DIGITALCOMMONS
@WAYNESTATE

Journal of Modern Applied Statistical Methods

Volume 9 | Issue 2 Article 11

11-1-2010

A Flexible Method for Testing Independence in Two-Way Contingency Tables

Peyman Jafari
Shiraz University of Medical Sciences, Shiraz, Iran

Noori Akhtar-Danesh
McMaster University, Hamilton, Ontario Canada

Zahra Bagheri
Shiraz University of Medical Sciences, Shiraz, Iran

Recommended Citation
Jafari, Peyman; Akhtar-Danesh, Noori; and Bagheri, Zahra (2010) "A Flexible Method for Testing Independence in Two-Way Contingency Tables," *Journal of Modern Applied Statistical Methods*: Vol. 9 : Iss. 2 , Article 11.
DOI: 10.22237/jmasm/1288584600
Available at: http://digitalcommons.wayne.edu/jmasm/vol9/iss2/11

Journal of Modern Applied Statistical Methods
November 2010, Vol. 9, No. 2, 443-451

A Flexible Method for Testing Independence in Two-Way Contingency Tables

Peyman Jafari
Shiraz University of Medical Sciences, Shiraz, Iran

Noori Akhtar-Danesh
McMaster University, Hamilton, Ontario Canada

Zahra Bagheri
Shiraz University of Medical Sciences, Shiraz, Iran

A flexible approach for testing association in two-way contingency tables is presented. It is simple, does not assume a specific form for the association and is applicable to tables with nominal-by-nominal, nominal-by-ordinal, and ordinal-by-ordinal classifications.

Key words: Monte-Carlo simulation, log-linear models, row-effect models.

Introduction

In many social and medical studies a crucial question is whether the categorical variables forming a contingency table are independent. Suppose that a sample of N observations is classified with respect to two categorical variables, one with r levels and the other with c levels. Using the notation in Table 1 for this two-dimensional table, n_{ij} denotes the observed frequency for cell (i, j), and $n_{i.}$ and $n_{.j}$ denote the row and column totals, respectively. Also, P_{ij} is estimated by $\hat{P}_{ij} = \frac{n_{ij}}{N}$.

Table 1: Notation for a Two-Way Contingency Table

Row Variable	Column Variable					Total
	1		j		c	
1	n_{11} p_{11}	...	n_{1j} p_{1j}	...	n_{1c} p_{1c}	$n_{1.}$ $p_{1.}$
i	n_{i1} p_{i1}	...	n_{ij} p_{ij}	...	n_{ic} p_{ic}	$n_{i.}$ $p_{i.}$
r	n_{r1} p_{r1}	...	n_{rj} p_{rj}	...	n_{rc} p_{rc}	$n_{r.}$ $p_{r.}$
Total	$n_{.1}$ $p_{.1}$	...	$n_{.j}$ $p_{.j}$	...	$n_{.c}$ $p_{.c}$	N 1

Log-linear models are a general approach for the analysis of contingency tables. The major advantages of log-linear models are that they provide a systematic approach to the analysis of complex multidimensional tables and estimate the magnitude of effects of interest; consequently, they identify the relative importance of different effects (Agresti, 2002). Let m_{ij} denote the expected frequencies in a two-way contingency table with nominal row and column classifications. In addition, let x and y represent the row and column variables, respectively. In the standard system of hierarchical log-linear models, there are two possible models. The saturated model

$$\log(m_{ij}) = \lambda + \lambda_i^x + \lambda_j^y + \lambda_{ij}^{xy} \tag{1}$$

Peyman Jafari is an Assistant Professor of Biostatistics in the Department of Biostatistics, Faculty of Medicine. His research interests include: sequential clinical trials, design and analysis of quality of life studies. Email: jafarip@sums.ac.ir. Noori Akhtar-Danesh, is an Associate Professor of Biostatistics in the Department of Epidemiology and Biostatistics at the School of Nursing. His research interests include: survival analysis (including analysis of recurrent data and competing risks), multilevel modeling and longitudinal data analysis, structural equation modeling, modeling risk factors of obesity and depression, and meta-analysis. Email: daneshn@mcmaster.ca. Zahra Bagheri is a Ph.D. student in the Department of Biostatistics. Her research interests include: Longitudinal studies and categorical mixed models. Email: zbagheri@sums.ac.ir.

has *rc* parameters and zero degree of freedom (d.f.). Hence, this model describes the data perfectly, however, it is not useful because it does not provide data reduction. The model only serves as a baseline for comparison with the independence model.

The independence model

$$\log(m_{ij}) = \lambda + \lambda_i^x + \lambda_j^y \qquad (2)$$

has $r + c - 1$ parameters and $(r-1)(c-1)$ d.f. for testing lack of fit. Thus, the hypothesis of independence can be tested by comparing the saturated and independence models. The deviation from independence can be measured by the likelihood ratio statistic (LR)

$$D_I = 2\sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij} \log\left(\frac{n_{ij}}{\hat{m}_{ij}}\right)$$

where $\hat{m}_{ij} = n_{i.}n_{.j} / N$ is the estimation of the expected frequency in the i^{th} category of the row and the j^{th} category of the column variable under the hypothesis of independence (H_0). If H_0 is true, D_I has an asymptotic Chi-square distribution with *(r-1)*(*c-1*) degrees of freedom.

The log-linear method presented has a number of limitations. First, it often has low power to detect departures from independence, especially when the dimension of the table increases (Davis, 1991). Second, it treats all classifications as nominal; therefore if the order of categories changes for a variable in any way, the fit remains the same (Agresti, 2002). Instead, if the row and column variables are both ordinal with known scores, the Linear-by-Linear association model can be used. On the other hand, when scoring is used only for one of the row or column variables, the row-effect or column-effect association model can be used (Agresti, 1984).

In practice it may not be possible to choose obvious scores for both the row and column categories. One alternative is Goodman's RC model, in which the row and column scores are treated as parameters to be estimated (Goodman, 1969). Although the RC model can be used if the two variables are nominal, which does not impose any restriction on the type of the variables, calculation of the conditional test of independence is complicated and the distribution of the test statistic is not Chi-square (Agresti, 2002). In all of these models the researcher needs to specify the functional form for the association and, if the association form is chosen incorrectly, then the power of the model will decrease.

It should be noted that, some methods used for testing interaction in two-way ANOVA can also be applied to two-way contingency tables for testing association (Alin & Kurt, 2006). For example, Davis (1991) tested association in two-way contingency tables based on Tukey's model (Tukey, 1949). Also Christensen (1990) tested interaction in log-linear and logit models for categorical data with the logit version of Mandel's models (Mandel, 1961). Milliken and Graybill (1970) established a two-stage fitting procedure using Tukey's model (Tukey, 1949). Recently, Kharati and Sadooghi (2007) have proposed a new method for testing interaction in two-way ANOVA.

In this study, the same method used by Kharati and Sadooghi (2007) will be applied for testing association in two-way contingency tables. It is a flexible approach for testing independence that does not assume a special form for the association model. The method was applied to detect association in tables with nominal-by-nominal and nominal-by-ordinal data.

Methodology

Row Effect Model

If either the row or the column variable (but not both of them) is ordinal, then a row-effect or column-effect model can be fitted (Agresti, 1984; Agresti, 2002). The row effects model has the form

$$\log(m_{ij}) = \lambda + \lambda_i^x + \lambda_j^y + \mu_i v_j. \qquad (4)$$

This model is appropriate for two-way tables with ordered columns, using scores $v_1 < v_2 < \ldots < v_c$. Because the rows are unordered, the model treats them as parameters and denotes them by μ_i. The μ_i's are called the row effects. This model has *r-1* more parameters than the independence model, which is a special case where $\mu_1 = \mu_2 = ... = \mu_r$.

The LR test of independence requires maximum likelihood (ML) estimates $\hat{m}_{ij}$ of expected cell frequencies under model (4). Let D_R denote the LR goodness of fit statistic for model (4) and let D_I denote the classical test of independence given by (2). A (*r*-1) (*c*-2) degrees of freedom test of $H_0 : \mu_1 = \mu_2 = ... = \mu_r$ can then be based on the LR statistic $D_{I|R} = D_I - D_R$.

We used the same method proposed by Kharati and Sadooghi (2007) for testing association in two-way contingency tables. Assume a $r \times c$ contingency table and simultaneously $r \geq 4$ (so the method excludes only 2×2, 3×2 and 3×3 tables). Divide the table according to the rows, into two sub-tables. The sub-tables are two contingency tables with $r_1 \times c$ and $r_2 \times c$ dimensions in which $r_1 + r_2 = r$. In the absence of any association in each sub-table, then the independence model

$$\log m_{ij} = \lambda + \lambda_i^x + \lambda_j^y \qquad (5)$$

can fit both datasets well. Let D_{I1} and D_{I2} denote the deviances for the two sub-tables, respectively. In generalized linear models if the response variables are normally distributed then D has a Chi-square distribution exactly. However, for responses with a Poisson distribution, the sampling distribution of D may have an approximate Chi-square distribution (Dobson, 2002). Therefore, under the independence log-linear model, D_{I1} and D_{I2} are independent and have approximate Chi-square distributions with df$_1$=$(r_1 - 1)(c - 1)$ and df$_2$= $(r_2 - 1)(c - 1)$ degrees of freedom, respectively. A new statistic for testing independence in two-way contingency tables is now defined.

If $t_1 = \frac{D_{I1}}{df_1}$ and $t_2 = \frac{D_{I2}}{df_2}$, then the new variable $F^* = \frac{\text{Max}(t_1, t_2)}{\text{Min}(t_1, t_2)}$ has the F distribution with d.f. = (df$_1$, df$_2$) where $t_1 > t_2$ or d.f. = (df$_2$, df$_1$) where $t_2 > t_1$. In the presence of any association, the F^* statistic tends to be large, thus, the hypothesis of no association if $F^* > F_\alpha(df_1, df_2)$ is rejected when $t_1 > t_2$ or $F^* > F_\alpha(df_2, df_1)$ where $t_2 > t_1$.

However, in this approach the most important question is how a table can be split into two separate tables. In some cases, based on a priori information, there may be a natural division of the table. In the absence of a-priori information, drawing a profile plot is suggested. Based on such a profile plot those lines which are parallel or have the same pattern will be put in the same group and the remaining in the other group. Additional details are provided in the examples and readers are also referred to Kharati and Sadooghi (2007) for more information.

Simulation Study

The programming for the Monte Carlo simulation was written in SAS version 9.1. The RANTBL function was used for generating and simulating contingency tables in SAS (Fan, Felsovalyi, Sivo & Keenan, 2002). Contingency table data may result from one of several possible sampling models. The test of independence discussed in this study is based on sampling in which a single random sample of size <u>*N*</u> is classified with respect to two characteristics simultaneously (Dobson, 2002). In the resulting contingency table, both sets of marginal total frequencies are random variables. The empirical power of each test was determined by simulating contingency tables under the dependence structure, and computing the proportion of times the independence hypothesis was rejected at a given significance level α. Under the dependence structure, P_{ij} is estimated by $\hat{P}_{ij} = \frac{n_{ij}}{N}$ (Table 1).

For each studied situation, 5,000 contingency tables were generated in which cell frequencies were drawn under the dependent structure. The influence of the total sample size (*N*) on the statistical properties of all tests was also evaluated. The choice of the proper total sample size for simulation depends on dimensions of the table. The power of the $D_{I|R}$ and F statistics for testing independence in two-way contingency tables (nominal-by-ordinal) were investigated and compared. In order to find the maximum F in each simulated table, all combinations of rows and columns to classify each table into two subtables were considered. The power of the D_I and *F* statistics for testing independence in two-way contingency tables (nominal-by- nominal) were also computed and compared.

Example 1: The Location of Prehistoric Artifact

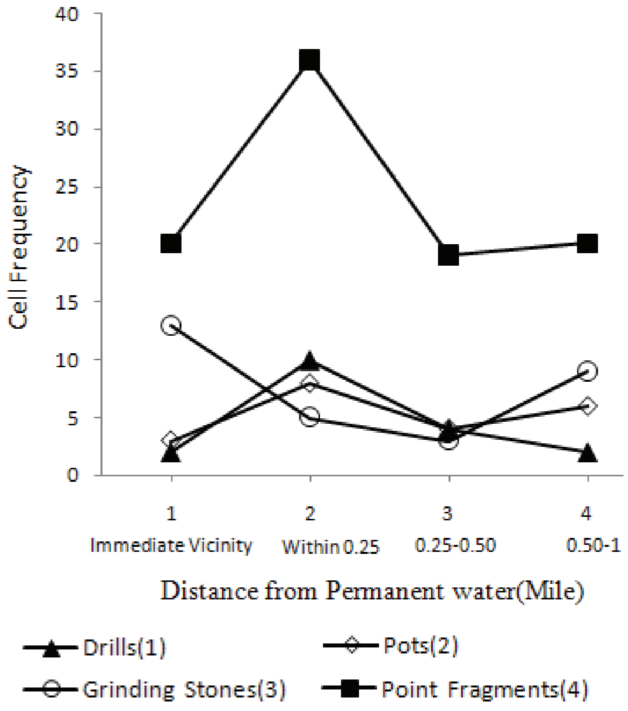
This example is based on the data provided in Simonoff (2003). As a result of archaeological excavations in Ruby Valley, Nevada, various prehistoric artifacts were discovered. Archaeologists were interested in the relationship between the type of artifacts found and the distance to permanent water, because the type of artifact discovered describes the type of site used by prehistoric hunters (Table 2). It was presumed that some tools were more difficult to move place to place and would thus be more likely to be discovered near permanent water. The following table is based on a subset of the artifacts discovered in Nevada (Simonoff, 2003).

In this example the row variable is nominal and the column variable is ordinal. Using the row-effect model (4), $D_R = 14.85$, $D_I = 16.26$ and $D_{I|R} = D_I - D_R$ =1.40. With respect to the asymptotic Chi-square distribution, χ^2_3 =7.815, there is no evidence of departure from independence. A similar result was obtained based on the F statistic. In the profile plot for these data (shown in Figure 1), the lines corresponding to rows 2, 4 are parallel. Thus, these rows were placed in the first sub-table and the remaining rows in the second sub-table. In this situation, F (3, 3) = 14.94 and P = .002 which is significant at the nominal level of 0.05. The result of our simulation showed that the F statistic is considerably more powerful than the row-effect model. The power of the F and $D_{I|R}$ are 0.43 and 0.15 respectively.

Table 2: Frequencies for Artifact Type and Distance from Permanent Water

Artifact Type	Distance from Permanent Water			
	Immediate Vicinity	Within 0.25 Miles	0.25-0.50 Miles	0.50-1 Miles
Drills	2	10	4	2
Pots	3	8	4	6
Grinding Stones	13	5	3	9
Point Fragments	20	36	19	20

Figure 1: Profile Plot of Data in Example 1



Example 2.1: Malignant Melanoma

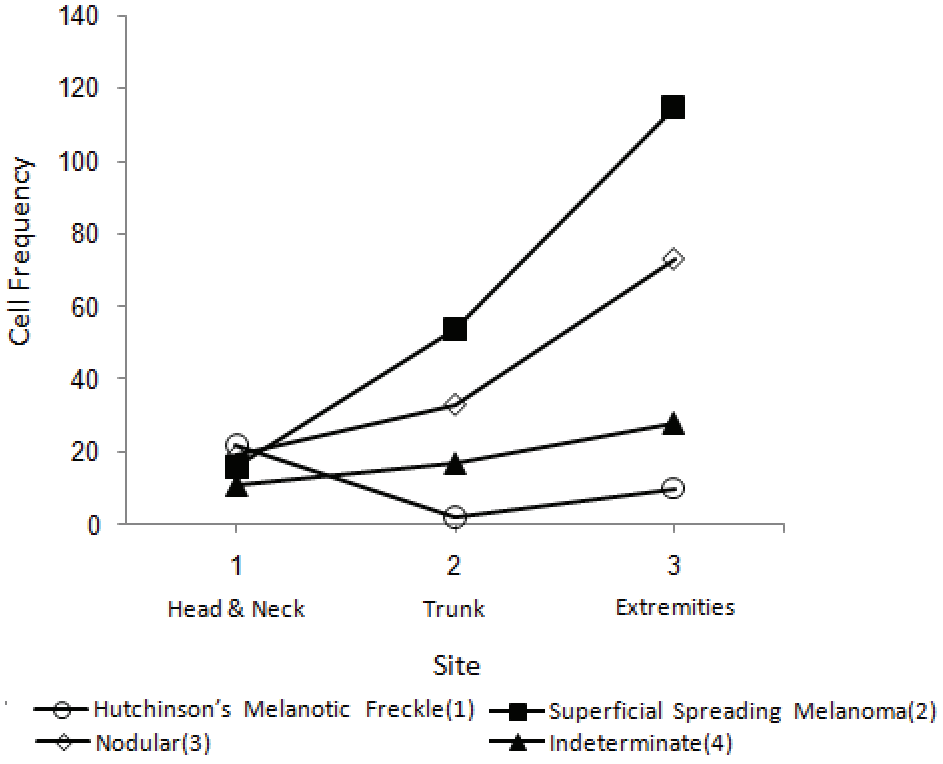
For the data in Table 3 the question of interest is whether there is any association between tumor type and site. These data are from a cross-sectional study of patients with a form of skin cancer called malignant melanoma (Dobson, 2002). For a sample of *N*=400 patients the site of the tumor and its histological type were recorded.

In testing the null hypothesis that tumor type and tumor site are independent, $D_I = 51.79$ and P < .001, which indicate that the association between type and site of tumor is highly significant. A similar result was obtained based on the proposed F statistic. In the profile plot for these data in Figure 2, the lines corresponding to rows 3 and 4 are nearly parallel which suggests that these rows can be placed in the one sub-table and the remaining rows in the other sub-table. The F statistic value for this division is statistically significant, F (2, 2) = 43.41, p = .02.

Table 3: Frequencies for Tumor Type and Site

Tumor Type	Site		
	Head and Neck	Trunk	Extremities
Hutchinson’s Melanotic Freckle	22	2	10
Superficial Spreading Melanoma	16	54	115
Nodular	19	33	73
Indeterminate	11	17	28

Figure 2: Profile Plot of Data in Malignant Melanoma Example 2.1



Example 2.2: Malignant Melanoma

Next, substitute the frequencies 2, 16, 115, 73 and 28 in the cells (1, 2), (2, 1), (2, 3), (3, 3) and (4, 3) by 18, 45, 60, 38 and 20, respectively. In this situation, the null hypothesis is tested again. The new results, based on the likelihood ratio statistic, show that there is no significant association between tumor site and tumor type, $D_I = 11.80$, $P = .067$. However, a different result was obtained based on the F statistic at the $\alpha=0.05$ level. In the profile plot for these data (Figure 3), the lines corresponding to rows 3, 4 are nearly parallel and close to each other. Therefore, these rows were placed in one table and the remaining rows in another table. The value of the F statistic for this division is highly significant, $F(2, 2) = 108.42$, $p < .01$.

Simulation Results

The results of the simulations showed that the power of the F and LR statistics in Malignant Melanoma Example 2.1 are 0.653 and 1, respectively, and in Malignant Melanoma Example 2.2 are 0.425 and 0.736, respectively.

This study also evaluated the influence of the total sample size (N) on the statistical properties for the above two examples. Table 4 shows the results of the estimation of power of the proposed F statistic and row-effect model ($D_{I|R}$) based on 5,000 simulated tables for the nominal-by-ordinal association model in Example 1. Table 5 shows these results for the proposed F statistic and the likelihood ratio statistic (D_I) based on 5,000 simulated tables for the nominal-by-nominal association model in Examples 2.1 and 2.2.

Table 4 shows that, for $N \leq 800$, especially when $N \leq 500$, the estimated power for the F statistic is considerably higher compared to the row-effect model ($D_{I|R}$). However for $N > 900$ the power of the row-effect model is dramatically higher than the F statistic.

Figure 3: Profile Plot of Data in Malignant Melanoma 2.2 Example

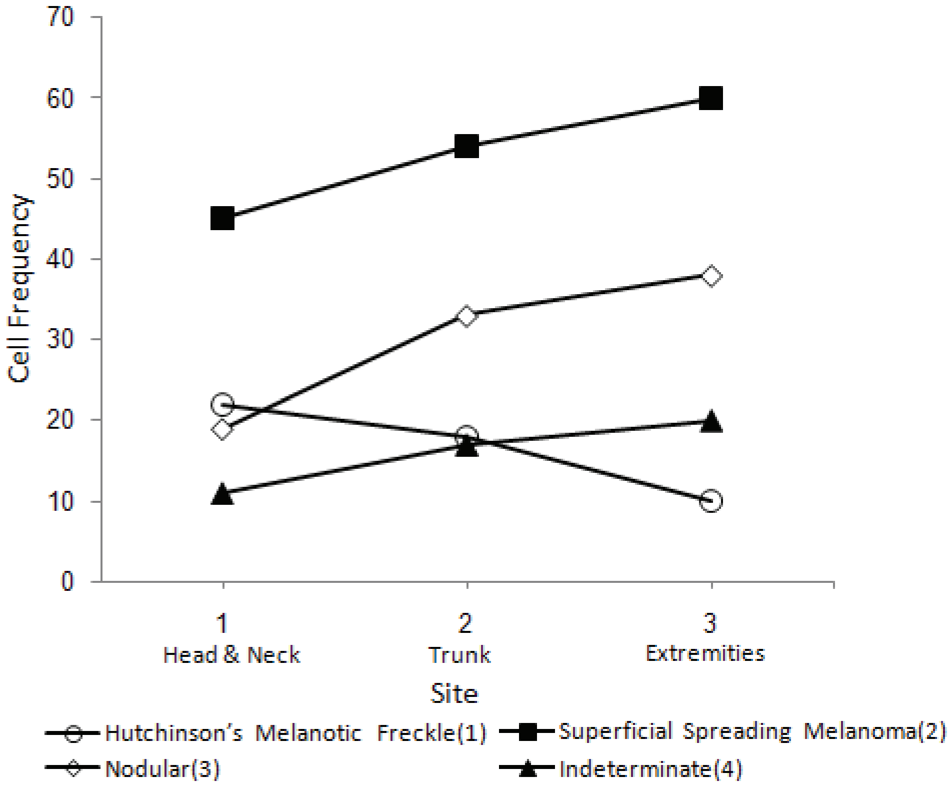


Table 4: Nominal-by-Ordinal Association: Estimation of Power for the F Statistic and the Row-Effect Model ($D_{I|R}$) Based on 5,000 Simulations in Example 1

N	F	$D_{I\|R}$
200	0.440	0.172
300	0.488	0.242
400	0.514	0.311
500	0.553	0.372
600	0.575	0.442
700	0.576	0.499
800	0.606	0.564
900	0.617	0.616
1,000	0.620	0.671
2,000	0.731	0.937

Table 5: Nominal-by-Nominal Association: Estimation of Power and for the F Statistic and the Likelihood Ratio Statistics (D_I) Based on 5,000 Simulations

Example 2.1		
N	F	D_I
200	0.532	0.976
300	0.602	0.999
400	0.641	1.000
500	0.669	1.000
600	0.697	1.000
700	0.716	1.000
800	0.736	1.000
900	0.759	1.000
1,000	0.769	1.000
2,000	0.872	1.000

Example 2.2		
N	F	D_I
200	0.343	0.474
300	0.387	0.649
400	0.455	0.806
500	0.490	0.903
600	0.525	0.947
700	0.564	0.979
800	0.583	0.988
900	0.610	0.996
1,000	0.629	0.998
2,000	0.772	1.000

Regarding Example 2.1, Table 5 shows that, for all *N*, the likelihood ratio statistic is considerably more powerful compared with the F statistic. When *N* increases, the power of the F statistic steadily increases, but the power of the likelihood ratio statistic converges to 1 for $N >$ 400. Conversely, in Example 2.2, although the power of the likelihood ratio statistic is higher than the power of the F statistic, the rate of power increase is lower compared to Example 2.1.

Conclusion

A new statistic is proposed for testing independence in two-way contingency tables by dividing a table into two sub-tables. This method has been constructed based on the independence model so there is no need to specify any functional form for the association terms. Therefore, it could be applicable to any type of contingency tables, including nominal-by-nominal, nominal-by-ordinal and ordinal-by-ordinal.

The idea of partitioning contingency tables was first introduced by Kullback, et al. (1962) and Lancaster (1951). They showed that the overall Chi-square statistic for a contingency table can always be partitioned into as many components as the table's degrees of freedom. The Chi-square value of each component corresponds to a particular 2×2 table arising from the original table, and each component is independent of the others. Consequently a detailed examination of departures from independence can be made, thus enabling identification of those categories responsible for a significant overall Chi-square value. However, in this article the same technique was used for partitioning contingency tables that was applied to two-way ANOVA by Kharati and Sadooghi (2007). In the present work, this method was used for analyzing nominal-by-nominal and nominal-by-ordinal data.

It is notable that in a two-way ANOVA data are assumed to be normally distributed and the proposed F for testing interaction has an exact F distribution which leads to a two-sided test for equality of variances. In this study the response variable had Poisson distribution, so the proposed one-sided test has an asymptotic F distribution. Profile plots were also used as a preliminary tool to divide one table into two separate tables, which was the first step before applying the proposed method. However, there are other graphical methods such as corresponding analysis (Blasius & Greenarce, 2006), mosaic (Friendly, 1998) and z-plot (Choulakian & Allard, 1998), all of which can be helpful to visualizing and screening contingency tables before conducting any formal statistical analysis.

The power of the F statistic was compared with D_I and $D_{I|R}$. In Example 1, in which the row and column were nominal and ordinal respectively, it was believed that the row-effect model would be the best method for testing the association between row and column. Surprisingly, the proposed F statistic worked much better than expected. The results showed that while $D_{I|R}$ could not find any association between rows and columns; the proposed F was strongly significant. In this case the power simulation showed that the F statistic is more powerful than $D_{I|R}$ (0.43 vs. 0.15). Also the simulation results in Table 4 showed that for $N \leq$ 500 the power of the F statistic was considerably higher than $D_{I|R}$. In this example, the results of the proposed F demonstrated that, despite the simplicity of its computations, it is more powerful than the row-effect model. These findings may encourage researchers to use the proposed F statistic for testing association in contingency tables.

In the Malignant Melanoma Example 2.1 when the row and column were nominal and there was a significant association between them, the simulation results showed that the D_I statistic was more powerful than F. In contrast, in the Malignant Melanoma Example 2.2, although D_I could not find any association between row and column, the proposed F was strongly significant. However, simulation showed that D_I was more powerful than F (0.76 vs. 0.44). In this case it should be noted that although the F statistic was often less powerful than the D_I, it was able to detect some special types of departures from the null hypothesis which could not be detected by D_I.

In conclusion, it is suggested that the F statistic serves as an alternative method for testing association in two-way contingency tables, in particular, if one variable is in ordinal scale. It is easy to use because it does not need any functional form for the association term. It is simple to compute and has good power. In addition to simplicity and flexibility, this test could be helpful in detecting the part of a table which contributes the association between row and column. It seems that, in some cases, this method enables us to detect an association in contingency tables that cannot be found by a row-effect model or likelihood ratio statistics.

References

Agresti, A. (1984). *Analysis of ordinal categorical data*. New York, NY: Wiley.

Agresti, A. (2002). *Categorical data analysis*. New York, NY: Wiley.

Alin, A., & Kurt, S. (2006). Testing non-additivity (interaction) in two-way ANOVA tables with no replication. *Statistical Methods in Medical Research*, *15*, 63-85.

Blasius, J., & Greenarce, M. (2006). Correspondence analysis and related methods in practice. In J. Blasius, & M. Greenarce, Eds., *Multiple correspondence analysis and related methods* (*Statistics in the Social and Behavioral Sciences*), 3-41. London: Chapman & Hall.

Choulakian, V., & Allard, J. (1998). The z-plot: a graphical procedure for contingency tables with an ordered response variable. In J.Blasius, & M. Greenarce, Eds., *Visualization of categorical data*, 99-106. San Diego, CA: Academic Press.

Christensen, R. (1990). *Log-linear models*. New York, NY: Springer-Verlag.

Davis, C. S. (1991). A one degree of freedom nominal association model for testing independence in two-way contingency tables. *Statistics in Medicine*, *10*, 1555-1563.

Dobson, A. (2002). *An Introduction to generalized linear models*. London: Chapman & Hall.

Fan, X., Felsovalyi, A., Sivo, S. A., & Keenan, S. C. (2002). *SAS for Monte Carlo studies a guide for quantitative researchers*. North Carolina: SAS Institute.

Friendly, M. (1998). Conceptual models for visualizing contingency table data. In J.Blasius, & M. Greenarce, Eds., *Visualization of categorical data*, 17-36. San Diego, CA: Academic Press.

Goodman, L. A. (1969). On partitioning chi-square and detecting partial association in three-way contingency tables. *Journal of the Royal Statistical Society*, *Series B*, *31*, 486-498.

Kharati, M., & Sadooghi, S. M. (2007). A new method for testing interaction in unreplicated two-way analysis of variance. *Communications in Statistics-Theory and Methods*, *36*, 2787-2803.

Kullback, S., Kupperman, M., & Ku, H. H. (1962). Tests for contingency tables and Markov chains. *Technometrics*, *4*, 573-608.

Lancaster, H. O. (1951). Complex contingency tables treated by partition of Chi-square. *Journal of the Royal Statistical Society*, *Series B*, *13*, 242-249.

Mandel, J. (1961). Non-additivity in two-way analysis of variance. *Journal of the American Statistical Association*, *56*, 878-888.

Milliken, G. A., & Graybill, F. A. (1970). Extensions of the general linear hypothesis model. *Journal of the American Statistical Association*, *65*, 797-807.

Simonoff, J. F. (2003). *Analyzing categorical data*. New York, NY: Springer.

Tukey, J. W. (1949). One degree of freedom for non-additivity. *Biometrics*, *5*, 232-242.