Disparity-based Robust Tests for Two Normal Populations

Sayar Karmakar, Ayanendranath Basu

In this paper, we try to address the issue of robustness in testing the equality of means of two normal populations with the same but unknown variance. When the model is misspecified or outliers are present in the data, the usual pooled two-sample 't-test' for this hypothesis performs poorly in terms of the empirical level and power. The t-test statistic is based on the Maximum Likelihood Estimator (MLE) of the parameter. Maximum likelihood estimation can also be viewed as a special case of Minimum Distance Estimation. In this paper we consider a few other density-based distances and contrast the performance of the corresponding minimum distance procedures with the likelihood based methods. In particular, we construct two new test statistics and test the hypothesis of equality of means using them. The tests with these new statistics lead to inference that is substantially more stable relative to the pooled two sample t-test. We demonstrate that these new test statistics perform better in the presence of outliers. We also consider the extension of some of our tests to construct a more robust version of the Analysis of Variance test. Finally, we present a relevant simulation study and discuss several real data examples to substantiate our claims.

Keywords ANOVA, Hellinger Distance, Minimum Disparity Estimation, Negative Exponential Disparity, Outliers, Robustness, t-test

1 Introduction

Parametric models are convenient ways of describing real data in terms of a small number of interpretable parameters. However, all models are approximations to reality and small deviations are never unexpected. Yet, at certain times, such small deviations can have a substantial effect on the classical tests of hypotheses. Such deviations are often manifested through, among other things, large outliers. Presence of outliers is a very common phenomenon that we face while analyzing real data. Sometimes, these outliers may also be introduced due to erroneous handling of the data, rather than through actual model misspecification. There are numerous occasions where a small proportion of

data seem to be far away from the 'data-cloud'. Yet we cannot subjectively ignore a few data points that appear to be geometrically well separated from the rest of the observations. Other cases of model misspecification including the inlier problem (See Lindsay (1994)) can also seriously harm the analysis. In such contexts, the issue of robust inference comes in very naturally. Testing equality of means of two populations is a very common testing frame. If we assume normality and equality of variances of the two populations then this allows us to do an exact pooled two sample t-test which has high power under the true model. But, its performance is severely compromised under model misspecification and the presence of outliers. It often leads to a severe inflation in the observed level, and can also lead to a drastic drop in power under contamination. These facts motivate us to look for new testing procedures which are relatively less affected by the presence of a few geometrically 'distant' datapoints. In order to provide an early motivation for our test statistics, we will start with a specific real example.

1.1 A Motivational Example

Koopmans (1987, Page 86) has presented a dataset in which the yearly measure of wastage (termed as 'run-up') of two cotton mills were reported.

Mill 1	0.12	1.01	−0.20	0.15	−0.30	−0.07	0.32	0.27
	−0.32	−0.17	0.24	0.03	0.35	−0.08	**2.94**	0.28
	1.30	**4.27**	0.14	0.30	0.24	0.13		
Mill 2	1.64	−0.60	−1.16	−0.13	0.40	1.70	0.38	0.43
	1.04	0.42	0.85	0.63	0.90	0.71	0.43	1.97
	0.30	0.76	**7.02**	0.85	0.60	0.29		

Table 1: Mill Data

We subjectively consider the observations written in bold to be outliers. We carry out a two sample pooled t-test on the entire data and the outliers-deleted data which leads to the following results.

	p-value
With outliers	0.3428
Without outlier	0.0308

Table 2: t-test p-values for the mill data example

This demonstrates that the presence of outliers can substantially affect a t-test in terms of its p-value and hence can cause a reversal in the statistical conclusion. In terms of the decision, the removal of outliers in this example changes a clear decision of acceptance (at 5% level) to a comfortable rejection. We will get back to this data example and illustrate the robustness that can be brought in to this hypothesis testing with our new statistics.

2 Density-based Minimum Distance Estimation

In parametric minimum distance estimation two broad classes of distances have been used in the literature, namely the distance between two distribution functions (e.g. the Kolmogorov Smirnov distance) and the distance between two densities (e.g. the chi square type distances). In this paper we will focus on density-based distances. Several minimum distance estimators including those based on the Hellinger distance and the Negative Exponential Disparity have been shown to perform significantly better than the MLE on the robustness count. see Lindsay (1994 [8]) and Basu et. al. (2011[2]) for a detailed description of the structural geometry which naturally leads to the robustness of these estimators. Many of the 'distances' that we consider in this paper are really divergences rather than metrics in the strict sense of the term. They represent discrepancies between density functions but are not necessarily symmetric or do not satisfy the triangle inequality. These include the class of disparities which will be our primary tool in this paper. In a loose sense, however, we will refer to these measures as 'statistical distances' or simply 'distances'.

2.1 The Mathematical Set-up

As we will focus on the normal distribution, we will describe our set-up in the case of a continuous model. Let $X_1, X_2, ..., X_n$ be a sequence of independent and identically distributed observations from a distribution G having density g with respect to the Lebesgue measure. The support is assumed to be the entire real line unless otherwise mentioned. The distribution G will be modeled by the parametric class of distributions $\mathcal{F}_\theta = \{F_\theta : \theta \in \Theta \subset \mathbb{R}^p\}$. As the data are discrete and the model is continuous, one cannot directly construct a distance between the data and the model densities. We will take recourse to kernel density estimation to produce a continuous density estimate representing the data generating density. One can then construct a distance between this density and the model density. This distance may be minimized to determine the corresponding minimum distance estimator, or to be used in tests of hypothesis.

2.2 The Disparity Measure

To introduce the disparity measure, let us consider the parametric set up of the previous section. Let G be the true, data generating distribution, and f_θ be the density function of the model distribution function F_θ. Let $X_1, X_2, ..., X_n$

be a random sample from G, and let

$$g_n^*(x) = \frac{1}{nh_n}\sum_{i=1}^{n} w\left(\frac{x - X_i}{h_n}\right) = \int \frac{1}{h_n} w\left(\frac{x-y}{h_n}\right) dG_n(y)$$

define a nonparametric density estimator of the unknown true density g. Here G_n is the empirical distribution function, $w(\cdot)$ is a smooth kernel function and h_n is the bandwidth. Let C be a thrice differentiable, strictly convex function on $[-1, \infty)$, satisfying $C(0) = 0$. The Pearson residual at the value x is defined by

$$\delta(x) = \frac{g_n^*(x)}{f_\theta(x)} - 1, \tag{1}$$

and the disparity measure generated by C between the densities g_n^* and f_θ is given by

$$\rho_C(g_n^*, f_\theta) = \int C(\delta(x)) f_\theta(x) dx.$$

We drop the differential dx and occasionally the argument x as well for brevity whenever there is no scope of confusion.

As C is convex it follows from Jensen's inequality that $\rho_C(g, f) \geq 0$ for any two densities g and f with respect to the same measure. In minimum disparity estimation, the estimator is the minimizer of $\rho_C(g_n^*, f_\theta)$ over the parameter space Θ. If we take $C(\delta) = (\delta + 1)\log(\delta + 1)$, the disparity is a version of Kullback-Liebler divergence and is denoted as the likelihood disparity. It has the form

$$\mathrm{LD}(g_n^*, f_\theta) = \int g_n^*(x) \log\left(\frac{g_n^*(x)}{f_\theta(x)}\right). \tag{2}$$

Under discrete models, the exact same approach works for the construction of the disparity. In addition, the discrete case has the advantage that now one has a natural density estimate d_n, which is the vector of relative frequencies obtained from the sample. In this case, the likelihood disparity

$$\mathrm{LD}(d_n, f_\theta) = \sum_x d_n(x) \log\left(\frac{d_n(x)}{f_\theta(x)}\right) \tag{3}$$

is a decreasing linear function of the log-likelihood so that the minimum disparity estimator corresponding to the likelihood disparity is actually the maximum likelihood estimator.

For the continuous case, minimizing $\mathrm{LD}(g_n^*, f_\theta)$ is equivalent to maximizing $\int \log(f_\theta(x)) dG_n^*(x)$ where G_n^* is the distribution that corresponds to the density g_n^*. As G_n^* is an estimate of G, an alternative is to simply substitute it with

G_n, which generates the usual log-likelihood (scaled by n)

$$\frac{1}{n}\sum_{i=1}^{n}\log f_\theta(X_i).$$

However, the LD is the only disparity where replacing G_n^* with G_n is possible in the continuous case. In case of the Hellinger distance, for example, the objective function being maximized is

$$\int (g_n^{*1/2} - f_\theta^{1/2})^2 = 2\left[1 - \int g_n^{*1/2} f_\theta^{1/2}\right]$$

and a kernel smoothed density estimate g_n^* is inevitably necessary.

While $V_1 = \int \log(f_\theta(x)) g_n^*(x)$ and $V_2 = \frac{1}{n}\sum_i \log f_\theta(X_i)$ are in general different, they lead to (asymptotically) equivalent results in the following sense. Let

$$\int u_\theta(x) g_n^*(x)dx = 0 \quad \text{and} \quad \frac{1}{n}\sum_i u_\theta(X_i) = 0$$

be the estimating equations resulting from the maximization of these two criterions where $u_\theta(x) = d/d\theta log f_\theta(x)$. Then, under routine conditions on the kernels (Beran,1977, Theorem 4), we have

$$\int u_\theta(x) g_n^*(x) = \frac{1}{n}\sum_i u_\theta(X_i) + o_p(n^{-1/2}), \tag{4}$$

and the estimators are easily seen to be equivalent. Thus the density-based minimum distance estimation procedure is a general one which includes maximum likelihood estimation as a special case.

2.3 The Properties of the C Function

The conditions imposed on C will be called the disparity conditions. Apart from the conditions mentioned in Section 2.2, we need certain other properties of the C function, which we describe below.

- For the analysis of disparity-based minimum distance estimators it is often beneficial to redefine the disparity to make the integrand in the expression of $\rho_C(g_n^*, f_\theta)$ non-negative. As we have already imposed the condition $C(0) = 0$ on the convex function C, non-negativity of the integrand is implied by the additional condition $C'(0) = 0$.

- For determining the asymptotic distribution of the disparity based goodness of fit statistic, it is useful to have the additional condition $C''(0) = 1$.

Both modifications can be done without changing the estimating properties of the disparity. It is easy to see that the minimization problem under these modifications is equivalent to the minimization of $\rho_{C^*}(g_n^*, f_\theta)$ where $C^*(\delta) = \frac{C(\delta)-C'(0)\delta}{C''(0)}$. This standardized form has both the additional properties mentioned above. The estimating properties are preserved since $\rho_{C^*}(g_n^*, f_\theta)$ is just a constant positive multiple of $\rho_C(g_n^*, f_\theta)$. So we can as well work with this standardized C^* function without any loss of generality. Henceforth, we will assume $C'(0) = 0, C''(0) = 1$ unless otherwise stated.

We also have an associated function, the Residual Adjustment Function (RAF) which has a crucial role in the estimation process. Under appropriate differentiability conditions, the minimum disparity estimator is obtained as the solution of the estimating equation

$$-\nabla\rho_C(g_n^*, f_\theta) = \int (C'(\delta)(\delta + 1) - C(\delta))\nabla f_\theta = 0,$$

where δ is as in equation (1) and∇ represents the gradient with respect to θ (Similarly, ∇_2 will represent the second derivative with respect to θ.) The RAF is the function $A(\delta) = C'(\delta)(\delta + 1) - C(\delta)$. Under our assumption on the C function, we will have

$$A(0) = 0 \text{ and } A'(0) = 1.$$

Without loss of generality, we assume that the RAF satisfies these two conditions. The residual adjustment function provides a very convenient construct for the geometrical description of the robustness of the minimum disparity estimators. To be robust, the function should provide a dampened response to increasingly positive δ. For the maximum likelihood estimating function $A(\delta)$ turns out to be δ itself so that the function is linear in δ. Hence the comparison of other minimum distance estimators with the MLE must focus on how other RAFs $A(\delta)$ depart from linearity. Ideally we want these RAFs to provide strong down-weighting for large positive δ.

2.4 Examples of Disparities

The following table gives a list of some useful (from the robustness viewpoint) disparity functions and their corresponding RAF's. A substantially larger list of disparities is provided in Basu. et. al (2011), which also contains an extended discussion of the properties of these distances.

Disparity	C function	RAF
LD	$(\delta+1)\log(\delta+1)-\delta$	δ
HD	$2((\delta+1)^{1/2}-1)^2$	$2((\delta+1)^{1/2}-1)$
PCS	$\frac{\delta^2}{2}$	$\delta+\frac{\delta^2}{2}$
NCS	$\frac{\delta^2}{2(\delta+1)}$	$1-\frac{1}{\delta+1}$
PD	$\frac{(\delta+1)^{\lambda+1}-(\delta+1)}{\lambda(\lambda+1)}-\frac{\delta}{\lambda+1}$	$\frac{(\delta+1)^{\lambda+1}-1}{\lambda+1}$
NED	$e^{-\delta}-1+\delta$	$2-(2+\delta)e^{-\delta}$

2.5 Notation

In this subsection, we present some notation that will facilitate the discussion. We will consider two normal population with possibly different means (μ_1 and μ_2) but a common variance (σ^2). We denote

$$\theta = (\mu_1, \mu_2, \sigma^2)'.$$

$$(\theta)_1 = (\mu_1, \sigma^2)' \text{ and } (\theta)_2 = (\mu_2, \sigma^2)'.$$

So whenever we put a subscript 1 after the bracketed 3×1 vector θ, we refer to the projection of the latter onto the 1^{st} and 3^{rd} co-ordinates of θ and for $(\theta)_2$ it will be the projection onto the 2^{nd} and the 3^{rd} components.

Let x be a 2×1 vector $(x_1, x_2)'$. Then we will denote

$$x^1 = \begin{pmatrix} x_1 \\ 0 \\ x_2 \end{pmatrix} \text{ and } x^2 = \begin{pmatrix} 0 \\ x_1 \\ x_2 \end{pmatrix}.$$

Similarly, if A is a 3×3 matrix,

$$A_1 = \begin{pmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{pmatrix} \text{ and } A_2 = \begin{pmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{pmatrix}.$$

where the subscripts denote the indicated components of A. Let B be a 2×2 matrix. Then

$$B^1 = \begin{pmatrix} b_{11} & 0 & b_{12} \\ 0 & 0 & 0 \\ b_{21} & 0 & b_{22} \end{pmatrix} \text{ and } B^2 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & b_{11} & b_{12} \\ 0 & b_{21} & b_{22} \end{pmatrix}.$$

3 The First Proposal

In this section and the next we will describe two new test statistics for testing the equality of means of two normal populations under the assumption of equal variances. The basic set up is as follows. We have two independent random samples $X_1, X_2, \ldots, X_{n_1}$ and $Y_1, Y_2, \ldots, Y_{n_2}$ from two distributions G_1 and G_2 (having densities g_1 and g_2) where G_i is modeled by a $N(\mu_i, \sigma^2)$ distribution. Thus the means are possibly different, but the variances are assumed to be equal. Under this assumption we want to test the composite null hypothesis

$$H_0 : \mu_1 = \mu_2 \quad \text{versus} \quad H_1 : \mu_1 \neq \mu_2.$$

We denote $\theta = (\mu_1, \mu_2, \sigma^2)$. We will state our results for a general disparity ρ_C. We will minimize the overall disparity ρ_O based on ρ_C which is defined as

$$\rho_O(g^*_{n_1}, g^*_{n_2}, \theta) = \frac{n_1}{n_1 + n_2}\rho_C(g^*_{n_1}, f_{(\theta)_1}) + \frac{n_2}{n_1 + n_2}\rho_C(g^*_{n_2}, f_{(\theta)_2})$$

where $g^*_{n_i}$ is the kernel smoothed density based on the sample from the i^{th} population and $f_{(\theta)_i}$ is the normal probability density function with the parameters $(\theta)_i$ as defined in Section 2.5. Here n_1 and n_2 represent the respective sample sizes with $n_1 + n_2 = n$. Both n_1 and n_2 tend to ∞ at a rate which guarantees that $\frac{n_1}{n_1+n_2} \to w \in (0, 1)$. Then our first test statistic T_1 is defined as

$$T_1 = \frac{2n_1n_2}{n_1 + n_2}\rho_C(f_{(\hat{\theta})_1}, f_{(\hat{\theta})_2})$$

where $\hat{\theta} = \arg\inf_{\theta\in\Theta} \rho_O(g^*_{n_1}, g^*_{n_2}, \theta) == (\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}^2)$.

3.1 Null Distribution of the Statistic T_1

We will need some preliminary results before we can arrive at the distribution of the statistic T_1. Note that the parameter space Θ of $\theta = (\mu_1, \mu_2, \sigma^2)$ is not compact in itself but in can be embedded within a compact space $\bar{\Theta}$ and $\rho_O(g_1, g_2, \theta)$ can be extended to a continuous function of θ on $\bar{\Theta}$, where g_1 and g_2 are the true data generating densities. This can be done by reparamtrizing μ_1, μ_2, σ^2 as $\alpha = (\alpha_1, \alpha_2, \alpha_3)$, $\mu_1 = \tan(\alpha_1)$, $\mu_2 = \tan(\alpha_2)$, $\sigma^2 = \tan(\alpha_3)$.

Henceforth we will assume that the densities g_1 and g_2 belong to the model family. We will let $\theta_0 = (\mu_{10}, \mu_{20}, \sigma^2_0)$ represent the true parameter, so that, according to our notation, $g_1 = f_{(\theta_0)_1}$ and $g_2 = f_{(\theta_0)_2}$. We will, henceforth, also assume that $\theta_0 = \arg\inf \rho_O(f_{(\theta_0)_1}, f_{(\theta_0)_2}, \theta)$. Under the identifiability of the normal densities this assumption is automatically true.

To complete the required derivations, we will show that $\hat{\theta}$ is a consistent estimator of θ_0, and demonstrate its asymptotic normality. We will then com-

bine these results appropriately to derive the asymptotic null distribution of the statistic T_1.

Lemma 3.1. ***(Consistency)*** *Suppose that $|C'(\cdot)|$ is bounded on $[-1, \infty)$. Also assume that $\theta_0 = \arg\inf \rho_O(f_{(\theta_0)_1}, f_{(\theta_0)_2}, \theta)$ is unique. If $g^*_{n_1} \xrightarrow{as} f_{(\theta_0)_1}$ in L_1 , and $g^*_{n_2} \xrightarrow{as} f_{(\theta_0)_2}$ then $\hat{\theta}$ is a consistent estimator of θ_0.*

Proof. First, suppose $g^*_{n_1}$ and $g^*_{n_2}$ are two fixed sequences of densities which tend to $f_{(\theta_0)_1}$ and $f_{(\theta_0)_2}$ in L_1. We will prove that this implies $\hat{\theta}$ goes to θ_0. Note that, in this case, as $g^*_{n_1}$ and $g^*_{n_2}$ are fixed, $\hat{\theta}$ is also a fixed sequence. In our context, as the two former convergences happen almost surely, so does the last convergence and hence $\hat{\theta} \xrightarrow{P} \theta_0$. Define,

$$\rho(t) = \rho_O(f_{(\theta_0)_1}, f_{(\theta_0)_2}, t)$$

and

$$\rho_n(t) = \rho_O(g^*_{n_1}, g^*_{n_2}, t).$$

We have

$$\theta_0 = \arg\inf \rho(t) \text{ and } \hat{\theta} = \arg\inf \rho_n(t).$$

Now,

$$|\rho_n(t) - \rho(t)| \leq \frac{n_1}{n_1 + n_2} \int |C(\delta_{n_1}) - C(\delta_1)| f_{(t)_1} + \frac{n_2}{n_1 + n_2} \int |C(\delta_{n_2}) - C(\delta_2)| f_{(t)_2}$$

where $\delta_{n_i} = g^*_{n_i}/f_{(t)_i} - 1$ and $\delta_i = f_{(\theta_0)_i}/f_{(t)_i} - 1$ for $i = 1, 2$. By the mean value theorem, there exist δ^*_1 and δ^*_2 satisfying

$$C(\delta_{n_i}) - C(\delta_i) = C'(\delta^*_i)(\delta_{n_i} - \delta_i)$$

where δ^*_i lies between δ_{n_i} and δ_i for $i = 1, 2$. Denote $K = \max_\delta |C'(\delta)|$. Then we have,

$$|\rho_n(t) - \rho(t)| \leq K \int [|g^*_{n_1} - f_{(\theta_0)_1}| + |g^*_{n_2} - f_{(\theta_0)_2}|]$$

for all $t \in \Theta$. Hence in this case we get, under the L_1 convergence of $g^*_{n_i}$ to $f_{(\theta_0)_i}$,

$$\sup_t |\rho_n(t) - \rho(t)| \to 0.$$

If $\rho(\theta_0) \geq \rho_n(\hat{\theta})$, then $\rho(\theta_0) - \rho_n(\hat{\theta}) \leq \rho(\hat{\theta}) - \rho_n(\hat{\theta})$, and if $\rho_n(\hat{\theta}) \geq \rho(\theta_0)$, then $\rho_n(\hat{\theta}) - \rho(\theta_0) \leq \rho_n(\theta_0) - \rho(\theta_0)$. Therefore, we have

$$|\rho_n(\hat{\theta}) - \rho(\theta_0)| \leq 2 \sup_t |\rho_n(t) - \rho(t)|,$$

which implies $\rho_n(\hat{\theta}) \to \rho(\theta_0)$ and hence $\rho(\hat{\theta}) \to \rho(\theta_0)$.

If $\hat{\theta}$ does not converge to θ_0, compactness of Θ ensures existence of a subsequence $\{\theta_m\} \subset \{\hat{\theta}\}$ such that $\theta_m \to \theta^* \neq \theta_0$, implying $\rho(\theta_m) \to \rho(\theta^*)$ by the continuity of $\rho(\cdot)$. This implies $\rho(\theta^*) = \rho(\theta_0)$ which contradicts the uniqueness of $\arg\inf \rho_O(f_{(\theta_0)_1}, f_{(\theta_0)_2}, \theta)$. This completes the proof. □

Lemma 3.2. ***(Normality)****: We assume that all the conditions stated in Park and Basu (2004, Theorem 3.4) hold with appropriate modifications in the notation in the context of our problem. Additionally, we assume $|C'(\delta)|$ is bounded in $[-1, \infty)$.*

Under these assumptions, we have the asymptotic convergence

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{D} N(0, I_{\theta_0}^{-1})$$

where

$$I_{\theta_0} = \begin{pmatrix} w\frac{1}{\sigma_0^2} & 0 & 0 \\ 0 & (1-w)\frac{1}{\sigma_0^2} & 0 \\ 0 & 0 & \frac{1}{2\sigma_0^4} \end{pmatrix}.$$

Remark Before we prove this, let us add a few remarks about the matrix I_{θ_0}. We note that, under the notation of Section 2.5,

$$I_{\theta_0} = w(I((\theta_0)_1))^1 + (1-w)(I((\theta_0)_2))^2$$

where

$$I((\theta_0)_1) = \left(\left(E\left[\frac{\partial^2}{\partial(\theta)_1^i \partial(\theta)_1^j} \log f_{(\theta)_1}(X)\right]\right)\right) = \begin{pmatrix} \frac{1}{\sigma_0^2} & 0 \\ 0 & \frac{1}{2\sigma_0^4} \end{pmatrix},$$

$$I((\theta_0)_2) = \left(\left(E\left[\frac{\partial^2}{\partial(\theta)_2^i \partial(\theta)_2^j} \log f_{(\theta)_2}(Y)\right]\right)\right) = \begin{pmatrix} \frac{1}{\sigma_0^2} & 0 \\ 0 & \frac{1}{2\sigma_0^4} \end{pmatrix}.$$

Here, $I((\theta_0)_1)$ is the 2×2 information matrix for a normal population with mean μ_1 and variance σ_0^2. Similarly for $I((\theta_0)_2)$. Now, we are ready to present the proof following the derivations in Park and Basu (2004) and Lemma 3.1.

Proof. Due to the condition (c) imposed on the kernel function in Theorem 3.4 in Park and Basu (2004), we have, $g_{n_i}^*(x) \xrightarrow{a.s} f_{(\theta_0)_i}(x)$ for every x and

$$\int |g_{n_i}^*(x) - f_{\theta_0}(x)| dx \to 0,$$

for $i = 1, 2$ and hence by Lemma 3.1, $\hat{\theta} \xrightarrow{P} \theta_0$.

For convenience let us write $\rho_O(\theta) = \rho_O(g^*_{n_1}, g^*_{n_2}, f_\theta)$. As $\hat{\theta}$ minimizes $\rho_O(\theta)$ over Θ, Taylor's theorem yields

$$\begin{aligned} 0 &= \nabla\rho_O(\widehat{\theta}) = \nabla\rho_O(\theta_0) + \nabla^2\rho_O(\theta^*)(\widehat{\theta} - \theta_0) \\ \sqrt{n}(\widehat{\theta} - \theta_0) &= (\nabla^2\rho_O(\theta^*))^{-1}(-\sqrt{n}\nabla\rho_O(\theta_0)), \end{aligned}$$

where θ^* is a point on the line segment joining θ_0 and $\hat{\theta}$. Now, as $\hat{\theta}$ is trapped in between $\hat{\theta}$ and θ_0, it also converges to θ_0 in probability. We will prove that,

(A1) $\nabla^2\rho_O(\theta^*) \xrightarrow{P} I_{\theta_0}$.

(A2) $-\sqrt{n}\nabla\rho_O(\theta_0) \xrightarrow{D} N(0, I_{\theta_0})$.

That proving these two suffices for the final conclusion is evident from Slutsky's theorem. To prove (A1) and (A2) we will follow the proofs of Equation (6) and Equation (7) in Theorem 3.4 in Park and Basu (2004).

Proof. **(A1)** Note that,

$$\nabla^2\rho_O(\theta^*) = \frac{n_1}{n_1 + n_2}(\nabla^2\rho_1((\theta^*)_1))^1 + \frac{n_2}{n_1 + n_2}(\nabla^2\rho_2((\theta^*)_2))^2.$$

As $\hat{\theta}^*$ is in between $\hat{\theta}$ and θ_0, θ^* converges in probability to θ_0. Consequently, $(\theta^*)_i$, for $i = 1, 2$, converges in probability to $(\theta_0)_i$. Then it follows from Park and Basu (2004, Theorem 3.4) that

$$\nabla^2\rho_1((\theta^*)_1) \xrightarrow{P} I((\theta_0)_1)$$

and

$$\nabla^2\rho_2((\theta^*)_1) \xrightarrow{P} I((\theta_0)_2).$$

Therefore,

$$\nabla^2\rho_O(\theta^*) = \frac{n_1}{n_1 + n_2}(I((\theta_0)_1))^1 + \frac{n_2}{n_1 + n_2}(I((\theta_0)_2))^2 + o_p(1) = I_{\theta_0} + o_p(1)$$

and this completes the proof of this part. □

Proof. **(A2)** Note that,

$$\begin{aligned} -\sqrt{n}\nabla\rho_O(\theta_0) &= -\sqrt{n}\left[\left(\frac{n_1}{n_1 + n_2}\nabla\rho_1((\theta_0)_1)\right)^1 + \left(\frac{n_2}{n_1 + n_2}\nabla\rho_1((\theta_0)_1)\right)^2\right] \\ &= \sqrt{\frac{n_1}{n_1 + n_2}}(-\sqrt{n_1}\nabla\rho_1((\theta_0)_1))^1 + \sqrt{\frac{n_1}{n_1 + n_2}}(-\sqrt{n_1}\nabla\rho_1((\theta_0)_1))^2. \end{aligned}$$

Applying Equation (7) from Park and Basu (2004, Theorem 3.4) we have

- $(-\sqrt{n_1}\nabla\rho_1((\theta_0)_1)) \xrightarrow{D} N(0, I((\theta_0)_1))$,
- $(-\sqrt{n_2}\nabla\rho_2((\theta_0)_2)) \xrightarrow{D} N(0, I((\theta_0)_2))$.

Also, these two are independent as $(-\sqrt{n_1}\nabla\rho_1((\theta_0)_1))$ and $(-\sqrt{n_2}\nabla\rho_2((\theta_0)_2))$ are related to observations from the first and second samples respectively. Therefore,

$$-\sqrt{n}\nabla\rho_O(\theta_0) \xrightarrow{D} N(0, I_{\theta_0}).$$

□

Hence, the result is proved. □

Result 3.3(Asymptotic null distribution) Suppose that, the assumptions stated in Park and Basu (2004, Theorem 3.4) hold as do the assumptions of Lemma 3.1. Then T_1 has an asymptotic χ^2 distribution with 1 degree of freedom.

Proof. Let $\alpha((\mu, \sigma^2)')$ denote $\rho_C(f_{(\mu,\sigma^2)'}, f_{(\hat{\theta})_2})$. We write T_1 as $2c_n\alpha((\hat{\mu}_1, \hat{\sigma}^2)')$ where, $c_n = \frac{n_1 n_2}{n_1+n_2}$. Now,

$$\alpha((\hat{\mu_1}, \hat{\sigma^2})') = \alpha((\hat{\mu_2}, \hat{\sigma^2})') + \nabla\alpha((\hat{\mu_2}, \hat{\sigma^2})')'(\hat{\mu}_1 - \hat{\mu}_2, 0)' + \frac{1}{2}(\hat{\mu}_1 - \hat{\mu}_2, 0)\nabla^2\alpha(\theta_2^*)(\hat{\mu}_1 - \hat{\mu}_2, 0)'$$

where θ_2^* lies between $(\hat{\mu}_1, \hat{\sigma}^2)'$ and $(\hat{\mu}_2, \hat{\sigma}^2)'$ and hence under null goes to $(\mu_0, \sigma_0^2)'$ in probability.

Now, from the definition of $\alpha((\mu, \sigma^2)')$ it follows that,

$$\alpha((\hat{\mu}_2, \hat{\sigma}^2)') = 0.$$

and

$$\nabla\alpha((\hat{\mu}_2, \hat{\sigma}^2)') = 0$$

where ∇ represents the gradient with respect to (μ, σ^2) and $\nabla\alpha((\hat{\mu}_2, \hat{\sigma}^2)')$ is the derivative of $\alpha(\cdot)$ evaluated at $(\hat{\mu}_1, \hat{\sigma}^2)'$. The latter equality follows as the minima of the function $\alpha(\cdot, \cdot)$ is attained at $(\hat{\mu}_2, \hat{\sigma}^2)'$.

Simplifying the above expression we obtain

$$T_1 = c_n(\hat{\mu}_1 - \hat{\mu}_2)'\frac{\partial^2}{\partial\mu^2}\alpha((\mu^*, \sigma^*)')(\hat{\mu}_1 - \hat{\mu}_2.)$$

Using assumption (a) of Theorem 3.4 of Park and Basu (2004), the boundedness of $C'(\delta)$ and the Dominated Convergence Theorem gives us

$$\frac{\partial^2}{\partial\mu^2}\alpha((\mu^*, \sigma^*)') = \frac{\partial^2}{\partial\mu^2}\alpha((\hat{\mu}_2, \hat{\sigma}^2)') + o_p(1) = \frac{1}{\hat{\sigma}^2} + o_p(1)$$

The latter equality is established by the following simple calculation.

$$\begin{aligned}
& \frac{\partial^2}{\partial\mu^2}\int C\left(\frac{f_{(\theta)_1}}{f_{\widehat{(\theta)_2}}}-1\right)f_{\widehat{(\theta)_2}}|_{(\theta)_1=\widehat{(\theta)_2}} \\
= & \int C''\left(\frac{f_{(\theta)_1}}{f_{\widehat{(\theta)_2}}}-1\right)\frac{(\frac{\partial}{\partial\mu}f_{(\theta)_1})^2}{f_{\widehat{(\theta)_2}}}+C'\left(\frac{f_{(\theta)_1}}{f_{\widehat{(\theta)_2}}}-1\right)\frac{\partial^2}{\partial\mu^2}f_{(\theta)_1}|_{(\theta)_1=\widehat{(\theta)_2}} \\
= & \int\frac{(\frac{\partial}{\partial\mu}f_{(\theta)_1})^2}{f_{\widehat{(\theta)_2}}}|_{(\theta)_1=\widehat{(\theta)_2}}\ [\text{As } C'(0)=0 \text{ and } C''(0)=1] \\
= & \int\left(\frac{\partial}{\partial\mu}\log f_{(\theta)_1}|_{(\theta)_1=\widehat{(\theta)_2}}\right)^2 f_{\widehat{(\theta)_2}} \\
= & E_{\widehat{(\theta)_2}}\left(\frac{x-\hat{\mu}_2}{\hat{\sigma}^2}\right)^2 \\
= & \frac{1}{\widehat{\sigma^2}}.
\end{aligned}$$

So,

$$T_1 = c_n\frac{(\hat{\mu}_1-\hat{\mu}_2)'(\hat{\mu}_1-\hat{\mu}_2)}{\widehat{\sigma^2}}+c_n(\hat{\mu}_1-\hat{\mu}_2)'(\hat{\mu}_1-\hat{\mu}_2)o_p(1).$$

Now, from Lemma 3.2, it follows that $\sqrt{n_1+n_2}(\hat{\mu}_1-\hat{\mu}_2)\xrightarrow{D}N(0,\sigma_0^2(\frac{1}{w}+\frac{1}{1-w}))$. That implies $\sqrt{n_1+n_2}\sqrt{w(1-w)}(\hat{\mu}_1-\hat{\mu}_2)\xrightarrow{D}N(0,\sigma_0^2)$. Now, it is trivial to show that as n_1,n_2 goes to ∞ in a way that $\frac{n_1}{n_1+n_2}\to w$ then

$$\frac{(n_1+n_2)w(1-w)}{\frac{n_1n_2}{n_1+n_2}}\to 1$$

So,

$$\sqrt{\frac{n_1n_2}{n_1+n_2}}(\mu_1-\mu_2)\xrightarrow{D}N(0,\sigma_0^2).$$

From Lemma 3.1,

$$\hat{\sigma}^2\xrightarrow{P}\sigma_0^2.$$

By Slutsky's theorem,

$$\frac{n_1n_2}{n_1+n_2}\frac{(\hat{\mu}_1-\hat{\mu}_2)'(\hat{\mu}_1-\hat{\mu}_2)}{\widehat{\sigma^2}}\xrightarrow{D}\chi_1^2$$

$$\frac{n_1n_2}{n_1+n_2}\frac{(\hat{\mu}_1-\hat{\mu}_2)'(\hat{\mu}_1-\hat{\mu}_2)}{\widehat{\sigma^2}}o_p(1)=O_p(1)o_p(1)=o_p(1).$$

So, we note that $c_n=\frac{n_1n_2}{n_1+n_2}$ is the correct multiplier. The last two equations with the help of Slutsky's theorem implies

$$2c_n\alpha((\hat{\mu}_1,\hat{\sigma}^2)')\xrightarrow{D}\chi_1^2$$

i.e. under H_0,

$$T_1 \sim \chi^2_1 \text{ asymptotically.}$$

□

3.2 Modifications for the inclusion of Hellinger Distance

For the Hellinger distance, as we can see, none of the three functions $|C'(\delta)|$, $A(\delta)$ and $A'(\delta)(\delta+1)$ are bounded on $[-1,\infty)$. Naturally, if we look at the subclass of disparities for which Park and Basu (2004) results are valid, it does not include Hellinger distance. But Hellinger distance remains one of the most popular distances used in minimum distance literature. So, in this subsection, we will discuss conditions which will allow the Hellinger distance to be included. **Consistency**: For the Hellinger disparity, we see that

$$\begin{aligned} |\rho_n(t)-\rho(t)| &\leq 4\left(\int |(g_{n_1}^{*1/2}-f_{(\theta_0)_1}^{1/2})|f_{(t)_1}^{1/2}+\int |(g_{n_2}^{*1/2}-f_{(\theta_0)_2}^{1/2})|f_{(t)_2}^{1/2}\right) \\ &\leq 4\left(\int |(g_{n_1}^{*1/2}-f_{(\theta_0)_1}^{1/2})|^2+\int |(g_{n_2}^{*1/2}-f_{(\theta_0)_2}^{1/2})|^2\right) \end{aligned}$$

where the last inequality follows from Cauch-Schwarz inequality.

Thus the required condition for $\sup_t |\rho_n(t)-\rho(t)|$ to converge to 0 is that $g_{n_i}^{*1/2} \to f_{(\theta_0)_i}^{1/2}$ in L_2 for $i=1,2$. Under the restrictions imposed on the kernel in Lemma 3.2 , the kernel smoothed densities $g_{n_i}^*$ indeed satisfy this (Beran,1977).

Normality: In Lemma 3.2, while we prove conclusion (A1), we require the derivation of Equation (6) as done in Park and Basu Theorem 3.4. A crucial step in that derivation was that $A(\delta^*)$ and $A'(\delta^*)(\delta^*+1)$ is bounded in $[-1,\infty)$. In the case of Hellinger disparity, $A(\delta^*)$ and $A'(\delta^*)(\delta^*+1)$ are both linear function of $\sqrt{\delta^*+1} = \sqrt{g_{n_i}^*/f_{(\theta^*)_i}}$ where $\theta^* = \theta_0 + o_p(1)$. Now, from the conditions imposed on the kernel $g_{n_i}^*(x) \overset{as}{\to} f_{(\theta_0)_i}(x)$ and $f_{(\theta^*)_i}(x) \overset{as}{\to} f_{(\theta_0)_i}(x)$. If we assume that $\sup_x(g_{n_i}^*(x)/f_{(\phi)_i}(x))$ is bounded where $\phi = \theta_0 + o_p(1)$ then we can carry out the proof as done in the proof of (6) in Theorem 3.4 from Park and Basu (2004).

For the proof of (A2) in Lemma 3.2, we have used Park and Basu Theorem 3.4 derivation of Equation (7). We note that the crucial step is to find a B such that

$$|A(r^2-1)-(r^2-1)| \leq B(r-1)^2 \text{ for all } r.$$

For the Hellinger disparity we can see that left hand side of the above inequality itself turns out to be $(r-1)^2$ and hence is trivially satisfied for any choice of $B \geq 1$. Thus, although $A'(\delta)$ and $A''(\delta)(\delta+1)$ are not bounded, we can still prove Equation(7) of Theorem 3.4 in Park and Basu (2004) automatically holds in a direct manner for the special case of Hellinger Distance.

The proof for Result 3.3 remains the same for the Hellinger distance under the additional assumptions required here.

3.3 Power Approximation of T_1

We will approximate the power of the test T_1 based on its asymptotic distribution. In this test we reject the null at level α if observed value of T_1 is greater than the chi-square $\alpha\%$ upper quantile. To find the power, we need to find out $P_\theta[T_1 > \chi^2_{1,\alpha}]$ where $\theta = (\mu_1\ \mu_2\ \sigma^2)'$. Here μ_1, μ_2 are not necessarily equal. Let $r(\theta) = \rho_C(f_{(\theta)_1}, f_{(\theta)_2})$. Then,

$$\sqrt{n}r(\widehat{\theta}) = \sqrt{n}r(\theta) + (\nabla_\theta r(\theta))'\sqrt{n}(\widehat{\theta}-\theta) + \frac{1}{2}\frac{1}{\sqrt{n}}[\sqrt{n}(\widehat{\theta}-\theta)'\nabla_{2\theta}r(\theta^*)\sqrt{n}(\widehat{\theta}-\theta)]$$

A careful check of the proof of Lemma 3.2 (Asymptotic normality) indicates that even under a general θ,

$$\sqrt{n}(\widehat{\theta}-\theta) \sim N(0, I(\theta)^{-1}).$$

Therefore, under θ, the third term of the above Taylor series expansion is $o_p(1)$ and the second term follows a normal distribution with mean 0 and variance $\nabla_\theta r(\theta)' I(\theta) \nabla_\theta r(\theta)$. If we call this variance σ^2_θ then asymptotically,

$$\sqrt{n}\left[\frac{T_1}{2c_n} - r(\theta)\right] \sim N(0, \sigma^2_\theta).$$

The approximated power is

$$1 - \Phi\left(\frac{\sqrt{n}(\frac{\chi^2_{1,\alpha}}{2c_n} - r(\theta))}{\sigma_\theta}\right).$$

4 The Second Statistic

In this section, the basic set-up of the problem remains the same as the previous one. We define $\hat{\theta}_0 = (\hat{\mu}_0, \hat{\mu}_0, \hat{\sigma}^2_0)'$ where

$$(\hat{\mu}_0, \hat{\sigma}^2_0) = \arg\inf_{\mu,\sigma^2} \frac{n_1}{n_1+n_2}\rho_C(g^*_{n_1}, f_{(\mu,\sigma^2)}) + \frac{n_2}{n_1+n_2}\rho_C(g^*_{n_2}, f_{(\mu,\sigma^2)})$$

where $g^*_{n_i}$ is the kernel-smoothed density for the i^{th} population and $f_{(\mu,\sigma^2)}$ is the normal probability density function (pdf) with mean μ and variance σ^2.

Here our test statistic is

$$T_2 = 2n(\rho_O(\widehat{\theta_0}) - \rho_O(\widehat{\theta})).$$

The intuition behind this test statistic is that if the null is true then $\widehat{\theta}$ and $\widehat{\theta_0}$ should be close. Hence the test statistic described above should not be too large. In the following subsections, we will show that, under H_0,

$$T_2 \xrightarrow{L} \chi^2_1.$$

The degree of freedom of the asymptotic chi-square distribution is also natural as there is just one restriction imposed by H_0.

4.1 Some Important Results

In order to prove the null distribution of T_2 we will need the following results listed below. Suppose that $\widehat{\theta}$ and $\widehat{\theta_0}$ denote the estimators for a fixed ρ and $\hat{\theta}_{MLE}$ and $\hat{\theta}_{0MLE}$ denote the corresponding maximum likelihood estimators. Then,

(B1) $\sqrt{n}(\widehat{\theta} - \theta_0) = I^{-1}_{\theta_0}(K_{\theta_0}) + o_p(1)$ where $K(\theta_0)$ does not depend on the disparity used.

(B2) $\sqrt{n}(\widehat{\theta} - \widehat{\theta_0}) = \sqrt{n}(\widehat{\theta}_{MLE} - \widehat{\theta}_{0MLE}) + o_p(1)$.

(B3) $\sqrt{n}(\widehat{\theta} - \widehat{\theta_0}) = O_p(1)$.

Proof. **(B1)** By a Taylor's series expansion,

$$\begin{aligned}\sqrt{n}(\widehat{\theta} - \theta_0) &= (\nabla_2\rho_O(\theta^*))^{-1}(-\sqrt{n}\nabla\rho_O(\theta_0)) \\ &= I^{-1}_{\theta_0}(-\sqrt{n}\nabla\rho_O(\theta_0)) + o_p(1)\end{aligned}$$

where the last equality follows from Equation (A1) in Lemma 3.2. Here, by the consistency of $\hat{\theta}$, we get $\theta^* = \theta_0 + o_p(1)$. Now,

$$\begin{aligned}-\sqrt{n}\nabla\rho_O(\theta_0) &= \sqrt{\frac{n_1}{n}}(-\sqrt{n_1}\nabla(\rho_C(\theta_0)_1))^1 + \sqrt{\frac{n_2}{n}}(-\sqrt{n_2}\nabla(\rho_C(\theta_0)_2))^2 \\ &= \sqrt{\frac{n_1}{n}}\left(-\sqrt{n_1}\int A(\delta_{n_1})\nabla f\right)^1 + \sqrt{\frac{n_2}{n}}\left(-\sqrt{n_2}\int A(\delta_{n_2})\nabla f\right)^2 \\ &= \sqrt{\frac{n_1}{n}}\left(-\sqrt{n_1}\int \delta_{n_1}\nabla f\right)^1 + \sqrt{\frac{n_2}{n}}\left(-\sqrt{n_2}\int \delta_{n_2}\nabla f\right)^2 + o_p(1) \\ &= \sqrt{\frac{n_1}{n}}\left(-\sqrt{n_1}\int u_{(\theta_0)_1} g^*_{n_1}\right)^1 + \sqrt{\frac{n_2}{n}}\left(-\sqrt{n_2}\int u_{(\theta_0)_2} g^*_{n_2}\right)^2 + o_p(1)\end{aligned}$$

where $\delta_{n_i} = g^*_{n_i}/f_{(\theta_0)_i} - 1$ for $i = 1, 2$. The penultimate equality follows from Equation (7) of Park and Basu (2004, Theorem 3.4)). The final expression obviously does not depend on the disparity used. We call this expression by K_{θ_0}. □

Proof. **(B2)**

From Result (B1) and Equation (4) it follows that,

$$\sqrt{n}(\widehat{\theta} - \widehat{\theta}_{MLE}) = o_p(1) \tag{5}$$

Under the null, we have a restriction that the first two co-ordinates of θ is same. So $\theta_0 = (\mu_0, \mu_0, \sigma_0^2)'$. We call the subset of parameter vectors in Θ which has its first two co-ordinates same by Θ_0. If we construct the vector of free parameters under the restrictions of the null, $\nu = (\mu, \sigma^2)$ then, $g(\nu) = (\mu, \mu, \sigma^2) \in \Theta_0$. Now, we can prove consistency and asymptotic normality of $\hat{nu}$ in a similar fashion and thus

$$\sqrt{n}(\widehat{\nu} - \widehat{\nu}_{MLE}) = o_p(1).$$

This allows us to prove,

$$\sqrt{n}(g(\widehat{\nu}) - g(\widehat{\nu}_{MLE})) = o_p(1).$$

In other words,

$$\sqrt{n}(\widehat{\theta_0} - \widehat{\theta}_{0MLE}) = o_p(1). \tag{6}$$

Equations (4) and (5) give us the Result (B2). The proofs of relevant results for $\hat{\nu}$ involve no essential additional difficulty and hence is omitted for brevity. □

Proof. **(B3)** We have already proved that

$$\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow{D} N(\mathbf{0}, I_{\theta_0}^{-1}).$$

And hence,

$$\sqrt{n}(\widehat{\theta} - \theta_0) = O_p(1).$$

Under this v-formulation, it can be shown that,

$$\sqrt{n}(\widehat{\theta_0} - \theta_0) = O_p(1)$$

If the above two are $O_p(1)$ then so are their difference. Hence, Result (B3) is established. □

4.2 Final Proof For The Null Distribution

An application of the Taylor series in $\hat{\theta}_0$ around $\hat{\theta}$ gives

$$\begin{aligned} T_2 &= 2n(\rho_O(\widehat{\theta_0}) - \rho_O(\widehat{\theta})) \\ &= 2n(\widehat{\theta_0} - \widehat{\theta})'\nabla\rho_O(\widehat{\theta}) + n(\widehat{\theta} - \widehat{\theta_0})'\nabla^2\rho_O(\theta^*)(\widehat{\theta} - \widehat{\theta_0}) \\ &= n(\widehat{\theta} - \widehat{\theta_0})'\nabla^2\rho_O(\theta^*)(\widehat{\theta} - \widehat{\theta_0}) \end{aligned}$$

where the last equality follows from the fact that $\nabla_O(\hat{\theta}) = 0$. Now, Result (B3), Result (A2) from Lemma (3.2) imply

$$T_2 = \left(\sqrt{n}(\widehat{\theta} - \widehat{\theta_0})\right)' (I_{\theta_0}) \left(\sqrt{n}(\widehat{\theta} - \widehat{\theta_0})\right) + o_p(1)$$

Another application of Result (B2) establishes that

$$T_2 = \left(\sqrt{n}(\widehat{\theta}_{MLE} - \widehat{\theta}_{0MLE})\right)' (I_{\theta_0}) \left(\sqrt{n}(\widehat{\theta}_{MLE} - \widehat{\theta}_{0MLE})\right) + o_p(1).$$

From Serfling (1980) we have

$$\left(\sqrt{n}(\widehat{\theta}_{MLE} - \widehat{\theta}_{0MLE})\right)' (I_{\theta_0}) \left(\sqrt{n}(\widehat{\theta}_{MLE} - \widehat{\theta}_{0MLE})\right) \xrightarrow{D} \chi^2_r$$

under the null , where r is the number of restriction($r = 1$ in our case).

5 Level and Power of the New Tests

In this section we will demonstrate how our test statistics perform under pure and contaminated data. First, we will be describing a few simulations. We will contrast the performance of our test statistics with the pooled t-test and Yuen's test (1974). We will show how small contamination of only 10 % can have a severe effect on the performance of the pooled t-test. In the literature, there already exists some robust tests for testing the hypothesis of equality of mean. We will compare our tests with Yuen's(1974) trimmed t-test. The latter test is a t-test on a symmetrically winsorized sample. We will consider two different winsorization proportions (10% and 5%) and look at their performances. We call these two statistics by W_1 and W_2 in our tables. The pair of sample sizes chosen are (30,50), (50,50) and (70,50). We are providing the results for the Hellinger distances here.

For each calculation, we have chosen the bi weight kernel with the bandwidth chosen according to the rules specified by Simpson(1989).

5.1 Level comparison

For performance in terms of the empirical level, we have generated two datasets from $N(0,1)$ and $N(0,1)$ with sample sizes specified above. Then we have discussed 11 cases. One of them is without any contamination. For the others, we have contaminated the second data in 10% proportion with the degenerate random variable y where y takes value from 0 to -9. The proportion of rejection in 500 replications is reported in the following table.

Table 3: Level comparison

Level			0	-1	-2	-3	-4	-5	-6	-7	-8	-9
30-50	T_1	0.07	0.068	0.088	0.124	0.16	0.124	0.1	0.098	0.102	0.092	0.086
	T_2	0.087	0.086	0.086	0.161	0.334	0.311	0.201	0.246	0.234	0.223	0.193
	W_1	0.154	0.144	0.204	0.262	0.296	0.308	0.312	0.314	0.306	0.304	0.3
	W_2	0.048	0.056	0.084	0.138	0.194	0.226	0.246	0.27	0.276	0.292	0.306
	t	0.05	0.052	0.062	0.108	0.154	0.202	0.23	0.248	0.254	0.28	0.28
50-50	T_1	0.062	0.048	0.076	0.144	0.16	0.13	0.08	0.092	0.1	0.09	0.086
	T_2	0.105	0.073	0.107	0.177	0.373	0.261	0.199	0.242	0.242	0.248	0.191
	W_1	0.162	0.144	0.184	0.248	0.272	0.288	0.292	0.288	0.284	0.282	0.276
	W_2	0.042	0.044	0.07	0.134	0.216	0.294	0.368	0.43	0.462	0.496	0.532
	t	0.06	0.046	0.066	0.15	0.216	0.312	0.366	0.422	0.444	0.482	0.504
70-50	T_1	0.074	0.068	0.1	0.152	0.186	0.102	0.072	0.092	0.094	0.09	0.096
	T_2	0.087	0.061	0.121	0.164	0.349	0.213	0.202	0.204	0.219	0.245	0.189
	W_1	0.16	0.168	0.232	0.32	0.34	0.34	0.34	0.34	0.336	0.33	0.318
	W_2	0.056	0.05	0.072	0.152	0.274	0.376	0.464	0.548	0.592	0.628	0.652
	t	0.04	0.038	0.064	0.16	0.316	0.404	0.5	0.586	0.628	0.658	0.674

From the above table, we can make the following remarks:

- For the new tests T_1 and T_2 we see that the observed level is increasing up to the contamination of -3 and then it slowly goes down. But for the trimmed t tests or usual pooled t test they are steadily increasing.

- Out of T_1 and T_2 we see that T_1 remains much closer to the asymptotic level 0.05. The statistic T_2 brings the observed level down to 0.19 whereas T_1 can bring it down to 0.08. We can also explain why does these level first increase and then decrease. Initially the contaminating point is not too far from the mean of the original population. So the 10 % contamination, sometimes, were not identified as outliers. But after a certain stage, as the contaminating discrete mass goes further, it becomes identified as outlier and hence the robust estimators down-weight the impact of those points and hence the level comes closer to the asymptotically true level.

- We also see that the existing robust tests behave quite similar to the t-test. naturally when the trimming proportion is only 5% it is much closer to the t test than the one with 10 % trimming. When the data is trimmed 10% from each side, most of the time the contaminating points go away and hence our results become more robust. But still the observed level cannot go down below .3 even when the outlier is very extreme. This is a serious drawback of an existing measure which is supposed to perform good in the presence of outliers.

- We also note that our new test statistics perform well in all of the above combination of sample sizes. But the trimmed t test and the poled t test have huge variation for the three combinations.

5.2 Power comparison

We generate the second data from $N(1,1)$. We keep the contaminating set-up the same.The choice of -9 was made so that the mean of the true data generating distribution becomes zero. Here, proportion of rejection will give us an idea about the asymptotic power. The following two tables report the performance of the tests based on the Hellinger distance.

Table 4: Power Comparison

Power		..	0	-1	-2	-3	-4	-5	-6	-7	-8	-9
30-50	T_1	0.988	0.954	0.912	0.828	0.888	0.958	0.966	0.972	0.97	0.978	0.978
	T_2	0.864	0.779	0.836	0.773	0.883	0.908	0.912	0.914	0.918	0.918	0.917
	W_1	0.994	0.982	0.938	0.906	0.87	0.816	0.792	0.75	0.734	0.71	0.692
	W_2	1	0.986	0.898	0.792	0.65	0.498	0.414	0.338	0.278	0.236	0.208
	t	0.994	0.974	0.918	0.72	0.504	0.312	0.192	0.122	0.088	0.06	0.04
50-50	T_1	0.994	0.978	0.956	0.92	0.946	0.986	0.984	0.984	0.986	0.99	0.99
	T_2	0.906	0.852	0.859	0.851	0.91	0.952	0.938	0.944	0.942	0.94	0.94
	W_1	1	0.996	0.988	0.928	0.878	0.83	0.784	0.75	0.732	0.7	0.68
	W_2	0.998	0.998	0.966	0.864	0.732	0.6	0.5	0.41	0.348	0.296	0.248
	t	1	0.994	0.962	0.858	0.634	0.434	0.282	0.2	0.138	0.108	0.066
70-50	T_1	1	0.988	0.958	0.94	0.964	0.99	0.992	0.996	0.996	0.994	0.992
	T_2	0.933	0.881	0.871	0.873	0.904	0.951	0.955	0.956	0.962	0.969	0.957
	W_1	1	0.998	0.98	0.944	0.908	0.844	0.788	0.758	0.724	0.712	0.686
	W_2	1	1	0.978	0.92	0.778	0.64	0.506	0.422	0.334	0.276	0.248
	t	1	1	0.972	0.88	0.706	0.504	0.364	0.23	0.16	0.124	0.11

We can make the following remarks about the above table

- We see that initially the power of T_1 and T_2 were below the pooled t and the trimmed t test but by the time the contaminating value is -3 or

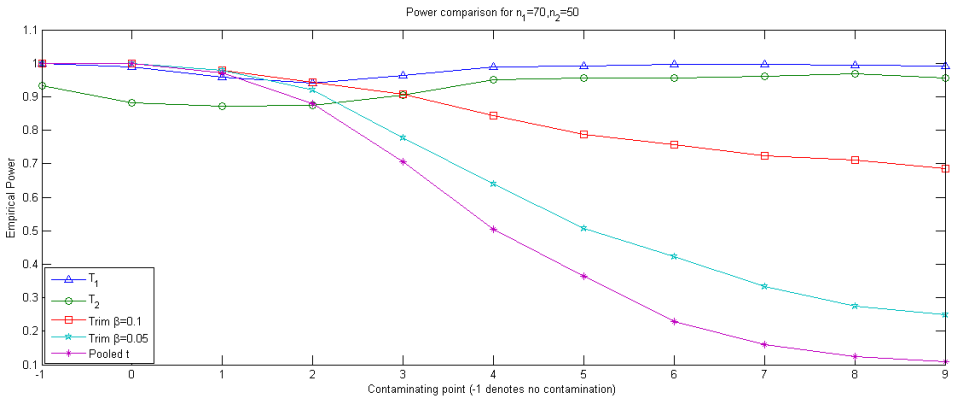

Figure 1: Power: n_1=70, n_2=50

smaller, the powers of pooled t and trimmed t go down steadily whereas the power for our tests increase steadily.

- Out of T_1 and T_2 we see that the power of T_1 is almost always better than that of T_2. But compared to other tests, both these tests are far more efficient.

- The 10 % trimmed test performs a little better than others but in the extreme case it cannot give a power more than 0.7 whereas T_1 and T_2 is giving a power of around 0.95 to 0.99.

- The 5% trimmed t test does not do as good as the 10% trimmed one but it is better than the pooled t-test. Pattern wise, both these trimmed tests are similar to the pooled t test.

- t-test performance is very poor in terms of robustness. Even in the contamination of only 10% the power can go as low as 0.04. The drop is very sharp too. This is a very serious concern about the pooled t-test. It is an exact test and it might be a little shocking that it is performing this poorly. This motivates us to look for an explanation. The problem is that in this setup the mean of the contaminated population and the true population were kept same and so there is not a huge effect of outlier on the numerator of the t statistic. But in the denominator, we have a S term which is the pooled estimate of the common standard deviation. When we have a big outlier it increases the standard deviation and hence the t statistic becomes smaller and hence it becomes harder to reject than usual. So, we are getting many false negatives in the case where outliers are present.

In the above discussions we see that in terms of both level and power, our new test statistics are better than the pooled t-test. Not only that, it is also

significantly better than an existing robust test. In both the cases of level and power we see that by the time the contaminating value becomes sufficiently small (-2 for level and -3 for power) the pattern for the new test statistic change where as the trimmed tests and the pooled t test show a very steady pattern. These simulation results consolidate our claim that these new test statistics provide significantly more robust test statistics in the case where the data points are vulnerable of being contaminated.

6 The Multiple Sample Problem: ANOVA

6.1 Natural extension of T_2 to multi-sample

We have already seen that T_1 is better than T_2 in terms of both level and power, although the difference is marginal. But T_2 enjoys the advantage of being extendable to a multi-sample mean comparison scenario but T_1 does not.

Given samples of sizes $n_1, n_2, ..., n_k$ from k different normal populations, we consider, under the common variance assumption, the statistic

$$T_2 = 2n[\rho_O(g^*_{n_1}, g^*_{n_2}, ...g^*_{n_k}, \hat{\theta}_0) - \rho_O(g^*_{n_1}, g^*_{n_2}, ...g^*_{n_k}, \hat{\theta}_0)]$$

where each entry has its own obvious meaning to test the null hypothesis $H_0 : \mu_1 = \mu_2 = ... = \mu_k$. Here

$$\rho_O(g^*_{n_1}, g^*_{n_1}, ..., g^*_{n_1}, \theta) = \frac{n_1}{n}\rho_C(g^*_{n_1} f_{(\theta_0)_1}) + \frac{n_2}{n}\rho_C(g^*_{n_2} f_{(\theta_0)_2}) + ...\frac{n_k}{n}\rho_C(g^*_{n_k} f_{(\theta_0)_k}).$$

The test is expected to perform better than the classical test of Analysis of variance.

6.2 Proofs are similar

The proof of the null distribution of the ANOVA statistic T_2 has no new ideas or tricks involved beyond the two sample case. The number of groups are simply extended to $k(k > 2)$ from 2. The number of restrictions in the null hypothesis $H_0 : \mu_1 = \mu_2 = ... = \mu_k$ is $k - 1$. Accordingly, the statistic has a χ^2_{k-1} distribution under this scenario. We do not provide a separate proof in this case.

7 Real data examples

In this section we will be discussing a few real data examples. These data are carefully chosen to illustrate the impact of outliers in statistical inference. For

each of the following data, we will be discussing it's source and nature. We will first describe the data and will show the outliers in bold letters. These outliers are chosen subjectively. We will delete the outliers from one or both the datasets. That will give us four different comparable datasets. In these, we will look at the p-values of the two new test statistic and contrast it with that of the pooled two-sample t test. We will also present the $\hat{\theta}$, $\hat{\theta}_0$, $\hat{\theta}_{MLE}$, $\hat{\theta}_{0MLE}$, in a separate table for each of these datasets. In all these computations, we use Hellinger distance as our C function.

7.1 Two Sample comparison datasets

Example 1 (Lake Data): These data are obtained from Balakrishnan (1985[1]). The data consist of the pollution levels of two lakes. The data, presented in Table 5 represented the values of $10(x - 20)$ where x is the actual pollution level.

Lake 1	−1.48	1.25	−0.51	0.46	0.6	**−4.27**	0.63	−0.14
	−0.38	1.28	0.93	0.51	1.11	−0.17	−0.79	−1.02
	−0.91	0.1	0.41	1.11				
Lake 2	1.32	1.81	−0.54	2.68	2.27	2.7	0.78	**−4.62**
	1.88	0.86	2.86	0.47	−0.42	0.16	0.69	0.78
	1.72	1.57	2.14	1.62				

Table 5: Lake Data

From Table 6, we see that the p-values for both the new statistics are quite small in all the four cases including the different combinations of presence and absence of outliers. This will lead to solid rejections of the null hypothesis even at 1% level of significance for each of the four cases and each of the two new statistics. However, for the t-test we see that both in the first and the third cases the p-values are more than 1% and in the third case it even exceeds 5%. This shows that the presence (or absence) of the outliers have a significant effect on the decision of the t-test.

Table 7 shows that the removal of the first outlier brings the means closer and leads to a larger p-value for the pooled t-test. however, outlier deletion or retention has negligible effects on either T_1 or T_2.

Outliers	T_1	T_2	pooled t-test
Both Present	0.000308	0.0003	0.027525
Both Removed	0.00045	0.000399	0.000523
1st Removed	0.001774	1E-05	0.050257
2nd Removed	0.000396	0.000334	0.00075

Table 6: p-values for the Lake Data Example for the different tests

Outliers	$\hat{\mu_1}$	$\hat{\mu_2}$	$\hat{\sigma}^2$
Both present			
$\hat{\theta}$	0.104247	1.328859	1.055237
$\hat{\theta}_0$	0.65597	0.65597	1.211732
$\hat{\theta}_{MLE}$	−0.064	1.0365	2.18983
$\hat{\theta}_{0MLE}$	0.48625	0.48625	2.444219
Both removed			
$\hat{\theta}$	0.133052	1.336587	1.024374
$\hat{\theta}_0$	0.663216	0.663216	1.168162
$\hat{\theta}_{MLE}$	0.157368	1.334211	0.858668
$\hat{\theta}_{0MLE}$	0.745789	0.745789	1.191057
1st removed			
$\hat{\theta}$	0.164888	1.336876	1.282064
$\hat{\theta}_0$	0.673556	0.673556	1.165059
$\hat{\theta}_{MLE}$	0.157368	1.0365	1.744319
$\hat{\theta}_{0MLE}$	0.608205	0.608205	1.897962
2nd removed			
$\hat{\theta}$	0.100967	1.332965	1.08062
$\hat{\theta}_0$	0.646353	0.646353	1.212074
$\hat{\theta}_{MLE}$	−0.064	1.334211	1.33831
$\hat{\theta}_{0MLE}$	0.617179	0.617179	1.804794

Table 7: Parameter Estimates for the Lake Data Example

Example 2 (Ozone data): Doksum and Sievers (1976) describes a data from a study designed to assess the effects of ozone on weight gains in rats. The experimental group consisted of 22 seventy-day-old rats kept in an ozone environment for 7 days. A control group of 23 rats , of the same age, was kept in an ozone-free environment. The weight gains are reported.
The data: The two data are given below:

Case	**41**	**38.4**	24.4	25.9	21.9	18.3	13.1	27.3
	28.5	**-16.9**	26.0	17.4	21.8	15.4	27.4	19.2
	22.4	17.7	26	29.4	21.4	26.6	22.7	
Control	10.1	6.1	20.4	7.3	14.3	15.5	-9.9	6.8
	28.2	17.9	-9.0	-12.9	14.0	6.6	12.1	15.7
	39.9	-15.9	**54.6**	-14.7	**44.1**	-9.0		

Table 8: Ozone data

Analyses: These data are also showing very similar feature to what we have seen in the Lake data.

Example 3 (Mill Data): We already provided the data in the 'motivational example' of section 1. We now present the analyses of that data set in the following two tables:

Outliers	T_1	T_2	pooled t-test
Both present	0	1.19E-07	0.019307
Both removed	1.11E-16	2.05E-09	5.44E-06
1st removed	2.45E-14	3.04E-10	0.012983
2nd removed	2.42E-07	5.36E-07	6.58E-05

Table 9: p-values for the Ozone Data Example for the different tests

Outliers	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}^2$
Both present			
$\hat{\theta}$	367.1096	14.68201	81.55534
$\hat{\theta}_0$	18.72241	18.72241	196.3416
$\hat{\theta}_{MLE}$	22.40435	11.00909	236.1146
$\hat{\theta}_{0MLE}$	16.83333	16.83333	263.8086
Both removed			
$\hat{\theta}$	22.54936	7.307674	13.34944
$\hat{\theta}_0$	14.46383	14.46383	108.0243
$\hat{\theta}_{MLE}$	22.64	5.452632	96.9561
$\hat{\theta}_{0MLE}$	14.26667	14.26667	170.0449
1st removed			
$\hat{\theta}$	30.74687	11.49846	39.15131
$\hat{\theta}_0$	17.11528	17.11528	129.9109
$\hat{\theta}_{MLE}$	22.64	11.00909	199.5915
$\hat{\theta}_{0MLE}$	16.54762	16.54762	229.6845
2nd removed			
$\hat{\theta}$	24.47243	6.818259	100.875
$\hat{\theta}_0$	18.07431	18.07431	137.4113
$\hat{\theta}_{MLE}$	22.40435	5.452632	143.419
$\hat{\theta}_{0MLE}$	14.73571	14.73571	212.7053
p-value	2.42E-07	5.36E-07	6.58E-05

Table 10: Parameter Estimates for the Ozone Data Example

Outliers	T_1	T_2	pooled t-test
Both present	0.000877	0.007016	0.353921
Both removed	0.001689	0.008475	0.034902
1st removed	0.001795	0.009225	0.061691
1st present	0.000796	0.007068	0.748713

Table 11: p-values for the Mill Data Example

Analyses: From Table 14, we see that the p-values for T_1 and T_2 changes very negligibly whereas the p-value for the pooled t test varies a lot. It also can directly affect the statistical conclusion if we remain ignorant about the outliers.

Table 15 shows that the removal of the second outlier brings the means closer and leads to a larger p-value for the pooled t-test.

Data 4: Outliers	$\hat{\mu}_1$	$\hat{\mu_2}$	$\hat{\sigma^2}$
Both present			
$\hat{\theta}$	0.158693	0.608719	0.188302
$\hat{\theta}_0$	0.372624	0.372624	0.291829
$\hat{\theta}_{MLE}$	0.497727	0.883182	1.775425
$\hat{\theta}_{0MLE}$	0.690455	0.690455	1.772144
Both removed			
$\hat{\theta}$	0.163374	0.608531	0.193234
$\hat{\theta}_0$	0.373513	0.373513	0.288536
$\hat{\theta}_{MLE}$	0.187	0.590952	0.332813
$\hat{\theta}_{0MLE}$	0.393902	0.393902	0.366494
1st removed			
$\hat{\theta}$	0.166465	0.610305	0.199147
$\hat{\theta}_0$	0.381235	0.381235	0.29692
$\hat{\theta}_{MLE}$	0.187	0.883182	1.30843
$\hat{\theta}_{0MLE}$	0.551667	0.551667	1.402917
2nd removed			
$\hat{\theta}$	0.155031	0.606647	0.181721
$\hat{\theta}_0$	0.353248	0.353248	0.273156
$\hat{\theta}_{MLE}$	0.497727	0.590952	0.856046
$\hat{\theta}_{0MLE}$	0.543256	0.543256	0.83827

Table 12: Parameter Estimates for the Mill Data Example

7.2 ANOVA datasets

Example 1 (Newcomb Data): In 1882, the astronomer and mathematician Simon Newcomb, measured the time required for a light signal to pass from his laboratory on the Potomac River to a mirror at the base of the Washington Monument and back, a distance of 744373 meters. Table 13 contains these measurements from three samples, as deviations from 24800 nanoseconds. For example, for the first observation, 28 represents that the 24828 nanoseconds that were spent for the light to travel the required 744373 meters. The data comprises three samples, of sizes 20, 20 and 26, respectively, corresponding to three different days. These data have been analyzed previously by a number of authors including Stigler (1973[12])

Day 1	28	26	33	24	34	**−44**	27
	16	40	**−2**	29	22	24	21
	25	30	23	29	31	19	
Day 2	24	20	36	32	36	28	25
	21	28	29	47	25	28	26
	30	32	36	26	30	22	
Day 3	36	23	27	27	28	27	31
	27	26	33	26	32	32	24
	39	28	24	25	32	25	29
	27	28	29	**16**	23		

Table 13: Newcomb Data

Outliers	T_2	ANOVA F test
Both present	0.413907	0.080485
Both removed	0.932318	0.462579
1st removed	0.982654	0.545247
2nd removed	0.48357	0.067259

Table 14: p-values for Newcomb Data

New Comb	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\sigma}^2$
Both present				
$\hat{\theta}$	26.62393	38.79901	33.18425	40.76972
$\hat{\theta}_0$	26.85875	26.85875	26.85875	30.98443
$\hat{\theta}_{MLE}$	21.75	28.55	27.84615	110.3952
$\hat{\theta}_{0MLE}$	26.21212	26.21212	26.21212	115.462
Both removed				
$\hat{\theta}$	26.13828	27.1628	24.876	20.14234
$\hat{\theta}_0$	23.5525	23.5525	23.5525	21.1874
$\hat{\theta}_{MLE}$	26.72222	28.55	28.32	24.21739
$\hat{\theta}_{0MLE}$	27.93651	27.93651	27.93651	23.9959
1st removed				
$\hat{\theta}$	25.3099	30.32547	22.12487	35.862
$\hat{\theta}_0$	21.19399	21.19399	21.19399	27.89625
$\hat{\theta}_{MLE}$	26.72222	28.55	27.84615	26.20072
$\hat{\theta}_{0MLE}$	27.75	27.75	27.75	25.84127
3rd removed				
$\hat{\theta}$	35.5367	27.69195	36.27236	47.26743
$\hat{\theta}_0$	26.85872	26.85872	26.85872	30.98437
$\hat{\theta}_{MLE}$	21.75	28.55	28.32	109.7682
$\hat{\theta}_{0MLE}$	26.36923	26.36923	26.36923	115.6115

Table 15: parameter estimates for the Newcomb Data

Analyses The removal of 1st outlier is having an impact on the p value of the two statistics. This is quite natural as 16 is not that far as -44 and -2 are from the data cloud.

With both the outliers and with the first outlier, the value of the usual F-test statistic is on the border line of rejection if we take our level of significance to be 0.05. However, the test T_2 in all the four cases cannot reject the test with higher p-values.

Example 2 (Football Data): The following data from The Sports Encyclopedia Pro Football represent weights (pounds) of a random sample of professional football players on the five teams of Dallas.

Team 1	2569	2928	2865	3844	3027	2336	3211	3037
Team 2	2074	2885	3378	3906	2782	3018	3383	3447
Team 3	2505	2315	2667	2390	3021	3085	3308	3231
Team 4	2838	2351	3001	2439	2199	3318	3601	3291
Team 5	**1532**	2552	3083	2330	2079	3366	2416	3100

Table 16: ANOVA 2nd Data: Footballers' weight

Outlier	T_2	ANOVA F test
Present	0.9653	0.283826
Deleted	0.9733	0.539731

Table 17: p-value table for the Football Data

Football Data	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\mu}_4$	$\hat{\mu}_5$	$\hat{\sigma}^2$
Both Present						
$\hat{\theta}$	28.05358	31.9085	31.13095	36.93937	23.79538	40.8915
$\hat{\theta}_0$	28.96059	28.96059	28.96059	28.96059	28.96059	20.2013
$\hat{\theta}_{MLE}$	29.77125	31.09125	28.1525	28.7975	25.5725	25.80014
$\hat{\theta}_{0MLE}$	28.677	28.677	28.677	28.677	28.677	26.63162
p-value	0.9653	0.283826				
5th removed						
$\hat{\theta}$	28.89423	35.87701	33.36418	22.53566	35.69094	43.26009
$\hat{\theta}_0$	28.83519	28.83519	28.83519	28.83519	28.83519	11.87717
$\hat{\theta}_{MLE}$	29.77125	31.09125	28.1525	28.7975	27.03714	23.02535
$\hat{\theta}_{0MLE}$	29.01949	29.01949	29.01949	29.01949	29.01949	22.51708
p-value	0.9733	0.539731				

Table 18: Parameter estimates for the Football Data

Analyses: We divide the original data by 100 and perform our analyses. We consider the last data point of the first column as the only outlier. We see that the usual F-test is having a drastic change in the p-value depending on whether the outlier is there or not. But T_2 is much more 'robust' in that decision. Additionally, it can be said that T_2 infers that the equality of two mean with more confidence than what is done by the usual F-test.

References

[1] Balakrishnan, N. and Tiku, M. L. (1985). *Robust Univariate Two- Way Classification* Biometrica Journal, **27** 123-138.

[2] Basu, A., Shioya, H., Park, C. (2011). *Statistical Inference: The Minimum Distance Approach*, Chapman & Hall/CRC, Boca Raton, Florida.

[3] Doksum, K. A., Sievers, G. L. (1976) Plotting With Confidence: Graphical Comparisons Of Two Populations, *Biometrika*, **63**, 421-434.

[4] Hampel, F. R., Ronchetti, E. , Rousseeuw, P. J. and Stahel, W. (1986). *Robust Statistics: The Approach Based on Influence Functions.* New York, USA: John Wiley & Sons.

[5] Harris, I. R. and Basu, A. (1994). Hellinger distance as a penalized log likelihood. *Communications in Statistics: Simulation and Computation*, **23**, 1097-1113.

[6] Huber, P. J. (1981). *Robust Statistics.* John Wiley & Sons.

[7] Koopmans, L. H. (1987). *Introduction to Contemporary Statistical Methods*, 2nd Ed. , Boston, Duxbury.

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

[9] Sarkar, S. and Basu, A. (1995). On Disparity Based Robust Tests For Two Discrete Populations.*Sankhya: The Indian Journal of Statistics*, **57B**, Pt. 3 353-364.

[10] Simpson D. G. (1989). Hellinger Deviance Tests: Efficiency, Breakdown Points, and Examples. *Journal of the American Statistical Association*, **84**, 107-113.

[11] Staudte, Robert G., Sheather, Simon J. (1990). *Robust Estimation and Testing* John Wiley & Sons, Inc., USA.

[12] Stigler, S. M. (1973). "Simon Newcomb, Percy Daniel and the history of robust estimation 1885-1920". *Journal of the American Statistical Association*, **68**, 872879.

[13] Tamura, R. N. and D. D. Boos (1986). Minimum Hellinger distance estimation for multivariate location and covariance. *Journal of the American Statistical Association*, **81**, 223-229.

[14] Tiku, M. L., Tan, W. Y., Balakrishnan, N., (1986). *Robust Inference* Dekker M. .

[15] Wilcox, Rand R. (2005). *Introduction to Robust Estimation and Hypothesis Testing*, Academic Press.