Semiparametric Inference Based on a Class of Zero-Altered Distributions

Sujit K. Ghosh and Honggie Kim

Department of Statistics, NC State University.

Institute of Statistics Mimeo Series #2589

Abstract

In modeling count data collected from manufacturing processes, economic series, disease outbreaks and ecological surveys, there are usually a relatively large or small number of zeros compared to positive counts. Such low or high frequencies of zero counts often require the use of under or over dispersed probability models for the underlying data generating mechanism. The commonly used models such as generalized or zero-inflated Poisson distributions can usually account for only the over dispersion, but such distributions are often found to be inadequate in modeling underdispersion because of the need for awkward parameter or support restrictions. This article introduces a flexible class of semiparametric zero-altered models which account for both under and over dispersion and includes other familiar models such as those mentioned above as special cases. Consistency and asymptotic normality of the dispersion parameter are derived under general conditions. Numerical support for the performance of the proposed method of inference is presented for the case of common discrete distributions.

Keywords: Asymptotic normality; Consistency; Overdispersion; Semiparametric inference; Underdisperson; Zero-altered distribution.

1 Introduction

Statistical methods for analyzing count data with too few or too many zeros are very important in various scientific fields including but not limited to industrial applications (e.g., Lambert, 1992), econometrics (e.g., Cameron and Trivedi, 1986) and biomedical applications (e.g., Heilbron and Gibson, 1990, Hall, 2000). Most of these applications were motivated by the observed overdispersion due to excess zero counts. For other interesting applications related to zero-inflated models see Dahiya and Gross (1973), Umbach (1981), Yip (1988), Gupta et al. (1996), Welsh et al. (1996), Gurmu (1997) and Hinde and Demetrio (1998). An overview of zero inflated models can be found in Ridout et al. (1998) and Tu (2002). However underdispersion can also be observed in practice (Famoye, 1993). As the case of underdispersion is relatively rare (but not inevitable) we present a data set that features underdispersion (see Table 1). Winkelmann and Zimmermann (1995) present many illustrations and applications. From their article we quote only a few areas of potential applications: the analysis of accident proneness (e.g., airline failures), labor mobility (the number of changes of employer), the demand for health-care services (as measured by the number of doctor consultations in a given time) and, in economic demography, total fertility (the number of births by a woman). Winkelmann and Zimmermann (1995) also provides a thorough review of related statistical inference for count data, mostly based on parametric models. Our proposed model includes most of such parametric models as special cases within a semiparametric framework.

As an illustration, consider the data set in Table 1 that aims modeling the function word counts (Bailey, 1990). The data show clear evidence of underdispersion. There are several other data sets as above which show evidence of underdispersion. We use the above mentioned data set as a motivating example to develop our models for count data with relatively low or high proportion of zeros. However, by no means, the proposed methodology is limited to this specific data set.

In this article, we propose a new class of zero-altered distributions that can account for both under and over dispersion. This is similar in spirit to the work of Castillo and Perez-Casany (2005) who have recently introduced a class of parametric generalizations of the Poisson distribution that can also account for both under and over dispersion. However we find the class of such weighted version of Poisson distributions (see Rao, 1965) to be somewhat inflexible in practice. For instance, closed form analytical expressions for the normalizing constants are not available in general, which leads to only implicit function representation of the moments of the distribution. Further only an approximate log-linear models can be used to fit such distributions as the closed form of the likelihood is not available. Although their models seem to include the class zero-modified distributions (Johnson et al., 1992), but in order to capture underdispersion the weighting parameter needs to be truncated by a bound that will depend on other parameters (e.g., in equation (4) of Castillo and Perez-Casany (2005), one requires $\epsilon > -p_0/(1-p_0)$, where p_0 is probability of a zero count under a parametric model, such as Poisson model). We propose a class of *semiparametric* models that avoids the above mentioned analytical and practical limitations.

We show that besides the flexibility of modeling both types of dispersion, the proposed distributions have several other advantages: (i) the support of the distributions is $\mathbb{N} = \{0, 1, 2, \ldots\}$ even when the distribution is underdispersed; (ii) the parameter value (which we denote by δ) that determines the nature of dispersion lies in the open interval (e.g. $\delta \in (-1, 1)$); (iii) asymptotic distributions of the estimator $\hat{\delta}$ and the test statistic under the null hypothesis $H_0 : \delta = 0$, both have normal distributions; and last but not the least (iv) the estimation methodology is based on a semiparametric model. It may be noted that absolute bounds of the parameter $\delta \in (-1, 1)$ makes it easier to formulate regression problems using suitable link functions.

Section 2 presents the general formulations of zero-altered models (Heilbron,

1994) with no assumption on the underlying equidispersed distribution. Section 3 presents statistical inference based on an M-estimation theory. Section 4 illustrates the procedure with real-life data sets presented at the beginning of this section and also presents results based on simulation studies. Section 5 concludes this study and addresses a few areas of future research.

2 Generalized Zero-Altered Distributions

Consider a non-degenerate random variable U taking values in $\mathbb{N} = \{0, 1, 2, \ldots\}$ with probability mass function $f_0(u)$ for $u \in \mathbb{N}$, i.e., $f_0(u) \geq 0$ and $\sum_{u=0}^{\infty} f_0(u) = 1$. Without loss of any generality we assume that $f_0(0) < 1$. We can modify any such probability mass function by altering its probability at $u = 0$, using a *dispersion parameter*, $\delta \in (-1, 1)$ by defining a new class of probability mass functions as follows:

$$f_\delta(x) = \begin{cases} \delta_+^2 + (1 - \delta^2) f_0(0) & \text{if } x = 0 \\ \left(1 - \delta_+^2 + \delta_-^2 \frac{f_0(0)}{1 - f_0(0)}\right) f_0(x) & \text{for } x = 1, 2, \ldots \end{cases} \tag{1}$$

Here, we have used the standard notations, $\delta_+ = \max\{\delta, 0\}$ and $\delta_- = \max\{-\delta, 0\}$ to denote the positive and negative part of δ, so that $\delta = \delta_+ - \delta_-$. Clearly, for any $|\delta| < 1$, the function $f_\delta(\cdot)$ in (1) is a probability mass function (pmf). Also notice that, $f_{\delta=0} = f_0$, so that f_0 is included in the class of distributions generated by $f_\delta(\cdot)$. Notice that in (1) instead of δ_+^2 we could have used δ_+ or more generally any smooth increasing function of δ_+. However our choice makes f_δ a smooth function of δ having a continuous first derivative with respect to δ, which will turn out to be useful in deriving the asymptotic inference (see Section 3).

In order to study the general properties of f_δ, let X denote a statistic with probability mass function f_δ, which we will denote by $X \sim f_\delta$, $\delta \in (-1, 1)$. First, notice that the representation in (1) is unique in the sense that if there were δ^* and f_0^*, such that $f_\delta(\cdot) = f_{\delta^*}(\cdot)$ and $f_0(0) - f_0^*(0)$, then $\delta = \delta^*$ and $f_0(\cdot) = f_0^*(\cdot)$. In other words, the parameterization defined by (1) is identifiable under a mild

restriction on $f_0(\cdot)$ (e.g., $f_0(0) = 0.5$ etc.). Next we show that f_δ can account for both over and under dispersion.

DEFINITION 1. *A random variable X having finite second moment is said to be* **underdispersed** *or* **overdispersed** *if $\mu > \sigma^2$ or $\mu < \sigma^2$, respectively, where $\mu = E[X]$ and $\sigma^2 = Var[X]$ denote the mean and variance of the random variable X, respectively. If X is neither under nor over dispersed, i.e., if $\mu = \sigma^2$, then X is said to be* **equidispersed**.

From here on, we assume that $E_{f_0}[U^2] < \infty$, i.e., $\sum_{u=1}^{\infty} u^2 f_0(u) < \infty$. It immediately follows from (1) that if $X \sim f_\delta$ then $E[X^2] < \infty$ for all $\delta \in (-1, 1)$. We now state the mean and variance of $X \sim f_\delta$ in terms of the mean $\mu_0 = E_{f_0}[U]$ and variance $\sigma_0^2 = Var_{f_0}[U]$ of the underlying random variable U.

LEMMA 1. *The mean μ and the variance σ^2 of $X \sim f_\delta$ is given by*

$$\mu = \omega(\delta)\mu_0 \tag{2}$$

$$\sigma^2 = \omega(\delta)\sigma_0^2 + \frac{1-\omega(\delta)}{\omega(\delta)}\mu^2, \tag{3}$$

where $\omega(\delta) = \left(1 - \delta_+^2 + \delta_-^2 \frac{f_0(0)}{1-f_0(0)}\right)$.

Lemma 1 easily follows from the fact that $f_\delta(x) = \omega(\delta) f_0(x)$ for $x \neq 0$, which in turn implies that $E[X] = \omega(\delta)E[U] = \omega(\delta)\mu_0$ and $E[X^2] = \omega(\delta)E[U^2] = \omega(\delta)(\sigma_0^2 + \mu_0^2)$. Based on Lemma 1 we can derive the following result:

THEOREM 1. *Suppose $X \sim f_\delta$, where f_δ is given by (1) and assume that the underlying random variable U is equidispersed, i.e., $\mu_0 = \sigma_0^2$. Then X is*

$$\begin{aligned} & \textit{underdispersed i.e., } \mu > \sigma^2 \quad \textit{if} \quad \delta < 0, \\ & \textit{equidispersed i.e., } \mu = \sigma^2 \quad \textit{if} \quad \delta = 0, \\ \textit{and} \quad & \textit{overdispersed i.e., } \mu < \sigma^2 \quad \textit{if} \quad \delta > 0. \end{aligned}$$

Proof: From (3) of Lemma 1 it follows that when U is equidispersed (i.e., $\mu_0 = \sigma_0^2$), $\sigma^2 - \mu = \frac{1-\omega(\delta)}{\omega(\delta)}\mu^2$ and

$$\frac{1-\omega(\delta)}{\omega(\delta)} = \begin{cases} \frac{1-\delta^2}{\delta^2} & \text{if } \delta > 0 \\ \frac{-\delta^2 f_0(0)}{1-(1-\delta^2)f_0(0)} & \text{if } \delta \leq 0. \end{cases}$$

Thus, $\sigma^2 - \mu$ has the same sign as δ. This completes the proof of Theorem 1.

The above result clearly indicates that δ acts as a measure of dispersion; a negative value ($\delta < 0$) indicates underdispersion while a positive value ($\delta > 0$) indicates overdispersion and a value of zero ($\delta = 0$) indicates equidispersion. Theorem 1 can be generalized when the underlying random variable U is not equidispersed, but then we can always obtain under and over dispersed models using the modification in (1). In this sense, we can restrict the underlying distribution to belong to a class of equidispersed distributions (e.g., a Possion distribution). Appendix A presents an extension of Theorem 1, when U is not necessarily equidispersed.

In general, any discrete distribution could be used for U, however to derive asymptotic inference and parsimony we restrict our attention to only an equidispersed discrete distribution for U. Moreover, given any random variable X with pmf $g(\cdot)$ and having a finite second moment, we can find δ and $f_0(\cdot)$ such that $f_\delta(\cdot) = g(\cdot)$, where $f_0(\cdot)$ satisfies $\mu_0 = \sigma_0^2$ (see Theorem 4 in the Appendix). Based on this semiparametric model for X we develop estimation and testing methodology based on the general theory of M-estimation (Boos and Stefanski, 2002) methods.

3 Statistical Inference

Let $X_1, \ldots, X_n$ be a independent and identically distributed (iid) random variables with common distribution $f_\delta(x)$ as given in (1) with the restriction that $\sum_{u=1}^{\infty} u^2 f_0(u) < \infty$ and that $\mu_0 = E[U] = Var[U] = \sigma_0^2$. Notice that no specific

distribution for U is assumed other than the fact that U is equidispersed. Let $\pi_0 = f_0(0)$ and $\omega(\delta, \pi_0) = 1 - \delta_+^2 + \delta_-^2 \frac{\pi_0}{1-\pi_0}$. Under the above assumption, it follows that $E[X] = \mu_0 \omega(\delta, \pi_0)$, $E[X(X-1)] = \mu_0^2 \omega(\delta, \pi_0)$ and $\Pr[X = 0] = \delta_+^2 + (1-\delta^2)\pi_0$. The above three facts lead to the following set of estimating equations:

$$\begin{aligned} \sum_{i=1}^{n} \psi_1(X_i, \boldsymbol{\theta}) &= \sum_{i=1}^{n} (X_i - \theta_3 \omega(\theta_1, \theta_2)) = 0 \\ \sum_{i=1}^{n} \psi_2(X_i, \boldsymbol{\theta}) &= \sum_{i=1}^{n} (X_i(X_i - 1) - \theta_3^2 \omega(\theta_1, \theta_2)) = 0 \text{ and} \\ \sum_{i=1}^{n} \psi_1(X_i, \boldsymbol{\theta}) &= \sum_{i=1}^{n} (I_{\{0\}}(X_i) - \theta_{1+}^2 - (1 - \theta_1^2)\theta_2) = 0 \end{aligned} \tag{4}$$

where $I_A(x)$ denotes the indicator function such that $I_A(x) = 1$ if $x \in A$, otherwise $I_A(x) = 0$ and $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$ where $\theta_1 = \delta, \theta_2 = \pi_0$ and $\theta_3 = \mu_0$

From here on we assume that there is a $\boldsymbol{\theta}_0 \in \Theta = (-1, 1) \times (0, 1) \times (0, \infty)$ such that $E_{\boldsymbol{\theta}_0}[\boldsymbol{\psi}(\boldsymbol{\theta}, X)] = 0$ if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, where $E_{\boldsymbol{\theta}_0}$ denotes the expectation with respect to $f(x|\boldsymbol{\theta}_0)$ and $\boldsymbol{\psi}(\cdot, \cdot) = (\psi_1(\cdot, \cdot), \psi_2(\cdot, \cdot), \psi_3(\cdot, \cdot))$. In other words, the true pmf of the data is given by $f(x|\boldsymbol{\theta}_0)$ (as defined in (1)) for some $\boldsymbol{\theta}_0 \in \Theta$. Notice that as the underlying pmf f_0 of U is specified only up to second moment, any pmf with finite second moment belongs to the class of our proposed models $\mathcal{M} = \{f(x|\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ (see Theorem 4 in the Appendix). We now develop an estimator of $\boldsymbol{\theta}_0$ (and in particular $\theta_{10} = \delta_0$) that is consistent and asymptotically normal for the model $\mathcal{M}$.

Let $\bar{X} = \sum_{i=1}^{n} X_i/n$ and $S^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2/(n-1)$ denote the sample mean and sample variance, respectively. Also let $P_0 = \sum_{i=1}^{n} I_{\{0\}}(X_i)/n$ denotes the proportion of zeros observed in the sample. It follows easily that the equation (4) is equivalent to solving:

$$\begin{aligned} \bar{X} &= \theta_3 \omega(\theta_1, \theta_2), \\ \frac{n-1}{n} S^2 - \bar{X} + \bar{X}^2 &= \theta_3^2 \omega(\theta_1, \theta_2) \text{ and} \\ P_0 &= \theta_{1+}^2 + (1 - \theta_1^2)\theta_2. \end{aligned} \tag{5}$$

Now, by solving (5) we obtain,

$$\begin{aligned} \hat{\delta} = \hat{\theta}_1 &= sgn(D)\sqrt{\frac{|D|}{|D| + R\bar{X}^2}}, \\ \hat{\pi_0} = \hat{\theta}_2 &= \frac{(P_0\bar{X}^2 - D(1-P_0))}{\bar{X}^2} \text{ and} \\ \hat{\mu_0} = \hat{\theta}_3 &= \frac{(D + \bar{X}^2)}{\bar{X}}, \end{aligned} \tag{6}$$

where $D = \frac{n-1}{n}S^2 - \bar{X}$ and $R = P_0/(1-P_0)$ if $D < 0$ and $R = 1$ otherwise. As expected, it follows from (6) that the sign of $\hat{\delta}$ is determined by the sign of D denoted by $sgn(D)$.

We can use the above estimator $\hat{\delta}$ as a test statistic to test the null hypothesis $H_0 : \delta = 0$ against two-sided or one-sided alternatives. Also we may obtain a confidence interval for δ by deriving the standard error (s.e.) of $\hat{\delta}$. As the form of the $\hat{\delta}$ is complicated, we have at least two options: (a) obtain a bootstrap distribution of $\hat{\delta}$ (Efron and Tibshirani, 1993) or (b) obtain the asymptotic distribution of $\hat{\delta}$ using the M-estimation formulation as given in (4). Although our main interest is in estimating the dispersion parameter δ, by solving (5) (or equivalently (4)), we have also obtained the estimates of μ_0 and π_0 (as given in 6)). We can use (nonparametric) bootstrap to make inference about these underlying (nuisance) parameters as well.

Clearly $\boldsymbol{\psi}(\cdot, x)$ is a smooth function of $\boldsymbol{\theta}$ having a continuous first derivative and also $\boldsymbol{\psi}(\boldsymbol{\theta}, \cdot)$ is a Borel measurable function for all $\boldsymbol{\theta} \in \Theta$, where Θ is an open set as defined above. Thus, it follows that the regularity conditions needed to derive the asymptotic distribution of $\hat{\boldsymbol{\theta}}$ (given by Huber, 1967) are satisfied in this case.

We now derive the so-called "$A = A(\boldsymbol{\theta})$ matrix" and the "$B = B(\boldsymbol{\theta})$ matrix" (see Boos and Stefanski, 2002) to derive the "sandwich estimator" of the asymp-

totic variance, $V = V(\boldsymbol{\theta}) = A^{-1}B(A^{-1})^T$, where

$$\begin{aligned} A(\boldsymbol{\theta}) &= -E_{\boldsymbol{\theta}}\left[\nabla_\theta \boldsymbol{\psi}(\boldsymbol{\theta}, X)\right] \\ B(\boldsymbol{\theta}) &= E_{\boldsymbol{\theta}}[\boldsymbol{\psi}(\boldsymbol{\theta}, X)\boldsymbol{\psi}(\boldsymbol{\theta}, X)^T] \end{aligned}$$

where $\nabla_\theta \boldsymbol{\psi}(\boldsymbol{\theta}, x)$ denotes gradient of $\boldsymbol{\psi}(\boldsymbol{\theta}, x)$ with respect to $\boldsymbol{\theta}$. Notice that $\boldsymbol{\psi}(\boldsymbol{\theta}, x)$ is of the form $a(x) + b(\boldsymbol{\theta})$ for some functions $a(\cdot)$ and $b(\cdot)$ and hence we do not have to compute the expected value of $\nabla_\theta \boldsymbol{\psi}(\boldsymbol{\theta}, x)$ to obtain the "A-matrix". Since $\psi_2(\boldsymbol{\theta}, x) = x(x-1) - \theta_3\omega(\theta_1, \theta_2)$ it follows that in order to have $B(\boldsymbol{\theta})$ finite we need to assume that X has finite fourth moment, i.e., we assume $E_{\boldsymbol{\theta}_0}[X^4] < \infty$ which in turn is equivalent to the assumption that $E[U^4] < \infty$. Notice that we need this latter assumption only for the asymptotic normality; the estimator $\hat{\boldsymbol{\theta}}$ is consistent even when the forth moment is not finite. We now state the main theorem that provides the consistency and asymptotic normality of the estimator $\hat{\boldsymbol{\theta}}$:

THEOREM 2. *Suppose $X_1, \ldots, X_n$ are iid $f(x|\boldsymbol{\theta})$ where $f(\cdot|\cdot)$ is given by (1) with $\theta_1 = \delta, \theta_2 = \pi_0$ and $\theta_3 = \mu_0$. Assume that $\sigma_0^2 < \infty$ and $\mu_0 = \sigma_0^2$. Then the M-estimator $\hat{\boldsymbol{\theta}}$ given by (6) is uniformly consistent.*

Further, if $E[U^4] < \infty$, then $\hat{\boldsymbol{\theta}}$ converges in distribution to a normal distribution. More precisely,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \to N(0, V(\boldsymbol{\theta}_0)), \tag{7}$$

where $V(\boldsymbol{\theta}_0) = A(\boldsymbol{\theta}_0)^{-1}B(\boldsymbol{\theta}_0)(A(\boldsymbol{\theta}_0)^{-1})^T$

Proof: It easily follows that $\boldsymbol{\psi}(\boldsymbol{\theta}, x)$ satisfies all the regularity conditions (B-1)-B(4) of Theorem 2 in Huber (1967) and hence the consistency follows from the general theory of M-estimation.

For asymptotic normality it is possible to show that the regularity conditions (N-1)-(N-4) of Theorem 3 in Huber (1967) are satisfied and hence by the corollary to the Theorem 3 in Huber (1967) the result follows. Some technical details are omitted.

We now derive the exact form of the "A-matrix". Let $a_{ij} = \frac{\partial \psi_i}{\partial \theta_j}(\boldsymbol{\theta}, x)$. As noted earlier a_{ij}'s do not depend on x and hence $A(\boldsymbol{\theta}) = ((a_{ij}(\boldsymbol{\theta})))_{3\times 3}$. It easily follows that,

$$\begin{aligned}
a_{11} &= -\theta_3 \frac{\partial \omega}{\partial \theta_1} = 2\theta_3 \left[\theta_{1+} + \theta_{1-}\frac{\theta_2}{1-\theta_2}\right], \\
a_{12} &= -\theta_3 \frac{\partial \omega}{\partial \theta_2} = -\theta_3 \left(\frac{\theta_{1-}}{1-\theta_2}\right)^2, \\
a_{13} &= -\omega(\theta_1, \theta_2), \\
a_{21} &= \theta_3 a_{11}, a_{22} = \theta_3 a_{12}, a_{23} = 2\theta_3 a_{13} \\
a_{31} &= -2(\theta_{1+} - \theta_1\theta_2), a_{32} = -(1-\theta_{1-})^2 \text{ and } a_{33} = 0
\end{aligned}$$

The closed of expression for the "B-matrix" can also be derived similarly involving up to the forth moments of U. If the distribution of U is specified parametrically (e.g., $U \sim Poisson(\lambda)$), then we can easily obtain a closed form expression for the "B-matrix" similar to the closed form expression of "A-matrix" as given above. However it turns out that for estimation (and to obtain confidence interval), we can estimate the A and B matrices using the following consistent estimators:

$$\begin{aligned}
\hat{A} = A(\hat{\boldsymbol{\theta}}) &= ((a_{ij}(\hat{\boldsymbol{\theta}}))) \\
\hat{B} = \hat{B}(\hat{\boldsymbol{\theta}}) &= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\psi}(\hat{\boldsymbol{\theta}}, X_i)\boldsymbol{\psi}(\hat{\boldsymbol{\theta}}, X_i)^T
\end{aligned}$$

and hence we can obtain $\hat{V} = \hat{A}^{-1}\hat{B}(\hat{A}^{-1})^T$ to obtain the standard errors of $\hat{\boldsymbol{\theta}}$. As $\hat{V}$ is a consistent estimator of $V(\boldsymbol{\theta}_0)$ (see Iverson and Randles, 1989) it follows from Theorem 2 (by an use of the Slutsky's Theorem) that $\sqrt{n}\hat{V}^{-1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ converges in distribution to a trivariate normal distribution with mean vector zero and variance covariance matrix an identity matrix of order 3. Hence, it follows that, $\sqrt{n}(\hat{\delta} - \delta_0) \sim AN(0, \mathbf{e}_1^T\hat{V}\mathbf{e}_1)$ where $\mathbf{e}_1 = (1,0,0)^T$.

Alternatively, as the forth moment condition can not be checked in general, we may prefer the use of the bootstrap method (Efron and Tibshirani, 1993) to make inference about δ using the closed form expression of $\hat{\delta}$ given by (6). Next, we

occurrences	5-word count	10-word count
0	45	27
1	49	44
2	6	26
> 2	0	3
$\bar{x}$	0.61	1.05
s^2	0.36	0.65

Table 1: Frequency of function word counts.

apply these results to a data set and study the performance of the method using a simulation study motivated by the real data application.

4 Simulation and Data Analysis

In this section, first we analyze the data listed in Table 1 to illustrate the methods described in Section 3. In addition to the asymptotic s.e. of the estimator $\hat{\delta}$ we also compute a bootstrap distribution of $\hat{\delta}$ and compare its value to the asymptotic s.e. obtained by the M-estimation theory. We conclude this section with a simulation study motivated by this real application to see the performance of the proposed estimates.

4.1 Application to function word count data

First, we analyze the data presented in Table 1 which presents a count of function words for $n = 100$ words. In studies aimed at characterizing an author's style, samples of m words are taken and the number of function words in each sample counted. Often binomial or Poisson distributions are assumed to hold for the proportions of function words. Table 1 shows the combined frequencies of the

	5-word count	10-word count
$\hat{\delta}$	-0.673	-0.705
Bootstrap s.e.	0.054	0.066
M-estimate s.e.	0.051	0.062
Bootstrap 95% C.I.	$(-0.764, -0.558)$	$(-0.815, -0.557)$
M-estimate 95% C.I.	$(-0.773, -0.573)$	$(-0.826, -0.583)$

Table 2: Estimates for word count data

articles "the", "a" and "an" in samples from McCauley's "Essay on Milton", taken from the Oxford edition of Macualey's (1923) literary essays. Non-overlapping samples were drawn from opening words of two randomly chosen lines from each of 50 pages of printed text, 10 word samples being simply extensions of 5 word samples. As $\bar{x} > s^2$, the data show clear evidence of underdispersion for both 5-word and 10-word counts.

Clearly if we use a model that represents only one type of dispersion (e.g., the regular Poisson, Negative Binomial or zero inflated Poisson distributions) it will not provide adequate fit to these samples. Also any parametric assumption on the distribution might as well influence the test for underdispersion. In this respect our proposed semi-parametric method provides the most flexibility in testing the hypothesis $H_0 : \delta = 0$.

The results are presented in Table 2. From this table it follows that the hypothesis of equidispersion can be rejected in favor of underdispersion using almost any level of significance. From Figure 1 it is also clearly evident that the bootstrap distributions of $\hat{\delta}$'s lie entirely to the left of zero, indicating a strong support for underdispersion. These bootstrap distributions and corresponding estimates are based on $B = 5000$ samples. We study the power of $\hat{\delta}$ in detecting the nature of dispersion using a simulation study motivated by this data sets.

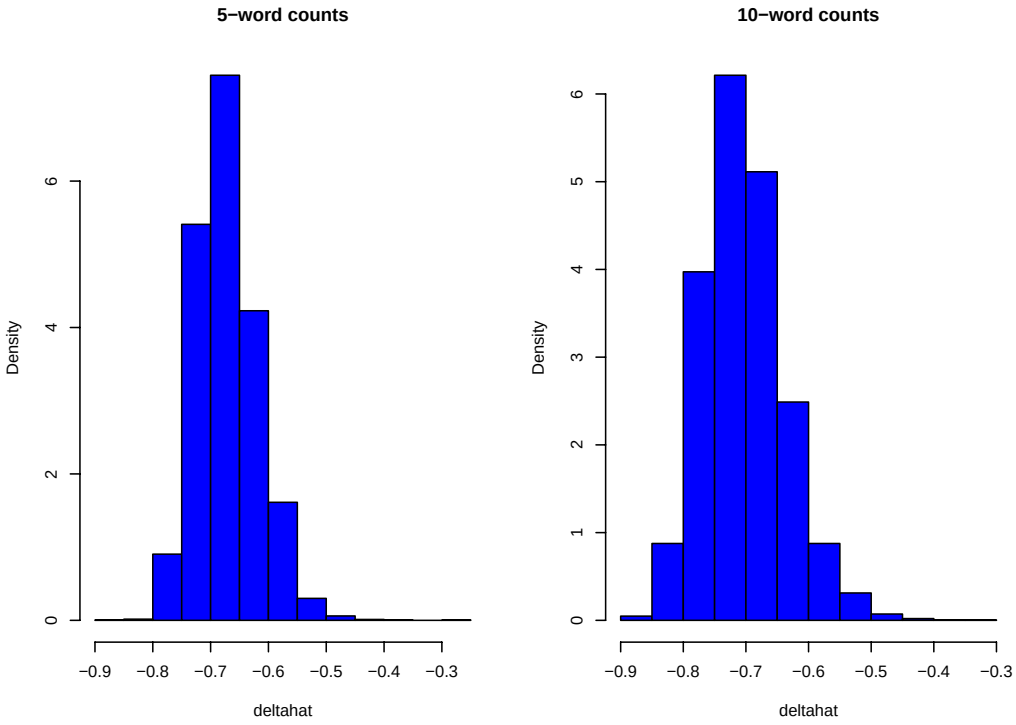

Figure 1: Bootstrap distribution of $\hat{\delta}$'s.

4.2 A Simulation study

In this section we generate data sets of size n from three discrete distributions: (a) $Binomial(m, \frac{\mu}{m})$ (an underdispersed distribution), (b) $Poisson(\mu)$ (an equidispersed distribution) and (c) $ZIP(p, \frac{\mu}{1-p})$ (an overdispersed distribution). Here $ZIP(p, \frac{\mu}{1-p})$ denotes a zero inflated Poisson (ZIP) distribution which is a special case of zero-altered Poisson distribution with $\delta = \sqrt{p}$ and $f_0 \sim Poi(\frac{\mu}{1-p})$ (see (1)). Notice that $\mu > 0$ is the mean of each of the three distributions. We fix $\mu = 1$ for all three distributions and choose $m = 3$ and $p = 2/3$ for non-equidispersed models; these choices being motivated by the above real data. Notice that by this choice the true value of σ^2/μ is 2/3 and 3/2 under binomial and zero inflated Poisson sampling, respectively. From the Theorem 4 in the Appendix A, it follows

	$Bin(3,1/3)$	$Poi(1)$	$ZIP(2/3,3/\sqrt{2})$
	$n=30$		
$\hat{\delta}$	-0.643	-0.156	0.764
$s.e.(\hat{\delta})$	0.185	0.347	0.225
$0 \in$ 95% C. I.	0.336	0.865	0.487
	$n=100$		
$\hat{\delta}$	-0.663	-0.079	0.805
$s.e.(\hat{\delta})$	0.084	0.288	0.053
$0 \in$ 95% C. I.	0.040	0.914	0.007
	$n=500$		
$\hat{\delta}$	-0.665	-0.037	0.814
$s.e.(\hat{\delta})$	0.033	0.208	0.020
$0 \in$ 95% C. I.	0.000	0.928	0.000

Table 3: Results based on simulation study

that the true value of δ is about 0.6647 and 0.8165 for the chosen binomial and ZIP distributions, respectively (see the examples in Appendix A).

We present results for $n = 30, 100$ and 500, which represents small, moderate and large sample sizes, respectively, based on $N = 1000$ Monte Carlo (MC) simulations runs. To compute the standard errors and 95% confidence intervals we used a bootstrap samples of size $B = 500$.

The results are summarized in numerically in Table 3 and graphically in Figure 2. In Table 3 we provide the average value (based on $N = 1000$ NC runs) of the $\hat{\delta}$ obtained by using (6). We also provide the standard error estimate and the proportion of times the 95% confidence intervals contained the null value $\delta = 0$. The standard error and 95% confidence interval (C.I.) estimates were obtained by the bootstrap method. It is clearly evident as the sample size increases the power of detecting the nature of dispersion increases. For instance notice that

when $n = 100$, percentage of 95% C.I.'s that contained the null value 0 is about 4% under binomial sampling and is about 0.7% under ZIP sampling, whereas it is about 91.4% under Poisson sampling.

More insights can be obtained from Figure 2, where we present the boxplot of the $\hat{\delta}$'s along with the average values of the lower and upped bound of the 95% C.I's indicated by the dashed lines. The vertical axes of each plot has been set the minimum and maximum of the lower and upper bounds of the 95% C.I.'s, respectively. It is clear that the proposed method works remarkably well to detect both underdispersion and overdispersion. Similar results were observed when repeated the simulation study with various other choices of the δ's and underlying distributions. As the procedure is semiparametric, the results were observed to be fairly robust against a variety of assumed underlying distributions. A R code to obtain $\hat{\delta}$ and associated C.I.'s can be obtained from the first author.

5 Conclusions

Zero-altered models have been shown to be useful for modeling outcomes of manufacturing processes and other situations where count data with too few or too many zeros are encountered. The proposed semi-parametric method of estimation provides a flexible yet simple framework to model any discrete distribution with finite second moment. Consistency and asymptotic normality of the proposed estimator makes it straightforward to apply the method in practice. Alternatively a bootstrap method is found to be suitable for data with small sample size.

In the presence of covariates, proposed generalized zero-altered models can be extended to include the effect of such predictor variables using suitable link functions on δ. For instance, as $\delta \in (-1, 1)$, a Fisher's z-transformation, given by $\eta = \log \frac{1+\delta}{1-\delta}$, might serve as an useful link function. Results from such models are under investigation by the authors and will be presented elsewhere.

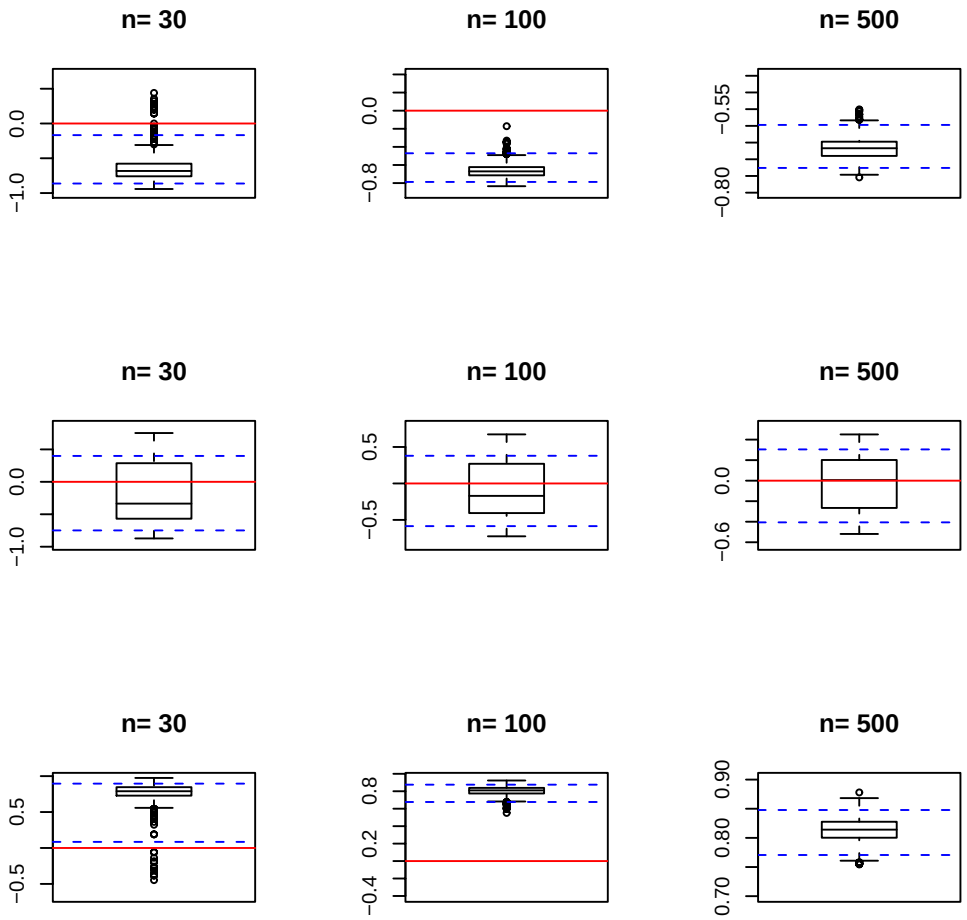

Figure 2: Performance of $\hat{\delta}$: (a) first row is based on binomial sampling, (b) second row is based on poisson sampling and (c) third row is based on zero inflated Poisson sampling (for details see Section 4.2)

Acknowledgments The second author appreciates hospitality provided by the Department of Statistics at NC State University, during his visit. The second author's work was supported by the Research Foundation of Chungnam National University in Korea..

References

Bailey, B.J.R. (1990). A model for function word counts, *Applied Statistics*, **39**, 107-114.

Boos, D. and Stefanski, L. A. (2002). The calculus of M estimation, *The American Statistician*, **56**, 29-38.

Cameron, A. C. and Trivedi, P. K. (1986). Econometric models based on count data: comparisons and applications of some estimators and tests, *Journal of Econometrics*, **1**, 29-53.

Castillo and Perez-Casany (2005). Overdispersed and underdispersed Poisson generalizations, *Journal of Statistical Planning and Inference*, **134**, 486-500.

Dahiya, R. C. and Gross, A. J. (1973). Estimating the zero class from a truncated Poisson sample, *Journal of American Statistical Association*, **68**, 731-733.

Efron, B. and Tibshirani, (1993). *An Introduction to Bootstrap*, Chapman and Hall, New York.

Famoye, F. (1993). Restricted generalized Poisson regression model, *Communications in Statistics - Theory and Methods*, **22**, 1335-1354.

Gupta, P. L. , Gupta, R.C., and Tripathi, R.C.(1996). Analysis of zero-adjusted count data, *Computational Statistics & Data Analysis*, **23**, 207-218.

Gurmu, S. (1997). Semiparametric estimation of hurdle regression models with an application to Medicaid utilization, *Journal of Applied Econometrics*, **12**, 225-242.

Hall, D.B. (2000). Zero-inflated Poisson and binomial regression with random effects: a case study, *Biometrics*, **56**, 1030-1039.

Heilbron, D.C. (1994). Zero-altered and other regression models for count data with added zeroes, *Biometrical Journal*, **36**, 531-547.

Heilbron, D. C., and Gibson, D. R. (1990). Shared needle use and health beliefs concerning AIDS: Regression modeling of zero-heavy count data, Poster session, 6th International conference on AIDS, San Francisco, CA.

Hinde, J. and Demetrio, C. (1998). Overdispersion: models and estimation, *Computational Statistics and Data Analysis*, **27**, 151-170.

Huber, P. J. (1967). The behaviour of maximum likelihood estimates under non-standard conditions, Proc. Fifth Berkeley Symp. Math. Statist. Prob., **1**, 221-233.

Iverson, h. K. and Randles, R. H. (1989). The effects on convergence of substituting estimates into U-statistics and other families of statistics, *Probability Theory and related Fields*, **81**, 453-471.

Johnson, N. L., Kotz, S., Kemp, A. W. (1992). *Univariate Discrete Distributions*, Wiley, New York.

Lambert, D. (1992). Zero-inflated Poisson regression with an application to defects in manufacturing, *Technometrics*, **34**, 1-14.

Rao, C. R. (1965). On discrete distributions arising out of ascertainment, *Sankhya Ser. A*, , 311-324.

Ridout, M., Demetrio, C. G. B., and Hinde, J. (1998). Models for count data with many zeros, International Biometric Conference, Cape Town.

Tu, W. (2002). Zero-inflated data, *Encyclopedia of Environmetrics*, **4**, 2387-2391.

Umbach, D. (1981). On inference for a mixture of Poisson and a degenerate distribution, *Communication Statistics, Series A*, **10**, 299-306.

Welsh, A, Cunningham, R., Donnelly, C. and Lindenmayer, D. (1996). Modeling the abundance of rare species - statistical models for count with extra zeros, *Ecological Modeling*, **88**, 297-308.

Winkelmann, R. and Zimmermann, K. F. (1995). Recent developments in count data modeling: Theory and applications, *Journal of Economic Surveys*, **9**, 1-24.

Yip, P.(1988). Inference about the mean of a Poisson distribution in the presence of a nuisance parameter, *Australian Journal of Statistics*, **30**, 299-306.

Appendix A: Additional Results

Here we present conditions for the under, equi and over dispersion of $X \sim f_\delta$ in terms of the first two moments of $U \sim f_0$. This extends the result given by Theorem 1.

THEOREM 3. *Suppose $X \sim f_\delta$, where f_δ is given by (1) and assume that the underlying random variable $U \sim f_0$ with $\sigma_0^2 < \infty$. The X is*

$$\begin{aligned} & \textit{under-dispersed i.e., } \mu > \sigma^2 \quad \textit{if} \quad \delta < -\delta_0 \\ & \textit{equi-dispersed i.e., } \mu = \sigma^2 \quad \textit{if} \quad \delta = \delta_0 \\ \textit{and} \quad & \textit{over-dispersed i.e., } \mu < \sigma^2 \quad \textit{if} \quad \delta > \delta_0 \end{aligned}$$

where the cut-off value δ_0 is a function of μ_0, σ_0^2 and π_0 given by

$$\delta_0 = \begin{cases} \sqrt{\frac{\mu_0 - \sigma_0^2}{\mu_0^2}} & \textit{if } \mu_0 > \sigma_0^2 \\ \sqrt{\frac{1-\pi_0}{\pi_0} \cdot \frac{\sigma_0^2 - \mu_0}{\mu_0^2}} & \textit{if } \mu_0 \leq \sigma_0^2 \end{cases}$$

Proof: From Lemma 1, it follows that

$$\sigma^2 - \mu = \frac{\omega(\delta)}{\mu_0^2} \left((1 - \omega(\delta)) - \frac{\mu_0 - \sigma_0^2}{\mu_0^2} \right).$$

As $\omega(\delta) > 0$ if $\delta < 1$, it follows that $\sigma^2 \geq \mu$ if and only if $1 - \omega(\delta) \geq (\mu_0 - \sigma_0^2)/\mu_0^2$. But notice that,

$$1 - \omega(\delta) = \begin{cases} \delta^2 & \text{if } \delta \geq 0 \\ -\delta^2 \frac{\pi_0}{1-\pi_0} & \text{if } \delta < 0 \end{cases}.$$

Hence the result follows.

Next we show that under a mild condition any discrete distribution having finite second moment can be represented by $f_\delta(\cdot)$ for some $\delta \in (-1, 1)$ where the underlying pmf $f_0(\cdot)$ can be chosen to be equidispersed.

THEOREM 4. *Let X be random variable with pmf $g(\cdot)$ such that $\mu = \sum_{x=0}^{\infty} xg(x)$ and $\sigma^2 = \sum_{x=0}^{\infty}(x-\mu)^2 g(x) < \infty$. Assume that $g(0) + g(1) < 1$. Then there is a unique $\delta \in (-1, 1)$ and pmf $f_0(\cdot)$ such that $f_\delta(x) = g(x)$ for all $x = 0, 1, 2 \ldots$, where $f_\delta(\cdot)$ is as given in (1) and $\mu_0 = \sigma_0^2$.*

Proof: Define $\omega = \left[1 + \frac{\sigma^2-\mu}{\mu^2}\right]^{-1}$. Notice that $\omega > 0$ as $E[X(X-1)] = \mu^2+\sigma^2-\mu > 0$ by the assumption. Define,

$$f_0(x) = \begin{cases} 1 - \frac{1-g(0)}{\omega} & \text{if } x = 0 \\ \frac{g(x)}{\omega} & \text{if } x = 1, 2 \ldots \end{cases} \tag{8}$$

Clearly, $f_0(\cdot)$ defined by (8) is a pmf and it satisfies $\mu_0 = \sigma_0^2$ where $\mu_0 = \sum_{u=0}^{\infty} u f_0(u)$ and $\sigma_0^2 = \sum_{u=0}^{\infty}(u-\mu_0)^2 f_0(u)$. Finally, δ can be obtained by solving the equation $\delta_+^2 - \delta_-^2 \frac{f_0(0)}{1-f_0(0)} = 1 - \omega$. Notice that $\omega \leq 1$ if and only if $\sigma^2 \geq \mu$. Thus, it follows that δ is given by

$$\delta = \begin{cases} \sqrt{1-\omega} = \sqrt{\frac{\sigma^2-\mu}{\sigma^2-\mu+\mu^2}} & \text{if } \omega \leq 1 \text{ (i.e. } \sigma^2 \geq \mu) \\ \sqrt{\frac{(\omega-1)(1-g(0))}{\omega-(1-g(0))}} = \sqrt{\frac{\mu-\sigma^2}{\mu-\sigma^2+\frac{g(0)}{1-g(0)}\mu^2}} & \text{if } \omega > 1 \text{ (i.e. } \sigma^2 < \mu). \end{cases}$$

This completes the proof.

Some Examples:

(i) Suppose $X \sim Bin(m, p)$, then δ is given by

$$\delta = \left[1 + \frac{m(1-p)^m}{1-(1-p)^m}\right]^{-1/2}$$

(ii) Suppose $X \sim Poi(\lambda)$, then $\delta = 0$.

(iii) Suppose $X \sim NegBin(m, p)$, then $\delta = 1/\sqrt{m+1}$.

(iv) Suppose $X \sim ZIP(p, \lambda)$, then $\delta = \sqrt{p}$.

The corresponding $f_0(\cdot)$ for all of the above examples can be obtained from (8).