Statistica Sinica **13**(2003), 555-570

SIZE AND POWER CONSIDERATIONS FOR TESTING LOGLINEAR MODELS USING ϕ-DIVERGENCE TEST STATISTICS

Noel Cressie[1], Leandro Pardo[2] and Maria del Carmen Pardo[2]

[1] *The Ohio State University and* [2] *Complutense University of Madrid*

Abstract: In this article, we assume that categorical data are distributed according to a multinomial distribution whose probabilities follow a loglinear model. The inference problem we consider is that of hypothesis testing in a loglinear-model setting. The null hypothesis is a composite hypothesis nested within the alternative. Test statistics are chosen from the general class of ϕ-divergence statistics. This article collects together the operating characteristics of the hypothesis test based on both asymptotic (using large-sample theory) and finite-sample (using a designed simulation study) results. Members of the class of power divergence statistics are compared, and it is found that the Cressie-Read statistic offers an attractive alternative to the Pearson-based and the likelihood ratio-based test statistics, in terms of both exact and asymptotic size and power.

Key words and phrases: Chi-squared distribution, contiguous alternatives, multinomial distribution, nested hypotheses, noncentral chi-squared distribution, power-divergence statistic.

1. Introduction

Categorical data analysis is an essential tool when the data are nominal. Even when the data are ordinal, it sometimes makes sense to categorize them into a discrete number $k > 1$, of classes (e.g., for stratification purposes or for assessing goodness of fit to a parametric family of distributions). In this article, we consider statistical models for categorical data whose parameter space Θ has dimension $t < k-1$.

Let $Y_1, \ldots, Y_n$ be a sample of size $n \geq 1$, with realizations from $\mathcal{X} = \{1, \ldots, k\}$ and independent and identically distributed (i.i.d.) according to a probability distribution $P(\theta_0)$. This distribution is assumed to be unknown, but belonging to a known family, $\mathcal{P} = \{P(\theta) = (p_1(\theta), \ldots, p_k(\theta))^T : \theta \in \Theta\}$, of distributions on $\mathcal{X}$ with $\Theta \subseteq R^t$, $t < k-1$. Thus the true value θ_0 of parameter $\theta = (\theta_1, \ldots, \theta_t)^T \in \Theta \subseteq R^t$ is fixed but unknown. We denote $P = (p_1, \ldots, p_k)^T$ and $\widehat{P} = (\widehat{p}_1, \ldots, \widehat{p}_k)^T$, with

$$\widehat{p}_j = \frac{X_j}{n} \quad \text{and} \quad X_j = \sum_{i=1}^{n} I_{\{j\}}(Y_i), \quad j = 1, \ldots, k. \tag{1}$$

Here and in the sequel, "T" denotes vector or matrix transpose. The statistic $(X_1, \ldots, X_k)$ is obviously sufficient for the statistical model under consideration and is multinomially distributed:

$$P(X_1 = x_1, \ldots, X_k = x_k) = \frac{n!}{x_1! \cdots x_k!} p_1(\theta)^{x_1} \times \cdots \times p_k(\theta)^{x_k}, \tag{2}$$

for integers $x_1, \ldots, x_k \geq 0$ such that $x_1 + \cdots + x_k = n$.

A class of models often used in (2) is the family of loglinear models:

$$p_u(\theta) = \exp\left(w_u^T \theta\right) / \sum_{v=1}^{k} \exp\left(w_v^T \theta\right); \quad u = 1, \ldots, k, \tag{3}$$

where the $k \times t$ matrix $W = (w_1, \ldots, w_k)^T$ is assumed to have full column rank $t < k-1$ and columns linearly independent of the $k \times 1$ column vector $(1, \ldots, 1)^T$. This will be the model we consider for the theoretical results in Sections 2 and 3, and for the simulation in Section 4.

The problem that has motivated our research involves a nested sequence of hypotheses,

$$H_l : P = P(\theta); \quad \theta \in \Theta_l; \quad l = 1, \ldots, m, \quad m \leq t < k-1, \tag{4}$$

where $\Theta_m \subset \Theta_{m-1} \subset \cdots \subset \Theta_1 \equiv \Theta \subseteq R^t$; $t < k-1$ and $\dim(\Theta_l) = d_l$; $l = 1, \ldots, m$, with

$$d_m < d_{m-1} < \cdots < d_1 = t. \tag{5}$$

In this framework, there is an integer m^* $(1 \leq m^* \leq m)$ for which H_{m^*} is true but H_{m^*+1} is not true. A common strategy for making inference on m^* (e.g., Read and Cressie (1988), p.42) is to test successively,

$$H_{Null} : H_{l+1} \quad \text{against} \quad H_{Alt} : H_l; \quad l = 1, \ldots, m-1, \tag{6}$$

where we continue to test as long as the null hypothesis is accepted, and we infer m^* to be the first l for which H_{l+1} is rejected as a null hypothesis. The full operating characteristics of this sequence of tests of nested hypotheses are not known. Our goal in this paper is to give comparative size and power results for individual tests in the sequence based on a general class of $\phi-$divergence test statistics.

Since the parameter values in $\{\Theta_l : l = 1, \ldots, m\}$ are generally unknown, most tests require their estimation. For example, if $\sum_{j=1}^{k} \widehat{p}_j \log p_j(\theta)$ is almost surely (a.s.) maximized over Θ_l at some $\widehat{\theta}^{\,(l)}$, then $\widehat{\theta}^{\,(l)}$ is the point maximum likelihood estimator (MLE). The MLE can equivalently be defined by the condition,

$$\widehat{\theta}^{\,(l)} = \arg\min_{\theta \in \Theta_l} D\left(\widehat{P}, P(\theta)\right) \quad \text{a.s.}, \tag{7}$$

where $D(P,Q) = \sum_{j=1}^{k} p_j \log \frac{p_j}{q_j}$ is the Kullback-Leibler divergence and $P = (p_1, \ldots, p_k)^T$, $Q = (q_1, \ldots, q_k)^T$. The definition (7) hints at a much more general inference framework based on divergence measures, which was investigated by Cressie and Pardo (2000). In the next several paragraphs, we give the essential details of the framework for estimation and hypothesis testing there.

Consider the $\phi-$divergence defined by Csiszár (1963) and Ali and Silvey (1966):

$$D_\phi(P,Q) \equiv \sum_{j=1}^{k} q_j \phi\left(\frac{p_j}{q_j}\right); \phi \in \Phi^*, \tag{8}$$

where Φ^* is the class of all convex functions $\phi(x)$, $x > 0$, such that at $x = 1$, $\phi(1) = 0$, $\phi''(1) > 0$, and at $x = 0$, $0\phi(0/0) = 0$ and $0\phi(p/0) = \lim_{u\to\infty} \phi(u)/u$. For every $\phi \in \Phi^*$ that is differentiable at $x = 1$, the function $\psi(x) \equiv \phi(x) - \phi'(1)(x-1)$ also belongs to Φ^*. Then we have $D_\psi(P,Q) = D_\phi(P,Q)$, and ψ has the additional property that $\psi'(1) = 0$. Because the two divergence measures are equivalent, we can consider the set Φ^* to be equivalent to the set $\Phi \equiv \Phi^* \cap \{\phi : \phi'(1) = 0\}$. In what follows, we give our theoretical results for $\phi \in \Phi$, but we often apply them to choices of functions in Φ^*.

For example, an important family of $\phi-$divergences in statistical problems is the power-divergence family:

$$\begin{aligned} &\phi_{(\lambda)}(x) = (\lambda(\lambda+1))^{-1}(x^{\lambda+1} - x); \quad \lambda \neq 0,\ \lambda \neq -1, \\ &\phi_{(0)}(x) = \lim_{\lambda\to 0} \phi_{(\lambda)}(x), \quad \phi_{(-1)}(x) = \lim_{\lambda\to -1} \phi_{(\lambda)}(x), \end{aligned} \tag{9}$$

introduced and studied by Cressie and Read (1984). We observe that the functions $\phi_{(\lambda)}(x)$ and $\psi_{(\lambda)}(x) \equiv \phi_{(\lambda)}(x) - (x-1)(\lambda+1)^{-1}$ define the same divergence measure. In the following, we denote the power-divergence measures by $I^\lambda(P,Q) \equiv D_{\phi_{(\lambda)}}(P,Q) = D_{\psi_{(\lambda)}}(P,Q)$.

Cressie and Read (1984) defined the minimum power-divergence estimator of $\theta \in \Theta \subseteq R^t$ as

$$\widehat{\theta}_{(\lambda)} \equiv \arg\min_{\theta\in\Theta} I^\lambda\left(\widehat{P}, P(\theta)\right) \quad \text{a.s.}, \tag{10}$$

and they studied its properties. Notice that $\widehat{\theta}_{(0)}$ is the MLE, as observed at (7). Other estimators (less well known than the MLE) that are members of the family of minimum power-divergence estimators are the minimum chi-squared estimator (Neyman (1949)) for $\lambda = 1$; the minimum modified chi-squared estimator (Neyman (1949)) for $\lambda = -2$; the modified MLE or minimum discrimination information estimator (Kullback (1985)) for $\lambda = -1$; the minimum Matusita distance (or Hellinger distance) estimator (Matusita (1954)) for $\lambda = -1/2$; and the minimum Cressie-Read distance estimator (Cressie and Read (1984)) for $\lambda = 2/3$.

Later, Morales, Pardo and Vajda (1995) considered the minimum ϕ-divergence estimator

$$\widehat{\theta}_{\phi} \equiv \arg\min_{\theta\in\Theta} D_{\phi}\left(\widehat{P}, P(\theta)\right) \quad \text{a.s.}, \tag{11}$$

and studied its properties. Furthermore, they showed that to test $H_{Null} : P = P(\theta)$; $\theta \in \Theta \subseteq R^t$, a natural (suitably normalized) test statistic, is

$$Q_{\phi_1,\phi_2} \equiv \frac{2n}{\phi_1''(1)} D_{\phi_1}\left(\widehat{P}, P\left(\widehat{\theta}_{\phi_2}\right)\right); \quad \phi_1,\ \phi_2 \in \Phi,$$

with H_{Null} rejected if this statistic is too large.

Cressie and Pardo (2000, 2002) extended this framework to test $H_{Null} : H_{l+1}$ against $H_{Alt} : H_l$; $l = 1, ..., m-1$, given by (4). To test H_{l+1} against H_l, they suggested using the test statistic

$$Q^{(l)}_{\phi_1,\phi_2} = \frac{2n}{\phi_1''(1)} D_{\phi_1}\left(P\left(\widehat{\theta}_{\phi_2}^{(l)}\right), P\left(\widehat{\theta}_{\phi_2}^{(l+1)}\right)\right); \quad \phi_1,\ \phi_2 \in \Phi, \tag{12}$$

where $\widehat{\theta}_{\phi_2}^{(l+1)}$ and $\widehat{\theta}_{\phi_2}^{(l)}$ are defined by (11), and the null hypothesis $H_{Null} : H_{l+1}$ is rejected if $Q^{(l)}_{\phi_1,\phi_2}$ is too large.

In the rest of this section and in the simulations of Section 4, we choose $\phi_2 \equiv \phi_{(0)}$, which corresponds to estimation of unknown parameters by maximum likelihood. Notice that $Q^{(l)}_{\phi_1,\phi_2}$ is one of a number of possible test statistics chosen by Cressie and Pardo (2000); (12) seems the most natural in this context because it generalizes the test statistic Q_{ϕ_1,ϕ_2}, which in turn generalizes the log-likelihood-ratio test.

The two most commonly used test statistics in (12) are the Pearson statistic, corresponding to $\phi_1 \equiv \phi_{(1)}$ given by (9), and the log-likelihood-ratio statistic, corresponding to $\phi_1 \equiv \phi_{(0)}$ given by (9) (e.g., Christensen (1997, p.338)). The asymptotic null distribution of both of these statistics is a central chi-squared distribution with $d_l - d_{l+1}$ degrees of freedom. Regarding the alternative, Haberman (1974) was the first to study the asymptotic distribution of the two previous statistics under contiguous alternative hypotheses (Section 3), establishing that the asymptotic distribution is non-centrally chi-squared distributed with $d_l - d_{l+1}$ degrees of freedom. Oler (1985) presented a systematic study of the contiguous alternative hypothesis in multinomial populations, obtaining as a particular case the asymptotic distribution for the loglinear models. Through simulations, she also studied how closely the noncentral chi-squared distributions agree with the exact sampling distributions. Fenech and Westfall (1988) presented an interesting analytic study of the noncentrality parameter in the case of loglinear models.

In what is to follow, we use the general inference framework based on divergence measures to make a coherent study of testing a composite null hypothesis against a composite alternative hypothesis, where the alternative can be contiguous or not. One of our goals is to determine which divergence measures have better size and power properties in small samples under different types of alternatives. In Section 2, we review some of the results obtained by Cressie and Pardo (2000) under the null hypothesis. In Section 3, we generalize the results of Oler (1985) for contiguous alternatives, to tests based on the $\phi-$divergence statistic (12). A sequence of loglinear models is the basis of a simulation given in Section 4. There, we use $\phi_1 = \phi_{(\lambda)}$ and $\phi_2 = \phi_{(0)}$ in tests based on (12), and compare various values of λ $(-2, -1, -1/2, 0, 2/3, 1, 2)$ for small to moderate values of n.

2. A Review of Distributional Properties Under the Null Hypothesis

For testing the nested hypotheses $\{H_l : l = 1, \ldots, m\}$ given by (4), we test $H_{Null} : H_{l+1}$ against $H_{Alt} : H_l$, using test statistic $Q^{(l)}_{\phi_1,\phi_2}$ given by (12); if it is too large, H_{Null} is rejected. When $Q^{(l)}_{\phi_1,\phi_2} > c$, we reject H_{Null} in (6), where c is specified so that the size of the test is α :

$$\Pr\left(Q^{(l)}_{\phi_1,\phi_2} > c \mid H_{l+1}\right) = \alpha; \quad \alpha \in (0, 1). \tag{13}$$

Cressie and Pardo (2000) show that under (2), (3), (4), and $H_{Null} : H_{l+1}$, the test statistic $Q^{(l)}_{\phi_1,\phi_2}$ converges in distribution to a chi-squared distribution with $d_l - d_{l+1}$ degrees of freedom $(\chi^2_{d_l-d_{l+1}})$; $l = 1, \ldots, m-1$. Thus

$$c = \chi^2_{d_l-d_{l+1}}(1-\alpha), \tag{14}$$

where $\Pr(\chi^2_f \leq \chi^2_f(p)) = p$. Notice that when $\phi_1 = \phi_2 = \phi_{(0)}$, given by (9), we obtain the usual likelihood-ratio test, and that when $\phi_1 = \phi_{(1)}$ and $\phi_2 = \phi_{(0)}$, we obtain the Pearson test statistic (e.g., Agresti (1990), Ch.6). It should be noted at this point that the asymptotic distribution theory of $Q^{(l)}_{\phi_{(\lambda)},\phi_{(0)}}$ under the null hypothesis, due to Cressie and Pardo (2000), can be generalized to a result under a sequence of contiguous alternative hypotheses. This theoretical result is proved in Section 3 and its accuracy is assessed in Section 4.

The choice of (14) in (13) only guarantees an asymptotic size-α test. In the case of the Pearson and loglikelihood ratio statistics, some corrections to (14) have been proposed. These have been discussed by Read and Cressie (1988), Ch.5, in the context of power-divergence statistics for testing goodness-of-fit. Here we use (14) but ask, in the finite-sample simulations given in Section 4, for what choices of λ in $Q^{(l)}_{\phi_1,\phi_2}$ is the relation (13) most accurately attained?

The asymptotic chi-squared approximation, $c = \chi^2_{d_l-d_{l+1}}(1-\alpha)$, is checked for a sequence of loglinear models in the simulation given in Section 4. We give a small illustration of those results now. Figure 1 shows *departures* of the exact size from the nominal size of $\alpha = 0.05$ for one particular choice (specified in Section 4) of H_{l+1} and H_l, for various choices of λ in $\phi_1 = \phi_{(\lambda)}$, and for small to large sample sizes ($n = 15,\ 20,\ 25,\ 35,\ 50, 100,\ 200$). Figure 1a represents nonpositive choices of λ, and Figure 1b represents nonnegative choices of λ. The positive values of λ perform the best.

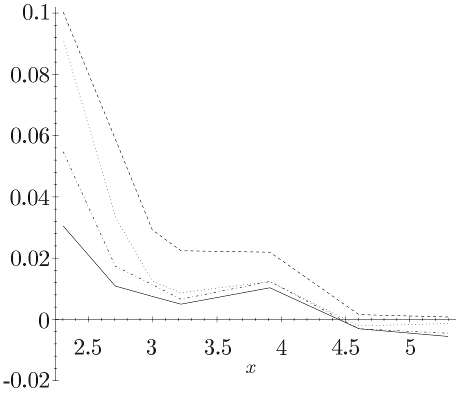

Figure 1a. (Exact size−Nominal size of 0.05) as a function of $x = \log n$. Shown are $\lambda = -2$ (dashed line), $\lambda = -1$ (dotted line), $\lambda = -1/2$ (dash-dotted line), and $\lambda = 0$ (solid line).

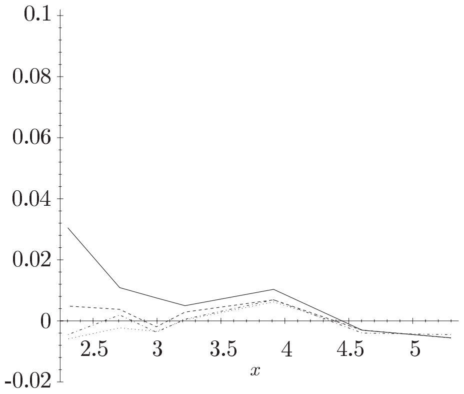

Figure 1b. (Exact size−Nominal size of 0.05) as a function of $x = \log n$. Shown are $\lambda = 0$ (solid line), $\lambda = 2/3$ (dashed line), $\lambda = 1$ (dotted line), and $\lambda = 2$ (dash-dotted line).

To test the nested sequence of hypotheses $\{H_l : l = 1, \ldots, m\}$ referred to in Section 1, we need an asymptotic independence result for the sequence of test statistics $Q^{(1)}_{\phi_1,\phi_2}$, $Q^{(2)}_{\phi_1,\phi_2}, \ldots, Q^{(m^*)}_{\phi_1,\phi_2}$, where m^* is the integer $1 \leq m^* \leq m$ for which H_{m^*} is true but H_{m^*+1} is not true. This result was not given by Cressie and Pardo (2000); we give it in the theorem below.

Theorem 1. *Suppose that data $(X_1, \ldots, X_k)$ are multinomially distributed according to the loglinear model* (3). *We first test, $H_{Null} : H_l$ against $H_{Alt} : H_{l-1}$, followed by $H_{Null} : H_{l+1}$ against $H_{Alt} : H_l$. Then, under the hypothesis H_l, the statistics $Q^{(l-1)}_{\phi_1,\phi_2}$ and $Q^{(l)}_{\phi_1,\phi_2}$ are asymptotically independent and chi-squared distributed on $d_{l-1} - d_l$ and $d_l - d_{l+1}$ degrees of freedom, respectively.*

Proof. See the Appendix.

3. Contiguous Alternative Hypotheses

In this section, we derive results for testing one of the H_{Null}, H_{Alt} pairs. Our ultimate goal in this paper is to compare various $\phi-$divergence statistics and give recommendations for those that are the most accurate and powerful.

The expectation is that a superior test statistic for individual tests in the sequence of nested hypotheses, leads to superior inference for (6) and its associated m^* (e.g., Oler (1985); Fenech and Westfall (1988); Christensen (1977)). However, to establish this is beyond the scope of this paper.

In general, theoretical results for the test statistic $Q^{(l)}_{\phi_1,\phi_2}$ under alternative hypotheses are not easy to obtain. An exception to this is when there is a contiguous sequence of alternatives that approach the null hypothesis H_{l+1} at the rate of $O(n^{-1/2})$. In Section 1, we reviewed early contributions of Haberman (1974), Oler (1985), and Fenech and Westfall (1988) to this theory. In this section, we generalize their results to tests based on the $\phi-$ divergence statistic $Q^{(l)}_{\phi_1,\phi_2}$ given by (12); the results below mirror those of Section 2, but under a sequence of contiguous alternatives.

Consider the multinomial probability vector

$$P_n(\theta) \equiv P(\theta) + s/\sqrt{n}, \quad \theta \in \Theta_{l+1},\ n \geq n_0 > 0, \tag{15}$$

where $s \equiv (s_1, \ldots, s_k)^T$ is a fixed $k \times 1$ vector such that $\sum_{j=1}^{k} s_j = 0$, and n is the total-count parameter of the multinomial distribution. As $n \to \infty$, the sequence of multinomial probabilities $\{P_n(\theta)\}_{n\in N}$ converges to a multinomial probability in H_{l+1} at the rate of $O(n^{-1/2})$. We call

$$H_{l+1,n} : P = P_n(\theta) = P(\theta) + s/\sqrt{n}, \quad \theta \in \Theta_{l+1},\ n \geq n_0 > 0, \tag{16}$$

a sequence of *contiguous alternative hypotheses,* here contiguous to the null hypothesis H_{l+1}.

Now consider testing $H_{Null} : H_{l+1}$ against $H_{Alt} : H_{l+1,n}$, using the test statistic $Q^{(l)}_{\phi_1,\phi_2}$ given by (12). The power of this test is,

$$\pi_n^{(l)} \equiv \ \Pr\left(Q^{(l)}_{\phi_1,\phi_2} > c\,|H_{l+1,n}\right). \tag{17}$$

In what is to follow, we show that under the alternative $H_{l+1,n}$, and as $n \to \infty$, $Q^{(l)}_{\phi_1,\phi_2}$ converges in distribution to a non-central chi-squared random variable with non-centrality parameter μ, where μ is given in Theorem 2, and $d_l - d_{l+1}$ degrees of freedom $(\chi^2_{d_l - d_{l+1},\mu})$. Consequently, as $n \to \infty$,

$$\pi_n^{(l)} \to \ \Pr\left(\chi^2_{d_l-d_{l+1},\mu} > c\right). \tag{18}$$

One way to prove these results is to use Le Cam's lemmas, in particular the third lemma (Hájek and Sidák (1967)). However, in the case of multinomial sampling, the results can be proved directly. The technique of the proof has already been used in Menéndez, Morales, Pardo and Zografos (1999), Pardo, Pardo and Zografos (2001), and for loglinear models by Oler (1985).

In Remark 1 of Cressie and Pardo (2000), it was established that the asymptotic expansion of the minimum $\phi-$divergence estimator about $\theta_0 \in \Theta_{l+1}$ is given by

$$\widehat{\theta}_{\phi}^{(l+1)} = \theta_0 + \left(W_{(l+1)}^T \Sigma_{P(\theta_0)} W_{(l+1)}\right)^{-1} W_{(l+1)}^T \Sigma_{P(\theta_0)} \operatorname{diag}(P(\theta_0)^{-1}) \left(\widehat{P} - P(\theta_0)\right) + o\left(\left\|\widehat{P} - P(\theta_0)\right\|\right), \tag{19}$$

where W_{l+1} is the loglinear-model matrix of explanatory variables under the null hypothesis H_{l+1}, and $\Sigma_{P(\theta_0)} = \operatorname{diag} P(\theta_0) - P(\theta_0)P(\theta_0)^T$.

Under the hypothesis given in (16), we have $\sqrt{n}(\widehat{P} - P(\theta_0)) = \sqrt{n}(\widehat{P} - P_n(\theta_0)) + s$, and hence $\sqrt{n}(\widehat{P} - P(\theta_0)) \xrightarrow[n\to\infty]{L} N(s, \Sigma_{P(\theta_0)})$, so $o(\|\widehat{P} - P(\theta_0)\|) = o(O_p(n^{-1/2})) = o_p(n^{-1/2})$. Therefore, we have established that under the contiguous hypothesis given in (16), and for $\theta_0 \in \Theta_{l+1}$,

$$\widehat{\theta}_{\phi}^{(l+1)} = \theta_0 + (W_{(l+1)}^T \Sigma_{P(\theta_0)} W_{(l+1)})^{-1} W_{(l+1)}^T \Sigma_{P(\theta_0)} \operatorname{diag}(P(\theta_0)^{-1})(\widehat{P} - P(\theta_0)) + o_p(n^{-1/2}). \tag{20}$$

This result will be important in the following theorem.

Theorem 2. *Suppose that* $(X_1, \ldots, X_k)$ *is multinomially distributed according to* (2) *and* (3). *The asymptotic distribution of the statistic* $Q_{\phi_1,\phi_2}^{(l)}$, *under the contiguous alternative hypotheses* (16), *is chi-squared with* $d_l - d_{l+1}$ *degrees of freedom and non-centrality parameter* $\mu = s^T \operatorname{diag}(P(\theta_0)^{-1/2})(A_{(l)} - A_{(l+1)}) \operatorname{diag}(P(\theta_0)^{-1/2})s$, *where* $s = (s_1, \ldots, s_k)^T$ *is defined in* (16) *and satisfies* $\sum\limits_{i=1}^{k} s_i = 0$, *and*

$$A_{(i)} = \operatorname{diag}(P(\theta_0)^{-1/2}) \Sigma_{P(\theta_0)} W_{(i)} (W_{(i)}^T \Sigma_{P(\theta_0)} W_{(i)})^{-1} W_{(i)}^T \Sigma_{P(\theta_0)} \operatorname{diag}(P(\theta_0)^{-1/2});$$
$i = l,\ l+1.$

Proof. By Theorem 3 in Cressie and Pardo (2000), we know that

$$Q_{\phi_1,\phi_2}^{(l)} = Z^t Z + n \times o\left(\left\|P\left(\widehat{\theta}_{\phi_2}^{(l+1)}\right) - P(\theta_0)\right\|^2 + \left\|P\left(\widehat{\theta}_{\phi_2}^{(l)}\right) - P(\theta_0)\right\|^2\right),$$

where $Z = \sqrt{n} \operatorname{diag}(P(\theta_0)^{-1/2})(P(\widehat{\theta}_{\phi_2}^{(l+1)}) - P(\widehat{\theta}_{\phi_2}^{(l)}))$. But $P(\widehat{\theta}_{\phi_2}^{(l+1)}) - P(\theta_0) = P(\widehat{\theta}_{\phi_2}^{(l+1)}) - P_n(\theta_0) + P_n(\theta_0) - P(\theta_0)$, and $\sqrt{n}(P(\widehat{\theta}_{\phi_2}^{(l+1)} - P(\theta_0)) = \sqrt{n}(P(\widehat{\theta}_{\phi_2}^{(l+1)}) - P_n(\theta_0)) + s$. Now it is clear that

$$\sqrt{n}(P(\widehat{\theta}_{\phi_2}^{(l+1)}) - P(\theta_0)) = \sqrt{n}\frac{\partial P_n(\theta_0)}{\partial \theta}\left(\widehat{\theta}_{\phi_2}^{(l+1)} - \theta_0\right) + s + o_p(1),$$

and by (20), $\sqrt{n}\frac{\partial P_n(\theta_0)}{\partial\theta}(\widehat{\theta}_{\phi_2}^{(l+1)} - \theta_0) = O_p(1)$. Then we have $P(\widehat{\theta}_{\phi_2}^{(l+1)}) - P(\theta_0) = O_p(n^{-1/2})$, and $\|P(\widehat{\theta}_{\phi_2}^{(l+1)}) - P(\theta_0)\|^2 = O_p(n^{-1})$. In a similar way, it can be obtained that $\|P(\widehat{\theta}_{\phi_2}^{(l)}) - P(\theta_0)\|^2 = O_p(n^{-1})$. Then $Q_{\phi_1,\phi_2}^{(l)} = Z^T Z + o_p(1)$.

From (20), we have, under the contiguous alternative hypotheses (16), that

$$Z = \sqrt{n}(A_{(l+1)} - A_{(l)}) \operatorname{diag}\left(P(\theta_0)^{-1/2}\right)\left(\widehat{P} - P(\theta_0)\right) + o_p(1).$$

Now $\sqrt{n}(\widehat{P}-P(\theta_0)) = \sqrt{n}(\widehat{P}-P_n(\theta_0))+s$, so that $\sqrt{n}(\widehat{P}-P(\theta_0))\xrightarrow[n\to\infty]{L} N(s,\Sigma_{P(\theta_0)})$, and hence $Z\xrightarrow[n\to\infty]{L} N(\delta,\Sigma^*)$, where $\delta = (A_{(l+1)} - A_{(l)}) \operatorname{diag}(P(\theta_0)^{-1/2})s$ and

$$\begin{aligned}\Sigma^* &= (A_{(l+1)} - A_{(l)}) \operatorname{diag}\left(P(\theta_0)^{-1/2}\right)\Sigma_{P(\theta_0)} \operatorname{diag}\left(P(\theta_0)^{-1/2}\right)(A_{(l+1)} - A_{(l)})\\ &= (A_{(l+1)} - A_{(l)})\left(I - \sqrt{P(\theta_0)}\sqrt{P(\theta_0)^T}\right)(A_{(l+1)} - A_{(l)}).\end{aligned}$$

Using the results in the proof of Theorem 1 (Appendix), it can be shown that $\Sigma^* = (A_{(l)} - A_{(l+1)})$, and it is a projection of rank $(d_l - d_{l+1})$.

If we establish that $\Sigma^*\delta = \delta$, the theorem follows from the lemma on p.63 of Ferguson (1996), because in this case the non-centrality parameter is given by $\mu = \delta^T\delta$.

Applying (i) and (ii) given in the proof of Theorem 1 (Appendix), we have

$$\begin{aligned}\Sigma^*\delta &= (A_{(l)} - A_{(l+1)})\delta = A_{(l)}\delta - A_{(l+1)}\delta\\ &= A_{(l)}(A_{(l+1)}-A_{(l)}) \operatorname{diag}\Big(P(\theta_0)^{-1/2}\Big)s - A_{(l+1)}(A_{(l+1)}-A_{(l)}) \operatorname{diag}\Big(P(\theta_0)^{-1/2}\Big)s\\ &= \delta.\end{aligned}$$

Then the non-centrality parameter is $\mu = \delta^T\delta = s^T \operatorname{diag}(P(\theta_0)^{-1/2})(A_{(l)} - A_{(l+1)}) \operatorname{diag}(P(\theta_0)^{-1/2})s$.

Remark 1. Theorem 2 can be used to obtain an approximation to the power function of (6), as follows. Write $P(\theta^{(l)}) = P(\theta^{(l+1)})+\frac{1}{\sqrt{n}}(\sqrt{n}(P(\theta^{(l)})-P(\theta^{(l+1)})))$, and define $P_n(\theta^{(l)}) \equiv P(\theta^{(l+1)})+\frac{1}{\sqrt{n}}s$, where $s = (\sqrt{n}(P(\theta^{(l)})-P(\theta^{(l+1)})))$. Then substitute s into the definition of μ, and finally μ into the right side of (18).

The asymptotic non-central chi-squared approximation for power is checked for finite samples in the simulation given in Section 4. Figure 2 shows departures of the exact power from the asymptotic power for one particular choice (specified in Section 4) of H_{l+1} and H_l, for various choices of λ in $\phi_1 = \phi_{(\lambda)}$, and for small to large sample sizes (n = 15, 20, 25, 35, 50, 100, 200). Figure 2a represents nonpositive choices of λ and Figure 2b represents nonnegative choices of λ. These figures need to be interpreted in light of associated exact sizes; see Section 4.

However, it is immediately apparent that from an asymptotic-approximation point of view, $\lambda = 2/3$ seems to perform the best, particularly for small and moderate sample sizes.

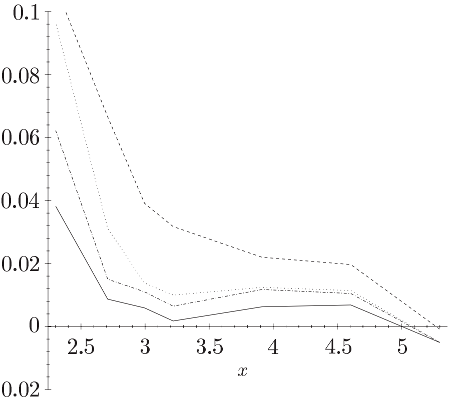

Figure 2a. (Exact power−Asymptotic power) as a function of $x = \log n$. Shown are $\lambda = -2$ (dashed line), $\lambda = -1$ (dotted line), $\lambda = -1/2$ (dash-dotted line), and $\lambda = 0$ (solid line).

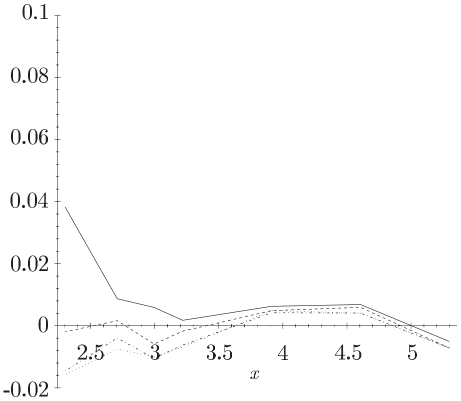

Figure 2b. (Exact power−Asymptotic power) as a function of $x = \log n$. Shown are $\lambda = 0$ (solid line), $\lambda = 2/3$ (dashed line), $\lambda = 1$ (dotted line), and $\lambda = 2$ (dash-dotted line).

Remark 2. If we consider the statistic $Q^{(l)}_{\phi_1,\phi_2}$ with $\phi_2(x) = \psi_{(0)}(x) = x \log x - x + 1$ and $\phi_1(x) = \psi_{(1)}(x) = (1-x)^2$, we obtain the classical Pearson statistic for testing loglinear models (e.g., Christensen (1997, p.338)). If we consider the statistic $Q^{(l)}_{\phi_1,\phi_2}$ with $\phi_2(x) = \psi_{(0)}(x) = x \log x - x + 1$ and $\phi_1(x) = \psi_{(0)}(x) = x \log x - x + 1$, we obtain the classical likelihood ratio statistic for testing loglinear models (e.g., Christensen (1997, p.338)). In this latter case, the result given by Theorem 2 was obtained for the first time by Oler (1985).

4. Simulation Study

We now describe briefly the finite-sample simulation study from which Figures 1 and 2 were obtained, and give new results that compare the powers of tests based on $\{Q^{(l)}_{\phi_{(\lambda)},\phi_{(0)}} : \lambda = -2,\ -1,\ -1/2,\ 0,\ 2/3,\ 1,\ 2\}$. Further details of the study can be found in an Ohio State University Technical Report by Cressie, Pardo, and Pardo (2001).

Consider a $2\times 2\times 2$ contingency table, so $k = 8$. We simulate data $X_1, \ldots, X_k$ from a multinomial distribution with sample size n and probability vector $P = (p_1, \ldots, p_k)^T$, where n and P are specified. The motivation for our simulation study comes from a similar one carried out by Oler (1985). For the moment, fix l

and consider the statistic $Q^{(l)}_{\phi_{(\lambda)},\phi_{(0)}}$ for testing $H_{Null} : H_{l+1}$ against $H_{Alt} : H_{l+1,n}$, and let $P_0 \in H_{Null}$ and $P_{1,n} \in H_{Alt}$, where $P_{1,n}$ is subscripted with n because its entries may depend on n. The essence of our simulation study is to obtain the exact probabilities,

$$\begin{aligned} \alpha_n^{(l)} &\equiv \Pr\left(Q^{(l)}_{\phi_{(\lambda)},\phi_{(0)}} > c \mid P_0\right) \\ \pi_n^{(l)} &\equiv \Pr\left(Q^{(l)}_{\phi_{(\lambda)},\phi_{(0)}} > c \mid P_{1,n}\right). \end{aligned} \tag{21}$$

In fact, $\alpha_n^{(l)}$ and $\pi_n^{(l)}$ are estimated using $N = 100,000$ simulations from the multinomial sampling schemes (n, P_0) and $(n, P_{1,n})$, respectively. For a given P_0 (see below), the various choices of n and $P_{1,n}$ represent the design of our simulation study. We choose $n = 15$, 20, 25, 35, 50, 100, 200, to represent small, moderate, and large sample sizes.

We simulate multinomial random vectors $(X_1, \ldots, X_k)$ and compute probabilities $\alpha_n^{(l)}$ for (n, P_0) and $\pi_n^{(l)}$ for $(n, P_{1,n})$. To see what happens for contiguous alternatives, we fix $P_1 \in H_l$ (see below) and define

$$P^*_{1,n} \equiv P_0 + (25/n)^{1/2}(P_1 - P_0). \tag{22}$$

Notice that $P^*_{1,25} = P_1$ and, as n increases, $P^*_{1,n}$ converges to P_0 at the rate $n^{-1/2}$; that is, $\{P^*_{1,n}\}$ is a sequence of contiguous alternatives. Our design for the simulation study is to choose $(n, P_{1,n})$ as both a fixed and a contiguous alternative, which we now give.

Contiguous alternatives: $\{(n, P^*_{1,n}) : n = 15, 20, 25, 35, 50, 100, 200\}$, where $P^*_{1,n}$ is given by (22) and P_1 is specified below.

Fixed alternatives: $\{(n, P_1) : n = 15, 20, 25, 35, 50, 100, 200\}$, where P_1 is specified below.

Notice that for $n < 25$, the contiguous alternatives are further from H_{Null} than are the fixed alternatives and that the two sequences share the alternative $(25, P_1)$. These choices allow reasonable coverage of the space of alternatives.

In the simulation study, we considered the same nested sequence of loglinear models considered by Oler (1985), and we chose what Oler called a "moderate value" for each main effect and a "small value" for the interactions. The four hypotheses we considered were:

$$H_1 : p_{ijk}(\theta) = \exp\{u+\theta_{1(i)}+\theta_{2(j)}+\theta_{3(k)}+\theta_{12(ij)}+\theta_{13(ik)}+\theta_{23(jk)}\}; \quad i, j, k = 1, 2$$

$$H_2 : p_{ijk}(\theta) = \exp\{u + \theta_{1(i)} + \theta_{2(j)} + \theta_{3(k)} + \theta_{12(ij)} + \theta_{13(ik)}\}; \quad i, j, k = 1, 2$$

$$H_3 : p_{ijk}(\theta) = \exp\{u + \theta_{1(i)} + \theta_{2(j)} + \theta_{3(k)} + \theta_{12(ij)}\}; \quad i, j, k = 1, 2$$

$$H_4 : p_{ijk}(\theta) = \exp\{u + \theta_{1(i)} + \theta_{2(j)} + \theta_{3(k)}\}; \quad i, j, k = 1, 2,$$

where $\exp(\theta_{1(1)}) = \exp(\theta_{2(1)}) = \exp(\theta_{3(1)}) = 5/6$ and $\exp(\theta_{12(11)}) = \exp(\theta_{13(11)}) = \exp(\theta_{23(11)}) = 9/10$. Here, $\exp(-u)$ is the normalizing constant and the subscripted $\theta-$terms add to zero over each of their indices.

In Section 2, we showed Figure 1 (Exact size $-$ Nominal size of 0.05) for $H_{Null} : H_4$, using the test statistic $Q^{(3)}_{\phi(\lambda),\phi(0)}$ and $c = \chi^2_1(0.95)$. In Section 3, we showed Figure 2 (Exact power $-$ Asymptotic power) for $H_{Null} : P_0 \in H_4$ and $H_{Alt} : P^*_{1,n}$, with $P_1 \in H_3$, using the test statistic $Q^{(3)}_{\phi(\lambda),\phi(0)}$ and $c = \chi^2_1(0.95)$.

In the simulation study, we compare members of the power-divergence family of test statistics. We use two basic criteria for a good performance. The first is good exact power and size for small to moderate sample sizes. For this, we consider $H_{Null} : P_0 \in H_{l+1}$ versus $H_{Alt} : (n, P_1)$, where $P_1 \in H_l$ and $n = 15$, 20, 25, 35; $l = 1, 2, 3$. The second is good agreement of exact and asymptotic probabilities for small to moderate sample sizes. For this, we consider $H_{Null} : P_0 \in H_{l+1}$ versus $H_{Alt} : (n, P^*_{1,n})$, where $P^*_{1,n}$ is given by (22), $P_1 \in H_l$, and $n = 15$, 20, 25, 35; $l = 1, 2, 3$. The complete results of the simulation study are given in Cressie, Pardo and Pardo (2001).

We have chosen to concentrate below on tests associated with $H_{Null} : P_0 \in H_3$, to illustrate the type of results obtained; see Table 1.

Table 1. Entries show results from the simulation study as a function of multinomial sample-size parameter (n) and power-divergence parameter (λ). The notation (a) corresponds to (Exact size) for testing $H_{Null} : P_0 \in H_3$. The notation (b) corresponds to (Exact power-Asymptotic power) for testing $H_{Null} : P_0 \in H_3$ versus $H_{Alt} : P^*_{1,n}$, where $P_1 \in H_2$. The notation (c) corresponds to (Exact power-Exact size) for testing $H_{Null} : P_0 \in H_3$ versus $H_{Alt} : P_1 \in H_2$.

		λ						
	n	-2	-1	$-1/2$	0	2/3	1	2
(a)	15	0.15232	0.14335	0.10416	0.08450	0.05575	0.04479	0.04620
	20	0.11111	0.08752	0.07375	0.06645	0.06017	0.05290	0.05661
	25	0.08060	0.06277	0.06065	0.05690	0.05005	0.04743	0.04727
	35	0.07112	0.06039	0.05498	0.05217	0.04961	0.04633	0.04658
(b)	15	0.10853	0.09573	0.06041	0.03651	-.00233	-.01807	-.01643
	20	0.06595	0.03407	0.02055	0.01431	0.00659	-.00399	-.00203
	25	0.04227	0.02056	0.01811	0.01385	-.00096	-.00519	-.00652
	35	0.02614	0.00575	0.00199	-.00178	-.00741	-.01211	-.01189
(c)	15	0.01810	0.01672	0.01978	0.01629	0.01354	0.01152	0.01113
	20	0.02771	0.02043	0.02361	0.02418	0.02313	0.02004	0.01807
	25	0.04384	0.03996	0.03963	0.03912	0.03116	0.02955	0.02838
	35	0.05895	0.05198	0.05361	0.05031	0.04721	0.04557	0.04557

First of all, we study the closeness of the exact size to the nominal size $\alpha = 0.05$. Following Dale (1986), consider the inequality,

$$|\text{logit}(1 - \alpha_n^{(l)}) - \text{logit}(1 - \alpha)| \leq e, \tag{23}$$

where $\text{logit}(p) \equiv \ln(p/(1-p))$. The two probabilities are considered to be "close" if they satisfy (23) with $e = 0.35$ and "fairly close" if they satisfy (23) with $e = 0.7$. Note that for $\alpha = 0.05$, $e = 0.35$ corresponds to $\alpha_n^{(l)} \in [0.0357, 0.0695]$, and $e = 0.7$ corresponds to $\alpha_n^{(l)} \in [0.0254, 0.0959]$. From Table 1, the statistics that satisfy (23) for $e = 0.35$ are those corresponding to $\lambda = 2/3$, 1, 2. For $e = 0.7$, only one extra statistic, that corresponding to $\lambda = 0$, is added.

Now consider the comparison of exact power and asymptotic power under a contiguous alternative. From Table 1, the test statistic corresponding to $\lambda = 2/3$ has the best behavior. We also consider the difference between exact power and exact size as a measure of how quickly the power curve increases from its probability of type I error. From Table 1, the increase in power is a little more for tests based on negative λ than for positive λ. This should be tempered with the fact that for negative λ the exact size is considered not even "fairly close". This trade-off between size behavior and power behavior is a classical problem in hypothesis testing.

The results given in Table 1 are illustrative of what happens for the null hypotheses H_2 and H_4. Due to space considerations, we cannot present them all here, however the interested reader can consult Figures 3, 4, 5, and 6 in Cressie, Pardo and Pardo (2001) for the complete results.

In what follows, we consider only the statistics that satisfy (23) with $e = 0.7$, and to discriminate between them we calculate

$$g_1(\lambda) \equiv \left| AP_{i,n}^{(l)}(\lambda) - SEP_{i,n}^{(l)}(\lambda) \right| \quad \text{and} \quad g_2(\lambda) \equiv \left(SEP_{i,n}^{(l)}(\lambda) - STS_{i,n}^{(l)}(\lambda) \right)^{-1},$$

where $AP_{i,n}^{(l)}(\lambda)$ is the asymptotic power, $SEP_{i,n}^{(l)}(\lambda)$ is the simulated exact power, and $STS_{i,n}^{(l)}(\lambda)$ is the simulated test size of the statistic $Q_{\phi(\lambda),\phi(0)}^{(l)}$; $l = 1, 2, 3$, under the alternative $i = F$ (fixed), C(contiguous), and $n = 15$, 20, 25, 35. Then, for a given l, we consider a statistic $Q_{\phi(\lambda_1),\phi(0)}^{(l)}$ to be better than a statistic $Q_{\phi(\lambda_2),\phi(0)}^{(l)}$ iff

$$g_1(\lambda_1) < g_1(\lambda_2) \quad \text{and} \quad g_2(\lambda_1) < g_2(\lambda_2). \tag{24}$$

In Figure 3, we plot $y = g_2(\lambda)$ versus $x = g_1(\lambda)$, for $l = 2$; from (24), we look for values of λ that are as close to $(0, 0)$ as possible in the (x, y) plane.

The points $(g_1(\lambda), g_2(\lambda))$ far away from $(0, 0)$ are those corresponding to smallest sample size $n = 15$, as expected. For this sample size, Table 1 shows that the exact size of the tests based on $Q_{\phi(0),\phi(0)}^{(l)}$ is too large in relation to that of $Q^{(l)}_{\phi(\lambda),\phi(0)}$, for $\lambda = 2/3$, 1, 2. For $n = 15$, the points $(g_1(2/3), g_2(2/3))$ are closer to $(0, 0)$ than the points $(g_1(1), g_2(1))$ and $(g_1(2), g_2(2))$. Thus, according to the criterion (24), the test based on $Q^{(l)}_{\phi(2/3),\phi(0)}$ is best for $n = 15$.

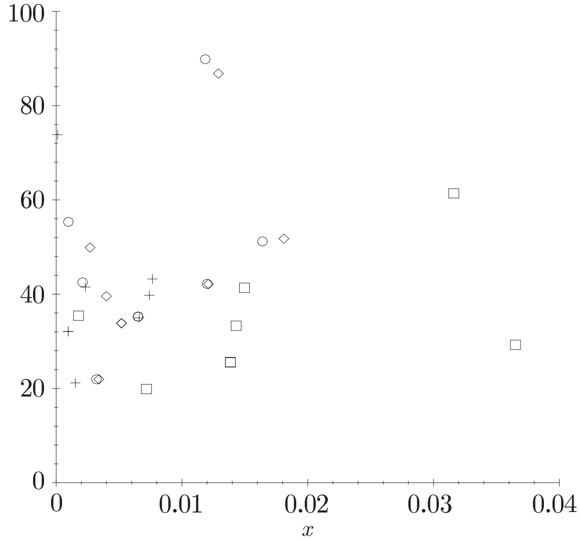

Figure 3. $y = g_2(\lambda)$ versus $x = g_1(\lambda)$ for $Q^{(2)}_{\phi(\lambda),\phi(0)}$. Shown are $\lambda = 0$ (Square), $\lambda = 2/3$ (Cross), $\lambda = 1$ (Diamond), and $\lambda = 2$ (Circle).

For $n = 20$, 25, 35, it can be seen that $Q^{(l)}_{\phi(2/3),\phi(0)}$ is better than $Q^{(l)}_{\phi(\lambda),\phi(0)}$, $\lambda = 1$, 2, according to (24). However, $Q^{(l)}_{\phi(2/3),\phi(0)}$ is not obviously better than $Q^{(l)}_{\phi(0),\phi(0)}$, since $g_1(2/3) < g_1(0)$ but $g_2(0) < g_2(2/3)$. Similar conclusions hold for $l = 1$ and $l = 3$; see Figure 6 in Cressie, Pardo and Pardo (2001).

From the simulation studies we have carried out, our conclusion is that the test based on $Q^{(l)}_{\phi(2/3),\phi(0)}$ is a very good, and often better, alternative to the tests based on the classical statistics $Q^{(l)}_{\phi(\lambda),\phi(0)}$ with $\lambda = 0$, 1.

Acknowledgements

The research in this paper was supported in part by a U.S. Environmental Protection Agency grant R827-257-01-0 (N. Cressie), and by a Spanish agency grant DGES BMF2000-0800 (L. Pardo and M. C. Pardo). We would like to thank the referees and editors for their helpful comments.

Appendix

Proof of Theorem 1. In a derivation similar to that given in Theorem 3 of Cressie and Pardo (2000), it can be established that $Q^{(l)}_{\phi_1,\phi_2} = \sqrt{n}(\widehat{P} - P(\theta_0))^T M_l^T M_l \sqrt{n}(\widehat{P} - P(\theta_0)) + o_p(1)$ and $Q^{(l-1)}_{\phi_1,\phi_2} = \sqrt{n}(\widehat{P} - P(\theta_0))^T M_{l-1}^T M_{l-1} \sqrt{n}$ $(\widehat{P} - P(\theta_0)) + o_p(1)$, where $M_i = (A_{(i+1)} - A_{(i)})\,\mathrm{diag}(P(\theta_0)^{-1/2})$ for $i = l-1, l$, $A_{(i)} \equiv \mathrm{diag}(P(\theta_0)^{-1/2})\Sigma_{P(\theta_0)} W_{(i)} (W_{(i)}^T \Sigma_{P(\theta_0)} W_{(i)})^{-1} W_{(i)}^T \Sigma_{P(\theta_0)}\,\mathrm{diag}(P(\theta_0)^{-1/2})$ for $i = l-1, l, l+1$, $\Sigma_{P(\theta_0)} = \mathrm{diag}(P(\theta_0)) - P(\theta_0)P(\theta_0)^T$, and $W_{(i)}$ is the matrix associated with the i-th loglinear model for $i = l-1, l, l+1$.

Now, because $\sqrt{n}(\widehat{P} - P(\theta_0)) \xrightarrow[n\to\infty]{L} N(0, \Sigma_{P(\theta_0)})$, from Theorem 4 in Searle (1971, p.59), $Q^{(l)}_{\phi_1,\phi_2}$ and $Q^{(l-1)}_{\phi_1,\phi_2}$, are asymptotically independent if $M_{l-1}^T M_{l-1}$ $\Sigma_{P(\theta_0)} M_l^T M_l = 0$. We have

$$\begin{aligned} &M_{l-1}^T M_{l-1} \Sigma_{P(\theta_0)} M_l^T M_l \\ &= M_{l-1}^T (A_{(l)} - A_{(l-1)}) \left(I - \sqrt{P(\theta_0)}\sqrt{P(\theta_0)^T} \right) (A_{(l+1)} - A_{(l)}) M_l \\ &= M_{l-1}^T (A_{(l)} - A_{(l-1)})(A_{(l+1)} - A_{(l)}) M_l, \end{aligned}$$

since $A_{(i)}\sqrt{P(\theta_0)} = 0$ for $i = l-1, l, l+1$.

But $\{A_{(i)} : i = l-1, l, l+1\}$ are orthogonal projection operators, and the column space of $W_{(i+1)}$ is a subspace of the column space of $W_{(i)}$. Thus (i) $A_{(i)}A_{(i+1)} = A_{(i+1)}A_{(i)} = A_{(i+1)}$ for $i = l-1, l$, and (ii) $A_{(i)}A_{(i)} = A_{(i)}$ for $i = l-1, l, l+1$.

We have $M_{l-1}^T M_{l-1} \Sigma_{P(\theta_0)} M_l^T M_l = 0$.

References

Agresti, A. (1990). *Categorical Data Analysis.* John Wiley, New York.

Ali, S.M. and Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *J. Roy. Statist. Soc. Ser. B* **286**, 131-142.

Christensen, R. (1997). *Log-Linear Model and Logistic Regression.* Springer-Verlag, New York.

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

Cressie, N. and Pardo, L. (2000). Minimum $\phi-$divergence estimator and hierarchical testing in loglinear models. *Statist. Sinica* **10,** 867-884.

Cressie, N. and Pardo, L. (2002). Model checking in loglinear models using ϕ-divergences and MLEs. *J. Statist. Plann. Inference* **103,** 437-453.

Cressie, N., Pardo, L. and Pardo, M.C. (2001). Size and power considerations for testing loglinear models using $\phi-$divergence test statistics. Technical Report No. 680. Department of Statistics, The Ohio State University, Columbus, OH.

Csiszár, I. (1963). Eine Informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. *Publications of the Mathematical Institute of Hungarian Academy of Sciences A* **8**, 85-108.

Dale, J. R. (1986). Asymptotic normality of goodness-of-fit statistics for sparse product multinomials. *J. Roy. Statist. Soc. Ser. B* **41**, 48-59.

Fenech, A. P. and Westfall, P. H. (1988). The power function of conditional log-linear model tests. *J. Amer. Statist. Assoc.* **83,** 198-203.

Ferguson, T. S. (1996). *A Course in Large Sample Theory.* John Wiley, New York.

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

Hájek, J. and Z. Sidák (1967). *Theory of Rank Tests.* Academic Press, New York.

Kullback, S. (1985). Kullback information. In *Encyclopedia of Statistical Sciences* **4** (Edited by S. Kotz and N. L. Johnson), 421-425. John Wiley, New York.

Matusita, K. (1954). On the estimation by the minimum distance method. *Ann. Inst. Statist. Math.* **5,** 59-65.

Menéndez, M. L., Morales, D., Pardo, L. and Zografos, K. (1999). Statistical inference for finite Markov chains based on divergences. *Statist. Probab. Lett.* **41**, 9-17.

Morales, D., Pardo, L. and Vajda, I. (1995). Asymptotic divergences of estimates of discrete distributions. *J. Statist. Plann. Inference* **48,** 347-369.

Neyman, J. (1949). Contribution to the theory of the χ^2 test. *Proceedings of the First Berkeley Symposium on Mathematical Statistics and Probability,* 239-275.

Oler, J. (1985). Noncentrality parameters in chi-squared goodness-of-fit analyses with an application to log-linear procedures. *J. Amer. Statist. Assoc.* **80,** 181-189.

Pardo, L., Pardo, M. C. and Zografos, K. (2001). Minimum phi-divergence estimator for homogeneity in multinomial populations. *Sankhyā*, A **63**, 72-92.

Read, T. R. C. and Cressie, N. A. C. (1988). *Goodness-of-fit Statistics for Discrete Multivariate Data.* Springer-Verlag, New York.

Searle, S. R. (1971). *Linear Models.* John Wiley, New York.

Department of Statistics, the Ohio State University, Columbus, Ohio 43210, U.S.A.

E-mail: ncressie@stat.ohio-state.edu

Department of Statistics and O.R., Complutense University of Madrid, 28040 Madrid, Spain.

E-mail: Leandro_Pardo@Mat.ucm.es

Department of Statistics and O.R. Complutense University of Madrid, 28040 Madrid, Spain.

E-mail: mcapardo@mat.ucm.es

(Received September 2001; accepted August 2002)