Estimate-based Goodness-of-Fit Test for Large Sparse Multinomial Distributions

by

Sung-Ho Kim, Heymi Choi, and Sangjin Lee

Applied Mathematics
Research Report
07-08
November 12, 2007

DEPARTMENT OF MATHEMATICAL SCIENCES

KAIST 한국과학기술원
Korea Advanced Institute of Science and Technology

Estimate-based Goodness-of-Fit Test for Large Sparse Multinomial Distributions

Sung-Ho Kim [a,*], Hyemi Choi [b], Sangjin Lee [a]

[a] *Department of Mathematical Sciences, Korea Advanced Institute of Science and Technology, Daejeon, 305-701, South Korea*

[b] *Department of Statistical Informatics, Chonbuk National University, Chonju, 561-756, South Korea*

Abstract

The Pearson's chi-squared statistic (X^2) does not in general follow a chi-square distribution when it is used for goodness-of-fit testing for a multinomial distribution based on sparse contingency table data. We explore properties of Zelterman's (1987) D^2 statistic and compare them with those of X^2 and we also compare these two statistics and the statistic (L_r) which is proposed by Maydeu-Olivares and Joe (2005) in the context of power of the goodness-of-fit testing when the given contingency table is very sparse. We show that the variance of D^2 is not larger than the variance of X^2 under null hypotheses where all the cell probabilities are positive, that the distribution of D^2 becomes more skewed as the multinomial distribution becomes more asymmetric and sparse, and that, as for the L_r statistic, the power of the goodness-of-fit testing depends on the models which are selected for the testing. A simulation experiment strongly recommends to use both D^2 and L_r for goodness-of-fit testing with large sparse contingency table data.

Key words: Test power, P-value, Sample-cell ratio, Model discrepancy, Normal approximation, Skewness

1 Introduction

A variety of statistics have been derived for goodness-of-fit testing such as the Pearson chi-squared statistic (X^2) (Pearson, 1900), the log likelihood ratio statistic (G^2), the Freeman-Tukey statistic, the Neyman modified X^2, and the modified G^2.

* Corresponding author.

Email addresses: shkim@amath.kaist.ac.kr (Sung-Ho Kim), hchoi@chonbuk.ac.kr (Hyemi Choi), Lee.Sang.jin@kaist.ac.kr (Sangjin Lee).

Cressie and Read (1984) proposed a family of power divergence statistics, in which all of the above statistics are embedded. All of these statistics are asymptotically chi-squared distributed with appropriate degrees of freedom (e.g. Fienberg, 1979; Horn, 1977; Lancaster, 1969; Moore, 1976; Watson, 1959).

Two most popular statistics among them are X^2 and G^2. Cochran (1952) gives a nice review of the early development of X^2 and concludes that there is little to distinguish it from G^2. The chi-squared approximation to X^2 and G^2 was studied by Read (1984), and the accuracy of these approximations was examined by Koehler and Larntz (1980) and Koehler (1986). Koehler (1986) demonstrated asymptotic normality of G^2 for testing the fit of log-linear models with closed form maximum likelihood estimates in sparse contingency tables and showed that the normal approximation can be much more accurate than the chi-squared approximation for G^2, but he admitted that the bias of the estimated moments is a potential problem for very sparse tables. A good review on goodness-of-fit tests for sparse contingency tables is provided in Koehler (1986).

Zelterman (1987) proposed a statistic D^2 and compared it with X^2 for testing goodness-of-fit (GOF) to sparse multinomial distributions, where

$$D^2 = X^2 - \sum^{*}(n_i/e_i),$$

where n_i is the count in the ith cell, e_i the estimate of the cell mean of the ith cell, and $\sum^*$ is the summation over all the cells such that $e_i > 0$. The D^2 statistic is not a member of the family of Cressie and Read's power divergence statistics.

The goal of this paper is to explore properties of the goodness-of-fit test using D^2 (or the D^2 test for short), in comparison with the goodness-of-fit test using X^2 (or the X^2 test for short), other than the asymptotic properties as derived in Zelterman (1987), to do a more rigorous comparative study on the goodness-of-fit tests using the statistics, D^2, X^2, and L_r, in the context of power, and to propose an approach to goodness-of-fit testing for a very sparse multinomial distribution.

This paper consists of six sections. Section 2 describes what happens to X^2 when the data size n is smaller than the number of cells (k) of a given contingency table and summarizes earlier works on asymptotic properties of X^2 with more attention paid to the case that $n < k$. Section 3 compares D^2 and X^2 in the context of variance, and it is shown that the variance of D^2 is not larger than that of X^2 under any null hypothesis of multinomial distribution. We carry out a Monte Carlo study in Section 4 to compare the goodness-of-fit tests by using D^2, X^2, and L_r in the context of test power. The Monte Carlo result strongly suggests that the D^2 test is at least as good as the X^2 test when $n < k$ and comparable with the L_r test. We then investigated normal approximation to the standardized D^2 curve in Section 5. Section 6 concludes the paper with some further remarks on the goodness-of-fit testing for sparse contingency tables.

2 A BRIEF REVIEW ON GOODNESS-OF-FIT TESTS USING X^2 AND SOME RELATED STATISTICS

Suppose that we have a contingency table with k cells and n observations, and consider a goodness-of-fit testing of a symmetric null hypothesis, i.e., all the cell probabilities are the same. We will, first of all, consider the two extreme cases that $n = 1$ and $n = 2$. Under the symmetric null hypothesis, $X^2 = k - 1$ when $n = 1$; and when $n = 2$, it is either $k - 1$ or $k/2 - 1$ according to whether a particular cell frequency equals 2 or not. It is interesting to note that when $n < k$, X^2 is strictly positive under any hypothesis for the contingency table. This simply shows that the asymptotic distribution of X^2 can not be a chi-square distribution when $n < k$. This is one of the reasons that the normal approximation was sought for in the literature when $n < k$.

While the literature on the asymptotic properties of X^2 abounds in regard to contingency tables with large enough data (e.g., Agresti (1990); Bishop et al. (1975)), a unifying result is yet to be explored as for the goodness-of-fit testing for sparse contingency tables. Results from the literature concerning Pearson's statistics are summarily listed below in regard to sparse multinomial problems.

(i) The X^2 test has some optimal local power properties in the symmetrical case when the number of cells is moderately large. Hence the X^2 test based on the traditional chi-squared approximation is preferred for the test of symmetry. For the null hypothesis of symmetry, the chi-squared approximation for X^2 is quite adequate at the 0.05 and 0.01 nominal levels for expected frequencies as low as 0.25 when $k \geq 3,\ \ n \geq 10,\ \ n^2/k \geq 10$ (Koehler and Larntz, 1980, p. 343)

(ii) The chi-squared approximation for X^2 produces inflated rejection levels for asymmetrical null hypotheses that contain many expected frequencies smaller than 1 (Koehler and Larntz, 1980, p. 344).

(iii) The adverse effects of small expected frequencies on the chi-squared approximation are generally less severe for X^2 than for other commonly used goodness-of-fit statistics (Koehler, 1986, p. 483).

(iv) The asymptotic power of the D^2 test is moderate and at least as great as that of the normalized X^2 (Zelterman, 1987, p. 628).

(v) For the symmetric null hypothesis, choosing the value of the family parameter λ of Cressie and Read's (1984) power divergence family of statistics in $[\frac{1}{3}, 1\frac{1}{2}]$ results in a test statistic whose size can be approximated accurately by a chi-squared distribution, provided that $k \leq 6$ and $n \geq 10$. When $\lambda \in [-5,\ 5] \setminus [\frac{1}{3}, 1\frac{1}{2}]$, the corrected chi-square approximation F_C is recommended where F_C is the corrected chi-square distribution for which the mean and variance agree to second order with the mean and variance of Cressie and Read's power divergence statistics, since it gives reasonably accurate approximate levels and is easy to compute. When both n and k are large (say, $n > 100$), then the

normal approximation should be used (Read, 1984, p. 935).

(vi) Berry and Mielke (1988) proposed non-asymptotic chi-square tests based on X^2 and D^2 that use Pearson Type *III* distribution approximation for testing large sparse contingency tables.

(vii) Maydeu-Olivares and Joe (2005) proposed a statistic, L_r, for goodness-of-fit testing using 2^J contingency table data which is given in a quadratic form based on multivariate moments up to order r and showed that X^2 belongs to this family of L_r. They showed that when r is small (e.g., 2 or 3), the L_r has better small-sample properties and are asymptotically more powerful than X^2 for some multivariate binary models.

3 COMPARISON OF VARIANCES

The favorable comparison (Zelterman (1987, p. 628)) of D^2 over the normalized version of X^2 in the context of asymptotic power in goodness-of-fit tests may be extended to the situations of small or moderate data sizes. To see this, we will have a closer look at the two statistics and compare their exact means and variances across multinomial distributions, which will give us an insight into the power comparison.

Suppose that the cell probability at the ith cell, p_i, is equal to p_{0i} under the null hypothesis (H_0) and let it be equal to p_{ai} under an alternative hypothesis (H_a). We will denote the cell means in the same manner, i.e., by m_{0i} under H_0 and by m_{ai} under H_a. Throughout the rest of the paper, we will assume that $\prod_i p_{0i} > 0$.

We can then re-express D^2 and X^2 as

$$X^2 = \sum_{i=1}^{k} (n_i - m_{0i})^2 / m_{0i},$$
$$D^2 = X^2 - \sum_i n_i / m_{0i}.$$

We list a couple of simple facts:

$$E(X^2) = \sum_{i=1}^{k} n p_i q_i / m_{0i}. \tag{1}$$
$$E(D^2) = E(X^2) - \sum_i p_i / p_{0i}. \tag{2}$$

The difference between the variance of D^2 and that of X^2 is given in

Theorem 1 *Under a multinomial assumption with the sample size n and a contingency table of k cells, we have*

$$\begin{aligned} V(D^2) &= V(X^2) + n\left(\sum_{i=1}^{k}\sum_{j=1}^{k}\frac{p_i p_j}{m_{0i}m_{0j}} - \sum_{i=1}^{k}\frac{p_i}{m_{0i}^2}\right) \\ &\quad +4n(n-1)\left(\sum_{i=1}^{k}\sum_{j=1}^{k}\frac{p_i p_j^2}{m_{0i}m_{0j}} - \sum_{i=1}^{k}\frac{p_i^2}{m_{0i}^2}\right) \\ &= V(X^2) + \sum_{i=1}^{k}\frac{p_i + 4(n-1)p_i^2}{m_{0i}}\sum_{j=1}^{k}\left(\frac{p_j}{p_{0j}} - \frac{1/k}{p_{0i}}\right). \end{aligned}$$

Proof: See Appendix A.

A couple of corollaries follow from this theorem.

Corollary 2 *Under H_0, $V(D^2)$ is simplified as*

$$\begin{aligned} V(D^2) &= V(X^2) + \frac{k^2}{n} - \sum_{i=1}^{k}\frac{1}{m_{0i}} \\ &\leq V(X^2), \end{aligned}$$

where the equality holds if and only if H_0 is symmetrical.

Corollary 3 *Suppose that $p_{ah} = 1$, for some $1 \leq h \leq k$ under an H_a. Then under the H_a, we have*

$$V(D^2) = V(X^2). \tag{3}$$

Let $\delta_1 = E(D^2) - E(X^2)$ and $\delta_2 = V(D^2) - V(X^2)$. Corollary 2 implies that $\delta_2 < 0$ in a near neighborhood of $(p_{01}, \ldots, p_{0k})$ when H_0 is not symmetrical. And judging from Corollary 3, we can see that δ_2 is bounded on the boundary of the $k-1$ dimensional simplex of $(p_1, \ldots, p_k)$ since δ_2 is continuous in $(p_1, \ldots, p_k)$.

4 SIMULATION EXPERIMENTS

To compare the X^2 and D^2 tests, we conducted simulation experiments with small samples, using a model of 10 binary variables whose model structure is represented as a directed acyclic graph (DAG). We will call this model Model 0 and denote it by M_0. Ten alternative models are considered for the goodness of fit test, which are determined by adding or deleting some edges from Model 0. They are displayed

in Figure 1 along with the true model, M_0, where each vertex represents a binary variable. A brief explanation of the DAG models is given in Appendix B.

For each model, sample sizes $n = 50, 200, 500$ are selected such that n/k is close to $1/20, 1/5, 1/2$. For each of the ten alternative models, say M^*, we obtain an approximate value of the p-value of the goodness-of-fit test as follows:

(i) Generate a sample $\{(n_1^{(0)}, \ldots, n_k^{(0)}) : \sum_{i=1}^k n_i^{(0)} = n\}$ of size n from M_0.

(ii) Estimate the cell probabilities for model M^* based on the sample values, $n_1^{(0)}$, $\cdots, n_k^{(0)}$.

(iii) Generate a sample of size n from model M^* based on the estimates of the cell probabilities that are obtained in (ii) and compute X^2 and D^2 using the sample values and the estimates obtained in (ii).

(iv) Repeat (iii) until t X^2 and D^2 values are obtained, and arrange the t X^2 values in an increasing sequence of $x_{(1)}, \cdots, x_{(t)}$ and similarly the D^2 values into the sequence of $d_{(1)}, \cdots, d_{(t)}$.

(v) Compute the values of X^2 and D^2 using the sample values, $n_1^{(0)}, \cdots, n_k^{(0)}$, and the estimates obtained in (ii), and then obtain the values of $p_x(X^2)$ and $p_d(X^2)$ which are defined below in expression (4).

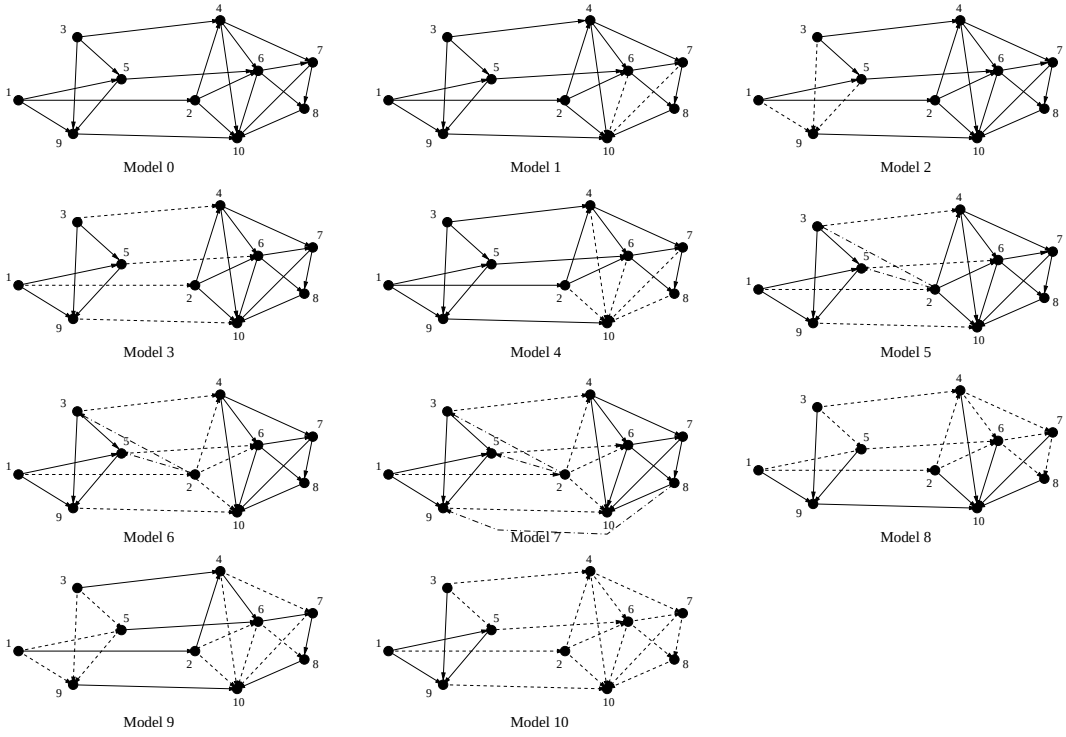


Fig. 1. Graphs of 11 DAG Models, Models 0(M_0), 1(M_1) through 10(M_{10}). Models $M_i,\ i = 1, 2, \cdots, 10$, are constructed by removing (dashed line) or adding (dot-dashed line) arrows to M_0.

We define a pseudo p-value of X^2, $p_x(X^2)$, as

$$p_x(X^2) = \begin{cases} \frac{1}{t+1} & \text{if } X^2 \geq x_{(t)} \\ 1 - \frac{i}{t+1} & \text{if } x_{(i)} \leq X^2 < x_{(i+1)},\ 1 \leq i < t\,. \\ 1 - \frac{0.5}{t+1} & \text{if } X^2 < x_{(1)} \end{cases} \tag{4}$$

A pseudo p-value of D^2, $p_d(D^2)$, is defined similarly as above except that X^2 and x are replaced with D^2 and d, respectively. As is apparent in (4), the pseudo p-value gets close to the actual p-value as t increases, where the actual p-value is for the goodness-of-fit test of model M^*. If $X^2 \geq x_{(1)}$, we have

$$0 \leq p_x - \text{(the actual p-value)} < \frac{1}{t+1},$$

and

$$|p_x - \text{(the actual p-value)}| < \frac{0.5}{t+1}$$

otherwise. The same inequalities hold for p_d also.

To compare the power of goodness-of-fit test between D^2 and X^2, we repeated the above procedure, (i) through (v), 500 times to obtain as many values of each of p_x and p_d with $t = 100$. Table 1 lists the proportions that p_x (p_d) values are less than $\alpha \in \{0.05, 0.1\}$ for three sample sizes(n) 50, 200, and 500.

In the table, we can see that the D^2 test is more powerful than the X^2 test. Model M_0 is of 10 binary variables, which means a contingency table of $2^{10} = 1,024$ cells. So the sample-cell ratios (n/k) considered in the table are $50/1024 \approx 0.05$, $200/1024 = 0.195$, and $500/1024 = 0.488$. The power increases as the sample size increases for both of the D^2 and the X^2 tests. However, it is interesting to see that the power does not change noticeably for some models with the X^2 test. As for models M_2 and M_4, the proportions that $p_x \leq 0.05$ is not larger than 0.1 while those for p_d increase up to 0.43 and 0.79, respectively. We can see the same phenomenon for the two models when $\alpha = 0.1$. We can see in the figures in Appendix C that p_x is less sensitive to the sample size than p_d.

We will denote by $\pi_x(\alpha)$ the proportion that $p_x \leq \alpha$ out of the 500 p_x values and similarly for $\pi_d(\alpha)$. From the π values of models M_2, M_3, and M_4, we can see that the π values do not change monotonically in the number of deleted edges from model M_0. It is indicated in the table that the π value is sensitive to how the edge deletion affects the inter-relationship among the variables in the model. For example, in model M_3, variables X_1, X_3, X_5, and X_9 are independent of the other 6 variables in the model, while no variable is isolated from the rest of the variables in models M_2 and M_4.

In Table 1, we can also see that the X^2 is less sensitive, than the D^2 test, to differences in model structure. For example, model M_4 is smaller than model M_1 by three

edges and this difference in model structure is reflected in $\pi_d(\alpha)$, $(\alpha = 0.05, 0.1)$, while $\pi_x(\alpha)$ values remain more or less the same.

One may expect that any addition of edges between the set of variables, X_1, X_3, X_5, X_9 and the set of the remaining 6 variables in model M_3 might decrease the π value. But this is not the case as the π values indicate for models M_3 and M_5. We see almost no difference in the π values between the two models. It is interesting to note that connecting the two sets of variables by adding inappropriate edges which are not found in the true model M_0 does not affect the π value. The same phenomenon occurs for models M_6 and M_7.

The simulation result strongly recommends the D^2 test in comparison with the X^2 test in the sense that the D^2 test is more sensitive to the sample size and the difference of the model structure between the true model and a selected model.

Table 1
The proportions of the p_x and p_d values that are less than or equal to $\alpha \in \{0.05, 0.1\}$ out of 500 replications for the p_x and p_d values. The values of the proportions are rounded to the two decimals.

$\alpha = \mathbf{0.05}$	Models		D^2			X^2		
Add	Delete	Model	$n = 50$	$n = 200$	$n = 500$	$n = 50$	$n = 200$	$n = 500$
0	0	0	0.00	0.00	0.00	0.00	0.02	0.01
0	-2	1	0.00	0.02	0.33	0.02	0.06	0.08
0	-3	2	0.00	0.04	0.43	0.00	0.03	0.06
0	-4	3	0.02	0.43	0.93	0.02	0.11	0.40
0	-5	4	0.00	0.13	0.79	0.03	0.05	0.09
2	-4	5	0.00	0.44	0.92	0.02	0.13	0.42
2	-7	6	0.04	0.75	0.99	0.06	0.29	0.71
3	-7	7	0.04	0.73	0.99	0.04	0.27	0.72
0	-12	8	0.35	0.97	1.00	0.08	0.69	0.96
0	-13	9	0.26	0.98	1.00	0.14	0.54	0.88
0	-16	10	0.49	1.00	1.00	0.20	0.79	0.97

$\alpha = \mathbf{0.1}$						
Model	$n = 50$	$n = 200$	$n = 500$	$n = 50$	$n = 200$	$n = 500$
0	0.00	0.00	0.00	0.01	0.03	0.04
1	0.00	0.06	0.49	0.06	0.11	0.16
2	0.01	0.10	0.52	0.02	0.05	0.13
3	0.03	0.57	0.96	0.04	0.23	0.62
4	0.01	0.27	0.91	0.06	0.10	0.17
5	0.01	0.55	0.95	0.03	0.24	0.63
6	0.11	0.85	1.00	0.09	0.45	0.83
7	0.09	0.82	1.00	0.07	0.40	0.86
8	0.48	0.99	1.00	0.16	0.79	0.98
9	0.44	0.99	1.00	0.25	0.69	0.97
10	0.66	1.00	1.00	0.31	0.86	0.99

However, when the sample size is too small (e.g., $n = 50$) relative to the cell size (e.g., $k = 1,024$), the D^2 test may not be very useful unless our interested models are very different from the true model M_0 as the models M_8, M_9, and M_{10}. In such a small-sample situation, it is desirable that we use both of the D^2 and the L_r tests. A comparison between these tests follows below.

Maydeu-Olivares and Joe (2005) suggest using L_r for $r = 1, 2, 3$ for large and sparse 2^J tables. From an additional simulation experiment, where we used D^2 and L_r for testing the goodness-of-fit of the set of the models in Figure 1, we could see that the D^2 test is more powerful than the L_r test for some of the models in the figure and vice versa for the others in the figure.

When the sample size was as small as 25, the power of the L_r test became unstable while it was not the case as for the D^2 test. For instance, the proportion that the p-value of the goodness-of-fit test using L_2 is less than 0.05 was 0.30, 0.03, and 0.24 for the models, M_8, M_9, M_{10}, respectively, out of 500 iterations, while they are 0.10, 0.05, and 0.16 with the D^2 test. Considering the structural discrepancies of these three models from M_0 (see the first three columns in Table 1), the proportions from the D^2 test reflect the structural discrepancies better than the L_2 test when $n = 25$. The proportions from the L_1 test were 0 or close to zero for all the (M_i, n) combinations with $i = 0, 1, \cdots, 10$, $n = 25, 50, 200, 500$ except the combinations, $(M_8, 200)$ and $(M_8, 500)$, for which the proportions were 0.05 and 0.19 respectively.

5 THE DISTRIBUTION OF THE STANDARDIZED D^2

Although the D^2 test is preferred, when the sample-cell ratio (n/k) is less than 1, to the X^2 test in the goodness-of-fit testing due to its sensitivity to the sample size and the discrepancy between model structures, it is important to investigate the dependency of its distribution upon the sample-cell ratio and the level of asymmetry (or non-uniformity) of the cell probabilities, $p_1, \cdots, p_k$, which is defined by

$$\lambda = \sum_{i=1}^{k} (p_i - \frac{1}{k})^2.$$

The first three moments of X^2 and D^2 are obtained under H_0 in Horn (1977) and Mielke and Berry (1985), respectively. The mean and the variance of D^2 are given under H_0, respectively, by

$$E_0(D^2) = -1 \text{ and } V_0(D^2) = 2(k-1)(1-n^{-1}),$$

which have nothing to do with the distribution under H_0, while

$$V_0(X^2) = 2(k-1)(1-n^{-1}) - \frac{k^2}{n} + \sum_{i=1}^{k} \frac{1}{m_{0i}}$$

is not (see the equation in Corollary 2.) But the skewnesses of X^2 and D^2 under H_0 are dependent upon the distribution under H_0 (e.g. Mielke and Berry, 1985, p. 792).

According to Mielke and Berry (1985), the skewness of D^2 is given by

$$\begin{aligned}\gamma = &\sqrt{\frac{2n}{n-1}} k^{-\frac{1}{2}} \{(2 - \frac{7}{n}) - (2 - \frac{6}{n}) k^{-1}\} (1 - \frac{1}{k})^{-\frac{3}{2}} \\ &+ \sqrt{\frac{2n}{n-1}} (k-1)^{-\frac{3}{2}} \sum_{i=1}^{k} p_i^{-1}/n. \end{aligned} \tag{5}$$

Denote the two terms in the right hand side of (5) by A and B, respectively. Then

$$A = \sqrt{\frac{2}{k}} \left(2 + \frac{1}{k} - \frac{6}{n}\right) + \frac{1}{\sqrt{k}} \left(O\Big(\frac{1}{kn}\Big) + O\Big(\frac{1}{k^2}\Big) + O\Big(\frac{1}{n^2}\Big)\right) \tag{6}$$

and, as for B, we will consider asymmetric distributions of p_i's as given by, for $0 < \epsilon < 1$ and $1 \le g < k$,

$$p_1 = \cdots = p_g = \frac{\epsilon}{k} \text{ and } p_{g+1} = \cdots = p_k = \frac{k - \epsilon g}{k(k-g)}. \tag{7}$$

Then we have

$$\sum_{i=1}^{k} p_i^{-1} = g\frac{k}{\epsilon} + (k-g)^2 \frac{1}{1 - \epsilon g/k}. \tag{8}$$

For the cell probabilities in (7), the level of asymmetry is given by

$$\lambda = \frac{g(1-\epsilon)^2}{k(k-g)}. \tag{9}$$

From (5), (8), and (9), we can see that as g, $1 \le g \le k$, increases and ϵ, $0 < \epsilon < 1$, becomes smaller, both γ and λ increases. As a matter of fact, γ increases as λ increases. In particular, when $s (s > 1)$ is chosen so that $g = k/s$ is an integer and k and n satisfy $k = cn$ for some real c, we have

$$B = c\sqrt{\frac{2}{k}} \cdot \left(1 + \frac{3+c}{2k}\right) \cdot \frac{1 + \epsilon(s-2)}{\epsilon(s-\epsilon)} + o\Big(\frac{1}{k^2}\Big). \tag{10}$$

Note that when $\epsilon = 1$, we have $p_1 = \cdots = p_k = 1/k$. Under the distribution as

given by (7), we have from (6) and (10) that

$$\gamma = \sqrt{\frac{2}{k}} \left\{ 2 + \frac{1-6c}{k} + c\left(1 + \frac{3+c}{2k}\right) \frac{1+\epsilon(s-2)}{\epsilon(s-\epsilon)} \right\} + o\left(\frac{1}{k^2}\right). \qquad (11)$$

The last expression implies that as we have more small p_i's (that is, a smaller s ($s > 1$) with a small ϵ value), the skewness (γ) of D^2 gets larger. In other words, as there are more small p_i's, the normal approximation to D^2 gets worse at the tail areas. This phenomenon is observed in Tables 2 and 3. It is also noteworthy in expression (11) that as the contingency table becomes sparser (i.e., $c = k/n$ gets larger), the distribution of D^2 gets more skewed.

Table 2 is obtained under symmetric multinomial distributions for $k = 2^{10}, 2^{14}$, with varying sample sizes (n) as given in the table. Under each symmetric distribution, 10,000 D^2 values were generated and percentiles were obtained from them, and after repeating this 100 times, the means and the variances of the percentiles were obtained.

Table 3 is obtained from the same Monte Carlo experiment as for Table 2 except that a variety of multinomial distributions were used, where each distribution is defined by the vector of the cell probabilities ($p_1, \cdots, p_k$). Here, each p_i was obtained as a ratio given by $u_i / \sum_{j=1}^{k} u_j$ where the random quantities, u_j's, were confined to lie uniformly between 0.05 and 0.95 to avoid too small cell probabilities.

Table 2

Three upper tail percentiles of the standardized D^2 under symmetric multinomials. d_α is the upper $100 \times \alpha$ percentile of the standardized D^2 distribution. 100 d_α values were generated for each pair of k and n and their mean, standard deviation (sd), and skewness measure were computed. $n/k = \infty$ refers to the standard normal distribution. γ is the skewness value as given by (5).

				mean			sd		
	n	n/k	γ	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
$k = 2^{10}$	2^5	2^{-5}	1.519	0.741	2.179	3.602	0.000	0.000	0.134
(1024)	2^6	2^{-4}	0.798	1.448	1.847	2.874	0.000	0.356	0.000
	2^7	2^{-3}	0.442	1.443	1.798	2.572	0.000	0.000	0.137
	2^8	2^{-2}	0.265	1.268	1.730	2.517	0.030	0.085	0.048
	2^9	2^{-1}	0.177	1.294	1.691	2.452	0.043	0.032	0.056
	∞	∞		**1.282**	**1.645**	**2.327**			
$k = 2^{14}$	2^6	2^{-8}	2.872	2.500	2.500	2.500	0.000	0.000	0.000
(16384)	2^7	2^{-7}	1.441	0.715	2.135	3.555	0.000	0.000	0.000
	2^8	2^{-6}	0.730	1.423	2.039	2.840	0.000	0.239	0.000
	2^9	2^{-5}	0.376	1.421	1.775	2.508	0.000	0.000	0.091
	2^{10}	2^{-4}	0.199	1.258	1.709	2.477	0.048	0.086	0.047
	2^{11}	2^{-3}	0.110	1.302	1.678	2.398	0.042	0.027	0.043
	2^{12}	2^{-2}	0.066	1.291	1.666	2.364	0.021	0.026	0.040
	2^{13}	2^{-1}	0.044	1.289	1.663	2.354	0.022	0.022	0.041

Table 3

Three upper tail percentiles of the standardized D^2. d_α is the upper $100 \times \alpha$ percentile of the standardized D^2 distribution. 100 d_α values were generated for each pair of k and n and their mean, standard deviation (sd), and skewness measure were computed. $n/k = \infty$ refers to the standard normal distribution.

			mean			sd		
	n	n/k	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
$k = 2^7$	2^4	2^{-3}	1.262	1.846	3.345	0.026	0.042	0.117
(=128)	2^5	2^{-2}	1.286	1.809	3.022	0.022	0.036	0.079
	2^6	2^{-1}	1.301	1.771	2.789	0.022	0.028	0.062
	2^7	1	1.306	1.749	2.675	0.019	0.030	0.059
	∞	∞	**1.282**	**1.645**	**2.327**			
$k = 2^{10}$	2^4	2^{-6}	1.390	1.908	3.999	0.029	0.048	0.166
(=1024)	2^5	2^{-5}	1.264	1.892	3.489	0.029	0.043	0.123
	2^6	2^{-4}	1.293	1.830	3.119	0.021	0.032	0.083
	2^7	2^{-3}	1.304	1.778	2.822	0.022	0.025	0.063
	2^8	2^{-2}	1.304	1.736	2.609	0.022	0.029	0.058
	2^9	2^{-1}	1.297	1.702	2.503	0.020	0.027	0.044
$k = 2^{14}$	2^5	2^{-9}	-0.174	-0.174	4.598	0.000	0.000	0.164
(=16384)	2^6	2^{-8}	1.365	1.908	4.002	0.021	0.046	0.136
	2^7	2^{-7}	1.265	1.901	3.469	0.023	0.043	0.105
	2^8	2^{-6}	1.300	1.836	3.071	0.024	0.036	0.081
	2^9	2^{-5}	1.307	1.778	2.769	0.022	0.027	0.056
	2^{10}	2^{-4}	1.305	1.725	2.555	0.024	0.029	0.052
	2^{11}	2^{-3}	1.298	1.692	2.452	0.020	0.026	0.041
	2^{12}	2^{-2}	1.290	1.667	2.396	0.018	0.023	0.043
	2^{13}	2^{-1}	1.299	1.674	2.383	0.019	0.023	0.043

We denote by d_α the upper $100 \times \alpha$-percentile of the curve of the standardized D^2 and by z_α for the standard normal curve. The d_α values are listed in Tables 2 and 3 for $\alpha = 0.01, 0.05, 0.1$. We can see in Table 2 that the d_α values appear, as expected, farther away from z_α as the skewness index (γ) increases unless the sample size is too small relative to the table size (k). For example, when $k = 2^{10}$, we see in Table 2 that the d_α values change abruptly as n comes down from 2^6 to 2^5; and a similar phenomenon takes place, when $k = 2^{14}$, as n comes down from 2^8 to 2^7 and 2^6. In particular, $d_{0.01}$, $d_{0.05}$, and $d_{0.1}$ are all equal to 2.500 when $(k, n) = (2^{14}, 2^6)$. It is due to an extreme sparseness of the contingency table. In the Monte Carlo experiment, all the 100 distributions that were generated under this sparseness yielded the same set of upper 0.99, 0.95, 0.9, 0.75, 0.5, 0.25, 0.1, 0.05, 0.01 percentile points, -0.351 for the first six points and 2.500 for the rest three. -0.351 corresponds to the case that the cell frequencies are at most 1 and 2.500 corresponds to the case that only one cell frequency is 2 and the others are at most 1. We will call such a phenomenon as this due to an extreme sparseness an hyper-sparseness phenomenon.

The skewness index (γ) is not given in Table 3 since different sets of p_i's may pro-

duce different index values for fixed values of k and n. However, we can see in the table that the $d_{0.01}$ values are larger on average for asymmetric multinomials than for symmetric multinomials. This phenomenon is a reflection of the expression (11). We also see another hyper-sparseness phenomenon in Table 3 at the row $(k, n) = (2^{14}, 2^5)$.

6 CONCLUDING REMARKS

Large scale modelling is not unusual nowadays in many research fields such as bio-sciences, cognitive science, management sciences, etc. In the simulation experiments, we considered Bayesian network models which are a useful tool for representing causal relationship among variables. When the variables are not causally related, log-linear modelling is an appropriate method for contingency table data.

Whether it is a Bayesian network model or a log-linear model, we can express a model in a factorized form. Suppose that a model for a set of categorical random variables, $X_v,\ v \in V$, is factorized as

$$P(x) = \prod_{A \in \mathcal{V}} P_A(x_A) \tag{12}$$

where $\mathcal{V}$ is a set of subsets of V. We will denote the marginal table on the subset of variables, $X_v,\ v \in A$, for $A \subset V$, by τ_A and call it the configuration on A. Then we can say that the configurations, $\tau_A,\ A \in \mathcal{V}$, are sufficient for the parameters of the model expressed in (12). This means that a large sparse contingency table may not cause a trouble in parameter estimation as long as none of the configurations are sparse. In this respect, the statistic, L_r, which is a function of multivariate moments of the variables $X_v,\ v \in V$, is a reasonable one for goodness-of-fit testing when the configurations in (12) satisfy that $|A| \leq r$. This is one of the reasons why the performance of the L_r test is subject to the models which are selected for the goodness-of-fit testing with a given contingency table.

The key idea behind the D^2 method for goodness-of-fit testing is the same as for the goodness-of-fit testing using the Pearson chi-square statistic, X^2, and for the testing using the likelihood ratio statistic (G^2). The latter two statistics represent a measure of distance between the data and the model which is selected by a model builder, where the measure is standardized by a chi-squared distribution. In the proposed D^2 method, we construct a reference distribution, instead of using the chi-squared distribution, to create a distance measure which is similar to the P-value of a test, such as p_x and p_d in (4). The simulation experiment strongly recommends to use the D^2 test (i.e., p_d values), in comparison with the X^2 test, for goodness-of-fit testing with a large sparse contingency table due to its high sensitivity to the sample size and model discrepancy.

This observation leads us to recommend using both of the D^2 and the L_r tests when the contingency table is large and sparse. Not knowing the true model for a given contingency table, it is desirable to use both of the tests, since
(i) the L_r test seems to be more powerful than the D^2 test when $|A| \leq r$ and the contingency table is of 2^J type and not too sparse,
(ii) the D^2 test seems to be more powerful when $|A| > r$, and
(iii) as the contingency table becomes sparser, the L_r test becomes unstable while the D^2 test remains stable.

APPENDIX A: PROOF OF THEOREM 1

The proof is a straightforward application of algebra.

$$\begin{aligned} V(D^2) &= V(X^2) - \sum \frac{n_i}{m_{0i}}) \\ &= V(X^2) + V(\sum \frac{n_i}{m_{0i}}) - 2\sum Cov(\frac{n_i}{m_{0i}},\ X^2). \end{aligned} \tag{13}$$

$$V(\sum \frac{n_i}{m_{0i}}) = n\sum \frac{p_i q_i}{m_{0i}^2} - n\sum_i \sum_j \frac{p_i p_j}{m_{0i} m_{0j}} + n\sum \frac{p_i^2}{m_{0i}^2}. \tag{14}$$

As for the last term in (13), we have

$$Cov(\frac{n_i}{m_{0i}},\ X^2) = E(X^2 \frac{n_i}{m_{0i}}) - E(X^2)E(\frac{n_i}{m_{0i}}). \tag{15}$$

After a simple algebra, we have

$$E(X^2) = n\sum \frac{p_i q_i}{m_{0i}} + n^2 \frac{p_i^2}{m_{0i}} - n, \tag{16}$$

$$\begin{aligned} E(X^2 \frac{n_i}{m_{0i}}) &= n(n-1)\sum_{i \neq j} \frac{p_i p_j}{m_{0i} m_{0j}} + n(n-1)(n-2)\sum_{i \neq j} \frac{p_i p_j^2}{m_{0i} m_{0j}} - n^2 \frac{p_i}{m_{0i}} \\ &\quad + n p_i \frac{1 + 3(n-1)p_i + (n-1)(n-2)p_i^2}{m_{0i}^2}. \end{aligned} \tag{17}$$

Substituting equations (16) and (17) into (15) and then (14) and (15) into equation (13) yields the desired result of the theorem. □

APPENDIX B: DAG MODEL OF CATEGORICAL VARIABLES

A DAG model of a set of variables, $X_v,\ v \in V$, is a probability model whose model structure can be represented by a directed acyclic graph, $\mathcal{G}$ say, which consists of vertices and directed edges or arrows between vertices. For a subset A of V, $X_A = (X_v)_{v \in A}$, and for two subsets A and B of V, we denote the conditional probability $P(X_B = x_B | X_A = x_A)$ by $P_{B|A}(x_B|x_A)$.

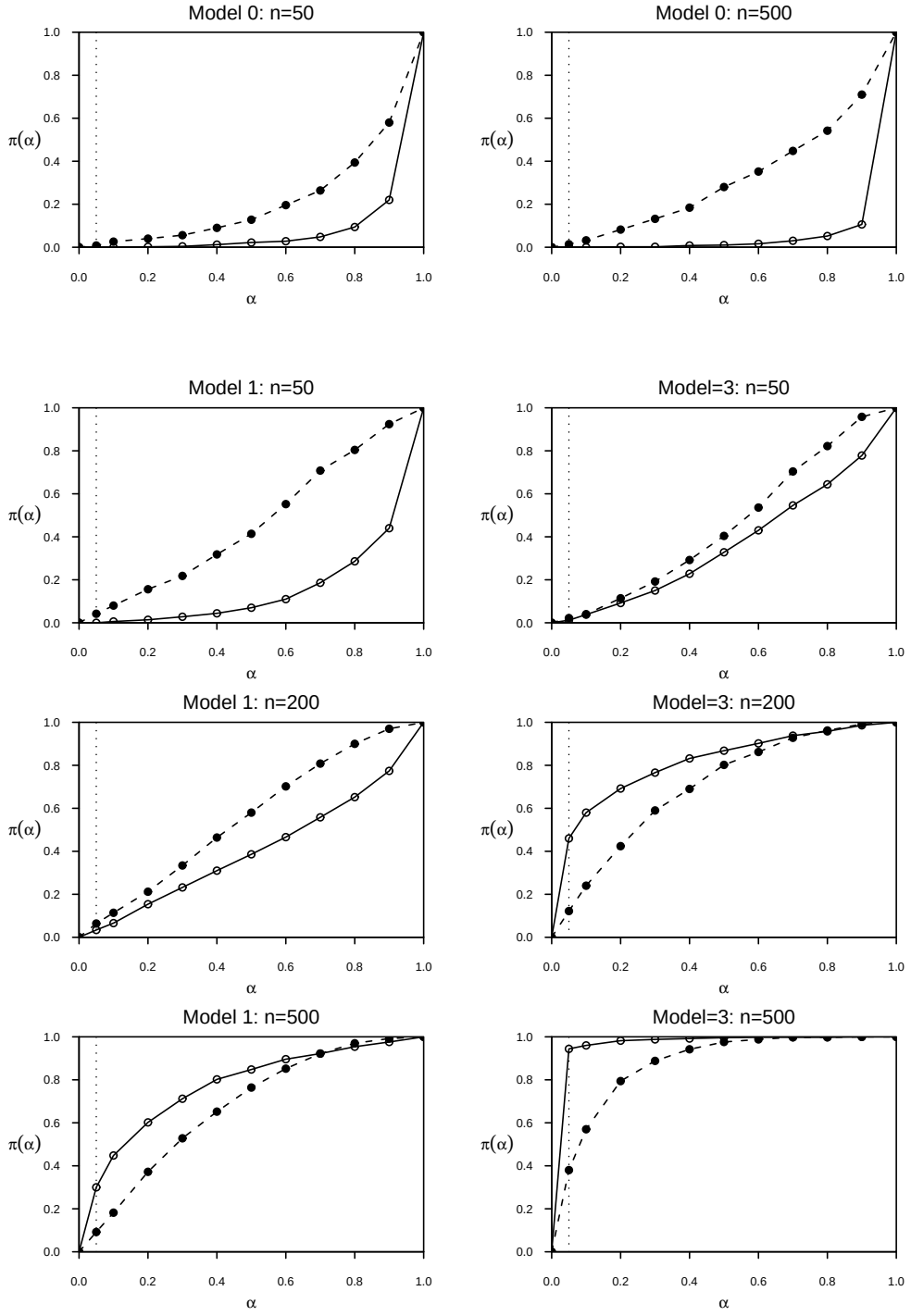
If there is an arrow from vertex u to v, we call u a parent vertex of v. Since a DAG is acyclic, every vertex in a DAG has a unique set of parent vertices. We denote by $pa(v)$ the set of the parent vertices of v. Then we can represent the joint probability of $(X_v)_{v \in V}$ as the product of conditional probabilities as in

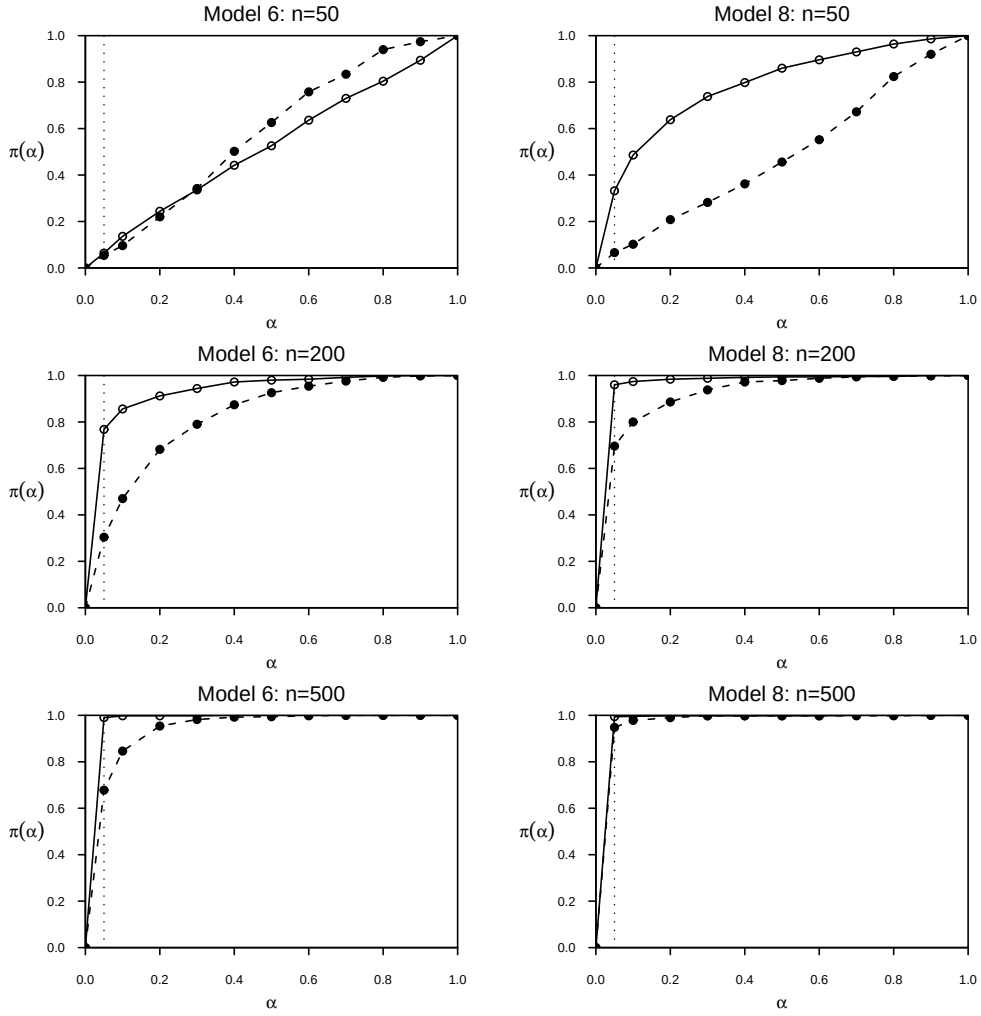
$$P(X_V = x_V) = \prod_{v \in V} P_{v|pa(v)}(x_v | x_{pa(v)})$$

where $pa(v) = \emptyset$ if vertex v has no parent vertex.

APPENDIX C: COMPARISON OF π_x AND π_d VALUES

Solid curves with circles (○) are for π_d and dashed curves with bullets (●) for π_x. The dotted vertical line represents that $\alpha = 0.05$.





References

Agresti, A., 1990. Categorical Data Analysis. John Wiley & Sons: New York.

Berry, K.J., Mielke, Jr. P.W., 1988. Monte Carlo comparisons of the asymptotic Chi-square and likelihood-ratio tests with the nonasymptotic Chi-square test for sparse $r \times c$ tables. Psychological Bulletin 103(2), 256-264.

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

Cochran, W.G., The χ^2 test of goodness-of-fit. Ann. Math. Statist. 23, 315-345.

Cressie, N., Read, T.R.C., 1984. Multinomial goodness-of-fit tests. J. Roy. Statist. Soc. Ser. B 46(3), 440-464.

Fienberg, S.E., 1979. The use of Chi-squared statistics for categorical data problems. J. Roy. Statist. Soc. Ser. B 41, 54-64.

Horn, S.D., 1977. Goodness-of-fit tests for discrete data: A review and an application to a health impairment scale. Biometrics 33, 237-248.

Johnson, N.L., Kotz, S., 1969. Distributions in Statistics: Discrete Distributions. Houghton Mifflin Company, New York.

Koehler, K.J., 1986. Goodness-of-fit tests for log-linear models in sparse contingency tables. J. Amer. Statist. Assoc. 81(394), 483-493.

Koehler, K.J., Larntz, K., 1980. An empirical investigation of goodness-of-fit statistics for sparse multinomials. J. Amer. Statist. Assoc. 75(370), 336-344.

Lancaster, H.O., 1969. The Chi-squared Distribution. Wiley, New York.

Maydeu-Olivares, A., Joe, H., 2005. Limited- and full-information estimation and goodness-of -fit testing in 2^n contingency tables: a unified framework. J. Amer. Statist. Assoc. 100(471), 1009-1020.

Mielke, Jr. P.W., 1997. Some exact and nonasymptotic analyses of discrete goodness-of-fit and r-Way contingency tables. In: Johnson NL, Balakrishman N (Eds), Advances in the Theory and Practice of Statistics: A Volume in Honor of Samuel Kotz, Wiley, New York, 179-192.

Mielke, Jr. P.W., Berry, K.J., 1985. Non-asymptotic inferences based on the chi-square statistic for r by c contingency tables. J. Statist. Plann. Inference 12, 41-45.

Mislevy, R.J., 1994. Evidence and inference in educational assessment. Psychometrika 59(4), 439-483.

Moore, D.S., Recent developments in chi-square tests for goodness-of-fit. Mimeo. (Dept. of Statistics, Purdue University, Lafayette, IN, 1976).

Pearson, K., 1900. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philosophy Magazine Series 50(5), 157-172.

Read, T.R.C., 1984. Small-sample comparisons for the power divergence goodness-of-fit statistics. J. Amer. Statist. Assoc. 79(388), 929-935.

Watson, G.S., 1959. Some recent results in chi-square goodness-of-fit tests. Biometrics 15, 440-468.

Wermuth, N., Lauritzen, S.L., 1983. Graphical and recursive models for contingency tables. Biometrika 70(3), 537-552.

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