5-2018

Estimation of Zero-Inflated Population Mean: A Bootstrapping Approach

Khyam Paneru

R. Noah Padgett

Hanfeng Chen
Bowling Green State University, hchen@bgsu.edu

Journal of Modern Applied Statistical Methods
May 2018, Vol. 17, No. 1, eP2557
doi: 10.22237/jmasm/1525133460

בס"ד

ISSN 1538 − 9472

Estimation of Zero-Inflated Population Mean: A Bootstrapping Approach

Khyam Paneru
University of Wisconsin-Whitewater
Whitewater, WI

R. Noah Padgett
Baylor University
Waco, TX

Hanfeng Chen
Bowling Green State University
Bowling Green, OH

A mixture model was adopted from the maximum pseudo-likelihood approach under complex sampling designs to estimate the mean of zero-inflated population. To overcome the complexity and assumptions of asymptotic distribution, the maximum pseudo-likelihood function was used, but a bootstrapping procedure was proposed as an alternative. Bootstrap confidence intervals consistently capture the true means of zero-inflated populations of the simulation studies.

Keywords: Zero-inflation, mixture model, maximum pseudo-likelihood, bootstrapping

Introduction

An underlying population that contains many zeros is very common. It has a large spike of zero-values with a proportion of non-zero values and is thus a zero-inflated population (ZIP). The name comes from a spike of zero values in a frequency distribution, and ZIP has a highly skewed distribution. In auditing, for example, most taxpayers may receive the correct refund (zero false refund is recorded), although others may request excessive refunds (non-zero false refunds are recorded). Similarly, zero-inflation is a common issue in modeling the abundance of rare species in ecological studies. Another example is modeling defect counts in a well-established manufacturing process. The process aims to make a lot of items with no defects (a zero-defect is recorded), but also has a distribution of defects that accompany the non-defects (Lambert, 1992). The importance of such applications is widespread, and developing meaningful and interpretable methods for estimation in ZIPs is important.

doi: 10.22237/jmasm/1525133460 | Accepted: November 18, 2017; Published: June 7, 2018.
Correspondence: Khyam Paneru, paneruk@uww.edu

In the construction of confidence intervals for the mean of a zero-inflated population, many techniques have been developed to overcome the issue of zero-inflation. Kvanli, Shen, and Deng (1998) proposed a two-component mixture model, a proportion of zeros mixed with a non-zero component that follows a known probability distribution. Zhou and Tu (2000) applied the method of likelihood ratio statistics and bootstrap techniques, though the likelihood ratio statistics rely on parametric assumptions. As a generalized approach of constructing confidence intervals for the zero-inflated population mean, Tian (2005) combined generalized *P*-value and a generalized confidence interval developed by Weerahandi (1993). That approach overcomes some issues of computational complexity, but failed under many data conditions. Other approaches constructing confidence intervals can be seen in J. Chen and Sitter (1999) and Taylor, Kupper, Rappaport, and Lyles (2001). Many of the existing methods for estimating the mean of ZIPs fail to account for data obtained through complex probability sampling designs in stratification and clustering. H. Chen, Chen, and Chen (2010) proposed a maximum pseudo-likelihood approach under complex probability sampling designs for the interval estimation.

Many of the cited approaches and techniques for the estimation of ZIP mean are expanded to construct regression models. Lambert (1992) introduced a method of zero-inflated Poisson regression to account for the over-dispersion of zeros in count data when the non-zeros follow a Poisson distribution. Cui and Yang (2009) expanded the modeling technique of zero-inflated Poisson regression with a zero-inflated generalized Poisson regression mixture model to account for the zero inflation and Poisson dispersion. Welsh, Cunningham, Donnelly, and Lindenmayer (1996) suggested the use of a mixture of Bernoulli and Poisson or negative binomial distributions to construct a zero-inflated regression model. Other extensions of zero-inflated regression techniques are presented in Ahmad et al. (2015), Purhadi, Dewi and Amaliana (2015), Fletcher, MacKenzie, and Villouta (2005), He, Tang, Wang, and Crits-Christoph (2014), and Loeys, Moerkerke, De Smet, and Buysse (2012).

Confidence intervals developed under the pseudo-likelihood approach have better coverage than existing methods when applied to data obtained through complex sampling designs (H. Chen et al., 2010). However, methods to produce confidence intervals under the pseudo-likelihood approach are mathematically and computationally complex. The pseudo-likelihood approach was extended into regression analysis in modeling ZIPs in Paneru and Chen (2014a) and applied to real data. Then, in Paneru and Chen (2014b), the methodology was further explained as a technical supplement.

The purpose of this study is to adopt the pseudo-likelihood function defined in H. Chen et al. (2010) and then propose a bootstrap procedure for constructing the confidence intervals of the mean that is mathematically, computationally, and intuitively simpler than existing methods. The aim is not to make comparisons to the existing complex methods; instead it is to present a simpler way of getting consistent results through bootstrapping techniques.

Methodology

Consider the concept of the two-component parametric mixture model from Kvanli et al. (1998) and the maximum pseudo-likelihood approach under complex sampling designs from H. Chen et al. (2010). In ZIPs, two components exist; one consisting solely of a proportion of zero values and a second component of non-zero values that adheres to some probability distribution. The two-component mixture model for zero-inflated population is defined by

$$\mathrm{h}\left(y;\alpha,\mu,\sigma\right)=\alpha\,\mathrm{f}\left(y;\mu,\sigma\right)\mathrm{I}_{(y\neq 0)}+\left(1-\alpha\right)\mathrm{I}_{(y=0)} \tag{1}$$

where α is the proportion of non-zeros, μ is the mean and σ the nuisance parameter of non-zero components, and I is the indicator function with value 1 if true and 0 if false. The parameter of interest is the mean of the mixture distribution h, i.e.

$$\theta=\alpha\mu$$

Consider a random subset s of n sampling units with values $y_1, y_2,\ldots, y_n$ obtained from a surveyed population with N units. These N units with values $y_1, y_2,\ldots, y_N$ are considered to be independently generated from the super population defined in the model (1). Let m be the number of zero values in n observed units. For the rest of the reading, assume that $m < n$ and arrange the samples as

$$\begin{aligned} y_i &\neq 0 \text{ for } i=1,2,\ldots,n-m \\ &= 0 \text{ for } i=n-m+1,n-m+2,\ldots,n \end{aligned}$$

If all N sampling units of the survey population are sampled, the log-likelihood function would be

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

As explained in H. Chen et al. (2010), consider the probability sampling design, where the random subset s of n sampling units is obtained from the surveyed population with the probability of inclusion π_i and sampling weight $w_i=\pi_i^{-1}$, i = 1, 2,…, n. Under this complex sampling design, the estimate of $l(\alpha, \mu, \sigma)$, called the pseudo-likelihood function, is defined by

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

where the sampling weights w_i, i = 1, 2,…, n are chosen such that $E(\hat{l})=l$.

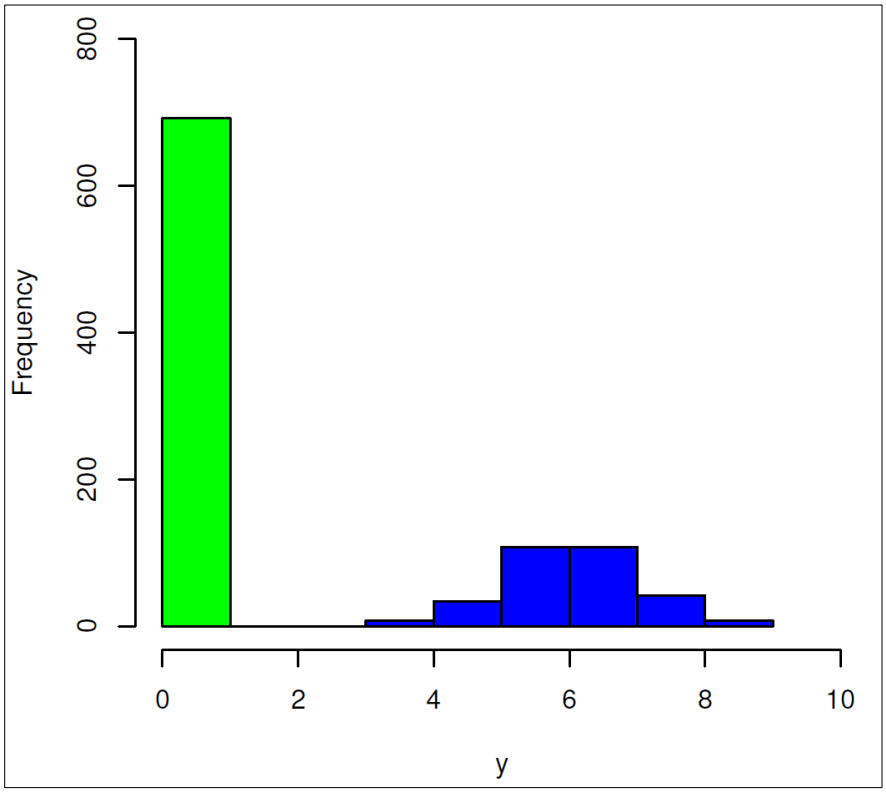

Figure 1. Histogram of a zero-inflated population as a mixture of zeros and normal population with μ = 6 and σ = 1

Application in Normal Model

For the mixture model (1), consider that the non-zero component ($y \neq 0$) follows normal distribution with mean μ and variance σ^2. So the probability density function $f(y; \mu, \sigma)I_{(y\neq 0)}$ is given by

$$f\left(y;\mu,\sigma\right)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2\sigma^2}(y_i-\mu)^2}$$

For instance, a histogram of 1000 randomly generated samples presented in Figure 1 gives a visual idea of a zero-inflated normal mixture model. The histogram contains 70% zero values and 30% non-zero values, where non-zero values follow a normal distribution with mean 6 and standard deviation 1. Note that the histogram is highly skewed to the right, owing to the large proportion of zeros, though the non-zero component is normal. However, in a real-life situation, the non-zero component may also be highly skewed, which brings extra skewness into the ZIP.

Maximum Pseudo-likelihood function and pseudo-likelihood estimates for the zero-inflated normal model are derived as follows:

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

where

$$w^+ = \sum_{i=1}^{n-m} w_i, \quad w^0 = \sum_{i=n-m+1}^{n} w_i, \quad C = \sum_{i=1}^{n-m} w_i \log\left(2\pi\right)^{-\frac{1}{2}}$$

As derived in H. Chen et al. (2010), taking first order derivatives with respect to α, μ, and σ^2, and setting

$$\frac{\partial \hat{l}}{\partial \alpha} = 0, \quad \frac{\partial \hat{l}}{\partial \mu} = 0, \quad \frac{\partial \hat{l}}{\partial \sigma^2} = 0$$

the maximum pseudo-likelihood estimates of the zero-inflated normal model are

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

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

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

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

$$\hat{\theta} = \hat{\alpha}\hat{\mu} \tag{5}$$

Bootstrap Confidence Intervals of θ: Mean of Zero-Inflated Population

The bootstrap technique for making statistical inferences was introduced by Efron (1979). “The basic ideas of statistics haven’t changed, but their implementation has. The modern computer lets us apply these ideas flexibly, quickly and easily with a minimum mathematical assumptions” (p. 2). The idea of the bootstrap technique is to simplify the complexity of calculation of traditional statistical theories by using the computing power of computer-based methods (Efron & Tibshirani, 1993). Further developments on bootstrap techniques can be found in Efron (1981a, 1981b, 1982).

Bootstrap methods can be either parametric or nonparametric. Parametric bootstrap methods involve sampling from a known probability distribution; in nonparametric bootstrap, the distribution is not specified. This paper uses the idea

of parametric methods where the non-zero component in mixture model (1) has a specified probability distribution. As an application in normal models, we assume that the non-zero component follows a normal distribution with mean μ and standard deviation σ, and use maximum pseudo-likelihood estimates accordingly as given in equations (2), (3), (4), and (5).

Under bootstrap methods, there are different approaches to obtain an approximate confidence interval of the parameter of interest. Commonly used bootstrap confidence intervals are standard normal bootstrap confidence interval, basic bootstrap confidence interval, percentile bootstrap confidence interval, bootstrap *t* confidence interval, and better bootstrap confidence interval. More detail and statistical computing of these different types of bootstrap confidence intervals can be found in Rizzo (2007).

Simulation Results

R is used for statistical computing of bootstrap estimates. 10,000 bootstrap replicates of $\hat{\theta}$ are computed under both parametric and nonparametric methods. Bootstrap replicates were computed at different values of α, proportion of non-zero component. For example, as presented in tables and figures below, $\alpha = 0.10$ means that the sample contains 10% observations with non-zero values and 90% observations with zero values.

For the simulations studies of the zero-inflated normal model, finite populations of size 10,000 ($N = 10{,}000$) are randomly generated and divided into four ($k = 4$) strata, each of size 2500. For a random sample of size $n = 135$, the inclusion probability for each stratum is set to $\pi_1 = 25/2500$, $\pi_2 = 30/2500$, $\pi_3 = 35/2500$, and $\pi_4 = 45/2500$. The corresponding weights for each stratum are set to $w_1 = 2500/25$, $w_2 = 2500/30$, $w_3 = 2500/35$, and $w_4 = 2500/45$, respectively. A random sample of size $n_j = np_j$, $j = 1,\ldots, 4$, is drawn from each stratum using simple random sampling without replacement. The strata sample sizes are: $n_1 = 25$, $n_2 = 30$, $n_3 = 35$, and $n_4 = 45$, respectively.

The simulation results are valid for a wide class of normal distributions. Results are presented from three different normal populations where the means and standard deviations differ significantly for the non-zero component. Results in Table 1 and Figure 2 assumed the non-zero component follows N($\mu = 100$, $\sigma = 10$), results in Table 2 and Figure 3 assume that the non-zero component follows N($\mu = 50$, $\sigma = 2$), and results in Table 3 and Figure 4 assumed the non-zero component follows N($\mu = 20$, $\sigma = 1$).

Table 1. Bootstrap confidence intervals of θ for given μ = 100 and σ = 10 at α = 0.05, 0.10, 0.20, and 0.50

Non-zero proportion α	True value θ	Nonparametric			Parametric		
		$LL_{0.025}$	$UL_{0.975}$	$\hat{\theta}$	$LL_{0.025}$	$UL_{0.975}$	$\hat{\theta}$
0.05	5	3.487	9.956	6.263	2.900	8.691	5.561
0.10	10	5.740	15.347	10.095	5.523	12.719	9.014
0.20	20	13.446	27.267	20.149	16.693	25.068	20.820
0.50	50	40.702	58.164	49.465	49.202	57.878	53.503

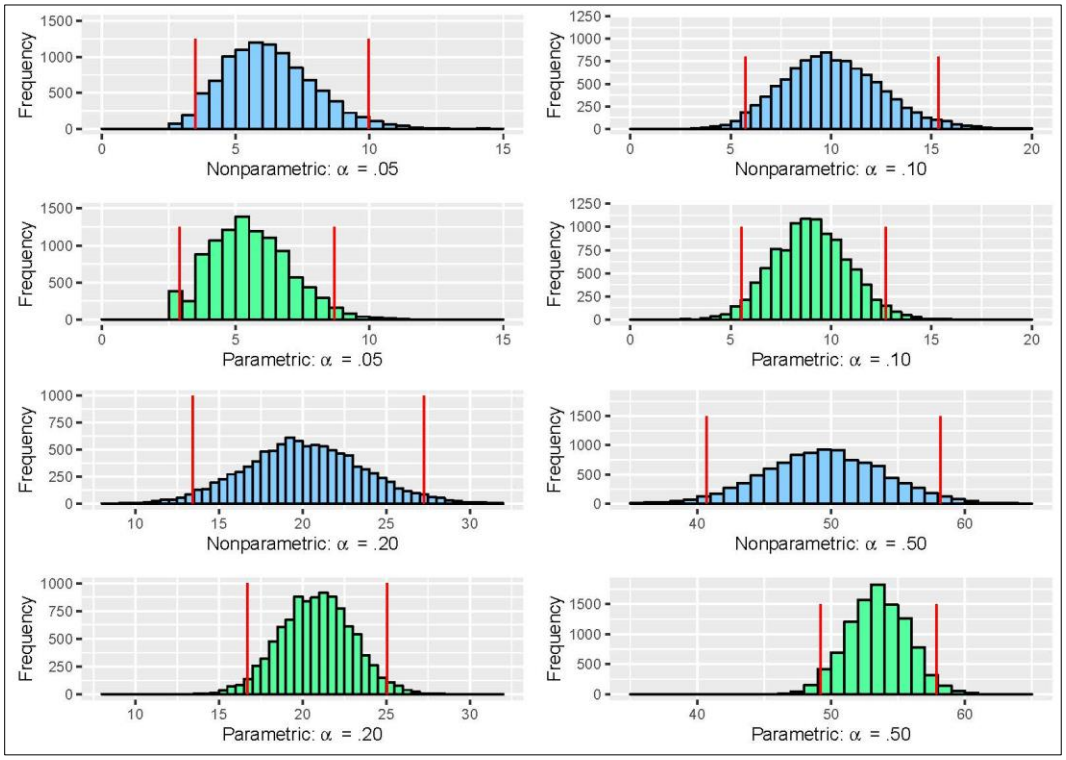

Figure 2. Bootstrap distribution of $\hat{\theta}$ for given μ = 100 and σ = 10 at α = 0.05, 0.10, 0.20, and 0.50

Table 2. Bootstrap confidence intervals of θ for given $\mu = 50$ and $\sigma = 2$ at $\alpha = 0.05$, 0.10, 0.20, and 0.50

Non-zero proportion α	True value θ	Nonparametric			Parametric		
		$LL_{0.025}$	$UL_{0.975}$	$\hat{\theta}$	$LL_{0.025}$	$UL_{0.975}$	$\hat{\theta}$
0.05	2.5	1.810	5.059	3.161	1.545	4.639	2.959
0.10	5	2.930	7.820	5.127	2.976	6.853	4.850
0.20	10	6.657	13.425	9.907	7.536	11.366	9.455
0.50	25	20.500	29.089	24.795	23.121	27.197	25.134

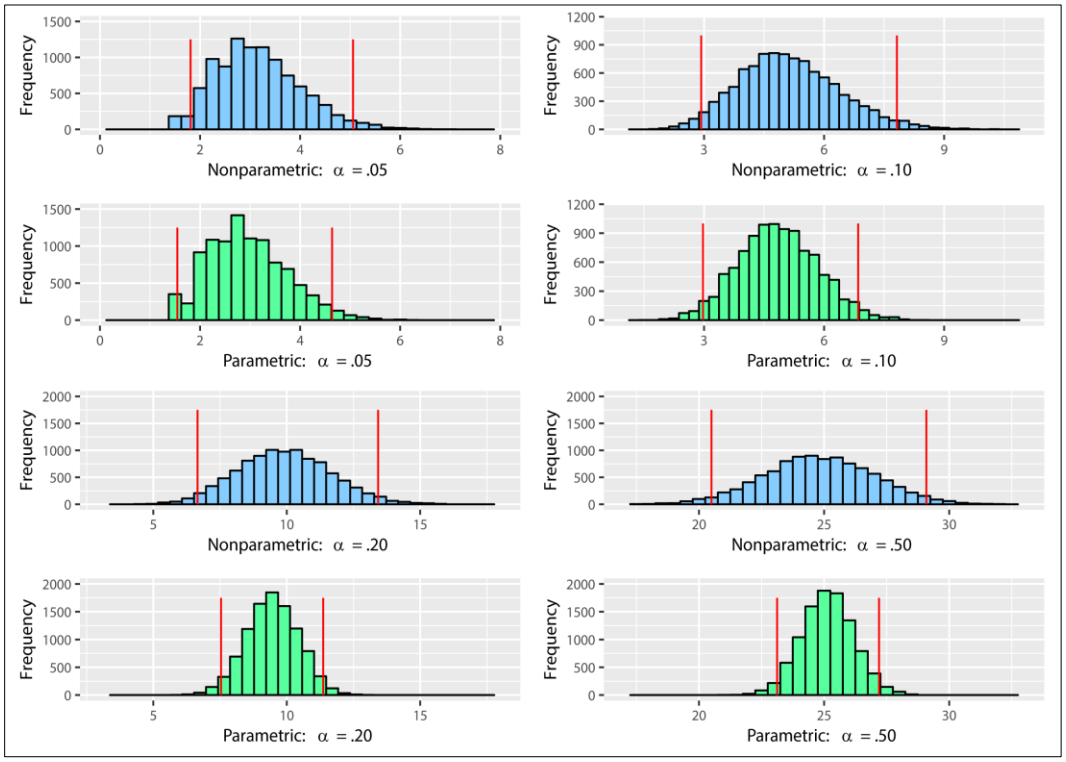

Figure 3. Bootstrap distribution of $\hat{\theta}$ for given $\mu = 50$ and $\sigma = 2$ at $\alpha = 0.05$, 0.10, 0.20, and 0.50

Table 3. Bootstrap confidence intervals of θ for given $\mu = 20$ and $\sigma = 1$ at $\alpha = 0.05$, 0.10, 0.20, and 0.50

Non-zero proportion α	True value θ	Nonparametric			Parametric		
		$LL_{0.025}$	$UL_{0.975}$	$\hat{\theta}$	$LL_{0.025}$	$UL_{0.975}$	$\hat{\theta}$
0.05	1	0.733	2.035	1.279	0.626	1.877	1.203
0.10	2	1.172	3.112	2.059	1.140	2.631	1.859
0.20	4	2.697	5.468	4.024	3.289	4.946	4.106
0.50	10	8.194	11.615	9.897	9.525	11.232	10.380

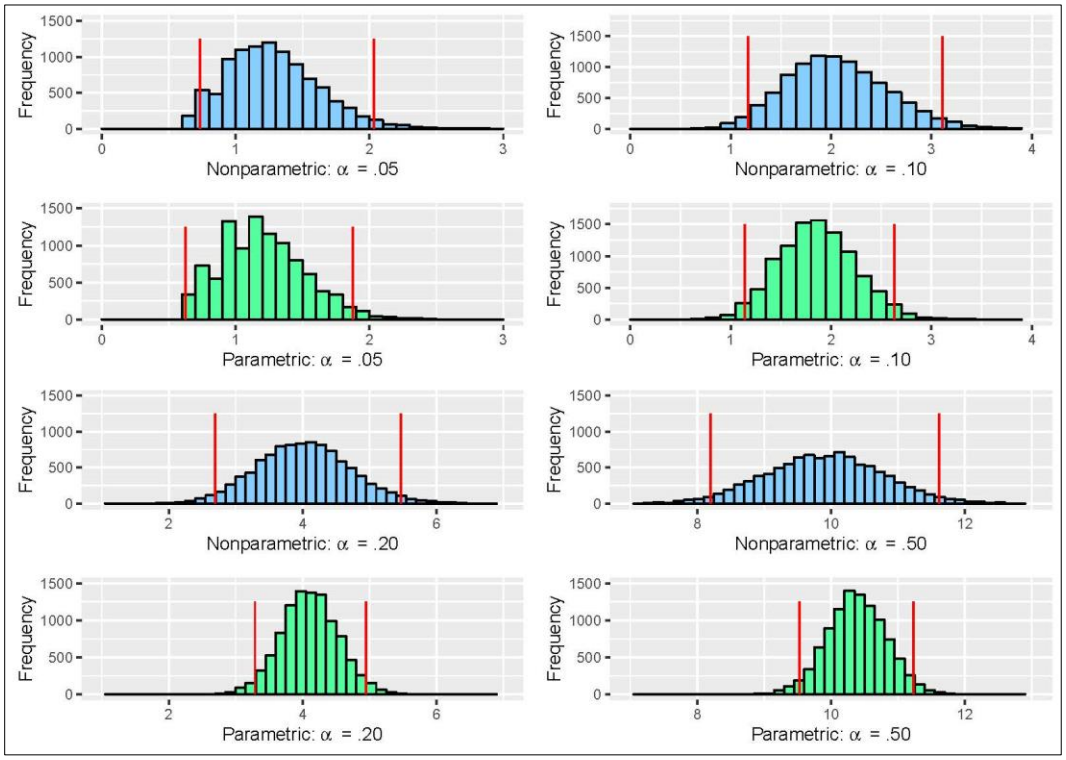

Figure 4. Bootstrap distribution of $\hat{\theta}$ for given $\mu = 20$ and $\sigma = 1$ at $\alpha = 0.05$, 0.10, 0.20, and 0.50

Conclusion

The asymptotic distribution of the pseudo-likelihood ratio statistic developed in H. Chen et al. (2010) has assumptions, and computing confidence intervals of the mean of a zero-inflated population are mathematically and computationally complex. The complexity arises as the method uses the complex probability

sampling designs in the two-component model. As an alternative to the maximum pseudo-likelihood ratio statistic and asymptotic distribution, we propose bootstrap methods to compute confidence intervals of a zero-inflated population mean. Using bootstrapping bypasses many of the assumptions of the asymptotic distribution and construction of confidence intervals is computationally simpler.

Use of the pseudo-likelihood function assumes a known probability distribution for the non-zero component, so the parametric bootstrap method is an appropriate approach. However, the non-parametric bootstrap method is employed to make a comparison. Confidence intervals based on the quantiles of the bootstrap distribution of $\hat{\theta}$ under both the parametric and non-parametric bootstrap method have been shown to consistently capture the true value of θ, the mean of the zero-inflated population. Under the assumption of a normally distributed non-zero component, the parametric bootstrap method gives narrower confidence intervals than the nonparametric bootstrap method.

References

Ahmad, W. M. A. W., Abdullah, S. A., Mokhtar, K., Aleng, N. A., Halim, N., & Ali, Z. (2015). Applications of zero inflated models for health sciences data. *Journal of Advanced Scientific Research, 6*(2), 39-44. Retrieved from http://www.sciensage.info/journal/1435391622JASR_3112142.pdf

Chen, H., Chen, J., & Chen, S.-Y. (2010). Confidence intervals for the mean of a population containing many zero values under unequal-probability sampling. *Canadian Journal of Statistics, 38*(4), 582-597. doi: 10.1002/cjs.10077

Chen, J., & Sitter, R. R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica, 9*(2), 385-406. Retrieved from http://www3.stat.sinica.edu.tw/statistica/j9n2/j9n25/j9n25.htm

Cui, Y., & Yang, W. (2009). Zero-inflated generalized Poisson regression mixture model for mapping quantitative trait loci underlying count trait with many zeros. *Journal of Theoretical Biology, 256*(2), 276-285. doi: 10.1016/j.jtbi.2008.10.003

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics, 7*(1), 1-26. doi: 10.1214/aos/1176344552

Efron, B. (1981a). Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika, 68*(3) 589-599. doi: 10.1093/biomet/68.3.589

Efron, B. (1981b). Nonparametric standard errors and confidence intervals. *Canadian Journal of Statistics, 9*(2), 139-158. doi: 10.2307/3314608

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*. Philadelphia, PA: Society for Industrial and Applied Mathematics. doi: 10.1137/1.9781611970319

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Boca Raton, FL: CRC press.

Fletcher, D., MacKenzie, D., & Villouta, E. (2005). Modelling skewed data with many zeros: A simple approach combining ordinary and logistic regression. *Environmental and Ecological Statistics, 12*(1), 45-54. doi: 10.1007/s10651-005-6817-1

He, H., Tang, W., Wang, W., & Crits-Christoph, P. (2014). Structural zeroes and zero-inflated models. *Shanghai Archives of Psychiatry, 26*(4), 236-242.

Kvanli, A. H., Shen, Y. K., & Deng, L. Y. (1998). Construction of confidence intervals for the mean of a population containing many zero values. *Journal of Business & Economic Statistics, 16*(3), 362-368. doi: 10.1080/07350015.1998.10524776

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics, 34*(1), 1-14. doi: 10.2307/1269547

Loeys, T., Moerkerke, B., De Smet, O., & Buysse, A. (2012). The analysis of zero-inflated count data: Beyond zero-inflated Poisson regression. *British Journal of Mathematical and Statistical Psychology, 65*(1), 163-180. doi: 10.1111/j.2044-8317.2011.02031.x

Paneru, K., & Chen, H. (2014a). Asymptotic distribution of pseudo-likelihood ratio statistic for zero-inflated generalized linear models under complex sampling designs. *Far East Journal of Theoretical Statistics, 49*(1), 41-60. Available from http://www.pphmj.com/abstract/8744.htm

Paneru, K., & Chen, H. (2014b). Regression analysis under complex probability sampling designs in presence of many zero-value responses. *Advances and Applications in Statistics, 40*(1), 1-29. Available from http://www.pphmj.com/abstract/8459.htm

Purhadi, Dewi, Y. S., & Amaliana, L. (2015). Zero inflated Poisson and geographically weighted zero-inflated Poisson regression model: application to

elephantiasis (*filariasis*) counts data. *Journal of Mathematics and Statistics, 11*(2), 52-60. doi: 10.3844/jmssp.2015.52.60

Rizzo, M. L. (2007). *Statistical computing with R*. New York, NY: Chapman and Hall/CRC Press. doi: 10.1201/9781420010718

Taylor, D. J., Kupper, L. L., Rappaport, S. M., & Lyles, R. H. (2001). A mixture model for occupational exposure mean testing with a limit of detection. *Biometrics, 57*(3), 681-688. doi: 10.1111/j.0006-341x.2001.00681.x

Tian, L. (2005). Inferences on the mean of zero-inflated lognormal data: The generalized variable approach. *Statistics in Medicine, 24*(20), 3223-3232. doi: 10.1002/sim.2169

Weerahandi, S. (1993). Generalized confidence intervals. *Journal of the American Statistical Association, 88*(423), 899-905. doi: 10.1080/01621459.1993.10476355

Welsh, A. H., Cunningham, R. B., Donnelly, C. F., & Lindenmayer, D. B. (1996). Modelling the abundance of rare species: Statistical models for counts with extra zeros. *Ecological Modelling, 88*(1-3), 297-308. doi: 10.1016/0304-3800(95)00113-1

Zhou, X.-H., & Tu, W. (2000). Confidence intervals for the mean of diagnostic test charge data containing zeros. *Biometrics, 56*(4), 1118-1125. doi: 10.1111/j.0006-341x.2000.01118.x