Tests of Hypotheses in Multiple Samples based on Penalized Disparities

Chanseok Park[1], Ayanendranath Basu[2] and Ian R. Harris[3]

Abstract

Robust analogues of the likelihood ratio test are considered for testing of hypotheses involving multiple discrete distributions. The test statistics are generalizations of the Hellinger deviance test of Simpson (1989) and disparity tests of Lindsay (1994), obtained by looking at a 'penalized' version of the distances; Harris and Basu (1994) suggest that the penalty be based on reweighting the empty cells. The results show that often the tests based on the ordinary and penalized distances enjoy better robustness properties than the likelihood ratio test. Also, the tests based on the penalized distances are improvements over those based on the ordinary distances in that they are much closer to the likelihood ratio tests at the null and their convergence to the χ^2 distribution appears to be dramatically faster; extensive simulation results show that the improvement in performance of the tests due to the penalty is often substantial in small samples.

Keywords: robustness, likelihood ratio test, blended weight Hellinger distance, overall disparity.

[1] Department of Statistics, Pennsylvania State University, University Park, PA 16802, U.S.A.

[2] Applied Statistics Unit, Indian Statistical Institute, Calcutta 700 035, India. He is currently visiting Department of Statistics, Pennsylvania State University.

[3] Department of Mathematics and Statistics, Northern Arizona University, Flagstaff, AZ 86011, USA.

1. Introduction

Consider a discrete parametric model with density $m_{\boldsymbol{\beta}}(x)$, indexed by an unknown $\boldsymbol{\beta} \in \mathbb{R}^p$. To test hypotheses of interest about the unknown parameter, Simpson (1989) proposed the Hellinger deviance test. Lindsay (1994) discussed the class of 'disparity' tests; disparities form a particular subclass of density based distances between the data and the parametric model. Several of the disparity tests, including the one based on the Hellinger distance enjoy much better robustness properties than the likelihood ratio test while being asymptotically equivalent to the latter when the null hypothesis is true. These tests can be extended quite easily to the case when random samples are available for two or more populations, each indexed by an unknown parameter vector, and one wishes to perform tests of hypotheses involving the parameters of the different populations. Simpson (1989) considered such an extension for the special case of the Hellinger distance; the general case of disparity tests was considered by Sarkar and Basu (1995).

While for most parametric models the convergence of the likelihood ratio test (LRT) statistic to the appropriate asymptotic χ^2 limit is relatively quick, for many of the popular disparities like the Hellinger distance the assumed χ^2 distribution may be a very poor approximation to the true null distribution of the disparity tests in small samples. For the Hellinger deviance test, this can be seen in the numbers reported by Simpson (1989, Table 3) for the Poisson model. Harris and Basu (1994) propose a penalized Hellinger distance obtained by reweighting the empty cells. Basu, Harris and Basu (1996) examine this distance in the context of hypothesis testing for a single population. In this paper we apply the penalized distances for testing of hypotheses in multiple populations; our results indicate that the penalty can improve the performance of the tests in a general class of disparities including the Hellinger distance.

The rest of the paper is organized as follows. Section 2 provides a review of the penalized disparities. The testing procedures in multiple samples based on the penalized disparities are introduced in Section 3. Section 4 presents an extensive empirical study to illustrate the performance of these tests in some discrete models. Section 5 presents concluding remarks.

2. Minimum Disparity Estimation and the Impact of Empty Cells

Consider a parametric family with countable support and density $m_{\boldsymbol{\beta}}(x)$, $\boldsymbol{\beta} \in \mathbb{R}^p$. For simplicity of presentation, we discuss our results in terms of a particular subfamily of disparities, the blended weight Hellinger distances (BWHD); our numerical results of Section 4 will also be based on the BWHD family. However the particular asymptotic results given below and in Section 3 would also hold for the general class of disparities.

Assume that we have a sample of size n from the true distribution and let $d(x)$ be the proportion of sample observations at the value x. The BWHD is a measure of discrepancy between d and $m_{\boldsymbol{\beta}}$, which can be written as a function of a parameter $\alpha \in [0,1]$ as:

$$\mathrm{BWHD}_\alpha(d, m_{\boldsymbol{\beta}}) = \frac{1}{2}\sum_x \left\{ \frac{d(x) - m_{\boldsymbol{\beta}}(x)}{\alpha\sqrt{d(x)} + \bar{\alpha}\sqrt{m_{\boldsymbol{\beta}}(x)}} \right\}^2, \ \bar{\alpha} = 1 - \alpha. \tag{2.1}$$

Note that for $\alpha = 0.5$ one gets $2\sum \left\{\sqrt{d(x)} - \sqrt{m_{\boldsymbol{\beta}}(x)}\right\}^2$, which is twice the squared Hellinger distance.

Given a particular model and a sample of size n, one can obtain estimates of the unknown parameter $\boldsymbol{\beta}$ by minimizing some member of the BWHD family. For $\alpha \in [1/3, 1]$, the estimates obtained by minimizing the above disparity enjoy better robustness properties compared to the maximum likelihood estimator (see Lindsay 1994); the latter is a minimizer of the likelihood disparity (LD),

$$\mathrm{LD}(d, m_{\boldsymbol{\beta}}) = \sum_x d(x) \log\big(d(x)/m_{\boldsymbol{\beta}}(x)\big). \tag{2.2}$$

The estimating equation of the maximum likelihood estimator, obtained by equating the derivative of the last equation with respect to $\boldsymbol{\beta}$ to zero, has the form

$$\sum_x \delta(x) \nabla m_{\boldsymbol{\beta}}(x) = 0,$$

where ∇ represents the derivative with respect to $\boldsymbol{\beta}$ and $\delta(x) = d(x)/m_{\boldsymbol{\beta}}(x) - 1$ is a standardized form of the residual which has been called the 'Pearson residual' in Lindsay (1994). The estimating equation of any other minimum disparity estimator has a similar form given by

$$\sum_x A(\delta(x)) \nabla m_{\boldsymbol{\beta}}(x) = 0,$$

where the function $A(\cdot)$ is specific to the particular disparity. This function can be centered and rescaled, without changing the estimating properties of the corresponding estimators, so that $A(0) = 0$ and $A'(0) = 1$. This is the reason why one considers the factor of 1/2 in equation (2.1), or twice squared Hellinger distance, rather than the squared Hellinger distance itself. Henceforth we will implicitly mean twice the squared distance when we refer to the Hellinger distance (HD). The centered and rescaled function A is called the residual adjustment function of the disparity.

Harris and Basu (1994) observed that one can generate a family of disparities through an adjustment of the weight of the empty cells in the HD. This family, called the family of penalized Hellinger distances (PHD) by Harris and Basu, has the form:

$$2 \sum_{d(x)\neq 0} \left\{ \sqrt{d(x)} - \sqrt{m_{\boldsymbol{\beta}}(x)} \right\}^2 + h \sum_{d(x)=0} m_{\boldsymbol{\beta}}(x). \tag{2.3}$$

Substituting $h = 2$ in the above equation returns the ordinary HD, and other members correspond to other values of h. In particular Harris and Basu noticed that $h = 1$ provided a member of the PHD family which was an attractive alternative to the ordinary HD. The choice of $h = 1$ makes the residual adjustment function of the PHD equal to that of the likelihood disparity, *i.e.* this reweights the empty cells so as to treat them in the same manner as the maximum likelihood estimator.

In the general case we can define the penalized blended weight Hellinger distance (PBWHD) as a function of α and h as:

$$\mathrm{PBWHD}_{\alpha,h}(d, m_{\boldsymbol{\beta}}) = \frac{1}{2} \sum_{d(x)\neq 0} \left\{ \frac{d(x) - m_{\boldsymbol{\beta}}(x)}{\alpha\sqrt{d(x)} + \bar{\alpha}\sqrt{m_{\boldsymbol{\beta}}(x)}} \right\}^2 + h \sum_{d(x)=0} m_{\boldsymbol{\beta}}(x), \tag{2.4}$$

where $\bar{\alpha} = 1 - \alpha$. In particular we will focus on $\mathrm{PBWHD}_{\alpha,1}$, for the same reason as in the case of the PHD. This reweighting of the empty cells can be particularly useful for disparities with large values of α (in which we are more interested for robustness purposes anyway), as the weight assigned to the empty cells increases with the value of α. In the extreme case of $\alpha = 1$ (this distance is also known as the Neyman's modified chi-square) the BWHD is not even defined if any of the cells are empty. The reweighting provides a simple solution to this problem.

Now we consider tests of hypotheses based on the penalized distances. Let

Ω_0 be a proper subset of $\mathbb{R}^p$, and suppose that the hypothesis

$$H_0 : \boldsymbol{\beta} \in \Omega_0, \tag{2.5}$$

is of interest. Let $\boldsymbol{\beta}_M^0$ and $\hat{\boldsymbol{\beta}}_M$ be the maximum likelihood estimators of $\boldsymbol{\beta}$ under the null hypothesis and without any restrictions respectively. It is well known that

$$2n\left\{\mathrm{LD}(d, m_{\boldsymbol{\beta}_M^0}) - \mathrm{LD}(d, m_{\hat{\boldsymbol{\beta}}_M})\right\}, \tag{2.6}$$

which equals negative of twice log likelihood ratio, has an asymptotic χ^2 distribution with degrees of freedom equal to the number of independent restrictions provided by the null hypothesis. Alternatively, consider the disparity test statistic generated by the $\mathrm{PBWHD}_{\alpha,h}$, given by

$$2n\left\{\mathrm{PBWHD}_{\alpha,h}(d, m_{\boldsymbol{\beta}_{\alpha,h}^0}) - \mathrm{PBWHD}_{\alpha,h}(d, m_{\hat{\boldsymbol{\beta}}_{\alpha,h}})\right\}, \tag{2.7}$$

where $\boldsymbol{\beta}_{\alpha,h}^0$ and $\hat{\boldsymbol{\beta}}_{\alpha,h}$ are the estimators obtained by minimizing the penalized distances in (2.4) under the null and without any restriction respectively. When $\alpha = 0.5$ and $h = 2$ this is the ordinary Hellinger deviance test statistic (Simpson 1989) and has the same asymptotic χ^2 distribution as the LRT statistic in (2.6) under the null. Since the penalty only reweights the empty cells, it follows from Lindsay (1994) that the other members of the family of tests defined in (2.7) also have the same asymptotic distribution as the LRT.

Lindsay (1994) shows that the robustness of the minimum disparity estimators as well as the disparity tests are related to the "estimation curvature of the disparity," defined as $A''(0)$, the second derivative of the residual adjustment function at zero. Since altering the weights of the empty cells (corresponding to $\delta = -1$) does not affect this curvature, the robustness of the tests based on (2.4) are expected to be minimally affected by the value of h in large samples.

Notice that these penalized disparities are also useful for testing goodness-of-fit. While in this paper we concentrated on robust tests of parametric hypothesis and not on goodness-of-fit, we mention the relevant result which will be utilized on a sequel paper dealing with the latter topic. Since the multinomial goodness-of-fit problems where we will be looking at the sample space is finite, and the total probability of the empty cells asymptotically go to zero, the following theorems are simple extensions of the results of Basu and Sarkar (1994).

For a sequence of n observations on a multinomial distribution with probability vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)$ and $\sum_{i=1}^{k} \pi_i = 1$. Let $\mathbf{x} = (x_1, \ldots, x_k)$ denote the vector of observed frequencies for k categories. Let $\mathbf{d} = \mathbf{x}/n = (x_1/n, \ldots, x_k/n)$ and $\boldsymbol{\pi}_0 = (\pi_{01}, \ldots, \pi_{0k})$ be a probability vectors with $\pi_{0i} > 0$ for each i and $\sum_{i=1}^{k} \pi_{0i} = 1$.

Theorem 2.1. *Under the simple null hypothesis $H_0 : \boldsymbol{\pi} = \boldsymbol{\pi}_0$, the test statistic $D_{\alpha,h} = 2n\mathrm{PBWHD}_{\alpha,h}(\mathbf{d}, \boldsymbol{\pi})$ has an asymptotic $\chi^2(k-1)$ distribution.*

Proof: Consider the multinomial random variable $\mathbf{X} = (X_1, \ldots, X_k)$ for which $\mathbf{x} = (x_1, \ldots, x_k)$ is an observed realization. Basu and Sarkar (1994, Theorem 3.1) showed that $2n\mathrm{BWHD}_\alpha(\mathbf{d}, \boldsymbol{\pi}) \xrightarrow{d} \chi^2(k-1)$ under the null. Hence it suffices to show that under the null

$$D_n = 2n\big\{\mathrm{BWHD}_\alpha(\mathbf{d}, \boldsymbol{\pi}) - \mathrm{PBWHD}_{\alpha,h}(\mathbf{d}, \boldsymbol{\pi})\big\} = o_p(1).$$

Since PBWHD and BWHD are different only at the empty cells, we have

$$D_n = nK_\alpha \sum_{i=1}^{k} \pi_{0i} I_i,$$

where $K_\alpha = 2\big|1/\{2(1-\alpha)^2\} - h\big|$ and $I_i = I(X_i = 0)$, the indicator function of the event $(X_i = 0)$. Now we have only to show $E(D_n) \to 0$ and $\mathrm{Var}(D_n) \to 0$ as

$n \to 0$. Since $E(I_i) = P(X_i = 0) = (1 - \pi_{0i})^n$, we have

$$E(D_n) = nK_\alpha \sum_{i=1}^{k} \pi_{0i} E(I_i) = nK_\alpha \sum_{i=1}^{k} \pi_{0i}(1 - \pi_{0i})^n.$$

Since $n^a r^n \to 0$ for $|r| < 0$ and k is finite, we have $nE(I_i) \to 0$ and so $E(D_n) \to 0$. Also

$$\begin{aligned}\mathrm{Var}(D_n) &= (nK_\alpha)^2 \Big\{ \sum_{i=1}^{k} \pi_{0i}^2 \mathrm{Var}(I_i) + 2 \sum_{i<j} \pi_{0i}\pi_{0j} \mathrm{Cov}(I_i, I_j) \Big\} \\ &\leq (nK_\alpha)^2 \Big\{ \sum_{i=1}^{k} E(I_i) + 2 \sum_{i<j} E(I_i) \Big\}.\end{aligned}$$

Since $n^2 E(I_i) \to 0$ and k is finite, $\mathrm{Var}(D_n) \to 0$ as $n \to \infty$. □

Theorem 2.2. *Let* $\Delta_k = \{\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k) \bigm| \pi_i > 0,\ i = 1, 2, \ldots, k;\ \sum_{i=1}^{k} \pi_i = 1\}$. *Define a parameter vector* $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_s) \in \mathbb{R}^s$, $s \leq k - 1$, *and the mapping* $f : \mathbb{R}^s \longrightarrow \Delta_k$ *such that for each parameter vector* $\boldsymbol{\theta}$ *there corresponds a probability vector* $\boldsymbol{\pi}$. *The hypothesis*

$$H_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0 \quad \textit{and} \quad H_0 : \boldsymbol{\pi} \in \boldsymbol{\Pi}_0, \tag{2.8}$$

are then equivalent when $\boldsymbol{\Pi_0} = f(\boldsymbol{\Theta}_0)$. *Suppose that the regularity conditions of Section* 4 *in* Basu and Sarkar (1994) *hold and* $\hat{\boldsymbol{\theta}}$ *is any BAN estimator of* $\boldsymbol{\theta}$ *and* $\hat{\boldsymbol{\pi}} = f(\hat{\boldsymbol{\theta}})$. *Then under* (2.8), $2n\mathrm{PBWHD}_{\alpha,h}(\mathbf{d}, \hat{\boldsymbol{\pi}})$ *converges in distribution to a* $\chi^2(k - s - 1)$ *random variable as* $n \to \infty$.

Proof: Basu and Sarkar (1994, Theorem 4.3) showed that $2n\mathrm{BWHD}_\alpha(\mathbf{d}, \hat{\boldsymbol{\pi}}) \xrightarrow{d} \chi^2(k - s - 1)$. Hence it suffices to show

$$D_n = 2n\big\{\mathrm{BWHD}_\alpha(\mathbf{d}, \hat{\boldsymbol{\pi}}) - \mathrm{PBWHD}_{\alpha,h}(\mathbf{d}, \hat{\boldsymbol{\pi}})\big\} = o_p(1).$$

Notice that I_i and $\hat{\boldsymbol{\pi}}$ are both non-negative random variables. We have

$$D_n = nK_\alpha \sum_{i=1} \hat{\pi}_{0i} I_i.$$

Notice that $E(\hat{\pi}_{0i} I_i)$, $E(\hat{\pi}_{0i}^2 I_i)$ and $E(\hat{\pi}_{0i}\hat{\pi}_{0j} I_i I_j)$ are all bounded above by $E(I_i)$. Hence applying the technique of the proof of Theorem 2.1, we have $E(D_n) \to 0$ and $\mathrm{Var}(D_n) \to 0$ as $n \to \infty$. □

3. Tests of Hypotheses in Multiple Samples

For simplicity of presentation, we discuss the two population case; the results are true for k populations, where $k \geq 2$. Suppose that random samples of size n_i are available from the population with density $m_{\boldsymbol{\beta}_i}(x)$, and let $d_i(x)$ be the proportion of sample observation with value x in the i-th sample, $i = 1, 2$. Let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^t, \boldsymbol{\beta}_2^t)^t \in \mathbb{R}^{2p}$ and $n = n_1 + n_2$. We will assume that $n_1/n \to a \in (0, 1)$, so that neither sample size asymptotically dominates the other. The hypothesis to be tested is:

$$H_0 : \boldsymbol{\beta} \in \Omega_0,$$

where Ω_0 is a proper subset of $\mathbb{R}^{2p}$.

Let ρ be an appropriate disparity and construct $\rho(d_i, m_{\boldsymbol{\beta}_i}), i = 1, 2$. Define the overall disparity $\rho_O(\boldsymbol{\beta})$ for the two samples taken together as

$$\rho_O(\boldsymbol{\beta}) = \frac{1}{n}\{n_1 \rho(d_1, m_{\boldsymbol{\beta}_1}) + n_2 \rho(d_2, m_{\boldsymbol{\beta}_2})\}.$$

Then the disparity test statistic for the hypothesis H_0 is given by the test statistic $2n\{\rho_O(\boldsymbol{\beta}^0) - \rho_O(\hat{\boldsymbol{\beta}})\}$ which has an asymptotic $\chi^2(r)$ distribution under the null hypotheses H_0 where r is the number of independent restrictions imposed by the null hypothesis, and $\boldsymbol{\beta}^0$ and $\hat{\boldsymbol{\beta}}$ are the minimizers of the overall disparities under the null hypothesis and without any restriction respectively (Simpson 1989; Sarkar and Basu 1995). For the likelihood disparity of (2.2), this statistic reduces to the LRT statistic.

4. Numerical Results

4.1. Simulation Results

In this section we present selected results from an extensive numerical study to compare the performance of the tests resulting from the penalized distances to those generated by the ordinary distances. The results presented in this section are for the Poisson model. However, numerical results obtained at the geometric model, not presented here, indicate that similar results hold in that model as well. All the simulation results presented in this paper are based on 5000 replications.

The first study involves random samples of sizes n_1 and n_2 drawn from two Poisson populations, with means $\beta_1 = \beta_2 = 5$. The tests studied are the LRT, the HD test, and the PHD test for different combinations of (n_1, n_2) values in testing the hypothesis $H_0 : \beta_1 = \beta_2$. The observed level for each testing procedure is determined as the proportion of test statistics exceeding the critical value of the assumed $\chi^2(1)$ limit. Hence, given a probability estimate $\hat{p}$ its estimated standard deviation may be computed as $[\hat{p}(1-\hat{p})/5000]^{1/2}$ (assuming binomial rejection frequencies) which can be no greater than $[0.5 \times 0.5/5000]^{1/2} \simeq 0.007$. The results, displayed in Table 4.1, show, as expected, that the levels of the likelihood ratio converge to the nominal levels faster than the other two test statistics. However, for each of the three nominal levels considered (results corresponding to nominal level 0.01, also computed in our simulations, have not been reported for brevity),

Table 4.1: Levels of the LRT, HD test and PHD test. Both populations are Poisson(5).

Nominal level	0.1			0.05		
(n_1, n_2)	LRT	HD	PHD	LRT	HD	PHD
(25, 25)	0.096	0.142	0.081	0.048	0.078	0.037
(25, 50)	0.105	0.148	0.092	0.052	0.080	0.041
(50, 50)	0.099	0.123	0.088	0.050	0.068	0.045
(50, 100)	0.104	0.141	0.100	0.055	0.073	0.051
(100, 100)	0.101	0.114	0.095	0.053	0.066	0.052

Table 4.2: Levels of the LRT, HD test and PHD test under contamination. Population 1 is Poisson(5). Population 2 is 0.9Poisson(5) + 0.1Poisson(15).

Nominal level	0.1			0.05		
(n_1, n_2)	LRT	HD	PHD	LRT	HD	PHD
(25, 25)	0.452	0.179	0.118	0.351	0.112	0.061
(25, 50)	0.515	0.175	0.131	0.420	0.108	0.067
(50, 50)	0.644	0.196	0.157	0.551	0.121	0.086
(50, 100)	0.719	0.201	0.177	0.635	0.129	0.108
(100, 100)	0.844	0.259	0.227	0.785	0.179	0.138

the observed levels of the PHD tests are closer to those of the LRTs compared to the HD test. The HD test apparently requires considerably large sample sizes for its observed level to be reasonably close to the nominal level, which was observed by Simpson (1989). To make the above observations graphically explicit we can choose $n_1 = n_2 = n$ and plot the observed levels of the three tests on the same graph. This is displayed in Figure 4.1(a) over a fine grid of values for n. It is clear that the PHD test follows the LRT quite closely, whereas the convergence of the observed levels of the HD test is considerably slower. Figure 4.1(b) provides the difference of the other two tests against the LRT. The difference between the PHD test and the LRT stays much closer to zero.

Table 4.2 presents the results of a similar study, where population 2 is now

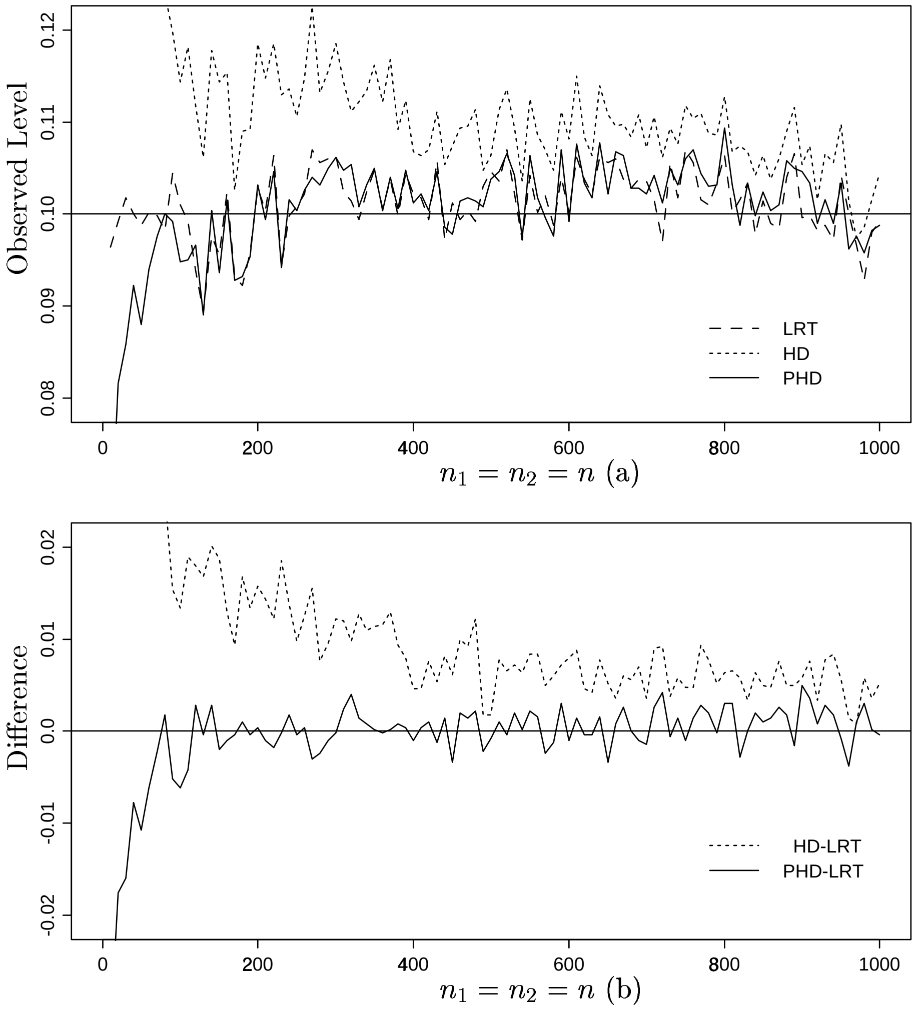

Figure 4.1: (a) Convergence of the observed levels of the LRT, the HD test, and the PHD test to the nominal level 10%. (b) Difference of the observed levels against the LRT.

the mixture 0.9Poisson(5)+0.1Poisson(15), where the second, smaller, component is considered to be a contamination. While the minimum disparity estimators are robust to data contamination, they are not invariant under it; therefore the asymptotic limit of a robust minimum disparity estimator, when data are generated from the above contaminated mixture, will be different from the target

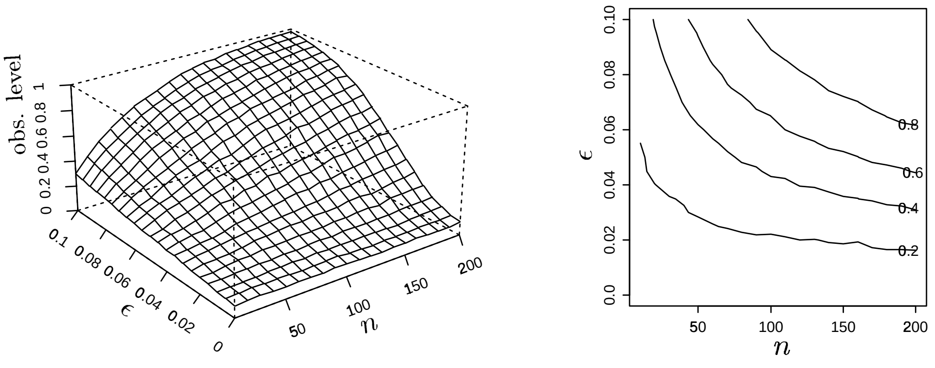

Figure 4.2: Observed levels of the LRT under contamination. Nominal level 10%.

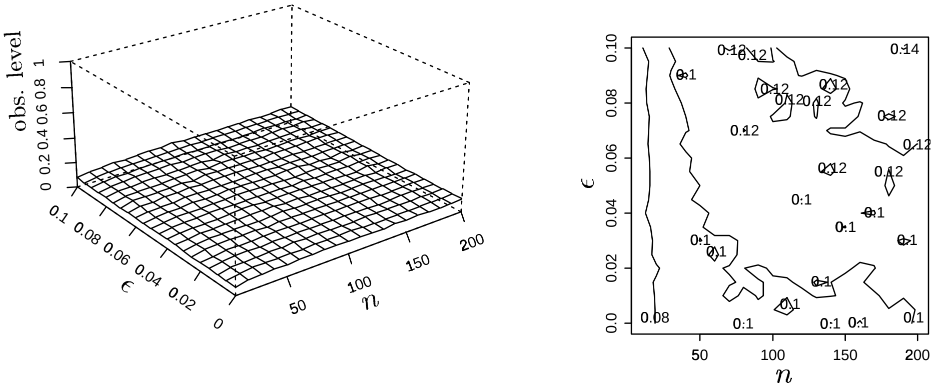

Figure 4.3: Observed levels of the PHD test under contamination. Nominal level 10%.

value of 5, although it will probably be closer to the target value than the mean of the maximum likelihood estimator, which in this case is 6. Thus as the sample size increases, the observed levels of the tests (which, strictly speaking, are observed powers) should approach 1. Suppose, however, we want to test whether the mean of the first population is equal to the mean of the larger component in the second population. That is, we are treating the smaller component in the second population as a contaminant rather than as an inherent part of the distribution generating the data. In this case one might want to construct a test which has a probability of rejection close to the nominal levels when the mean of the first population is actually equal to the mean of the larger component of the second population. By this criterion, the HD and PHD tests are superior to the LRT. The observed difference between the first two tests is partly due to their true critical values being different in small samples when the null hypothesis is true (the HD test having a larger critical value). The robustness of the PHD test compared to the LRT in this situation is graphically represented in Figures 4.2 and 4.3, using several values of the sample sizes with $n_1 = n_2 = n$. In this case the second sample is drawn from the $(1-\epsilon)$Poisson(5) + ϵPoisson(15) mixture, $\epsilon \in [0, 0.1]$. The three dimensional plot in Figure 4.2 provides the observed levels (at nominal level 10%) of the LRT statistics; as ϵ increases, the observed levels of the test statistics increase quickly, as evidenced by the sharp rise in the observed level around the far corner of the cube. In comparison, the surface representing the levels of the PHD test (Figure 4.3) is much flatter and closer to the nominal level. The same effect is visible in the contour plots corresponding to the three dimensional graphics.

Next we study the power of these tests. Table 4.3 presents the power of these tests when the first sample is drawn from Poisson(10) and the second one from Poisson(11). Note that a large part of the difference in the observed power of the three procedures in small samples is due to the fact that the actual critical values for the tests can be quite different from the χ^2 critical values. In particular, the HD has too inaccurate a level in small samples; its true critical values are often substantially higher than the χ^2 critical values (see Table 4.1).

Table 4.3: Powers of the LRT, HD test, and PHD test. Population 1 is Poisson(10). Population 2 is Poisson(11).

Nominal level	0.1			0.05		
(n_1, n_2)	LRT	HD	PHD	LRT	HD	PHD
(25, 25)	0.300	0.332	0.238	0.197	0.237	0.144
(25, 50)	0.356	0.404	0.305	0.243	0.298	0.194
(50, 50)	0.464	0.483	0.415	0.343	0.369	0.283
(50, 100)	0.562	0.608	0.530	0.435	0.491	0.397
(100, 100)	0.699	0.709	0.675	0.586	0.600	0.555

Table 4.4: Powers of the LRT, HD test, and PHD test under contamination. Population 1 is Poisson(10). Population 2 is $0.9\text{Poisson}(11) + 0.1\text{Poisson}(1)$.

Nominal level	0.1			0.05		
(n_1, n_2)	LRT	HD	PHD	LRT	HD	PHD
(25, 25)	0.182	0.287	0.184	0.109	0.196	0.102
(25, 50)	0.151	0.322	0.222	0.089	0.223	0.133
(50, 50)	0.176	0.381	0.307	0.108	0.281	0.204
(50, 100)	0.157	0.443	0.381	0.094	0.341	0.268
(100, 100)	0.171	0.530	0.496	0.104	0.421	0.384

Table 4.4 displays the results where the observations in the second sample come from $0.9\text{Poisson}(11) + 0.1\text{Poisson}(1)$. The presence of the second component as a contamination causes the LRT to lose power rapidly, whereas the other two tests hold their power much better. These observations are illustrated in Figures 4.4 and 4.5 using $n_1 = n_2 = n$. A look at the ϵ edge of the cubes shows that as ϵ increases, the power of the LRT sharply falls, whereas the PHD test shows a much smaller loss in power. This effect is also seen in the contour plots.

In Tables 4.5 and 4.6 we study the effect of the penalty on general members of the BWHD family for several values of α. The results show that as α moves towards 1, the penalty seems to provide a bigger improvement in the observed small sample levels of the tests. Expectedly, the penalty does not improve the

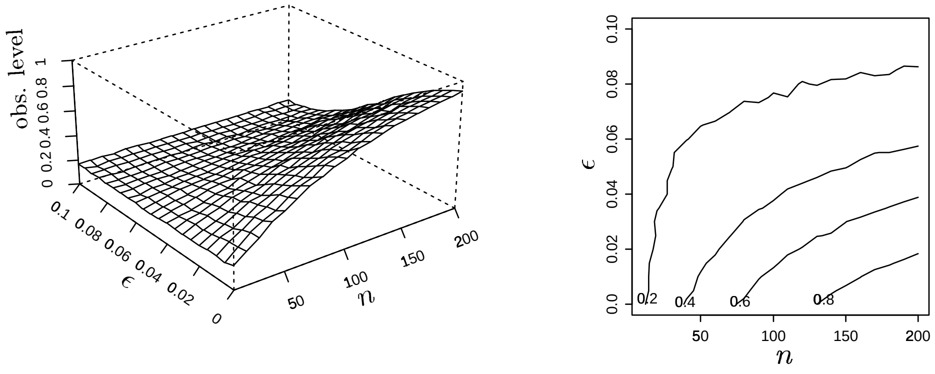

Figure 4.4: Observed powers of the LRT under contamination. Nominal level 10%.

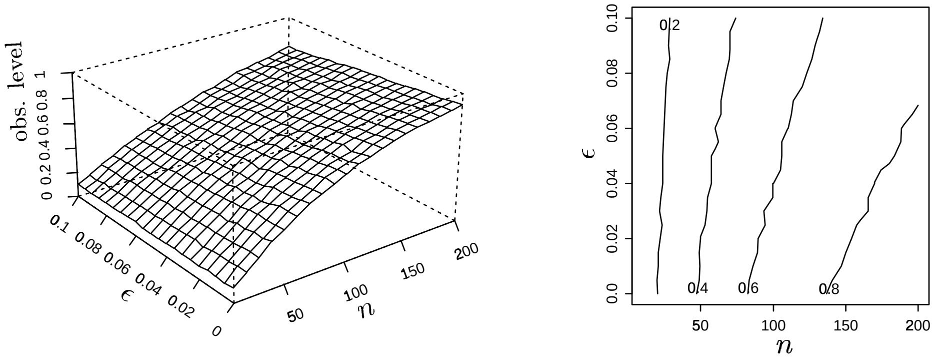

Figure 4.5: Observed powers of the PHD test under contamination. Nominal level 10%.

situation at $\alpha = 0$. Simulations were performed (Tables 4.7 and 4.8) to determine the powers of these tests at $\beta_1 = 10$ and $\beta_2 = 11$ (as in Table 4.3) and when population 2 is 0.9Poisson(11) + 0.1Poisson(1) (as in Table 4.4). The findings are similar to the results in Tables 4.3 and 4.4. The high power for some of the ordinary BWHD tests in Tables 4.7 and 4.8 are the consequences of the true critical values of the test being severely underestimated by the χ^2 approximation.

When they are converted to more legitimate level α tests (via the penalty), $\alpha = 1/3$ performs the best in terms of power (except for $\alpha = 0$, for which the empty cell correction fails to provide any improvement making the apparently high power values meaningless). The robust tests perform well in retaining their powers under contamination (compare Tables 4.7 and 4.8). For $\alpha \geq 1/3$, the actual distances generate higher observed powers compared to the penalized distances, mainly because the robust tests based on the actual distances have levels which are too inaccurate in small samples.

In general, the comparison between values of α may be summarized as follows. It appears that if there is no contamination, the tail of the distribution is best approximated by a χ^2 distribution when α is close to $1/3$. As α moves away in either direction, the level gets inflated — which can be largely corrected by the empty cell penalty for larger values of α, but not for smaller values of α. Thus tests with penalized distances with α in the range $[1/3, 1)$ appear to be the meaningful set for our purpose. Values of α close to $1/3$ seem to give the best power for pure data, while larger values are better at preserving the power for contaminated data. On the whole we believe values of α close to 0.5 can give the best compromise, and feel that the PHD test can be a very attractive practical data analysis tool.

Table 4.5: Levels of the disparity tests for several members of the BWHD and PBWHD family. Both populations are Poisson(5).

Nominal level		0.1		0.05	
(n_1, n_2)	α	BWHD	PBWHD	BWHD	PBWHD
(25, 25)	0	0.231	0.251	0.167	0.179
	1/3	0.100	0.093	0.051	0.047
	0.5	0.142	0.081	0.078	0.037
	0.7	0.358	0.086	0.268	0.039
	0.9	0.791	0.117	0.752	0.061
(25, 50)	0	0.235	0.248	0.163	0.177
	1/3	0.109	0.104	0.054	0.049
	0.5	0.148	0.092	0.080	0.041
	0.7	0.342	0.094	0.261	0.044
	0.9	0.772	0.126	0.727	0.065
(50, 50)	0	0.217	0.230	0.149	0.158
	1/3	0.100	0.097	0.051	0.050
	0.5	0.123	0.088	0.068	0.045
	0.7	0.297	0.102	0.217	0.055
	0.9	0.751	0.144	0.705	0.090
(50, 100)	0	0.219	0.232	0.148	0.158
	1/3	0.109	0.104	0.058	0.055
	0.5	0.141	0.100	0.073	0.051
	0.7	0.304	0.118	0.220	0.061
	0.9	0.747	0.157	0.697	0.096
(100, 100)	0	0.200	0.210	0.140	0.147
	1/3	0.101	0.100	0.055	0.052
	0.5	0.114	0.095	0.066	0.052
	0.7	0.231	0.110	0.156	0.057
	0.9	0.716	0.150	0.667	0.082

Table 4.6: Levels of the disparity tests for several members of the BWHD and PBWHD family under contamination. Population 1 is Poisson(5). Population 2 is 0.9Poisson(5) + 0.1Poisson(15).

Nominal level		0.1		0.05	
(n_1, n_2)	α	BWHD	PBWHD	BWHD	PBWHD
(25, 25)	0	0.862	0.868	0.838	0.845
	1/3	0.214	0.210	0.134	0.131
	0.5	0.179	0.118	0.112	0.061
	0.7	0.373	0.105	0.301	0.053
	0.9	0.796	0.132	0.759	0.068
(25, 50)	0	0.872	0.875	0.853	0.857
	1/3	0.232	0.232	0.153	0.154
	0.5	0.175	0.131	0.108	0.067
	0.7	0.371	0.113	0.281	0.055
	0.9	0.784	0.151	0.744	0.083
(50, 50)	0	0.969	0.971	0.962	0.964
	1/3	0.302	0.303	0.209	0.207
	0.5	0.196	0.157	0.121	0.086
	0.7	0.340	0.137	0.250	0.076
	0.9	0.757	0.174	0.713	0.109
(50, 100)	0	0.979	0.980	0.975	0.976
	1/3	0.339	0.341	0.247	0.252
	0.5	0.201	0.177	0.129	0.108
	0.7	0.323	0.154	0.243	0.091
	0.9	0.743	0.194	0.696	0.128
(100, 100)	0	0.999	0.999	0.999	0.999
	1/3	0.469	0.469	0.361	0.358
	0.5	0.259	0.227	0.179	0.138
	0.7	0.347	0.180	0.260	0.106
	0.9	0.754	0.205	0.709	0.129

Table 4.7: Powers of the disparity tests for several members of the BWHD and PBWHD family under contamination. Population 1 is Poisson(10). Population 2 is Poisson(11).

Nominal level		0.1		0.05	
(n_1, n_2)	α	BWHD	PBWHD	BWHD	PBWHD
(25, 25)	0	0.486	0.507	0.400	0.422
	1/3	0.304	0.291	0.200	0.191
	0.5	0.332	0.238	0.237	0.144
	0.7	0.513	0.206	0.437	0.122
	0.9	0.826	0.202	0.794	0.120
(25, 50)	0	0.528	0.547	0.440	0.463
	1/3	0.362	0.352	0.251	0.239
	0.5	0.404	0.305	0.298	0.194
	0.7	0.579	0.276	0.493	0.169
	0.9	0.837	0.270	0.805	0.171
(50, 50)	0	0.586	0.601	0.499	0.522
	1/3	0.467	0.461	0.344	0.335
	0.5	0.483	0.415	0.369	0.283
	0.7	0.595	0.395	0.518	0.277
	0.9	0.837	0.412	0.806	0.307
(50, 100)	0	0.649	0.669	0.555	0.583
	1/3	0.568	0.560	0.446	0.434
	0.5	0.608	0.530	0.491	0.397
	0.7	0.701	0.513	0.627	0.384
	0.9	0.854	0.511	0.823	0.402
(100, 100)	0	0.742	0.755	0.663	0.680
	1/3	0.700	0.697	0.586	0.582
	0.5	0.709	0.675	0.600	0.555
	0.7	0.747	0.664	0.675	0.552
	0.9	0.865	0.673	0.840	0.574

Table 4.8: Powers of the disparity tests for several members of the BWHD and PBWHD family under contamination. Population 1 is Poisson(10). Population 2 is 0.9Poisson(11) + 0.1Poisson(1).

Nominal level		0.1		0.05	
(n_1, n_2)	α	BWHD	PBWHD	BWHD	PBWHD
(25, 25)	0	0.832	0.833	0.801	0.803
	1/3	0.196	0.184	0.116	0.107
	0.5	0.287	0.184	0.196	0.102
	0.7	0.506	0.183	0.429	0.101
	0.9	0.832	0.194	0.798	0.108
(25, 50)	0	0.889	0.889	0.862	0.864
	1/3	0.190	0.181	0.115	0.109
	0.5	0.322	0.222	0.223	0.133
	0.7	0.529	0.237	0.448	0.148
	0.9	0.821	0.262	0.784	0.173
(50, 50)	0	0.925	0.922	0.907	0.908
	1/3	0.233	0.224	0.151	0.143
	0.5	0.381	0.307	0.281	0.204
	0.7	0.555	0.355	0.479	0.250
	0.9	0.843	0.399	0.813	0.302
(50, 100)	0	0.968	0.967	0.960	0.958
	1/3	0.244	0.234	0.159	0.153
	0.5	0.443	0.381	0.341	0.268
	0.7	0.593	0.449	0.516	0.330
	0.9	0.795	0.495	0.765	0.384
(100, 100)	0	0.984	0.984	0.978	0.977
	1/3	0.312	0.305	0.216	0.208
	0.5	0.530	0.496	0.421	0.384
	0.7	0.635	0.589	0.560	0.473
	0.9	0.804	0.638	0.771	0.537

4.2. An Example

In a biological test concerning chemical mutagenicity (see Woodruff *et al.*, 1984), male flies were either exposed to a particular dose of a chemical or to control conditions. The responses were the number of daughter flies of these males having a recessive lethal mutation. One such data sets was analyzed by Simpson (1989, Table 5). The responses are modeled as random samples from a Poisson distribution with mean θ_1 (control), and θ_2 (exposed) respectively.

For testing $H_0 : \theta_1 \geq \theta_2$ against $H_1 : \theta_1 < \theta_2$, a signed disparity is appropriate. Given the ordinary disparity test statistic d_n, this signed disparity statistic is given by $d_n^{1/2}\text{sign}(\hat{\theta}_2 - \hat{\theta}_1)$ where $\hat{\theta}_1$ and $\hat{\theta}_2$ are the minimum disparity estimators of the parameters; for both the HD and the PHD, the signed disparity test is asymptotically equivalent to the signed LRT. For the full data and the reduced data (after removing the two large observations from the treated group) the signed disparities and the associated p-values are given in Table 4.9.

Table 4.9: The signed disparity statistics and their p-values for the Drosophila Data.

Distance	All observations		Outliers Deleted	
	Disparity	p-value	Disparity	p-value
LD	2.595	0.002	1.099	0.136
HD	0.698	0.243	0.743	0.229
PHD	0.707	0.240	0.750	0.227

Notice that the presence or absence of the two large counts has little effect on the robust methods. The null hypothesis, that the mean number of recessive lethal daughters for the control group is larger than that in the treated group is supported in either case. The conclusions, however, are opposite when one uses the signed LRT. Notice also that the HD and the PHD give very similar results, indicating that the robustness property of the test has not been compromised in this case by the use of the penalty.

5. Concluding Remarks

Testing of hypotheses is a fundamental paradigm in statistics. The LRTs which are widely used for parametric inference and have several asymptotic optimality properties are not, in general, robust to outliers. When dealing with real data, robustness to outliers is a major concern. Hampel *et al.* (1986 p. 28) comment that "1–10% gross errors in routine data seem to be more the rule rather than the exception." The Hellinger deviance test of Simpson (1989) and the disparity tests of Lindsay (1994) provide robust alternatives to the LRT. However, many of the robust tests require very large sample sizes for the χ^2 approximation to be reasonably valid; using this approximation in samples of small and moderate sizes may lead to results which are too inaccurate. In this paper we have discussed the use of penalized disparities which can significantly improve the performance of the tests in small samples when multiple samples are involved. Within the BWHD family, the improvement is substantial for disparities with large values of α, in which we are more interested for robustness purposes.

Acknowledgment. A substantial part of this work was done when all three authors were at the University of Texas at Austin.

References

Basu, A., Harris, I. R., and Basu, S. (1996). Tests of hypotheses in discrete models based on the penalized Hellinger distance. *Statistics and Probability Letters*, **27**, 376–373.

Basu, A. and Sarkar, S. (1994). On disparity based goodness-of-fit tests for multinomial models. *Statistics and Probability Letters*, **19**, 307–312.

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

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

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

Huber, P. J. (1981). *Robust Statistics.* John Wiley & Sons, New York.

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

Sarkar, S. and Basu, A. (1995). On disparity based robust tests for two discrete populations. *Sankhyā B*, **57**, 353–364.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics.* John Wiley & Sons, New York.

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

Simpson, D. G. (1989). Hellinger deviance test: efficiency, breakdown points, and examples. *Journal of the American Statistical Association*, **84**, 107–113.

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

Woodruff, R. C., Mason, J. M., Valencia, R., and Zimmering, A. (1984). Chemical mutagenesis testing in drosophila — I: Comparison of positive and negative control data for sex-linked recessive lethal mutations and reciprocal translocations in three laboratories. *Environmental Mutagenesis*, **6**, 189–202.