Computational Statistics & Data Analysis 5 (1987) 9–21
North-Holland

An empirical investigation of some effects of sparseness in contingency tables

Alan AGRESTI and Ming-Chung YANG

Department of Statistics, University of Florida, Gainesville, FL 32611, USA

Received 21 April 1986

Abstract: A simulation study investigates some effects of having 'sparse' categorical data, for which the ratio of the sample size to the number of cells is relatively small. In this study, the true cell proportions satisfy the uniform association model for ordinal variables. Conclusions include the following: (1) For direct testing of the model, the distribution of the Pearson goodness-of-fit statistic is closer to the asymptotic chi-squared distribution than is the distribution of the likelihood-ratio statistic. (2) For comparing two unsaturated loglinear models (such as in testing independence under the assumption that a particular model holds), it is usually preferable to compare likelihood-ratio statistics rather than Pearson statistics. (3) A beneficial aspect of sparseness is that the power of certain single-degree-of-freedom test statistics tends to increase as the table becomes more sparse, for a fixed sample size. (4) The common practice of adding constants to empty cells can cause havoc with the distribution of the Pearson statistic.

Keywords: Independence, Likelihood-ratio statistic, Ordinal variables, Pearson chi-squared statistic, Uniform association model.

1. Introduction

Contingency tables are said to be *sparse* when the ratio of the sample size to the number of cells is relatively small. For Poisson or multinomial sampling models, standard asymptotic theory is valid as the cell expected frequencies go to infinity, for a fixed number of cells. These asymptotic approximations may be quite poor for sparse tables, even if the total sample size is quite large. For examples of investigations into the adequacy of chi-squared approximations for goodness-of-fit statistics for multinomial distributions, see Larntz [15] and Koehler and Larntz [16] and the references in those articles.

The dangers inherent in analyzing sparse contingency tables have been well advertised. Less attention has been given to ways that sparseness may not be harmful, or indeed may even be beneficial, to inference-making for categorical data. This article considers two related analyses for which this can be the case.

Results are given of a simulation study that investigates the effects of sparseness on these analyses. In this study, the effects are studied by increasing the number of cells in the table, for a fixed sample size and underlying distribution.

Let $\{n_i\}$ denote observed cell counts and let $\{\hat{m}_i\}$ denote maximum likelihood (ML) estimates of corresponding expected values $\{m_i\}$ for a particular model. Let

$$G^2(M) = 2\sum n_i \log(n_i/\hat{m}_i) \quad \text{and} \quad X^2(M) = \sum (n_i - \hat{m}_i)^2/\hat{m}_i$$

denote the likelihood-ratio and Pearson statistics for testing the goodness of fit of model M. Let M_1 and M_2 denote two models such that M_1 is a special case of M_2. The statistic $G^2(M_1 \mid M_2) = G^2(M_1) - G^2(M_2)$ is used to test the fit of M_1, given that M_2 holds. It is commonly employed for comparing models in model-building procedures. For loglinear models, from Bishop et al. [1, p. 126],

$$G^2(M_1|M_2) = 2\sum n_i \log(\hat{m}_{i2}/\hat{m}_{i1}) = 2\sum \hat{m}_{i2}\log(\hat{m}_{i2}/\hat{m}_{i1}),$$

where $\{\hat{m}_{ik}\}$ refers to model M_k, $k = 1, 2$.

Suppose we want to test a hypothesis that can be expressed as the condition that some model M_1 holds. In practice it is usually possible to imbed the model in a slightly more complex model M_2 that reflects the pattern of departures from the hypothesis that one expects. The first purpose of this paper is to show that, even for fairly sparse data, the standard asymptotic approximation for the statistic $G^2(M_1 \mid M_2)$ can hold quite well. The estimates $\{\hat{m}_{i2}\}$ are functions of cell counts in the lower-dimensional marginal tables that are the minimal sufficient configurations for model M_2, and these tables are much less sparse than the full table $\{n_i\}$. Thus, although the chi-squared approximation for the statistic $G^2(M_1)$ may be quite poor for sparse tables, it is possible that the null distribution of $G^2(M_1 \mid M_2)$ may be well approximated by the usual reference chi-squared distribution. Several studies have noted that $X^2(M)$ follows a chi-squared distribution more closely than does $G^2(M)$, for sparse tables. However, $X^2(M_1) - X^2(M_2)$ depends on the cell counts as well as the sufficient statistics for model M_2, so its behavior is more uncertain. The simulation study in this article investigates the behaviors of $G^2(M_1 \mid M_2)$ and $X^2(M_1) - X^2(M_2)$, for a particular pair of models, as a function of the degree of sparseness in the table.

The second purpose of the paper is to note that sparseness need not adversely affect inferences involving certain single-degree-of-freedom statistics that describe characteristics of the entire table. As the number of cells increases, improved normal approximations resulting from increasing the number of approximately additive effects of similar order of magnitude on the statistic may counterbalance the increase in sparseness. The specific types of statistics we have in mind here involve ordinal variables, for which it is usually plausible to hypothesize underlying continuous variables. The refinement of the measurement scales, though increasing the degree of sparseness, can improve the agreement between the descriptive and inferential conclusions reached with ordinal categorical data methods and with analogous methods for continuous variables. Asymptotic

approximations usually are valid for smaller sample sizes in the continuous case, so sparseness may even be beneficial in this respect. In addition, refining the measurement scales can result in improved power for detecting associations, through eliminating 'tied' pairs of observations that provide no information about the direction of association.

The simulation study described in the next section considers these two types of analyses in the context of an important association model for ordinal variables. As a by-product of this study, it is noted that the standard practice of adding a small constant to cells in sparse tables before fitting models can have a severe impact on asymptotic approximations for goodness-of-fit statistics.

The empirical results obtained in this study are not surprising or even especially innovative, as conjectures such as the one regarding $G^2(M_1 \mid M_2)$ already appear in the literature (see, e.g., Haberman [12, p. 326] and Imrey et al. [13]). However, some of these results illustrate the utility of the limit theorems in Haberman [11] for sparse tables, and suggest that not enough attention has been paid to that paper in the contingency table literature.

2. Sparse uniform association

Let t denote the number of cells in the contingency table, and let n denote the total sample size. The issues discussed in Section 1 will be considered in this article only for two-way tables, with $t = rc$ cells. We illustrate our arguments using two loglinear models for the r-by-c table, the independence (I) model

$$\log m_{ij} = \lambda + \lambda_i^X + \lambda_j^Y, \tag{2.1}$$

and the linear-by-linear association model

$$\log m_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \beta u_i v_j \tag{2.2}$$

for ordinal variables, in which fixed monotone scores $\{u_i\}$ and $\{v_j\}$ are assigned to the rows and to the columns. We shall use the version of model (2.2) having scores $\{u_i = i\}$ and $\{v_j = j\}$, called the uniform association (U) model [6] because it implies uniformity of local odds ratios $\{m_{ij}m_{i+1,j+1}/m_{i,j+1}m_{i+1,j} = \exp(\beta)\}$. Goodness-of-fit statistics have degrees of freedom $\mathrm{df} = (r-1)(c-1)$ for the I model and $\mathrm{df} = (r-1)(c-1) - 1$ for the U model.

We chose model (2.2) as the alternative to independence because it corresponds to a family of distributions that contains a discrete analog of the bivariate normal distribution as a special case [8]. In particular, the U model describes well a bivariate normal distribution that has been categorized by selecting cutpoints that are equally-spaced. Thus, through the addition of a single parameter to model (2.1), we obtain a model that describes departures from independence of the type often expected for ordinal variables.

Sufficient statistics for fitting the U model are the row totals $\{n_{i+}\}$, the column totals $\{n_{+j}\}$, and $\Sigma\Sigma u_i v_j n_{ij}$ (or, equivalently, the correlation). The

likelihood-ratio statistic $G^2(\mathrm{I}\,|\,\mathrm{U}) = G^2(\mathrm{I}) - G^2(\mathrm{U})$ for testing independence, assuming that the U model holds, depends on the data only through these sufficient statistics. The first argument in Section 1 is that the null distribution of $G^2(\mathrm{I}\,|\,\mathrm{U})$ should be reasonably well approximated by the χ_1^2 distribution if the $\{m_{i+}\}$ and $\{m_{+j}\}$ are not particularly small, even though the distribution of $G^2(\mathrm{I})$ may be poorly approximated by the $\chi^2_{(r-1)(c-1)}$ distribution.

Secondly, we consider inference regarding the association parameter β in model (2.2). Let $\hat{\beta}$ denote the ML estimate of β and let $\hat{\sigma}/\sqrt{n}$ denote its estimated asymptotic standard error obtained from the inverse of the estimated information matrix. We shall see that the normal approximation for the asymptotic distribution of $\sqrt{n}(\hat{\beta} - \beta)/\hat{\sigma}$ is not adversely affected by sparseness. Also, we observe that when the U model holds with $\beta \neq 0$, the power of the test of independence ($\beta = 0$) based on the statistic $z_{\mathrm{U}} = \sqrt{n}\,\hat{\beta}/\hat{\sigma}$ tends to increase as the degree of sparseness increases, for a fixed underlying distribution. That is, the improved precision of measurement can result in an increased noncentrality value of $n\beta^2/\sigma^2$. In this regard, it follows from Cox and Hinkley [3, pp. 322–324] that we should expect similar results from two other tests. Specifically, this 'ML test' has the same asymptotic efficacy for local alternatives as the likelihood-ratio test based on $G^2(\mathrm{I}\,|\,\mathrm{U})$ or a 'score test' based on the derivative of the log likelihood evaluated at the null hypothesis. For the linear-by-linear association model, it is easily seen that the score is

$$S = \sum\sum u_i v_j p_{ij} - \Big(\sum u_i p_{i+}\Big)\Big(\sum v_j p_{+j}\Big),$$

where $p_{ij} = n_{ij}/n$. A related statistic for testing independence is $z_{\mathrm{S}} = \sqrt{n}\,S/\hat{\sigma}_{\mathrm{S}}$, where

$$\hat{\sigma}_{\mathrm{S}}^2 = \sum\sum (u_i - \hat{\mu}_X)^2 (v_j - \hat{\mu}_Y)^2 p_{ij} - S^2$$

with $\hat{\mu}_X = \sum u_i p_{i+}$ and $\hat{\mu}_Y = \sum v_j p_{+j}$. Here, $\hat{\sigma}_{\mathrm{S}}^2$ is the estimated asymptotic variance of $\sqrt{n}\,[S - E(S)]$, where

$$E(S) = [(n-1)/n]\Big[\sum\sum u_i v_j \pi_{ij} - \Big(\sum u_i \pi_{i+}\Big)\Big(\sum v_j \pi_{+j}\Big)\Big],$$

with $\pi_{ij} = E(p_{ij}) = m_{ij}/n$.

For our simulation study we used population cross-classification tables that satisfy the U model perfectly and correspond to underlying normal distributions with correlations $\rho = 0$ and $\rho = 0.2$. For sample sizes $n = 50$ and $n = 100$, and for table sizes 2×3, 4×4, 6×6, and 10×10, we generated 5000 multinomial distributions. The purpose was to consider, for fixed n, how increasing the sparseness (through increasing $t = rc$) affected the sampling distributions of various statistics. For the underlying marginal normal $\mathrm{N}(\mu, \sigma)$ distributions, the cutpoints were selected at μ when $r = 2$, at $\mu \pm 0.6\sigma$ when $c = 3$, at μ and $\mu \pm 0.8\sigma$ when $r = c = 4$, at μ, $\mu \pm 0.6\sigma$, $\mu \pm 1.2\sigma$ when $r = c = 6$, and at μ, $\mu \pm 0.4\sigma$, $\mu \pm 0.8\sigma$, $\mu \pm 1.2\sigma$, $\mu \pm 1.6\sigma$ when $r = c = 10$. For each randomly generated table we calculated $G^2(\mathrm{I})$, $G^2(\mathrm{U})$, $G^2(\mathrm{I}\,|\,\mathrm{U})$, corresponding Pearson statistics, and the squares of z_{U}, $z_{\mathrm{Ua}} = \sqrt{n}(\hat{\beta} - \beta)/\hat{\sigma}$, z_{S}, and $z_{\mathrm{Sa}} = \sqrt{n}(S -$

$ES)/\hat{\sigma}_S$. Upon completion of the 5000 simulations, we analyzed the asymptotic approximations in the upper tail for these statistics, by giving sample proportion estimates for probabilities of exceeding the $100(1-\alpha)$ percentage point of the relevant chi-squared distribution, for $\alpha = 0.01$, 0.05, 0.10, and 0.25. We used the GGMTN routine in IMSL for random generation of multinomial distributions, on the IBM 3081 mainframe computer. Assuming the adequacy of the generator, the standard error for these estimates is about 0.006 when the true proportions are about 0.25, and it is about 0.0014 when the true proportions are about 0.01.

3. Results

The results of this investigation are reported in the following subsections.

3.1. Effect of sparseness on G^2

Consider first the case in which independence holds. Table 1 gives the performance of $G^2(\mathrm{I})$, $G^2(\mathrm{U})$, and $G^2(\mathrm{I}\,|\,\mathrm{U})$, for the four table sizes and the two sample sizes. For goodness-of-fit testing of a completely specified multinomial distribution, Koehler and Larntz [14] observed that the distribution of G^2 is generally not well approximated by its asymptotic chi-squared distribution when $n/t < 5$. Similar results are observed here for the I and U models. The $G^2(\mathrm{I})$ and $G^2(\mathrm{U})$ goodness-of-fit statistics behave adequately at both sample sizes for 2×3 tables, but deteriorate dramatically for larger tables. These statistics are highly liberal except for the 10×10 case with $n = 50$, when they are highly conservative. This is consistent with the observation made by Koehler and Larntz that the

Table 1
Proportion of times likelihood-ratio statistic exceeds chi-squared percentage point, when there is independence

Statistic	$r \times c$	$n = 50$, $\alpha =$				$n = 100$, $\alpha =$			
		0.01	0.05	0.10	0.25	0.01	0.05	0.10	0.25
$G^2(\mathrm{I})$	2× 3	0.013	0.051	0.107	0.261	0.011	0.058	0.108	0.258
	4× 4	0.027	0.102	0.173	0.371	0.016	0.075	0.131	0.304
	6× 6	0.024	0.126	0.236	0.501	0.031	0.129	0.224	0.445
	10×10	0.000	0.000	0.002	0.028	0.013	0.086	0.188	0.464
$G^2(\mathrm{U})$	2× 3	0.010	0.050	0.102	0.255	0.011	0.055	0.104	0.265
	4× 4	0.030	0.102	0.178	0.374	0.016	0.074	0.137	0.303
	6× 6	0.024	0.125	0.237	0.506	0.031	0.128	0.227	0.449
	10×10	0.000	0.000	0.003	0.029	0.013	0.085	0.187	0.462
$G^2(\mathrm{I}\,|\,\mathrm{U})$	2× 3	0.013	0.054	0.108	0.263	0.011	0.054	0.103	0.250
	4× 4	0.011	0.054	0.111	0.267	0.012	0.052	0.100	0.241
	6× 6	0.016	0.060	0.109	0.249	0.012	0.054	0.103	0.257
	10×10	0.009	0.055	0.104	0.254	0.010	0.052	0.109	0.253

mean and variance of G^2 tend to be smaller than the chi-squared moments when $n/t < 0.5$ and larger when $n/t > 1$.

Unlike $G^2(\mathrm{I})$ and $G^2(\mathrm{U})$, $G^2(\mathrm{I}\,|\,\mathrm{U})$ behaves quite well for all table sizes, even when n is only 50. This is not surprising, since the $\{m_{i+}\}$ or $\{m_{+j}\}$ are of moderate size for these table sizes. More generally, whenever the difference between the dimensions of two models converges to a constant d as $t \to \infty$, Theorem 4 in Haberman [11] suggests that $G^2(M_1 \mid M_2)$ may have a limiting null χ^2_d distribution even if t grows proportionally to n.

3.2. *Effect of sparseness on* X^2

Other investigators (e.g. Haberman [12, p. 325] and Larntz [14]) have indicated that the Pearson statistic performs better than the likelihood-ratio statistic for sparse tables. For direct tests of a model, this was true in this study as well. Table 2 contains results for $X^2(\mathrm{I})$ and $X^2(\mathrm{U})$, when there is independence. The Pearson statistics were adequate whenever $n/t \geq 1$; that is, for all cases except the 10×10 table with $n = 50$. Unlike $G^2(\mathrm{I}\,|\,\mathrm{U})$, however, $X^2(\mathrm{I}) - X^2(\mathrm{U})$ depends on the cell counts as well as the sufficient statistics, and its behavior resembled that of $X^2(\mathrm{I})$ and $X^2(\mathrm{U})$. Its performance was poorer than that of $G^2(\mathrm{I}\,|\,\mathrm{U})$ for the most sparse tables.

One would expect the Pearson statistic of the form $X^2(M_1 \mid M_2) = \Sigma(\hat{m}_{i2} - \hat{m}_{i1})^2/\hat{m}_{i1}$ (Haberman [10, p. 108]), to perform better than $X^2(M_1) - X^2(M_2)$ for

Table 2
Proportion of times Pearson statistic exceeds chi-squared percentage point, when there is independence

Statistic	$r \times c$	$n = 50$, $\alpha =$				$n = 100$, $\alpha =$			
		0.01	0.05	0.10	0.25	0.01	0.05	0.10	0.25
$X^2(\mathrm{I})$	2× 3	0.008	0.046	0.100	0.255	0.009	0.055	0.106	0.255
	4× 4	0.010	0.052	0.104	0.265	0.008	0.048	0.103	0.267
	6× 6	0.008	0.043	0.093	0.255	0.010	0.048	0.095	0.253
	10×10	0.010	0.036	0.070	0.188	0.010	0.047	0.096	0.245
$X^2(\mathrm{U})$	2× 3	0.009	0.047	0.099	0.254	0.011	0.054	0.104	0.265
	4× 4	0.009	0.052	0.107	0.273	0.009	0.047	0.106	0.270
	6× 6	0.009	0.046	0.095	0.258	0.009	0.049	0.096	0.258
	10×10	0.012	0.041	0.074	0.185	0.011	0.048	0.093	0.250
$X^2(\mathrm{I}) - X^2(\mathrm{U})$	2× 3	0.010	0.048	0.101	0.257	0.010	0.051	0.099	0.249
	4× 4	0.009	0.049	0.097	0.242	0.012	0.045	0.091	0.235
	6× 6	0.016	0.060	0.107	0.231	0.011	0.052	0.105	0.245
	10×10	0.077	0.159	0.222	0.344	0.034	0.097	0.149	0.282
$X^2(\mathrm{I}\,|\,\mathrm{U})$	2× 3	0.007	0.051	0.101	0.253	0.010	0.050	0.102	0.254
	4× 4	0.010	0.053	0.108	0.267	0.012	0.050	0.099	0.240
	6× 6	0.015	0.058	0.108	0.249	0.011	0.054	0.104	0.258
	10×10	0.011	0.059	0.108	0.257	0.011	0.052	0.109	0.252

comparing models for sparse tables. Theorem 4 in [11] suggests that $X^2(M_1 \mid M_2)$ is asymptotically equivalent to $G^2(M_1 \mid M_2)$ for certain sequences of sparse tables in which M_1 holds, when the difference between the dimensions of the two models converges. Table 2 indicates that for these simulations $X^2(M_1 \mid M_2)$ behaves much better than $X^2(M_1) - X^2(M_2)$ and, in fact, very much like $G^2(M_1 \mid M_2)$.

3.3. *Effects of sparseness on power*

Table 3 contains estimated powers for $X^2(\mathrm{I})$, $X^2(\mathrm{I}\,|\,\mathrm{U})$, $G^2(\mathrm{I}\,|\,\mathrm{U})$, z_{U}, and z_{S} when the U model holds and there is an underlying normal distribution having correlation 0.2. We report $X^2(\mathrm{I})$ rather than $G^2(\mathrm{I})$ for the direct test of the independence model, since the poor null approximation for $G^2(\mathrm{I})$ implies that power comparisons with it are less meaningful. The statistics $X^2(\mathrm{I}\,|\,\mathrm{U})$, $G^2(\mathrm{I}\,|\,\mathrm{U})$, z_{U}, and z_{S} behave very much alike, the apparent slight reduction in power for z_{U} explained by its null approximation being slightly more conservative (see Table 4). Note that these statistics become *more* powerful for sparser tables (i.e., as t

Table 3
Estimated powers, when the U model holds and there is an underlying bivariate normal distribution with correlation 0.2

Statistic	$r \times c$	$n = 50$, $\alpha =$				$n = 100$, $\alpha =$			
		0.01	0.05	0.10	0.25	0.01	0.05	0.10	0.25
$X^2(\mathrm{I})$	2× 3	0.033	0.131	0.224	0.432	0.088	0.234	0.349	0.552
	4× 4	0.022	0.095	0.168	0.371	0.049	0.162	0.265	0.482
	6× 6	0.015	0.070	0.138	0.333	0.032	0.125	0.203	0.416
	10×10	0.014	0.054	0.101	0.235	0.026	0.091	0.153	0.341
$X^2(\mathrm{I}\,|\,\mathrm{U})$	2× 3	0.053	0.177	0.279	0.473	0.130	0.305	0.417	0.621
	4× 4	0.086	0.245	0.348	0.553	0.193	0.416	0.540	0.736
	6× 6	0.099	0.263	0.376	0.583	0.249	0.474	0.611	0.774
	10×10	0.116	0.282	0.396	0.598	0.276	0.509	0.634	0.795
$G^2(\mathrm{I}\,|\,\mathrm{U})$	2× 3	0.056	0.180	0.282	0.473	0.135	0.308	0.420	0.621
	4× 4	0.094	0.249	0.350	0.552	0.197	0.418	0.541	0.735
	6× 6	0.103	0.263	0.375	0.579	0.253	0.475	0.610	0.771
	10×10	0.116	0.279	0.391	0.593	0.273	0.505	0.631	0.790
$\sqrt{n}\,\hat{\beta}/\hat{\sigma}_{\hat{\beta}}$	2× 3	0.035	0.159	0.263	0.469	0.116	0.295	0.411	0.619
	4× 4	0.054	0.220	0.330	0.547	0.171	0.401	0.529	0.734
	6× 6	0.065	0.229	0.353	0.573	0.212	0.457	0.599	0.771
	10×10 [a]	0.117	0.283	0.388	0.618	0.248	0.501	0.637	0.805
$\sqrt{n}\,S/\hat{\sigma}_{\mathrm{S}}$	2× 3	0.071	0.195	0.290	0.477	0.147	0.316	0.423	0.623
	4× 4	0.106	0.260	0.355	0.559	0.206	0.420	0.543	0.737
	6× 6	0.114	0.275	0.386	0.590	0.255	0.483	0.611	0.774
	10×10	0.117	0.285	0.396	0.599	0.282	0.503	0.631	0.796

[a] Due to computing expense, this case is based on 1000 simulations.

Table 4
Estimated proportion of times statistic exceeds normal critical value, when U model holds and there is an underlying bivariate normal distribution with correlation ρ

Statistic	$r \times c$	$n = 50$, $\alpha =$				$n = 100$, $\alpha =$			
		0.01	0.05	0.10	0.25	0.01	0.05	0.10	0.25
$\sqrt{n}(\hat{\beta} - \beta)/\hat{\sigma}_{\hat{\beta}}$									
$\rho = 0$	2× 3	0.006	0.044	0.096	0.257	0.010	0.049	0.098	0.250
	4× 4	0.005	0.042	0.096	0.259	0.009	0.046	0.095	0.238
	6× 6	0.007	0.047	0.097	0.248	0.009	0.050	0.095	0.252
	10×10 [a]	0.008	0.042	0.087	0.240	0.008	0.046	0.089	0.236
$\rho = 0.2$	2× 3	0.006	0.045	0.089	0.261	0.007	0.049	0.097	0.259
	4× 4	0.008	0.048	0.096	0.262	0.007	0.050	0.100	0.249
	6× 6	0.006	0.042	0.089	0.240	0.006	0.043	0.097	0.253
	10×10 [a]	0.017	0.060	0.115	0.257	0.014	0.060	0.105	0.262
$\sqrt{n}[S - E(S)]/\hat{\sigma}_S$									
$\rho = 0$	2× 3	0.016	0.060	0.115	0.267	0.014	0.056	0.105	0.254
	4× 4	0.014	0.059	0.113	0.275	0.014	0.053	0.103	0.245
	6× 6	0.017	0.061	0.112	0.254	0.014	0.058	0.106	0.255
	10×10	0.011	0.056	0.106	0.254	0.010	0.051	0.109	0.256
$\rho = 0.2$	2× 3	0.015	0.059	0.109	0.266	0.013	0.057	0.106	0.263
	4× 4	0.018	0.066	0.116	0.274	0.012	0.055	0.108	0.254
	6× 6	0.015	0.057	0.108	0.258	0.011	0.053	0.105	0.265
	10×10	0.013	0.059	0.112	0.258	0.016	0.058	0.113	0.268

[a] Due to computing expense, this case is based on 1000 simulations.

increases, for fixed n), whereas $X^2(\mathrm{I})$ loses power. Thus, sparseness is actually advantageous for these four statistics.

For local alternatives to independence, the statistics in this study have asymptotic noncentral chi-squared distributions. Das Gupta and Perlman [4] showed that for fixed noncentrality, the power of chi-squared statistics increases as the degrees of freedom decrease. When the U model holds, $G^2(\mathrm{I}\,|\,\mathrm{U})$ and $G^2(\mathrm{I})$ have the same noncentrality, so we expect $G^2(\mathrm{I}\,|\,\mathrm{U})$ (for which the noncentrality is focused on a single degree of freedom) to be more powerful than $G^2(\mathrm{I})$. The same remark applies to $X^2(\mathrm{I}\,|\,\mathrm{U})$ and $X^2(\mathrm{I})$. As t increases for the normal underlying distribution, the noncentrality increases and is still focused on a single degree of freedom for z_{U}^2, z_{S}^2, $X^2(\mathrm{I}\,|\,\mathrm{U})$, and $G^2(\mathrm{I}\,|\,\mathrm{U})$, resulting in increased power. On the other hand, df also increases for $X^2(\mathrm{I})$ or $G^2(\mathrm{I})$, more than off-setting the increase in noncentrality. The single-degree-of-freedom statistics would be inappropriate when the U model fits very poorly, such as when the association is non-monotonic. In practice, though, many associations are nearly monotonic in some sense, and it is important to investigate how the power behaves if the U model holds only approximately. This was considered in a Ph.D. dissertation by A. Kezouh at the University of Florida (1984), who studied these tests when model (2.2) actually holds with monotone but unequally-spaced scores. He

concluded that statistics such as $G^2(\mathrm{I}\,|\,\mathrm{U})$ still tend to outperform greatly ones such as $G^2(\mathrm{I})$, except when the true table is 'close' to independence. The relative advantage of $G^2(\mathrm{I}\,|\,\mathrm{U})$ tended to increase as the strength of association increased and as r and c increased.

Table 4 shows that the asymptotic approximations for the distributions of z_{Ua} and z_{Sa} did not deteriorate as sparseness increased, for either value of ρ. Hence, confidence intervals for association parameters that describe characteristics of the entire table (such as β in the U model, or the correlation) retain their usefulness for sparse tables. This result for $\hat{\beta}$ can be regarded as an illustration of Theorem 1 in Haberman [11], which indicates that functionals of $\{\hat{m}_i\}$ can have asymptotic normal distributions even if the number of cells grows at the same rate as the sample size.

3.4. Effects of adding cell constants

Sparse tables typically contain many empty cells. This can cause problems with existence of estimates for loglinear model parameters or cell probabilities, problems with severe bias in estimation of descriptive statistics such as odds ratios, problems with the performance of computational algorithms, as well as problems with asymptotic approximations of chi-squared statistics (see, e.g., Brown and Fuchs, [2]). Thus, it is common practice for researchers to add a small constant to cell counts before conducting the analysis. Goodman [5] suggests adding 0.5 to each cell count before computing model parameter estimates. Weighted least squares solutions require all cell counts to be positive, so Grizzle et al. [9] suggest adding to each empty cell the inverse of the number of categories of the response variable. Bishop et al. [1, p. 401] indicate that it is generally accepted practice to add 0.5 to each cell count of large, sparse tables, though they instead recommend Bayes or empirical Bayes approaches whereby a prior distribution induces the smoothing.

Our simulations have indicated that adding constants to cells in sparse tables can cause havoc with the distribution of $X^2(M)$ statistics. Adding a constant to each cell or to each empty cell represents a smoothing towards independence, resulting in a conservative influence on the statistics. To illustrate, Table 5 shows the effect of adding $1/c$ to every cell. This approach makes $X^2(\mathrm{I})$ far too conservative when n/t is less than about 5, the effect becoming very severe for larger tables. The $1/c$ adjustment results in an improvement for the $G^2(\mathrm{I})$ statistic for those situations in which it was highly liberal for the unadjusted table. However, as n/t decreases the adjustment becomes overly severe, particularly for cases where $G^2(\mathrm{I})$ is itself conservative for the original table. The statistics $G^2(\mathrm{U})$ and $X^2(\mathrm{U})$ behaved much like $G^2(\mathrm{I})$ and $X^2(\mathrm{I})$, respectively, and are not reported here. Again, $G^2(\mathrm{I}\,|\,\mathrm{U})$ and $X^2(\mathrm{I}\,|\,\mathrm{U})$ behaved alike and were much better than $G^2(\mathrm{I})$ and $X^2(\mathrm{I})$, particularly for very sparse tables. However, these statistics performed more poorly than when they were applied to the unadjusted table (compare with Table 1).

Effects on the statistics of adding $1/c$ only to the empty cells were very similar,

Table 5
Proportion of times statistic exceeds chi-squared percentage point when there is independence and constant $1/c$ added to all cells

Statistic	$r \times c$	$n = 50$, $\alpha =$				$n = 100$, $\alpha =$			
		0.01	0.05	0.10	0.25	0.01	0.05	0.10	0.25
$G^2(\mathrm{I})$	2× 3	0.009	0.046	0.095	0.244	0.008	0.051	0.104	0.255
	4× 4	0.007	0.045	0.095	0.254	0.008	0.053	0.103	0.264
	6× 6	0.000	0.009	0.027	0.118	0.007	0.041	0.092	0.248
	10×10	0.000	0.000	0.000	0.000	0.000	0.001	0.004	0.023
$X^2(\mathrm{I})$	2× 3	0.006	0.042	0.088	0.238	0.007	0.048	0.100	0.251
	4× 4	0.004	0.027	0.066	0.196	0.006	0.037	0.084	0.235
	6× 6	0.001	0.007	0.024	0.089	0.004	0.023	0.054	0.166
	10×10	0.000	0.000	0.000	0.003	0.001	0.005	0.011	0.040
$G^2(\mathrm{I}\mid\mathrm{U})$	2× 3	0.010	0.048	0.099	0.250	0.009	0.049	0.100	0.250
	4× 4	0.007	0.041	0.090	0.240	0.011	0.046	0.093	0.230
	6× 6	0.006	0.034	0.078	0.208	0.008	0.045	0.087	0.238
	10×10	0.004	0.022	0.054	0.176	0.003	0.031	0.072	0.209
$X^2(\mathrm{I}\mid\mathrm{U})$	2× 3	0.008	0.046	0.095	0.247	0.008	0.048	0.098	0.249
	4× 4	0.006	0.039	0.088	0.240	0.010	0.045	0.092	0.230
	6× 6	0.005	0.033	0.077	0.207	0.008	0.043	0.086	0.237
	10×10	0.004	0.021	0.055	0.175	0.003	0.031	0.072	0.210

producing slightly more conservative results for the very sparse tables. The effects were far more severe when a larger constant, such as 0.5, was added to every cell. For instance, for 6×6 tables with $n = 50$, the estimated true tail probabilities for $X^2(\mathrm{I})$ are 0.000, 0.000, 0.001, and 0.010, corresponding to the nominal values 0.01, 0.05, 0.10, and 0.25, respectively.

4. Conclusions and recommendations

The study reported here was limited in scope, and one cannot use it to make sweeping generalizations about the analysis of sparse data. However, certain tentative conclusions and further conjectures are suggested by the results in Tables 1–5.

4.1. Behavior of G^2 and X^2 statistics

Table 1 is consistent with previous findings that $G^2(M)$ behaves poorly for sparse tables. This statistic is likely to behave even more poorly when there is more variation in the $\{m_i\}$ than encountered in this study. For goodness-of-fit testing of a specified multinomial, Koehler and Larntz [14] showed that a standardized version of G^2 is well approximated by the normal distribution for very sparse tables. For testing the fit of a model, it is also likely that a normal

limiting distribution will give better approximations than the usual reference chi-squared distribution, for sparse data. McCullagh [16] reviewed ways of handling sparse tables, and he presented a normal approximation for G^2 that may be a useful alternative. However, its use is computationally intensive, and it assumes that the dimension of the model parameter vector is fixed as the size of the table increases.

Tables 1 and 2 are also consistent with previous studies that noted that the Pearson statistic $X^2(M)$ behaves much better than $G^2(M)$ for sparse tables. In this study, the asymptotic approximation for $X^2(M)$ was adequate for n/t as small as 1. The size of n/t that produces adequate approximations tends to decrease as t increases. For instance, Koehler and Larntz suggest the guideline $n > \sqrt{10t}$ (i.e., $n/t > \sqrt{10/t}$) for using X^2 for goodness-of-fit testing of the uniform multinomial probabilities $(1/t, \ldots, 1/t)$. In testing models for which there is considerable variability in cell probabilities, $n > 10\sqrt{t}$ might be a more reasonable guideline for use of $X^2(M)$. As Koehler and Larntz note, it is hopeless to expect one rule to cover all cases, but further research with other models and other choices of underlying distributions may help to suggest appropriate guidelines of this type.

4.2. Behavior of model-comparison statistics

The adequacy of the asymptotic distribution of $G^2(M_1 \mid M_2)$ or $X^2(M_1 \mid M_2)$ is likely to be governed by the sufficient marginal configuration for M_2 that is farthest from its asymptotic distribution. For instance, the statistic $G^2(\mathrm{I} \mid \mathrm{U})$ will be influenced by the most sparse marginal distribution, so it should usually behave well if $n > 5[\max(r, c)]$, when $\max(r, c)$ is relatively large. On the other hand, $X^2(M_1) - X^2(M_2)$ may be inadequate whenever $X^2(M_1)$ is. Thus, even though the distribution of $X^2(M_1)$ may be closer to chi-squared than that of $G^2(M_1)$, usually one would prefer $G^2(M_1) - G^2(M_2)$ to $X^2(M_1) - X^2(M_2)$ for comparing two models or for testing a hypothesis by imbedding it within a model.

4.3. Sparseness and power

Table 3 shows that increasing the numbers of categories, for a fixed sample size, tends to improve the power of statistics designed to detect associations between ordinal variables. We conjecture that statistics based on $\hat{\beta}$ (when it exists) or the score S retain their inferential usefulness regardless of the degree of sparseness. For instance, suppose we consider a sequence of categorizations of an underlying bivariate normal distribution for which cutpoints are equally-spaced and equal-interval scores are assigned to the rows and columns. As $r = c \to \infty$ with $\max\{\pi_{1+}, \ldots, \pi_{r+}, \pi_{+1}, \ldots, \pi_{+c}\} \to 0$, the sampling distribution of S behaves like that of the sample correlation for the underlying continuous distribution. (Note that S is the sample correlation in the table, at each stage, when scores are chosen such that marginal standard deviations equal 1.) Also, $\hat{\beta}$ has an

approximate functional relationship with the correlation in this case (see [7]).

This conjecture does not apply to statistics for which df increases when categorizations are refined, as illustrated by $X^2(\mathrm{I})$ in Table 3. Also, in practice, sparseness often results from increasing the number of variables rather than from increasing the numbers of category levels (particularly for nominal variables), so this conjecture has limited applicability.

4.4. Adding cell constants

In using $X^2(M)$ to test the fit of a model, we observed that it can be risky to add a constant to the cells. Even the addition of only 0.25 to the cells of a 4×4 table with $n = 50$ has a marked conservative influence on the distribution of $X^2(\mathrm{I})$, for instance. This shrinkage towards independence also applies to model parameter estimates. For the 10×10 table with underlying correlation 0.2, the parameter β in the U model equals 0.036. When $n = 50$, the expected value of $\hat{\beta}$ is approximately 0.039 for the original table, but it is only 0.026 when 0.1 is added to each empty cell before one fits the model. If adding a constant is necessary to ensure existence of estimates, it may be preferable to select a very small constant, and it is wise to try constants of various sizes to assess the dependence of the result on that choice. In doing this, there may still be problems with weighted least squares estimation, since relatively more weight is given to cells when the cell proportion estimate (and resulting variance estimate) decreases.

In Tables 1 and 2 we noted that $G^2(M)$ behaves much more poorly than $X^2(M)$ when n/t is approximately in the range 1 to 10, and Table 5 suggests that the addition of a constant can improve the asymptotic approximation of $G^2(M)$ for n/t between about 2 and 10. It would be useful if future research could establish guidelines for the choice of this constant, as a function of n and t.

4.5. Generalizations

One conclusion from this research is the following: If we wish to test a hypothesis for sparse categorical data, it is wise to imbed that hypothesis as a special case of an unsaturated loglinear model, so that a statistic of the form $G^2(M_1 \mid M_2)$ or $X^2(M_1 \mid M_2)$ can be used. We illustrated this by imbedding the independence hypothesis in the uniform association model for two ordinal variables, but the idea extends quite generally. For instance, suppose we wish to test whether X and Y are conditionally independent, given Z, in an r-by-c-by-k table. Let (XZ, YZ) denote the loglinear model corresponding to this condition. The distribution of $G^2[(XZ, YZ)]$ may be poorly approximated by the $X^2_{k(r-1)(c-1)}$ distribution, particularly if $n < 5rck$. However, suppose that we test conditional independence under the assumption that the no three-factor interaction model (denoted by (XY, XZ, YZ)) holds. The statistic $G^2[(XZ, YZ) \mid (XY, XZ, YZ)]$ should behave well if the two-dimensional margins are not particularly sparse, say $n > 5[\max(rc, rk, ck)]$. If $r = c = 2$, this is a single-degree-of-freedom statistic

that may (like the Mantel–Haenszel statistic) behave adequately even if k is quite large. Of course, if model (XY, XZ, YZ) fits poorly, it is inappropriate to test the fit of model (XZ, YZ) in any case.

References

[1] Y.M.M. Bishop, S.E. Fienberg and P.W. Holland, *Discrete Multivariate Analysis* (MIT Press, Cambridge, MA, 1975).

[2] M.B. Brown and C. Fuchs, On maximum likelihood estimation in sparse contingency tables, *Comput. Statist. Data Anal.* **1** (1983) 3–15.

[3] D.R. Cox and D.V. Hinkley, *Theoretical Statistics* (Chapman and Hall, London, 1974).

[4] S. Das Gupta and M.D. Perlman, Power of the noncentral F-test: effect of additional variates on Hotelling's T^2-test, *J. Amer. Statist. Assoc.* **79** (1974) 174–180.

[5] L.A. Goodman, The multivariate analysis of qualitative data: Interactions among multiple classifications, *J. Amer. Statist. Assoc.* **65** (1970) 226–256.

[6] L.A. Goodman, Simple models for the analysis of association in cross-classifications having ordered categories. *J. Amer. Statist. Assoc.* **74** (1979) 537–552.

[7] L.A. Goodman, Association models and the bivariate normal distribution in the analysis of cross-classifications having ordered categories, *Biometrika* **68** (1981) 347–355.

[8] L.A. Goodman, The analysis of cross-classified data having ordered and/or unordered categories: Association models, correlation models, and asymmetry models for contingency tables with or without missing entries, *Ann. Statist.* **13** (1985) 10–69.

[9] J.E. Grizzle, C.F. Starmer and G.G. Koch, Analysis of categorical data by linear models, *Biometrics* **25** (1969) 489–504.

[10] S.J. Haberman, *The Analysis of Frequency Data* (University of Chicago Press, 1974).

[11] S.J. Haberman, Log-linear models and frequency tables with small expected cell counts, *Ann. Statist.* **5** (1977) 1148–1169.

[12] S.J. Haberman, *Analysis of Qualitative Data. Vol. 1: Introductory Topics* (Academic Press, New York, 1978).

[13] P.B. Imrey, G.G. Koch and M.E. Stokes, Categorical data analysis: Some reflections on the log linear model and logistic regression. Part II: Data analysis. *Internat. Statist. Rev.* **50** (1982) 35–63.

[14] K. Koehler and K. Larntz, An empirical investigation of goodness-of-fit statistics for sparse multinomials. *J. Amer. Statist. Assoc.* **75** (1980) 336–344.

[15] K. Larntz, Small-sample comparison of exact levels for chi-squared goodness-of-fit statistics, *J. Amer. Statist. Assoc.* 73 (1978) 253–263.

[16] P. McCullagh, The conditional distribution of goodness-of-fit statistics for discrete data. *J. Amer. Statist. Assoc.* **81** (1986) 104–107.