



Minimum Hellinger Distance Estimation for the Analysis of Count Data

Author(s): Douglas G. Simpson

Reviewed work(s):

Source: *Journal of the American Statistical Association*, Vol. 82, No. 399 (Sep., 1987), pp. 802-807

Published by: [American Statistical Association](#)

Stable URL: <http://www.jstor.org/stable/2288789>

Accessed: 08/03/2012 12:52

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at <http://www.jstor.org/page/info/about/policies/terms.jsp>

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.



American Statistical Association is collaborating with JSTOR to digitize, preserve and extend access to *Journal of the American Statistical Association*.

<http://www.jstor.org>

Minimum Hellinger Distance Estimation for the Analysis of Count Data

DOUGLAS G. SIMPSON*

Minimum Hellinger distance (MHD) estimation is studied in the context of discrete data. The MHD estimator is shown to provide an effective treatment of anomalous data points, and its properties are illustrated using short-term mutagenicity test data. Asymptotic normality for a discrete distribution with countable support is derived under a readily verified condition on the model. Breakdown properties of the MHD estimator and an outlier screen are compared. Count data occur frequently in statistical applications. For instance, in chemical mutagenicity studies, which comprise an important step in the identification of environmental carcinogens, much of the resultant data are counts. Woodruff, Mason, Valencia, and Zimmering (1984) reported anomalous counts in the sex-linked recessive lethal test in *drosophila*. These outliers can have a substantial impact on the experimental conclusions. MHD estimation provides a means for reliable inference when modeling count data that are prone to outliers. The MHD fit gives little weight to counts that are improbable relative to the model. On the other hand, the MHD estimator is asymptotically equivalent to the maximum likelihood estimator when the model is correct. This latter result, long known for a parametric multinomial model with finite support, is extended here to models with countable support. The breakdown point provides a quantification of outlier resistance. Roughly, it is the smallest proportion of outliers in the data that can cause an arbitrarily large shift in the estimate (Donoho and Huber 1983). Here the MHD estimator is shown to have an asymptotic breakdown point of $\frac{1}{2}$ at the model.

KEY WORDS: Asymptotic efficiency; Breakdown point; Discrete probability model; Minimum-distance estimate; Outlier screen.

1. INTRODUCTION

Minimum-distance estimation is appealing for parametric inference when the model is suspected to be inexact; the estimated quantity can always be interpreted as an index of the best-fitting parametric distribution (see Holm 1976, p. 159; Parr 1981, p. 1208). Moreover, this intuitive interpretation can be useful for explaining the estimates to nonstatistical researchers. The purpose of this article is to demonstrate that minimum Hellinger distance (MHD) estimation provides an effective means for analyzing count data that are prone to outliers.

Established properties of the MHD estimator are particularly promising. This estimator has long been known to be first-order efficient in a parametric multinomial model (Rao 1963). More recently, Beran (1977) argued that an MHD estimator for a continuous parametric model should be robust, and he showed that the estimator is asymptotically efficient for a model with compact support. Stather (1981) and Tamura and Boos (1986) gave partial extensions to models with noncompact support. Tamura and Boos (1986) also showed that an affine-invariant MHD estimator for multivariate location and covariance has a

breakdown point of at least $\frac{1}{4}$ at the model. This result is important because it is independent of the dimension of the parameter space. In contrast, the breakdown point of an affine-invariant M estimator is at most $(d + 1)^{-1}$, where d is the dimension (Maronna 1976).

This article focuses on MHD estimation for discrete data, where the model is allowed to have countably infinite support. Section 2 compares MHD estimation and the widely used maximum-likelihood approach. Section 3 presents asymptotic results concerning MHD estimation for discrete data. The asymptotic distribution theory differs for discrete and continuous models because of the need to estimate the probability density function. Section 4 gives an example involving mutagenicity test data. Section 5 compares the breakdown properties (sensitivity to masking) of the MHD estimator and a sequential outlier screen. An improved breakdown bound of $\frac{1}{2}$ is obtained for the MHD estimator.

2. DEFINITION AND COMPARISON WITH MAXIMUM LIKELIHOOD

Consider first the general framework of estimation in a parametric class of distributions $\Gamma = \{F_\theta, \theta \in \Theta\}$. Assume that $\Theta \subset R^d$ and that Γ is dominated. Denote by $\|\cdot\|_p$ the L^p norm

$$\|h\|_p = \left\{ \int |h|^p \right\}^{1/p}. \quad (2.1)$$

The integral in (2.1) is with respect to the dominating measure for Γ . If $\|h\|_p < \infty$, then h is said to be in L^p . A vector-valued function will be said to be in L^p if its components are.

Let f_θ denote the density for F_θ , and let f_n denote a nonparametric density estimate based on a random sample Y_1, \dots, Y_n . An MHD estimate of θ minimizes the Hellinger distance $\|f_n^{1/2} - f_\theta^{1/2}\|_2$. If $f_n \geq 0$ and $\int f_n = 1$, then direct calculation gives

$$\|f_n^{1/2} - f_\theta^{1/2}\|_2^2 = 2 - 2 \int f_n^{1/2} f_\theta^{1/2}. \quad (2.2)$$

For count data take f_n to be the empirical density function

$$f_n(x) = N_x/n, \quad x = 0, 1, \dots, \quad (2.3)$$

where N_x is the frequency of x among Y_1, \dots, Y_n (see Sec. 3). By (2.2), the MHD estimate then maximizes

$$\rho_{n,\theta} = \sum_{x=0}^{\infty} f_n^{1/2}(x) f_\theta^{1/2}(x).$$

* Douglas G. Simpson is Assistant Professor, Department of Statistics and Institute for Environmental Studies, University of Illinois, Champaign, IL 61820. This work was partially supported by National Science Foundation Contract DMS 8400602. The author thanks Barry H. Margolin for stimulating this research and Raymond J. Carroll and David Ruppert for helpful suggestions.

This yields the (standardized) estimating equation

$$\rho_{n,\theta}^{-1} \sum_{x=0}^{\infty} f_n^{1/2}(x) f_{\theta}^{1/2}(x) l_{\theta}(x) = 0, \quad (2.4)$$

where $l_{\theta}(x)$ is the gradient of $\log f_{\theta}(x)$.

Contrast (2.4) with the maximum likelihood estimating equation

$$\sum_{x=0}^{\infty} f_n(x) l_{\theta}(x) = 0. \quad (2.5)$$

Equations (2.4) and (2.5) agree in the limit when F_{θ} obtains. On the other hand, the effect of a deviation from F_{θ} is blunted in (2.4), because the expectation of l_{θ} is with respect to $\rho_{n,\theta}^{-1} f_n^{1/2} f_{\theta}^{1/2}$ rather than f_n . If, for instance, F_{θ} has finite Fisher information, then $f_{\theta}^{1/2}(x) |l_{\theta}(x)| \rightarrow 0$ as $x \rightarrow \infty$; an improbable count has little impact on the MHD estimator.

3. ASYMPTOTIC PROPERTIES

Firmly establishing the asymptotic distribution of the MHD estimator is important for making theoretical comparisons with other estimators and for making valid approximate inferences. In the discrete case, previous derivations of the limiting distribution required the model to have finite support (Rao 1963) or the underlying distribution to satisfy complicated conditions (Stather 1981). Here an alternative derivation is given under a readily verified condition on the model.

First, it is necessary to discuss the consistency of the MHD estimator, that is, the continuity of the MHD functional (see Beran 1977). Write

$$H(\theta; G) = \|g^{1/2} - f_{\theta}^{1/2}\|_2^2, \quad (3.1)$$

where g is the density of G (or the absolutely continuous part of G relative to Γ). The MHD functional T solves

$$H(T(G); G) = \min_{t \in \Theta} H(t; G)$$

if a solution exists. The MHD estimate is given by $T_n = T(F_n)$, where F_n is the distribution corresponding to f_n .

Beran (1977) characterized the existence and continuity of T in the continuous case for compact Θ . The result also applies if Θ is embedded in a compact space $\bar{\Theta}$ and $H(\cdot; G)$ is continuous on $\bar{\Theta}$. Such embeddings can be complicated, however, in multiparameter situations, and the distance does not always extend continuously.

Consider, for example, the two-parameter negative binomial model

$$f_{\theta}(x) = \frac{\Gamma(x + c^{-1})}{x! \Gamma(c^{-1})} \left(\frac{cm}{1 + cm} \right)^x \left(\frac{1}{1 + cm} \right)^{c^{-1}},$$

$$x = 0, 1, \dots, \quad (3.2)$$

where $\theta = (m, c)$, $0 < m < \infty$, and $0 < c < \infty$ (Collings and Margolin 1985). Note that $c = 0$ corresponds to the Poisson density with mean m . As $m \rightarrow \infty$ with $c = 0$, $H(\theta; G) \rightarrow 2$ for each fixed G . On the other hand, it can be shown that $f_{\theta}(0) \rightarrow 1$ as $m \rightarrow \infty$ with m/c fixed, so $H(\theta;$

$G) \rightarrow 2 - 2g^{1/2}(0)$ in this case. Hence $H(\cdot; G)$ does not extend continuously to the limit points of Θ , and a compactification of Θ via its conformal mapping onto a sphere (Apostol 1957, p. 11) would fail to satisfy the conditions of Beran's (1977) theorem 1.

The following theorem extends Beran's existence and continuity result. Let \mathfrak{G} denote the class of distributions G for which

$$\inf_{t \in \Theta - C} H(t; G) > H(\theta^*; G) \quad (3.3)$$

for some compact $C \subset \Theta$ and $\theta^* \in C$. If Θ is compact, then $C = \Theta$, and \mathfrak{G} includes any distribution that is not singular with respect to Γ . [If G and Γ are mutually singular, then $H(\theta; G) \equiv 2$ and $G \notin \mathfrak{G}$.]

Theorem 1. Suppose $f_{\theta}(x)$ is continuous in θ for each x . Then for each $G \in \mathfrak{G}$, (a) $T(G)$ exists, and (b) if $T(G)$ is unique, then $\|g_n^{1/2} - g^{1/2}\|_2 \rightarrow 0$ implies that $T(G_n) \rightarrow T(G)$ as $n \rightarrow \infty$.

Outline of Proof. (a) is immediate from (3.3) and the continuity of $H(\cdot; G)$ on C . To establish (b), first show that if $\|g_n^{1/2} - g^{1/2}\|_2 \rightarrow 0$ then $G_n \in \mathfrak{G}$ eventually, so $T(G_n)$ exists, and show that $T(G_n) \in C$ eventually. Then apply Beran's (1977) theorem 1 with $\Theta = C$.

Corollary. Suppose Γ is identifiable (i.e., $\theta_1 \neq \theta_2$ implies that $\|f_{\theta_1}^{1/2} - f_{\theta_2}^{1/2}\|_2 > 0$), suppose

$$\inf_{t \in \Theta - C} \|f_t^{1/2} - f_{\theta}^{1/2}\|_2 > 0$$

for some compact $C \subset \Theta$, and suppose $f_{\theta}(x)$ is continuous in $\theta \in C$ for each x . Then $\|g_n^{1/2} - f_{\theta}^{1/2}\|_2 \rightarrow 0$ implies that $T(G_n) \rightarrow \theta$ as $n \rightarrow \infty$.

For the negative binomial (3.2), it can be shown that

$$\lim_{n \rightarrow \infty} \inf_{\theta \in C_n} H(t; G) = 2 - 2g^{1/2}(0),$$

with $C_n = \{\theta = (m, c) : n^{-1} \leq m \leq n, 0 \leq c \leq n\}$. Hence T is continuous at G if $H(\theta^*; G) < 2 - 2g^{1/2}(0)$ for some θ^* . If G is in the model family [i.e., $g(x) = f_{\theta}(x)$], then this reduces to $f_{\theta}(0) < 1$, which is satisfied for all θ .

Observe that if G is a count distribution and f_n is given by (2.3), then $\|f_n^{1/2} - g^{1/2}\|_2^2 \leq \|f_n - g\|_1 \rightarrow 0$ almost surely, by the strong law of large numbers and a version of Scheffé's theorem (see Devroye and Györfi 1985, p. 10). If $T(G)$ is unique, consistency of T_n at G then follows via Theorem 1.

To derive the asymptotic normality of the MHD estimator, we impose smoothness conditions on the model. As in Beran (1977), let $s_{\theta} = f_{\theta}^{1/2}$. Suppose for θ interior to Θ that s_{θ} is twice differentiable in L^2 ; that is, suppose \dot{s}_{θ} ($d \times 1$) and \ddot{s}_{θ} ($d \times d$) are in L^2 and satisfy

$$\|s_t - s_{\theta} - \dot{s}_{\theta}'(t - \theta)\|_2 = o(|t - \theta|) \quad (3.4)$$

and

$$\{\dot{s}_t - \dot{s}_{\theta} - \ddot{s}_{\theta}(t - \theta)\} / |t - \theta| \rightarrow 0 \quad (3.5)$$

(componentwise) in L^2 as $|t - \theta| \rightarrow 0$. Here $|a| = \max(|a_1|, \dots, |a_d|)$. Observe that (with probability 1) $\dot{s}_{\theta} =$

$\frac{1}{2}f_{\theta}^{1/2}l_{\theta}$ and $\ddot{s}_{\theta} = \frac{1}{4}f_{\theta}^{1/2}(l_{\theta}l'_{\theta} + 2\dot{l}_{\theta})$, where \dot{l}_{θ} is the gradient of l_{θ} .

Recall that $T(G)$ minimizes $H(t; G) = 2 - 2 \int s_t g^{1/2}$. By the Cauchy-Schwarz inequality, conditions (3.4) and (3.5) imply that H is twice differentiable:

$$\begin{aligned}\dot{H}(t; G) &= -2 \int \dot{s}_t g^{1/2} = - \int l_t f_t^{1/2} g^{1/2}, \\ \ddot{H}(t; G) &= -2 \int \ddot{s}_t g^{1/2} = - \int (\tfrac{1}{2} l_t l'_t + \dot{l}_t) f_t^{1/2} g^{1/2}.\end{aligned}$$

The MHD estimate T_n is then a zero of $\dot{H}(t; F_n)$. In the discrete case considered here, F_n is the empirical distribution function.

Theorem 2. Let G and Γ be supported on $\{0, 1, 2, \dots\}$, and let Y_1, Y_2, \dots be independent observations from G . Assume (3.4) and (3.5). Suppose $\dot{H}(t; G)$ has a zero θ interior to Θ , $\dot{H}(\theta; G)$ is nonsingular, and $\dot{s}_{\theta} \in L^1$. Then $T_n \rightarrow \theta$ in probability implies that $n^{1/2}(T_n - \theta) \rightarrow Z \sim N_d(0, V_{\theta})$ in law as $n \rightarrow \infty$, where $V_{\theta} = \frac{1}{4}\ddot{H}(\theta; G)^{-1}i(\theta)\ddot{H}(\theta; G)^{-1}$ and $i(\theta) = 4 \int \dot{s}_{\theta}\dot{s}'_{\theta} = \int l_{\theta}l'_{\theta}f_{\theta}$. If $G \equiv F_{\theta}$, then $V_{\theta} = i(\theta)^{-1}$. The proof is in the Appendix.

In the proof more is shown, namely,

$$n^{1/2}(T_n - \theta) = n^{-1/2}i(\theta)^{-1} \sum_{i=1}^n l_{\theta}(Y_i) + o_p(1)$$

when $G = F_{\theta}$, so the MHD estimator is asymptotically equivalent to the maximum likelihood estimator.

Note that in Theorem 2 the bulk of the conditions are on the model, not the actual distribution. The most stringent condition, that $\dot{s}_{\theta} \in L^1$, is satisfied by the Poisson, negative-binomial (3.2), and log-series models.

Stather (1981) required the existence of a sequence $\{k_n\}$ for which

$$\sum_{x \leq k_n} \dot{s}_{\theta}(x)g(x)^{-1/2} = o(n^{1/2}), \quad \sum_{x > k_n} g(x) = o(n^{-1}),$$

as $n \rightarrow \infty$, where $\theta = T(G)$ and g is the density for G . Under general conditions this condition implies $\dot{s}_{\theta} \in L^1$ when $g = f_{\theta}$ (see Simpson 1985).

4. AN EXAMPLE: THE SEX-LINKED RECESSIVE LETHAL TEST

To complement the asymptotic results of the previous section, it is useful to consider the performance of the MHD estimator on chemical mutagenicity data. In the sex-linked recessive lethal test in drosophila (fruit flies), groups of male flies are exposed to different doses of a chemical to be screened. Each male is then mated with unexposed females. One observes the number of daughter flies carrying a recessive lethal mutation on the X chromosome. Details of the experimental protocol were given by Woodruff, Mason, Valencia, and Zimmering (1984).

Table 1 summarizes 124 control runs from this experiment (data supplied by J. Mason). For these data, roughly 100 daughters were sampled from each male. The table gives frequencies of frequencies, among the 124 runs, of

males having 0, 1, 2, . . . recessive lethal daughters. The frequencies of zero frequencies, in Table 1, for counts larger than 1 indicate that spontaneous mutations are rare. A striking feature of these data, however, is the occasional occurrence of exceptionally large counts. For instance, one male is reported to have produced 91 recessive lethal daughters.

Woodruff et al. (1984) referred to these exceptional counts as "clusters." They conjectured that, unlike the majority of the recessive lethals, which result from mutations during meiosis, a cluster results "from a single spontaneous premeiotic event" (p. 195). Consequently, they advocated the exclusion of observations identified as clusters, and they employed a sequential outlier screen.

Table 2 shows four of the experimental runs summarized in Table 1. For each run, Poisson models were fit using three methods—maximum likelihood (averaging), averaging after outlier deletion (see Sec. 5), and MHD estimation. Jackknife standard errors, shown in parentheses, provide nonparametric estimates of precision. See Efron (1982, pp. 13–19) for a description of the jackknife estimate of the standard deviation. The MHD estimates were computed by direct minimization of (3.1) using IMSL routine ZXGSP.

For Day 27 and the first run of Day 177, all of the estimates are nearly the same and the corresponding fitted frequencies agree quite well with the observed frequencies. The estimates differ markedly, however, for the remaining runs, which contain exceptional counts relative to the Poisson model. In particular, the maximum likelihood fit to the observed frequencies is poor and the precision estimates are inflated. On the other hand, the fits obtained via outlier deletion or MHD estimation provide reasonable summaries of the major portions of the data.

Note that an outlier screen explicitly brands certain observations as outliers. MHD estimation provides similar diagnostic information; improbable counts are readily identified as unusual because of their small fitted frequencies.

One could, of course, develop alternative analyses specific to these data. The fitting of a Poisson model is used here to illustrate the potential usefulness of an MHD fit to count data that are prone to outliers.

Table 1. Frequencies in 124 Runs of the *Drosophila* Assay of Frequencies of Recessive Lethal Counts for Individual Males

No. of males	No. of recessive lethal daughters							
	0	1	2	3	4	5	6	≥7
0	0	7	57	99	121	122	121	121
1	0	6	39	21	3	2	3	(9, 13, 91)*
2	0	13	16	4				
3	0	15	9					
4	0	13	1					
5–9	1	50	2					
10–19	25	19						
20–39	81	1						
≥40	17							

* Actual counts.

Table 2. Fits of a Poisson Model to *Drosophila* Data Using Maximum Likelihood (ML), Maximum Likelihood With Outlier Deletion (ML + D), and Minimum Hellinger Distance (MHD) Estimation

Day		Recessive lethal count						$\hat{\theta}$	Standard error
		0	1	2	3	4	≥ 5		
27	Obs.	25	4	0	0	0	0		
	ML, ML + D	25.26	3.48	.24	.01	—	—	.138	.065
	MHD	25.64	3.15	.19	—	—	—	.123	.053
28	Obs.	23	3	0	1	1	0		
	ML	19.59	7.00	1.25	.15	.01	—	.357	.18
	ML + D	24.95	2.88	.17	.01	—	—	.115	.064
	MHD	24.70	3.01	.19	—	—	—	.125	.072
177	Obs.	25	9	1	0	0	0		
	ML, ML + D	25.56	8.03	1.26	.13	.01	—	.314	.090
	MHD	25.86	7.83	1.19	.12	—	—	.303	.11
177	Obs.	23	7	3	0	0	1(91)		
	ML	1.60	4.88	7.47	7.61	5.82	6.62	3.059	2.67
	ML + D	22.93	9.03	1.78	.23	.02	—	.394	.11
	MHD	23.63	8.60	1.56	.19	.02	—	.364	.10

NOTE: A dash indicates a frequency less than .01.

5. BREAKDOWN AND MASKING

Outlier deletion is a popular and often easily implemented method for treating anomalous data points. Therefore, it is worth comparing the MHD approach and the outlier screen used by Woodruff et al. (1984). The comparison here centers on sensitivity to masking, which amounts to a study of the breakdown points of the two procedures (see Donoho and Huber 1983).

Woodruff et al. (1984) were concerned with the possibility of erroneous large counts. Their sequential outlier screen excludes the largest count, at each stage, if it is too extreme relative to a Poisson model. Specifically, let $F(\cdot, \mu)$ denote the Poisson distribution function with mean μ , and let $y_{(n)}$ and \bar{y}_n denote, respectively, the observed maximum and mean. The rule then is to use \bar{y}_n to estimate μ provided that $F(y_{(n)}; \bar{y}_n)^n \leq 1 - \alpha_0$; otherwise omit $y_{(n)}$, set $n = n - 1$, and repeat (J. Mason 1985, personal communication). The quantity α_0 is a prespecified small probability, for example, $\alpha_0 = .01$ or $.001$.

Consider the effect of adding large contaminants c_1, \dots, c_k to data $y = (y_1, \dots, y_n)$. For simplicity take $c_1 = \dots = c_k = c$. The mean of the augmented sample is $m(c, k) = (n\bar{y}_n + kc)/(n + k)$. Masking occurs when

$$F(c; m(c, j))^{n+j} \leq 1 - \alpha_0 \text{ for some } j \in \{1, \dots, k\}, \quad (5.1)$$

that is, when at least one contaminant is retained. Note that (5.1) is equivalent to

$$1 - F(c; m(c, k)) \geq 1 - (1 - \alpha_0)^{1/(n+k)}, \quad (5.2)$$

because the left side of (5.1) is decreasing in j . A breakdown of the procedure in the sense of Donoho and Huber (1983) would occur if (5.2) were to hold as $c \rightarrow \infty$. But $1 - F(c; m(c, k)) = O(c^{-1})$ by Chebyshev's inequality, so breakdown does not occur for fixed n and k ; all outliers are excluded if they are sufficiently large.

An outlier screen would presumably be most useful for

treating the less obvious moderate outliers; the finite sample breakdown model is uninformative in this example. To model the effect of moderate outliers, consider instead the following asymptotic setup. As $n \rightarrow \infty$ let $\bar{y}_n \rightarrow m_0 < \infty$ and $k/(n + k) \rightarrow \varepsilon \in (0, 1)$. If $c_n \rightarrow \infty$, then $m(c_n, k) \sim \varepsilon c_n \rightarrow \infty$. The procedure will then break down in the limit if there is a sequence $\{c_n\}$, $c_n \rightarrow \infty$, for which (5.2) holds eventually. One such sequence is given by

$$c_n = \max\{x : F(x; m_0) \leq \gamma^{1/(n+k)}\},$$

where $0 < \gamma < 1 - \alpha_0$. This same sequence works for any $\varepsilon > 0$, so the asymptotic breakdown point of the outlier screen is 0.

For the MHD estimator we obtain a general lower bound for the breakdown point at a distribution G . Consider the contamination model

$$H_n = (1 - \varepsilon)G + \varepsilon K_n, \quad (5.3)$$

where $\{K_n\}$ is some sequence of distributions. The data-contamination model considered previously is recovered when G is an empirical distribution of y , K_n is a point mass distribution at c_n , and ε is the proportion of contaminants. For distributions H and K having densities h and k , define $\rho(H, K) = \int h^{1/2}k^{1/2}$.

Theorem 3. Let $\hat{\rho} = \max\{\rho(G, F_t), t \in \Theta\}$, and suppose the maximum occurs interior to Θ . Let $\rho^* = \lim_{M \rightarrow \infty} \sup_{|t| > M} \rho(G, F_t)$. If

$$\varepsilon < (\hat{\rho} - \rho^*)^2/[1 + (\hat{\rho} - \rho^*)^2], \quad (5.4)$$

Table 3. Effect of Adding Sixes to 20 Observations With Mean 1

k	$m(c, k)$	$F(c, m(c, k))^{n+k}$	$\hat{\mu}$	
			ML + D	MHD
0	1.0	.998	1.0	.95
1	1.24	.994	1.0	.99
2	1.45	.983	1.45	1.01

NOTE: ML + D represents maximum likelihood with outlier deletion. MHD represents minimum Hellinger distance.

then there is no sequence of the form (5.3) for which $|T(H_n) - T(G)| \rightarrow \infty$ as $n \rightarrow \infty$. Here T denotes the MHD functional.

Proof. Suppose $|T(H_n) - T(G)| \rightarrow \infty$, where H_n is as in (5.3). It will be shown that this implies that ε exceeds the bound in (5.4). Let $\theta = T(G)$, the maximizing value of t in $\rho(G, F_t)$. There must be a sequence $\{\theta_n\}$ with $|\theta_n| \rightarrow \infty$ for which $\rho(H_n, F(\cdot, \theta_n)) > \rho(H_n, F(\cdot, \theta))$, infinitely often. Now $\rho(H_n, F(\cdot, \theta)) \geq (1 - \varepsilon)^{1/2} \hat{\rho}$ and $\rho(H_n, F(\cdot, \theta_n)) \leq (1 - \varepsilon)^{1/2} \rho(G, F(\cdot, \theta_n)) + \varepsilon^{1/2}$; moreover, $\rho(G, F(\cdot, \theta_n)) \leq \rho^* + \delta$ eventually for every $\delta > 0$. Hence it must be that $(1 - \varepsilon)^{1/2} \hat{\rho} \leq (1 - \varepsilon)^{1/2} \rho^* + \varepsilon^{1/2}$; that is, (5.4) is not satisfied.

For the special case that Γ is the Poisson family and G is a Poisson distribution, $\hat{\rho} = 1$ and $\rho^* = 0$, so the asymptotic breakdown point is $\frac{1}{2}$. Note that frequently $\rho^* = 0$ and that in general $\hat{\rho}$ depends on the fit of the model to G ($\hat{\rho} = 1$ means perfect fit) (see Simpson 1985; Tamura and Boos 1986).

As an example of the effect of masking, suppose the original data consist of 20 observations with frequencies $N_0 = 6$, $N_1 = 9$, $N_2 = 4$, and $N_3 = 1$. The mean is $\bar{y}_{20} = 1$. Table 3 shows the effect of adding sixes to the data. If $\alpha_0 = .01$, then the outlier screen fails to exclude more than one of the contaminants; the estimate of μ jumps from 1.0 at $k = 1$ to 1.45 at $k = 2$, a change of roughly two "standard errors." In contrast, the effect of the contaminants on the MHD estimate is minimal.

6. DISCUSSION

MHD estimation has a number of desirable features for the analysis of count data. The MHD estimator discounts anomalous data points in a smooth manner, effectively rejecting highly improbable counts. On the other hand, the estimator enjoys the theoretical advantage of first-order efficiency for a correctly specified model. Like maximum likelihood (and unlike some other approaches) the robust-estimation MHD is "automatic"; the model determines the estimator. This is advantageous for scientific applications (see Efron 1986).

APPENDIX: PROOF OF THEOREM 2

The following lemma is central to the proof.

Lemma. Let F be supported on $\{0, 1, 2, \dots\}$, and let Y_1, Y_2, \dots be independent observations from G . Assume that (3.4) holds at θ , a zero of $\dot{H}(\theta; G)$. If $\dot{s}_\theta \in L^1$, then

$$-\dot{H}(\theta; F_n) = n^{-1} \sum_{i=1}^n \dot{s}_\theta(Y_i) g(Y_i)^{-1/2} + o_p(n^{-1/2}). \quad (\text{A.1})$$

Proof. Write $\dot{s}(x, t)$ for $\dot{s}_t(x)$. The right-side sum in (A.1) is $\sum_{x=0}^{\infty} \dot{s}(x, \theta) g(x)^{-1/2} f_n(x)$, omitting terms with $g(x) = 0$. Let $R_n = -\dot{H}(\theta; F_n) - \sum \dot{s}(x, \theta) g(x)^{-1/2} f_n(x)$. Then

$$\begin{aligned} R_n &= 2 \sum_{x=0}^{\infty} \dot{s}(x, \theta) \{f_n^{1/2}(x) - g^{1/2}(x)\} \\ &\quad - \sum \dot{s}(x, \theta) g(x)^{-1/2} \{f_n(x) - g(x)\} \\ &= - \sum_{x=0}^{\infty} \dot{s}(x, \theta) g(x)^{-1/2} \{f_n^{1/2}(x) - g^{1/2}(x)\}^2, \end{aligned}$$

because $\dot{H}(\theta; G) = 0$. Let R_{ni} denote the i th component of R_n , and let $s_i(x, t)$ denote the i th component of $\dot{s}(x, t)$. Then

$$\begin{aligned} E\{n^{1/2} R_{ni}\} &\leq \sum_{x=0}^{\infty} |s_i(x, \theta)| g(x)^{-1/2} \\ &\quad \times n^{1/2} E\{f_n^{1/2}(x) - g^{1/2}(x)\}^2. \end{aligned} \quad (\text{A.2})$$

Now $E\{f_n^{1/2}(x) - g^{1/2}(x)\}^2 \leq E\{f_n(x) - g(x)\} \leq [E\{f_n(x) - g(x)\}^2]^{1/2} = n^{-1/2} g^{1/2}(x) \{1 - g(x)\}^{1/2}$, so the summand in (A.2) is dominated by $|s_i(x, \theta)|$. To show that (A.2) converges to 0, it is sufficient, by the Lebesgue dominated convergence theorem, to show that $E Y_n^2(x) \rightarrow 0$, where $Y_n(x) = n^{1/4} [f_n^{1/2}(x) - g^{1/2}(x)]$.

Now $Y_n(x) \rightarrow 0$ in law as $n \rightarrow \infty$, and

$$E|Y_n(x)|^{2(1+\varepsilon)} \leq [g(x)\{1 - g(x)\}]^{(1+\varepsilon)/2} < \infty$$

for $0 < \varepsilon < 1$. Hence $\{Y_n^2(x)\}$ is uniformly integrable (Serfling 1980, p. 14), and the desired convergence holds. The result follows by Markov's inequality.

Proof of Theorem 2. Under the conditions given,

$$0 = \dot{H}(T_n; F_n) = \dot{H}(\theta; F_n) + \ddot{H}(\theta; F_n)(T_n - \theta) + o(|T_n - \theta|)$$

and $|\ddot{H}(\theta; F_n) - \ddot{H}(\theta; G)| \leq 2L\|f_n^{1/2} - g^{1/2}\|_2$, where L is the maximum L^2 norm of the components of \dot{s}_θ . Hence

$$T_n - \theta + o(|T_n - \theta|) = \{-\ddot{H}(\theta; G)^{-1} + o_p(1)\} \ddot{H}(\theta; F_n) \quad (\text{A.3})$$

as $n \rightarrow \infty$. By the previous lemma and a multivariate central limit theorem,

$$-n^{1/2} \dot{H}(\theta; F_n) \rightarrow Y \sim N_d(0, \frac{1}{4} i(\theta)) \quad (\text{A.4})$$

in law as $n \rightarrow \infty$. The result follows from (A.3) and (A.4).

When $G = F_\theta$, a standard argument shows that $\ddot{H}(\theta; F_\theta) = -\frac{1}{2} i(\theta)$, so $V_\theta = i(\theta)^{-1}$.

[Received April 1986. Revised October 1986.]

REFERENCES

- Apostol, T. M. (1957), *Mathematical Analysis*, Reading, MA: Addison-Wesley.
- Beran, R. J. (1977), "Minimum Hellinger Distance Estimates for Parametric Models," *The Annals of Statistics*, 5, 445-463.
- Collings, B. J., and Margolin, B. H. (1985), "Testing Goodness of Fit for the Poisson Assumption When Observations Are Not Identically Distributed," *Journal of the American Statistical Association*, 80, 411-418.
- Devroye, L., and Györfi, L. (1985), *Nonparametric Density Estimation: The L_1 View*, New York: John Wiley.
- Donoho, D. L., and Huber, P. J. (1983), "The Notion of Breakdown Point," in *A Festschrift for Erich L. Lehmann*, eds. P. Bickel, K. Doksum, and J. L. Hodges, Jr., Belmont, CA: Wadsworth.
- Efron, B. (1982), "The Jackknife, the Bootstrap, and Other Resampling Plans" (CBMS-NSF Regional Conference Series in Applied Mathematics, No. 38), Philadelphia: Society for Industrial and Applied Mathematics.
- (1986), "Why Isn't Everyone a Bayesian?," *The American Statistician*, 40, 1-5.
- Holm, S. (1976), Discussion of "Another Look at Robustness," by Peter Bickel, *Scandinavian Journal of Statistics*, 3, 158-161.
- Maronna, R. A. (1976), "Robust M -Estimators of Multivariate Location and Scatter," *The Annals of Statistics*, 4, 51-67.
- Parr, W. C. (1981), "Minimum Distance Estimation: A Bibliography," *Communications in Statistics, Part A—Theory and Methods*, 10, 1205-1224.
- Rao, C. R. (1963), "Criteria of Estimation in Large Samples," *Sankhyā*, Ser. A, 25, 189-206.
- Serfling, R. J. (1980), *Approximation Theorems of Mathematical Statistics*, New York: John Wiley.

- Simpson, D. G. (1985), "Some Contributions to Robust Inference for Discrete Probability Models," Mimeo Series 1594, University of North Carolina, Institute of Statistics.
- Stather, C. R. (1981), "Robust Statistical Inference Using Hellinger Distance Methods," unpublished Ph.D. dissertation, LaTrobe University (Australia), Dept. of Mathematical Statistics.
- Tamura, R., and Boos, 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, S. (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 Mutagenicity*, 6, 189–202.