Statistica Sinica **10**(2000), 867-884

MINIMUM ϕ-DIVERGENCE ESTIMATOR AND HIERARCHICAL TESTING IN LOGLINEAR MODELS

Noel Cressie and Leandro Pardo

The Ohio State University and Complutense University of Madrid

Abstract: In this paper we consider inference based on very general divergence measures, under assumptions of multinomial sampling and loglinear models. We define the minimum ϕ-divergence estimator, which is seen to be a generalization of the maximum likelihood estimator. This estimator is then used in a ϕ-divergence goodness-of-fit statistic, which is the basis of two new statistics for solving the problem of testing a nested sequence of loglinear models.

Key words and phrases: Asymptotic distributions, Framinghan heart study, multinomial distribution, nested hypotheses, power divergence, Renyi divergence.

1. Introduction

We consider the problem of statistical inference for multinomial data, and we devise a general theory for parameter estimation and goodness-of-fit. Further, we show how the theory can be applied to the problem of inference on a nested sequence of loglinear models. Consider a sample $Y_1, \ldots, Y_n$ of size $n \in N$ 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 $\Theta \subset R^t$ $(t < k-1)$. In other words, the true value θ_0 of parameter $\theta = (\theta_1, \ldots, \theta_t)^T \in \Theta$ is assumed to be unknown. We denote $\widehat{P} = (\widehat{p}_1, \ldots, \widehat{p}_k)^T$ with

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

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

$$P(X_1 = x_1, \ldots, X_k = x_k) = \frac{n!}{x_1! \cdots x_k!} p_1(\theta_0)^{x_1} \times \cdots \times p_k(\theta_0)^{x_k}, \tag{2}$$

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

If $\sum_{j=1}^k \widehat{p}_j \log p_j(\theta)$ is almost surely (a.s.) maximized over Θ at some $\widehat{\theta}$, then $\widehat{\theta}$ is the point maximum likelihood estimator (MLE). The MLE can equivalently be defined by,

$$\widehat{\theta} = \arg\min_{\theta\in\Theta} D(\widehat{P}, P(\theta)) \text{ a.s.}, \tag{3}$$

where $D(P,Q) = \sum_{j=1}^k p_j \log \frac{p_j}{q_j}$ is the Kullback-Leibler divergence and $P = (p_1,\ldots,p_k)^T$, $Q = (q_1,\ldots,q_k)^T$. This divergence is a particular case of the ϕ-divergence defined by Csiszár (1967):

$$D_\phi(P,Q) = \sum_{j=1}^k q_j \phi\Big(\frac{p_j}{q_j}\Big); \phi \in \Phi^*, \tag{4}$$

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 ϕ-divergences in statistical problems is the power divergence family,

$$\begin{aligned} &\phi_{(\lambda)}(x) = (\lambda(\lambda+1))^{-1}(x^{\lambda+1} - x);\ \lambda \neq 0, \lambda \neq -1, \\ &\phi_{(0)}(x) = \lim_{\lambda\to 0} \phi_{(\lambda)}(x),\ \phi_{(-1)}(x) = \lim_{\lambda\to -1} \phi_{(\lambda)}(x), \end{aligned} \tag{5}$$

introduced and studied by Cressie and Read (1984). We can 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 $I^\lambda(P,Q) \equiv D_{\phi_{(\lambda)}}(P,Q) = D_{\psi_{(\lambda)}}(P,Q)$.

As a generalization of the MLE $\widehat{\theta}$, Cressie and Read (1984) considered the minimum power-divergence estimator $\widehat{\theta}_{(\lambda)} \equiv \arg\min_{\theta\in\Theta} I^\lambda(\widehat{P}, P(\theta))$, and studied its properties. Notice that when $\lambda \to 0$, the MLE is obtained and, for $\lambda = 1$, the minimum chi-squared estimator is obtained.

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

$$\widehat{\theta}_\phi = \arg\min_{\theta\in\Theta} D_\phi(\widehat{P}, P(\theta)), \tag{6}$$

and studied its properties. For testing whether the data are generated by a probability distribution contained in $\mathcal{P}$, it is natural then to use as goodness-of-fit statistic, $D_\phi(\widehat{P}, P(\widehat{\theta}_\phi))$. It is a generalization of the likelihood-ratio statistic

and Pearson's chi-squared statistic (Cressie and Read (1984); Morales, Pardo and Vajda (1995)).

In what is to follow, we assume that $P(\theta)$ belongs to the general class of loglinear models. That is, we assume

$$p_u(\theta) = \exp(w_u^T\theta)/\sum_{v=1}^{k}\exp(w_v^T\theta);\ u = 1,\ldots,k, \tag{7}$$

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 $(1,\ldots,1)^T$.

Then we are interested in the properties of $\widehat{\theta}_\phi$ given by (6) and how $\widehat{\theta}_\phi$ might be used in (4) to test a nested sequence of hypotheses,

$$H_l : \theta \in \Theta_l;\ l = 1,\ldots,m,\ m \le t < k-1, \tag{8}$$

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

$$d_m < d_{m-1} < \cdots < d_1 \le t. \tag{9}$$

Our strategy is to test successively the hypotheses H_{l+1} against H_l; $l = 1,\ldots,m-1$, as null and alternative hypotheses respectively. We continue to test as long as the null hypothesis is accepted and choose the loglinear model Θ_l according to the first l for which H_{l+1} is rejected (as a null hypothesis) in favour of H_l (as an alternative hypothesis). This strategy is quite standard for nested models (Read and Cressie (1988, p.42)). The nesting occurs naturally because of the hierarchical principle, which says that interactions should not be fitted unless the corresponding main effects are present (e.g., Collett (1994, p.78)).

In Section 2, we derive asymptotic results for minimum divergence estimators under multinomial sampling and loglinear model assumptions (7). In Section 3, the main result for hierarchical testing in loglinear models using general divergence measures is given. In Section 4, we present an example to demonstrate how the results of Sections 2 and 3 can be applied in practice.

2. Asymptotic Results for Minimum ϕ-divergence Estimators under the Loglinear Model

In this section, we present some asymptotic results for the minimum ϕ-divergence estimator under the loglinear model (7).

Theorem 1. *If $\widehat{\theta}_\phi$ is the minimum ϕ-divergence estimator for the loglinear model, $p_u(\theta) = \exp(w_u^T\theta)/\sum_{v=1}^{k}\exp(w_v^T\theta)$; $u = 1,\ldots,k$, we have*

$$n^{1/2}(\widehat{\theta}_\phi - \theta_0) \xrightarrow[n\to\infty]{L} N\Big(0, (W^T\Sigma_{P(\theta_0)}W)^{-1}\Big),$$

where $\Sigma_P \equiv diag(P) - PP^T$ *and* "$\xrightarrow[n\to\infty]{L}$" *denotes convergence in law* (*or distribution*).

Proof. By Morales, Pardo and Vajda (1995), we know that

$$n^{1/2}(\widehat{\theta}_\phi - \theta_0) \xrightarrow[n\to\infty]{L} N(0, I_F(\theta_0)^{-1}),$$

where $I_F(\theta_0)$ is the Fisher Information matrix, given by

$$I_F(\theta_0) = \Big(\sum_{j=1}^{k} \frac{1}{p_j(\theta_0)} \frac{\partial p_j(\theta_0)}{\partial \theta_r} \frac{\partial p_j(\theta_0)}{\partial \theta_s} \Big)_{r,s=1,\ldots,t} = A(\theta_0)^T A(\theta_0),$$

with $A(\theta_0) = diag(P(\theta_0)^{-1/2}) \frac{\partial P(\theta_0)}{\partial \theta}$ being a $k \times k$ diagonal matrix times a $k \times t$ matrix $\frac{\partial P(\theta_0)}{\partial \theta} \equiv \Big(\frac{\partial p_j(\theta_0)}{\partial \theta_r} \Big)_{\substack{j=1,\ldots,k \\ r=1,\ldots,t}}$.

For the loglinear models we have $\frac{\partial p_j(\theta_0)}{\partial \theta_r} = p_j(\theta_0) w_{rj} - p_j(\theta_0) \sum_{v=1}^{k} w_{rv} p_v(\theta_0)$. Then $\frac{\partial P(\theta_0)}{\partial \theta} = (diag(P(\theta_0)) - P(\theta_0)P(\theta_0)^T)W = \Sigma_{P(\theta_0)} W$, and hence $A(\theta_0) = diag(P(\theta_0)^{-1/2}) \Sigma_{P(\theta_0)} W$. Finally, $I_F(\theta_0) = A(\theta_0)^T A(\theta_0) = W^T \Sigma_{P(\theta_0)} W$, and we have

$$n^{1/2}(\widehat{\theta}_\phi - \theta_0) \xrightarrow[n\to\infty]{L} N(0, (W^T \Sigma_{P(\theta_0)} W)^{-1}).$$

Remark 1. By Morales, Pardo and Vajda (1995), we know that the asymptotic expansion of the minimum ϕ-divergence estimator of θ_0 under $H_0 : \theta \in \Theta_0$ is,

$$\widehat{\theta}_\phi = \theta_0 + I_F(\theta_0)^{-1} A(\theta_0)^T diag(P(\theta_0)^{-1/2})(\widehat{P} - P(\theta_0)) + o(\|\widehat{P} - P(\theta_0)\|),$$

where we recall that $\widehat{P} = (X_1/n, \ldots, X_k/n)^T$.

Then, using the theorem above, we see that for loglinear models,

$$\widehat{\theta}_\phi = \theta_0 + I_F(\theta_0)^{-1} W^T \Sigma_{P(\theta_0)} diag(P(\theta_0)^{-1})(\widehat{P} - P(\theta_0)) + o(\|\widehat{P} - P(\theta_0)\|).$$

Another interesting result, useful later in this paper, is the following.

Theorem 2. *If* $\widehat{\theta}_\phi$ *is the minimum* ϕ*-divergence estimator for the loglinear model,* $p_u(\theta) = \exp(w_u^T \theta) / \sum_{v=1}^{k} \exp(w_v^T \theta)$; $u = 1, \ldots, k$, *then*

$$n^{1/2}(P(\widehat{\theta}_\phi) - P(\theta_0)) \xrightarrow[n\to\infty]{L} N(0, \Sigma_{P(\theta_0)} W (W^T \Sigma_{P(\theta_0)} W)^{-1} W^T \Sigma_{P(\theta_0)}).$$

Proof. We know that $(P(\widehat{\theta}_\phi) - P(\theta_0)) = \frac{\partial P(\theta_0)}{\partial \theta}(\widehat{\theta}_\phi - \theta_0) + o(\|\widehat{\theta}_\phi - \theta_0\|)$, and because $n^{1/2} o(\|\widehat{\theta}_\phi - \theta_0\|) = o_p(1)$, the random variables $n^{1/2}(P(\widehat{\theta}_\phi) - P(\theta_0))$ and $n^{1/2} \frac{\partial P(\theta_0)}{\partial \theta}(\widehat{\theta}_\phi - \theta_0)$ have the same asymptotic distribution. Using the

previous theorem, we have $n^{1/2}(P(\widehat{\theta}_\phi) - P(\theta_0)) \xrightarrow[n\to\infty]{L} N(0, \Sigma^0)$, where $\Sigma^0 = \frac{\partial P(\theta_0)}{\partial \theta}(W^T \Sigma_{P(\theta_0)} W)^{-1}(\frac{\partial P(\theta_0)}{\partial \theta})^T$. But, in the proof of Theorem 1, we saw that $\frac{\partial P(\theta_0)}{\partial \theta} = \Sigma_{P(\theta_0)} W$, and hence the proof is complete.

3. Testing a Hierarchical Sequence of Loglinear Models

Assuming that the multinomial data $\{X_j : j = 1, \ldots, k\}$ follow the loglinear model (7), we now consider the problem of testing the nested sequence of hypotheses given by (8) and (9). The suggested hypothesis-testing strategy is to test successively the null hypothesis H_{l+1} versus H_l; $l = 1, \ldots, m-1$, choosing the first l for which H_{l+1} is rejected in favor of H_l. To solve this problem, we make use of the following results.

Theorem 3. *Suppose that data $(X_1, \ldots, X_k)$ is multinomially distributed according to* (2) *and* (7). *Consider the nested sequence of hypotheses given by* (8) *and* (9). *Choose functions $\phi_1, \phi_2 \in \Phi$. Then, for testing hypotheses, $H_0 : H_{l+1}$ against $H_1 : H_l$, the asymptotic null distribution of the test statistic,*

$$T^{(l)}_{\phi_1,\phi_2} \equiv \frac{2n}{\phi_1''(1)} D_{\phi_1}(P(\widehat{\theta}^{(l+1)}_{\phi_2}), P(\widehat{\theta}^{(l)}_{\phi_2})) \tag{10}$$

is a chi-squared distribution with $d_l - d_{l+1}$ degrees of freedom, $l = 1, \ldots, m-1$. In (10), $\widehat{\theta}^{(l)}_{\phi_2}$ *and* $\widehat{\theta}^{(l+1)}_{\phi_2}$ *are the minimum ϕ_2-divergence estimators under the models H_l and H_{l+1}, respectively, where the minimum ϕ-divergence estimators are defined by* (6).

Proof. The second-order expansion of $D_{\phi_1,\phi_2} \equiv D_{\phi_1}(P(\widehat{\theta}^{(l+1)}_{\phi_2}), P(\widehat{\theta}^{(l)}_{\phi_2}))$ about $(P(\theta_0), P(\theta_0))$ gives

$$\begin{aligned}
D_{\phi_1,\phi_2} = {} & D_{\phi_1}(P(\theta_0), P(\theta_0)) \\
& + \sum_{j=1}^{k} \Big(\frac{\partial D_{\phi_1}(P,Q)}{\partial p_j}\Big)_{(P(\theta_0),P(\theta_0))} (p_j(\widehat{\theta}^{(l+1)}_{\phi_2}) - p_j(\theta_0)) \\
& + \sum_{j=1}^{k} \Big(\frac{\partial D_{\phi_1}(P,Q)}{\partial q_j}\Big)_{(P(\theta_0),P(\theta_0))} (p_j(\widehat{\theta}^{(l)}_{\phi_2}) - p_j(\theta_0)) \\
& + \frac{1}{2} \sum_{j=1}^{k} \Big(\frac{\partial^2 D_{\phi_1}(P,Q)}{\partial p_j^2}\Big)_{(P(\theta_0),P(\theta_0))} (p_j(\widehat{\theta}^{(l+1)}_{\phi_2}) - p_j(\theta_0))^2 \\
& + \frac{1}{2} \sum_{j=1}^{k} \Big(\frac{\partial^2 D_{\phi_1}(P,Q)}{\partial q_j^2}\Big)_{(P(\theta_0),P(\theta_0))} (p_j(\widehat{\theta}^{(l)}_{\phi_2}) - p_j(\theta_0))^2 \\
& + \sum_{i=1}^{k} \sum_{j=1}^{k} \Big(\frac{\partial^2 D_{\phi_1}(P,Q)}{\partial p_i \partial q_j}\Big)_{(P(\theta_0),P(\theta_0))} (p_i(\widehat{\theta}^{(l+1)}_{\phi_2}) - p_i(\theta_0))(p_j(\widehat{\theta}^{(l)}_{\phi_2}) - p_j(\theta_0))
\end{aligned}$$

$$+o(\|P(\widehat{\theta}_{\phi_2}^{(l+1)}) - P(\theta_0)\|^2 + \|P(\widehat{\theta}_{\phi_2}^{(l)}) - P(\theta_0)\|^2).$$

Recall that $\phi_1(1) = 0$, by assumption, and hence it is easy to see that $D_{\phi_1}(P(\theta_0), P(\theta_0)) = 0$. Furthermore,

$$\sum_{j=1}^{k} \Big(\frac{\partial D_{\phi_1}(P,Q)}{\partial p_j}\Big)_{(P(\theta_0),P(\theta_0))} (p_j(\widehat{\theta}_{\phi_2}^{(l+1)}) - p_j(\theta_0)) = 0.$$

This is because $D_{\phi_1}(P,Q) = \sum_{j=1}^{k} q_j \phi_1(\frac{p_j}{q_j})$, and hence $(\frac{\partial D_{\phi_1}(P,Q)}{\partial p_j})_{(P(\theta_0),P(\theta_0))} = \phi_1'(1)$.

Similar calculations yield second partial derivatives. Thus,

$$T_{\phi_1,\phi_2}^{(l)} = \frac{2n}{\phi_1''(1)} D_{\phi_1}(P(\widehat{\theta}_{\phi_2}^{(l+1)}), P(\widehat{\theta}_{\phi_2}^{(l)}))$$

can be written as,

$$\begin{aligned}
T_{\phi_1,\phi_2}^{(l)} &= n \sum_{j=1}^{k} \frac{1}{p_j(\theta_0)} (p_j(\widehat{\theta}_{\phi_2}^{(l+1)}) - p_j(\theta_0))^2 \\
&\quad + n \sum_{j=1}^{k} \frac{1}{p_j(\theta_0)} (p_j(\widehat{\theta}_{\phi_2}^{(l)}) - p_j(\theta_0))^2 \\
&\quad - 2n \sum_{j=1}^{k} \frac{1}{p_j(\theta_0)} (p_j(\widehat{\theta}_{\phi_2}^{(l+1)}) - p_j(\theta_0))(p_j(\widehat{\theta}_{\phi_2}^{(l)}) - p_j(\theta_0)) \\
&\quad + n\ o(\|P(\widehat{\theta}_{\phi_2}^{(l+1)}) - P(\theta_0)\|^2 + \|P(\widehat{\theta}_{\phi_2}^{(l)}) - P(\theta_0)\|^2) \\
&= (n^{1/2} diag(P(\theta_0)^{-1/2})(P(\widehat{\theta}_{\phi_2}^{(l+1)}) - P(\widehat{\theta}_{\phi_2}^{(l)})))^T \cdot \\
&\quad \cdot (n^{1/2} diag(P(\theta_0)^{-1/2})(P(\widehat{\theta}_{\phi_2}^{(l+1)}) - P(\widehat{\theta}_{\phi_2}^{(l)}))) \\
&\quad + n\ o(\|P(\widehat{\theta}_{\phi_2}^{(l+1)}) - P(\theta_0)\|^2 + \|P(\widehat{\theta}_{\phi_2}^{(l)}) - P(\theta_0)\|^2).
\end{aligned}$$

We know from Theorem 2 that, under the loglinear model (7) and the null hypothesis H_{l+1}, $n^{1/2}(P(\widehat{\theta}_{\phi_2}^{(l+1)}) - P(\theta_0)) \xrightarrow[n\to\infty]{L} N(0, \Sigma^*)$, with $\Sigma^* = \Sigma_{P(\theta_0)} W_{(l+1)} (W_{(l+1)}^T \Sigma_{P(\theta_0)} W_{(l+1)})^{-1} W_{(l+1)}^T \Sigma_{P(\theta_0)}$, and $W_{(l+1)}$ the loglinear model matrix of explanatory variables under the null hypothesis H_{l+1}.

Then $\|P(\widehat{\theta}_{\phi_2}^{(l+1)}) - P(\theta_0)\|^2 = O_p(n^{-1})$, and because it is assumed that $\theta_0 \in \Theta_{l+1} \subset \Theta_l$, we also have that $\|P(\widehat{\theta}_{\phi_2}^{(l)}) - P(\theta_0)\|^2 = O_p(n^{-1})$. Consequently, $n\, o(\|P(\widehat{\theta}_{\phi_2}^{(l+1)}) - P(\theta_0)\|^2 + \|P(\widehat{\theta}_{\phi_2}^{(l)}) - P(\theta_0)\|^2) = o_p(1)$, and hence the asymptotic distribution of the statistic $T_{\phi_1,\phi_2}^{(l)}$ is the same as the asymptotic distribution of the random variable $Z^T Z$, where $Z \equiv n^{1/2} diag(P(\theta_0)^{-1/2})(P(\widehat{\theta}_{\phi_2}^{(l+1)}) - P(\widehat{\theta}_{\phi_2}^{(l)}))$.

Now, from Theorem 2, $(P(\widehat{\theta}^{(i)}_{\phi_2})-P(\theta_0)) = \Sigma_{P(\theta_0)}W_{(i)}(\widehat{\theta}^{(i)}_{\phi_2}-\theta_0)+o(\|\widehat{\theta}^{(i)}_{\phi_2}-\theta_0\|)$, $i = l, l+1$. Then, using Remark 1, we obtain

$$\begin{aligned} P(\widehat{\theta}^{(l+1)}_{\phi_2}) - P(\widehat{\theta}^{(l)}_{\phi_2}) = &\left\{\Sigma_{P(\theta_0)}W_{(l+1)}(W^T_{(l+1)}\Sigma_{P(\theta_0)}W_{(l+1)})^{-1}W^T_{(l+1)}\Sigma_{P(\theta_0)}\right. \\ &\left.-\Sigma_{P(\theta_0)}W_{(l)}(W^T_{(l)}\Sigma_{P(\theta_0)}W_{(l)})^{-1}W^T_{(l)}\Sigma_{P(\theta_0)}\right\} \\ &\cdot diag(P(\theta_0)^{-1})(\widehat{P} - P(\theta_0)) \\ &+o(\|\widehat{\theta}^{(l+1)}_{\phi_2} - \theta_0\|) - o(\|\widehat{\theta}^{(l)}_{\phi_2} - \theta_0\|). \end{aligned}$$

If we denote

$$A_{(i)} \equiv diag(P(\theta_0)^{-1/2})\Sigma_{P(\theta_0)}W_{(i)}(W^T_{(i)}\Sigma_{P(\theta_0)}W_{(i)})^{-1}W^T_{(i)}\Sigma_{P(\theta_0)}diag(P(\theta_0)^{-1/2}), \quad i = l, l+1,$$

which is a symmetric matrix, then

$$\begin{aligned} &diag(P(\theta_0)^{-1/2})(P(\widehat{\theta}^{(l+1)}_{\phi_2}) - P(\widehat{\theta}^{(l)}_{\phi_2})) \\ &= (A_{(l+1)}-A_{(l)})diag(P(\theta_0)^{-1/2})(\widehat{P}-P(\theta_0)) + o(\|\widehat{\theta}^{(l+1)}_{\phi_2}-\theta_0\|)-o(\|\widehat{\theta}^{(l)}_{\phi_2} - \theta_0\|). \end{aligned}$$

Thus, $n^{1/2}diag(P(\theta_0)^{-1/2})(P(\widehat{\theta}^{(l+1)}_{\phi_2}) - P(\widehat{\theta}^{(l)}_{\phi_2})) \xrightarrow[n\to\infty]{L} N(0, \Sigma^*)$, where

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

and $P(\theta_0)^{1/2} = (p_1(\theta_0)^{1/2}, \ldots, p_k(\theta_0)^{1/2})^T$. Then, because $(P(\theta_0)^{1/2})^T diag(P(\theta_0)^{-1/2})\Sigma_{P(\theta_0)}W_{(i)} = 0$, $i = l, l+1$, we see that the expression above is $\Sigma^* = (A_{(l+1)}-A_{(l)})(A_{(l+1)}-A_{(l)}) = A_{(l)} - A_{(l+1)}$. Now the matrix $(A_{(l)} - A_{(l+1)})$ is symmetric and idempotent with $trace\ (A_{(l)} - A_{(l+1)}) = d_l - d_{l+1}$, and hence Z^TZ is asymptotically a chi-squared distribution with $d_l - d_{l+1}$ degrees of freedom.

Finally then, $T^{(l)}_{\phi_1,\phi_2} \equiv \frac{2n}{\phi_1''(1)}D_{\phi_1}(P(\widehat{\theta}^{(l+1)}_{\phi_2}), P(\widehat{\theta}^{(l)}_{\phi_2})) \xrightarrow[n\to\infty]{L} \chi^2_{d_l-d_{l+1}}$.

Remark 2. There are important measures of divergence that are not possible to write as ϕ-divergences, for instance, the divergence measures given by Battacharya, Renyi, and Sharma and Mittal. However, such measures can be written in the following form: $D_{\phi,h}(P,Q) = h(D_\phi(P,Q))$, where h is a differentiable increasing function mapping from $[0,\infty)$ onto $[0,\infty)$, with $h(0) = 0$ and $h'(0) > 0$, and $\phi \in \Phi$. In the following table, we present these divergence measures.

Divergence	$h(x)$	$\phi(x)$
Renyi	$\frac{1}{r(r-1)}\log(r(r-1)x+1); r\neq 0,1$	$\frac{x^r-r(x-1)-1}{r(r-1)}; r\neq 0,1$
Sharma-Mittal	$\frac{1}{s-1}\{(1+r(r-1)x)^{\frac{s-1}{r-1}}-1\}; s,r\neq 1$	$\frac{x^r-r(x-1)-1}{r(r-1)}; r\neq 0,1$
Battacharya	$-\log(-x+1)$	$-x^{1/2}+\frac{1}{2}(x+1)$

In the case of Renyi's divergence, we have

$$D^{(r)}(P,Q)=\frac{1}{r(r-1)}\log(\sum_{j=1}^{k}p_j^r q_j^{1-r}),\ r\neq 0,1,$$

and limiting cases for $r=0$ and $r=1$. That is, $D^{(1)}(P,Q)\equiv\lim_{r\to 1}D^{(r)}(P,Q)=\sum_{j=1}^{k}p_j\log\frac{p_j}{q_j}$, which is the Kullback-Leibler divergence. Similarly, $D^{(0)}(P,Q)=\sum_{j=1}^{k}q_j\log\frac{p_j}{q_j}=D^{(1)}(Q,P)$.

Theorem 4. *Under the assumptions given in Theorem* 3*, the asymptotic null distribution of the test statistic*

$$T^{(l)}_{\phi_1,\phi_2,h_1,h_2}\equiv\frac{2n}{\phi_1''(1)h_1'(0)}h_1(D_{\phi_1}(P(\widehat{\theta}^{(l+1)}_{\phi_2,h_2}),P(\widehat{\theta}^{(l)}_{\phi_2,h_2})))$$

is a chi-squared distribution with d_l-d_{l+1} degrees of freedom, $l=1,\ldots,m-1$, where $\widehat{\theta}^{(l)}_{\phi_2,h_2}$ and $\widehat{\theta}^{(l+1)}_{\phi_2,h_2}$ are the minimum (ϕ_2,h_2)-divergence estimators under the models H_l and H_{l+1}, respectively, defined by $\widehat{\theta}^{(l)}_{\phi_2,h_2}\equiv\arg\min_{\theta\in\Theta_l}h_2(D_{\phi_2}(\widehat{P},P(\theta)))$, $l=1,\ldots,m$.

Proof. Using a similar approach to that given in the proof of Theorem 3, it can be established that $D_{\phi_1}(P(\widehat{\theta}^{(l+1)}_{\phi_2,h_2}),P(\widehat{\theta}^{(l)}_{\phi_2,h_2}))=\frac{\phi_1''(1)}{2n}\widetilde{Z}^T\widetilde{Z}+o_p(n^{-1})$, where $\widetilde{Z}^T\widetilde{Z}$ is asymptotically a chi-squared distribution with d_l-d_{l+1} degrees of freedom.

Further, because $h_1(x)=h_1(0)+h_1'(0)x+o(x)$, we have $T^{(l)}_{\phi_1,\phi_2,h_1,h_2}=\widetilde{Z}^T\widetilde{Z}+o_p(1)\xrightarrow[n\to\infty]{L}\chi^2_{d_l-d_{l+1}}$.

In the particular case of Renyi's divergence, $h_1(x)=\frac{1}{r(r-1)}\log(r(r-1)x+1)$ and $\phi_1(x)=\frac{x^r-r(x-1)-1}{r(r-1)}$, with $r\neq 0,1$. Then, from Theorem 4, we have $\frac{2n}{r(r-1)}\log\sum_{j=1}^{k}p_j(\widehat{\theta}^{(l+1)}_{\phi_2,h_2})^r p_j(\widehat{\theta}^{(l)}_{\phi_2,h_2})^{1-r}\xrightarrow[n\to\infty]{L}\chi^2_{d_l-d_{l+1}}$.

Theorem 5. *Under the assumptions of Theorems* 3 *and* 4*, the asymptotic null distribution of each of the test statistics,*

$$\widetilde{T}^{(l)}_{\phi_1,\phi_2}=\frac{2n}{\phi_1''(1)}D_{\phi_1}(P(\widehat{\theta}^{(l)}_{\phi_2}),P(\widehat{\theta}^{(l+1)}_{\phi_2})) \tag{11}$$

and

$$\widetilde{T}^{(l)}_{\phi_1,\phi_2,h_1,h_2} = \frac{2n}{\phi_1''(1)h_1'(0)} h_1(D_{\phi_1}(P(\widehat{\theta}^{(l)}_{\phi_2,h_2}), P(\widehat{\theta}^{(l+1)}_{\phi_2,h_2}))), \tag{12}$$

is a chi-squared distribution with $d_l - d_{l+1}$ degrees of freedom.

Proof. We consider the function $\varphi(x) = x\phi_1(x^{-1})$. It is clear that $\varphi(x) \in \Phi, T^{(l)}_{\varphi,\phi_2} = \widetilde{T}^{(l)}_{\phi_1,\phi_2}$, and $T^{(l)}_{\varphi,\phi_2,h_1,h_2} = \widetilde{T}^{(l)}_{\phi_1,\phi_2,h_1,h_2}$. Then the result follows directly from Theorems 3 and 4.

A special case occurs if we put $h_1(x) = h_2(x) = x, \phi_1(x) = -\log x + x - 1$, $\phi_2(x) = x\log x - (x-1)$ in the statistic $T^{(l)}_{\phi_1,\phi_2}$ given in (10), or $h_1(x) = h_2(x) = x, \phi_1(x) = \phi_2(x) = x\log x - (x-1)$ in the statistic $\widetilde{T}^{(l)}_{\phi_1,\phi_2}$ given in (11). In this case, we obtain the classical likelihood ratio test statistic (Agresti (1996, p.197), Christensen (1997, p.322)) and the well known result,

$$2n\sum_{j=1}^{k} p_j(\widehat{\theta}^{(l)}) \log \frac{p_j(\widehat{\theta}^{(l)})}{p_j(\widehat{\theta}^{(l+1)})} \xrightarrow[n\to\infty]{L} \chi^2_{d_l - d_{l+1}},$$

where $\widehat{\theta}^{(i)}$ is the maximum likelihood estimator of θ under the model H_i ($\theta \in \Theta_i$); $i = l, l+1$.

Another important case occurs if we put $h_1(x) = h_2(x) = x, \phi_1(x) = \frac{1}{x}(1-x)^2, \phi_2(x) = x\log x - (x-1)$ in the statistic $T^{(l)}_{\phi_1,\phi_2}$ given in (10), or $h_1(x) = h_2(x) = x, \phi_1(x) = (1-x)^2, \phi_2(x) = x\log x - (x-1)$ in the statistic $\widetilde{T}^{(l)}_{\phi_1,\phi_2}$ given in (11). Then we obtain a Pearson-type test statistic given in Agresti (1996, p.197), and the result,

$$n\sum_{j=1}^{k} \frac{(p_j(\widehat{\theta}^{(l)}) - p_j(\widehat{\theta}^{(l+1)}))^2}{p_j(\widehat{\theta}^{(l+1)})} \xrightarrow[n\to\infty]{L} \chi^2_{d_l - d_{l+1}}.$$

Theorem 6. *Suppose that data $(X_1, \ldots, X_k)$ are multinomially distributed according to* (2) *and* (7). *Consider the nested sequence of hypotheses given by* (8) *and* (9). *Choose functions $\phi_1, \phi_2 \in \Phi$. Then, for testing hypotheses, $H_0 : H_{l+1}$ against $H_1 : H_l$; $l = 1, \ldots, m-1$, the test statistics,*

$$S^{(l)}_{\phi} = \frac{2n}{\phi''(1)}\{D_{\phi}(\widehat{P}, P(\widehat{\theta}^{(l+1)}_{\phi})) - D_{\phi}(\widehat{P}, P(\widehat{\theta}^{(l)}_{\phi}))\}, \tag{13}$$

and

$$S^{(l)}_{\phi,h} = \frac{2n}{\phi''(1)h'(0)}\{h(D_{\phi}(\widehat{P}, P(\widehat{\theta}^{(l+1)}_{\phi,h}))) - h(D_{\phi}(\widehat{P}, P(\widehat{\theta}^{(l)}_{\phi,h})))\}, \tag{14}$$

are nonnegative and their asymptotic null distribution is chi-squared with $d_l - d_{l+1}$ degrees of freedom, $l = 1, \ldots, m-1$, where h is a differentiable increasing function mapping from $[0, \infty)$ onto $[0, \infty)$, with $h(0) = 0$ and $h'(0) > 0$.

Proof. It is clear that $S_\phi^{(l)} \geq 0$, $l = 1, \ldots, m-1$, because $D_\phi(\widehat{P}, P(\widehat{\theta}_\phi^{(l+1)})) = \inf_{\theta \in \Theta_{l+1}} D_\phi(\widehat{P}, P(\theta)) \geq \inf_{\theta \in \Theta_l} D_\phi(\widehat{P}, P(\theta)) = D_\phi(\widehat{P}, P(\widehat{\theta}_\phi^{(l)}))$. Furthermore, from the results of Menéndez, Morales and Pardo (1997), under H_{l+1}, we have that $S_\phi^{(l)} \xrightarrow[n\to\infty]{L} \chi^2_{d_l - d_{l+1}}$. In a similar way, the results for $S_{\phi,h}^{(l)}$ can be established.

Remark 3. The asymptotic result of Theorem 6 can be generalized further to include a ϕ_1 for divergence D_{ϕ_1}, and a ϕ_2 for estimation $\widehat{\theta}_{\phi_2}^{(i)}$. That is, the statistic

$$S_{\phi_1,\phi_2}^{(l)} \equiv \frac{2n}{\phi_1''(1)} \{D_{\phi_1}(\widehat{P}, P(\widehat{\theta}_{\phi_2}^{(l+1)})) - D_{\phi_1}(\widehat{P}, P(\widehat{\theta}_{\phi_2}^{(l)}))\} \xrightarrow[n\to\infty]{L} \chi^2_{d_l - d_{l+1}}$$

under H_{l+1}.

The special case of $\phi_1(x) = (1-x)^2$, $\phi_2(x) = x \log x - (x-1)$ yields a statistic based on the difference of Pearson X^2 statistics with maximum likelihood estimation used to obtain the expected frequencies (e.g., Agresti (1996, p.197)), namely $n\{D_{(1-x)^2}(\widehat{P}, P(\widehat{\theta}^{(l+1)})) - D_{(1-x)^2}(\widehat{P}, P(\widehat{\theta}^{(l)}))\}$.

However, the nonnegativity of $S_{\phi_1,\phi_2}^{(l)}$ does not hold when $\phi_1 \neq \phi_2$. Thus, for the case above, considered by Agresti, the difference of the Pearson X^2 statistics is not necessarily nonnegative. Since it is common to use maximum likelihood estimation (that is, $\phi_2(x) = x \log x - (x-1)$), the statistic $S_{\phi_1,\phi_2}^{(l)}$, where $\phi_1 \neq \phi_2$, is not all that interesting to us. Instead, we concentrate on the statistics $T_{\phi_1,\phi_2,h_1,h_2}^{(l)}$ and $\widetilde{T}_{\phi_1,\phi_2,h_1,h_2}^{(l)}$ given in Theorems 4 and 5, respectively, which are nonnegative for all choices of $\phi_1, \phi_2 \in \Phi$.

4. Application of Hierarchical Testing

In this section, we present an example of loglinear modeling of contingency tables to demonstrate how Theorem 3 and its generalizations can be applied in practice. We choose interesting cases (see below) for ϕ_1, h_1, ϕ_2, h_2 in the statistic, $T_{\phi_1,\phi_2,h_1,h_2}^{(l)} \equiv \frac{2n}{\phi_1''(1)h_1'(0)} h_1(D_{\phi_1}(P(\widehat{\theta}_{\phi_2,h_2}^{(l+1)}), P(\widehat{\theta}_{\phi_2,h_2}^{(l)})))$, to demonstrate its versatility.

During the period 1948 to August 1952, community members of Framingham, Masachusetts had their cholesterol and systolic blood pressure measured. The data analyzed here are for 1329 males aged 40-59 who showed no heart disease at the time of initial measurement. During a six-year follow-up, 92 of the original 1329 males developed clinically manifest coronary heart disease. The data are given in Table I. This example has been analyzed by Cornfield (1962) and Medak and Cressie (1991).

We see that there are $k = 2 \times 4 \times 4 = 32$ categories. Although it is a slight abuse of the notation used above, we use triple subscripts "ijk" to denote the

categories: subscript "i" is for presence or absence of coronary heart disease, subscript "j" is for level of systolic blood pressure, and subscript "k" is for level of serum cholesterol.

Table I.

Coronary Heart Disease: Present ($i = 1$)				
	Systolic Blood Pressure (mm Hg)			
Serum Cholesterol (mg/100cc)	($j = 1$) < 127	($j = 2$) 127-146	($j = 3$) 147-166	($j = 4$) ≥ 167
< 200 ($k = 1$)	2	3	3	4
200-219 ($k = 2$)	3	2	0	3
220-259 ($k = 3$)	8	11	6	6
≥ 260 ($k = 4$)	7	12	11	11
Coronary Heart Disease: Absent ($i = 2$)				
	Systolic Blood Pressure (mm Hg)			
Serum Cholesterol (mg/100cc)	($j = 1$) < 127	($j = 2$) 127-146	($j = 3$) 147-166	($j = 4$) ≥ 167
< 200 ($k = 1$)	117	121	47	22
200-219 ($k = 2$)	85	98	43	20
220-259 ($k = 3$)	119	209	68	43
≥ 260 ($k = 4$)	67	99	46	33
				Grand Total 1329

Following Medak and Cressie (1991), consider the following hierarchy of loglinear models:

$$H_1 : \log(p_{ijk}(\theta)) = u + \theta_{1(i)} + \theta_{2(j)} + \theta_{3(k)} + \theta_{12(ij)} + \theta_{13(ik)} + \theta_{23(jk)}$$
$$H_2 : \log(p_{ijk}(\theta)) = u + \theta_{1(i)} + \theta_{2(j)} + \theta_{3(k)} + \theta_{12(ij)} + \theta_{13(ik)}$$
$$H_3 : \log(p_{ijk}(\theta)) = u + \theta_{1(i)} + \theta_{2(j)} + \theta_{3(k)} + \theta_{12(ij)}$$
$$H_4 : \log(p_{ijk}(\theta)) = u + \theta_{1(i)} + \theta_{2(j)} + \theta_{3(k)}$$
$$H_5 : \log(p_{ijk}(\theta)) = u + \theta_{1(i)} + \theta_{3(k)}$$
$$H_6 : \log(p_{ijk}(\theta)) = u + \theta_{3(k)}$$
$$H_7 : \log(p_{ijk}(\theta)) = u,$$

where $i = 1, 2$; $j = 1, 2, 3, 4$; and $k = 1, 2, 3, 4$. Here, $\exp(-u)$ is the normalizing constant and the subscripted θ-terms add to zero over each of their indices. Write θ as the collection of these subscripted terms; then the models $H_1, \ldots, H_7$ can be written as $H_l : \theta \in \Theta_l$, $l = 1, \ldots, 7$, respectively.

We consider the minimum power-divergence estimator given by $\widehat{\theta}_{\phi_2,h_2}$, where $\phi_2(x) = \phi_{(\lambda)}(x) \equiv (\lambda(\lambda+1))^{-1}(x^{\lambda+1} - x), h_2(x) = x$. Cressie and Read (1984) introduced this family and showed that not only were the familiar likelihood-based ($\lambda = 0$) and Pearson-based ($\lambda = 1$) estimators and test statistics included, but that a new member, $\lambda = 2/3$, demonstrated superior performance in some

settings. For estimating parameters in a mixture of normal distributions, Pardo (1999) showed that the best member of the family was $\lambda = 2/3$.

For the goodness-of-fit statistic, we choose Renyi's divergence given by, $\phi_1(x) = \phi_{(r)}(x) \equiv \frac{x^r - r(x-1) - 1}{r(r-1)}, h_1(x) = \frac{1}{r(r-1)} \log(r(r-1)x + 1)$, $r \neq 0, 1$, and limiting cases for $r = 0, 1$. Renyi's divergence has been used succesfully for testing composite hypotheses in a normal model, in Morales, Pardo and Vajda (1997).

Therefore, the statistic for testing the nested sequence of hypotheses is given by,

$$T^{(l)}_{\phi_{(r)},\phi_{(\lambda)},h_1,h_2} = \frac{2n}{r(r-1)} \log \sum_{j=1}^{k} p_j(\widehat{\theta}^{(l+1)}_{(\lambda)})^r p_j(\widehat{\theta}^{(l)}_{(\lambda)})^{1-r}, \quad r \neq 0, 1,$$

and limiting cases for $r = 0$ or 1. For example, for $r = 1$,

$$T^{(l)}_{\phi_{(1)},\phi_{(\lambda)},h_1,h_2} = 2n \sum_{j=1}^{k} p_j(\widehat{\theta}^{(l+1)}_{(\lambda)}) \log \frac{p_j(\widehat{\theta}^{(l+1)}_{(\lambda)})}{p_j(\widehat{\theta}^{(l)}_{(\lambda)})},$$

where $l = 1, \ldots, m-1$. In the analysis that follows, we consider $\lambda = 0, 2/3, 1$ in ϕ_2 and $r = 1, 2$ in ϕ_1.

$\underline{\lambda = 0}$

In the following table we give the minimum power-divergence estimator for the case $\lambda = 0$ (i.e., the MLE) and the model H_1. This estimator has been obtained using the statistics package Statgraphics Plus (1993).

MODEL H_1

Factor		$\widehat{\theta}^{(1)}_{\phi_{(0)}}$	**Factor**		$\widehat{\theta}^{(1)}_{\phi_{(0)}}$	**Factor**		$\widehat{\theta}^{(1)}_{\phi_{(0)}}$	**Factor**		$\widehat{\theta}^{(1)}_{\phi_{(0)}}$
			θ_{ij}				21	0.21226		32	0.01286
$\theta_{1(i)}$				11	-0.21140		22	0.31617		33	-0.05763
	1	-1.31713		12	-0.23215		23	-0.06875		34	0.07857
	2	1.31713		13	0.05477		24	-0.45968		41	-0.16907
$\theta_{2(j)}$				14	0.38878	θ_{jk}				42	-0.08602
	1	0.17515		21	0.21140		11	0.22180		43	0.01803
	2	0.43751		22	0.23215		12	0.11373		44	0.23706
	3	-0.19092		23	-0.05477		13	-0.11440			
	4	-0.42174		24	-0.38878		14	-0.22113			
$\theta_{3(k)}$			θ_{ik}				21	-0.01894			
	1	-0.23167		11	-0.21226		22	-0.04057			
	2	-0.52367		12	-0.31617		23	0.15400			
	3	0.42836		13	0.06875		24	-0.09449			
	4	0.32698		14	0.45968		31	-0.03380			

In the following table, we present the expected frequencies obtained using the estimator $\widehat{\theta}^{(1)}_{\phi_{(0)}}$ given in the table above.

Coronary Heart Disease: Present ($i = 1$)				
	Systolic Blood Pressure (mm Hg)			
Serum Cholesterol	($j = 1$)	($j = 2$)	($j = 3$)	($j = 4$)
(mg/100cc)	< 127	127-146	147-166	≥ 167
< 200 ($k = 1$)	3.6	3.6	2.5	2.4
200-219 ($k = 2$)	2.1	2.3	1.8	1.8
220-259 ($k = 3$)	6.5	10.8	6.2	7.4
≥ 260 ($k = 4$)	7.8	11.3	9.5	12.4
Coronary Heart Disease: Absent ($i = 2$)				
	Systolic Blood Pressure (mm Hg)			
Serum Cholesterol	($j = 1$)	($j = 2$)	($j = 3$)	($j = 4$)
(mg/100cc)	< 127	127-146	147-166	≥ 167
< 200 ($k = 1$)	115.5	120.4	47.5	23.6
200-219 ($k = 2$)	85.9	97.7	41.2	21.2
220-259 ($k = 3$)	120.5	209.2	67.8	41.6
≥ 260 ($k = 4$)	66.2	99.7	47.5	31.6
				Grand Total 1329

In a similar way, we have obtained the corresponding values for the models $H_2, \ldots, H_7$, although they are not given here.

Now, for $\lambda = 0$, we present the values of the statistics $T^{(l)}_{\phi_{(1)},\phi_{(0)},h_1,h_2}$ and $T^{(l)}_{\phi_{(2)},\phi_{(0)},h_1,h_2}$, as well as the critical points to be used for the selection of an appropriate model.

$\mathbf{H}_{l+1}$ v. $\mathbf{H}_l$	$d_l - d_{l+1}$	$T^{(l)}_{\phi_{(1)},\phi_{(0)},h_1,h_2}$	$T^{(l)}_{\phi_{(2)},\phi_{(0)},h_1,h_2}$	$\chi^2_{d_l-d_{l+1},0.01}$
2 versus 1	9	18.8821	18.5993	21.6676
3 versus 2	3	32.3665	34.0836	11.3462
4 versus 3	3	21.2012	19.6293	11.3462
5 versus 4	3	348.3705	357.1923	11.3462
6 versus 5	1	1792.9995	1796.7157	6.6385
7 versus 6	3	77.3392	70.9508	11.3462

We can see, using the model-selection criterion given above, that we must choose the model H_2. That is, given the presence/absence of coronary heart disease, cholesterol and blood pressure are conditionally independent.

To obtain minimum power-divergence estimators of θ's for $\lambda \neq 0$, we use the IMSL subroutine ZXMIN based on a quasi-Newton method of minimization. Good starting values are crucial for the iterative procedures. Given a model for H_l, we computed the MLE's of θ's. Then, for each lambda, we generated 2000 starting points from a d_l-dimensional hypercube centered on the MLE's. Averaging the resulting minimum power-divergence estimators avoids difficulties with bad starting values.

$\underline{\lambda = 1}$

For comparison and illustration, we give the minimum power-divergence estimator for the case $\lambda = 1$ and the models H_1 and H_4.

MODEL H_1

Factor		$\widehat{\theta}^{(1)}_{\phi_{(1)}}$	**Factor**		$\widehat{\theta}^{(1)}_{\phi_{(1)}}$	**Factor**		$\widehat{\theta}^{(1)}_{\phi_{(1)}}$	**Factor**		$\widehat{\theta}^{(1)}_{\phi_{(1)}}$
			θ_{ij}				21	0.21282		32	0.01383
$\theta_{1(i)}$				11	-0.21061		22	0.31545		33	-0.05738
	1	-1.31671		12	-0.23150		23	-0.06917		34	0.07836
	2	1.31671		13	0.05526		24	-0.45919		41	-0.16760
$\theta_{2(j)}$				14	0.38685	θ_{jk}				42	-0.08633
	1	0.17528		21	0.21061		11	0.21827		43	0.01981
	2	0.43689		22	0.23150		12	0.11334		44	0.23412
	3	-0.19030		23	-0.05526		13	-0.11537			
	4	-0.42187		24	-0.38684		14	-0.21979			
$\theta_{3(k)}$			θ_{ik}				21	−0.01975			
	1	-0.23150		11	-0.21291		22	−0.04084			
	2	-0.52371		12	-0.31545		23	0.15329			
	3	0.42806		13	0.06917		24	−0.09269			
	4	0.32714		14	0.45919		31	-0.03446			

MODEL H_4

Factor		$\widehat{\theta}^{(4)}_{\phi_{(1)}}$
$\theta_{1(i)}$		
	1	-1.16226
	2	1.16226
$\theta_{2(j)}$		
	1	0.32255
	2	0.62154
	3	-0.27516
	4	-0.66893
$\theta_{3(k)}$		
	1	-0.01078
	2	-0.24357
	3	0.35180
	4	0.09744

Now, for $\lambda = 1$, we present the values of the statistics $T^{(l)}_{\phi_{(1)},\phi_{(1)},h_1,h_2}$, and $T^{(l)}_{\phi_{(2)},\phi_{(1)},h_1,h_2}$, as well as the critical points to be used for the selection of an appropriate model.

$\mathbf{H}_{l+1}$ v. $\mathbf{H}_l$	$d_l - d_{l+1}$	$T^{(l)}_{\phi_{(1)},\phi_{(1)},h_1,h_2}$	$T^{(l)}_{\phi_{(2)},\phi_{(1)},h_1,h_2}$	$\chi^2_{d_{l+1}-d_l,0.01}$
2 versus 1	9	19.1562	18.7796	21.6676
3 versus 2	3	38.5508	45.7661	11.3462
4 versus 3	3	29.9479	30.9946	11.3462
5 versus 4	3	327.1631	331.6542	11.3462
6 versus 5	1	1732.8761	1732.8202	6.6385
7 versus 6	3	91.4849	86.7172	11.3462

Notice that similar results are obtained here as for $\lambda = 0$, leading to the same choice of model, namely H_2.

$\underline{\lambda = 2/3}$

For comparison and illustration, we give the minimum power-divergence estimators for the case $\lambda = 2/3$ and the model H_1 and H_5.

MODEL H_1

Factor		$\widehat{\theta}^{(1)}_{\phi_{(2/3)}}$	**Factor**		$\widehat{\theta}^{(1)}_{\phi_{(2/3)}}$	**Factor**		$\widehat{\theta}^{(1)}_{\phi_{(2/3)}}$	**Factor**		$\widehat{\theta}^{(1)}_{\phi_{(2/3)}}$
			θ_{ij}				21	0.21291		32	0.01383
$\theta_{1(i)}$				11	-2.10608		22	0.31542		33	-0.05773
	1	-1.31671		12	-0.23150		23	-0.06917		34	0.07837
	2	1.31671		13	0.05526		24	-0.45919		41	-0.16760
$\theta_{2(j)}$				14	0.38685	θ_{jk}				42	-0.08633
	1	0.17528		21	0.21061		11	0.22182		43	0.01981
	2	0.43689		22	0.23150		12	0.11333		44	0.23412
	3	-0.19030		23	-0.05263		13	-0.11537			
	4	-0.42187		24	-0.38685		14	-0.21979			
$\theta_{3(k)}$			θ_{ik}				21	-0.01975			
	1	-0.23150		11	-0.21291		22	-0.04084			
	2	-0.52371		12	-0.31545		23	0.15329			
	3	0.42806		13	0.06917		24	-0.09269			
	4	0.32715		14	0.45919		31	-0.03446			

MODEL H_5

Factor		$\widehat{\theta}^{(5)}_{\phi_{(2/3)}}$
$\theta_{1(i)}$		
	1	-1.27839
	2	1.27839
$\theta_{3(k)}$		
	1	0.00129
	2	-0.23751
	3	0.38131
	4	-0.45101

Now, for $\lambda = 2/3$, we present the values of the statistics $T^{(l)}_{\phi_{(1)},\phi_{(2/3)},h_1,h_2}$ and $T^{(l)}_{\phi_{(2)},\phi_{(2/3)},h_1,h_2}$, as well as the critical points to be used for the selection of an appropriate model.

$\mathbf{H}_{l+1}$ v. $\mathbf{H}_l$	$d_l - d_{l+1}$	$T^{(l)}_{\phi_{(1)},\phi_{(2/3)},h_1,h_2}$	$T^{(l)}_{\phi_{(2)},\phi_{(2/3)},h_1,h_2}$	$\chi^2_{d_{l+1}-d_l,0.01}$
2 versus 1	9	19.1562	18.7796	21.6676
3 versus 2	3	34.7833	39.6591	11.3462
4 versus 3	3	26.7233	26.4015	11.3462
5 versus 4	3	332.6503	338.4426	11.3462
6 versus 5	1	1756.0330	1755.9999	6.6385
7 versus 6	3	85.6067	81.0606	11.3462

Once again, the model H_2 is chosen from among the sequence of nested models hypotheses, $H_1, \ldots, H_7$.

Remark 4. We can also solve the nested-hypothesis problem with the statistic $S^{(l)}_{\phi}$ given in (13) of Theorem 6. If we consider $\phi(x) = \phi_{(\lambda)}(x) \equiv (\lambda(\lambda+1))^{-1}(x^{\lambda+1} - 1)$, then for testing H_{l+1} versus H_l, we obtain the statistic,

$$S^{(l)}_{\phi_{(\lambda)}} = \frac{2n}{\lambda(\lambda+1)}\Big\{\sum_{j=1}^{k} \widehat{p}_j \Big(\Big(\frac{\widehat{p}_j}{p_j(\widehat{\theta}^{(l+1)}_{\phi_{(\lambda)}})}\Big)^{\lambda} - \Big(\frac{\widehat{p}_j}{p_j(\widehat{\theta}^{(l)}_{\phi_{(\lambda)}})}\Big)^{\lambda}\Big)\Big\}.$$

When $\lambda = 2/3$, we obtain the following values of the statistic $S^{(l)}_{\phi_{(2/3)}}$, as well as the critical points to be used for the selection of an appropriate model. Once again, model H_2 is chosen.

$\mathbf{H}_l$ v. $\mathbf{H}_{l+1}$	$d_l - d_{l+1}$	$S^{(l)}_{\phi_{(2/3)}}$	$\chi^2_{d_{l+1}-d_l,0.01}$
2 versus 1	9	19.860	21.6676
3 versus 2	3	31.018	11.3462
4 versus 3	3	30.129	11.3462
5 versus 4	3	304.275	11.3462
6 versus 5	1	1170.346	6.6385
7 versus 6	3	187.891	11.3462

It should be noted that the choice of different test statistics yields different values but no difference in model choice. This is to be hoped for, however it is not guaranteed in every problem and for every choice of ϕ and h. Asymptotically, the statistics have the same distribution, but in finite samples their performances will differ. In future work we will compare important family members, *inter alia* those considered here, through a simulation study.

5. Conclusions

Inference for categorical data is an important problem whose history goes back at least as far as Karl Pearson's work in the late nineteenth century on goodness-of-fit statistics. In the 1920s, Ronald Fisher advocated likelihood-based methods and, for much of the rest of the twentieth century, the pros and cons of both were debated. In the recent past, it has been realized that inferences can be based on the more general notion of divergence between discrete probability measures; see Section 1 for a brief literature review. The purpose of this paper has been to build on this reseach. Assuming multinomial sampling and a loglinear model, we present the most general asymptotic results possible for minimum-divergence estimation, divergence-based goodness-of-fit testing, and choosing a model from a nested sequence of hypotheses. Any estimator or statistic that is divergence-based is covered by our results. We apply them in interesting special cases (e.g., the minimum power-divergence estimator coupled with Renyi's goodness-of-fit statistic) to a data set concerned with heart disease, cholesterol, and blood pressure.

Acknowledgement

The research in this paper was supported in part by NATO grant No. CRG970442.

References

Agresti, A. (1996). *An Introduction to Categorical Data Analysis.* John Wiley, New York.

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

Collett, D. (1994). *Modelling Survival Data in Medical Research.* Chapman and Hall, London.

Cornfield, J. (1962). Joint dependence of risk of coronary heart disease on serum cholesterol and systolic blood pressure: a discriminant function analysis. *Federation Proceedings* **21**, 58-61.

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

Csiszar, I. (1967). Information type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica* **2**, 105-113.

Medak, F. M. and Cressie, N. (1991). Hierarchical testing of parametric models using the power-divergence family of test statistics. Statistical Laboratory Preprint, No. 91-14, Iowa State University, Ames, IA.

Menéndez, M. L., Morales, D. and Pardo, L. (1997). ϕ-divergences and nested models. *Appl. Math. Lett.* **10**, 129-132.

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

Morales, D., Pardo, L. and Vajda, I. (1997). Some new statistics for testing hypotheses in parametric models. *J. Multivariate Anal.* **62**, 137-168.

Pardo, M. C. (1999). Estimation of parameters for a mixture of normal distributions on the basis of the Cressie and Read divergence. *Comm. Statist. Simulation Comput.* **28**, 115-130.

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

Statgraphics Plus (1993). Statistical Graphics System by Statistical Graphics Corporation: Reference Manual, Version 7 for DOS. Manugistics Inc. Rockville, MD.

Department of Statistics, The Ohio State University, Columbus, OH 43210-1247, U.S.A.

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

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

E-mail: Leandro_Pardo@Mat.ucm.es

(Received October 1998; accepted September 1999)