Sains Malaysiana 45(10)(2016): 1573–1578

Power Divergence Statistics under Quasi Independence Model for Square Contingency Tables

(Statistik Pencapahan Kuasa Model Kuasi Ketakbersandaraan untuk Jadual Kontingensi Segi Empat Sama)

SERPIL AKTAŞ*

ABSTRACT

In incomplete contingency tables, some cells may contain structural zeros. The quasi-independence model, which is a generalization of the independence model, is most commonly model used to analyze incomplete contingency tables. Goodness of fit tests of the quasi-independence model are usually based on Pearson chi square test statistic and likelihood ratio test statistic. In power divergence statistics family, the selection of power divergence parameter is of interest in multivariate discrete data. In this study, a simulation study is conducted to evaluate the performance of the power divergence statistics under quasi independence model for particular power divergence parameters in terms of power values.

Keywords: Power divergence family; square contingency tables; structural zeros

ABSTRAK

Dalam jadual kontingensi tidak lengkap, sesetengah sel boleh mengandungi struktur sifar. Model kuasi ketakbersandaran yang merupakan suatu generalisasi daripada model ketakbersandaran adalah model yang paling biasa digunakan untuk menganalisis jadual kontingensi yang tidak lengkap. Ujian kebagusan penyuaian model kuasi ketakbersandaran biasanya berdasarkan statistik ujian khi kuasa dua Pearson dan ujian statistik nisbah kebolehjadian. Dalam keluarga statistik pencapahan kuasa, pemilihan parameter pencapahan kuasa adalah penting dalam data diskret multivariat. Dalam penyelidikan ini, suatu kajian simulasi dijalankan untuk menilai prestasi statistik pencapahan kuasa di bawah parameter model kuasi ketakbersandaran untuk parameter pencapahan kuasa daripada segi nilai kuasa tertentu.

Kata kunci: Jadual kontinjensi segi empat sama; kuasa keluarga pencapahan; struktur sifar

INTRODUCTION

For the contingency tables, we are interested in whether the variables are independent of one another. For incomplete contingency tables, *structural zeros* occur where the cells are theoretically impossible to observe a value. A cell with a structural zero has an expected value of zero. Therefore, they do not contribute to the likelihood function or model fitting. A contingency table containing structural zeros is referred as an incomplete table. Therefore, the usual chi-square tests cannot be applied directly (Bishop et al. 1975; Fienberg 1980; Haberman 1979). Quasi independence (QI) model gives better fit than ordinary independence model. Goodness-of-fit tests summarize the discrepancy between the observed values and the expected values under the model. The primary problem is specification of the most suitable test statistic when implementing the test. Studies indicate that none of the test statistics has a clear advantage over any others. The results also suggested that none of the test statistics completely dominate the other and that the choice of which test to use depends on the nature of the alternative hypothesis. Although all the members of the family of power-divergence statistics converge asymptotically to a chi-square distribution, their small-sample accuracy is not guaranteed. Read and Cressie (1988) presented an approach to goodness-of-fit testing in multinomial models through the family of power divergences. Cressie and Read (1984) suggested an alternative to the Pearson-based and the likelihood ratio-based test statistics, in terms of both exact and asymptotic size and power. Despite the broad family of power-divergence statistics, the likelihood ratio statistic is treated as if the only alternative to Pearson's X^2 statistic for testing independence or homogeneity in analysis of contingency table (García-Pérez & Núñez-Antón 2009).

This paper aimed to compare the power divergence statistics for various λ values with respect to their power values under the QI model in square contingency tables where observations are cross-classified by two variables with the same categories. In square contingency tables, on diagonal cells often lack of fit of the independence model. The hypothesis might be whether the rest of the table satisfies the independence model off diagonal cells. This leads to the quasi-independence model which fits much better than independence. A square table satisfies quasi-independence if the row and and column variables are independent of each other in off-diagonal cells. We herein only concentrate on the QI model. We will illustrate the results of comparative analysis of the power of goodness-of-fit tests on a simulation study. An illustrative example of unaided vision data set is analyzed as well.

POWER DIVERGENCE FAMILY

Cressie and Read (1984) developed a class of goodness-of-fit test statistics referred the family of power divergence statistics. All members of the statistics are particular members of the power divergence family. The divergences within this family are indexed by a single parameter $\lambda \in \mathbb{R}$. Let X_i denote a random vector of counts having multinomial distribution; $\hat{p}$ the maximum likelihood estimator (MLE) of p under H_0 ; n is the total number of observation.

For various λ values, power divergence statistics are defined as:

For $\lambda = 1$, Pearson's X^2 :

$$\mathrm{PD}(1) = \sum_{i=1}^{k} \frac{\left(x_i - n\hat{p}_i\right)^2}{n\hat{p}_i}, \tag{1}$$

for $\lambda = 0$, Likelihood ratio, G^2 :

$$\mathrm{PD}(0) = 2\sum_{i=1}^{k} X_i \log \frac{x_i}{n\hat{p}_i}, \tag{2}$$

for $\lambda = -1/2$, Freeman-Tukey's F^2 :

$$\mathrm{PD}(-1/2) = 4\sum_{i=1}^{k} \left(\sqrt{X_i} - \sqrt{n\hat{p}_i}\right)^2, \tag{3}$$

for $\lambda = -1$, Neyman's modified X^2 :

$$\mathrm{PD}(-1) = 2\sum_{i=1}^{k} n\hat{p}_i \log \frac{n\hat{p}_i}{x_i}, \tag{4}$$

for $\lambda = -2$, modified G^2:

$$\mathrm{PD}(-2) = \sum_{i=1}^{k} \frac{\left(n\hat{p}_i - x_i\right)^2}{x_i}. \tag{5}$$

Lawal (1993) proposed a version of test:

$$T^2 = \sum_{i=1}^{k} \frac{\sum_i \left[\left(X_i - 1/2\right) - n\hat{p}_i\right]^2}{n\hat{p}_i}, \tag{6}$$

Zelterman (1987) proposed the test statistics below,

$$T^2 = D^2 + k + \frac{1}{4}\sum_i \frac{1}{n\hat{p}_i}, \tag{7}$$

where,

$$D^2 = \sum_{i=1}^{k} \frac{\sum_i \left[\left(X_i - n\hat{p}_i\right) - x_i\right]^2}{n\hat{p}_i}.$$

Cressie and Read (1984) claimed that $\lambda = 2/3$ is a very good choice between X^2 ($\lambda = 1$) and G^2 ($\lambda = 0$) for testing whether the observed multinomial variables are sufficiently close to their null expected values.

$$\mathrm{PD}(\lambda) = 2N \tfrac{9}{10} \sum_{i=1}^{k} x_i \left[\left(\frac{x_i}{n\hat{p}_i}\right)^{2/3} - 1\right]. \tag{8}$$

All statistics given above are distributed as chi-square distribution with k-1 degrees of freedom under H_0.

Assume that number of cells fixed, the multinomial probabilities π_{0i} are completely specified ($\lambda \neq 0, -1$).

$$2nI^{\lambda}\left(\tfrac{X}{n};\Pi_0\right) = \frac{2n}{\lambda(\lambda+1)} \sum_{i=1}^{k} \pi_{0i} \left\{\left(1 + \frac{X_i - n\pi_{0i}}{n\pi_{0i}}\right)^{\lambda+1} - 1\right\}.$$

Writing $V_i = \dfrac{\left(X_i - n\pi_{oi}\right)}{n\pi_{oi}}$ and expanding in a Taylor series,

$$\frac{2n}{\lambda(\lambda+1)} \sum_{i=1}^{k} \pi_{0i} \left\{(\lambda+1)V_i + \frac{\lambda(\lambda+1)}{2} V_i^2 + 0_p\left(V_i^3\right)\right\} = 2n \left\{\sum_{i=1}^{k} \frac{\pi_{0i} V_i^2}{2} + 0_p(1/n)\right\}.$$

Under the model H_o. A similar results hold for λ=0, -1 by a Taylor series expansion.

$$2nI^{\lambda}\left(\tfrac{X}{n};\Pi_0\right) = 2nI^{1}\left(\tfrac{X}{n};\Pi_0\right) + 0_p(1);\ \lambda \in \mathbb{R}.$$

Thus each power divergence statistic has asymptotically has the Pearson chi-squared distribution with k-1 degrees of freedom under the null hypothesis (Cressie & Read 1984). Although all the members of the family of power-divergence statistics converge asymptotically to a chi-square distribution but, their small-sample accuracy has not been guaranteed in the studies.

PENALIZED POWER DIVERGENCE FAMILY

In addition to the various models referred to previously, Basu and Basu (1998) considered the empty cell penalty for the family of power-divergence measures.

This penalty yields a weight on the empty cells in contingency tables by the family of power-divergence measures,

$$I_h^{\lambda}\left(X,\pi_0\right) = \sum x_i > 0 \left[\frac{X_i}{\lambda(\lambda+1)}\left\{\left(\frac{X_i}{n\hat{p}_i}\right)^{\lambda} - 1\right\} + \frac{1}{\lambda+1}\left(n\hat{p}_i - X_i\right)\right] + \frac{1}{1+\lambda}\sum_{X_i=0}^{k} n\hat{p}_i, \tag{9}$$

$$I_h^{\lambda}\left(X,\pi_0\right) = \sum x_i > 0\left[\frac{X_i}{\lambda(\lambda+1)}\left\{\left(\frac{X_i}{n\hat{p}_i}\right)^{\lambda} - 1\right\} + \frac{1}{\lambda+1}\left(n\hat{p}_i - X_i\right)\right]$$
$$+h\sum_{X_i=0}^{k} n\hat{p}_i, \tag{10}$$

where h represents the penalty weight.

The left side of the equations corresponds to the nonempty cells.

$2I_h^l\left(X,\pi_0\right)$ has an asymptotic chi-squared distribution with k-1 degrees of freedom under the null hypothesis.

Testing Independence for Ordinal Data

When the variables of interest are ordinal in nature, then we would better to reflect this ordinal nature of the variable in appropriate modeling techniques (Lawal 2003). Pearson and Likelihood chi-squares do not take into account any ordering in the classification (Bishop et al. 1975). If the ordering of the rows or columns in the table is interchanged, the value of the statistics is expected not to change. When rows and/or columns of classification are ordered, more powerful tests exist and more information can be attained from the data structure using these tests. Mantel-Haenszel chi-square is one of the alternatives for analyzing ordered categorical data.

MANTEL-HAENSZEL CHI-SQUARE

Suppose a RxR contingency table having ordinal row and column variables. We test the null hypothesis of linear association such as,

H_0: No linear association.

Mantel-Haenszel chi-square, also called the *Mantel-Haenszel test for linear association* or *linear by linear association chi-square*, unlike ordinary and likelihood ratio chi-square, is an ordinal measure of significance. Under the Ho is true, M^2 has approximately chi-square distribution with one degree of freedom (Agresti 2002).

Mantel-Haenszel chi-square is defined by:

$$M^2 = \frac{\left[LL - E(LL)\right]^2}{Var(LL)}, \tag{11}$$

where LL, $\Sigma_i \Sigma_j u_i v_j n_{ij}$. Expected value and variance of LL can be calculated as:

$$E(LL) = \frac{\left(\sum_i u_i n_i\right)\left(\sum_j v_j n_j\right)}{n_{..}} \tag{12}$$

$$Var(LL) = \frac{1}{n_{..}-1}\left[\sum_i u_i^2 n_{i.} - \frac{\left(\sum_i u_i n_{i.}\right)^2}{n_{..}}\right]\left[\sum_i v_i^2 n_{.j} - \frac{\left(\sum_i v_i n_{.j}\right)^2}{n_{..}}\right]. \tag{13}$$

When dealing with ordinal data and when there is a positive or negative linear association between variables, M^2 has power advantage over X^2 and G^2 can be defined as,

$$M^2 = (n-1)r^2.$$

The formula for the linear-by-linear association involves the Pearson product-moment correlation coefficient, r and the total number of cases, n.

$$r = \frac{\sum_i \sum_j (u_i - \overline{u})(v_j - \overline{v}) n_{ij}}{\sqrt{\left[\sum_i \sum_j (u_i - \overline{u})^2 n_{ij}\right]\left[\sum_i \sum_j (v_j - \overline{v})^2 n_{ij}\right]}}. \tag{14}$$

$M = \sqrt{(n-1)}\,r$ is approximately distributed as N(0,1).

Quasi-Independence Model

The QI model is a generalization of the model of independence for a two-dimensional contingency table (Goodman 1968). It states that a particular subset of cells satisfies the independence structure. It is often used for square tables in which the cells are not falling on the main diagonal (Goodman 1979; Ireland et al. 1965; McCullagh 1978). The QI model is represented as log-linear form as,

$$\log m_{ij} = \mu + \lambda_i^X + \lambda_j^Y + \delta\, \mathrm{I}(i = j). \tag{15}$$

$$I(i=j) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}.$$

The likelihood equations,

$$\hat{m}_{i.} = n_{i.}$$
$$\hat{m}_{.j} = n_{.j}$$
$$\hat{m}_{ii} = n_{ii} \qquad \text{i,j=1,…,R.}$$

π_{ij} denoting the probability that an observation falls in the ith row and jth column of the table. QI hypothesis is written as,

$$H_0: \pi_{ij} = \alpha_i \beta_j. \tag{16}$$

As a discrepancy measure from the hypothesis of quasi-independence, we use the likelihood ratio statistic,

$$G^2 = 2\sum_i \sum_j n_{ij} \log \frac{n_{ij}}{\hat{m}_{ij}}, \tag{17}$$

where $\hat{m}_{ij}$ is the MLE of the expected parameter m_{ij}.

Degrees of freedom for testing the null hypothesis of QI model would be (R-1)(R-1) - # of strucural zeros (Goodman 1968). We shall employ QI model for analysis of square contingency tables which contain structural zeros on the main diagonal.

Simulation Study

Random samples were drawn under bivariate normal distribution, then the random numbers were transformed into contingency tables with equal-interval frequency tables. The normal distribution parameters are assigned as: $\mu_1=\mu_2=25$; $\sigma_1 = \sigma_2 = 1$; $\rho = 0.90$. Sample sizes are set to N=100 ; 500 ; 1000. Dimensions of tables are assigned as R=3, 5 and 8. For each table, power divergence statistics, penalized divergence statistics and M^2 were calculated. Power-divergence goodness-of-fit statistics were compared in terms of power under different index parameters under QI model. The power of a test which is defined as 'the probability of rejecting the null hypothesis given that the alternative hypothesis is true' is calculated. Note that the entries on the main diagonal are considered as the structural zeros.

An example of the simulated tables of size 100 and $\rho\neq0$ is given in Table 1.

TABLE 1. An 5×5 simulated table

X/Y	1	2	3	4	5
1	26	19	1	0	7
2	2	11	5	3	4
3	0	1	6	6	0
4	0	0	0	4	1
5	1	1	0	0	2

Power divergence statistics for Table 1 under QI model are displayed in Table 2. Except Neyman's modified chi-square, all statistics are statistically significant. One would expect to reject the null hypothesis in accordance with the high correlation structure.

TABLE 2. Power divergence statistics under the QI model

The power divergence statistics	df	Values
Pearson's X^2	11	45.51*
Likelihood ratio, G^2	11	37.18*
Freeman-Tukey's F^2	11	48.53*
Neyman's modified X^2	11	4.47
Modified G^2	11	20.76*
Penalized Power Divergence	11	48.74*
Zelterman's T^2	11	55.20*
Lawal's T^2	11	52.92*
M^2	1	20.28*

*$p<0.05$

Power of test results under QI model for n=100 in Table 3 shows that, the power is relatively highest for Penalized divergence statistics for $\lambda=1$, $\lambda=2$ and M^2. This result also holds for all sample sizes.

TABLE 3. Power of tests under QI model, n=100

	Dimension		
Power divergence statistics	3	5	8
λ=1	0.616699	0.693148	0.590650
λ=0	0.611792	0.649621	0.541102
λ=1a	0.553765	0.540096	0.535649
λ=1b	0.574813	0.582186	0.550867
λ=-1/2	0.369101	0.308677	0.375487
λ=-1	0.633919	0.670489	0.538411
λ=-2	0.666222	0.654866	0.607676
λ=2/3	0.549850	0.548449	0.544504
λ=1	0.749237	0.751568	0.64968
λ =2	0.727064	0.717762	0.63107
Mantel-Haenszel test			
M^2	0.725160	0.725440	0.684419

Mantel-Haenszel test for linear association or *linear by linear association* is a good maximization of power.

Power of tests under QI model of size 500 is displayed in Table 4. Pearson and Penalized Divergence Statistics test yield usually higher power for $\alpha = 0.5$.

Mantel Haenzsel statistic gives higher power as Penalized Divergence Statistics.

TABLE 4. Power of tests under QI model, n=500

	Dimension		
Power divergence statistics	3	5	8
λ =1	0.756322	0.792848	0.729496
λ =0	0.630843	0.743532	0.690375
λ =-1/2	0.352873	0.385308	0.375381
λ =1a	0.480257	0.633615	0.623288
λ =1b	0.328909	0.617782	0.55018
λ =-1	0.683263	0.657375	0.519137
λ =-2	0.712969	0.720398	0.724523
λ =2/3	0.649689	0.663829	0.671563
Penalized divergence dtatistics			
λ =1	0.751943	0.720232	0.741675
λ =2	0.750172	0.709503	0.712008
Mantel-Haenszel test			
M^2	0.762975	0.749909	0.7589065

TABLE 5. Power of tests under QI model, n=1000

	Dimension		
Power divergence statistics	3	5	8
λ =1	0.748561	0.808799	0.813393
λ =0	0.728113	0.754047	0.723373
λ =1a	0.643914	0.689401	0.622276
λ =1b	0.640287	0.613645	0.616447
λ =-1/2	0.433711	0.416161	0.454684
λ =-1	0.639581	0.660814	0.608923
λ =-2	0.647336	0.658839	0.654328
λ =2/3	0.705241	0.714728	0.72907
Penalized divergence statistics			
λ =1	0.826069	0.818051	0.815517
λ =2	0.803843	0.805907	0.891738
Mantel-Haenszel test			
M^2	0.811049	0.860078	0.825683

Power of test for penalized divergence statistics is a decreasing function of λ which confirms the results of this study. When the sample size n=1000, power is the highest (Table 5).

NUMERICAL EXAMPLE

Data in Table 6 reports the unaided vision of 4746 students aged 18 to about 25 including about 10% woman in Faculty of Science and Technology, Science University of Tokyo in Japan examined in April 1982 (Tomizawa 1985).

We fit QI model to data and calculated Power Divergence Statistics for various λ values (Table 7). The results suggested that QI model would seem not adequate to represent the data.

Note that degrees of freedom for independence model is (R-1)(R-1)=3×3=9. Therefore, degrees of freedom for quasi-independence model would be (R-1)(R-1)- # of structural zeros=3×3-4=5.

TABLE 6. Unaided vision data of 4746 students in Japan

	Left eye grade				
Right eye grade	Lowest (1)	Second (2)	Third (3)	Highest (4)	Total
Lowest (1)	1429	249	25	20	1723
	1429	(151.19)	(81.56)	(61.25)	
Second (2)	185	660	124	64	1033
	(118.82)	660	(145.16)	(109.01)	
Third (3)	23	114	221	149	507
	(68.38)	(154.87)	(221)	(62.74)	
Highest (4)	22	40	130	1291	1483
	(42.79)	(96.92)	(52.28)	(1291)	
Total	1659	1063	500	1524	4746

TABLE 7. Power divergence statistics under various λ values

The power divergence statistics	df	Values	The power divergence statistics	df	Values
λ=1	5	507.371	λ=5	5	1817.660
λ=0	5	475.646	λ=-4	5	1361.715
λ=-1/2	5	481.988	λ=-1/3	5	479.622
λ=-1	5	504.456	M^2	1	3602.305
λ=-2	5	611.654	λ=1a	5	507.270
λ=2/3	5	490.269	λ=1b	5	507.270
λ=-3/2	5	545.973			
λ=4	5	1159.376			

Conclusion

The approximation to chi-squared distribution of X^2 and G^2 can be poor for sparse tables. For smaller n and sparser tables, it is advisable to use X^2 rather than G^2. G^2 is usually poorly approximated by the chi-squared distribution when $n/N < 5$ (Agresti 2002). The P-values for G^2 may be too large or too small. Asymptotic results may not apply in small-sample situations and the exact significance of a goodness-of-fit statistic may potentially be stated (Agresti 2002).

Several correction terms have been proposed to improve the accuracy of the asymptotic distribution by Cressie and Read (1984). Among the power divergence family members, the accuracy of the asymptotic distribution seems to be optimal for Pearson's X^2 statistic based on simulation results.

When the statistics corresponding to $\lambda = 2$ or 3 are penalized with a large penalty weight, but there is a substantial gain in power large positive values of λ lead to high power. Our results showed that penalization improves the power properties of ordinary power-divergence test statistics. In this case $\lambda = 1$ (Pearson's X^2) results in optimal efficiency and large values of λ perform poorly in comparison.

The simulated power of the Freeman-Tukey test statistic is generally shown to be relatively less than the power of all the other investigated test statistics. There is generally no improvement in the simulated power for the Power Divergence test statistic with $\lambda={}^2/_3$ over the alternatives.

M^2 is more powerful and tends to be about the same size as G^2 and X^2 but only has df=1 rather than df=(R-1) (R-1). Sampling distribution of the test statistics for small sample sizes, are better approximated for those with smaller df.

M^2 detects a specific type of association and can summarize it in terms of $df = 1$ parameter. M^2 is more powerful because it approximately has the same value as X^2 and G^2 but with only $df = 1$ rather than $(R - 1)(R - 1)$, thus smaller p-value. For small to moderate sample size, the sampling distribution of M^2 are better approximated than for X^2 and G^2; this in general holds for distributions with smaller df's. However, it is difficult to make general recommendations as to the most powerful goodness-of-fit test statistic for the specific alternative distributions used in this study.

References

Agresti, A. 2002. *Categorical Data Analysis*. New York: John Wiley & Sons.

Basu, A. & Basu, S. 1998. Penalized minimum disparity methods for multinomial models. *Stat. Sin.* 8: 841-860.

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

Cressie, N. & Read, T.R.C. 1984. Multinomial goodness-of-fit tests. *Journal of the Royal Statistic Society, Series B* 46: 440-464.

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

García-Pérez, M.A. & Núñez-Antón, V. 2009. Accuracy of power-divergence statistics for testing independence and homogeneity in two-way contingency tables. *Communications in Statistics - Simulation and Computation* 38(3): 503-512.

Goodman, L. 1979. A multiplicative models for square contingency tables with ordered categories. *Biometrika* 61: 207-214.

Goodman, L.A. 1968. The analysis of cross-classied data: Independence, quasi- Independence, and interactions in contingency tables with or without missing entries. *J. Amer. Statist. Assoc.* 63: 1091-1131.

Haberman, S.J. 1979. *Analysis of Qualitative Data*. Volume 2 New Developments. New York: Academic Press.

Ireland, C.T., Ku, H. & Kullback S. 1965. Symmetry and marginal homogenity of a R*R contingency table. *JASA* 64: 1323-1341.

Lawal, H.B. 2003. *Categorical Data Analysis with SAS and SPSS Applications*. New Jersey: Lawrence Erlbaum Associates Inc.

Lawal, H.B. 1993. Association symmetry and diagonal models for occupational mobility and other similar square contingency tables having ordered categorical variables. *Biometrical J.* 35: 193-206.

McCullagh, P. 1978. A class of parametric models for the analysis of square contingency tables with ordered categories. *Biometrika* 65: 413-418.

Read, T.R.C. & Cressie, N. 1988. *Goodness-of-Fit Statistics for Discrete Multivariate Data*. New York: Springer-Verlag.

Tomizawa, S. 1985. Analysis of data in square contingency tables with ordered categories using the conditional symmetry model and its decomposed models. *Environmental Health Perspectives* 63: 235-239.

Zelterman, D. 1987. Goodness-of-fit tests for large sparse multinomial distributions. *J. Amer. Statist. Assoc.* 82(398): 624-629.

Department of Statistics, Hacettepe University
06800 Beytepe, Ankara
Turkey

*Corresponding author; email: serpilaltunay@gmail.com

Received: 10 May 2015
Accepted: 7 March 2016