Sankhya : The Indian Journal of Statistics
1995, Volume 57, Series B, Pt. 3, pp. 353-364

ON DISPARITY BASED ROBUST TESTS FOR TWO DISCRETE POPULATIONS*

By SAHADEB SARKAR
Oklahoma State University
and
AYANENDRANATH BASU[1]
University of Texas at Austin

SUMMARY. For discrete two sample problems disparity tests based on minimum disparity estimation (Lindsay 1994) are considered. The likelihood ratio test can be obtained as a disparity test by using the likelihood disparity. It is shown that the asymptotic distribution of the disparity tests under composite null hypotheses is chi-square. In general, several disparity tests are more robust against outliers than the likelihood ratio test. A Monte Carlo study illustrates these points in Poisson populations for the Hellinger distance test.

1. INTRODUCTION

Beran (1977) showed that one can simultaneously obtain asymptotic efficiency and robustness by using the minimum Hellinger distance estimator. As robust M-estimators typically lose some efficiency at the model to achieve their robustness, Beran's method was an improvement over them. Several authors have continued this line of research including Tamura and Boos (1986), Simpson (1987) and Lindsay (1994). Tests of hypotheses based on the Hellinger and related distances were considered by Simpson (1989), Basu (1993), Lindsay (1994) and Basu and Sarkar (1994a). These tests are asymptotically equivalent to the likelihood ratio test (Neyman and Pearson 1928, Wilks 1938) at the model and

Paper received. May 1994; revised January 1995.

AMS (1991) *subject classification.* Primary 62F03, 62F35; Secondary 62F05.

Key words and phrases. Blended weight Hellinger distance, disparity tests, Hellinger distance, likelihood ratio test, minimum disparity estimation, Poisson distribution, outliers, robustness.

* The research of the first author was supported by a Grant from the College of Arts and Sciences at Oklahoma State University and the research of the second author was supported by the URI Summer Research Grant, University of Texas at Austin. The authors wish to thank Professor Bruce G. Lindsay for kindly suggesting this problem. The authors also wish to thank an anonymous referee and the Co-Editor for many helpful suggestionss.

[1] Presently at the Indian Statistical Institute, Calcutta.

at contiguous alternatives but have far better robustness properties than the latter when outliers are present in the data.

In this paper we extend these ideas to develop testing procedures when two discrete populations are involved. The results easily extend to three or more populations. We are currently studying the corresponding problems for the continuous models, theory for which is somewhat more complex as it involves kernel density estimation. In Section 2 we briefly discuss minimum disparity estimation and disparity tests studied in single population cases. Section 3 describes the null hypothesis and introduces the disparity tests in the two populations case, and establishes the limiting chi-square distribution of the tests under the null hypothesis. In Section 4 we provide some simulation results for Poisson populations showing that the disparity test based on the Hellinger distance is far more robust against outlying observations than the likelihood ratio test. We present some concluding remarks in Section 5.

2. Minimum disparity estimation and disparity tests in one population

Let $\{m_\beta(x)\}$ represent a family of probability mass functions having a countable support and indexed by $\beta = (\beta^1, \ldots, \beta^k)'$. Given a sample of size n $\{X_1, X_2, \ldots, X_n\}$ from this distribution, let $d(x)$ represent the observed proportion of X_i's taking the value x. Let $\delta(x) = [d(x) - m_\beta(x)]/m_\beta(x)$ represent the "Pearson" residual at the value x. Let G be a convex function with $G(0) = 0$. Then, the nonnegative "disparity" measure ρ corresponding to G is defined as

$$\rho(d, m_\beta) = \Sigma_x G(\delta(x)) m_\beta(x). \qquad \ldots (2.1)$$

When there is no scope for confusion, we will write $\rho(d, m_\beta)$ simply as $\rho(\beta)$. A value of β that minimizes (2.1) is called a minimum disparity estimate. When $G(\delta) = (\delta + 1)log(\delta + 1)$, the disparity

$$LD(d, m_\beta) = \Sigma_x d(x)[log(d(x)) - log(m_\beta(x))] \qquad \ldots (2.2)$$

is called the likelihood disparity, and its minimizer is the maximum likelihood estimator (MLE) of β, because the likelihood disparity is the negative of the log likelihood divided by n plus a factor free from parameters so that maximizing the likelihood is equivalent to minimizing the likelihood disparity. Note that (2.2) is a form of Kullback-Leibler divergence. On the other hand, $G(\delta) = [(\delta + 1)^{1/2} - 1]^2$ generates the squared Hellinger distance. Other examples of disparities include the Pearson's chi-square, Neyman's chi-square, the power divergence family (Cressie and Read 1984), the blended weight Hellinger distance family (Lindsay 1994; Basu and Sarkar 1994b) and the negative expoential disparity (Basu and Sarkar 1994c; Lindsay 1994).

Let ∇ represent the vector gradient with respect to β. Under differentiability of the model, the minimum disparity estimating equations have the form

$$-\nabla\rho = \Sigma_x A(\delta(x))\nabla m_\beta(x) = 0, \qquad \ldots(2.3)$$

where $A(\delta) = (\delta+1)[G^{(1)}(\delta)] - G(\delta)$, and $G^{(1)}(\delta)$ denotes the first derivatives of $G(\delta)$. The function $A(\delta)$ is an increasing function on $[-1,\infty)$ and can be redefined, without changing the estimating properties of the disparity, so that $A(0) = 0$ and $A^{(1)}(0) = 1$, where $A^{(1)}(\delta)$ denotes the first derivative of $A(\delta)$. This function $A(\delta)$ is called the residual adjustment function of the disparity and plays a leading role in determining the theoretical properties of the estimators. For the likelihood disparity (LD) the residual adjustment function is linear with $A(\delta) = \delta$. However, the residual adjustment function of a disparity like the Hellinger distance (for which $A(\delta) = 2[(\delta+1)^{1/2} - 1]$, after the above standardization) can significantly downweight the effect of a large Pearson residual. In this sense the residual adjustment function has an interpretation similar to the ψ-function in M-estimation. A minimum disparity estimator is more robust than the MLE if its residual adjustment function downweights an x value with a large positive $\delta(x)$ relative to the residual adjustment function of the likelihood disparity. On the other hand, negative Pearson residuals represent sparse data where one would expect more observations under the model. The x-values where this occurs can be called "Pearson inliers". A disparity like the negative exponential disparity (Basu and Sarkar 1994c; Lindsay 1994) can downweight Pearson inliers relative to the MLE.

The curvature parameter A_2 of a disparity is the second derivative of its residual adjustment function evaluated at zero. This parameter plays an important role in determining the trade-off between robustness and efficiency. Disparities with large negative values of the curvature parameter generate more robust estimators, whereas $A_2 = 0$ implies second order efficiency in the sense of Rao (1961). Similarly, disparity tests with large negative values of A_2 provide stability to the level and power of the tests under contamination, whereas $A_2 = 0$ usually leads to more powerful tests. For the likelihood disparity $A_2 = 0$, and for the Hellinger distance $A_2 = -1/2$.

Let $T = T_\rho$ represent the minimum disparity functional obtained by minimizing the disparity measure ρ with respect to β. Consider testing the simple null hypothesis $\beta = \beta_*$ against some suitable alternative. For this the disparity test statistic corresponding to the disparity measure ρ is defined as $D_\rho = -2n[\rho(T_\rho) - \rho(\beta_*)]$, and under null hypothesis D_ρ has an asymptotic χ^2 distribution with degrees of freedom equal to the dimensions of β (Lindsay 1994). When ρ equals the likelihood disparity, D_ρ equals the negative of twice log likelihood ratio.

The two population case of the hypothesis testing problem using the minimum disparity approach can be solved by generalizing the method of Lindsay (1994) appropriately. In order to motivate the extension of the one sample

case to the two sample situation, we now present in some detail the derivation of the asymptotic distribution of the disparity test statistic in the one sample case under a composite null hypothesis. Consider the null hypothesis $H_0 : \beta \in \mathcal{B}_0$, where $\mathcal{B}_0$ is a subset of the parameter space $\mathcal{B}$. Let r be the number of independent restrictions imposed by the null hypothesis $H_0 : \beta \in \mathcal{B}_0$. We assume that the specification of $\mathcal{B}_0$ can be expressed as a transformation $\beta^i = g_i(\nu^1, \ldots, \nu^{k-r}), i = 1, \ldots, k$, where $\nu = (\nu^1, \ldots, \nu^{k-r})'$ ranges through an open subset of $\Re^{k-r}$. We also assume that g_i possesses continuous first partial derivatives.

Let $u_\beta(x) = \nabla \, log m_\beta(x)$, the maximum likelihood score function. Let ∇_i represent gradient with respect to β^i, the i-th component of β, and $u_i(x) = u_i(x, \beta) = \nabla_i log m_\beta(x)$. Similarly, ∇_{ij} and u_{ij} will represent the second partial derivatives with respect to β^i and β^j and u_{ijk} will represent the third partial derivatives. We let β_0 denote the true parameter value, and let $\delta_0(x)$ denote $\delta(x)$ when $m_\beta = m_{\beta_0}$. We assume the following regularity conditions (Lindsay 1994).

Assumption I. $Var(u_{\beta_0}(X))$ is finite

Assumption II. The residual adjustment function $A(\delta)$ is such that $A^{(1)}(\delta)$ and $[A^{(2)}(\delta)](1 + \delta)$ are bounded by C and D, say, on $[-1, \infty)$.

Assumption III. $\Sigma_x[m_{\beta_0}(x)]^{1/2} \mid u_i(x; \beta_0) \mid, \Sigma_x[m_{\beta_0}(x)]^{1/2} \mid u_{ij}(x; \beta_0) \mid$ and $\Sigma_x[m_{\beta_0}(x)]^{1/2} \mid u_i(x; \beta_0) \| u_j(x; \beta_0) \mid$ are all finite for all i, j.

Assumption IV. β_0 is the unique minimizer of $\rho(m_{\beta_0}, m_\beta)$ with respect to β.

Assumption V. The conditions on pages 409 and 429 of Lehmann (1983) are satisfied and there exists $M_{ijk}(x)$, $M_{ij,k}(x)$, and $M_{i,j,k}(x)$ that dominate in absolute value $u_{ijk}(x; \beta), u_{ij}(x; \beta)u_k(x; \beta)$ and $u_i(x; \beta)u_j(x; \beta)u_k(x; \beta)$ respectively for all β in a neighborhood of β_0 and that are uniformly bounded in expectation E_β for all β in some, possibly smaller, open neighborhood of β_0.

Let $D_\rho = -2n[\rho(\widehat{\beta}_n) - \rho(\beta_n^*)]$ be the minimum disparity test statistic where $\widehat{\beta}_n, \beta_n^*$ are the minimum disparity estimates of β_0 without any restriction and under the null hypothesis repsectively. Let $\widehat{\beta}_{nL}$ and β_{nL}^* be the MLE's of β_0 without any restriction and under the null hypothesis respectively. We show that when the null hypothesis is true the limiting distribution of D_ρ is $\chi^2(r)$. First we present the following :

Lemma 2.1. *(i)* $-n^{1/2}\nabla\rho(\beta_0) = n^{-1/2}\Sigma_{i=1}^n u_{\beta_0}(X_i) + o_p(1)$. *(ii)* $\nabla_{ij}\rho(\beta_0) = I_{\beta_0}(i, j) + o_p(1)$, *where* $I_{\beta_0}(i, j)$ *represents the* (i, j)*-th element of* $\mathbf{I}_{\beta_0}$, *the Fisher information matrix. (iii) The minimum disparity estimator* $\widehat{\beta}_n$ *is a consistent estimator of* β_0, *and* β_n^* *is a consistent estimator of* β_0 *when the null hypothesis is true. (iv)* $n^{1/2}(\widehat{\beta}_{nL} - \beta_{nL}^*) = n^{1/2}(\widehat{\beta}_n - \beta_n^*) + o_p(1)$.

Proof of (i). Since $-n^{1/2}\nabla\rho(\beta_0) = n^{-1/2}\Sigma_{i=1}^n u_{\beta_0}(X_i) + n^{1/2}\Sigma[A(\delta_0(x)) - \delta_0(x)]\nabla m_{\beta_0}(x)$, it suffices to prove that

$$E \mid n^{1/2}\Sigma[A(\delta_0(x)) - \delta_0(x)]\nabla m_{\beta_0}(x) \mid \to 0. \qquad \ldots(2.4)$$

Let

$$Y_n(x) = n^{1/2}[(\frac{d(x)}{m_{\beta_0}(x)})^{1/2} - 1]^2.$$

It can be shown that (see Lemma 24, Lemma 25 and Lemma 23 respectively of Lindsay (1994))

$$E[Y_n(x)] \leq E(\mid \delta_0(x) \mid)n^{1/2} \leq [m_{\beta_0}(x)]^{-1/2}, \qquad \ldots(2.5)$$

$$\lim_{n\to\infty} E[Y_n(x)] = 0, \qquad \ldots(2.6)$$

and

$$\mid A(\delta_0(x)) - \delta_0(x) \mid \leq B[(\frac{d(x)}{m_{\beta_0}(x)})^{1/2} - 1]^2 \qquad \ldots(2.7)$$

for some positive constant B. By (2.7) $E \mid n^{1/2}\Sigma[A(\delta_0(x)) - \delta_0(x)]\nabla m_{\beta_0}(x) \mid$ is bounded by $BE[\Sigma Y_n(x) \mid \nabla m_{\beta_0}(x) \mid]$. Then (2.4) follows from (2.5), (2.6) and Assumption III.

Proof of (ii). Note that $\nabla_{ij}\rho(\beta_0) = \Sigma A^{(1)}(\delta_0(x))(1+\delta_0(x))u_i(x)u_j(x)m_{\beta_0}(x)$ $- \Sigma A(\delta_0(x))\nabla_{ij}m_{\beta_0}(x)$. The first term

$$\mid \Sigma A^{(1)}(\delta_0(x))(1 + \delta_0(x))u_i(x)u_j(x)m_{\beta_0}(x) - \Sigma u_i(x)u_j(x)m_{\beta_0}(x) \mid$$

$$\leq (C + D)\Sigma \mid \delta_0(x)u_i(x)u_j(x)m_{\beta_0}(x) \mid \qquad \ldots(2.8)$$

by Assumption II. Then, by (2.5) and Assumption III, the expectation of the right hand side of (2.8) goes to 0. It follows by Markov's inequality that

$$\mid \Sigma A^{(1)}(\delta_0(x))(1 + \delta_0(x))u_i(x)u_j(x)m_{\beta_0}(x) - I_{\beta_0}(i,j) \mid = o_p(1).$$

Similarly, using the first order Taylor series expansion of $A(\delta)$ around $\delta = 0$ it can be shown that $\Sigma A(\delta_0(x))\nabla_{ij}m_{\beta_0}(x)$ converges in probability to 0. This completes the proof of (ii).

Proof of (iii). The proof of consistency of $\widehat{\beta}_n$ follows from arguments similar to those of Lehmann (1983, pp. 430 - 432) used to prove consistency of the MLE, with $-\rho(\beta)$ in place of n^{-1} log likelihood function, and from using (i), (ii) and Assumptions (IV) and (V).

Suppose that the null hypothesis is true. By assumption $\mathcal{B}_0$ can be expressed as a transformation $\beta^i = g_i(\nu^1, \ldots, \nu^{k-r}), i = 1, \ldots, k$. Let

$$\mathbf{D}_\nu = \left[\frac{\partial g_i}{\partial \nu_j}\right]_{k\times(k-r)}.$$

Let $\widehat{\nu}_n$ be the minimum disparity estimator of the true parameter ν_0, defined by $\beta_0 = g(\nu_0)$ under the ν-formulation of the model. Then, it follows from the arguments for consistency of $\widehat{\beta}_n$ that $\widehat{\nu}_n$ is a consistent estimator of ν_0 and $\beta_n^* = g(\widehat{\nu}_n) = (g_1(\widehat{\nu}_n), \ldots, g_k(\widehat{\nu}_n))'$ is a consistent estimator of β_0.

Proof of (iv). It follows from (i), (ii), (iii) and Assumption V that

$$n^{1/2}(\widehat{\beta}_n - \beta_0) = (\mathbf{I}_{\beta_0})^{-1} n^{1/2}[n^{-1}\Sigma_i^n u_{\beta_0}(X_i)] + o_p(1)$$

for all minimum disparity estimators including the MLE. Thus, $n^{1/2}(\widehat{\beta}_{nL} - \widehat{\beta}_n) = o_p(1)$, and $n^{1/2}(\beta_{nL}^* - \beta_n^*) = o_p(1)$ in particular. ▱

From the proof above it also follows that $n^{1/2}(\widehat{\beta}_n - \beta_0) \xrightarrow{L} N(0, I_{\beta_0}^{-1})$, and $n^{1/2}(\beta_n^* - \beta_0) \xrightarrow{L} \mathbf{N}(\mathbf{0}, \mathbf{D}_{\nu 0}\mathbf{J}_{\nu_0}^{-1}\mathbf{D}_{\nu 0}')$ under the null hypothesis, where $\mathbf{J}_{\nu 0}$ is the information matrix under the ν-formulation and $\xrightarrow{L}$ denotes convergence in distribution as $n \to \infty$. Therefore, $n^{1/2}(\widehat{\beta}_n - \beta_n^*) = O_p(1)$. Serfling (1980, Theorem 4.4.4) shows that $n(\widehat{\beta}_{nL} - \beta_{nL}^*)' \mathbf{I}_{\beta 0}(\widehat{\beta}_{nL} - \beta_{nL}^*) \xrightarrow{L} \chi^2(r)$. Now using Lemma 2.1 - (ii) and a Taylor series expansion of $2n\rho(\beta_n^*)$ around $\widehat{\beta}_n$, we have

$$-2n[\rho(\widehat{\beta}_n) - \rho(\beta_n^*)] = n(\widehat{\beta}_n - \beta_n^*)'\mathbf{I}_{\beta_0}(\widehat{\beta}_n - \beta_n^*) + n(\widehat{\beta}_n - \beta^*)'[\partial^2\rho(\tilde{\beta}_n)/\partial\beta\partial\beta' - \mathbf{I}_{\beta_0}](\widehat{\beta} - \beta^*)$$

where $\partial^2\rho(\tilde{\beta}_n)/\partial\beta\partial\beta'$ is the matrix of second partial derivatives evaluated at $\tilde{\beta}_n$ a point between $\widehat{\beta}_n$ and β_n^*. Therefore, using $n^{1/2}(\widehat{\beta}_n - \beta_n^*) = O_p(1)$ and Lemma 2.1 - *(ii), (iv)* we have $-2n[\rho(\widehat{\beta}_n) - \rho(\beta_n^*)] \xrightarrow{L} \chi^2(r)$.

3. Disparity tests in two populations

Let $(X_1, \ldots, X_{n1})$ and $(Y_1, \ldots, Y_{n2})$ be random samples from populations having probability mass functions $m_{\theta_1}(x)$ and $m_{\theta_2}(y)$ respectively, where θ_1 and θ_2 are $k_1 \times 1$ and $k_2 \times 1$ parameter vectors respectively. Let $n = n_1 + n_2$ and let $\theta = (\theta_1', \theta_2')' = (\theta^1, \ldots, \theta^k)'$ be the combined vector of parameters of the two populations. Note that $k \leq k_1 + k_2$ since θ_1 and θ_2 may have some common parameters. We assume that the two random samples are independent, and that n increases to infinity with $n_1^{-1}n_2 \to c$, $0 < c < \infty$, i.e., neither samples size asymptotically dominates the other. We specify a null hypothesis H_0 to be tested as $H_0 : \theta \in \Theta_0$, where Θ_0 is a subset of the parameter space $\Theta \subset \Re^k$ and Θ_0 is determined by a set of $r \leq k$ restrictions given by equations $R_i(\theta) = 0, 1 \leq i \leq r$. For example, for $k = 2$, we might have $H_0 : \theta \in \Theta_0 = \{\theta = (\theta_1, \theta_2)' : \theta_1 = \theta_2\}$. In this case, $r = 1$ and the function $R_1(\theta)$ may be defined as $\theta_1 - \theta_2$.

Let $\mathbf{I}_\theta$ be the $k \times k$ matrix whose (i, j)-th element is given by

$$\frac{1}{1+c}E\left[\frac{\partial^2}{\partial\theta^i\partial\theta^j}logm_{\theta_1}(X)\right] + \frac{c}{1+c}E\left[\frac{\partial^2}{\partial\theta^i\partial\theta^j}logm_{\theta_2}(Y)\right]. \qquad \ldots(3.1)$$

Note that in case θ_1 and θ_2 have no common parameters, the matrix $\mathbf{I}_\theta$ is a block diagonal matrix with two blocks which are equal to $\frac{1}{1+c}I_{\theta_1}$ and $\frac{c}{1+c}I_{\theta_2}$, where $\mathbf{I}_{\theta_i}$ is the Fisher information matrix corresponding to $m_{\theta_i}(x), i = 1, 2$.

In the sequel, let θ_0 denote the true parameter value. Let $d_i(z), i = 1, 2$, be the proportion of sample observations having the value z in the i-th sample and let $\rho(\theta_i) = \rho(d_i, m_{\theta_i})$ denote the corresponding disparity. Let the overall disparity $\rho_O(\theta)$ for the two samples taken together be defined by

$$\rho_O(\theta) = n^{-1}[n_1\rho(\theta_1) + n_2\rho(\theta_2)], \qquad \ldots(3.2)$$

and the disparity test statistic is defined by

$$-2n[\rho_O(\widehat{\theta}_n) - \rho_O(\theta^*_n)] \qquad \ldots(3.3)$$

where $\widehat{\theta}_n = (\widehat{\theta}^1_n, \widehat{\theta}^2_n, \ldots, \widehat{\theta}^k_n)'$ is a vector at which ρ_O is minimized over Θ and similarly $\theta^*_n = (\theta^{*1}_n, \theta^{*2}_n, \ldots, \theta^{*k}_n)'$ is a vector at which ρ_O is minimized over Θ_0. Let $\widehat{\theta}^x_n, \widehat{\theta}^y_n$ denote the estimates of θ_1 and θ_2 components respectively.

Note that the overall disparity is defined as a weighted average of the disparities for the individual samples, instead of the ordinary average with equal weights. There are two reasons. First, this takes into account different sample sizes available for the two populations. Second, with this definition of overall disparity the likelihood disparity test coincides with the likelihood ratio test. This makes it possible to investigate other disparity tests in relation to the likelihood ratio test by direct comparison. When ρ is the likelihood disparity, let $LD_O(\theta)$ denote the overall disparity (3.2), and let $\widehat{\theta}_{nL} = (\widehat{\theta}^1_{nL}, \ldots, \widehat{\theta}^k_{nL})'$ denote a vector that minimizes $LD_O(\theta)$ over Θ with $\widehat{\theta}^x_{nL}$ and $\widehat{\theta}^y_{nL}$ representing the estimates of θ_1 and θ_2 components respectively. Similarly, let $\theta^*_{nL} = (\theta^{*1}_{nL}, \ldots, \theta^{*k}_{nL})'$ denote a vector that minimizes $LD_O(\theta)$ over Θ_0 with θ^{*x}_{nL} and θ^{*y}_{nL} representing the estimates of θ_1 and θ_2 components respectively. Note that

$$\begin{aligned}
&LD_O(\widehat{\theta}_{nL}) - LD_O(\theta^*_{nL}) \\
&\quad = n^{-1}[n_1 LD(d_1, \widehat{\theta}^x_{nL}) + n_2 LD(d_2, \widehat{\theta}^y_{nL}) - n_1 LD(d_1, \theta^{*x}_{nL}) - n_2 LD(d_2, \theta^{*y}_{nL})] \\
&\quad = -n^{-1}[n_1 \sum_x d_1(x) log m_{\widehat{\theta}^x_{nL}}(x) + n_2 \sum_x d_2(x) log m_{\widehat{\theta}^y_{nL}}(x)] \\
&\qquad + n^{-1}[n_1 \sum_x d_1(x) log m_{\theta^{*x}_{nL}}(x) + n_2 \sum_x d_2(x) log m_{\theta^{*y}_{nL}}(x)] \\
&\quad = n^{-1} log \left\{ \left[\prod_{i=1}^{n_1} m_{\theta^{*x}_{nL}}(X_i) \prod_{j=1}^{n_2} m_{\theta^{*y}_{nL}}(Y_j) \right] / \left[\prod_{i=1}^{n_1} m_{\widehat{\theta}^x_{nL}}(X_i) \prod_{j=1}^{n_2} m_{\widehat{\theta}^y_{nL}}(Y_j) \right] \right\},
\end{aligned} \qquad \ldots(3.4)$$

which shows that the likelihood disparity test is equivalent to the usual likelihood ratio test. Another justification for defining the overall disparity as in equation (3.2) is provided by Simpson (1989, Example 6.2), where a disparity equivalent to (3.2) for the Hellinger distance case has been used.

For the results presented next we assume that Assumptions I-V hold both for the families m_{θ_1} and m_{θ_2}. When ρ is the likelihood disparity this means that the the regularity conditions for Theorem 4.4.4 of Serfling (1980) are satisfied. Following the proof of Lemma 2.1 - *(ii)* it can be seen that the matrix of second partial derivatives of the overall disparity converges to the overall matrix $\mathbf{I}_{\theta_0}$ in

probability. Replacing the disparity $\rho(\beta)$ by $\rho(\theta)$, and $\mathbf{I}_\beta$ by $\mathbf{I}_\theta$ in the proof of Lemma 2.1 - (iii), we see that $\widehat{\theta}_n$ is a consistent estimator of θ_0.

The null hypothesis H_0 imposes r restrictions on θ. We assume that the parameter space can be described through a parameter ν, composed of $k-r$ independent parameters such that $\nu = (\nu^1, \ldots, \nu^{k-r})'$, i.e., $\theta = g(\nu)$ where g is a function from $\Re^{k-r}$ to $\Re^k$. Thus $\theta_n^* = g(\widehat{\nu}_n)$, where $\widehat{\nu}_n$ is the minimum disparity estimator of the parameter in the ν-formulation of the model. Let $\mathbf{D}_\nu$ and $\mathbf{J}_\nu$ be defined as in Section 2, and let ν_0 be the true value of ν.

Define $u_\theta^i(x) = \nabla log m_{\theta_i}(x), i = 1, 2$. Using Assumption V and arguments similar to those used in the proof of Lemma 2.1 - $(i), (ii)$ it can be seen that $n^{1/2}(\widehat{\theta}_n - \theta_0) = \mathbf{I}_{\theta_0}^{-1} n^{1/2}[\frac{1}{n}\Sigma_{i=1}^{n_1} u_{\theta_0}^1(X_i) + \frac{1}{n}\Sigma_{j=1}^{n_2} u_{\theta_0}^2(Y_j)] + o_p(1)$. This establishes that $n^{1/2}(\widehat{\theta}_{nL} - \widehat{\theta}_n) = o_p(1)$ for all minimum disparity estimators $\widehat{\theta}_n$, and $n^{1/2}(\widehat{\theta}_n - \theta_0) \xrightarrow{L} \mathbf{N}(\mathbf{0}, \mathbf{I}_{\theta_0}^{-1})$. Similar arguments establish that under the null hypothesis $n^{1/2}(\theta_{nL}^* - \theta_n^*) = o_p(1)$ for all minimum disparity estimators θ_n^*, and $n^{1/2}(\theta_n^* - \theta_0) \xrightarrow{L} \mathbf{N}(\mathbf{0}, \mathbf{D}_{\nu 0}\mathbf{J}_{\nu 0}^{-1}\mathbf{D}_{\nu 0}')$.

Combining these results we also get

$$n^{1/2}(\widehat{\theta}_{nL} - \theta_{nL}^*) = n^{1/2}(\widehat{\theta}_n - \theta_n^*) + o_p(1), \qquad \ldots(3.5)$$

and

$$n^{1/2}(\widehat{\theta}_{nL} - \theta_{nL}^*) = O_p(1),\ n^{1/2}(\widehat{\theta}_n - \theta_n^*) = O_p(1). \qquad \ldots(3.6)$$

Theorem 3.1. *Under the null hypothesis,* $-2n[LD_O(\widehat{\theta}_{nL}) - LD_O(\theta_{nL}^*)]$ *converges in distribution to* $\chi^2(r)$.

Proof. Let $b_\theta = (R_1(\theta), \ldots, R_r(\theta)')$. Then by an application of the multivariate delta method (Serfling 1980, Theorem 3.3A) the limiting distribution of $n^{1/2}\mathbf{b}_{\widehat{\theta}_{nL}}$ is $\mathbf{N}(\mathbf{b}_{\theta_0}, \mathbf{C}_{\theta_0}\mathbf{I}_{\theta_0}^{-1}\mathbf{C}_{\theta_0}')$, where

$$\mathbf{C}_\theta = \left[\frac{\partial R_i}{\partial \theta_j}\right]_{r \times k}$$

and $\mathbf{C}_{\theta_0}$ is $\mathbf{C}_\theta$ evaluated at $\theta = \theta_0$. By Theorem 3.5 of Serfling (1980) one gets

$$n(\mathbf{b}_{\widehat{\theta}_{nL}})'(\mathbf{C}_{\theta_0}\mathbf{I}_{\theta_0}^{-1}\mathbf{C}_{\theta_0}')^{-1}(\mathbf{b}_{\widehat{\theta}_{nL}}) \xrightarrow{L} \chi^2(r). \qquad \ldots(3.7)$$

Since the second derivative of the likelihood disparity evaluated at θ_0 converges to $\mathbf{I}_{\theta_0}$, we get

$$-2n[LD_O(\widehat{\theta}_{nL}) - LD_O(\theta_{nL}^*)] = n(\widehat{\theta}_{nL} - \theta_{nL}^*)'\mathbf{I}_{\theta_0}(\widehat{\theta}_{nL} - \theta_{nL}^*) + o_p(1). \qquad \ldots(3.8)$$

However, since under the null hypothesis $\mathbf{b}_{\theta_{nL}^*} = 0$, we have

$$\mathbf{b}_{\widehat{\theta}_{nL}} = \mathbf{b}_{\widehat{\theta}_{nL}} - \mathbf{b}_{\theta_{nL}^*} = \mathbf{C}_\theta(\widehat{\theta}_{nL} - \theta_{nL}^*) + o_p(|\ \widehat{\theta}_{nL} - \theta_{nL}^*\ |).$$

Since $n^{1/2}(\hat{\theta}_{nL} - \theta^*_{nL}) = O_p(1)$, (3.7) thus reduces to

$$n(\hat{\theta}_{nL} - \theta^*_{nL})'\mathbf{C}'_{\theta_0}(\mathbf{C}_{\theta_0}\mathbf{I}^{-1}_{\theta_0}\mathbf{C}'_{\theta_0})^{-1}\mathbf{C}_{\theta_0}(\hat{\theta}_{nL} - \theta^*_{nL}) + o_p(1).$$

But $\mathbf{C}'_{\theta_0}(\mathbf{C}_{\theta_0}\mathbf{I}^{-1}_{\theta_0}\mathbf{C}'_{\theta_0})^{-1}\mathbf{C}_{\theta_0} = \mathbf{I}_{\theta_0}$ (Serfling 1980, Theorem 4.4.4), which establishes that the right hand side of equation (3.8) converges to a $\chi^2(r)$ distribution. ▱

Next result establishes that each disparity test is asymptotically equivalent to the likelihood ratio test under the null hypothesis, which together with Theorem 3.1 establishes the limiting distribution of the disparity tests to be $\chi^2(r)$ under the null hypothesis.

Theorem 3.2. *Under the null hypothesis, for any general disparity measure* ρ, $(-2n[LD_O(\hat{\theta}_{nL}) - LD_O(\theta^*_{nL})] + 2n[\rho_O(\hat{\theta}_n) - \rho_O(\theta^*_n)])$ *converges to zero in probability as* $n \to \infty$.

Proof. As in the proof of Theorem 4.4.4. of Serfling (1980, first equation), we have $-2n[LD_O(\hat{\theta}_{nL}) - LD_O(\theta^*_{nL})] = n(\hat{\theta}_{nL} - \theta^*_{nL})'\mathbf{I}_{\theta_0}(\hat{\theta}_{nL} - \theta^*_{nL}) + o_p(1)$ and the limiting distribution of $n(\hat{\theta}_{nL} - \theta^*_{nL})'\mathbf{I}_{\theta_0}(\hat{\theta}_{nL} - \theta^*_{nL})$ is $\chi^2(r)$. Similarly, by the Taylor series expansion of $2n\rho_O(\theta^*_n)$ around $\hat{\theta}_n$ (using $\partial\rho_O(\hat{\theta}_n)/\partial\theta = 0$) we have

$$-2n[\rho_O(\hat{\theta}_n) - \rho_O(\theta^*_n)]$$

$$= n(\hat{\theta}_n - \theta^*_n)'\mathbf{I}_{\theta_0}(\hat{\theta}_n - \theta^*_n) + n(\hat{\theta}_n - \theta^*_n)'\left[\partial^2\rho_O(\tilde{\theta})/\partial\theta\partial\theta' - \mathbf{I}_{\theta_0}\right](\hat{\theta}_n - \theta^*_n)$$

where $\partial^2\rho_O(\tilde{\theta}_n)/\partial\theta\partial\theta'$ is the matrix of second partial derivatives of ρ_O evaluated at $\theta = \tilde{\theta}_n$, a point lying between $\hat{\theta}_n$ and θ^*_n. Since $(\tilde{\theta}_n - \theta_0) = o_p(1)$, so is $[\partial^2\rho_O(\tilde{\theta}_n)/\partial\theta\partial\theta' - \mathbf{I}_{\theta_0}]$. The result then follows from (3.5) and (3.6). ▱

4. Example

In this section we present some numerical evidence that illustrates that the Hellinger distance based disparity test is a far more robust alternative to the likelihood ratio test in the presence of outliers. For our example we consider Poisson populations. This simulation study was performed using FORTRAN on a Sun workstation at the University of Texas at Austin.

The data were generated from two Poisson distributions with parameters θ_1 and θ_2. The hypothesis of interest here is $H_0 : \theta_1 = \theta_2$. The asymptotic distribution of the dispartity tests for this hypothesis is $\chi^2(1)$. For the purpose of our experiment we took $\theta_1 = \theta_2 = 5$. All the computations presented here are based on five thousand replications, the same set of samples being used for the calculation of the two test statistics.

We generated samples of sizes n_1 and n_2 for different combinations of (n_1, n_2); samples of equal sizes $(n_2 = n_1)$ and of unequal sizes $(n_2 = 2n_1)$ were chosen for

$n_1 = 25$, 50 and 100. For each of these cases we computed the empirical levels of the likelihood ratio and the Hellinger distance tests as the proportions of test statistics exceeding the $\chi^2(1)$ critical values. We used 0.10, 0.05 and 0.01 values for the nominal level α. The results are presented in Table 1.

Table 1. EMPIRICAL LEVELS FOR THE TESTS UNDER UNCONTAMINATED DATA

(n_1, n_2)		Likelihood Ratio Test			Hellinger Distance Test		
	α	0.10	0.05	0.01	0.10	0.05	0.01
(25, 25)		0.0960	0.0480	0.0092	0.1416	0.0778	0.0216
(50, 50)		0.0987	0.0498	0.0118	0.1230	0.0681	0.0198
(100, 100)		0.0934	0.0506	0.0094	0.1135	0.0656	0.0149
(25, 50)		0.1049	0.0517	0.0093	0.1476	0.0778	0.0210
(50, 100)		0.1036	0.0545	0.0109	0.1409	0.0727	0.0168
(100, 200)		0.0935	0.0508	0.0092	0.1128	0.0588	0.0146

To study the robustness of the tests we contaminated samples 1 and 2 at the values u_1 and u_2 using contaminating proportions ϵ_1 and ϵ_2 respectively, i.e., instead of using the uncontaminated data $\{d_1(x)\}$ and $\{d_2(y)\}$, we use $\{d_{1,\epsilon_1}(x)\}$ and $\{d_{1,\epsilon_2}(y)\}$ defined by

$$d_{1,\epsilon_1}(x) = (1-\epsilon_1)d_1(x) + \epsilon_1 I_{u_1}(x), \quad 0 < \epsilon_1 < 1,$$

$$d_{1,\epsilon_2}(y) = (1-\epsilon_1)d_2(y) + \epsilon_2 I_{u_2}(y), \quad 0 < \epsilon_2 < 1,$$

where $\mathbf{I}_u$ denotes the indicator function at the value u. We looked at three different cases : (a) $n_1 = n_2$, $\epsilon_1 = 0.10$, $u_1 = 15$, no contamination in the second sample; (b) $n_1 = n_2$, $\epsilon_1 = 0.10$, $u_1 = 10$, $\epsilon_2 = 0.15$, $u_2 = 15$; (c) $n_2 = 2n_1$, $\epsilon_1 = 0.10$, $u_1 = 10$, $\epsilon_2 = 0.15$, $u_2 = 15$. Note that a Poisson (5) random variable takes the values 10 and 15 with approximate probabilities 0.0181 and 0.0002 respectively. These probability values are sufficiently small for us to study the contamination effects at the points $u_1 = 10$ and $u_2 = 15$. The empirical levels using the contaminated data are shown in Table 2.

Table 2. EMPIRICAL LEVELS FOR THE TESTS UNDER CONTAMINATED DATA

$(n_1, n_2, \epsilon_1, \epsilon_2, u_1, u_2)$		Likelihood Ratio Test			Hellinger Distance Test		
	α	0.10	0.05	0.01	0.10	0.05	0.01
(25, 25, 0.10, 0, 15, -)		0.6822	0.5526	0.3016	0.1776	0.1074	0.0334
(50, 50, 0.10, 0, 15, -)		0.9126	0.8442	0.6414	0.1662	0.0944	0.0302
(100, 100, 0.10, 0, 15, -)		0.9966	0.9914	0.9522	0.1716	0.1026	0.0350
(25, 25, 0.10, 0.15, 10, 15)		0.8218	0.7076	0.4198	0.1690	0.0996	0.0290
(50, 50, 0.10, 0.15, 10, 15)		0.9840	0.9552	0.8240	0.1666	0.0972	0.0258
(100, 100, 0.10, 0.15, 10, 15)		1.0000	0.9998	0.9968	0.1706	0.1022	0.0292
(25, 50, 0.10, 0.15, 10, 15)		0.9056	0.8314	0.5764	0.1478	0.0830	0.0250
(50, 100, 0.10, 0.15, 10, 15)		0.9960	0.9888	0.9388	0.1398	0.0798	0.0240
(100, 200, 0.10, 0.15, 10, 15)		1.0000	1.0000	1.0000	0.1492	0.0852	0.0220

The results clearly show the strong robustness properties of the Hellinger distance test relative to the likelihood ratio test. This point has also been observed in the empirical study of Basu and Sarkar (1994c). The level of the likelihood ratio test is not affected much if both the samples are contaminated at the same value at the same proportion. This is not unexpected because this perturbs the estimates of θ_1 and θ_2 roughly by the same amount. We noticed this in our simulations but have not presented those numbers here for brevity.

5. Concluding remarks

Disparity based test in the single population situation have been studied earlier by Simpson (1989), Basu (1993), Basu and Sarkar (1994c) and Lindsay (1994). The present paper extends the above works to the case of general disparity based robust tests for two populations. Extension of the results to the case of k populations, $k \geq 3$, is straightforward if $n^{-1}n_i \to c_i, 0 < c_i < 1$, for each $i = 1, \ldots, k$, where n_i is the sample size for the i-th population and $n = (n_1 + n_2 + \ldots + n_k)$. Our numerical example above demonstrates the robustness properties of the Hellinger distance test. However, the Hellinger distance is just one of several disparities that are known to produce robust estimators and tests in parametric models. For example, several other members of the blended weight Hellinger distance family and the negative exponential disparity can produce estimators and tests which are competitive with the Hellinger distance based statistics. In the single population case, the robustness of such estimators and tests statistics. In the single population case, the robustness of such estimators and tests were demonstrated by Basu and Sarkar (1994c) for the normal models.

In this paper we have considered robust disparity based tests for discrete models. The extension of this theory to continuous models requires additional tools like kernel density estimation methods. Such an extension will be of great value in many practical situations. For example, it will be very useful to determine robust tests for the equality of several means in the one way analysis of variance model.

References

Basu, A. (1993). Minimum disparity estimation : Applications to robust tests of hypotheses. Technical Report # 93-10, Center for Statistical Sciences, University of Texas at Austin, Austin, TX 78712.

Basu, A. and Sarkar, S. (1994a). Minimum disparity estimation in the errors-in-variables model. *Statistics & Probability Letters*. **20**, 69-73.

– – (1994b). On disparity based goodness-of-fit tests for multinomial models. *Statistics & Probability Letters* **19**, 307-312.

– – – (1994c). The trade-off between robustness and efficiency and the effect of model smoothing in minimum disparity inference. *J. Statist. Comput. Simul.*, **50**, 173-185.

BERAN, R.J. (1977). Minimum Hellinger distance estimates for parametric models. *Ann. Statist.* **5**, 445-463.

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

LEHMANN, E.L. (1983). *Theory of Point Estimation.* Wiley : New York.

LINDSAY, B.G. (1994). Efficiency versus robustness : The case for minimum Hellinger distance and related methods. *Ann. Statist.* **22**, 1081-1114.

NEYMAN, J. and PEARSON, E.S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika*, Ser. A. **20**, 175-240.

RAO, C.R. (1961). Asymptotic efficiency and limiting information. Proc. *Fourth Berkeley Symposium*, **1**, 531-546.

SERFLING, R.J. (1980). *Approximation Theorems of Mathematical Statistics.* New York : John Wiely & Sons.

SIMPSON, D.G. (1987). Minimum Hellinger distance estimation for the analysis of count data. *J. Amer. Statist. Assoc.* **82**, 802-807.

– – – – (1989). Hellinger deviance test : Efficiency, breakdown points and examples *J. Amer. Statist. Assoc.* **84**, 107-113.

TAMURA, R.N. and BOOS, D.D. (1986). Minimum Hellinger distance estimation for multivariate location and covariance. *J. Amer Statist. Assoc.* **81**, 223-229.

WILKS, S.S. (1938). The large sample distribution of the likelihood ratio test for testing composite hypothesis. *Ann. Math. Statist.* **9**, 60-62.

DEPARTMENT OF STATISTICS
OKLAHOMA STATE UNIVERSITY
STILWATER, OK 74078
USA

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF TEXAS AT AUSTIN
AUSTIN, TX 78712
USA