Estimation of Zero-Inflated Population Mean with Highly Skewed Nonzero Component: A Bootstrapping Approach

Khyam Paneru [1,*], R. Noah Padgett [2], Hanfeng Chen [3]

[1]*Department of Mathematics, The University of Tampa, Tampa, FL, USA*
[2]*Department of Educational Psychology, Baylor University, Waco, TX, USA*
[3]*Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, OH, USA*

Abstract   This paper adopts a bootstrap procedure in the maximum pseudo-likelihood method under probability sampling designs. It estimates the mean of a population that is a mixture of excess zero and a nonzero skewed sub-population. Simulations studies show that the bootstrap confidence intervals for zero-inflated log-normal population consistently capture the true mean. The proposed method is applied to a real-life data set.

1. Introduction

An underlying population that contains many zeros appears in a wide range of literature, from social sciences to natural sciences. Data with too many zero values are common in several applications such as insurance, reliability, meteorology, ecology, auditing, and manufacturing; Paneru [15] cited examples of data with too many zero values in various disciplines. In meteorology, for example, a geographical region may have many days with no rainfall (zero precipitation is recorded); the rest of the days are recorded with a value of positive precipitation. Similarly, in a production process, a manufacturing plant aims to produce a large proportion of non-defective items (a zero-defect is recorded) and a small proportion of defective items where the number of defects follows a probability distribution [11, 14]. In general, zero-inflated data has a large spike of zero values with a smaller proportion of nonzero values and is thus called a zero-inflated population (ZIP). These examples suggest that a zero-inflated population could consist of a nonzero sub-population with continuous or count data.

Excess zeros and skewness in the zero-inflated population are commonly encountered phenomena that limit traditional methods to estimate the population mean. Similar issues happen in developing regression models for the zero-inflated data. Estimating the population mean based on normal approximation or the central limit theorem (CLT) gives poor estimates because of apparent skewness caused by zero values, skewness in the population's nonzero component, or both. Different approaches and techniques are developed to overcome the issue of zero inflation. In a two-component mixture model, the population is split into zero and nonzero components, where

---

*Correspondence to: Khyam Paneru (Email: kpaneru@ut.edu). Department of Mathematics, The University of Tampa. 401 W Kennedy Blvd, Tampa, FL 33606, USA.

the nonzero component follows a known probability distribution [10]. Under parametric assumptions, likelihood ratio statistics and bootstrap techniques are applied to construct confidence intervals for the mean of diagnostic test charge data containing many zeros [26]. Other approaches to construct confidence intervals of zero-inflated population mean include a pseudo empirical likelihood approach in complex sampling surveys [2], a mixture model of true zero exposures, and a log-normal distribution with left censoring [22]. The zero-inflated log-normal distribution is important in real-life as many populations are highly skewed, such as rainfall, air contamination exposure, and diagnostic test charge; and different approaches exist to compare the mean of such populations [27, 23, 25, 26, 12]. Satter and Zhao [20] proposed empirical and adjusted empirical likelihood methods to construct nonparametric confidence intervals for the mean of a zero-inflated population. Satter and Zhao [21] further extended their empirical methods to the Jackknife empirical likelihood method to estimate the difference in means of two zero-inflated skewed populations. Pittman *et al.* [19] describe different methods for analyzing overdispersed zero-inflated count data and illustrate their application to cigarette and marijuana smoking data. Furthermore, modeling of zero-inflated count data and applications have been studied by many researchers, e.g., [24, 9, 3, 13, 4]. Although, many of the existing methods for estimating the mean of ZIPs do not address the problem when the data is produced through complex probability sampling designs such as stratified sampling and cluster sampling, which are commonly applied to survey sampling and designs. Chen *et al.* [1] proposed a maximum pseudo-likelihood approach under complex probability sampling designs for interval estimation of zero-inflated population mean. Paneru and Chen [16, 17] extended the pseudo-likelihood approach to develop regression models for the zero-inflated data.

Although the pseudo-likelihood approach addresses complex sampling designs and shows confidence intervals have better coverage than existing methods, the method reliant on pseudo-likelihood ratio test statistic to produce confidence intervals under the pseudo-likelihood approach is mathematically and computationally complex. To overcome the complexity, Paneru *et al.* [18] adopted a pseudo-likelihood approach in a bootstrap procedure, an alternative method for constructing interval estimates for the mean of the zero-inflated population. The bootstrap procedure is mathematically, computationally, and intuitively more straightforward than existing methods. Paneru *et al.* [18] provide the application in the normal model, assuming that the nonzero component is normally distributed.

This paper extends Paneru *et al.* [18] to highly skewed data where a nonzero component brings extra skewness in a zero-inflated population. This paper describes the two-component mixture model, maximum pseud-likelihood function, and derivation of parameter estimates in the zero-inflated log-normal mixture model. It explains the bootstrapping procedure used, along with the results of the simulations. It also demonstrates the application to actual data, the Canadian Labour Force Survey-2000 detailing extra work hours.

2. Methodology

This paper follows the concept of the two-component parametric mixture model from [10], the maximum pseudo-likelihood approach developed in [1], and the bootstrap procedure adopted in [18]. In ZIPs, two components exist; one consists solely of zero values and the second component of nonzero values that adheres to some known probability distribution. The two-component mixture model for the zero-inflated population is defined by

$$h(y; \alpha, \mu, \sigma) = \alpha f(y; \mu, \sigma) I_{(y \neq 0)} + (1 - \alpha) I_{(y=0)}, \tag{1}$$

where α is the proportion of nonzero values, μ is the mean and σ is the nuisance parameter of the nonzero component, and I is the indicator function. The parameter of interest, the mean of the zero-inflated population, is

$$\theta = \alpha\mu.$$

Consider a survey population of N units with values $y_1, \cdots, y_N$ be independently generated from the model (1). The log-likelihood function of all N sampling units of survey population is

$$l(\alpha, \mu, \sigma) = \sum_{i=1}^{N} \log h(y_i; \alpha, \mu, \sigma).$$

Let s be a random subset of n sampling units with values $y_1, \cdots, y_n$ selected from the surveyed population. Let $m(<n)$) be the number of zero values in n observed units arranged as

$$y_i \quad \neq 0 \quad \text{for } i = 1, \cdots, n - m$$
$$= 0 \quad \text{for } i = n - m + 1, \cdots, n.$$

As explained in [1], consider a probability sampling design, where the random subset s of n sampling units is obtained from the surveyed population with inclusion probability $\pi_i, i = 1, \cdots, n$. Under the probability sampling design, the pseudo-likelihood function, an estimate of the log-likelihood function, $l(\alpha, \mu, \sigma)$, is defined by

$$\hat{l}(\alpha, \mu, \sigma) = \sum_{i \in s} w_i \log h(y_i; \alpha, \mu, \sigma),$$

where the sampling weights $w_i = 1/\pi_i, i = 1, \cdots, n$ are chosen such that $E(\hat{l}) = l$.

In real-world applications, zero-inflated data is complex to analyze due to a spike at zero with addition skewness in nonzero component. To address the issue, this paper considers the nonzero component to follow a log-normal distribution in the two-component model (1). The pdf of the log-normal component of the model (1) is

$$f(y; \mu, \sigma) = \frac{1}{y\sqrt{2\pi\sigma^2}} e^{-\frac{(\ln y - \xi)^2}{2\sigma^2}}, \ y > 0,$$

where $\mu = E(Y|Y > 0) = \exp(\xi + \frac{\sigma^2}{2})$. The pseudo-likelihood function and the estimation of parameters for the zero-inflated log-normal population is given below.

$$
\begin{aligned}
\hat{l}(\alpha, \mu, \sigma) &= \sum_{i \in s} w_i \log h(y_i; \alpha, \mu, \sigma) \\
&= \sum_{i \in s} w_i \log\{\alpha f(y_i; \mu, \sigma) I_{(y_i \neq 0)} + (1 - \alpha) I_{(y_i = 0)}\} \\
&= \sum_{i=1}^{n-m} w_i \log\{\alpha f(y_i; \mu, \sigma) I_{(y_i \neq 0)}\} + \sum_{i=n-m+1}^{n} w_i \log\{(1 - \alpha) I_{(y_i = 0)}\} \\
&= \sum_{i=1}^{n-m} w_i \log \alpha + \sum_{i=1}^{n-m} w_i \log\left\{\frac{1}{y_i\sqrt{2\pi\sigma^2}} e^{-\frac{(\ln y_i - \xi)^2}{2\sigma^2}}\right\} + \sum_{i=n-m+1}^{n} w_i \log(1 - \alpha) \\
&= \sum_{i=1}^{n-m} w_i \log \alpha + \sum_{i=n-m+1}^{n} w_i \log(1 - \alpha) + \sum_{i=1}^{n-m} w_i \log\left\{\frac{1}{y_i\sigma\sqrt{2\pi}}\right\} - \sum_{i=1}^{n-m} \frac{w_i(\ln y_i - \xi)^2}{2\sigma^2} \\
&= w^+ \log \alpha + w^0 \log(1 - \alpha) - w^+ \log \sigma - \sum_{i=1}^{n-m} \frac{w_i(\ln y_i - \xi)^2}{2\sigma^2} + C,
\end{aligned}
$$

where

$$w^+ = \sum_{i=1}^{n-m} w_i \ \text{and} \ w^0 = \sum_{i=n-m+1}^{n} w_i.$$

Pseudo-Likelihood Estimates of Parameters:

By Setting up $\dfrac{\partial \hat{l}}{\partial \alpha} = 0$, the estimate of the proportion of non-zeros, α, is given by

$$\hat{\alpha} = \frac{w^+}{w^+ + w^0}.$$

We derive the estimate of ξ by setting up $\dfrac{\partial \hat{l}}{\partial \xi} = 0$. So,

$$\frac{1}{\sigma^2}\left[\sum_{i=1}^{n-m} w_i \ln y_i - \sum_{i=1}^{n-m} w_i \xi\right] = 0$$

$$\sum_{i=1}^{n-m} w_i \ln y_i - w^+ \xi = 0.$$

Therefore,

$$\hat{\xi} = \frac{1}{w^+}\sum_{i=1}^{n-m} w_i \ln y_i.$$

Similarly, the estimate of σ^2 is derived as below.

$$
\begin{aligned}
\frac{\partial \hat{l}}{\partial \sigma} &= \frac{\partial}{\partial \sigma}\left[w^+ \log \alpha + w^0 \log(1-\alpha) - w^+ \log \sigma - \sum_{i=1}^{n-m}\frac{w_i(\ln y_i - \xi)^2}{2\sigma^2} + C\right] \\
&= 0 + \frac{\partial}{\partial \sigma}\left[-w^+ \log \sigma - \sum_{i=1}^{n-m}\frac{w_i(\ln y_i - \xi)^2}{2\sigma^2}\right] + 0 \\
&= -\frac{w^+}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n-m} w_i(\ln y_i - \xi)^2.
\end{aligned}
$$

Then, by setting $\dfrac{\partial \hat{l}}{\partial \sigma} = 0$,

$$\hat{\sigma}^2 = \frac{1}{w^+}\sum_{i=1}^{n-m} w_i(\ln y_i - \xi)^2.$$

In summary, the pseudo-likelihood estimates of the parameters under the log-normal model are given by

$$\hat{\alpha} = \frac{w^+}{w^+ + w^0} \tag{2}$$

$$\hat{\xi} = \frac{1}{w^+}\sum_{i=1}^{n-m} w_i \ln y_i \tag{3}$$

$$\hat{\sigma}^2 = \frac{1}{w^+}\sum_{i=1}^{n-m} w_i(\ln y_i - \xi)^2 \tag{4}$$

$$\hat{\mu} = \exp\left(\hat{\xi} + \frac{\hat{\sigma}^2}{2}\right) \tag{5}$$

Thus, the pseudo-likelihood estimate of the mean of the zero-inflated log-normal population is given by

$$\hat{\theta} = E(Y) = \hat{\alpha}\hat{\mu} \tag{6}$$

3. Bootstrap Estimation and Simulation Results

The bootstrap technique was introduced by Efron [5], and further developments are found in Efron [6, 7]. The bootstrap technique's logic is that computing-based methods' power can overcome the complexity of standard statistical approaches [8]. This paper uses the bootstrap approach where the data follow a specified two-component mixture model as detailed in the model (1). This paper also presents the simulation results of parametric and nonparametric bootstrap estimates. We conduct simulation studies to calculate percentile bootstrap confidence intervals of $\theta = \alpha\mu$, the mean of zero-inflated populations. It uses the empirical distribution of bootstrap replicates as the reference sampling distribution of $\hat{\theta}$, pseudo-likelihood estimation under complex probability sampling designs given in equation (6) and $B(= 10,000)$ bootstrap replicates $\hat{\theta}_1, \cdots, \hat{\theta}_B$ of $\hat{\theta}$ are computed. Based on the bootstrap replicates, 95% bootstrap percentile intervals are computed with the quantiles $\hat{\theta}_{0.025}$ and $\hat{\theta}_{0.975}$ as lower and upper bounds, respectively. Different values of α (proportion of nonzero observations) are considered in simulations.

For the simulation studies of the zero-inflated log-normal model, finite populations of size $10,000(N = 10,000)$ are randomly generated and divided into four $(k = 4)$ strata of sizes $500, 1500, 3000$, and 5000. For a random sample of size $n = 500$, the inclusion probability for each stratum is set to $\pi_1 = 140/500, \pi_2 = 120/1500, \pi_3 = 130/3000$, and $\pi_4 = 110/5000$. The corresponding weights for each stratum are set to $w_1 = 500/140, w_2 = 1500/120, w_3 = 3000/130$ and $w_4 = 5000/110$, respectively. A random size $n_j, j = 1, \cdots, 4$, is drawn from each stratum using simple random sampling without replacement.

In the first population condition, the nonzero component is generated using $\mu = 100$ and $\sigma = 30$. Table 1 and Figure 1 present the resulting bootstrap CIs and distributions, respectively. We find that the parametric bootstrap method results in narrower intervals across all levels of nonzero proportion, α. At low levels α, bootstrap percentile intervals are less centered on the population value compared to those at high levels of α. In the second population condition, the nonzero component is generated using $\mu = 50$ and $\sigma = 25$. Table 2 and Figure 2 present the resulting bootstrap CIs and distributions, respectively. Results are similar to the first condition. A notable difference is that the nonparametric bootstrap CIs are more heavily weighted towards zero. The tendency toward zero meant all CI for the nonparametric approach is shifted to the left of the population value even though the intervals still contained the parameter of interest. In the final population condition, the nonzero component is generated using $\mu = 20$ and $\sigma = 15$, and results are presented in Table 3 and Figure 3. Results are similar to the previous conditions. For $\alpha = 0.50$, we observe a narrow overlap between the nonparametric and parametric bootstrap CIs which can be seen in the bootstrap distributions presented in Figure 3. Caution may be needed in interpreting bootstrap intervals in populations with a small mean.

In tables 1, 2, and 3 presented below, $\hat{\theta}$ is the pseudo-maximum likelihood estimate from simulated data, $LL_{.025}$ and $UL_{.975}$ represents the lower-limit and upper-limit, respectively, of the confidence intervals, and $\bar{X}$ is the bootstrap sample mean. Similarly, in figures 1, 2, and 3 displayed below, vertical dashed lines represent the true value of θ at each level of α.

Table 1. 95% Bootstrap confidence intervals of θ for given $\mu = 100$ (location $= 4.56$) and $\sigma = 25$ (shape $= 0.29$) at $\alpha = 0.05, 0.10, 0.20$, and 0.50.

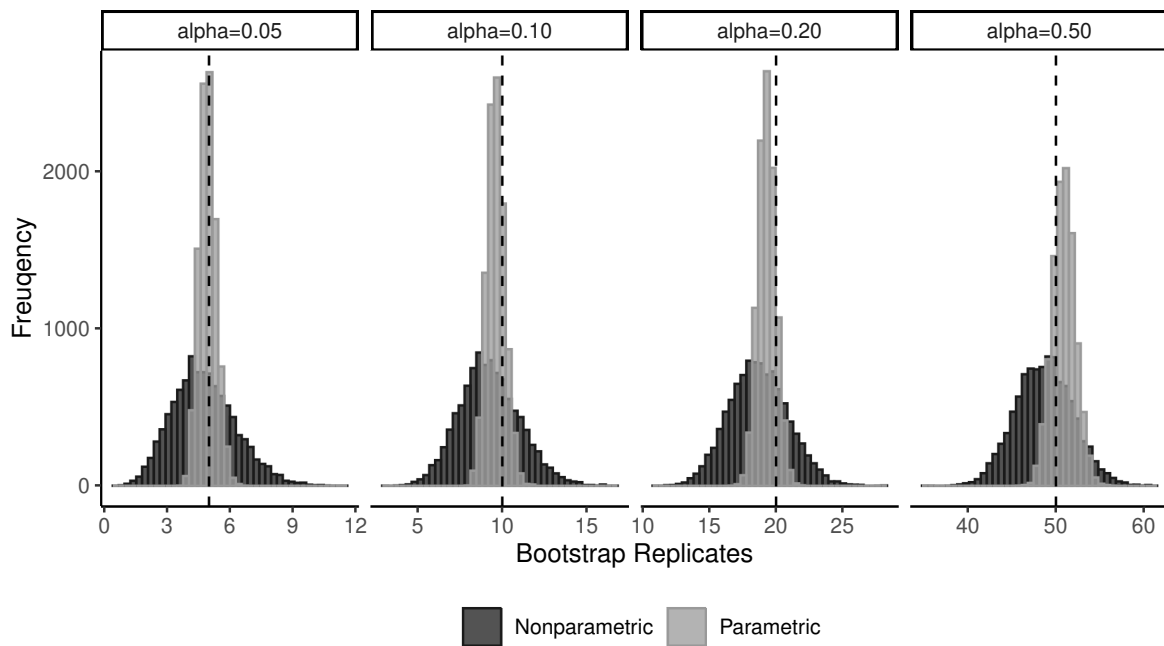|			Nonparametric			Parametric		
α	True Value θ	$\hat{\theta}$	$LL_{.025}$	$UL_{.975}$	$\bar{X}$	$LL_{.025}$	$UL_{.975}$	$\bar{X}$
0.05	5	3.924	1.634	6.612	3.878	3.305	5.307	4.209
0.10	10	8.936	6.169	11.898	8.926	8.414	10.574	9.440
0.20	20	18.940	14.303	24.012	18.960	18.321	21.708	19.937
0.50	50	48.252	42.183	54.518	48.288	47.982	52.553	50.186

Figure 1. Bootstrap distribution of $\hat{\theta}$ for given $\mu = 100$ (location $= 4.56$) and $\sigma = 25$ (shape $= 0.29$) at $\alpha = 0.05, 0.10, 0.20$, and 0.50.

Table 2. 95% Bootstrap confidence intervals of θ for given $\mu =50$ (location=3.80) and $\sigma =25$ (shape=0.47) at $\alpha = 0.05, 0.10, 0.20$, and 0.50.

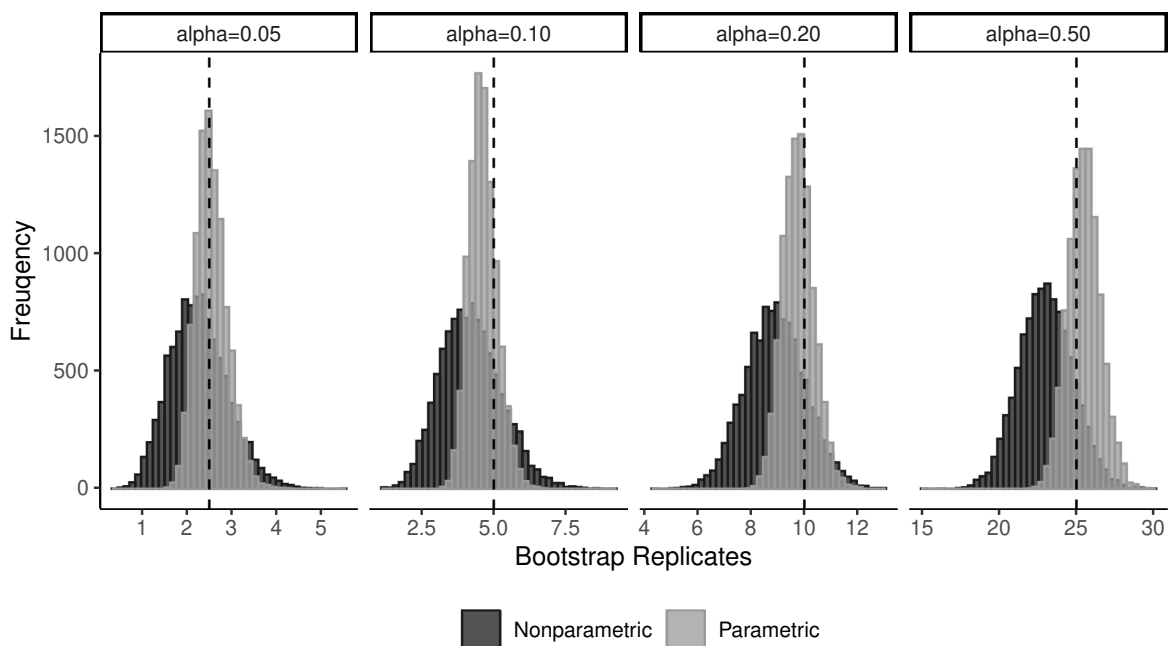| α	True Value θ	$\hat{\theta}$	Nonparametric			Parametric		
			$LL_{.025}$	$UL_{.975}$	$\bar{X}$	$LL_{.025}$	$UL_{.975}$	$\bar{X}$
0.05	2.5	2.231	1.222	3.386	2.222	1.966	3.246	2.541
0.10	5	4.365	2.857	6.033	4.368	4.056	5.430	4.698
0.20	10	9.308	7.049	11.754	9.320	9.223	12.274	10.671
0.50	25	22.708	19.322	26.370	22.721	23.380	27.537	25.398

Figure 2. Bootstrap distribution of $\hat{\theta}$ for given $\mu = 50$ (location $= 3.80$)and $\sigma = 25$ (shape $= 0.47$) at $\alpha = 0.05, 0.10, 0.20$, and 0.50.

Table 3. 95% Bootstrap confidence intervals of θ for given $\mu = 20$ (location=2.77) and $\sigma = 15$ (shape = 0.67) at $\alpha = 0.05, 0.10, 0.20,$ and 0.50.

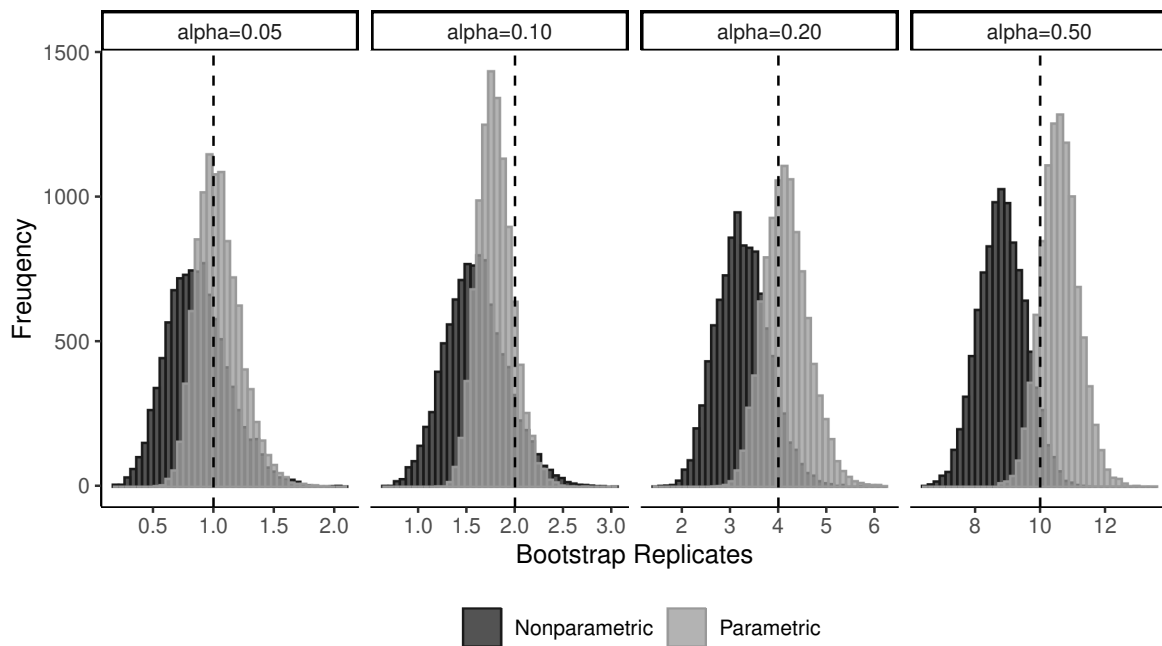| α	True Value θ	$\hat{\theta}$	Nonparametric			Parametric		
			$LL_{.025}$	$UL_{.975}$	$\bar{X}$	$LL_{.025}$	$UL_{.975}$	$\bar{X}$
0.05	1	0.862	0.416	1.416	0.861	0.732	1.323	0.990
0.10	2	1.676	1.016	2.427	1.674	1.579	2.402	1.945
0.20	4	3.502	2.548	4.534	3.497	3.607	5.307	4.373
0.50	10	8.936	7.498	10.492	8.940	9.687	12.049	10.825

Figure 3. Bootstrap distribution of $\hat{\theta}$ for given $\mu = 20$ (location = 2.77), $\sigma = 15$ (shape= 0.67) at $\alpha = 0.05, 0.10, 0.20,$ and 0.50.

In the simulation study, the coverage rates of bootstrap confidence intervals were investigated using the population described above. For conditions, we varied the proportion of nonzero values (α) among 0.05, 0.1, 0.2, and 0.5. For sample size, we varied the number of cases randomly sampled from each stratum. The estimated coverage rates for nonparametric percentile CIs are given in Table 4; the coverage rates increase consistently well when the proportion of non-zero values (α) gets higher (e.g., when $\alpha = 0.5$, recovery was at approximately nominal levels).

Table 4. Coverage rates of 95% percentile bootstrap confidence intervals of θ

μ	n_j	Proportion of nonzero values (α)			
		$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.20$	$\alpha = 0.50$
20	50	0.852	0.874	.905	.942
	100	0.897	0.913	.940	.940
	200	0.910	0.935	.934	.948
50	50	0.891	0.936	.942	.939
	100	0.923	0.932	.939	.949
	200	0.939	0.946	.947	.959
100	50	0.916	0.936	.949	.939
	100	0.932	0.925	.949	.944
	200	0.939	0.940	.951	.950

4. Application

Due to the complexity of applying the pseudo-likelihood function detailed in the methodology, we apply the bootstrapping method to the Canadian Labour Force Survey – 2000 for Ontario [1]. We looked at employees' overtime hours, which can either be paid or unpaid hours worked outside the main job's work schedule. The data provides information of 17,415 participants; overtime hours are highly skewed. As shown in Table 5, the six strata vary greatly in means and sizes, indicating that inclusion probability is highly associated with the response.In the data, only 3,677 respondents have nonzero overtime hours with a mean of 92.15, which results in the nonzero proportion $\alpha = 0.211$ and the mean $\theta = \alpha\mu = 19.46$. As stated in [1], we also view the data set as the underlying population with $\theta = 19.46$ for the illustration purpose. A random sample of sizes 110, 110, 70, 70, 40, and 40 was taken from each of the six strata, respectively; the inclusion probabilities are listed in Table 5 . A similar approach is taken as in the simulation, pseudo-likelihood estimates are generated, and nonparametric and parametric bootstrap confidence intervals are constructed. The point estimate is 21.09, and the 95% percentile confidence intervals are (14.79, 28.19) and (16.12, 25.90) for the nonparametric and parametric methods, respectively. Figure 4 gives the bootstrap distributions for both techniques, parametric and nonparametric. *Note*. The vertical dashed lines in Figure 4 represent the 95% CIs of each bootstrap distribution of $\hat{\theta}$ and the middle (red) line represents the population mean of interest.

The paper illustrates a bootstrap procedure (both mathematically and computationally simple) to the maximum pseudo-likelihood approach under complex sampling designs developed in [1], it is worth comparing the confidence intervals under the bootstrap procedure, and the existing approach developed in [1]. 95% confidence interval of $\theta(= 19.46)$ under the existing approach [1] is estimated to be (9.47, 24.16). Bootstrap distribution of $\hat{\theta}$ estimates that 95% percentile confidence intervals are (14.79, 28.19) and (16.12, 25.90), respectively, for nonparametric and

parametric methods. The application shows that the new simpler approach gives a consistent and more symmetric confidence interval as compared to the existing method.

Table 5. Summary characteristics and inclusion probabilities for the six strata $P_1, ..., P_6$ in the Canadian Labour Force Survey–2000 for Ontario

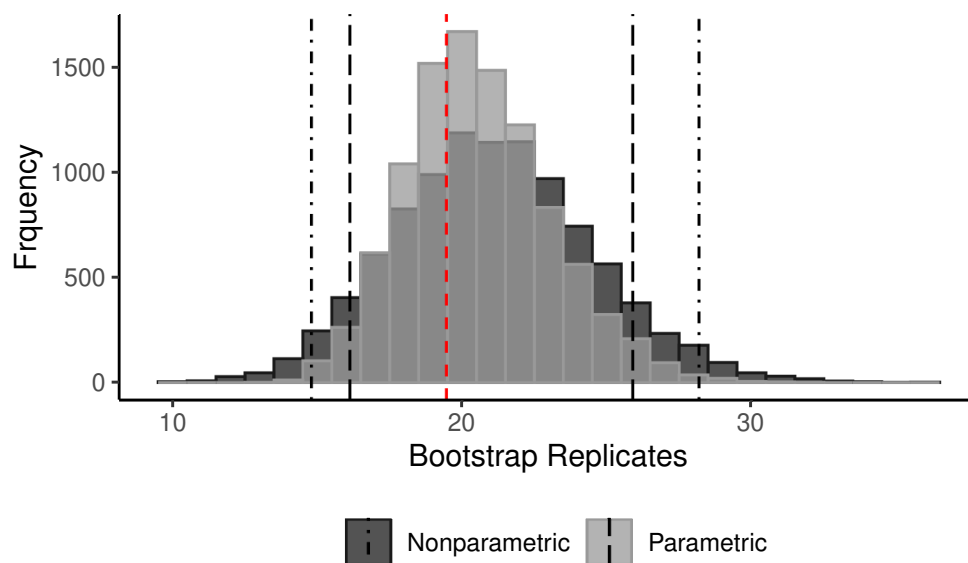|	P_1	P_2	P_3	P_4	P_5	P_6
Mean (Over time hours)	23.14	20.15	14.25	19.42	23.61	11.45
Size	9910	1220	206	554	656	4869
Sample Size	110	110	70	70	40	40
Inclusion Probability	0.011	0.090	0.340	0.126	0.061	0.008

Figure 4. Bootstrap distribution of $\hat{\theta}$ for Canadian Labour Force-2000, $\theta = 19.46$.

5. Conclusion

Paneru et al. [18] describe the bootstrap method as an alternative approach to compute confidence intervals of a zero-inflated population mean; the approach overcomes the complexity of maximum pseudo-likelihood ratio statistic and the asymptotic distribution. Paneru et al. [18] only provide an application to the normal model where the nonzero component is symmetrically distributed. However, it is common to work with data that is not normally distributed in a real-life situation. This paper extends the bootstrap approach where the nonzero part in the two-component model follows a log-normal distribution. Since the pseudo-likelihood function used in the methodology assumes a known probability distribution for the nonzero component, we recommend parametric bootstrap for the application. We also compute nonparametric bootstrap confidence intervals for the comparison. Under the assumption of a nonzero log-normal component, the parametric bootstrap method gives narrower confidence intervals than the nonparametric bootstrap method. Simulations results show that both bootstrap confidence intervals, parametric and nonparametric, consistently capture the true population mean. The bootstrap method

was applied to the real data of the Canadian Labour Force Survey – 2000 for Ontario. Although, as with the application to the Labour Force data, the pseudo-likelihood function is constrained by the sample sizes and inclusion probabilities. Caution should be kept in mind when selecting sample sizes, and careful consideration of the variation of inclusion probabilities allows for the constructed bootstrap distributions to be symmetrical.

Acknowledgement

REFERENCES

1. H. Chen, J. Chen, and S. Y. Chen, *Confidence intervals for the mean of a population containing many zero values under unequal-probability sampling*, Canadian Journal of Statistics, vol. 38, no. 4, pp. 582–597, 2010.
2. J. Chen, and R. R. Sitter, *A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys*, Statistica Sinica, vol. 9, pp. 385–406, 1999.
3. Y. Cui, and W. Yang, *Zero-inflated generalized Poisson regression mixture model for mapping quantitative trait loci underlying count trait with many zeros*, Journal of Theoretical Biology, vol. 256, no. 2, pp. 276–285, 2009.
4. Y. S. Dewi, and L. Amaliana, *Zero inflated Poisson and geographically weighted zero-inflated Poisson regression model: Application to elephantiasis (filariasis) counts data*, Journal of Mathematics and Statistics, vol. 11, no. 2, pp. 52, 2015.
5. B. Efron, *Bootstrap methods: Another look at the jackknife*, The Annuals of Statistics, vol. 7, no. 1, pp. 1–26, 1979.
6. B. Efron, *Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods*, Biometrika, vol. 63, no. 3, pp. 589–599, 1981.
7. B. Efron, *The Jackknife, the Bootstrap and Other Resampling Plans*, Society for Industrial and Applied Mathematics, 1982.
8. B. Efron, and R. J. Tibshirani, *An Introduction to the Bootstrap*, CRC press, 1994.
9. D. Fletcher, D. MacKenzie, and E. Villouta, *Modeling skewed data with many zeros: A simple approach combining ordinary and logistic regression*, Environmental and Ecological Statistics, vol. 12, no. 1, pp. 45–54, 2005.
10. A. H. Kvanli, Y. K. Shen, and L. Y. Deng, *Construction of confidence intervals for the mean of a population containing many zero values*, Journal of Business & Economic Statistics, vol. 16, no. 3, pp. 362–368, 1998.
11. D. Lambert, *Zero-inflated Poisson regression, with an application to defects in manufacturing*, Technometrics, vol. 34, no. 1, pp. 1–14, 1992.
12. L. Liu, Y. C. T. Shih, R. L. Strawderman, D. Zhang, B. A. Johnson, and H. Chai, *Statistical analysis of zero-inflated nonnegative continuous data: A review*, Statistical Science, vol. 34, no. 2, pp. 253–279, 2019.
13. T. Loeys , B. Moerkerke, O. De Smet, and A. Buysse, *The analysis of zero-inflated count data: Beyond zero-inflated Poisson regression*, British Journal of Mathematical and Statistical Psychology, vol. 65, no. 1, pp. 163–180, 2012.
14. T. Mahmood, and M. Xie, *Models and monitoring of zero-inflated processes: The past and current trends*, Quality and Reliability Engineering International, vol. 35, no. 8, pp. 2540–2557, 2019.
15. K. Paneru, *Regression analysis for zero inflated population under complex sampling designs* (Doctoral dissertation), Bowling Green State University, Bowling Green, OH, USA, 2013.
16. K. Paneru, and H. Chen, *Asymptotic distribution of pseudo-likelihood ratio statistic for zero-inflated generalized linear models under complex sampling designs*, Far East Journal of Theoretical Statistics, vol. 49, no. 1, pp. 41–60, 2014.
17. K. Paneru, and H. Chen, *Regression analysis under complex probability sampling designs in presence of many zero-value responses*, Advances and Applications in Statistics, vol. 40, no. 1, pp. 1–29, 2014.
18. K. Paneru, and H. Chen, *Estimation of zero-inflated population mean: A bootstrapping approach*, Journal of Modern Applied Statistical Methods, vol. 17, no. 1, pp. 1–14, 2018.
19. B. Pittman, E. Buta, S. Krishnan-Sarin, S. S. O'Malley, T. Liss, and R. Gueorguieva, *Models for analyzing zero-inflated and overdispersed count data: an application to cigarette and marijuana use*, Nicotine & Tobacco Research, vol. 22, no. 8, pp. 1390–1398, 2020.
20. F. Satter, and Y. Zhao, *Nonparametric interval estimation for the mean of a zero-inflated population*, Communications in Statistics-Simulation and Computation, vol. 49, no. 8, pp. 2059–2067, 2020.
21. F. Satter, and Y. Zhao, *Jackknife empirical likelihood for the mean difference of two zero-inflated skewed populations*, Journal of Statistical Planning and Inference, vol. 211, pp. 414–422, 2021.
22. D. J. Taylor, L. L. Kupper, S. M. Rappaport, and R. H. Lyles, *A mixture model for occupational exposure mean testing with a limit of detection*, Biometrics, vol. 57, no. 3, pp. 681–688, 2001.
23. L. Tian, *Inferences on the mean of zero-inflated log-normal data: The generalized variable approach*, Statistics in Medicine, vol. 24, no. 20, pp. 3223–3232, 2005.
24. A. H. Welsh, R. B. Cunningham, C. F. Donnelly, and D. B. Lindenmayer, *Modeling the abundance of rare species: statistical models for counts with extra zeros*, Ecological Modelling, vol. 88, no. 1-3, pp. 297–308, 1996.

25. Z. Xiao-Hua, and W. Tu, *Comparison of esveral independent population means when their samples contain log-normal and possibly zero observations*, Biometrics, vol. 55, no. 2, pp. 45–651, 1999.
26. Z. Xiao-Hua, and W. Tu, *Confidence intervals for the mean of diagnostic test charge data containing zeros*, Biometrics, vol. 56, no. 4, pp. 1118–1125, 2000.
27. M. Zidan, J. C. Wang, and M. Niewiadomska-Bugaj, *Comparison of k independent, zero-heavy log-normal distributions*, Canadian Journal of Statistics, vol. 39, no. 4, pp. 690–702, 2011.