ON χ^2 TESTS FOR CONTINGENCY TABLES FROM COMPLEX SAMPLE SURVEYS WITH FIXED CELLS AND MARGINAL DESIGN EFFECTS

Ha H. Nguyen, Merck & Co and Charles Alexander, US Bureau of the Census
Ha H. Nguyen, Merck Sharp and Dohme Research Laboratories, P.O. Box 2000, Rahway, NJ 07065

Goodness of fit tests and tests for independence of hierarchical log-linear models are studied for the special case where samples are obtained from complex survey designs with fixed cell and marginal design effects. The asymptotic null distribution of the χ^2 test is derived based on the results of Rao and Scott (1981, 1984). It is shown that, as a first order approximation, the usual χ^2 test divided by the common cell and marginal design effect has a χ^2 distribution.

A. Goodness of Fit Test

Background

Suppose that we have a discrete population with k categories where $\Pr(Y = i) = p_i$; $\sum_{i=1}^{k} p_i = 1$. Assume that we draw n observations from this population using a specific sampling design $\mathcal{P}(S)$. Furthermore, assume that the sample contains n_1 observations in the first category, ..., n_k observations in the k^{th} category ($\sum_{i=1}^{k} n_i = n$).

It is well known that the null hypothesis

$$H_0: \quad p_i = p_{0i} \quad i = 1, \dots k$$

for given p_{0i} can be tested using the Wald statistic

$$X_w^2 = n(\hat{\mathbf{p}} - \mathbf{p}_0)'V^{-1}(\hat{\mathbf{p}} - \mathbf{p}_0) \tag{1}$$

where
$\mathbf{p}_0 = (p_{01},\ p_{02}\ \dots,\ p_{0k-1})'$, $\hat{\mathbf{p}} = (\hat{p}_1, \hat{p}_2 \dots, \hat{p}_{k-1})'$ $= \frac{1}{n}(n_1,\ n_2,\ \dots,\ n_{k-1})'$ is an estimate of $\mathbf{p}$ based on S and $\frac{V}{n}$ is the covariance matrix of $\hat{\mathbf{p}}$.

For large n, under H_0, the Wald statistic will have asymptotically a χ^2 distribution with $k-1$ degrees of freedom. However, this statistic requires the knowledge of V which may not be readily available, especially for complex sample designs.

Alternatively, the hypothesis H_0 can also be tested using the Pearson Chi-squared statistic

$$X_p^2 = n(\hat{\mathbf{p}} - \mathbf{p}_0)'P_0^{-1}(\hat{\mathbf{p}} - \mathbf{p}_0) \tag{2}$$

where $P_0 = \operatorname{diag}(\mathbf{p}_0) - \mathbf{p}_0\mathbf{p}_0'$.

We note that, under H_0, this statistic can be computed easily from the summary table. Under simple random sampling, (1) and (2) are equivalent since $V = var(\hat{\mathbf{p}}) = P_0$. However, for general sampling designs, under H_0, the Pearson statistic will asymptotically be distributed as a linear combination of $k-1$ independent χ^2 random variables with 1 degree of freedom, that is, $X_p^2 \sim \sum_{i=1}^{k-1} \delta_i \chi_1^2$ where the δ_i's are eigenvalues of $P_0^{-1}V$.

Case of constant design effects

For a given sampling design $\mathcal{P}(S)$, the design effect (*deff*) of an estimate T is defined as the ratio of the variances of the estimate under the design $\mathcal{P}(S)$ and under simple random sampling:

$$deff(T) = \frac{var(T)}{var_{srs}(T)}$$

where var_{srs} denotes the variance under simple random sampling; $deff$ reflects the effect of the design on the variance of the estimate when compared to simple random sampling.

Suppose that the design effects of the p_i's are constant, * that is,

$$\operatorname{var}(\hat{p}_i) = \delta \operatorname{var}_{srs}(\hat{p}_i) = \delta p_{0i}(1 - p_{0i})/n \quad i = 1, 2, \dots, k$$

then the expected value of X_p^2 is

$$\begin{aligned} E\left(X_p^2\right) &= \sum_{i=1}^{k-1} \delta_i\, E\left(\chi_1^2\right) \\ &= \sum_{i=1}^{k-1} \delta_i\, 1 \\ &= \operatorname{trace}(P_0^{-1}V) \\ &= \sum_{i=1}^{k} \frac{v_{ii}}{p_{0i}} \qquad \text{(see Appendix 1)} \\ &= \sum_{i=1}^{k} \frac{\delta p_{0i}(1 - p_{0i})}{p_{0i}} \\ &= \delta \sum_{i=1}^{k} (1 - p_{0i}) \\ &= \delta(k - 1) \end{aligned}$$

Hence, $\frac{X_p^2}{\delta}$ has the same first moment as a χ^2 random variable with $k-1$ degrees of freedom. Thus, when the design effects for cells are assumed to be constant, the Pearson statistic will have, as a first order approximation, the distribution of a χ^2 random variable with

* It can be shown that except for the case $k = 2$, these assumptions do not imply that $var(\hat{\mathbf{p}})$ is equal to δP_0 (P_0 being the variance of $\hat{\mathbf{p}}$ under simple random sampling).

$k-1$ degrees of freedom. In other words, for complex sample designs with fixed design effects, tests that are based on X_p^2 can be approximately done using a χ^2 table.

B. Log-Linear Model

In categorical data analysis, when the observations have more than one characteristic of interest, it is often the case that we would like to study how these characteristics interrelate. The study of these associations and interactions can be nicely formulated using a log-linear model.

Notation and background

Suppose that we have an r-dimensional contingency table with independent variables x_1, x_2, ..., x_r, each having respectively v_1, v_2, ..., v_r categories. When $r = 3$, the indices i, j, k can be used to denote a given cell in the table. For example $\pi_{i,j,k}$ will denote the probability that an observation will be in the cell i, j, k. This notation can be generalized by using a single symbol, usually θ, to denote the complete set of subscripts. Thus, π_θ will be the probability that an observation will be in an elementary cell θ.

In this paper we will only consider hierarchical models as defined by Birch (1963). This means that the cell probabilities are permitted to be log-linearly related in such a way that a suitable set of marginals, usually called the minimal set of fitted marginals, is sufficient for the parameters. Tables of sums of non elementary cells will be called configurations and will be denoted by the letter C (Bishop et al. (1975)). For example, in a three-way contingency table, the table of partial sums $x_{ij+} = \sum_k x_{i,j,k}$, obtained by summing over the third variable, will be denoted by C_{12}. As the third variable has been removed by summing, the subscripts of C refer only to the remaining two variables. Configurations corresponding to the minimal set of fitted marginals, as defined above, will be called the sufficient configurations.

Bishop et al. (1975, page 68) outlined a method to derive sufficient configurations for comprehensive, unsaturated and hierarchical models. For such models, if the sufficient configurations are given, it is trivial to write down the log-likelihood function, $\log m_\theta$. Also, it can be shown that the number of independent parameters in the model can be expressed in terms of the numbers of cells in the sufficient configurations.

Indeed, when only C_θ is the sufficient configuration of the model, it is clear that the number of independent variables in the model is equal to the number of cells in C_θ. In other words, if $\mathcal{U}_\theta$ is the set of all linearly independent u-terms whose subscripts are subsets of θ (which is, in this case, the set of all linearly independent parameters for the model) then the cardinality of $\mathcal{U}_\theta$ is equal to the number of cells in C_θ. This result implies that if the sufficient configurations are C_{θ_i}, $i = 1, \ldots, k$ and the $\mathcal{U}_{\theta_i}$'s sets are defined as above then $\mathcal{U} = \mathcal{U}_{\theta_1} \cup \mathcal{U}_{\theta_2} \ldots \cup \mathcal{U}_{\theta_I}$ will be the set of all linearly independent parameters and the cardinality of $\mathcal{U}$ can be found using the inclusion-exclusion principle. For instance, the three dimensional contingency table with no three factor effect, $\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)}$ with $i = 1, \ldots, I$, $j = 1, \ldots, J$, $k = 1, \ldots, K$, has C_{12}, C_{13} and C_{23} as sufficient configurations. Let $\mathcal{U}$ be the set of all linearly independent parameters then

$$\begin{aligned}
\mathrm{card}(\mathcal{U}) &= 1 + (I-1) + (J-1) + (K-1) + \\
&\quad (I-1)(J-1) + (I-1)(K-1) + \\
&\quad (J-1)(K-1) \\
&= IJ + IK + JK - I - J - K + 1 \\
&= \mathrm{card}(\mathcal{U}_{12}) + \mathrm{card}(\mathcal{U}_{13}) + \mathrm{card}(\mathcal{U}_{23}) - \\
&\quad \mathrm{card}(\mathcal{U}_{12} \cap \mathcal{U}_{13}) - \mathrm{card}(\mathcal{U}_{12} \cap \mathcal{U}_{23}) - \\
&\quad \mathrm{card}(\mathcal{U}_{13} \cap \mathcal{U}_{23}) + \mathrm{card}(\mathcal{U}_{12} \cap \mathcal{U}_{13} \cap \mathcal{U}_{23}).
\end{aligned}$$

The formula for the number of independent variables will be simpler if the hierarchical log-linear model is decomposable. A hierarchical model with sufficient configurations C_{θ_i}, $i = 1, \ldots, I$ is decomposable if and only if the class $\{\theta_i\}$ can be ordered in such a way that each θ_i is composed of one set of elements which are missing in all θ_s for $s > i$ and one set ϕ_i which is contained in some θ_τ, for some $\tau > i$ (Haberman (1974, chapter5), Sundberg (1975)). In other words, we have

$$\theta_i = \theta_i^* \cup \phi_i$$

with

$$\theta_i^* \cap \phi_i = \emptyset,\ \theta_i^* \cap \bigcup_{j>i} \theta_j = \emptyset \text{ and } \phi_i \subset \theta_s \text{ for some } s > i. \tag{*}$$

Furthermore, it is a fact that if such an ordering is possible, a version may be found in which any prescribed set is the last one. For example,

(i) The three way contingency table with sufficient configuration C_{12}, C_{13} and C_{23} is not decomposable since the subscripts of any C can not be decomposed into two disjoint subsets satisfying (*).

(ii) The seven dimensional hiearchical log-linear model with sufficient configurations C_{123}, C_{124}, C_{235}, C_{136} and C_{57} is decomposable. An ordering of the θ_i which has $\{5, 7\}$ as the last set is

$$\{1,2,\underline{4}\},\ \{1,3,\underline{6}\},\ \{\underline{1},2,3\},\ \{2,3,\underline{5}\},\ \{5,\underline{7}\}$$

where the underlined elements do not belong to any set that follows. An ordering that has $\{1, 3, 6\}$ as the last set is

$$\{5,\underline{7}\},\ \{2,3,\underline{5}\},\ \{1,2,\underline{4}\},\ \{1,\underline{2},3\},\ \{1,\underline{3},\underline{6}\}$$

Usually, to obtain a particular ordering, it will be easier to start with the last set and work backwards.

Results under multinomial sampling

For completeness, let's first state some basic results about the log-linear model under multinomial sampling. The standard results for these models are given in Bishop et al. (1975) and Fienberg (1980). We will follow closely the notation in Rao and Scott's paper (1984). Let $\underline{\pi} = (\pi_1, \ldots, \pi_T)^T$ be a vector of cell proportions; $\sum_{i=1}^{k} \pi_i = 1$. We observe $\mathbf{n} = (n_1, \ldots, n_T)^T$ the counts in each cell from a random sample, so that $\mathbf{n}$ has a multinomial distribution ($\sum n_i = n$). Let $\mathbf{p} = \mathbf{n}/n$ 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

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

where $\mathbf{X}$ is a known $T \times r$ matrix of full rank r ($\leq T-1$) and $\mathbf{X}'\mathbf{1} = 0$, $\mathbf{1}$ is a T-vector of 1's. If $r = T - 1$, we have a saturated model.

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

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

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

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

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

in distribution.

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

We consider the problem of testing

$$H_0 : \underline{\theta}_2 = \mathbf{0},$$

against the alternative

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

Let $\underline{\hat{\theta}}_1$, $\underline{\hat{\theta}}_2$, $\underline{\hat{\pi}}$, etc. be the maximum likelihood estimates under the full model H_1. Alternatively, let $\underline{\hat{\hat{\theta}}}_1$, $\underline{\hat{\hat{\pi}}}$, denote the estimates under H_0. The likelihood ratio statistic for the above hypothesis is

$$G^2 = 2n \sum \hat{p}_t \log(\hat{p}_t/\hat{\hat{\pi}}_t) - 2n \sum \hat{p}_t \log(\hat{p}_t/\hat{\pi}_t).$$

Under H_0, this statistic has asymptotically a χ^2 distribution with u degrees of freedom. This statistic is also asymptotically equivalent to the Pearson statistic

$$W_p = n(\hat{\pi} - \hat{\hat{\pi}})^T \hat{D}_\pi^{-1} (\hat{\pi} - \hat{\hat{\pi}})$$

and the Wald statistic

$$W_w = n\hat{\hat{\theta}}_\mathbf{2}^T X_2^T \hat{P} X_2 \hat{\hat{\theta}}_2.$$

Results for other sampling schemes

We still assume that the cell proportions, $\underline{\pi}$, satisfy $\underline{\mu} = \log \underline{\pi} = u(\underline{\theta}_1, \underline{\theta}_2)\mathbf{1} + \mathbf{X}_1\underline{\theta}_1 + \mathbf{X}_2\underline{\theta}_2$ but we now have $n^{1/2}(\mathbf{p} - \underline{\pi}) \to N(\mathbf{0}, V)$, where $\mathbf{p}$ is a survey estimate and V is the corresponding covariance matrix of $\mathbf{p}$. Rao and Scott (1984) showed that under general sampling designs, the test statistic X^2 ($=G^2 = W_p = W_w$) has asymptotically the distribution of the sum of weighted independent chi-squared variables with 1 degree of freedom,

$$X^2 \sim \sum_{i=1}^{u} \delta_i W_i$$

where the W_i's are independent χ_1^2 random variables and the δ_i's (all greater than 0) are the eigenvalues of

$$(\tilde{\mathbf{X}}_2^T P \tilde{\mathbf{X}}_2)^{-1}(\tilde{\mathbf{X}}_2^T V \tilde{\mathbf{X}}_2)$$

where

$$\tilde{\mathbf{X}}_2 = (\mathbf{I} - \mathbf{X}_1(\mathbf{X}_1^T P \mathbf{X}_1)^{-1}\mathbf{X}_1^T P)\mathbf{X}_2$$
$$P = D_\pi - \pi\pi^T, \qquad D_\pi = diag(\underline{\pi}).$$

Rao and Scott also showed that under H_0,

$$\begin{aligned}
E(X^2) &= E(G^2) \\
&= E(G_2^2) - E(G_1^2) \\
&= \sum_{i=1}^{u} \delta_i(1) \\
&= u\delta_. \\
&= tr\big((X^T P X)^{-1}(X^T V X)\big) - \\
&\quad\; tr\big((X_1^T P X_1)^{-1}(X_1^T V X - 1)\big)
\end{aligned}$$

(G_1^2 and G_2^2 being the loglikelihood ratio under H_1 and H_0 respectively). Hence, as a first order approximation, $\frac{X^2}{\delta_.}$ can be regarded as a χ^2 with u degrees of freedom.

They also noted that when the models H_0 and H_1 admit explicit solutions for $\underline{\hat{\pi}}$ and $\underline{\hat{\hat{\pi}}}$, we have an alternative method of computing $\delta_.$. Hierarchical log-linear models have closed form expressions for the maximum likelihood estimate for π only for decomposable models (Haberman (1974), Sundberg(1975)).

Let $C_{\theta_1}, \ldots, C_{\theta_I}$ be the sufficient configurations for the model H_1 where the θ_i's are ordered according to the decomposability criterion, $\phi_t = \theta_t \cap (\cup_{s>t}\theta_s)$ then the mle of π_θ under H_1 is

$$\hat{\pi}_\theta = \frac{\prod_{i=1}^{I} \pi_{\theta_i}}{\prod_{j=1}^{I-1} \pi_{\phi_j}} \frac{1}{\prod_{k \in Z - \bigcup_t \theta_t} v_k}$$

where $Z = \{1, 2, \ldots, r\}$ (Haberman (1974) and Sundberg (1975)).

In the above formula, θ denotes an arbitrary cell, and π_{θ_i}, π_{ϕ_j} are marginal totals of π_θ summed over indices not in θ_i, ϕ_j respectively. For example, in a five way table, for $\theta = \{2,4\}$, $\pi_\theta = \sum_{i_1,i_3,i_5} \pi_{i_1 i_2 i_3 i_4 i_5}$

Results when the cell and marginal design effects are equal

Using a Taylor expansion, we have the following approximation for G_1^2,

$$E(G_1^2) = \sum_\theta (1-\pi_\theta)d_\theta - \sum_i \sum_{\theta_i}(1-\pi_{\theta_i})d_{\theta_i} + \sum_j \sum_{\phi_j}(1-\pi_{\phi_j})d_{\phi_j}$$

(Rao and Scott, 1984) where the d_θ's, d_{θ_i}'s, d_{ϕ_j}'s are the cell and marginal design effects. When the cell and design effects are all equal to δ, the above expression reduces to

$$\begin{aligned}
E(G_1^2) &= \delta\Big(\sum_\theta (1-\pi_\theta) - \sum_i \sum_{\theta_i}(1-\pi_{\theta_i}) + \sum_j \sum_{\phi_j}(1-\pi_{\phi_j})\Big) \\
&= \delta\Big((T-1) - \sum_{i=1}^{T}(\#\text{ cells in } C_{\theta_i} - 1) + \\
&= \quad \sum_{j=1}^{I-1}(\#\text{ cells in } C_{\phi_j} - 1)\Big) \\
&= \delta(T - \sum_{i=1}^{T}\#\text{ cells in } C_{\theta_i} + \sum_{j=1}^{I-1}\#\text{ cells in } C_{\phi_j}) \\
&= \delta(T - \#\text{ independent parameters in } H_1)
\end{aligned}$$

Similarly,

$$E(G_2^2) = \delta(T - \#\text{ independent parameters in } H_0).$$

Thus,

$$\begin{aligned}
E(G^2) &= u\delta_. \\
&= E(G_2^2) - E(G_1^2) \\
&= \delta(\#\text{ ind. parameters in } H_1 - \\
&\qquad \#\text{ ind. parameters in } H_0) \\
&= \delta u \\
\text{or} \quad \delta_. &= \delta
\end{aligned}$$

Hence, under H_0, $\frac{X^2}{\delta_.} = \frac{X^2}{\delta}$ has asymptotically a χ^2 distribution with u degrees of freedom, where u is the difference of the number of independent parameters in the 2 models.

References

Birch, N.W. (1963) Maximum likelihood in three-way contingency tables. *Journal of the Royal Statistics Society, Series B*, **25**, 220-233.

Bishop, Y.M., Fienberg, S.E., Holland, P.W. (1975). *Discrete Multivariate Analysis, Theory and Practice*, Cambridge, Massachusetts: MIT Press.

Fienberg, S.E. (1980). *The Analysis of Cross Classified Data*, Cambridge, Massachusetts: MIT Press.

Haberman, S.J. (1974). *The Analysis of Frequency Data*, Chicago, Illinois: University of Chicago Press.

Rao, J.N.K. and Scott, A.J. (1981). The analysis of categorical data from complex sample surveys: chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association*,**76**, 221-230.

Rao, J.N.K. and Scott, A.J. (1984). On Chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *Annals of Statistics*, **12**, 46-60.

Sundberg, R. (1975). Some results about decomposable (or Markov-type) models for multidimensional contingency tables: distribution of marginals and partitioning of tests. *Scandinavian Journal of Statistics*, **2**, 71-79.

Appendix 1

Since $P_0 = \text{diag}(\hat{\mathbf{p}}_0) - \hat{\mathbf{p}}_0\hat{\mathbf{p}}_0'$, the inverse P_0^{-1} is

$$P_0^{-1} = (\text{diag}(\mathbf{p}_0))^{-1} + \frac{1}{p_{0k}}\mathbf{1}\mathbf{1}'$$

Hence,

$$P_0^{-1}V = (\text{diag}(\mathbf{p}_0))^{-1}V + \frac{1}{p_{0k}}\mathbf{1}\mathbf{1}'V$$

$$\begin{aligned}
&\text{trace}(P_0^{-1}V) \\
&= \text{trace}(\text{diag}(\mathbf{p}_0))^{-1}V) + \text{trace}(\frac{1}{p_{0k}}\mathbf{1}\mathbf{1}'V) \\
&= \sum_{i=1}^{k-1}\frac{v_{ii}}{p_{0i}} + \frac{1}{p_{0k}}\sum_{i=1}^{k-1}\sum_{j=1}^{k-1} v_{ij} \\
&= \sum_{i=1}^{k-1}\frac{v_{ii}}{p_{0i}} + \frac{1}{p_{0k}}\text{var}(1 - p_{01} - \ldots - p_{0k-1}) \\
&= \sum_{i=1}^{k-1}\frac{v_{ii}}{p_{0i}} + \frac{1}{p_{0k}}v_{kk} \\
&= \sum_{i=1}^{k}\frac{v_{ii}}{p_{0i}}
\end{aligned}$$