Записки научных
семинаров ПОМИ
Том 363, 2009 г.

K. Mattheou, A. Karagrigoriou

ON NEW DEVELOPMENTS IN DIVERGENCE STATISTICS

Abstract. In this paper, we discuss measures of divergence and focus on a recently introduced measure of divergence, the so called BHHJ measure (Basu et. al, 1998). A general class of such measures is introduced and goodness of fit tests for multinomial populations are presented. Simulations are performed to check the appropriateness of the proposed test statistics.

1. Introduction

A measure of divergence is used as a way to evaluate the distance or divergence between any two populations or functions. Let f_1 and f_2 be two probability density functions which may depend or not on an unknown parameter of fixed finite dimension. The most well known measure of (directed) divergence is the Kullback–Leibler divergence which is given by

$$I_X^{KL}(f_1, f_2) = \int f_1 \log(f_1/f_2) d\mu = E_{f_1}\left[\log\left(\frac{f_1}{f_2}\right)\right],$$

for a measure μ which, for the continuous case, is the Lebesgue measure, and a random variable X with absolutely continuous distribution. This notation covers not only the continuous case but also a discrete setting where the measure μ is a counting measure. Indeed, for the discrete case, the divergence is meaningful for the probability mass functions f_1 and f_2 whose support is a subset of the support S_μ, finite or countable, of the counting measure μ that satisfies $\mu(x) = 1$ for $x \in S_\mu$, and 0 otherwise.

So, for the above divergence and for the subsequent ones consider that, if k is a measurable function, the expectation of $k(X)$ is given by:

$$E_f\left[k\left(X\right)\right] = \begin{cases} \displaystyle\int k\left(x\right) f\left(x\right) dx & \text{if } \mu \text{ is the Lebesgue measure,} \\ \displaystyle\sum_{x \in S_\mu} k\left(x\right) f\left(x\right) & \text{if } \mu \text{ is the counting measure.} \end{cases}$$

As generalizations of the Kullback–Leibler measure, the additive and nonadditive directed divergences of order a were introduced by Rényi (1961) and Csiszár (1963). An extension of these divergences was given by Liese and Vajda (1987), for all $a \neq 0, 1$:

$$I_X^{R_{lv},\alpha}(f_1, f_2) = \frac{1}{\alpha(\alpha-1)} \log \int f_1^{\alpha} f_2^{1-\alpha} d\mu$$
$$= \frac{1}{\alpha(\alpha-1)} \log E_{f_1} \left[\left(\frac{f_1(X)}{f_2(X)} \right)^{\alpha-1} \right], \quad \alpha \neq 0, 1.$$

Furthermore, the Matusita measure (see [Matusita, (1967)]) given by

$$I_X^M(f_1, f_2) = \int (\sqrt{f_1} - \sqrt{f_2})^2 d\mu$$

is the square of the well known Hellinger distance.

Another measure of divergence is the measure of Kagan (1963) which is known as Pearson's X^2 and is given by

$$I_X^{\mathrm{Ka}}(f_1, f_2) = \frac{1}{2} \int (1 - f_1/f_2)^2 f_2 d\mu.$$

Csiszár's measure of information (see [Csiszár (1963), Ali, and Silvey, (1966)] is a general divergence-type measure, known also as φ-divergence based on a convex function φ and defined by

$$I_X^{C,\varphi}(f_1, f_2) = \int \varphi(f_1/f_2) f_2 d\mu = E_{f_2} \left[\left(\frac{f_1}{f_2} \right) \right], \quad \varphi \in \Phi^*$$

where Φ^* is the class of all convex functions φ on $[0, \infty)$ such that $\varphi(1) = 0$, $0\varphi(0/0) = 0$, and $0\varphi(u/0) = \lim_{u\to\infty} \varphi(u)/u$, for $u > 0$.

Observe that Csiszár's measure reduces to Kullback–Leibler divergence if $\varphi(u) = u \log u$ or $\varphi(u) = u \log u - u + 1$. If $\varphi(u) = (1/2)(1-u)^2$ or $\varphi(u) = (1 - \sqrt{u})^2$ Csiszár's measure yields the Kagan and the Matusita divergence, respectively. More examples of φ-functions and the measures we obtain based on these functions are given in Pardo (2006).

A well known generalization of measures of divergence is the family of power divergences introduced independently by Cressie and Read (1984) and Liese and Vajda (1987) which is given by

$$I_X^{CR}(f_1, f_2) = \frac{1}{\lambda(\lambda+1)} \int f_1(z) \left[\left(\frac{f_1(z)}{f_2(z)} \right)^{\lambda} - 1 \right] dz, \quad \lambda \in R,$$

where for $\lambda = 0, -1$ is defined by continuity and reduces to Kullback–Leibler divergence. Note also that this divergence is a member of the Csiszár's family of measures.

1.1. The BHHJ Measure of Divergence

One of the most recently proposed measures of divergence is the BHHJ power divergence between f and g (Basu et *al.*, 1998) which is denoted by BHHJ, indexed by a positive parameter a, and defined as:

$$I_X^a(g, f) = \int \Big\{ f^{1+a}(z) - \Big(1 + \frac{1}{a}\Big) g(z) f^a(z) + \frac{1}{a} g^{1+a}(z) \Big\} dz, \quad a > 0. \qquad (1)$$

This family of measures was proposed by Basu et *al.* (1988) for the development of a minimum divergence estimating method for robust parameter estimation. The index a controls the trade-off between robustness and asymptotic efficiency of the parameter estimators which are the quantities that minimize (1). It should be also noted that the BHHJ family reduces to the Kullback–Leibler divergence for a tending to 0 and as it can be easily seen, to the square of the standard L_2 distance between f and g for $a = 1$. As a result, for $a = 0$ the family, as an estimating method, reduces to the traditional maximum likelihood estimation while for $a = 1$ becomes the mean squared error estimation. In the former case the resulting estimator is efficient but not robust while in the latter the method results in a robust but inefficient estimator. The authors observed that for values of a close to 0 the resulting estimators have strong robust features without a big loss in efficiency relative to the maximum likelihood estimating method. As a result one is interesting in small values of $a \geqslant 0$, say between zero and one, although values larger than one are also allowed. One should be aware though of the fact that the estimating method becomes less and less efficient as the index a increases.

It is interesting to note that the BHHJ measure can be considered as a special case of the Bregman divergence (Jones and Byrne, 1990; Csiszár, 1991) which has the general form

$$\int \Big[H\{g(z)\} - H\{f(z)\} - \{g(z) - f(z)\} H'\{f(z)\} \Big] dz,$$

where H is a convex function. Observe that a Taylor series expansion of the integrand of the Bregman divergence when f is close to g gives $\frac{1}{2}(f - g)^2 H''(f)$. If ones wants the Bregman divergence to reduce to the square of the L_2 distance for $a = 1$ (and consequently to the mean squared error estimating method) then $H''(f) \propto f^{a-1}$ for some $a \geqslant 0$ so that $H(f) \propto f^{a+1}$ in which case the Bregman divergence reduces to (1).

Some motivation for the form of the BHHJ divergence can be obtained by looking at the location model, where $\int f_\theta^{1+a}(z)dz$ is independent of any parameter. In this case, the proposed estimators maximize $\sum_{i=1}^n f_\theta^a(X_i)$, with the corresponding estimating equations having the form

$$\sum_{i=1}^{n} u_\theta\left(X_i\right) f_\theta^a\left(X_i\right) = 0, \tag{2}$$

where $u_\theta\left(z\right) = \partial \log f_\theta\left(z\right)/\partial\theta$ is the maximum likelihood score function. This can be viewed as a weighted version of the efficient maximum likelihood score equation. When $a > 0$, (2) provides a relative-to-the-model downweighting for outlying observations; observations that are wildly discrepant with respect to the model will get nearly zero weights. In the fully efficient case $a = 0$, all observations, including very severe outliers, get weights equal to one.

We generalize now the family (1) to a more general family of the following form that involves a general function $\Phi(\cdot)$.

Definition 1.1. *For a general function $\Phi \in \mathcal{G}$ and for $a > 0$, we define the divergence between two functions f and g by*

$$I_X^a\left(g,f\right) = E_g\Big(g^a(X)\Phi\Big(\frac{f(X)}{g(X)}\Big)\Big) = \int g^{1+a}\left(z\right)\Phi\Big(\frac{f(z)}{g(z)}\Big)d\mu, \tag{3}$$

where μ represents the Lebesgue measure and $\mathcal{G}$ is the class of all convex functions Φ on $[0,\infty)$ such that $\Phi(1) = 0$, $\Phi'(1) = 0$ and $\Phi''\left(1\right) \neq 0$. In the expression of $I_X^a\left(g,f\right)$, we assume the conventions

$$0\Phi\left(0/0\right) = 0 \;\; \textit{and} \;\; 0\Phi\left(u/0\right) = \lim_{u\to\infty}\Phi\left(u\right)/u \;\textit{for}\;\; u > 0.$$

The BHHJ measure of Basu et. al (1998) can be obtained from the above general BHHJ family if the function Φ takes the special form

$$\Phi\left(u\right) = u^{1+a} - \left(1 + \frac{1}{a}\right)u^a + \frac{1}{a}\cdot \tag{4}$$

Expression (3) covers not only the continuous case presented in (1) but also a discrete setting where the measure μ is a counting measure. Indeed, for the discrete case, the divergence in (3) is meaningful for probability mass functions f and g whose support is a subset of the support S_μ, finite or countable, of the counting measure μ that satisfies $\mu(x) = 1$ for $x \in S_\mu$ and 0 otherwise.

Consider now two multinomial distributions $P = (p_1, \ldots, p_m)$ and $Q = (q_1, \ldots, q_m)$ with sample space $\Omega = \{x : p(x) \cdot q(x) > 0\}$ where $p(x)$ and $q(x)$ are the probability mass functions of the two distributions. Then the discrete version of the Cressie and Read measure is given by

$$I_X^{CR}(P, Q) = \frac{1}{\lambda(\lambda+1)} \sum_{i=1}^{m} p_i \left[\left(\frac{p_i}{q_i} \right)^{\lambda} - 1 \right], \quad \lambda \in R, \quad \lambda \neq 0, -1. \tag{5}$$

The above measure was introduced by Cressie and Read (1984) for goodness of fit tests for multinomial distributions. Observe that the family includes important and well known test statistics like the Pearson's X^2 statistic (for $\lambda = 1$), the loglikelihood ratio statistic (for $\lambda \to 0$) and the Freeman–Tukey statistic (for $\lambda = -1/2$). Cressie and Read (1984) devoted their work to the analytic study of the asymptotic properties of the above measure and found that the $\lambda = 2/3$ case constitutes an excellent and compromising alternative between the traditional $\lambda \to 0$ (loglikelihood ratio test) and $\lambda = 1$ (Pearson's X^2 test) cases.

The discrete version of Csiszár's measure is given in a similar fashion, by

$$d_c = \sum_{i=1}^{m} q_i \varphi(p_i/q_i).$$

The discrete Csiszár's measure has been used by Zografos et *al.* (1990) for purposes analogous to the ones of the discrete Cressie and Read measure, namely, for goodness of fit tests for multinomial distributions.

In what follows we extend the class of measures of divergence (3) to a discrete setting analogous to the above discrete versions of Csiszár's or Cressie and Read's measures for multinomial distributions.

Definition 1.2. *For two discrete distributions $P = (p_1, \ldots, p_m)$ and $Q = (q_1, \ldots, q_m)$ with sample space $\Omega = \{x : p(x) \cdot q(x) > 0\}$, where $p(x)$ and $q(x)$ are the probability mass functions of the two distributions, the discrete version of the general BHHJ family of divergence measures with a general function Φ as in Definition* 1.1 *and $a > 0$ is given by*

$$d_a \equiv d_a(Q, P) = E_q\left(q^a(X) \Phi\left(\frac{p(X)}{q(X)} \right) \right) \equiv \sum_{i=1}^{m} q_i^{1+a} \Phi\left(\frac{p_i}{q_i} \right) \tag{6}$$

which for Φ *as in* (4) *becomes the discrete BHHJ measure given by*

$$d_a \equiv d_a\,(Q,P) = \sum_{i=1}^{m} p_i^{1+a} - \left(1+\frac{1}{a}\right)\sum_{i=1}^{m} q_i p_i^a + \frac{1}{a}\sum_{i=1}^{m} q_i^{1+a}. \tag{7}$$

For $a \to 0$ the measure reduces to the Kullback–Leibler divergence while for $\Phi(u) = \varphi(u)$ and for $a = 0$ we obtain the Csiszár's φ divergence.

The measures described above play a significant role in statistical inference and have several applications. The aim of this paper is to present some recent developments on measures of divergences. In particular, in Sec. 2 we propose some test statistics for goodness of fit tests for multivariate populations while in Sec. 3 simulation results are presented.

2. Goodness of Fit Tests

The statistical analysis and in particular the testing of models for discrete multivariate data has been given considerable attention during the last 30 years. The books of Cox (1970), Agresti (1984), and Cressie and Read (1988) are focusing on various aspects of model development. The usual assumption is that the adequacy of a model can be tested by one of the traditional goodness-of-fit tests, namely the Pearson's X^2 or the loglikelihood ratio test. Note that both of these tests are special cases of the Cressie and Read measure of divergence introduced in (5). Indeed in a discrete setting and for $\lambda = 1$ the Cressie and Read measure reduces to $\sum_{i=1}^{m} \frac{(p_i - q_i)^2}{q_i}$ which multiplied by $2n$ is the Pearson's X^2 test where p_i plays the role of the observed frequency and q_i the role of the expected one. Furthermore, the loglikelihood ratio test statistic (also known as Kullback-Leibler measure) $2n \sum_{i=1}^{m} p_i \log\left(\frac{p_i}{q_i}\right)$ can be deduced from the Cressie and Read measure for $\lambda \to 0$.

In this section, we focus on a discrete setting and provide some distributional properties of the estimator of the general BHHJ family of measures which is shown to be weakly consistent. These results are then used for establishing a goodness of fit test for multinomial distributions based on the general BHHJ family of divergence measures.

Definition 2.3. *Let f be a function with continuous derivatives of second order defined on the set* $S_k = \{(s_1, s_2) : 0 < s_i < \infty,\ i = 1, 2\}$. *Then the* f*-dissimilarity is defined to be*

$$d_a = d_f\,(Q,P) = \sum_{j=1}^{m} f\,(p_j, q_j),$$

where p_j, q_j, $j = 1, ..., m$ are the parameters from the multinomial distributions $M(N_p, P)$, $P = (p_1, p_2, ..., p_m)$ and $M(N_q, Q)$, $Q = (q_1, q_2, ..., q_m)$ and f is a continuous convex, homogeneous function.

For different functions f we have specific dissimilarity measures. For example for $f(p,q) = q\varphi(p/q)$ we have the Csiszár's measure and for $f(p,q) = q^{1+a}\Phi(p/q)$ we have the general BHHJ family of measures while for Φ as in (4) we have the discrete BHHJ measure. Observe that the estimator of d_a is

$$\widehat{d}_a = d_f\left(\widehat{Q}, \widehat{P}\right) = \sum_{j=1}^{m} f\left(\widehat{p}_j, \widehat{q}_j\right).$$

For the general BHHJ family of measures the estimator of the f-dissimilarity is given by

$$\widehat{d}_a = \sum_{j=1}^{m} \widehat{q}_j^{1+a}\Phi\left(\frac{\widehat{p}_j}{\widehat{q}_j}\right). \tag{8}$$

where $\widehat{p}_j = \frac{x_j}{N_p}$, $\widehat{q}_j = \frac{y_j}{N_q}, j = 1,...,m$, and $X = (x_1, ..., x_m), Y = (y_1, ..., y_m)$ are random observations from $M(N_p, P)$ and $M(N_q, Q)$.

Observe that in case one of the two distributions is known then the obvious notation applies, namely, $\widehat{d}_a = d_f\left(Q, \widehat{P}\right) = \sum_{j=1}^{m} f\left(\widehat{p}_j, q_j\right)$ if Q is known and $\widehat{d}_a = d_f\left(\widehat{Q}, P\right) = \sum_{j=1}^{m} f\left(p_j, \widehat{q}_j\right)$ if P is known.

Goodness of fit tests using measures of divergence such as Csiszár's have been extensively investigated [Zografos et *al.*, 1990; Morales et *al.*, 1997; Pardo, 1999, etc.].

If we have to examine whether the data $(n_1, n_2, ..., n_m)$ come from a multinomial distribution $M(N, P_0)$, where $P_0 = (p_{10}, p_{20}, ..., p_{m0})$ and $N = \sum_{i=1}^{m} n_i$, a well known test statistic is the chi-square goodness of fit test statistic. We define now for any function Φ such that $\Phi'(1) = 0$ and $\Phi''(1) \neq 0$, a new statistic for the above goodness of fit test:

$$X_a^2 \equiv \frac{2N\left(\widehat{d}_a - \Phi(1)\sum_{i=1}^{m} p_{i0}^{1+a}\right)}{\Phi''(1)} \tag{9}$$

which for $\Phi(u)$ as in (4) constitutes the test statistic associated with the BHHJ divergence. Observe that for the purpose of goodness of fit tests we use

$$\widehat{d}_a = \sum_{i=1}^{m} {q_i}^{1+a} \Phi\left(\frac{\widehat{p}_i}{q_i}\right) \tag{19}$$

with $q_i = p_{i0}$.

In what follows we establish the asymptotic distributions of the estimator $\widehat{d}_a$ (Corollary 2.1 and the test statistic (9) under appropriate null and alternative hypotheses (Theorem 2.3 and Corollary 2.2).

Theorem 2.1. *Let $g : \Re^k \to \Re$ a function of the form*

$$g(x_1, x_2, ..., x_m) = \sum_{i=1}^{m} q_i^{1+a} \Phi(x_i/q_i),$$

with $\Phi(u)$ any function such that $\Phi'(1) = 0$ and $\Phi''(1) \neq 0$ and q_i known. Then

$$\sqrt{N}\left[g(\widehat{p}_1, ..., \widehat{p}_m) - g(p_1, ..., p_m)\right] \xrightarrow{L} N\left(0, \sigma_a^2\right)$$

where

$$\sigma_a^2 = \left\{ \sum_{j=1}^{m} p_j \left[q_j^a \Phi'\left(\frac{p_j}{q_j}\right)\right]^2 - \left[\sum_{j=1}^{m} p_j q_j^a \Phi'\left(\frac{p_j}{q_j}\right)\right]^2 \right\}$$

and $\widehat{p}_i = \frac{x_i}{N}, i = 1, .., m$.

Proof. Since $X = (x_1, x_2, ..., x_m)$ is a random observation from the multinomial distribution $M(N, P)$, $P = (p_1, p_2, ..., p_m)$ and $\widehat{p}_i = \frac{x_i}{N}, i = 1, .., m$ it follows that (see, e.g., Serfling, 1980, p. 108–109),

$$\sqrt{N}(\widehat{p}_1 - p_1, \widehat{p}_2 - p_2, ..., \widehat{p}_m - p_m) \xrightarrow{L} N(0, \Sigma),$$

where the variance-covariance matrix is given by $\Sigma = [\sigma_{ij}]_{mxm}$,

$$\sigma_{ij} = \begin{cases} p_i(1 - p_i), & i = j, \\ -p_i p_j, & i \neq j. \end{cases}$$

The theorem is derived by applying the well known Delta method to the case under investigation (for a similar result see Rao, 1973, p. 387) with

$$\sigma_a^2 = \sum_{i=1}^{m} \sum_{j=1}^{m} \sigma_{ij} \frac{\partial g}{\partial p_i} \frac{\partial g}{\partial p_j},$$

where $\frac{\partial g}{\partial p_k} = q_k^a \Phi'(p_k/q_k)$, $k = 1, 2, \ldots, m$. Indeed, in this case we have

$$\begin{aligned}\sigma_a^2 &= \sum_{i=1}^{m} p_i(1-p_i)\left[q_i^a \Phi'\Big(\frac{p_i}{q_i}\Big)\right]^2 - \sum\sum\nolimits_{i\neq j} p_i p_j \left[q_i^a \Phi'\Big(\frac{p_i}{q_i}\Big)\right]\left[q_j^a \Phi'\Big(\frac{p_j}{q_j}\Big)\right] \\ &= \sum_{i=1}^{m} p_i \left[q_i^a \Phi'\Big(\frac{p_i}{q_i}\Big)\right]^2 - \sum_{i=1}^{m} p_i^2 \left[q_i^a \Phi'\Big(\frac{p_i}{q_i}\Big)\right]^2 \\ &\qquad - \sum\sum\nolimits_{i\neq j} p_i p_j \left[q_i^a \Phi'\Big(\frac{p_i}{q_i}\Big)\right]\left[q_j^a \Phi'\Big(\frac{p_j}{q_j}\Big)\right]\end{aligned}$$

and the result is immediate. □

Corollary 2.1. *Let d_a as in (6), $\widehat{d_a}$ as in (10), and any function Φ such that $\Phi'(1) = 0$ and $\Phi''(1) \neq 0$ with $q_i \equiv p_{i0}$, $i = 1, \ldots, m$. Then*

$$\sqrt{N}\left[\widehat{d_a} - d_a\right] \xrightarrow{L} N\left(0, \sigma_a^2\right),$$

where

$$\sigma_a^2 = \left\{\sum_{j=1}^{m} p_j \left[p_{j0}^a \Phi'\left(\frac{p_j}{p_{j0}}\right)\right]^2 - \left[\sum_{j=1}^{m} p_j p_{j0}^a \Phi'\left(\frac{p_j}{p_{j0}}\right)\right]^2\right\}.$$

Proof. It follows immediately from the previous theorem. □

Consider the hypothesis

$$H_0 : p_i = p_{i0} \ vs.\ H_1 : p_i = p_{ib},\ i = 1, \ldots, m.$$

Suppose that the null hypothesis indicates that $p_i = p_{i0}$, $i = 1, 2, \ldots, m$ when in fact it is $p_i = p_{ib}$, $\forall i$. As it is well known if p_{i0} and p_{ib} are fixed then as n tends to infinity then the power of the test tends to 1. In order to examine the situation when the power is not close to 1, we must make it continually harder for the test as n increases. This can be done by allowing the alternative hypothesis steadily closer to the null hypothesis. As a result we define a sequence of alternative hypotheses as follows

$$H_{1,n} : p_i = p_{in} = p_{i0} + d_i/\sqrt{n}, \quad \forall i \tag{11}$$

which is known as Pitman transition alternative or Pitman (local) alternative or local contiguous alternative to the null hypothesis $H_0 : p_i = p_{i0}$.

In vector notation the null hypothesis and the local contiguous alternative hypotheses take the form

$$H_0 : p = p_0 \ v.s. \ H_{1,n} : p = p_n = p_0 + d/\sqrt{n},$$

where $p = (p_1, \dots, p_m)'$, $p_n = (p_{1n}, p_{2n}, \dots, p_{mn})'$, and $d = (d_1, \dots, d_m)'$ is a fixed vector such that $\sum_{i=1}^{m} d_i = 0$. Observe that as n tends to infinity the local contiguous alternative converges to the null hypothesis at the rate $O(n^{-1/2})$.

In order to derive the asymptotic distribution of the test statistic (9) under the local contiguous alternatives $H_{i,n}$, we first obtain the asymptotic distribution of $\widehat{p}_i$ the maximum likelihood estimator of p_i.

Theorem 2.2. *Under the local contiguous alternative hypotheses* (11), *we have*

$$\sqrt{n}(\widehat{p} - p_n) \xrightarrow{L} N(0, \Sigma) \quad and \quad \sqrt{n}(\widehat{p} - p_0) \xrightarrow{L} N(d, \Sigma),$$

where $\widehat{p} = (\widehat{p}_1, \dots, \widehat{p}_m)'$ *and* Σ *as in the proof of Theorem* 2.1.

Proof. Observe that when indeed $p_i = p_{in}$, $\forall i$ and $\widehat{p}_i$ the maximum likelihood estimator of p_i then

$$\sqrt{n}\frac{(\widehat{p}_i - p_{in})}{\sqrt{p_{in}(1 - p_{in})}} \xrightarrow{L} N(0, 1).$$

Observe also that

$$\sqrt{\frac{p_{in}}{p_{i0}}} = \sqrt{1 + \frac{p_{in} - p_{i0}}{p_{i0}}} = \sqrt{1 + \frac{d_i}{\sqrt{n}p_{i0}}},$$

which converges to 1 as $n \to \infty$. In a similar fashion one can easily show that

$$\sqrt{\frac{1 - p_{in}}{1 - p_{i0}}} = \sqrt{1 - \frac{d_i}{\sqrt{n}(1 - p_{i0})}}$$

which converges also to 1 as $n \to \infty$. As a result,

$$\sqrt{n}\frac{(\widehat{p}_i - p_{in})}{\sqrt{p_{in}(1 - p_{in})}} \cdot \frac{\sqrt{p_{in}(1 - p_{in})}}{\sqrt{p_{i0}(1 - p_{i0})}} = \sqrt{n}\frac{(\widehat{p}_i - p_{in})}{\sqrt{p_{i0}(1 - p_{i0})}} \xrightarrow{L} N(0, 1).$$

It is easily seen that

$$\sqrt{n}(\widehat{p}_i - p_{i0}) = \sqrt{n}(\widehat{p}_i - p_{in}) + \sqrt{n}(\widehat{p}_{in} - p_{i0}) = \sqrt{n}(\widehat{p}_i - p_{in}) + d_i.$$

Hence,

$$\sqrt{n}(\widehat{p}_i - p_{i0}) \xrightarrow{L} N(d_i, p_{i0}(1 - p_{i0})).$$

The conclusion for the m-dimensional vector parameter is straight forward if we take into consideration Serfling (1980, pp. 108–109). □

Note that under the null hypothesis $H_0 : p_i = p_{i0}$ we have

$$\sqrt{n}(\widehat{p}_i - p_{i0}) \xrightarrow{L} N(0, p_{i0}(1 - p_{i0})).$$

We define now the noncentral chi-square distribution.

Definition 2.4. *If $X_1, \dots, X_m$ are independent random variables with $X_i \sim N(\xi_i, 1)$, the distribution of $\sum_{i=1}^{m} X_i^2$ is noncentral chi-square with m degrees of freedom and noncentrality parameter $\delta = \sum_{i=1}^{m} \xi_i^2$. In matrix notation we say that if $X \sim N(\xi, I)$ then $X'X \sim \mathcal{X}^2_{m,\delta}$, with $\delta = \xi'\xi$ where $X = (X_1, \dots, X_m)'$, $\xi = (\xi_1, \dots, \xi_m)'$ and I the mxm identity matrix.*

The following lemma from Hunter (2002, p. 72) which will be used later is presented below without proof. The lemma provides conditions for the noncentral chi-square distribution but applies also to the chi-square distribution when ξ is taken to be 0. In what follows $W = \sum_{i=1}^{m} \frac{N}{p_{i0}} \left(\frac{n_i}{N} - p_{i0}\right)^2$.

Lemma 2.1. *Suppose that $X \sim N(\xi, Q)$ where Q is a projection matrix of rank $r \leqslant m$ and $Q\xi = \xi$. Then, $X'X \sim \mathcal{X}^2_{r,\xi'\xi}$.*

Theorem 2.3. *Let $(n_1, ..., n_m) \sim M(N, P)$ with $P = (p_1, ..., p_m)$ and p_i, $i = 1, ..., m$ unknown parameters. Under the local contiguous alternative hypotheses $H_{i,n} : p_i = p_{in}$, $i = 1, ..., m$ we have:*

- $\left(\min_i p_{i0}^a\right) W \prec_{\text{st}} \sum_{i=1}^{m} \frac{N p_{i0}^a}{p_{i0}} \left(\frac{n_i}{N} - p_{i0}\right)^2 \prec_{\text{st}} \left(\max_i p_{i0}^a\right) W$;
- $X_a^2 - \sum_{i=1}^{m} \frac{N p_{i0}^a}{p_{i0}} \left(\frac{n_i}{N} - p_{i0}\right)^2 \xrightarrow{P} 0$ *and*
- *the distribution of (9) is estimated to be approximately $c\mathcal{X}^2_{m-1,\delta}$, with $c = 0.5(\min_i p_{i0}^a + \max_i p_{i0}^a)$,*

where $\mathcal{X}^2_{m-1,\delta}$ is the noncentral chi-square distribution with $m-1$ degrees of freedom and noncentrality parameter $\delta = \sum_{i=1}^{m} \frac{d_i^2}{p_{i0}}$ and $\prec_{\text{st}}$ the symbol for stochastic ordering.

Proof. The Taylor expansion of Φ in an open ball $_\varepsilon(p_i/p_{i0})$ of radius ε around the point p_i/p_{i0}, $i = 1, 2, ..., m$, is given by:

$$\Phi\left(\frac{\widehat{p}_i}{p_{i0}}\right) = \Phi\left(\frac{p_i}{p_{i0}}\right) + \left(\frac{\widehat{p}_i}{p_{i0}} - \frac{p_i}{p_{i0}}\right)\Phi'\left(\frac{p_i}{p_{i0}}\right)$$
$$+ \frac{1}{2}\left(\frac{\widehat{p}_i}{p_{i0}} - \frac{p_i}{p_{i0}}\right)^2 \Phi''\left(\frac{p_i}{p_{i0}}\right) + o\left(\left(\widehat{\frac{p_i}{p_{i0}}} - \frac{p_i}{p_{i0}}\right)\right)^2.$$

Multiplying both sides of the above relation by Np_{i0}^{1+a}, and taking the sum of both sides for $i = 1, 2, ..., m$ we get

$$\sum_{i=1}^{m} Np_{i0}^{1+a}\Phi\left(\frac{\widehat{p}_i}{p_{i0}}\right) = \sum_{i=1}^{m} Np_{i0}^{1+a}\Phi\left(\frac{p_i}{p_{i0}}\right)$$
$$+ \sum_{i=1}^{m} Np_{i0}^{1+a}\left(\frac{\widehat{p}_i}{p_{i0}} - \frac{p_i}{p_{i0}}\right)\Phi'\left(\frac{p_i}{p_{i0}}\right)$$
$$+ \frac{1}{2}\sum_{i=1}^{m} Np_{i0}^{1+a}\left(\frac{\widehat{p}_i}{p_{i0}} - \frac{p_i}{p_{i0}}\right)^2 \Phi''\left(\frac{p_i}{p_{i0}}\right)$$
$$+ \sum_{i=1}^{m} Np_{i0}^{1+a} o\left(\left(\widehat{\frac{p_i}{p_{i0}}} - \frac{p_i}{p_{i0}}\right)\right)^2.$$

which for $p_i = p_{i0}$ becomes:

$$N\widehat{d}_a - N\Phi(1)\sum_{i=1}^{m} p_{i0}^{1+a} - \frac{1}{2}\Phi''(1)\sum_{i=1}^{m}\frac{Np_{i0}^a}{p_{i0}}\left(\frac{n_i}{N} - p_{i0}\right)^2$$
$$= N\sum_{i=1}^{m} p_{i0}^a\left(\frac{n_i}{N} - p_{i0}\right)\Phi'(1) + \sum_{i=1}^{m} N\frac{p_{i0}^a}{p_{i0}} o((\widehat{p}_i - p_{i0}))^2. \qquad (12)$$

where $\widehat{p} = (n_1/N, \ldots, n_m/N)'$ and $p_0 = (p_{10}, \ldots, p_{m0})'$. But

$$\sum_{i=1}^{m} N\frac{p_{i0}^a}{p_{i0}} o((\widehat{p}_i - p_{i0}))^2 \leq \max_i\left\{\frac{p_{i0}^a}{p_{i0}}\right\}\sum_{i=1}^{m} No((\widehat{p}_i - p_{i0}))^2$$
$$= \max_i\left\{\frac{p_{i0}^a}{p_{i0}}\right\} \cdot N \cdot o(||\widehat{p} - p_0||)^2 = o_P(1) \qquad (13)$$

since $\sqrt{N}(\widehat{p}-p_0) \xrightarrow{L} N(d,\Sigma)$ where Σ as in the proof of Theorem 2.1 (see Serfling, 1980, pp. 108-109). From (12) and (13), we conclude that

$$\frac{2N\left(\widehat{d}_a-\Phi\left(1\right)\sum\limits_{i=1}^{m}p_{i0}^{1+a}\right)}{\Phi''\left(1\right)}-\sum_{i=1}^{m}\frac{Np_{i0}^{a}}{p_{i0}}\left(\frac{n_i}{N}-p_{i0}\right)^2 \xrightarrow{P} 0.$$

Observe that

$$\left(\min_i p_{i0}^a\right) W \prec_{\rm st} \sum_{i=1}^{m}\frac{Np_{i0}^{a}}{p_{i0}}\left(\frac{n_i}{N}-p_{i0}\right)^2 \prec_{\rm st} \left(\max_i p_{i0}^a\right) W.$$

Let P a diagonal matrix with diagonal elements the inverses of the elements of the vector p_0. Then, we have

$$W=N(\widehat{p}-p_0)'P(\widehat{p}-p_0)=\left(\sqrt{N}\left(P^{1/2}(\widehat{p}-p_0)\right)'\right)\left(\sqrt{N}\left(P^{1/2}(\widehat{p}-p_0)\right)\right)$$

so that

$$\sqrt{N}\left(P^{1/2}(\widehat{p}-p_0)\right) \xrightarrow{L} N(P^{1/2}d, P^{1/2}\Sigma P^{1/2}).$$

Lemma 2.1 can now be applied provided that the matrix $P^{1/2}\Sigma P^{1/2}$ is of rank $m-1$ and that $(P^{1/2}\Sigma P^{1/2})\cdot(P^{1/2}d)=P^{1/2}d$.

For the first condition we have

$$P^{1/2}\Sigma P^{1/2}=P^{1/2}[P^{-1}-p_0p_0']P^{1/2}=I-P^{1/2}p_0p_0'P^{1/2}=I-\sqrt{p_0}\sqrt{p_0}'$$

which clearly is symmetric with trace equal to $m-1$. The sum of its eigenvalues is also equal to $m-1$ since for symmetric matrices the trace and the sum of the eigenvalues coincide. Furthermore, since $\sqrt{p_0}'\sqrt{p_0}=1$ we have that

$$\begin{aligned}(I-\sqrt{p_0}\,\sqrt{p_0'})&(I-\sqrt{p_0}\,\sqrt{p_0'})\\ &=I-2\sqrt{p_0}\,\sqrt{p_0'}+\sqrt{p_0}\,\sqrt{p_0'}\,\sqrt{p_0}\,\sqrt{p_0'}=I-\sqrt{p_0}\,\sqrt{p_0'}\end{aligned}$$

and hence, the matrix $P^{1/2}\Sigma P^{1/2}$ is a projection matrix with implies that its eigenvalues are all equal to 0 or 1. As a result, there are $m-1$ eigenvalues equal to 1.

The second condition is easily established since

$$P^{1/2}\Sigma Pd = P^{1/2}[P^{-1} - p_0 p_0']Pd = P^{1/2}[d - p_0(1)'d],$$

where the second term vanishes since $(1)'d = \sum_{i=1}^{m} d_i = 0$, $\Sigma = P^{-1} - p_0 p_0'$ the covariance matrix appearing in the proof of Theorem 2.1 and (1) an m-dimensional vector with elements equal to 1.

As a result, under the local contiguous alternative hypotheses $H_{i,n}$ and as $N \to \infty$ we observe the noncentral distribution, namely,

$$\sum_{i=1}^{m} \frac{N}{p_{i0}} \left(\frac{n_i}{N} - p_{i0}\right)^2 \xrightarrow{L} \mathcal{X}^2_{m-1,\delta},$$

where the noncentrality parameter δ is given by $\delta = (P^{1/2}d)'P^{1/2}d = d'Pd$. Hence, the asymptotic distribution of the test statistic (9) under the contiguous alternatives $H_{i,n}$ is $c\mathcal{X}^2_{m-1,\delta}$ where $c = 1/2(\min_i p^a_{i0} + \max_i p^a_{i0})$.
□

The following corollary follows as a natural consequence of the above Theorem and is furnished without a proof. It provides the asymptotic distribution of the test statistic under the null hypothesis $H_0 : p_i = p_{i0}$.

Corollary 2.2. *Let $(n_1, ..., n_m) \sim M(N, P)$ with $P = (p_1, ..., p_m)$ and p_i, $i = 1, ..., m$ unknown parameters. Under the null hypothesis $H_0 : p_i = p_{i0}$, $i = 1, ..., m$ we have:*

- $\left(\min_i p^a_{i0}\right) W \prec_{\text{st}} \sum_{i=1}^{m} \frac{N p^a_{i0}}{p_{i0}} \left(\frac{n_i}{N} - p_{i0}\right)^2 \prec_{\text{st}} \left(\max_i p^a_{i0}\right) W$;
- $X^2_a - \sum_{i=1}^{m} \frac{N p^a_{i0}}{p_{i0}} \left(\frac{n_i}{N} - p_{i0}\right)^2 \xrightarrow{P} 0$ *and*
- *the distribution of* (9) *is estimated to be approximately* $c\mathcal{X}^2_{m-1}$, *with* $c = 0.5(\min_i p^a_{i0} + \max_i p^a_{i0})$

where $\mathcal{X}^2_{m-1}$ is the chi-square distribution with $m-1$ degrees of freedom and $\prec_{\text{st}}$ the symbol for stochastic ordering.

Observe that in the theorem above we assume that $\Phi'(1) = 0$. This assumption is necessary if the test statistic used is the one given by (9). It is easy to see and it will be evident immediately after the Theorem 2.4 that this assumption is satisfied not only for the discrete BHHJ measure but also for all measures covered by the Csiszár's family of measures. If though one selects a function Φ which does not satisfy this assumption then the appropriate test statistic has to be defined. It is not difficult to see that in such a case (12) is the main expression affected since the first term on the right-hand side of the expression does not vanish. The resulting test statistic will be given by

$$\Psi_a^2 \equiv \frac{2N\left(\widehat{d}_a - \Phi(1)\sum_{i=1}^{m} p_{i0}^{1+a} - \sum_{i=1}^{m} p_{i0}^{a}\left(\frac{n_i}{N} - p_{i0}\right)\Phi'(1)\right)}{\Phi''(1)}. \tag{14}$$

It should be noted though that for values of a close to zero the last term in the numerator of (14) vanishes since $\sum_{i=1}^{m} p_{i0}^{a}\left(\frac{n_i}{N} - p_{i0}\right) \approx 0$.

Due to the above theorems the power of the test under the fixed alternative hypothesis $H_1 : p_i = p_{ib}$ and the local contiguous alternative hypotheses (11) can be easily obtained. For the case of the local contiguous alternative hypotheses, the power is given by

$$\gamma_n = P(X^2 > \mathcal{X}_{m-1,\alpha}^2 | p_i = p_{in}, i = 1, ..., m) = P(\mathcal{X}_{m-1,\delta}^2 > \mathcal{X}_{m-1,\alpha}^2).$$

For the fixed alternative hypothesis the power is given in the theorem below:

Theorem 2.4. *The power of the test* $H_0 : p_i = p_{i0}$ *vs* $H_a : p_i = p_{ib}$, $i = 1, ..., m$ *using the test statistic* (9) *is approximately equal to:*

$$\gamma_a = P\left(Z \geqslant \frac{\Phi''(1)\, c\mathcal{X}_{m-1,\alpha}^2 + 2N\Phi(1)\sum_{i=1}^{m} p_{i0}^{1+a} - 2Nd_a}{2\sqrt{N}\sigma_a}\right), \tag{15}$$

where Z *a standard Normal random variable,* $\mathcal{X}_{m-1,\alpha}$ *the* α*-percentile of the* $\mathcal{X}_{m-1}^2$ *distribution, and* σ_a^2 *as in Corollary* 2.1 *with* $p_i = p_{ib}$.

Proof. By definition, the power is given by

$$\gamma_a = P\left(X_a^2 \geqslant c\mathcal{X}_{m-1,\alpha}^2 \middle| p_i = p_{ib}, i = 1, ..., m\right)$$

$$= P\left(\widehat{d}_a \geqslant (2N)^{-1}\Phi''(1)\, c\mathcal{X}_{m-1,\alpha}^2 + \Phi(1)\sum_{i=1}^{m} p_{i0}^{1+a} \middle| p_i = p_{ib}, i = 1, ..., m\right).$$

From Corollary 2.1 with $p_j = p_{jb}$, $j = 1, ..., m$, we have

$$\sqrt{N}\sigma_a^{-1}\left[\widehat{d}_a - d_a\right] \xrightarrow{L} N(0,1).$$

The result is immediate. □

Note that for the BHHJ test corresponding to the measure given in (6) and (4) we have $\Phi''(1) = 1 + a$ and $\Phi(1) = \Phi'(1) = 0$ so that the BHHJ statistic corresponding to the goodness of fit test of Theorem 2.3 is given by

$$X_a^2 \equiv \frac{2N\widehat{d}_a}{1+a} \tag{16}$$

while its power is given by

$$\gamma_a = P\left(Z \geqslant \frac{(1+a)c\mathcal{X}^2_{m-1,\alpha} - 2Nd_a}{2\sqrt{N}\sigma_a}\right). \tag{17}$$

Note also that the Csiszár's statistic corresponding to the goodness of fit test of Theorem 2.3 is given by

$$X_c^2 \equiv \frac{2N\left(\widehat{d}_c - \varphi(1)\right)}{\varphi''(1)} \tag{18}$$

while its power is given by

$$\gamma_c = P\left(Z \geqslant \frac{\varphi''(1)\,\mathcal{X}^2_{m-1,\alpha} + 2N\varphi(1) - 2Nd_c}{2\sqrt{N}\sigma_a}\right), \tag{19}$$

where $d_c = \sum_{i=1}^{m} p_{i0}\varphi(p_i/p_{i0})$ and $\widehat{d}_c = \sum_{i=1}^{m} p_{i0}\varphi(\widehat{p}_i/p_{i0})$. For the usual Kullback–Leibler, Kagan and Cressie and Read measures we can easily see that $\varphi(1) = 0$ and $\varphi''(1) = 1$ so that the power is simplified into the form

$$\gamma_c = P\left(Z \geqslant \frac{\mathcal{X}^2_{m-1,\alpha} - 2Nd_c}{2\sqrt{N}\sigma_a}\right) \tag{20}$$

where

$$\sigma_a^2 = \sum_{i=1}^{m} p_{ib}\left[\varphi'\left(\frac{p_{ib}}{p_{i0}}\right)\right]^2 - \left[\sum_{j=1}^{m} p_{ib}\varphi'\left(\frac{p_{ib}}{p_{i0}}\right)\right]^2$$

and $\varphi'(x) = \log x$ (Kullback–Leibler), $\varphi'(x) = x - 1$ (Kagan), and $\varphi'(x) = \frac{1}{\lambda}(x^{\lambda} - 1)$ (Cressie and Read). For the Matusita measure it is not difficult to provide the appropriate expressions for the test statistic and the power since we can easily see that $\varphi(1) = 0$, $\varphi'(x) = 1 - x^{-1/2}$ and $\varphi''(1) = 1/2$.

3. Simulations

For checking the accuracy of the proposed BHHJ test simulated results using trinomial distributions are presented in the present section. In order to understand the behavior of the BHHJ test we compare it with four other tests, namely the goodness of fit tests based on the Kullback measure (KL), the Kagan measure, the Matusita measure (Mat), and the Cressie and Read measure with $\lambda = 2/3$ (CR).

Our analysis is based on the equiprobable null hypothesis $H_0 : p_i = 1/k$, $i = 1, \dots, k$ which is extensively used in the literature, primarily in small-sample studies. For this hypothesis, we consider the set of alternatives given by

$$H_1 : p_i = \begin{cases} \{1 - \eta/(k-1)\}/k, & i = 1, ..., k-1, \\ (1+\eta)/k, & i = k, \end{cases}$$

where $-1 < \eta < k - 1$. Note that for $\eta > 0$ a bump alternative and for $\eta < 0$ a dip alternative is obtained. Based on 10000 simulations for $k = 3$, Table 1 provides the powers for the KL, CR ($\lambda = 2/3$), Matusita, Kagan and BHHJ ($\alpha = 0.01$) test statistics for small ($n = 25$) and moderate ($n = 150$) sample sizes and for dip ($\eta = -0.7$ and -0.6) and bump ($\eta = +0.4$ and $+.06$) alternatives. Our results show that the BHHJ test statistic is superior since it performs as well as the other tests for bump alternatives and is the most powerful among all competing tests for dip alternatives.

Table 1. Power calculations for the equiprobable null hypothesis ($k = 3$ & $a = 0.05$)

	$\eta = -0.7$		$\eta = -0.6$		$\eta = +0.4$		$\eta = +0.6$	
Test	$n = 25$	150	$n = 25$	150	$n = 25$	150	$n = 25$	150
KL	.181	.682	.296	.927	.240	.857	.454	.997
Kagan	.137	.661	.227	.916	.218	.862	.439	.997
Mat	.180	.682	.296	.927	.241	.857	.455	.996
CR	.138	.671	.228	.920	.218	.856	.440	.997
BHHJ	.181	.692	.296	.931	.241	.853	.455	.997

References

1. A. Agresti, *Analysis of Ordinal Categorical Data*, John Wiley (1984).
2. S. M. Ali, S. D. Silvey, *A general class of coefficients of divergence of one distribution from another.* — J. R. Statist. Soc. B, **28** (1966), 131–142.
3. A. Basu, I. R. Harris, N. L. Hjort, M. C. Jones, *Robust and efficient estimation by minimising a density power divergence.* — Biometrika **85** (1998), 549–559.
4. D. R. Cox, *The Analysis of Binary Data*, London, Methuen (1970).
5. N. Cressie, T. R. C. Read, *Multinomial goodness-of-fit tests.* — J. R. Statist. Soc. B **5** (1984), 440–464.
6. N. Cressie, T. R. C. Read, *Goodness-of-Fit Statistics for Discrete Multivariate Data.* Springer Verlag, New York (1988).
7. I. Csiszar, *Eine informationstheoretische ungleichung and ihre anwendung auf den beweis der ergodizitat von markoffischen ketten.* — Magyar Tud. Akad. Mat. Kutato Int. Kozl. **8** (1963), 85–108.
8. I. Csiszar, *Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems.* — Ann. Statist. **19** (1991), 2032–2066.
9. D. Hunter, *Lecture notes in Asymptotic Tools*, Penn-State Univ (2002).
10. L. K. Jones, C. L. Byrne, *General entropy criteria for inverse problems, with applications to data compression, pattern classification, and cluster analysis.* — IEEE Trans. Info. Theory **36** (1990), 23–30.
11. A. M. Kagan, *On the theory of Fisher's amount of information.* — Dokl. Akad. Nauk SSSR **151** (1963), 277–278.
12. F. Liese, I. Vajda, *Convex Statistical Distances.* Teubner, Leipzig (1987).
13. K. Matusita, *On the notion of affinity of several distributions and some of its applications.* — Ann. Inst. Statist. Math. **19** (1967), 181–192.
14. D. Morales, L. Pardo, I. Vajda, *Some new statistics for testing hypotheses in parametric models.* — J. Multivariate Anal. **62 (1)** (1997), 137–168.
15. L. Pardo, *Generalized divergence measures: statistical applications.* — In: Encyclopedia of Microcomputers, Marcel–Dekker (1999), pp. 163–191.
16. L. Pardo, *Statistical inference based on divergence measures.* Chapman and Hall/CRC (2006).
17. C. R. Rao, *Linear Statistical Inference and Its Applications.* John Wiley (1973).
18. A. Renyi, *On measures of entropy and information.* — In: Proc. 4th Berkeley Symp. on Math. Statist. Prob. **1**, Univ. California Press (1961), pp. 547–561.
19. R. J. Serfling, *Approximations Theorems of Mathematical Statistics.* John Wiley (1980).
20. K. Zografos, K. Ferentinos, T. Papaioannou, *Φ-divergence statistics: Sampling properties, multinomial goodness of fit and divergence tests.* — Commun. Statist. Theory, Methods **19(5)** (1990), 1785–1802.

Department of Mathematics and Statistics,
University of Cyprus, Nicosia, Cyprus,
E-mail: alex@ucy.ac.cy

Поступило 10 декабря 2008 г.