Journal of the Korean Statistical Society (1998), 27: 2, pp 197–203

The Chi-squared Test of Independence for a Multi-way Contingency Table with All Margins Fixed †

Cheolyong Park[1]

ABSTRACT

To test the hypothesis of complete or total independence for a multi-way contingency table, the Pearson chi-squared test statistic is usually employed under Poisson or multinomial models. It is well known that, under the hypothesis, this statistic follows an asymptotic chi-squared distribution. We consider the case where all marginal sums of the contingency table are fixed. Using conditional limit theorems, we show that the chi-squared test statistic has the same limiting distribution for this case.

Keywords: Conditional limit theorems; Complete independence; Multinomial models; Poisson models

1. INTRODUCTION

Consider a $d_1 \times d_2 \times \cdots \times d_r$ multi-way contingency table where the sample size is n and the i-th variable has d_i possible categories. For each cell $\pi = (\pi_1, \pi_2, \ldots, \pi_r)$ where $\pi_j = 1, 2, \ldots, d_j$ for $j = 1, 2, \ldots, r$, the cell frequency belonging to π is denoted by x_π. Let $\mathbf{x}^{(i)} = (x_1^{(i)}, \ldots, x_{d_i}^{(i)})^T$ be the vector of marginal sums of the i-th variable; i.e. $x_j^{(i)}$ is the number of observations that belongs to the j-th category of the i-th variable. For the hypothesis of complete independence, we can use the Pearson chi-squared test statistic

$$X = \sum_{\pi} \frac{(x_\pi - np_{n\pi})^2}{np_{n\pi}},$$

where $p_{n\pi} = \prod_{j=1}^{r} \left(x_{\pi_j}^{(j)}/n \right)$ is an estimator of the cell probability p_π. It is well known that, under the hypothesis of complete independence, the limiting

†The present research has been conducted by the Bisa Research Grant of Keimyung University in 1997.

[1]Department of Statistics, Keimyung University, Taegu, 704-701, Korea

distribution of X is a chi-squared distribution with $\prod_{j=1}^{r} d_j - 1 - \sum_{i=1}^{r}(d_i - 1)$ degrees of freedom.

We consider the case where all marginal sums of the contingency table are fixed. We will show that the limiting distribution of X is the same chi-squared distribution for this case. There has not been much study in this context. Roy and Mitra(1956) considered several configurations of the variables in two or three dimensional contingency tables and heuristically derived the limiting distributions of chi-squared test statistics. Alalouf(1987) used the Central Limit Theorem for finite populations to prove that the result holds for the bivariate case ($r = 2$). Park(1995) provided an intuitively appealing way of proof based on conditional limit theorems to show the same result. He used the fact that the joint distribution of the cell frequencies with both margins fixed is the same as the conditional distribution of a multinomial given both marginal sums equal to the fixed margins. Although Alalouf's proof is self-contained, it does not utilize the above statistical fact and so require a less intuitive mathematical approach.

We can utilize the same fact for multivariate case: The joint distribution of the cell frequencies x_π with all margins $\mathbf{x}^{(i)}$ fixed, given by

$$\left(\frac{n!}{\prod_\pi x_\pi!}\right) \Big/ \left(\prod_{i=1}^{r}\prod_{j=1}^{d_i}\frac{n!}{x_j^{(i)}!}\right), \tag{1.1}$$

is the same as the conditional distribution of the cell counts y_π from a multinomial model, given by

$$\frac{n!}{\prod_\pi y_\pi!}\prod_\pi\left(\prod_{j=1}^{r} p_{\pi_j}^{(j)}\right)^{y_\pi}, \tag{1.2}$$

given that $\mathbf{y}^{(i)} = \mathbf{x}^{(i)}$, where $\mathbf{y}^{(i)}$ is the vector of marginal sums of the i-th variable from (1.2) and $p_j^{(i)}$ is the probability of belonging to the j-th category of the i-th variable. We will use the above statistical fact and the conditional limit theorems in exponential families by Holst(1981) to prove that the result also holds for the multivariate case ($r \geq 3$): While deriving the result we will use the multivariate conditional limit theorem in Park(1995), which is an extension of a univariate conditional limit theorem in Holst(1981).

In Section 2, we will derive the asymptotic joint distribution of the cell frequencies and the limiting distribution of X using the conditional limit theorems. In Section 3, we will discuss some points in the proofs of the main results in

Section 2 and a possible application to the test of complete independence for a continuous multivariate distribution.

2. MAIN RESULTS

In this section, we will derive the asymptotic joint distribution of the cell frequencies and the limiting distribution of the chi-squared test statistic. Before deriving the distributions, we will define some notations which are needed for the presentation of main results.

For easy representation of results we assume the vector $\mathbf{x} = (x_\pi)$ of the cell frequencies is arranged in a standard order; the index π_1 of the first variable changes from 1 to d_1 slowest and π_2 of the second variable changes from 1 to d_2 second slowest, and so forth. Also, for the convenience of notation, we define

$$\mathbf{p}_n = (p_{n\pi}) = \mathbf{p}_n^{(1)} \otimes \mathbf{p}_n^{(2)} \otimes \cdots \otimes \mathbf{p}_n^{(r)},$$

where $\mathbf{p}_n^{(i)} = \mathbf{x}^{(i)}/n$ and $\otimes$ is the direct or Kronecker product (see p.265 of Searle(1982) for details). For a given vector $\mathbf{y} = (y_1, y_2, \ldots, y_m)^T$, we define the diagonal matrix $D(\mathbf{y})$ and the vector of square root values $\sqrt{\mathbf{y}}$ to be

$$D(\mathbf{y}) = \mathrm{diag}(y_1, y_2, \ldots, y_m), \ \sqrt{\mathbf{y}} = (\sqrt{y_1}, \sqrt{y_2}, \ldots, \sqrt{y_m})^T.$$

Finally, we will define a 'd'(deleted) notation such that $\mathbf{y}_d = (y_1, y_2, \ldots, y_{m-1})^T$ for a given vector $\mathbf{y} = (y_1, y_2, \ldots, y_m)^T$; i.e. $\mathbf{y}_d$ is the vector obtained by deleting last element of $\mathbf{y}$.

Theorem 2.1. *Suppose that* $\mathbf{x}$ *is drawn from the distribution in* (1.1). *Let* $\mathbf{p}_n^{(i)} \to \mathbf{q}^{(i)}$ *as* $n \to \infty$ *for all* $i = 1, 2, \ldots, r$, *where* $\mathbf{q}^{(i)} = (q_1^{(i)}, \ldots, q_{d_i}^{(i)})^T$ *satisfies* $0 < q_j^{(i)} < 1$ *for all* $j = 1, 2, \ldots, d_i$ *and* $\sum_j q_j^{(i)} = 1$. *Let* $\mathbf{q} = (q_\pi) = \mathbf{q}^{(1)} \otimes \mathbf{q}^{(2)} \otimes \cdots \otimes \mathbf{q}^{(r)}$ *so that* $\mathbf{p}_n \to \mathbf{q}$ *as* $n \to \infty$. *Then the limiting joint distribution of the cell frequencies is*

$$\{D(n\mathbf{p}_n)\}^{-1/2}(\mathbf{x} - n\mathbf{p}_n) \xrightarrow{\mathcal{L}} N(0, A^*),$$

where

$$A^* = I + (r-1)\sqrt{\mathbf{q}}\sqrt{\mathbf{q}}^T - \sum_{i=1}^{r} E^{(i)}\{E^{(i)}\}^T$$

and

$$E^{(i)} = \sqrt{\mathbf{q}^{(1)}} \otimes \cdots \otimes I_{d_i} \otimes \cdots \otimes \sqrt{\mathbf{q}^{(r)}}$$

with I_{d_i} in the i-th position of the Kronecker products for each $i = 1, 2, \ldots, r$.

Proof: From the statistical fact in Section 1, we observe that

$$\mathcal{L}(\mathbf{x}) = \mathcal{L}\left(\mathbf{y} \,|\, \mathbf{y}^{(i)}/n = \mathbf{p}_n^{(i)}, \text{ for } i = 1, 2, \ldots, r\right),$$

where $\mathbf{y}$ is drawn from the distribution in (1.2). Now we will verify the assumptions A1-A3, A4′, A5′, and A6 of Corollary 1 in Park(1995). Assumption A1 requires that (1.2) is a regular exponential family, which is satisfied with the canonical sufficient statistics equal to $\mathbf{y}_d^{(i)}, i = 1, 2, \ldots, r$. Assumption A2 requires that θ_n need to be the maximum likelihood estimator of θ, a natural parameter for the distribution (1.2). This assumption is satisfied when s_n, the vector of fixed values the canonical sufficient statistics are assumed to take, is the vector of $(\mathbf{p}_n^{(1)})_d$, $(\mathbf{p}_n^{(2)})_d$, $\ldots$, $(\mathbf{p}_n^{(r)})_d$ combined and θ_n is the natural parameter corresponding to $\mathbf{p}^{(i)} = \mathbf{p}_n^{(i)}$ for all i, where $(\mathbf{p}_n^{(i)})_d$ is the vector obtained by deleting the last element of $\mathbf{p}_n^{(i)}$, and $\mathbf{p}^{(i)} = (p_1^{(i)}, \ldots, p_{d_i}^{(i)})^T$ is the vector of marginal probabilities for the i-th variable $\mathbf{y}^{(i)}$. Assumptions A3 requires that the covariance matrix of the vector of $\mathbf{y}_d^{(i)}, i = 1, \ldots r$ combined is positive definite, Assumption A4′ requires that the variance of any linear combination of $\mathbf{y}_d^{(i)}, i = 1, \ldots, r$ is finite when $n = 1$, and A5′ is needed for continuous distributions, all of which are trivially satisfied. Assumption A6 requires that θ_n converges to a number θ_0, which is satisfied when θ_0 is the natural parameter corresponding to $\mathbf{p}^{(i)} = \mathbf{q}^{(i)}$ for all i.

Next define $\mathbf{z} = \{D(n\mathbf{q})\}^{-1/2}\mathbf{y}$ and $\mathbf{z}^{(i)} = \{D(\mathbf{q}_d^{(i)})\}^{-1/2}\mathbf{y}_d^{(i)}$ for each i. Then, by the corollary, we have

$$\{D(n\mathbf{p}_n)\}^{-1/2}(\mathbf{x} - n\mathbf{p}_n) \xrightarrow{\mathcal{L}} N(0, A - BC^{-1}B^T),$$

where

$$A = \mathrm{Cov}_{\theta_0}(\mathbf{z}) \;, B = \mathrm{Cov}_{\theta_0}(\mathbf{z}, \mathbf{w}) \;, C = \mathrm{Cov}_{\theta_0}(\mathbf{w}),$$

with $\mathbf{w}$ the vector of $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(r)}$ combined. It is well known that

$$A = I - \sqrt{\mathbf{q}}\sqrt{\mathbf{q}}^T,$$

$$C = \mathrm{diag}(I - \sqrt{\mathbf{q}_d^{(1)}}\sqrt{\mathbf{q}_d^{(1)}}^T, \ldots, I - \sqrt{\mathbf{q}_d^{(r)}}\sqrt{\mathbf{q}_d^{(r)}}^T),$$

and $B = (B_1, B_2, \ldots, B_r)$ with $B_i = E_d^{(i)} - \sqrt{\mathbf{q}}\sqrt{\mathbf{q}_d^{(i)}}^T$ for each i, where $E_d^{(i)}$ is the matrix formed by deleting the last column of $E^{(i)}$. Since

$$(I - \sqrt{\mathbf{q}_d^{(i)}}\sqrt{\mathbf{q}_d^{(i)}}^T)^{-1} = I + \sqrt{\mathbf{q}_d^{(i)}}\sqrt{\mathbf{q}_d^{(i)}}^T / q_{d_i}^{(i)}$$

for each i, we have

$$BC^{-1}B^T = \sum_{i=1}^{r} B_i \left(I + \sqrt{\mathbf{q}_d^{(i)}}\sqrt{\mathbf{q}_d^{(i)}}^T / q_{d_i}^{(i)}\right) B_i^T.$$

Let $\mathbf{e}_i$ be the last column vector of $E^{(i)}$ for each i. Then it is easy to show that

$$E^{(i)}\{E^{(i)}\}^T = E_d^{(i)}\{E_d^{(i)}\}^T + \mathbf{e}_i\mathbf{e}_i^T$$

and

$$\sqrt{\mathbf{q}} = E^{(i)}\sqrt{\mathbf{q}^{(i)}} = E_d^{(i)}\sqrt{\mathbf{q}_d^{(i)}} + \sqrt{q_{d_i}^{(i)}}\,\mathbf{e}_i$$

for each i.

Combining these relations and doing some simple calculation, we can show that

$$B_i \left(I + \sqrt{\mathbf{q}_d^{(i)}}\sqrt{\mathbf{q}_d^{(i)}}^T / q_{d_i}^{(i)}\right) B_i^T = E^{(i)}\{E^{(i)}\}^T - \sqrt{\mathbf{q}}\sqrt{\mathbf{q}}^T, \tag{2.1}$$

which completes the proof. □

Corollary 2.1. *Under the assumptions of Theorem 2.1, the limiting distribution of* $X = (\mathbf{x} - n\mathbf{p}_n)^T\{D(n\mathbf{p}_n)\}^{-1}(\mathbf{x} - n\mathbf{p}_n)$ *is a chi-squared distribution with* $\prod_{j=1}^{r} d_j - 1 - \sum_{i=1}^{r}(d_i - 1)$ *degrees of freedom.*

Proof: It is sufficient to show that A in Theorem 2.1 is an idempotent matrix of rank $\prod_{j=1}^{r} d_j - 1 - \sum_{i=1}^{r}(d_i - 1)$. Since $\{E^{(i)}\}^T E^{(i)} = I_{d_i}$ and $\{E^{(i)}\}^T\sqrt{\mathbf{q}} = \sqrt{\mathbf{q}^{(i)}}$ for each i, and since $\{E^{(i)}\}^T E^{(j)} = \sqrt{\mathbf{q}^{(i)}}\sqrt{\mathbf{q}^{(j)}}^T$ for $i \neq j$, it is easily verified that $\sqrt{\mathbf{q}}\sqrt{\mathbf{q}}^T$ and $E^{(i)}\{E^{(i)}\}^T - \sqrt{\mathbf{q}}\sqrt{\mathbf{q}}^T$ $(i = 1, 2, \ldots, r)$ are mutually orthogonal idempotent matrices of ranks 1 and $d_i (i = 1, 2, \ldots, r)$, respectively. From these results, it is easy to show that A is an idempotent matrix and that it is of rank $\mathrm{tr}(A) = \prod_{j=1}^{r} d_j - 1 - \sum_{i=1}^{r}(d_i - 1)$. □

3. REMARKS

Firstly, we will interpret the components of the asymptotic covariance matrix A^* and then show a heuristic derivation of A^*. A projection matrix $\sqrt{\mathbf{q}}\sqrt{\mathbf{q}}^T$ is the component corresponding to the loss of degrees of freedom due to fixing the sample size n and a projection matrix $E^{(i)}\{E^{(i)}\}^T - \sqrt{\mathbf{q}}\sqrt{\mathbf{q}}^T$ is the component due to fixing the margins of the i-th variable. Also, we can heuristically derive

the result in Theorem 2.1 easily by using generalized inverses (g-inverses). By using the (Moore) g-inverse, it is easy to show that

$$\begin{aligned} & (E^{(i)} - \sqrt{\mathbf{q}}\sqrt{\mathbf{q}^{(i)}}^T)\,(I - \sqrt{\mathbf{q}^{(i)}}\sqrt{\mathbf{q}^{(i)}}^T)^-\,(E^{(i)} - \sqrt{\mathbf{q}}\sqrt{\mathbf{q}^{(i)}}^T)^T \\ = \;& E^{(i)}\{E^{(i)}\}^T - \sqrt{\mathbf{q}}\sqrt{\mathbf{q}}^T \end{aligned} \tag{3.1}$$

since $(I - \sqrt{\mathbf{q}^{(i)}}\sqrt{\mathbf{q}^{(i)}}^T)^- = I - \sqrt{\mathbf{q}^{(i)}}\sqrt{\mathbf{q}^{(i)}}^T$ for each i. Also, by using a reflexive g-inverse, the equivalence between (2.1) and (3.1) is directly established since

$$(I - \sqrt{\mathbf{q}^{(i)}}\sqrt{\mathbf{q}^{(i)}}^T)^- = \begin{pmatrix} (I - \sqrt{\mathbf{q}_d^{(i)}}\sqrt{\mathbf{q}_d^{(i)}}^T)^{-1} & 0 \\ 0 & 0 \end{pmatrix}.$$

Secondly, the result can be used for testing complete or total independence of variables from a continuous multivariate distribution. By discretizing the i-th continuous variable into a categorical variable with d_i categories of (at least approximately) *equal size* for each i, we can test the hypothesis of complete independence by the chi-squared test. Even though there has been lots of study on nonparametric tests of bivariate independence, very little has been written on testing multivariate independence. Some tests are actually testing pairwise independence (see Blum, Kieffer, and Rosenblatt(1961) and Puri, Sen, and Gokhale(1970) among others) and some tests are not easy to use in practice since test statistics or their limiting distributions are not easy to compute. The chi-squared test statistic is easy to compute, nonparametric, and has a well-known limiting distribution. Moreover, it does not focus on pairwise independence but on multivariate independence involving three or more variables.

REFERENCES

Alalouf, I.S. (1987). "The Chi-Squared Test with Both Margins Fixed," *Communications in Statistics - Theory and Methods* **16**, 29-43.

Blum, J.R., Kieffer, J., and Rosenblatt, M. (1961). "Distribution Free Tests of Independence based on the Sample Distribution Function," *Annals of Mathematical Statistics* **32**, 485-98.

Holst, L. (1981). "Some Conditional Limit Theorems in Exponential Families," *The Annals of Probability* **9**, 818-30.

Park, C. (1995). "Some Remarks on the Chi-Squared Test with Both Margins Fixed," *Communications in Statistics - Theory and Methods* **24**, 653-61.

Puri, M.L., Sen, P.K., and Gokhale, D.V. (1970). "On a Class of Rank Order Tests for Independence in Multivariate Distributions," *Sankhyā Series A* **32**, 271-98.

Roy, S.N. and Mitra, S.K. (1956). "An Introduction to Some Nonparametric Generalizations of Analysis of Variance and Multivariate Analysis," *Biometrika* **43**, 361-76.

Searle, S.R. (1982). *Matrix Algebra Useful for Statistics,* John Wiley & Sons, Inc., New York.