum_{i=1}^{k} \{(p_i - \pi_{io})^2/\pi_{io}\},$$

which is the familiar Pearson chisquare test. Under H_o this is asymptotically χ^2_{k-1}. The likelihood ratio test for this problem is given by

$$LR_1 = 2n \sum_{i=1}^{k} p_i \log(p_i/\pi_{io}).$$

Since $2p_i \log(p_i/\pi_{io})$ is asymptotically equivalent to $2(p_i - \pi_{io}) + (p_i - \pi_{io})^2/\pi_{io}$ under H_o, we see that the likelihood ratio test is asymptotically equivalent to the Pearson chisquare statistic under H_o.

Another possible test for this hypothesis is derived by defining the vector of logs, $\underset{\sim}{\mu}_o = \log \underset{\sim}{\pi}_o$ and $\hat{\underset{\sim}{\mu}} = \log \underset{\sim}{p}$. Now under the null hypotheses $\hat{\underset{\sim}{\mu}} - \underset{\sim}{\mu}_o$ is asymptotically equivalent to $\underset{\sim}{D}_{\pi_o}^{-1}(\underset{\sim}{p} - \underset{\sim}{\pi}_o)$. Therefore, $n^{1/2}(\hat{\underset{\sim}{\mu}} - \underset{\sim}{\mu}_o) \rightarrow N(\underset{\sim}{0}, \underset{\sim}{D}_{\pi_o}^{-1} - \underset{\sim}{1}\underset{\sim}{1}^T)$ under H_o and the Wald statistic is

$$W_2 = (\hat{\underset{\sim}{\mu}} - \underset{\sim}{\mu}_o)^T \left[\underset{\sim}{D}_{\pi_o} + (\underset{\sim}{\pi}_o\underset{\sim}{\pi}_o^T/\pi_{ko})\right](\hat{\underset{\sim}{\mu}} - \underset{\sim}{\mu}_o)$$

$$\doteq \sum_{i=1}^{k} \pi_{io}(\hat{\mu}_i - \mu_{io})^2,$$

where $\mu_{ko} = \log \pi_{ko}$ and $\hat{\mu}_k = \log p_k$.

This approximation is obtained by noting that under H_o

$$\pi_{ko}(\hat{\mu}_k - \mu_{ko}) \doteq p_k - \pi_{ko}$$

$$= -(\underline{p} - \underline{\pi}_o)^T \underline{1} - (\underline{\mu} - \underline{\mu}_o)^T \underline{\pi}_o .$$

Note that W_2 is also asymptotically equivalent to the Pearson chisquare test under H_o.

2.2 Other Sampling Schemes

These results for W_1, W_2 and LR_1 are well-known. The question of interest to us here is the implication of the more general assumption that $n^{\frac{1}{2}}(\underline{p} - \underline{\pi}) \to N(\underline{0}, \underline{V})$, where $\underline{V}$ is not necessarily equal to $\underline{P}$. Here $\underline{p}$ is a survey estimate of $\underline{\pi}$ and may depend on sampling weights and other adjustment factors. This situation often arises in sampling under a complex sample design. We assume that $\hat{\underline{V}}$ is a consistent estimate of $\underline{V}$. There are two approaches which we shall consider here. The first is to construct the appropriate Wald statistic for the given sample design. This would be

$$W_3 = n(\underline{p} - \underline{\pi}_o)^T \hat{\underline{V}}^{-1}(\underline{p} - \underline{\pi}_o),$$

where the rank of $\hat{\underline{V}}$ is $k-1$ so that W_3 is asymptotically χ^2_{k-1} under H_o.

An alternative approach would be to use W_1, W_2 or LR_1 directly as a test statistic. Now from multivariate normal theory, we know that the distribution of $n(\underline{p} - \underline{\pi}_o)^T \underline{P}_o^{-1}(\underline{p} - \underline{\pi}_o)$ is that of $\Sigma\, \delta_i Z_i^2$, where $\{Z_i^2\}$ are independent χ^2_1 random variable and $\underline{\delta} = (\delta_1, \ldots, \delta_{k-1})^T$ are the eigenvalues of $\underline{P}_o^{-1}\underline{V}$; see Johnson and Kotz (1970, pg. 150). This result was shown by Rao and Scott (1981), who call the δ_i's generalized design effects. We note that for $k = 2$, we have $\delta = n\sigma_p^2/\{\pi_o(1 - \pi_o)\}$, where $\sigma_p^2 = V\{p\}$. This is the usual design effect for p under H_o.

2.3 Approximations

Now, in general, the distribution function for linear combinations of χ^2_1 random variables is cumbersome, although their moments are easily obtained. Rao and Scott (1981) have suggested two approximations to obtain the significance levels. The first is to approximate the distribution as being proportional to a χ^2_{k-1} random variable, the proportionality constant being determined by equating the mean of the approximating distribution to that of the theoretical distribution. This results in the approximation

$$\sum_{i=1}^{k-1} \delta_i Z_i^2 \doteq \{\sum_{i=1}^{k-1} \delta_i/(k-1)\}\chi^2_{k-1} \tag{2.2}$$

Now,

$$\Sigma\, \delta_i = \mathrm{tr}(\underline{P}_o^{-1}\underline{V})$$

$$= \sum_{i=1}^{k} v_{ii}/\pi_{io}$$

$$= \sum_{i=1}^{k} d_i(1 - \pi_{io}),$$

which depends only on the cell design effects $\{d_i\}$, where v_{ii} is the $i-th$ diagonal element of $\hat{\underline{V}}$ and $d_i = v_{ii}/[\pi_{io}(1 - \pi_{io})]$. This approximation is particularly convenient when the full covariance matrix is not known, but the cell design effects are given. This is often the case for official published data.

Table 1

Age Distribution Among Those Consuming 1-6 Drinks Per Week.
Census Age Distribution for Canada (1978-9)

	Age							
Census Distribution	15-19 .133	20-24 .127	25-34 .218	35-44 .152	45-54 .140	55-64 .115	65 + .115	Total 1.000
Distribution of those consuming 1-6 drinks/week	.117	.150	.264	.175	.148	.093	.053	1.000
Design Effect	1.4	1.2	2.2	1.1	0.6	1.1	1.0	

Example 1

For the Canada Health Survey (1978-9), a stratified multi-stage household survey, data was derived for the age distribution among those consuming one to six drinks per week, based on a sample of 5,204 persons, aged 15 years and over. A description of the survey may be found in "The Health of Canadians" (Statistics Canada Catalogue No. 82-538).

The data, taken from Hidiroglou and Rao (1981), are presented in Table 1. The raw value for W_1 is 298. This is reduced to 248 by taking the approximation given by (2.2). For these data, the post-stratification adjustments for age and sex lead to small design effects.

A second approximation to the distribution of $\Sigma\, \delta_i\, Z_i^2$, suggested by Rao and Scott (1981), is the Satterthwaite (1946) approximation: $\Sigma\, \delta_i\, Z_i^2 \approx a\chi_\nu^2$. To obtain a and ν, it is necessary to compute

$$\Sigma\, \delta_i^2 = \mathrm{tr}\{(\underset{\sim}{P}_o^{-1}\hat{\underset{\sim}{V}})^2\}$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{k} v_{ij}^2/(\pi_{io}\pi_{jo}).$$

However, this depends on all the terms of the matrix $\hat{\underset{\sim}{V}}$. The important point, though, is that some adjustment to the multinomial test statistic is necessary to obtain the appropriate significance level.

An alternative approximation, suggested by Fellegi (1980), is to divide the statistic $n(p-\underset{\sim}{\pi}_o)^T\,\underset{\sim}{P}_o^{-1}(p-\underset{\sim}{\pi}_o)$ by the average design effect, $\bar{d}$, instead of the weighted average given in (2.2). The effect of this on the data in Table 1 is that the adjusted chisquare value is 243, which is comparable to Rao and Scott's (1981) approximation.

3. TESTS OF INDEPENDENCE IN A TWO-WAY TABLE

3.1 Multinomial Sampling

We now suppose that the categories of the multinomial distribution can be cross-classified into an $r\times c$ table, where for the bivariate observation (Y_1, Y_2) we have $\Pr(Y_1 = i, Y_2 = j) = \pi_{ij}$; $\Sigma_{i=1}^{r}\Sigma_{j=1}^{c}\pi_{ij} = 1$. We denote $\pi_{i\cdot} = \Sigma_{j=1}^{c}\pi_{ij}$ and $\pi_{\cdot j} = \Sigma_{i=1}^{r}\pi_{ij}$. We denote $\underset{\sim}{\pi} = (\pi_{11}, \ldots, \pi_{1c}, \ldots, \pi_{r1}, \ldots, \pi_{rc})^T$, $\underset{\sim}{\pi}_R = (\pi_{1\cdot}, \ldots, \pi_{(r-1)\cdot})^T$, $\underset{\sim}{\pi}_C = (\pi_{\cdot 1}, \ldots, \pi_{\cdot(c-1)})^T$, $\underset{\sim}{P}_R = \underset{\sim}{D}_{\pi_R} - \underset{\sim}{\pi}_R\underset{\sim}{\pi}_R^T$, $\underset{\sim}{P}_C = \underset{\sim}{D}_{\pi_C} - \underset{\sim}{\pi}_C\underset{\sim}{\pi}_C^T$. We observe the random vector $\underset{\sim}{n}$ from the multinomial distribution, where $E\{\underset{\sim}{n}\} = n\underset{\sim}{\pi}$. We let $p = \underset{\sim}{n}/n$, $p_{i\cdot} = \sum_j p_{ij}$, and $p_{\cdot j} = \sum_i p_{ij}$.

We wish to test the hypothesis of independence

$$H_o: \pi_{ij} - \pi_{i+}\pi_{+j} = 0 \text{ for } 1 \leq i \leq r-1; 1 \leq j \leq c-1,$$

against the alternative

$$H_1: \pi_{ij} - \pi_{i+}\pi_{+j} \neq 0 \text{ for some } (i, j).$$

If we construct $h_{ij} = p_{ij} - p_{i+}p_{+j}$, for $1 \leq i \leq r-1$ and $1 \leq j \leq c-1$, then under multinomial sampling under H_o, the asymptotic covariance matrix for $\underset{\sim}{h} = (h_{11}, \ldots, h_{1,c-1}, \ldots, h_{r-1,c-1})^T$ is $\underset{\sim}{P}_R \otimes \underset{\sim}{P}_C$, where $\otimes$ denotes the direct matrix product operation. Hence, the Wald statistic under H_o becomes

$$W_4 = \underset{\sim}{h}^T(\hat{\underset{\sim}{P}}_C^{-1} \otimes \hat{\underset{\sim}{P}}_R^{-1})\underset{\sim}{h}$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{c}(p_{ij} - p_{i+}p_{+j})^2/(p_{i+}p_{+j}),$$

the familiar chisquare test with $(r-1)(c-1)$ degrees of freedom.

Another test, which is asymptotically equivalent to W_4 under H_o, is the likelihood ratio test given by

$$LR_2 = 2n\left[\sum_{i=1}^{r}\sum_{j=1}^{c} p_{ij}\log p_{ij} - \sum_{i=1}^{r} p_{i+}\log p_{i+} - \sum_{j=1}^{c} p_{+j}\log p_{+j}\right].$$

An alternative approach for this problem, which is a special case of the methods described in Grizzle, Starmer and Koch (1969) is to consider a Wald statistic based on

$$\{f_{ij} = \log p_{ij} - \log p_{i+} - \log p_{+j}; \text{ for } 1 \leq i \leq r-1, \text{ and } 1 \leq j \leq c-1\}.$$

The asymptotic covariance matrix for $\underset{\sim}{f} = (f_{11}, \ldots, f_{1,c-1}, \ldots, f_{r-1,c-1})^T$ is $(\underset{\sim}{D}_{\pi_R}^{-1} - \underset{\sim}{1}\underset{\sim}{1}^T) \otimes (\underset{\sim}{D}_{\pi_C}^{-1} - \underset{\sim}{1}\underset{\sim}{1}^T)$. Therefore the Wald statistic becomes

$$W_5 = \underset{\sim}{f}^T\left[(\hat{\underset{\sim}{D}}_{\pi_R} + \frac{\underset{\sim}{\pi}_R\underset{\sim}{\pi}_R^T}{p_{r+}}) \otimes (\hat{\underset{\sim}{D}}_{\pi_C} + \frac{\underset{\sim}{\pi}_C\underset{\sim}{\pi}_C^T}{p_{+c}})\right]\underset{\sim}{f}$$

Now under H_o we note that f_{ij} is asymptotically equivalent to

$$\frac{p_{ij}}{\pi_{i+}\pi_{+j}} - \frac{p_{i+}}{\pi_{i+}} - \frac{p_{+j}}{\pi_{+j}} + 1,$$

so that $\sum_{i=1}^{r}\pi_{i+}f_{ij} \doteq \sum_{j=1}^{c}\pi_{+j}f_{ij} \doteq 0$. Using this approximation W_5 becomes

$$W_5' = \sum_{i=1}^{r}\sum_{j=1}^{c} p_{i+}p_{+j}f_{ij}^2.$$

It should be noted that under H_o, the statistics W_4, LR_2 and W_5 are all asymptotically equivalent to

$$\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(p_{ij} - \pi_{i+}\pi_{+j})^2}{\pi_{i+}\pi_{+j}} - \sum_{i=1}^{r}\frac{(p_{i+} - \pi_{i+})^2}{\pi_{i+}} - \sum_{j=1}^{c}\frac{(p_{+j} - \pi_{+j})^2}{\pi_{+j}} \tag{3.1}$$

This result will prove useful in Section 3.3.

3.2 Other Sampling Schemes

Now, relaxing the assumption that n is multinomial, we assume instead that $n^{\frac{1}{2}}(p - \pi) \to N(0, V)$ where p is a survey estimate which may depend on sampling weights and other adjustment factors. For this case, Shuster and Downing (1976) and Fellegi (1980) suggest that we construct the Wald statistic based on $\{h_{ij} = p_{ij} - p_{i.}p_{.j}\}$. If we let J_a be the $(a-1) \times a$ matrix given by

$$J_a = [I \quad 0] \tag{3.2}$$

and let $\hat{H} = (J_r - \pi_R 1^T) \otimes (J_c - \pi_C 1^T) - (\pi_R 1^T \otimes \pi_C 1^T)$

then the Wald statistics is

$$W_6 = h^T(\hat{H}\hat{V}\hat{H}^T)^{-1}h,$$

which under H_0 is asymptotically $\chi^2_{(r-1)(c-1)}$.

Alternatively, we could construct a Wald statistic based on $\{f_{ij} = \log p_{ij} - \log p_{i.} - \log p_{.j}\}$. This is a special case of the log-linear model approach to be discussed in Section 4. We define $(r-1) \times r$ and $(c-1) \times c$ matrices as follows:

$$\hat{E}_R = \left[\hat{D}_{\pi_r}^{-1} \quad 0\right], \quad \hat{E}_C = \left[\hat{D}_{\pi_c}^{-1} \quad 0\right].$$

We let $\hat{F} = (\hat{E}_R - \hat{E}_R 1 1^T) \otimes (\hat{E}_C - \hat{E}_C 1 1^T) - (\hat{E}_R 1 1^T \otimes \hat{E}_C 1 1^T)$.

The appropriate Wald statistics is

$$W_7 = f^T(\hat{F}\hat{V}\hat{F}^T)^{-1}f.$$

Now, analogously to the goodness of fit problem in Section 2, Rao and Scott (1981) have considered null distributions of the test statistics based on W_3, LR_2 and W_5, which are all asymptotically equivalent to the null distribution of (3.1). We see, therefore, that the null distribution is the same as $\sum_{i=1}^{(r-1)(c-1)} \delta_i Z_i^2$, where $\{Z_i^2\}$ are independent χ_1^2 and the δ_i's are the eigenvalues of

$$(P_R^{-1} \otimes P_C^{-1})(HVH^T).$$

Cowan and Binder (1978) investigated the properties of the eigenvalue from a simple two-stage self-weighting design for a 2×2 table. They found that the eigenvalue increases as the degree of independence of the cell proportions within the primary sampling units decreased.

3.3 Approximations

An approximation for the distribution of $\Sigma\, \delta_i Z_i^2$ is

$$\Sigma\, \delta_i Z_i^2 \approx \frac{\Sigma\, \delta_i}{(r-1)(c-1)} \chi^2_{(r-1)(c-1)},$$

as in (2.2). Since the statistic is asymptotically equivalent to(3.1) under H_0, by computing the mean of (3.1) we obtain

$$\Sigma\, \delta_i = \sum_{i=1}^{r}\sum_{j=1}^{c} d_{ij}(1 - \pi_{i+}\pi_{+j}) - \sum_{i=1}^{r} d_i^{(r)}(1 - \pi_{i+}) - \sum_{j=1}^{c} d_j^{(c)}(1-\pi_{+j}),$$

where d_{ij} is the cell design effect; $d_i^{(r)}$ and $d_j^{(c)}$ are the row and column margin design effects, respectively. This particularly simple expression was obtained by Rao and Scott (1983). Fellegi (1980) suggested an alternative approximation as:

$$\left(\sum_{i=1}^{r}\sum_{j=1}^{c} d_{ij}/rc\right) \chi^2_{(r-1)(c-1)}$$

Example 2

In Table 2, we give a 4×2 table from the Canada Health Survey, which cross-classifies drug use (four categories; 0, 1, 2, 3+ drug classes in a 2-day period) and sex (male, female). Here $n = 31{,}668$.

The raw value for W_4 is 774. Rao and Scott's (1981) adjustment reduces this to 437. Fellegi's (1980) adjustment reduces this to 327. The Wald statistics, W_6, is 538. Hidiroglou and Rao (1981) found that the Rao and Scott (1981) approximation performs quite well relative to the Satterthwaite (1946) approximation which is based on the complete covariance matrix.

LOG-LINEAR MODELS

4.1 Multinomial Sampling

We now extend the results of the previous section to more general cross-classifications of the multinomial distribution. The standard results for these models are given in Bishop, Fienberg

Table 2

Variety of Drugs Taken by Sex for Canada (1978-79)

		Number of Drug Varieties				
Sex		0	1	2	3 ±	Total
Male	Proportion	0.293	0.134	0.048	0.021	0.496
	Design Effect	1.56	3.37	1.15	1.38	0.00*
Female	Proportion	0.228	0.159	0.072	0.045	0.504
	Design Effect	3.59	3.13	2.85	1.96	0.00*
Total	Proportion	0.521	0.293	0.120	0.066	1.000
	Design Effect	6.03	6.46	1.65	2.57	

* Because of age-sex post-stratification, these design effects are zero.

and Holland (1975) and Fienberg (1980). We have $\underline{\pi} = (\pi_1, \ldots, \pi_k)^T$ is a vector of cell proportions; $\Sigma_{i=1}^{k} \pi_i = 1$. We observe $\underline{n} = (n_1, \ldots, n_k)^T$, the counts in each cell from a random sample, so that $\underline{n}$ has a multinomial distribution ($\Sigma\, n_i = n$). We let $\underline{p} = \underline{n}/n$ and define

$$\underline{\mu} = \log \underline{\pi}.$$

The log-linear model assumes that for a parameter vector $\underline{\theta} = (\theta_1, \ldots, \theta_t)^T$, we have

$$\underline{\mu}(\underline{\theta}) = u(\underline{\theta})\underline{1} + \underline{X}\underline{\theta},$$

where $\underline{X}$ is a known $k \times t$ matrix of full rank and $\underline{X}^T\underline{1} = \underline{0}$. Note that $t \leq k-1$. If $t = k-1$, we have the saturated model.

The maximum likelihood estimate for $\underline{\theta}$ is given by solving

$$\underline{X}^T(\underline{p} - \hat{\underline{\pi}}) = \underline{0}, \tag{4.1}$$

where $\hat{\underline{\pi}} = \underline{\pi}(\hat{\underline{\theta}})$. Now, asymptotically we have

$$\hat{\underline{\pi}} - \underline{\pi} \doteq \underline{P}\underline{X}(\hat{\underline{\theta}} - \underline{\theta}),$$

where $\underline{P} = \underline{D}_{\pi} - \underline{\pi}\underline{\pi}^T$. From (4.1), we then obtain

$$\hat{\underline{\theta}} - \underline{\theta} \doteq (\underline{X}^T\underline{P}\underline{X})^{-1}\underline{X}^T(\underline{p} - \underline{\pi})$$

and

$$\hat{\underline{\pi}} - \underline{\pi} \doteq \underline{P}\underline{X}(\underline{X}^T\underline{P}\underline{X})^{-1}\underline{X}^T(\underline{p} - \underline{\pi}).$$

Since $n^{1/2}(\underline{p} - \underline{\pi}) \rightarrow N(\underline{0}, \underline{P})$ we obtain

$$n^{1/2}(\hat{\underline{\theta}} - \underline{\theta}) \rightarrow N[\underline{0}, (\underline{X}^T\underline{P}\underline{X})^{-1}]$$

$$n^{1/2}(\hat{\underline{\pi}} - \underline{\pi}) \rightarrow N[\underline{0}, \underline{P}\underline{X}(\underline{X}^T\underline{P}\underline{X})^{-1}\underline{X}^T\underline{P}].$$

Suppose now that the linear expression $\underline{X}\underline{\theta}$ can be decomposed as $\underline{X}_1\underline{\theta}_1 + \underline{X}_2\underline{\theta}_2$ where $\underline{X}_1$ and $\underline{X}_2$ are full rank, $\underline{X}_1$ is $k \times r$, $\underline{X}_2$ is $k \times s$, $\underline{\theta}_1$ is $r \times 1$ and $\underline{\theta}_2$ is $s \times 1$, where $r + s = t$.

We consider the problem of testing

$$H_o: \underline{\theta}_2 = \underline{0},$$

against the alternative

$$H_1: \underline{\theta}_2 \neq \underline{0}.$$

We use $\hat{\underline{\theta}}_1$, $\hat{\underline{\theta}}_2$, $\hat{\underline{\pi}}$, etc. to denote the estimates under the full model H_1. Alternatively, we let $\tilde{\underline{\theta}}_1$, $\tilde{\underline{\pi}}$, to denote estimates under H_o.

Now,

$$n^{1/2}(\hat{\underline{\theta}}_2 - \underline{\theta}_2) \rightarrow N[\underline{0}, (\tilde{\underline{X}}_2^T\underline{P}\tilde{\underline{X}}_2)^{-1}]$$

where

$$\tilde{\underline{X}}_2 = [\underline{I} - \underline{X}_1(\underline{X}_1^T\underline{P}\underline{X}_1)^{-1}\underline{X}_1^T\underline{P}]\underline{X}_2 \tag{4.2}$$

so that the Wald statistic is

$$W_8 = n\hat{\underline{\theta}}_2^T \tilde{\underline{X}}_2^T \hat{\underline{P}} \tilde{\underline{X}}_2 \hat{\underline{\theta}}_2.$$

Under H_o, this is asymptotically equivalent to the Pearson chisquare statistic

$$n(\hat{\underline{\pi}} - \hat{\tilde{\underline{\pi}}})^T \hat{\underline{D}}_\pi^{-1} (\hat{\underline{\pi}} - \hat{\tilde{\underline{\pi}}}),$$

or the likelihood ratio test

$$LR_3 = 2n \sum_{i=1}^{k} p_i \log(\hat{\pi}_i / \hat{\tilde{\pi}}_i).$$

Under H_o, these statistics are asymptotically χ^2_s.

4.2 Other Sampling Schemes

We still assume that the cell proportions, $\underline{\pi}$, satisfy $\underline{\mu} = \log \underline{\pi} = u(\underline{\theta}_1, \underline{\theta}_2)\underline{1} + \underline{X}_1\underline{\theta}_1 + \underline{X}_2\underline{\theta}_2$ but we now have $n^{1/2}(\underline{p} - \underline{\pi}) \to N(\underline{0}, \underline{V})$, where $\underline{p}$ is a survey estimate.

Rao and Scott (1983) suggest the following Wald statistic for testing $\underline{\theta}_2 = \underline{0}$. We let $\underline{C}$ be any $k \times s$ matrix with $\underline{C}^T\underline{X}_1 = \underline{0}$, $\underline{C}^T\underline{1} = \underline{0}$ and $\underline{C}^T\underline{X}_2$ nonsingular. For example if $\underline{X}_1^T\underline{X}_2 = \underline{0}$ then $\underline{C} = \underline{X}_2$ is convenient. Now the hypothesis is equivalent to $\underline{C}^T\underline{\mu} = \underline{0}$. We have

$$\underline{C}^T(\hat{\underline{\mu}} - \underline{\mu}) \doteq \underline{C}^T\underline{D}_\pi^{-1}(\hat{\underline{\pi}} - \underline{\pi})$$

$$\doteq \underline{C}^T\underline{X}(\underline{X}^T\underline{P}\underline{X})^{-1}\underline{X}^T(\underline{p} - \underline{\pi}),$$

where $\underline{\pi}$ is obtained from (4.1), based on the survey estimate, $\underline{p}$.

We therefore have the Wald statistics

$$W_9 = n\hat{\underline{\mu}}^T\underline{C}[\underline{C}^T\underline{X}(\underline{X}^T\hat{\underline{P}}\underline{X})^{-1}(\underline{X}^T\hat{\underline{V}}\underline{X})(\underline{X}^T\hat{\underline{P}}\underline{X})^{-1}\underline{X}^T\underline{C}]^{-1}\underline{C}^T\hat{\underline{\mu}}.$$

Similar results were also given in Binder (1983). If under H_1, the model is saturated ($r+s = k-1$), then $\underline{p} = \underline{\pi}$ and we obtain

$$W_9 = n\hat{\underline{\mu}}^T\underline{C}[\underline{C}^T\hat{\underline{D}}_\pi^{-1}\hat{\underline{V}}\hat{\underline{D}}_\pi^{-1}\underline{C}]^{-1}\underline{C}^T\hat{\underline{\mu}}.$$

Rao and Scott (1984) show that if we use $\hat{\underline{P}}$ instead of $\hat{\underline{V}}$ in W_9 then these are asymptotically equivalent to the likelihood ratio or Pearson χ^2 test statistics. They also show that the likelihood ratio test statistics is distributed as $\sum_{i=1}^{s} \delta_i Z_i^2$ under H_o, where $\{Z_i^2\}$ are independent χ^2_1 and $\{\delta_i\}$ are the eigenvalues of

$$(\tilde{\underline{X}}_2^T\underline{P}\tilde{\underline{X}}_2)^{-1}(\tilde{\underline{X}}_2^T\underline{V}\tilde{\underline{X}}_2), \tag{4.3}$$

for $\tilde{\underline{X}}_2$ defined in (4.2).

4.3 Approximations

As before, we approximate the null distribution

$$\sum_{i=1}^{s} \delta_i Z_i^2 \approx \left(\frac{\sum \delta_i}{s}\right) \chi^2_s.$$

This involves computing the trace of (4.3). Rao and Scott (1984) show that if the model admits explicit solutions for both $\hat{\underline{\pi}}$ and $\hat{\tilde{\underline{\pi}}}$, then the approximation depends on the matrix $\underline{V}$ only through cell design effects and marginal design effects. This observation is particularly convenient when ony the estimated design effects for the cell proportions and margins are available, as is often the case for published tables.

Example 3

Hidiroglou and Rao (1983) considered all direct estimates from the three-way table: Drug use (5 categories: 0, 1, 2, 3, 4+ drug classes in a 2 day period) × Age (4 categories; 0-14, 15-44, 45-64, 65+) × Sex (male, female), taken from the Canada health Survey. We give the results for testing whether Age and Sex are independent in each drug category (n = 31,668). This is equivalent to the hypothesis

$$H_0\colon \pi_{ijk} = \pi_{ij+}\frac{\pi_{i+k}}{\pi_{i++}}.$$

Using Bishop, Fienberg and Holland's (1975) notation, where $\log \pi_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)}$, the null hypothesis is equivalent to

$$H_0\colon u_{23(jk)} = u_{123(ijk)} = 0 \quad \text{for all } (i, j, k).$$

The raw chisquare value is 23 based on 15 degrees of freedom. The average eigenvalue is 1.39, so that the approximation reduces the chisquare value to 16. Whereas the unadjusted chisquare value would lead the analyst to reject the hypothesis at the 10% level, the approximation indicates that h_0 cannot be rejected even at the 30% level.

5. LOGISTIC REGRESSION MODELS

5.1 Multinomial Sampling

We now consider a logistic regression model for the conditional distribution of a binary response variable y given the vector $\underline{x}$ of independent variables. In particular, this conditional distribution is

$$\Pr(y_i \mid \underline{x}_i) = \pi(\underline{x}_i)^{y_i}\,[1 - \pi(\underline{x}_i)]^{1-y_i},$$

where $y_i \in \{0, 1\}$.

For the logistic regression model, we have

$$\log\left\{\frac{\pi(\underline{x}_i)}{1 - \pi(\underline{x}_i)}\right\} = \underline{x}_i^T\,\underline{\theta},$$

where $\underline{\theta}$ is an unknown vector of parameters.

We note that if $\underline{x}_i$ is a categorical vector of 0's and 1's, this is a special case of a log-linear model as described in Section 4. Here we allow $\underline{x}_i$ to be arbitrary. The extension to the case of k-categories for the y-variable is straight-forward, it is also possible to generalize the model to

$$\log\left\{\frac{\pi(\underline{x}_i)}{1 - \pi(\underline{x}_i)}\right\} = f(\underline{x}_i^T\,\underline{\theta}),$$

for a known function $f(\cdot)$, but we do not discuss this here.

Now, the maximum likelihood estimate for $\underline{\theta}$ is given by

$$\underline{X}^T(\underline{y} - \hat{\underline{\pi}}) = \underline{0}$$

where $\underline{y} = (y_1, \ldots, y_n)^T$, $\hat{\underline{\pi}} = [\hat{\pi}(\underline{x}_1), \ldots, \hat{\pi}(\underline{x}_n)]^T$ and $\underline{X} = [\underline{x}_1 \ldots \underline{x}_n]^T$.

Under suitable regularity conditions, we have

$$n^{1/2}(\hat{\underline{\theta}} - \underline{\theta}) \to N[0,\ n\,(\underline{X}^T\underline{\Delta}\underline{X})^{-1}], \text{ where } \underline{\Delta} = \underline{D}_\pi(\underline{I} - \underline{D}_\pi).$$

If we have $\underset{\sim}{X}\underset{\sim}{\theta} = \underset{\sim}{X}_1\underset{\sim}{\theta}_1 + \underset{\sim}{X}_2\underset{\sim}{\theta}_2$ and consider testing the hypothesis

$$H_0\colon \underset{\sim}{\theta}_2 = \underset{\sim}{0}$$
$$H_1\colon \underset{\sim}{\theta}_2 \neq \underset{\sim}{0},$$

we obtain the Wald statistic

$$W_{10} = n\,\hat{\underset{\sim}{\theta}}_2^T(\hat{\underset{\sim}{X}}_2^T\,\underset{\sim}{\Lambda}\,\hat{\underset{\sim}{X}}_2)\,\hat{\underset{\sim}{\theta}}_2$$

where

$$\hat{\underset{\sim}{X}}_2 = [I - \underset{\sim}{X}_1(\underset{\sim}{X}_1^T\underset{\sim}{\Lambda}\underset{\sim}{X}_1)^{-1}\underset{\sim}{X}_1^T\underset{\sim}{\Lambda}]\underset{\sim}{X}_2.$$

The likelihood ratio test here is

$$LR_4 = 2\sum_{i=1}^{n}\left[y_i\log\left(\frac{\hat{\pi}_i}{\hat{\bar{\pi}}_i}\right) + (1 - y_i)\log\left\{\frac{(1 - \hat{\pi}_i)}{(1 - \hat{\bar{\pi}}_i)}\right\}\right]$$

which is asymptotically equivalent to W_{10} under H_0.

5.2 Other Sampling Schemes

Suppose now that $n^{-1/2}\underset{\sim}{X}^T(\underset{\sim}{y} - \underset{\sim}{\pi}) \to N(\underset{\sim}{0}, \underset{\sim}{V})$ and that $\hat{\underset{\sim}{V}}$ is a consistent estimator of $\underset{\sim}{V}$. Here $\underset{\sim}{y}$ is not necessarily a vector of 0's and 1's, but may in fact depend on the sampling weights and other adjustment factors. Estimating $\underset{\sim}{V}$ is usually possible since $\underset{\sim}{X}^T(\underset{\sim}{y} - \underset{\sim}{\pi})$ is the sum of random observations and most sample designs admit a consistent estimator of the sum of (not necessarily independent) observations. To estimate $\underset{\sim}{V}$ we use $\hat{\underset{\sim}{\pi}}$ instead of $\underset{\sim}{\pi}$ in the estimate. Since asymptotically

$$(\hat{\underset{\sim}{\theta}} - \underset{\sim}{\theta}) \doteq (\underset{\sim}{X}^T\underset{\sim}{\Lambda}\underset{\sim}{X})^{-1}\underset{\sim}{X}^T(\underset{\sim}{y} - \underset{\sim}{\pi}),$$

we have that

$$n^{1/2}(\hat{\underset{\sim}{\theta}} - \underset{\sim}{\theta}) \to N[\underset{\sim}{0}, n^2(\underset{\sim}{X}^T\underset{\sim}{\Lambda}\underset{\sim}{X})^{-1}\underset{\sim}{V}(\underset{\sim}{X}^T\underset{\sim}{\Lambda}\underset{\sim}{X})^T];$$

see Binder (1983) for a detailed justification of this result. Now, a Wald statistic may be constructed from the estimated covariance matrix for $\underset{\sim}{\theta}_2$.

Table 3

Logistic Regression Model for Explaining Use of Physician Services

Variable	Type	d.f.	Wald Statistic
Age	Categorical	4	19.232
Sex	Categorical	1	12.494
Age-Sex Interactions	Categorical	4	36.001
Family Income	Categorical	5	14.642
Occupation	Categorical	3	8.614
Occupation-Sex Interactions	Categorical	3	11.501
Marital Status	Categorical	3	45.752
Medical History	Categorical	2	36.700
Number of Health Problems	Quantitative	1	81.554
Drug Use	Categorical	2	272.175
Number of Accidents	Quantitative	2	106.372
Number of Disability Days	Quantitative	2	29.052
Community Size	Categorical	2	11.751
Provincial Physician - Population Ratios	Quantitative	1	0.540

Example 4

A logistic regression model was fit on 20,726 respondents from the Canada health survey to explain use or non-use of physician services over a 12-month period. In total it was estimated that 77% of the population visited a physician at least once. The results are summarized in Table 3. For more complete details, see Binder (1983). The logistic model seemed to fit the data very well.

5.3 Qualitative Explanatory Variables

The theory of this section was obtained by G. Roberts in an unpublished manuscript (Carleton University). Here the explanatory variables are all qualitative. We label the domains, $\{1, \ldots, I\}$. We let p_i be the survey estimate of the i-th domain proportion and $\hat{N}_i$ is the estimate of the size of the i-th domain, N_i. Under the model, the expected proportion in the i-th domain is f_i, where

$$\log \{f_i/(1-f_i)\} = \underline{a}_i^T\underline{\theta},$$

for $\underline{a}_i$ known and $\underline{\theta}$ an unknown parameter. We define $\underline{A} = [\underline{a}_1, \ldots, \underline{a}_I]^T$ and let $\underline{D}_{\hat{N}} = \text{diag}\{\hat{N}_1, \ldots, \hat{N}_I\}$.

Under the model, the survey estimator of $\underline{f} = (f_1, \ldots, f_I)^T$ is given by $\hat{\underline{f}}$, the solution to

$$\underline{A}^T\underline{D}_{\hat{N}}(\underline{p} - \hat{\underline{f}}) = \underline{0}. \tag{5.1}$$

Since asymptotically

$$\hat{\underline{\theta}} - \underline{\theta} \doteq (\underline{A}^T\underline{\Delta}\underline{A})^{-1}\underline{A}^T\underline{D}_{\hat{N}}(\underline{p} - \underline{f}),$$

where $\underline{\Delta} = \text{diag}\{N_1 f_1(1-f_1), \ldots, N_I f_I(1-f_I)\}$, we have

$$n^{1/2}(\hat{\underline{\theta}} - \underline{\theta}) \to N[\underline{0}, (\underline{A}^T\underline{\Delta}\underline{A})^{-1}\underline{A}^T\underline{D}_{\hat{N}}\underline{V}_p\underline{D}_{\hat{N}}\underline{A}(\underline{A}^T\underline{\Delta}\underline{A})^{-1}]$$

whenever $n^{1/2}(\underline{p} - \underline{f}) \to N(\underline{0}, \underline{V}_p)$.

Under independent binomial sampling, the covariance matrix reduces to $(N/n)(\underline{A}^T\underline{\Delta}\underline{A})^{-1}$, where n is the sample size.

The likelihood ratio test for testing goodness of fit is

$$LR_s = 2(n/\hat{N})\sum_{i=1}^{I}\hat{N}_i[p_i\log(p_i/\hat{f}_i) + (1-p_i)\log\{(1-p_i)/(1-\hat{f}_i)\}],$$

where n is the sample size and $\hat{N} = \Sigma\,\hat{N}_i$. Under H_0 this is asymptotically equivalent to

$$W_{11} = (n/\hat{N})\sum_{i=1}^{I}\hat{N}_i(p_i - \hat{f}_i)^2/[f_i(1-f_i)].$$

In general, the distribution of LR_s will be that of $\Sigma\,\delta_i Z_i^2$, where $\{Z_i\}$ are independent χ_1^2, and $\{\delta_i\}$ are the eigenvalues of $\hat{N}^{-1}\underline{D}_{\hat{N}}[\underline{\Delta}^{-1} - \underline{A}(\underline{A}^T\underline{\Delta}\underline{A})^{-1}\underline{A}^T]\underline{D}_{\hat{N}}\underline{V}_p\underline{D}_{\hat{N}}[\underline{\Delta}^{-1} - \underline{A}(\underline{A}^T\underline{\Delta}\underline{A})^{-1}\underline{A}^T]\underline{\Delta}\underline{D}_{\hat{N}}^{-1}$. By taking the expectation of W_{11}, and approximating

$$W_{11} \approx \frac{\Sigma\,\delta_i}{I-s}\chi^2_{I-s}$$

where $s = \text{rank}(\underline{A})$, we obtain

$$\Sigma\,\delta_i = (n/\hat{N})\sum_{i=1}^{I}\hat{N}_i v_{ii}^{(p)}/\{f_i(1-f_i)\}$$

where $v_{ij}^{(r)} = V\{p_i - \hat{f}_i\}$. The $\{v_{ij}^{(r)}\}$ may be computed using the relationship $p - \hat{f} \doteq [I - \text{diag}\{f_i(1-f_i)\} A (A^T \Delta A)^{-1} A^T D_N] (p - f)$.

Example 5

The data from the October 1980 Canadian Labour Force Survey was used to fit logistic (logit) models for the probability of being employed. The sample consisted of males aged 15-64 who were in the labour force and not full time students. A logit model, quadratic in age and in education, was fitted. Age-group levels were formed by dividing the interval [15, 64] into ten groups with the jth age-group being the interval $[10 + 5j, 14 + 5j]$, $j = 1, 2, \ldots, 10$. The midpoint of each age-group was used as the value of the age for all persons in that age-group. Six levels of education were formed by assigning to each person a value based on the median years of schooling. Age by education classification led to the formation of 60 cells.

Let π_i = Pr{an individual in the ith cell is employed}, $i = 1, 2, \ldots, 60$. We assume that $0 < \pi_i < 1$. Hence $1 - \pi_i$ represents the probability that the individual in the ith cell is unemployed. The model, considered for fit, was

$$\ln \frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 a_i + \beta_2 a_i^2 + \beta_3 d_i + \beta_4 d_i^2, \qquad (1)$$

$$i = 1, 2, \ldots, 60$$

where a_i and d_i are the age and education variable values for the individuals in the ith cell.

Using the survey estimates p_i of π_i, the values of Pearson's statistic W_{11} and the likelihood ratio statistic LR_5 were computed as $W_{11} = 98.94$ and $LR_5 = 101.20$. The upper 5% point of the chi-square distribution, with 55 degrees of freedom, is 73.31. Using these values of W_{11} or LR_5 we would reject the model 1. These values of W_{11} or LR_5, however, are appropriate only if the sample was a random sample.

The estimate average eigenvalue, $\Sigma \hat{\delta}_i/55$, for testing goodness of fit for this data is 1.88. This would reduce W_{11} to 52.63 and LR_5 to 53.83. Hence, with this adjustment, we find that the data are consistent with the model (1).

The use of the Wald statistic, $(p - \hat{f})^T [\hat{V}^{(r)}]^- (p - \hat{f})$, for testing the goodness of fit was also considered. Here we use the g-inverse of $\hat{V}^{(r)}$ since the matrix is singular. Some perturbation to the estimates of p_i, when $p_i = 1$, was necessary for computing the Wald statistic. It was found that the Wald statistic was unstable for our problem. Minor perturbations in the estimates of p led to considerable change in the value of the Wald statistic. Also the value of the Wald statistic is very large here due to instability in the estimated covariance matrix involved in its calculation. The Wald statistic is at least 30 times larger than our adjusted Chi-squared values.

6. SOFTWARE CONSIDERATIONS

Advancement of computer technology has made data collection, storage and retrieval operations easy and efficient. Powerful generalized software systems, such as TPL, STATPAK and ESTIMATION SYSTEM, have been used to produce cell estimates and some of their variances fairly easily to users and analysts. As well a number of commercially available packages such as BMDP, SPSS and SAS are powerful analytic tools in certain contexts. However, the ability to perform analysis such as those described in this paper are limited. For example, in situations involving hypothesis testing or statistical inference, these packages assume that the data to be analyzed come from surveys with simple random samples.

At present, an integrated software package, similar to the ones mentioned above, but designed for analyses of the type of data discussed in this paper, is not available. As a result, the researcher requiring a quick solution to his problem is usually forced to use existing statistical packages which may not be appropriate.

The alternatives are

- use existing packages with modifications
- use existing stand-alone software
- write customized programs
- use combinations of the above.

For the analyses given in this paper, modifications to the MINI CARP program (Hidiroglou, Fuller and Hickman; 1980) were incorporated to obtain the results in Examples, 1, 2 and 3. For Example 4, a combination of PL/1 and SAS programs were developed. The analysis of the Labour Force Survey data (Example 5) used a combination of customized programs and SAS.

For the above alternatives, some practical drawbacks have been experienced. they include:

(a) If an existing package is to be modified, intimate knowledge of the package is often required;
(b) Identical information may have to be duplicated on separate data files, as these alternatives are not integrable like generalized systems;
(c) Compared to an integrated "user-friendly" package, these alternatives lack elegance and operational efficiency as software;
(d) Comprehensive documentation is not generally available for specially written programs limiting the availability of software.

Work is now ongoing to develop SAS based procedures for performing many of these analyses. Our ultimate goal is similar to that proposed by Shah (1981); namely, the development of an integrated software package for survey data analysis. This is a goal worth striving for, if we are to avoid the frustrations now being experienced by researchers who are faced with either developing their own software or using existing software which could lead to erroneous results and conslusions.

7. DISCUSSION

We have examined a number of problems which arise when fitting models to categorical data which have been collected under complex sampling designs. The basic approach has been to derive the appropriate Wald statistic for the fitted model or to use the test statistic which is motivated from multinomial-type sampling designs and find a suitable approximation to its null distribution.

We have not addressed the issue as to whether one should really be taking a model-based or design-based approach to begin with. Instead, we have concentrated on design-based inferences.

To put this issue into focus, let us reconsider the test of independence in a two-way contingency table. The question of independence arises if we are interested in whether knowing the value of variable Y_1 affects our knowledge about variable Y_2. If it does not, for all the individuals in the population, then we say the variables are independent. However, if we also know the value of Y_3, it may turn out that Y_1 and Y_2 are no longer independent. This is particularly important when Y_3 is a design variable (such as geographic stratum). Since design variables are usually known for all sampled individuals, we have one of two options: (a) we can say that the question of independence is no longer relevant, or (b) we can marginalize out Y_3, and say that we are only interested in Y_1 and Y_2, unconditionally. Assuming that we take approach (b), the resulis of this paper seem appropriate. In some cases it may be possible to test if Y_1 and Y_2 are conditionally independent given Y_3.

There is a further difficulty, however. Suppose we are interested in the cell proportions π_{ij} from a finite population of size N. If we were to take a census from this population, it is highly unlikely that we would obtain $\pi_{ij} = \pi_{i+}\pi_{+j}$ exactly. The best that we could hope for is that some measure of association such as $N \sum\sum (\pi_{ij} - \pi_{i+}\pi_{+j})^2/\pi_{i+}\pi_{+j}$ is small. Note that even

under a super-population model of exact independence, we would not expect this measure of association to be zero. Perhaps, we should instead be testing hypotheses such as

$$H_0: \text{Measure of Association} \leq C$$

$$H_1: \text{Measure of Association} > C.$$

Further research is needed in this area. However, for practical circumstances where the sampling fraction is not large, the methods given in this paper are suitable.

REFERENCES

ALTHAM, P.A.E. (1976). Discrete vriable analysis for idividuals grouped into families. *Biometrika*, 63, 263-269.

BINDER D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

BISHOP, Y.M.M., FIENBERG, S.E. and HOLLAND, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, Massachusetts: MIT Press.

BRIER, S.S. (1978). Discrete Data Models with Random Effects. Technical Report, University of Minnesota, School of Statistics.

COHEN, J.E. (1976). The distribution of the chi-squared statistic under cluster sampling from contingency tables. *Journal of the American Statistical Association*, 71, 665-670.

COWAN, J. and BINDER, D.A. (1978). The effect of a two-stage sample design on tests of independence in a 2 by 2 table. *Survey Methodology*, 4, 16-28.

FAY, R.E. (1985). A jackknifed chi-squared test for complex samples. *Journal of American Statistical Association*, 80, 148-157.

FELLEGI, I.P. (1980). Approximate tests of independence and goodness of fit based on stratified multistage samples. *Journal of American Statistical Assocociation*, 71, 665-670.

FIENBERG, S.E. (1980). *The Analysis of Cross Classified Data*, (2nd ed.). Cambridge, Massachusetts: MIT Press.

GRIZZEL, J.E., STARMER, C.F. and KOCH, G.G. (1969). Analysis of categorical data by linear models. *Biometrics*, 25, 489-504.

HIDIROGLOU, M.A., FULLER, W.A. and HICKMAN, R.D. (1980). MINICARP: A program for estimating simple descriptive statistics and their variances for multi-stage stratified designs. Iowa State University: Ames, Iowa.

HIDIROGLOU, M.A. and RAO, J.N.K. (1981). Chisquare tests for the analysis of categorical data from the Canada Health Survey. Paper presented at the International Statistical Institute Meetings, Buenos Aires, 1981.

HIDIROGLOU, M.A. and RAO, J.N.K. (1983). Chi-square tests for the analysis of three-way contingency tables from the Canada Health Survey. Technical Report, Statistics Canada.

IMREY, P.B., KOCH, G.G. and STOKES (1981). Categorical data analysis: Some reflections on the log linear model and logistic regression. part I: Historical and Methodological Ovewview. *International Statistical Review*, 49, 265-283.

JOHNSON,N.L. and KOTZ, S. (1970). *Continuous Univariate Distributions*. Boston: Houghton Mofflin.

KOCH, G.G., FREEMAN, D.H. JR., and FREEMAN, J.L. (1975). Strategies in the Multivariate Analysis of Data from Complex Surveys. *International Statistical Review*, 43, 59-78.

RAO, J.N.K. and SCOTT, A.J. (1981). The analysis of categorical data from complex sample surveys: Chi-square tests for goodness of fit and independence in two-way tables. *Journal of American Statistical Association*, 76, 221-30.

RAO, J.N.K. and SCOTT, A.J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *The Annals of Statistics*, 12, 48-60.

SATTERTHWAITE, F.E. (1946). An approximate distribution of estimates of variance components. *Biometrics*, 2, 110-114.

SHAH, B.V. (1981). Development of survey data analysis software. Research Triangle Institute, Research Triangle Park, North Carolina.

SHUSTER, J.J. and DOWNING, D.J. (1976). Two-way contingency tables for complex sampling schemes. *Biometrika*, 63, 271-276.