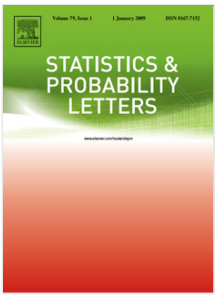
Hypothesis testing for two discrete populations based on the Hellinger distance

A. Basu, A. Mandal, L. Pardo



Hypothesis testing for two discrete populations based on the Hellinger distance

A. Basu, A. Mandal, L. Pardo

PII: S0167-7152(09)00394-0
DOI: 10.1016/j.spl.2009.10.008
Reference: STAPRO 5547

To appear in: *Statistics and Probability Letters*

Received date: 19 August 2008
Revised date: 3 April 2009
Accepted date: 13 October 2009

Please cite this article as: Basu, A., Mandal, A., Pardo, L., Hypothesis testing for two discrete populations based on the Hellinger distance. *Statistics and Probability Letters* (2009), doi:10.1016/j.spl.2009.10.008

Hypothesis Testing for two Discrete Populations based on the Hellinger Distance

A. Basu[1], A. Mandal[2] and L. Pardo[3]

[1]Bayesian and Interdisciplinary Research Unit, Indian Statistical Institute, India

[2]Applied Statistics Unit, Indian Statistical Institute, India

[3]Department of Statistics and O.R. I, Complutense University of Madrid, Spain

April 3, 2009

Abstract

Our interest is in the problem where independent samples are drawn from two different discrete populations, possibly with a common parameter. The goal is to test hypothesis about the parameters involved in this two sample situation. A number of tests are developed for the above purpose based on the Hellinger distance and penalized versions of it. The asymptotic distribution of the test statistics are derived. Extensive simulation results are provided, which illustrate the theory developed and the robustness of the methods.

April 3, 2009

keywords and phrases: Minimum Hellinger distance estimator; Empty cell penalty; Asymptotic distributions.

1 Introduction

Let $X_1, \ldots, X_{m_1}$ and $Y_1, \ldots, Y_{m_2}$ be two independent random samples from two discrete populations X and Y with common support $\mathcal{X} = \{x_0, x_1, \ldots\}$ and probability mass functions $f_{\theta_1}(x) = \Pr_{\theta_1}(X = x)$ and $f_{\theta_2}(x) = \Pr_{\theta_2}(Y = x)$, $x \in \mathcal{X}$, $\theta_i \in \Theta \subset \mathbb{R}$, $i = 1, 2$, respectively. We denote by $n_1(x)$ ($n_2(x)$) the number of elements in the sample $X_1, \ldots, X_{m_1}$ ($Y_1, \ldots, Y_{m_2}$) that coincide with $x \in \mathcal{X}$. We are interested in performing tests of hypothesis involving both θ_1 and θ_2. In order to keep the exposition short and notation simple, we have assumed θ_1 and θ_2 to be scalar parameters in this paper. The multiparameter cases can be handled by extensions of essentially the same ideas. In this paper, therefore, we will restrict ourselves to the problem of testing

$$H_0 : \theta_1 = \theta_2 \tag{1}$$

on the basis of some new statistics introduced in this paper. The test statistics considered here are based on the Hellinger distance between two different probability vectors. The unknown parameters are estimated by minimizing the Hellinger distance between the data and the model probability vectors, or a penalized version of it.

The statistics are introduced in Section 2; their asymptotic distributions are also derived in this section. The performance of the proposed tests are demonstrated numerically in Section 3.

2 Hellinger Distance: Estimation and Testing

The (twice) squared Hellinger distance between the probability vectors

$$\boldsymbol{d}_i = (d_i(x_0), \ldots, d_i(x_j), \ldots) = \left(\frac{n_i(x_0)}{m_i}, \ldots, \frac{n_i(x_j)}{m_i}, \ldots \right), \ i = 1, 2, \tag{2}$$

and

$$\boldsymbol{f}_{\theta_i} = (f_{\theta_i}(x_0), \ldots, f_{\theta_i}(x_j), \ldots), \ i = 1, 2, \tag{3}$$

is given by

$$HD\left(\boldsymbol{d}_i, \boldsymbol{f}_{\theta_i}\right) = 2\sum_{j=0}^{\infty} \left(d_i^{1/2}(x_j) - f_{\theta_i}^{1/2}(x_j) \right)^2, \ i = 1, 2, \tag{4}$$

and the minimum Hellinger distance estimator of θ_i is defined as the value $\widehat{\theta}_H^i$ of $\Theta \subset \mathbb{R}$ satisfying

$$\widehat{\theta}_H^i = \arg\min_{\theta_i} HD\left(\boldsymbol{d}_i, \boldsymbol{f}_{\theta_i}\right). \tag{5}$$

See Beran (1977), Simpson (1987, 1989) and Basu et al. (1997) for more details on this method of estimation.

It has been empirically observed that the minimum Hellinger distance estimator often performs poorly in small samples – compared to the maximum likelihood estimator – when the data generating distribution is correctly specified by the parametric model. To avoid this problem, one of the suggestions is to use the penalized Hellinger distance (*e.g.* Harris and Basu 1994; Basu Harris and Basu 1996; Basu and Basu 1998). In our context the penalized Hellinger distance between the probability vectors $\boldsymbol{d}_i$, defined in (2), and $\boldsymbol{f}_{\theta_i}$, defined in (3), is given by

$$PHD\left(\boldsymbol{d}_i, \boldsymbol{f}_{\theta_i}\right) = \sum_{j \in A_i}^{\infty} \left(d_i^{1/2}(x_j) - f_{\theta_i}^{1/2}(x_j) \right)^2 + h \sum_{j \in A_i^C}^{\infty} f_{\theta_i}(x_j) \ , \tag{6}$$

where h is a real, positive number and

$$A_i = \{j : d_i(x_j) > 0\} \ \text{ and } \ A_i^C = \{j : d_i(x_j) = 0\}.$$

As in the definition given in (5), the minimum penalized Hellinger distance estimator of θ_i, $i = 1, 2$, is given by

$$\widehat{\theta}_{PH}^i = \arg\min_{\theta_i} PHD\left(\boldsymbol{d}_i, \boldsymbol{f}_{\theta_i}\right). \tag{7}$$

While the penalized Hellinger distance is defined for any real, positive h, values around $h = 1$ appear to be preferable for small sample efficiency; for the penalized distances $h = 1$ will be our default value. The rationale of this choice is that $h = 1$ makes the weight on the empty cells identical with that applied by likelihood based methods. Note that $h = 2$ generates the ordinary Hellinger distance.

As the probability of the empty cells eventually goes to zero it is intuitive that the results based on the ordinary and penalized Hellinger distance will provide equivalent asymptotic results. Mandal et al. (2008) prove that for any fixed h, the methods based on the ordinary Hellinger distance and the penalized Hellinger distance have the same asymptotic inference properties in the one sample problem.

Based on (5) and (6) it seems natural to consider test statistics of the type

$$HD\left(\boldsymbol{f}_{\widehat{\theta}_H^1}, \boldsymbol{f}_{\widehat{\theta}_H^2}\right) \tag{8}$$

and

$$HD\left(\boldsymbol{f}_{\widehat{\theta}_{PH}^1}, \boldsymbol{f}_{\widehat{\theta}_{PH}^2}\right) \tag{9}$$

for testing the null hypothesis in (1).

The idea of considering test statistics of the type (8) and (9) has been used before in the literature. Kupperman (1957) considered, for the first time, the test statistic

$$2\frac{m_1 m_2}{m_1+m_2} D_{Kull}(\boldsymbol{f}_{\widehat{\theta}_1}, \boldsymbol{f}_{\widehat{\theta}_2}) \ , \tag{10}$$

where $D_{Kull}(\boldsymbol{f}_{\widehat{\theta}_1}, \boldsymbol{f}_{\widehat{\theta}_2})$ is the Kullback-Leibler divergence between $\boldsymbol{f}_{\widehat{\theta}_1}$ and $\boldsymbol{f}_{\widehat{\theta}_2}$. Its expression is given by

$$D_{Kull}(\boldsymbol{f}_{\widehat{\theta}_1}, \boldsymbol{f}_{\widehat{\theta}_2}) = \sum_{j=0}^{\infty} f_{\widehat{\theta}_1}(x_j) \log \frac{f_{\widehat{\theta}_1}(x_j)}{f_{\widehat{\theta}_2}(x_j)} \ .$$

For more details about Kullback-Leibler divergence see Kullback (1985). The symbols $\widehat{\theta}_1$ and $\widehat{\theta}_2$ represent the maximum likelihood estimators of θ_1 and θ_2, respectively. Kupperman (1957) established that the asymptotic distribution of the test statistic given in (10) is a chi-square with one degree of freedom. This result was extended by Salicru et al (1994) by considering the family of ϕ-divergence test statistics

$$2\frac{m_1 m_2}{m_1+m_2} D_\phi(\boldsymbol{f}_{\widehat{\theta}_1}, \boldsymbol{f}_{\widehat{\theta}_2})$$

where $D_\phi(\boldsymbol{f}_{\widehat{\theta}_1}, \boldsymbol{f}_{\widehat{\theta}_2})$ is the phi-divergence or phi-disparity between $\boldsymbol{f}_{\widehat{\theta}_1}$ and $\boldsymbol{f}_{\widehat{\theta}_2}$. Its expression is given by,

$$D_\phi(\boldsymbol{f}_{\widehat{\theta}_1}, \boldsymbol{f}_{\widehat{\theta}_2}) = \sum_{j=0}^{\infty} f_{\widehat{\theta}_2}(x_j) \phi\left(\frac{f_{\widehat{\theta}_1}(x_j)}{f_{\widehat{\theta}_2}(x_j)}\right), \qquad \phi \in \Phi^* \tag{11}$$

where Φ^* is the class of all convex functions $\phi(x)$, $x \geq 0$, such that, $\phi(1) = 0$ and $\phi''(1) \neq 0$. In (11) we shall assume the conventions $0\phi(0/0) = 0$ and $0\phi(p/0) = p \lim_{u\to\infty} \phi(u)/u$, for $p > 0$. Let $\phi \in \Phi^*$ be differentiable at $x = 1$, then the function $\psi(x) \equiv \phi(x) - \phi'(1)(x-1)$ also belongs to Φ^* and has the additional property that $\psi'(1) = 0$. This property, together with the convexity, implies that $\psi(x) \geq 0$, for any $x \geq 0$. Further, $D_\psi\left(\boldsymbol{f}_{\widehat{\theta}_1}, \boldsymbol{f}_{\widehat{\theta}_2}\right) = D_\phi\left(\boldsymbol{f}_{\widehat{\theta}_1}, \boldsymbol{f}_{\widehat{\theta}_2}\right)$.

In particular if we replace $\phi(x) = -4\left(\sqrt{x} - \frac{1}{2}(x+1)\right)$ in (11), we get

$$D_\phi(\boldsymbol{f}_{\widehat{\theta}_1}, \boldsymbol{f}_{\widehat{\theta}_2}) = HD\left(\boldsymbol{f}_{\widehat{\theta}_1}, \boldsymbol{f}_{\widehat{\theta}_2}\right).$$

For more details about ϕ-divergences see Pardo (2006) and Lindsay (1994). These divergences have also been referred to as disparities in the literature. Also see Sarkar and Basu (1995) who

considered a linear combination of divergences with weights proportional to their sample sizes in constructing an overall divergence involving two independent samples to test statistical hypotheses of the type given in equation (1).

The joint likelihood function based on $X_1, \ldots, X_{m_1}$ and $Y_1, \ldots, Y_{m_2}$ is given by

$$L(\theta_1, \theta_2) = \prod_{j=0}^{\infty} f_{\theta_1}(x_j)^{m_1(x_j)} f_{\theta_2}(x_j)^{m_2(x_j)}$$

and if we denote by $\widetilde{\theta}$ the maximum likelihood estimate of the common parameter under the hypothesis $\theta_1 = \theta_2$, the log likelihood ratio test statistic is given by

$$LRT = 2\left[\log L\left(\widehat{\theta}_1, \widehat{\theta}_2\right) - \log L\left(\widetilde{\theta}, \widetilde{\theta}\right)\right] \tag{12}$$

whose asymptotic distribution is a chi-square with one degree of freedom.

Now, we denote,

$$\boldsymbol{d}^* = \frac{1}{m_1 + m_2}\Big(n_1(x_0), n_2(x_0), \ldots, n_1(x_j), n_2(x_j), \ldots\Big) \tag{13}$$

and

$$\boldsymbol{f}_{\theta_1,\theta_2} = \frac{1}{m_1 + m_2}\Big(m_1 f_{\theta_1}(x_0), m_2 f_{\theta_2}(x_0), \ldots, m_1 f_{\theta_1}(x_j), m_2 f_{\theta_2}(x_j), \ldots\Big).$$

It is easy to see that the log likelihood ratio statistic in (12) is identical to

$$LRT = 2(m_1 + m_2)\left(D_{Kull}\left(\boldsymbol{d}^*, \boldsymbol{f}_{\widehat{\theta}_1,\widehat{\theta}_2}\right) - D_{Kull}\left(\boldsymbol{d}^*, \boldsymbol{f}_{\widetilde{\theta},\widetilde{\theta}}\right)\right) \tag{14}$$

and

$$LRT = 2(m_1 + m_2) D_{Kull}\left(\boldsymbol{f}_{\widehat{\theta}_1,\widehat{\theta}_2}, \boldsymbol{f}_{\widetilde{\theta},\widetilde{\theta}}\right) + o_p(1)\ . \tag{15}$$

The last equality follows using the same arguments that in Bishop (1975, page 525).

Formula (14) suggests the consideration of the following test statistic based on Hellinger distance in order to test the null hypothesis in (1):

$$2(m_1 + m_2)\left(HD\left(\boldsymbol{d}^*, \boldsymbol{f}_{\widehat{\theta}_H^1,\widehat{\theta}_H^2}\right) - HD\left(\boldsymbol{d}^*, \boldsymbol{f}_{\widetilde{\theta}_H,\widetilde{\theta}_H}\right)\right),$$

where $\widehat{\theta}_H^i$ was defined in (5) and $\widetilde{\theta}_H$ is defined by

$$\widetilde{\theta}_H = \arg\min_{\theta} HD\left(\boldsymbol{d}^*, \boldsymbol{f}_{\theta,\theta}\right).$$

Based on minimum penalized Hellinger distance estimators we can consider

$$2(m_1 + m_2)\left(PHD\left(\boldsymbol{d}^*, \boldsymbol{f}_{\widehat{\theta}_{PH}^1,\widehat{\theta}_{PH}^2}\right) - PHD\left(\boldsymbol{d}^*, \boldsymbol{f}_{\widetilde{\theta}_{PH},\widetilde{\theta}_{PH}}\right)\right),$$

where $\widehat{\theta}_{PH}^i$ was defined in (7). This statistic has been considered by Simpson in (1989).

Finally formula (15) suggest to us the test statistic

$$2(m_1 + m_2) HD\left(\boldsymbol{f}_{\widehat{\theta}_H^1,\widehat{\theta}_H^2}, \boldsymbol{f}_{\widetilde{\theta}_H,\widetilde{\theta}_H}\right).$$

In the following Theorem we present the asymptotic distribution of the six test statistics introduced here.

Theorem 1 *Let each of the sample sizes m_1 and m_2 go to infinity at a rate such that the limiting value of $m_1(m_1+m_2)^{-1}$ belongs to the open interval $(0,1)$. Then each of the following test statistics*

$$
\begin{aligned}
T_H &= 2\frac{m_1 m_2}{m_1+m_2} HD\left(\boldsymbol{f}_{\widehat{\theta}_H^1}, \boldsymbol{f}_{\widehat{\theta}_H^2}\right), \\
T_{PH} &= 2\frac{m_1 m_2}{m_1+m_2} HD\left(\boldsymbol{f}_{\widehat{\theta}_{PH}^1}, \boldsymbol{f}_{\widehat{\theta}_{PH}^2}\right), \\
S_H &= 2\left(m_1+m_2\right)\left(HD\left(\boldsymbol{d}^*, \boldsymbol{f}_{\widehat{\theta}_H^1, \widehat{\theta}_H^2}\right) - HD\left(\boldsymbol{d}^*, \boldsymbol{f}_{\widetilde{\theta}_H, \widetilde{\theta}_H}\right)\right), \\
S_{PH} &= 2\left(m_1+m_2\right)\left(PHD\left(\boldsymbol{d}^*, \boldsymbol{f}_{\widehat{\theta}_{PH}^1, \widehat{\theta}_{PH}^2}\right) - PHD\left(\boldsymbol{d}^*, \boldsymbol{f}_{\widetilde{\theta}_{PH}, \widetilde{\theta}_{PH}}\right)\right), \\
T_H^* &= 2\left(m_1+m_2\right) HD\left(\boldsymbol{f}_{\widehat{\theta}_H^1, \widehat{\theta}_H^2}, \boldsymbol{f}_{\widetilde{\theta}_H, \widetilde{\theta}_H}\right), \\
T_{PH}^* &= 2\left(m_1+m_2\right) HD\left(\boldsymbol{f}_{\widehat{\theta}_{PH}^1, \widehat{\theta}_{PH}^2}, \boldsymbol{f}_{\widetilde{\theta}_{PH}, \widetilde{\theta}_{PH}}\right),
\end{aligned}
$$

asymptotically has a chi-square distribution with one degree of freedom under the null hypothesis. Here $\widehat{\boldsymbol{\theta}}_H = \left(\widehat{\theta}_H^1, \widehat{\theta}_H^2\right)^T$ is the unrestricted minimizer of $HD\left(\boldsymbol{d}^*, \boldsymbol{f}_{\theta_1,\theta_2}\right)$ over $\theta_1, \theta_2 \in \Theta$, while $\widetilde{\boldsymbol{\theta}}_H = \left(\widetilde{\theta}_H, \widetilde{\theta}_H\right)^T$ represents the minimizer under the null hypothesis (1). The corresponding minimizers of the penalized Hellinger distance are denoted by $\widehat{\boldsymbol{\theta}}_{PH} = \left(\widehat{\theta}_{PH}^1, \widehat{\theta}_{PH}^2\right)^T$ and $\widetilde{\boldsymbol{\theta}}_{PH} = \left(\widetilde{\theta}_{PH}, \widetilde{\theta}_{PH}\right)^T$.

Proof. Assume that the null hypothesis given in equation (1) is true. Let $I(\theta_1)$ be the Fisher information, *i.e.* $I(\theta_1) = E\left[\frac{\partial}{\partial\theta_1}\log f_{\theta_1}(X)\right]^2$. A second order Taylor expansion of $HD\left(f_{\widehat{\theta}_H^1}, \boldsymbol{f}_{\widehat{\theta}_H^2}\right)$ (see page 443 in Pardo 2006) gives

$$
HD\left(\boldsymbol{f}_{\widehat{\theta}_H^1}, f_{\widehat{\theta}_H^2}\right) = \frac{1}{2}\left(\widehat{\theta}_H^1 - \widehat{\theta}_H^2\right)^2 I\left(\theta_1\right) + o_p\left(\left(\widehat{\theta}_H^1 - \theta_1\right)^2\right) + o_p\left(\left(\widehat{\theta}_H^1 - \theta_2\right)^2\right).
$$

By Simpson (1987) we know that

$$
\begin{aligned}
\sqrt{m_1}\left(\widehat{\theta}_H^1 - \theta_1\right) &\xrightarrow[m_1\to\infty]{L} N\left(0, I^{-1}(\theta_1)\right) \\
\sqrt{m_2}\left(\widehat{\theta}_H^2 - \theta_1\right) &\xrightarrow[m_2\to\infty]{L} N\left(0, I^{-1}(\theta_1)\right).
\end{aligned}
$$

Therefore,

$$
\sqrt{\frac{m_1 m_2}{m_1+m_2}}\left(\widehat{\theta}_H^1 - \widehat{\theta}_H^2\right) \xrightarrow[m_1,m_2\to\infty]{L} N\left(0, I^{-1}(\theta_1)\right)
$$

and

$$
T_H = 2\frac{m_1 m_2}{m_1+m_2} HD\left(\boldsymbol{f}_{\widehat{\theta}_H^1}, \boldsymbol{f}_{\widehat{\theta}_H^2}\right) \xrightarrow[m_1,m_2\to\infty]{L} \chi_1^2 .
$$

By Mandal et al. (2008) we have

$$
\begin{aligned}
\sqrt{m_1}\left(\widehat{\theta}_{PH}^1 - \theta_1\right) &\xrightarrow[m_1\to\infty]{L} N\left(0, I^{-1}(\theta_1)\right) \\
\sqrt{m_2}\left(\widehat{\theta}_{PH}^2 - \theta_1\right) &\xrightarrow[m_2\to\infty]{L} N\left(0, I^{-1}(\theta_1)\right).
\end{aligned}
$$

Therefore,

$$T_{PH} = 2\frac{m_1 m_2}{m_1+m_2} HD\left(\boldsymbol{f}_{\widehat{\theta}^1_{PH}}, \boldsymbol{f}_{\widehat{\theta}^2_{PH}}\right) \underset{m_1,m_2\to\infty}{\xrightarrow{L}} \chi^2_1 \ .$$

We denote $\boldsymbol{\theta}_1 = (\theta_1, \theta_1)^T$. If we consider the function $g(\theta_1, \theta_2) = \theta_1 - \theta_2$ the null hypothesis (1) can be written by $g(\theta_1, \theta_2) = 0$. We denote by $\boldsymbol{B} = \left(\frac{\partial g}{\partial \boldsymbol{\theta}}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}_1} = (1, -1)$. It is well known (see, for instance, Sen and Singer 1993) that

$$\sqrt{n}(\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}) = \mathcal{I}_F^{-1}(\boldsymbol{\theta}_1)\boldsymbol{B}^T\left(\boldsymbol{B}\mathcal{I}_F^{-1}(\boldsymbol{\theta}_1)\boldsymbol{B}^T\right)^{-1}\boldsymbol{B}\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_1) + o_p(1), \tag{16}$$

where $\widehat{\boldsymbol{\theta}} = \left(\widehat{\theta}_1, \widehat{\theta}_2\right)^T$ is the unrestricted maximum likelihood estimator, $\widetilde{\boldsymbol{\theta}} = \left(\widetilde{\theta}, \widetilde{\theta}\right)^T$ is maximum likelihood estimator restricted to the null hypothesis and

$$\mathcal{I}_F(\boldsymbol{\theta}_1) = \begin{pmatrix} \frac{m_1+m_2}{m_1} I(\theta_1) & 0 \\ 0 & \frac{m_1+m_2}{m_2} I(\theta_1) \end{pmatrix}.$$

The following results are routine extensions of the approach of Sarkar and Basu (1995):

$$S_H = \sqrt{m_1+m_2}\left(\widehat{\boldsymbol{\theta}}_H - \widetilde{\boldsymbol{\theta}}_H\right)^T \mathcal{I}_F(\boldsymbol{\theta}_1)\sqrt{m_1+m_2}\left(\widehat{\boldsymbol{\theta}}_H - \widetilde{\boldsymbol{\theta}}_H\right) + o_p(1),$$

$$\sqrt{m_1+m_2}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) = \sqrt{m_1+m_2}\left(\widehat{\boldsymbol{\theta}}_H - \boldsymbol{\theta}\right) + o_p(1),$$

and

$$\sqrt{m_1+m_2}\left(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) = \sqrt{m_1+m_2}\left(\widetilde{\boldsymbol{\theta}}_H - \boldsymbol{\theta}\right) + o_p(1).$$

Putting the above together with equation (16), we have

$$\sqrt{m_1+m_2}(\widetilde{\boldsymbol{\theta}}_H - \widehat{\boldsymbol{\theta}}_H) = \mathcal{I}_F^{-1}(\boldsymbol{\theta}_1)\boldsymbol{B}^T\left(\boldsymbol{B}\mathcal{I}_F^{-1}(\boldsymbol{\theta}_1)\boldsymbol{B}^T\right)^{-1}\boldsymbol{B}\sqrt{m_1+m_2}(\widehat{\boldsymbol{\theta}}_H - \boldsymbol{\theta}_1) + o_p(1).$$

Now

$$\begin{aligned} S_H &= \sqrt{m_1+m_2}\left(\widehat{\boldsymbol{\theta}}_H - \widetilde{\boldsymbol{\theta}}_H\right)^T \mathcal{I}_F(\boldsymbol{\theta}_1)\sqrt{m_1+m_2}\left(\widehat{\boldsymbol{\theta}}_H - \widetilde{\boldsymbol{\theta}}_H\right) + o_p(1) \\ &= \sqrt{m_1+m_2}(\widehat{\boldsymbol{\theta}}_H - \boldsymbol{\theta}_1)^T\boldsymbol{B}^T\left(\boldsymbol{B}\mathcal{I}_F^{-1}(\boldsymbol{\theta}_1)\boldsymbol{B}^T\right)^{-1}\boldsymbol{B}\mathcal{I}_F^{-1}(\boldsymbol{\theta}_1)\mathcal{I}_F(\boldsymbol{\theta}_1)\mathcal{I}_F^{-1}(\boldsymbol{\theta}_1)\boldsymbol{B}^T \\ &\times \left(\boldsymbol{B}\mathcal{I}_F^{-1}(\boldsymbol{\theta}_1)\boldsymbol{B}^T\right)^{-1}\boldsymbol{B}\sqrt{m_1+m_2}(\widehat{\boldsymbol{\theta}}_H - \boldsymbol{\theta}_1) + o_p(1) \\ &= \sqrt{m_1+m_2}(\widehat{\boldsymbol{\theta}}_H - \boldsymbol{\theta}_1)^T\boldsymbol{B}^T\left(\boldsymbol{B}\mathcal{I}_F^{-1}(\boldsymbol{\theta}_1)\boldsymbol{B}^T\right)^{-1}\boldsymbol{B}\sqrt{m_1+m_2}(\widehat{\boldsymbol{\theta}}_H - \boldsymbol{\theta}_1) + o_p(1) \ . \end{aligned}$$

Taking into account that

$$\sqrt{m_1+m_2}(\widehat{\boldsymbol{\theta}}_H - \boldsymbol{\theta}_1) \underset{m_1,m_2\to\infty}{\xrightarrow{L}} \mathcal{N}\left(\mathbf{0}, \mathcal{I}_F(\boldsymbol{\theta}_1)^{-1}\right)$$

and using Serfling (1980), Theorem 4.4.4, we have

$$S_H \underset{m_1,m_2\to\infty}{\xrightarrow{L}} \chi^2_1 \ .$$

The asymptotic distribution of S_{PH} follows the same steps that the proof of S_H taking into account that

$$\sqrt{m_1+m_2}(\widehat{\boldsymbol{\theta}}_H-\boldsymbol{\theta}_1)=\sqrt{m_1+m_2}(\widehat{\boldsymbol{\theta}}_{PH}-\boldsymbol{\theta}_1)+o_p(1)\ .$$

Finally the asymptotic distributions of T_H^* and T_{PH}^* follow in a similar way because a second Taylor expansion gives,

$$T_H^*=\sqrt{m_1+m_2}\left(\widehat{\boldsymbol{\theta}}_H-\widetilde{\boldsymbol{\theta}}_H\right)^T\mathcal{I}_F(\boldsymbol{\theta}_1)\sqrt{m_1+m_2}\left(\widehat{\boldsymbol{\theta}}_H-\widetilde{\boldsymbol{\theta}}_H\right)+o_p(1)\ ,$$

$$T_{PH}^*=\sqrt{m_1+m_2}\left(\widehat{\boldsymbol{\theta}}_{PH}-\widetilde{\boldsymbol{\theta}}_{PH}\right)^T\mathcal{I}_F(\boldsymbol{\theta}_1)\sqrt{m_1+m_2}\left(\widehat{\boldsymbol{\theta}}_{PH}-\widetilde{\boldsymbol{\theta}}_{PH}\right)+o_p(1)\ .$$

The techniques presented here for discrete models apply, in principle, to continuous models as well. However, this may require additional accessories, such as kernel density estimation, for the densities to be compatible when constructing the divergences. This makes the approach considerably more complicated. We hope to take up the issue of continuous models in a separate, future paper. ■

3 Simulation Study

To investigate the performance of the tests developed in the previous section, we present here the results of an extensive simulation study based on the Poisson distribution. In this connection it is useful to check whether the test statistics T_H, T_{PH}, T_H^* and T_{PH}^* have simplified expressions under the Poisson model. We note that a direct calculation gives

$$HD\left(\boldsymbol{f}_{\theta_1},\boldsymbol{f}_{\theta_2}\right)=4\left(1-\exp\left(-\frac{\theta_1+\theta_2}{2}\right)\exp\left(\sqrt{\theta_1\theta_2}\right)\right)$$

where $\boldsymbol{f}_{\theta_i}$, $i=1,2$ represents the probability vector of a Poisson random variable with parameter θ_i. Some straightforward algebra based on the above gives

$$T_H=8\frac{m_1m_2}{m_1+m_2}\left(1-\exp\left(-\frac{\widehat{\theta}_H^1+\widehat{\theta}_H^2}{2}\right)\exp\left(\sqrt{\widehat{\theta}_H^1\widehat{\theta}_H^2}\right)\right)\ ,$$

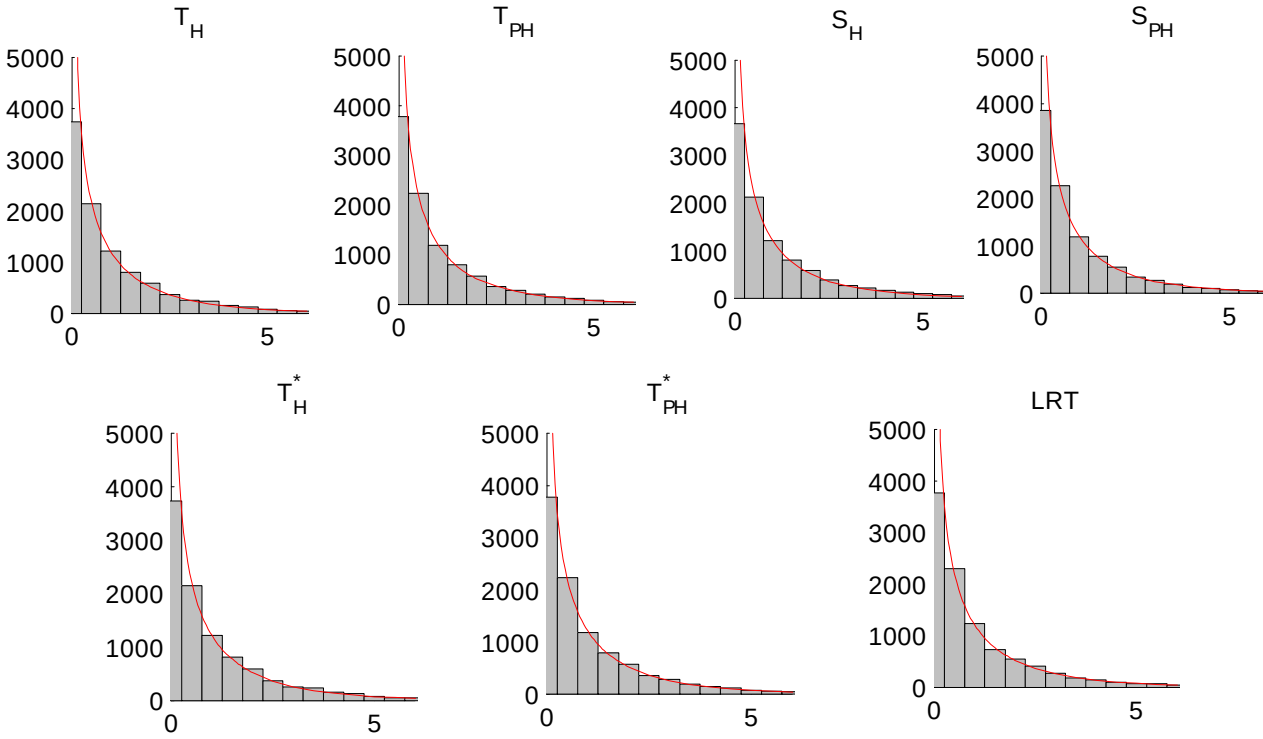
$$T_{PH}=8\frac{m_1m_2}{m_1+m_2}\left(1-\exp\left(-\frac{\widehat{\theta}_{PH}^1+\widehat{\theta}_{PH}^2}{2}\right)\exp\left(\sqrt{\widehat{\theta}_{PH}^1\widehat{\theta}_{PH}^2}\right)\right)\ ,$$

$$\begin{aligned}T_H^*=8\Bigg\{&m_1\left(1-\exp\left(-\frac{\widehat{\theta}_H^1+\widetilde{\theta}_H}{2}\right)\exp\left(\sqrt{\widehat{\theta}_H^1\widetilde{\theta}_H}\right)\right)\\&+m_2\left(1-\exp\left(-\frac{\widehat{\theta}_H^2+\widetilde{\theta}_H}{2}\right)\exp\left(\sqrt{\widehat{\theta}_H^2\widetilde{\theta}_H}\right)\right)\Bigg\}\ ,\end{aligned}$$

and

$$T^*_{PH} = 8\left\{m_1\left(1-\exp\left(-\frac{\widehat{\theta}^1_{PH}+\widetilde{\theta}_{PH}}{2}\right)\exp\left(\sqrt{\widehat{\theta}^1_{PH}\widetilde{\theta}_{PH}}\right)\right)\right.$$
$$\left.+m_2\left(1-\exp\left(-\frac{\widehat{\theta}^2_{PH}+\widetilde{\theta}_{PH}}{2}\right)\exp\left(\sqrt{\widehat{\theta}^2_{PH}\widetilde{\theta}_{PH}}\right)\right)\right\}.$$

Figure 1: Histograms of the six test statistics and the LRT where m_1=m_2=100, λ_1=λ_2 and ε=0



Data for the first sample are generated from the $(1-\epsilon)\boldsymbol{f}_{\lambda_1}+\epsilon\boldsymbol{f}_{\lambda_{1c}}$ mixture, where our target parameter is λ_1 and $100\epsilon\%$ data are coming from a contaminating population with parameter λ_{1c}. The second sample data are generated from the $\boldsymbol{f}_{\lambda_2}$ distribution. Assuming that the samples come from pure Poisson distributions with densities f_{θ_1} and f_{θ_2} respectively, λ_1, λ_2 unknown, we are interested in testing $H_0 : \theta_1 = \theta_2$ against the alternative that they are not equal. (We will use the "θ" symbol for the unknown values of the parameters involved in the hypotheses, and the "λ" symbol to denote the actual distributions from which the data have been generated.) All the

tests are performed at 5% level of significance at common values of m_1 and m_2 chosen as 20, 30, 50, 75, 100 and 150. All the results are based on 10000 replications. Assuming binomial rejection frequencies, the estimate of the standard deviation given a probability estimate $\hat{p}$ may be computed from $[\hat{p}(1-\hat{p})/10000]^{1/2}$. Hence the error will be no greater than $[0.5(1-0.5)/10000]^{1/2} = 0.005$.

In Figure 1, the histograms of the six test statistics and the likelihood ratio test statistic are plotted with the theoretical curve – the $\chi^2(1)$ density – overlaid. Here the sample sizes for the both populations are 100, $\epsilon = 0$, $\lambda_1 = \lambda_2$ and the common value of the parameter is 5. Although there are fine differences between the statistics, it is clear from the figure that all the test statistics approximate their asymptotic null distribution quite well.

Table 1: Comparison of the observed levels of the six tests and the LRT at nominal level 0.05 where $\lambda_1 = \lambda_2 = 5$ and $\epsilon = 0$.

$m_1 = m_2$	T_H	T_{PH}	S_H	S_{PH}	T_H^*	T_{PH}^*	LRT
20	0.0612	0.0532	0.0742	0.0350	0.0650	0.0557	0.0473
30	0.0655	0.0563	0.0786	0.0396	0.0685	0.0586	0.0520
50	0.0586	0.0527	0.0672	0.0427	0.0596	0.0539	0.0487
75	0.0603	0.0566	0.0666	0.0480	0.0613	0.0573	0.0509
100	0.0597	0.0555	0.0683	0.0503	0.0608	0.0565	0.0527
150	0.0582	0.0535	0.0627	0.0505	0.0588	0.0542	0.0508

Seven tables are constructed and presented here. In the first table, the observed levels (the proportion of statistics exceeding the chi-square critical value) of all the six tests at all the six sample sizes are presented, together with the corresponding levels of the likelihood ratio test. Here $\epsilon = 0$, so both samples represent pure Poisson data. While the observed levels of the likelihood ratio test match the nominal level very closely, most of the other tests are somewhat liberal (S_{PH} being the exception). Except for S_{PH} the observed probabilities of rejection are higher (quite substantially for some small samples) than the nominal level. The penalty seems to have a major effect, however, and the penalized statistics appear to generate observed levels which are significantly closer to the nominal levels compared to the ordinary ones. The statistic S_{PH} is very conservative for small samples, but for sample sizes of 75 or larger and S_{PH} appear to quite close to the likelihood ratio test.

In Table 2, the effect of the contaminant on the level is studied. Here $\lambda_1 = \lambda_2 = 5$, but $\epsilon = 0.05$, so that the first sample is generated by a mixture of two Poissons, with the contaminating smaller component having a mean of 15. It is clearly seen that all the six statistics based on the Hellinger and penalized Hellinger distances largely discount the contaminating component, but the effect of the latter on the likelihood ratio test is quite disastrous. Note that the observed levels become worse for all the methods as the sample size increases, since fixed amounts of contamination have greater impact in larger samples.

In Table 3 we look at the power of the tests when both samples represent pure Poisson data, but the first sample comes from a distribution with mean 5, and the second comes from a distribution with mean 6. The powers of all the statistics are quite competitive with those of the likelihood ratio test. Some of the tests have higher observed power than the likelihood ratio test, an artifact

Table 2: Comparison of the observed levels of the six tests and the LRT at nominal level 0.05 where $\lambda_1 = \lambda_2 = 5$, $\epsilon = 0.05$ and $\lambda_{1c} = 15$.

$m_1 = m_2$	T_H	T_{PH}	S_H	S_{PH}	T_H^*	T_{PH}^*	LRT
20	0.0839	0.0700	0.0911	0.0429	0.0905	0.0747	0.1476
30	0.0802	0.0715	0.0863	0.0484	0.0842	0.0749	0.1797
50	0.0853	0.0784	0.0891	0.0567	0.0871	0.0802	0.2286
75	0.0913	0.0867	0.0919	0.0698	0.0934	0.0884	0.2982
100	0.1006	0.0940	0.0999	0.0787	0.1026	0.0952	0.3673
150	0.1138	0.1019	0.1105	0.0874	0.1154	0.1033	0.4688

Table 3: Comparison of the observed powers of the six tests and the LRT at nominal level 0.05 where $\lambda_1 = 5$, $\lambda_2 = 6$ and $\epsilon = 0$.

$m_1 = m_2$	T_H	T_{PH}	S_H	S_{PH}	T_H^*	T_{PH}^*	LRT
20	0.2853	0.2686	0.3124	0.2088	0.2934	0.2773	0.2706
30	0.3989	0.3927	0.4301	0.3363	0.4056	0.3994	0.3920
50	0.5695	0.5696	0.5931	0.5325	0.5746	0.5736	0.5733
75	0.7364	0.7419	0.7524	0.7209	0.7394	0.7435	0.7450
100	0.8487	0.8524	0.8595	0.8408	0.8501	0.8532	0.8559
150	0.9590	0.9615	0.9624	0.9586	0.9592	0.9615	0.9616

of their higher observed levels at the null under the true model.

In Table 4, we study the power of the tests under contamination. The values of λ_1 and λ_2 are the same as in Table 3, but the first sample comes from a mixture of Poissons, with the contaminating component having a mean of 15 and a weight of 5%. This leads to a severe loss in the power in the likelihood ratio test, but all the other tests hold their levels pretty well.

The empirical critical values of the tests at 5% level of significance are given in Table 5. The values of λ_1 and λ_2 are taken to be 5 and ϵ equal zero. The theoretical chi-square critical value in this case is 3.8415. The empirical critical values of all the tests (except S_{PH}) are higher than their theoretical critical values, although the degree of inflation is much smaller for the penalized distances. The high power of some of the statistics in Table actually due to the true critical values being much higher than that of the chi-square density with 1 degrees of freedom. All the empirical critical values approach the theoretical value as the sample size increases; this happens much faster for the penalized distances.

In Table 6 and 7 we have used the same data as was used in Table 3 and 4 respectively. But now the empirical critical values (as presented in Table 5) have been used to determine the observed power instead of the theoretical critical values. When $\lambda_1 = 5$, $\lambda_2 = 6$ and $\epsilon = 0$ we can see that

Table 4: Comparison of the observed powers of the six tests and the LRT at nominal level 0.05 where $\lambda_1 = 5,\ \lambda_2 = 6$, $\epsilon = 0.05$ and $\lambda_{1c} = 15$.

$m_1 = m_2$	T_H	T_{PH}	S_H	S_{PH}	T_H^*	T_{PH}^*	LRT
20	0.2611	0.2391	0.2773	0.1772	0.2714	0.2476	0.1539
30	0.3438	0.3260	0.3588	0.2646	0.3516	0.3337	0.1831
50	0.4724	0.4632	0.4804	0.4085	0.4771	0.4683	0.2325
75	0.5952	0.5960	0.5972	0.5556	0.5981	0.5992	0.2883
100	0.7162	0.7312	0.7126	0.7008	0.7193	0.7331	0.3566
150	0.8556	0.8682	0.8496	0.8522	0.8570	0.8697	0.4561

Table 5: Comparison of empirical critical values of the six tests and the LRT at nominal level 0.05 where $\lambda_1 = \lambda_2 = 5$, and $\epsilon = 0$ (here the chi-square critical value is 3.8415).

$m_1 = m_2$	T_H	T_{PH}	S_H	S_{PH}	T_H^*	T_{PH}^*	LRT
20	4.4331	3.9548	4.8967	3.3039	4.5926	4.0533	3.7896
30	4.2640	3.9608	4.6493	3.4807	4.3635	4.0349	3.7185
50	4.2050	4.0283	4.5314	3.6427	4.2470	4.0652	3.8910
75	4.1069	3.9385	4.3513	3.7271	4.1474	3.9683	3.8795
100	4.0501	3.9225	4.2609	3.7786	4.0795	3.9484	3.8265
150	3.9963	3.8587	4.1468	3.7674	4.0136	3.8699	3.8159

the powers of all the tests are very close to those of the likelihood ratio test and for $\lambda_1 = 5$, $\lambda_2 = 6$, $\epsilon = 0.05$ and $\lambda_{1c} = 15$ all the tests perform much better than the likelihood ratio test.

On the whole, all the six proposed tests appear to do quite well in terms of their ability to hold their levels and powers under contamination. Considering the entire evidence, T_{PH} and T_{PH}^* appear to be the most desirable statistics in terms of their closeness to the likelihood ratio test, attained power, and robustness against contaminations.

It may take more extensive studies to determine whether the amount of penalty applied here ($h = 1$) is optimal in terms of the desirable properties, or whether another choice can do better. However, the choice of $h = 1$ makes the method identical to the likelihood ratio test in terms of their treatment of the empty cells. In addition, a choice of $h = 1/2$ (results not reproduced here for brevity) does not appear to produce any appreciable change in the results produced here. Thus we believe the choice $h = 1$ is a sensible choice for our purpose.

Acknowledgements: This work was partially supported by Grant MTM2009-10072. The authors would like to thank the referees for critically reading the paper and making useful suggestions.

Table 6: Comparison of the observed powers of the six tests and the LRT using the empirical critical values at nominal level 0.05 where $\lambda_1 = 5,\ \lambda_2 = 6$, $\epsilon = 0$.

$m_1 = m_2$	T_H	T_{PH}	S_H	S_{PH}	T_H^*	T_{PH}^*	LRT
20	0.2404	0.2587	0.2344	0.2584	0.2401	0.2596	0.2745
30	0.3580	0.3812	0.3567	0.3783	0.3569	0.3811	0.4032
50	0.5353	0.5509	0.5327	0.5562	0.5367	0.5513	0.5688
75	0.7140	0.7331	0.7124	0.7303	0.7137	0.7333	0.7411
100	0.8359	0.8471	0.8351	0.8452	0.8359	0.8466	0.8571
150	0.9559	0.9613	0.9559	0.9604	0.9560	0.9613	0.9624

Table 7: Comparison of the observed powers of the six tests and the LRT using the empirical critical values at nominal level 0.05 where $\lambda_1 = 5,\ \lambda_2 = 6$, $\epsilon = 0.05$ and $\lambda_{1c} = 15$.

$m_1 = m_2$	T_H	T_{PH}	S_H	S_{PH}	T_H^*	T_{PH}^*	LRT
20	0.2161	0.2311	0.2028	0.2192	0.2170	0.2325	0.1565
30	0.3108	0.3158	0.2953	0.2993	0.3105	0.3166	0.1894
50	0.4419	0.4440	0.4234	0.4292	0.4439	0.4462	0.2300
75	0.5740	0.5876	0.5545	0.5671	0.5747	0.5884	0.2856
100	0.6998	0.7240	0.6829	0.7060	0.7011	0.7247	0.3574
150	0.8480	0.8673	0.8317	0.8563	0.8488	0.8678	0.4581

References

[1] Basu, A. and Basu, S. (1998). Penalized minimum disparity methods for multinomial models. *Statistica Sinica*, **8**, 841–860.

[2] Basu, A., Harris, I. R. and Basu, A. (1996). Tests of hypothesis in discrete models based on the penalized Hellinger distance. *Statistics and Probability Letters*, **27**, 367–373.

[3] Basu, A., Harris, I.R., Basu, S., (1997). *Minimum distance estimation: the approach using density based distances*. In: Maddala, G.S., Rao, C.R. (Eds.), Handbook of Statistics, Vol. **15**, Robust Inference. Elsevier Science, New York, NY, pp. 21–48.

[4] Beran, R. (1977). Minimum Hellinger distance estimates for parametric models. *Annals of Statistics,* **5,** 445–463.

[5] Bishop, M. M., Fienberg, S. E. and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice.* MIT Press, Cambridge, Mass.

[6] Harris, I.R. and A. Basu (1994). Hellinger distance as a penalized log likelihood, *Comm. Statist. Simul. Comput.* **23**, 1097–1113.

[7] Kullback, S. (1985). Kullback information, in *Encyclopedia of Statistical Sciences,* Volume 4 (S. Kotz and N. L. Johnson, Eds.), 421–425. John Wiley & Sons, New York.

[8] Kupperman, M. (1957). Further application to information theory to multivariate analysis and statistical inference. *Ph. D. Dissertation,* George Washington University.

[9] Kupperman, M. (1958). Probability hypothesis and information statistics in sampling exponential class populations. *Annals of Mathematical Statistics,* **29,** 571–574.

[10] Lindsay, B. G. (1994). Efficiency versus robustness: the case for minimum Hellinger distance and related methods. *Annals of Statistics,* **22,** 1081–1114.

[11] Mandal, A., Basu, A. and Pardo, L. (2008). Minimum Hellinger Distance Inference and the Empty Cell Penalty: Asymptotic Results. Technical Report No. ASD/2008/3, Indian Statistical Institute.

[12] Pardo, L. (2006). *Statistical Inference Based on Divergence Measures*. Chapman & Hall/CRC.

[13] Salicrú, M., Morales, D., Menéndez, M. L. and Pardo, L. (1994). On the applications of divergence type measures in testing statistical hypotheses. *Journal of Multivariate Analysis,* **51**, 372–391.

[14] Sarkar, S., Basu, A. (1995). On disparity based robust tests for two discrete populations. *Sankhya,* **57**, 353–364.

[15] Sen, P. K. and Singer, J. M. (1993). *Large Sample Methods in Statistics.* Chapman & Hall.

[16] Serfling, R. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

[17] Simpson, D. G. (1987). Minimum Hellinger distance estimation for analysis of count data. *Journal of the American Statistical Association,* **82,** 802–807.

[18] Simpson, D. G. (1989). Hellinger deviance tests: Efficiency, breakdown points and examples. *Journal of the American Statistical Association,* **84,** 107–113.