


RESEARCH

Open Access



Generalized fiducial inference on the mean of zero-inflated Poisson and Poisson hurdle models

Yixuan Zou¹, Jan Hannig^{2*}  and Derek S. Young³

*Correspondence:

jan.hannig@unc.edu

²Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA
Full list of author information is available at the end of the article

Abstract

Zero-inflated and hurdle models are widely applied to count data possessing excess zeros, where they can simultaneously model the process from how the zeros were generated and potentially help mitigate the effects of overdispersion relative to the assumed count distribution. Which model to use depends on how the zeros are generated: zero-inflated models add an additional probability mass on zero, while hurdle models are two-part models comprised of a degenerate distribution for the zeros and a zero-truncated distribution. Developing confidence intervals for such models is challenging since no closed-form function is available to calculate the mean. In this study, generalized fiducial inference is used to construct confidence intervals for the means of zero-inflated Poisson and Poisson hurdle models. The proposed methods are assessed by an intensive simulation study. An illustrative example demonstrates the inference methods.

Keywords: Count data, Coverage probability, Data dispersion, Generalized confidence intervals, Zero-truncated poisson

1 Introduction

The Poisson distribution is arguably one of the most commonly used models for count data. As such, a large number of inferential tools are available for Poisson-based models, such as for the ratio of two Poisson rates (Gu et al. 2008), Poisson regression models (Cameron and Trivedi 1990), and Poisson point processes (Itô 2015). Assuming the Poisson as an underlying distribution for parametric modeling can be a fairly strong assumption since one must be willing to posit that their data are equi-dispersed. In practice, count data almost ubiquitously demonstrate over-dispersion, which can be attributed to, for example, (spatio-)temporal dependency, unexplained heterogeneity, and/or excess zeros (Cameron and Trivedi 2013).

One of the earliest papers to address the problem of excess zeros was (Mullahy 1986), who proposed a two-part model that permits a more flexible data-generating process: zeros are from a binomial distribution while positive values are from a truncated distribution. Such a model can accommodate under- and over-dispersion. The model using a

zero-truncated Poisson is often called the Poisson hurdle (PH) model. Later, the seminal paper of (Lambert 1992) extended this phenomenon of excess zeros to the count regression setting, but also framed the problem differently with respect to *how* the zeros were generated. Specifically, a certain number of zeros are expected to be generated according to the assumed count distribution (*random zeros*) while the excess zeros are assumed to be generated from a separate, degenerate process (*structural zeros*). This framework results in a zero-inflated model, which is a two-component mixture model with one component for the assumed count distribution and the second component a degenerate distribution at zero. In the work of (Lambert 1992), the development was in the context of zero-inflated Poisson (ZIP) regression models. Regardless, both PH and ZIP models accommodate the notion of excess zeros in a Poisson setting, but how the zeros are generated is treated differently under the two models. Moreover, both models tend to have comparable performance regarding goodness-of-fit measures, which underscores how the application should provide the guidance in determining the way the zeros are generated.

The more complex data setting posed by zero-inflation opens the door to additional inference considerations, many coupled with their own challenges. For example, there is a bevy of score tests developed for testing the presence of zero-inflation in various count data settings; cf. van den Broek (1995); Janaskul and Hinde (2002); Janaskul and Hinde (2008); Cao et al. (2014); Todem et al. (2018). Bhattacharya et al. (2008) used a general Bayesian setup for detecting if zero-inflation is present in the data, however, it is challenging to justify the selection of the prior distribution. Score-based tests are also available for testing the presence of overdispersion, which can be caused by zero-inflation; cf. (Ridout et al. 2001; Hall and Berenhaut 2002; Deng and Paul 2005). With the exception of large-sample-based approaches for constructing confidence intervals on regression parameters in zero-inflated regression and hurdle regression models, there is no panacea for constructing reliable, accurate confidence intervals for other parameters in their non-regression counterparts, such as the population mean of univariate ZIP and PH distributions.

Deriving confidence intervals for a more complex data setting, like the presence of excess zeros, is challenging in the frequentist setting. Typically, one resorts to normal-based theory, but finite sample properties can be highly unreliable. In particular, Waguespack et al. (2020) assessed Wald-based confidence intervals for the ZIP mean. Their simulation results showed more liberal results for smaller n . As an alternative, they proposed constructing a bootstrap-based confidence interval for the ZIP mean, which had coverage probabilities much closer to the nominal level. They also conducted a signed likelihood ratio test (SLRT) for testing the ZIP mean, which controlled the type I error rate satisfactorily. Bayesian approaches suffer from the challenge to justify the selection of the prior distribution, just like we noted with the work of Bhattacharya et al. (2008) earlier. Alternatively, one can consider fiducial inference as proposed by (Fisher 1935). Fiducial inference struggled to gain popularity among statisticians because of perceived deficiencies in the general approach. However, later works have developed more sophisticated procedures coupled with rigorous theory to mitigate such criticisms, all while reflecting the core tenets of the fiducial paradigm. For example, (Weerahandi 1995) introduced generalized confidence intervals (GCIs) constructed by generalized pivotal quantities (GPQs), and (Hannig et al. 2006) further established the connection between GCI and

the fiducial argument of Fisher. For the purposes of our study, we turn to generalized fiducial distributions as they often lead to attractive solution with asymptotically correct frequentist coverage levels. Moreover, many simulation studies have shown that generalized fiducial solutions have very good small sample properties; see, for example, (Hannig 2009) and (E et al. 2008). There is also some work on fiducial approaches for discrete distributions. For example, (Mathew and Young 2013) developed fiducial tolerance intervals for functions of discrete random variables, while (Hannig et al. 2016) presented an extensive summary about computing the generalized fiducial distribution for parameters of some common discrete distributions. In this paper, we shall consider using the fiducial inference for the mean of ZIP and PH distributions.

This paper is organized as follows. In Section 2, we give a brief sketch of generalized fiducial inference, with emphasis on the discrete data setting. In Section 3, we derive the respective fiducial distributions of the ZIP mean and PH mean. In Section 4, we present a numerical study to illustrate the good coverage probabilities of GCIs for the ZIP mean and PH mean constructed using fiducial inference, and demonstrate that the fiducial test has comparable performance to the SLRT when conducting the ZIP mean test. An analysis of urinary tract infection data is presented in Section 5. In Section 6, we make some concluding remarks.

2 Generalized fiducial inference

The aim of generalized fiducial inference is to define a distribution for parameters of interest that contains all of the information from data. Therefore, inference on the parameters can be made through this distribution. The tenet of generalized fiducial inference is to switch the role of the parameters and the data. We now briefly explain the philosophy of generalized fiducial inference.

Suppose that data Y are generated through the structural equation $Y = G(\xi, U)$, where ξ is a vector of parameters and U is some random variable with a known distribution independent of the parameter ξ . The structural equation can be regarded as a data generation process where the noise process U and the signal ξ will produce observed data Y . Hence, the distribution of Y can be determined via the structural equation given a fixed parameter ξ and the distribution U . After the data Y are observed, we can switch the position of the data and parameters by solving the structural equation conditioned on that the solution to that equation exists. Thus, we can get $\xi = Q(Y, U)$. For more details regarding this setup, we refer to (Hannig 2009).

2.1 Generalized fiducial inference on discrete data

Let Y now be a discrete random variable with the distribution function $F(\cdot|\theta)$. We know that if $U \sim \mathcal{U}(0, 1)$, data following the distribution $F(\cdot|\theta)$ can be generated through $Y = F^{-1}(U|\theta)$, where $F^{-1}(a|\theta) = \inf\{y : a \leq F(y|\theta)\}$ is the inverse function. According to the philosophy of generalized fiducial inference, we need to solve the data generating equation to get the parameter as a function of the data and a known random distribution. Assume for each fixed y , the distribution is a nonincreasing function of θ . It follows that $Q_y^+(u) = \sup\{\theta : F(y|\theta) = u\}$ and $Q_y^-(u) = \inf\{\theta : F(y_-|\theta) = u\}$ exist and satisfy $F(y|Q_y^+(u)) = F(y_-|Q_y^-(u)) = u$. Moreover, the closure of the inverse image is $\bar{Q}_y(u) = [Q_y^-(u), Q_y^+(u)]$, where $F(y_-|\theta)$ is the left limit of the distribution function. Hannig et al. (2016) chose a 50-50 mixture of the upper and lower bound as the generalized

fiducial distribution for the parameter, so that the fiducial sample of the parameter has 50% chance from either the upper or lower bound.

For the fiducial distributions of multiple parameters, a two-stage method can be applied based on the minimal sufficient statistics. Let the parameters of interest be $\xi = (\xi_1, \xi_2)$. Assume that the following two conditions hold:

- 1 If ξ_2 is known, there is a statistic $\mathcal{S}_1 = \mathcal{S}_1(\xi_2)$ that has an invertible pivotal relationship with ξ_1 .
- 2 A statistic \mathcal{S}_2 exists that \mathcal{S}_2 and ξ_2 have an invertible pivotal relationship.

Then we can obtain the fiducial distribution of ξ_2 followed by the fiducial distribution of ξ_1 given that ξ_2 is known.

3 Fiducial distributions for poisson data with excess zeros

3.1 Fiducial distribution of ZIP mean

The ZIP distribution has probability mass function

$$p(x|\pi, \lambda) = \pi I_{\{0\}}(x) + (1 - \pi) \frac{\lambda^x e^{-\lambda}}{x!} I_{\{\mathbb{N}\}}(x),$$

where $I_{\{A\}}(z)$ is the indicator function that z belongs to the set A . The following proposition establishes the minimal sufficient statistic for a ZIP distribution:

Proposition 3.1 Let $X = (X_1, X_2, \dots, X_n)^T$ be a random sample from a ZIP distribution. Denote the sum of the random sample as S and the number of zeros of the random sample as K , where $S = \sum_{i=1}^n X_i$ and $K = \sum_{i=1}^n I_{\{0\}}(X_i)$. Consequently the minimal sufficient statistic is (S, K) .

Proof First we need to prove (S, K) is sufficient. Let $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$ be the realizations of X . The joint density of (X_1, \dots, X_n) is

$$\begin{aligned} p(\mathbf{x}|\pi, \lambda) &= \prod_{i=1}^n \left\{ \pi I_{\{0\}}(x_i) + (1 - \pi) \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} I_{\{\mathbb{N}\}}(x_i) \right\} \\ &= \prod_{i=1}^n \left\{ \left[\pi + (1 - \pi)e^{-\lambda} \right] I_{\{0\}}(x_i) + (1 - \pi) \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} I_{\{\mathbb{N}^+\}}(x_i) \right\} \\ &= \prod_{i=1}^n \left\{ \left[\pi + (1 - \pi)e^{-\lambda} \right]^{I_{\{0\}}(x_i)} \left[(1 - \pi) \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \right]^{(1 - I_{\{0\}}(x_i))} \right\} \\ &= \left(\frac{\pi + (1 - \pi)e^{-\lambda}}{(1 - \pi)e^{-\lambda}} \right)^{\sum_{i=1}^n I_{\{0\}}(x_i)} [(1 - \pi)e^{-\lambda}]^n \lambda^{\sum_{i=1}^n x_i (1 - I_{\{0\}}(x_i))} \\ &\quad \times \prod_{i=1}^n \left(\frac{1}{x_i!} \right)^{(1 - I_{\{0\}}(x_i))} \\ &= \left(\frac{\pi + (1 - \pi)e^{-\lambda}}{(1 - \pi)e^{-\lambda}} \right)^{\sum_{i=1}^n I_{\{0\}}(x_i)} [(1 - \pi)e^{-\lambda}]^n \lambda^{\sum_{i=1}^n x_i} \\ &\quad \times \prod_{i=1}^n \left(\frac{1}{x_i!} \right)^{(1 - I_{\{0\}}(x_i))}, \end{aligned}$$

where $\mathbb{N}^+ = \mathbb{N} \setminus \{0\}$. According to the factorization theorem, (S, K) is sufficient.

Now we want to show (S, K) is minimal sufficient. Assume that we have another sample $Y = (Y_1, \dots, Y_n)^T$ with corresponding realizations $y = (y_1, \dots, y_n)^T$. The ratio of the two density functions is

$$\frac{p(\mathbf{x}|\pi, \lambda)}{p(\mathbf{y}|\pi, \lambda)} = \left(\frac{\pi + (1 - \pi)e^{-\lambda}}{(1 - \pi)e^{\lambda}} \right)^{\sum_{i=1}^n I_{\{0\}}(x_i) - \sum_{i=1}^n I_{\{0\}}(y_i)} \lambda^{\sum_{i=1}^n x_i - \sum_{i=1}^n y_i} f(\mathbf{x}, \mathbf{y}),$$

where f is a function that does not depend on the parameters. The ratio is free of (π, λ) if and only if $(\sum_{i=1}^n x_i, \sum_{i=1}^n I_{\{0\}}(x_i)) = (\sum_{i=1}^n y_i, \sum_{i=1}^n I_{\{0\}}(y_i))$. Hence (S, K) is minimal sufficient. \square

It immediately follows that $K \sim \text{Binomial}(n, \pi + (1 - \pi)e^{-\lambda})$, and $(S | K = k)$ has the same distribution as $\sum_{i=1}^{n-k} Y_i$, where Y_i are independent $\text{Poisson}(\lambda)$ random variables conditioned on the event $\{Y_i \geq 1\}$. We also need the following proposition regarding sums of zero-truncated Poisson distributions:

Proposition 3.2 *Let Y_1, Y_2, \dots, Y_m be independent $\text{Poisson}(\lambda)$ random variables conditioned on the event $\{Y_i \geq 1\}$. Then*

$$\begin{aligned} P\left(\sum_{j=1}^m Y_j = k\right) &= \frac{\lambda^k}{k! (e^{\lambda} - 1)^m} \sum_{j=0}^m \binom{m}{j} (-1)^{m-j} j^k \\ &= \frac{\lambda^k m! S(k, m)}{k! (e^{\lambda} - 1)^m} I_{\{m, m+1, \dots\}}(k), \end{aligned}$$

where $S(k, m) = \frac{1}{m!} \sum_{j=0}^m \binom{m}{j} (-1)^{m-j} j^k$ is the Stirling number of the second kind.

Proof The proof follows by mathematical induction, and can be found in Springael and Van Nieuwenhuyse (2006). \square

Denote the distribution function of the sum of m zero-truncated $\text{Poisson}(\lambda)$ by $F_1(k|m, \lambda)$, the distribution of Poisson with mean parameter λ by $F_P(k|\lambda)$, and the distribution function of $\text{Binomial}(n, p)$ random variables by $F_B(k|n, p)$. It follows that

$$F_1(k | m, \lambda) = P\left(\sum_{j=1}^m Y_j \leq k\right) = \sum_{j=1}^m (-1)^{m-j} \binom{m}{j} \frac{e^{\lambda j}}{(e^{\lambda} - 1)^m} F_P(k|\lambda j).$$

We will use the inverse distribution functions as a data generating equation:

$$K = F_B^{-1}(U_1|n, \pi + (1 - \pi)e^{-\lambda}) \text{ and } S = F_1^{-1}(U_2|n - K, \lambda),$$

where U_1, U_2 are independent $\mathcal{U}(0, 1)$. When $K = n$, the value of S is set as 0.

After observing k and s , and inverting the data generating equation, we see that

$$B_{k, n-k+1}(U_1^*) \leq \pi + (1 - \pi)e^{-\lambda} \leq B_{k+1, n-k}(U_1^*) \text{ and } H_{n-k, s-1}(U_2^*) \leq \lambda \leq H_{n-k, s}(U_2^*),$$

where $B_{a,b}(u)$ is the quantile function of the $\text{Beta}(a, b)$ distribution evaluated at u and $H_{m,s}(u)$ is the solution (in λ) of the equation $F_1(s | m, \lambda) = u$. Thus, the sample from the fiducial distribution is obtained by sampling (U_1^*, U_2^*) , and using the above inequalities to solve for π and λ . Consequently, when the parameter of interest is $\mu = (1 - \pi)\lambda$, the mean of the ZIP distribution, we have

$$\frac{H_{n-k, s-1}(U_2^*)(1 - B_{k+1, n-k}(U_1^*))}{1 - e^{-H_{n-k, s-1}(U_2^*)}} \leq \mu \leq \frac{H_{n-k, s}(U_2^*)(1 - B_{k, n-k+1}(U_1^*))}{1 - e^{-H_{n-k, s}(U_2^*)}},$$

if $k < n$. When $k = n$ then $0 \leq \mu \leq \infty$.

Finally, we need to select a representative region for the fiducial sample. Following (Hannig et al. 2016), we choose a 50-50 mixture of the upper and lower bound. In the case of $k = n$, this results in a 50-50 mixture of 0 and ∞ .

The algorithm for constructing a fiducial confidence interval for a ZIP mean is implemented as follows:

Algorithm 1

- 1 Let $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$ be a sample of size n from a ZIP distribution with Poisson parameter λ and binomial parameter π . Calculate the number of zeros $k = \sum_{i=1}^n I_{\{0\}}(x_i)$ and the sum of the sample $s = \sum_{i=1}^n x_i$.
- 2 Generate a realization U_1^* and U_2^* independently from $\mathcal{U}(0, 1)$. Then, calculate the realizations

$$M_L^* = \begin{cases} \frac{H_{n-k,s-1}(U_2^*)(1-B_{k+1,n-k}(U_1^*))}{1-e^{-H_{n-k,s-1}(U_2^*)}}, & \text{if } k < n \\ 0, & \text{if } k = n \end{cases}$$

$$M_U^* = \begin{cases} \frac{H_{n-k,s}(U_2^*)(1-B_{k,n-k+1}(U_1^*))}{1-e^{-H_{n-k,s}(U_2^*)}}, & \text{if } k < n \\ \infty, & \text{if } k = n, \end{cases}$$

where $B_{a,b}(u)$ is the quantile function of the Beta(a, b) distribution evaluated at u and $H_{m,s}(u)$ is the solution (in λ) of the equation $F_1(s | m, \lambda) = u$, where $F_1(s | m, \lambda)$ is the distribution function of the sum of m zero-truncated Poisson(λ).

- 3 Repeat step 2 B times, yielding $2B$ fiducial samples of the ZIP mean μ , denoted by $M_{L1}^*, \dots, M_{Ln}^*, M_{U1}^*, \dots, M_{Un}^*$. The $1 - \alpha$ two-sided GCI of the ZIP mean will be the lower and upper $\alpha/2$ quantiles of the fiducial samples.
-

3.2 Fiducial distribution of PH mean

The derivation of the fiducial distribution for the PH model follows that for the ZIP model *mutatis mutandis*. The PH distribution has probability mass function:

$$p(x|\pi, \lambda) = \pi I_{\{0\}}(x) + (1 - \pi) \frac{\lambda^x e^{-\lambda}}{x! (1 - e^{-\lambda})} I_{\{\mathbb{N}^+\}}(x).$$

Note that after reparameterization, a ZIP distribution characterized by the binomial parameter μ_1 and the Poisson parameter λ_1 can be expressed as a PH distribution that is characterized by the binomial parameter $\mu_2 = \mu_1 + (1 - e^{-\lambda_1})$ and the truncated Poisson parameter $\lambda_2 = \lambda_1$. Hence, the likelihoods of the two models are equivalent. The selection of the model should be based on how zeros are generated. Note that this equivalency does not hold in the ZIP regression and PH regression settings. In those settings, the likelihoods are based on a conditional distribution (i.e., Y given some covariates) for determining the estimates of the regression parameters. While the final likelihoods will typically be similar, they will not be equal.

The following proposition establishes the minimal sufficient statistic for a PH distribution:

Proposition 3.3 Let $X = (X_1, X_2, \dots, X_n)^T$ be a random sample from a PH distribution. Denote the sum of the random sample as S and the number of zeros of the random sample as K , where $S = \sum_{i=1}^n X_i$ and $K = \sum_{i=1}^n I_{\{0\}}(X_i)$. Consequently the minimal sufficient statistic is (S, K) .

Proof The proof is identical to the proof of Proposition 3.1, and is therefore omitted. \square

Immediately, we see that $K \sim \text{Binomial}(n, \pi)$, and $(S \mid K = k)$ has the same distribution as $\sum_{i=1}^{n-k} Y_i$, where Y_i are independent $\text{Poisson}(\lambda)$ random variables conditioned on the event $\{Y_i \geq 1\}$. We will then use the inverse distribution functions as a data generating equation

$$K = F_B^{-1}(U_1 | n, \pi) \text{ and } S = F_1^{-1}(U_2 | n - K, \lambda),$$

where U_1, U_2 are independent $\mathcal{U}(0, 1)$. When $K = n$ the value of S is again set as 0.

After observing k and s , and inverting the data generating equation, we see that

$$B_{k,n-k+1}(U_1^*) \leq \pi \leq B_{k+1,n-k}(U_1^*) \text{ and } H_{n-k,s-1}(U_2^*) \leq \lambda \leq H_{n-k,s}(U_2^*),$$

where $B_{a,b}(u)$ and $H_{m,s}(u)$ are as defined for the ZIP setting. Thus, the sample from the fiducial distribution is obtained by sampling (U_1^*, U_2^*) and using the above inequalities to solve for π and λ . Consequently, when the parameter of interest is $\mu = \frac{(1-\pi)\lambda}{1-e^{-\lambda}}$, the mean of the PH distribution, we have

$$\frac{H_{n-k,s-1}(U_2^*)(1 - B_{k+1,n-k}(U_1^*))}{1 - e^{-H_{n-k,s-1}(U_2^*)}} \leq \mu \leq \frac{H_{n-k,s}(U_2^*)(1 - B_{k,n-k+1}(U_1^*))}{1 - e^{-H_{n-k,s}(U_2^*)}},$$

if $k < n$. When $k = n$ then we again have $0 \leq \mu \leq \infty$. Thus, it turns out that the fiducial distribution of the mean of the ZIP and the mean of the PH are the same.

Finally, just as in the ZIP setting, the selection of a representative region for the fiducial sample is to choose a 50-50 mixture of the upper and lower bound. In the case of $k = n$, this again results in a 50-50 mixture of 0 and ∞ . The algorithm for constructing a GCI for a PH mean is the same as the ZIP setting, so it is omitted here.

4 Simulation study

We next assess the performance of the GCI just presented through an extensive simulation study. We also compare our results to the bootstrap confidence intervals constructed using the approach in Waguespack et al. (2020). Note that we do not include a comparison with the likelihood-based (i.e., Wald-based) confidence intervals since Waguespack et al. (2020) already demonstrated the relative superior performance of the bootstrap confidence intervals. The sample sizes used to assess the finite sample performance of the GCI include $n \in \{15, 30, 100\}$. For the parameters, the mixture proportion π is selected from $\{0.2, 0.5, 0.8\}$ and the mean λ of the Poisson distribution is selected from $\{1, 5\}$. The simulation settings for the PH distribution are the same as for the ZIP distribution: sample sizes $n \in \{15, 30, 100\}$, mixture proportions $\pi \in \{0.2, 0.5, 0.8\}$, and mean of the Poisson distribution $\lambda \in \{1, 5\}$. Note that when $\pi = 0$ in the ZIP setting or $\pi = e^{-\lambda}$ in the PH setting, the data are actually simulated from the Poisson distribution $\text{Poisson}(\lambda)$. Moreover, we demonstrate the performance of our approach when there is no under-/over-dispersion in the data. Specifically, the same values of n and λ are considered, but no

mixing proportion is present (or equivalently $\pi = 0$). The number of Monte Carlo samples for our simulations is set to 10,000 and the number of fiducial samples used is 1000. We also drew 10,000 bootstrap samples to construct the bootstrap confidence intervals. For each simulation scenario, we estimated the probability $Q(X) = P(\mathcal{R}_M(X) < M|X)$, where M is the mean of the distribution. If the generalized fiducial inference were exact, then $Q(X)$ should follow a standard uniform distribution, which could be examined through $Q - Q$ plots. In the results that follow, coverage probabilities and the median widths of the GCIs are reported. We report the median widths due to using a 50-50 mixture of 0 and ∞ as the fiducial distribution of λ , which has an expected value of ∞ . Furthermore, we compare type I error rates and the power for both the SLRT proposed by Waguespack et al. (2020) and our fiducial test for testing the ZIP mean. The simulation is set up similarly as in Waguespack et al. (2020) to enable a side-by-side comparison: sample sizes $n \in \{30, 40, 50\}$, mixture proportions $\pi \in \{0.1, 0.3, 0.5\}$, and mean of the Poisson distribution $\lambda \in \{1, 1.3, 1.6, 2\}$.

The first set of simulation results is for the ZIP distributions. The results are given in Table 1. As we can see, for the different simulation scenarios, the coverage probabilities for the bootstrap confidence intervals are typically liberal, especially for the sample sizes less than 100. Meanwhile, the coverage probabilities for the GCIs are all noticeably closer to the nominal level except for the sample size $n = 15$ and $\lambda = 1$. Under such a setting, the simulated samples are almost all zeros, thus compromising the inference. Even though the GCIs are conservative in this setting, they are closer to the nominal level compared to the more liberal bootstrap confidence intervals. The median widths are, of course, narrower for the bootstrap confidence intervals, but that is a result of the procedure being noticeably liberal relative to the nominal level. The $Q - Q$ plots for the different sample sizes are given in Fig. 1. As the sample size n increases, the agreement between the actual p -value and the nominal p -value improves.

The second set of simulation results is for the PH distributions. The results are given in Table 2. We again obtain similar results as in the ZIP setting for different simulation scenarios. The coverage probabilities are close to nominal for the GCIs, whereas the bootstrap confidence intervals are again noticeably liberal. In fact, the setting with sample size $n = 15$ and $\lambda = 1$ appears to be doing slightly better here in the PH setting compared to the analogous results in the ZIP setting. The $Q - Q$ plots for different sample sizes are given in Fig. 2. The same asymptotic behavior identified in the ZIP setting is also observed from the present simulation results; specifically, as the sample size n increases, the agreement between the actual p -value and the nominal p -value improves.

The third set of simulation results we consider is for the Poisson distribution. The results are given in Table 3. As noted earlier, the Poisson is just a special case of the ZIP and PH distributions such that there are no excessive zeros. Again, for different simulation scenarios, the coverage probabilities are close to nominal. For $n = 15$ and $n = 30$, there is a clear improvement using the GCIs compared to the bootstrap confidence intervals, but for large n , the procedures are comparable. This illustrates that regardless if zero-inflation is present in the data, our method is still appropriate for constructing a confidence interval of the mean. The $Q - Q$ plots for different sample sizes are given in Fig. 3. Only moderate discrepancies are noticeable when the sample size is small ($n = 15$) or moderate ($n = 30$); however, the tail behavior appears to be very good. Since we want to construct a 95% confidence interval, it is not a concern as long as the tails are accurate,

Table 1 Estimated coverage probabilities and median widths for the GCIs and bootstrap confidence intervals for the means generated from different ZIP distributions used in our simulation study

<i>n</i>	λ	π	Bootstrap		Fiducial	
			Cov. Prob.	Med. Width	Cov. Prob.	Med. Width
15	1	0.2	0.905	0.867	0.961	1.037
		0.5	0.902	0.800	0.978	0.973
		0.8	0.853	0.400	0.980	0.933
	5	0.2	0.926	2.733	0.959	2.839
		0.5	0.918	2.800	0.952	2.915
		0.8	0.880	2.067	0.974	2.455
30	1	0.2	0.920	0.667	0.953	0.716
		0.5	0.912	0.567	0.960	0.649
		0.8	0.878	0.367	0.976	0.521
	5	0.2	0.934	1.967	0.948	2.002
		0.5	0.936	2.067	0.952	2.079
		0.8	0.918	1.533	0.957	1.636
100	1	0.2	0.943	0.380	0.950	0.384
		0.5	0.941	0.330	0.950	0.343
		0.8	0.936	0.230	0.958	0.248
	5	0.2	0.950	1.100	0.952	1.101
		0.5	0.944	1.150	0.947	1.147
		0.8	0.943	0.870	0.952	0.876

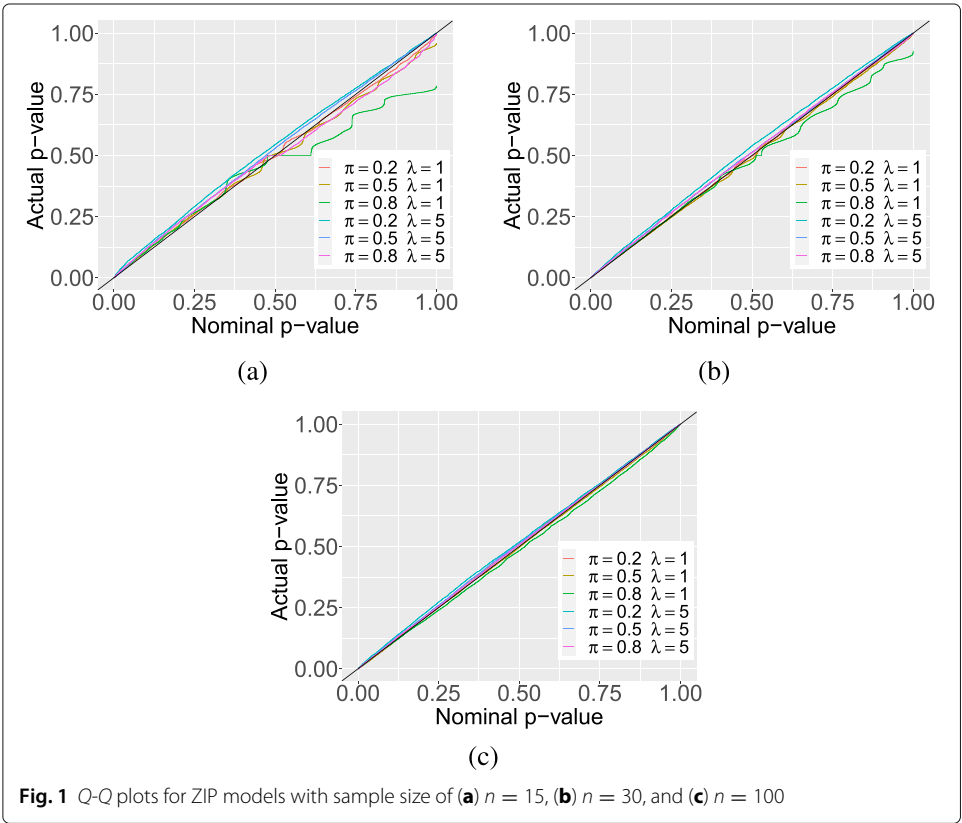


Table 2 Estimated coverage probabilities and median widths for the GCIs and bootstrap confidence intervals for the means generated from different PH distributions used in our simulation study

n	λ	π	Bootstrap		Fiducial	
			Cov. Prob.	Med. Width	Cov. Prob.	Med. Width
15	1	0.2	0.914	0.867	0.962	1.023
		0.5	0.918	0.867	0.965	1.035
		0.8	0.870	0.600	0.979	0.925
	5	0.2	0.922	2.733	0.956	2.828
		0.5	0.920	2.867	0.954	2.933
		0.8	0.853	2.067	0.968	2.443
30	1	0.2	0.931	0.667	0.953	0.708
		0.5	0.930	0.667	0.954	0.715
		0.8	0.891	0.467	0.974	0.573
	5	0.2	0.941	1.967	0.956	2.003
		0.5	0.937	2.067	0.950	2.080
		0.8	0.911	1.533	0.954	1.647
100	1	0.2	0.943	0.370	0.946	0.378
		0.5	0.944	0.380	0.951	0.384
		0.8	0.938	0.280	0.956	0.292
	5	0.2	0.949	1.100	0.950	1.099
		0.5	0.942	1.150	0.946	1.149
		0.8	0.936	0.870	0.946	0.881

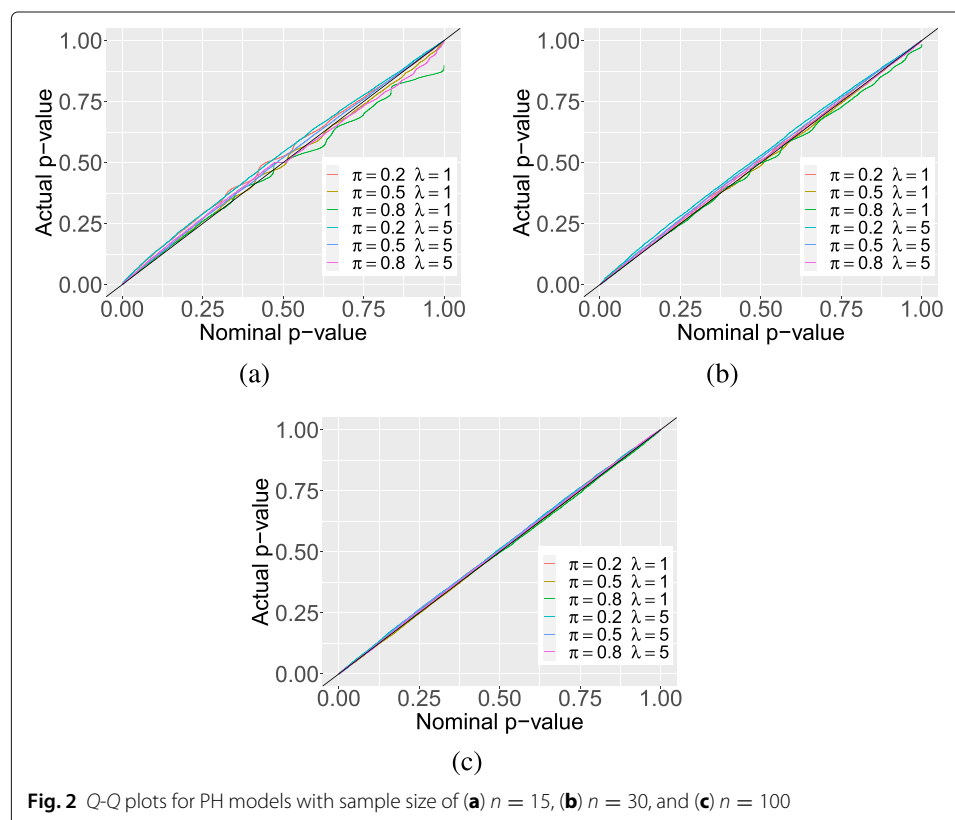
**Fig. 2** Q-Q plots for PH models with sample size of (a) $n = 15$, (b) $n = 30$, and (c) $n = 100$

Table 3 Estimated coverage probabilities and median widths for the GCI and bootstrap confidence intervals for the means generated from different Poisson distributions used in our simulation study

<i>n</i>	λ	Bootstrap		Fiducial	
		Cov. Prob.	Med. Width	Cov. Prob.	Med. Width
15	1	0.897	0.933	0.960	1.055
	5	0.913	2.133	0.962	2.442
30	1	0.926	0.700	0.952	0.728
	5	0.934	1.567	0.958	1.667
100	1	0.945	0.390	0.951	0.392
	5	0.948	0.870	0.956	0.885

which is confirmed by our results in Table 3. The asymptotic behavior is also observed from the simulation results: as the sample size n increases, the agreement between the actual p -value and the nominal p -value improves.

The last set of simulation results is to estimate the type I error rates and power of the SLRT and fiducial test for testing the following for the ZIP mean under $\alpha = 0.05$: $H_0 : \mu \leq \mu_0$ versus $H_a : \mu > \mu_0$. The μ_0 is assumed to be $1 - \pi$ and the true ZIP mean is $\mu = (1 - \pi)\lambda$. Therefore, the type I error rate could be estimated when $\lambda = 1$. The type I error rates and power of the SLRT and fiducial test are reported in Table 4. Generally, the performance of the two tests are almost identical. The type I error rates are all close to the nominal level for both tests while the power is increasing as λ or the sample size gets larger.

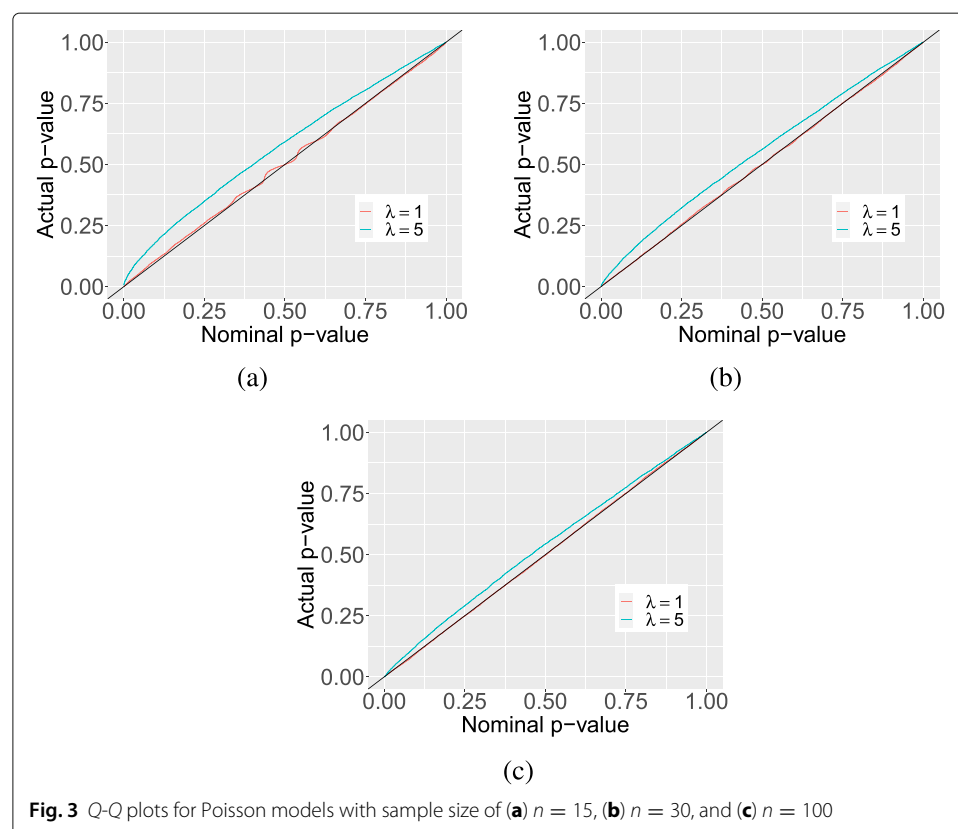


Table 4 Type I error rates and power of the SLRT and fiducial test for testing the ZIP mean under $\alpha = 0.05$: $H_0 : \mu \leq \mu_0$ versus $H_a : \mu > \mu_0$; $\mu = (1 - \pi)\lambda$; $\mu_0 = (1 - \pi)$

λ	n	π	SLRT	Fiducial test
1	30	0.1	0.048	0.046
		0.3	0.048	0.050
		0.5	0.048	0.046
	40	0.1	0.049	0.053
		0.3	0.048	0.048
		0.5	0.046	0.050
	50	0.1	0.045	0.050
		0.3	0.046	0.053
		0.5	0.045	0.045
1.3	30	0.1	0.364	0.386
		0.3	0.295	0.296
		0.5	0.217	0.223
	40	0.1	0.479	0.471
		0.3	0.355	0.356
		0.5	0.259	0.270
	50	0.1	0.517	0.552
		0.3	0.408	0.422
		0.5	0.291	0.298
1.6	30	0.1	0.799	0.810
		0.3	0.639	0.636
		0.5	0.461	0.470
	40	0.1	0.890	0.894
		0.3	0.745	0.741
		0.5	0.560	0.562
	50	0.1	0.943	0.947
		0.3	0.820	0.821
		0.5	0.637	0.640
2	30	0.1	0.984	0.982
		0.3	0.906	0.905
		0.5	0.732	0.737
	40	0.1	0.997	0.996
		0.3	0.960	0.959
		0.5	0.838	0.833
	50	0.1	0.999	0.999
		0.3	0.983	0.983
		0.5	0.899	0.904

5 Application: urinary tract infection data

We construct GCIs for a ZIP distribution fit to data on urinary tract infections (UTIs). Note that our fiducial approach will generate the same GCI no matter if the data follows a ZIP or PH distribution. Therefore, the UTI data could also serve as an example for constructing a GCI for a PH distribution. These data came from $n = 98$ HIV-infected men who were treated by the Department of Internal Medicine at the Utrecht University Hospital in the Netherlands. The frequency of times those patients had a UTI was recorded as X . The frequency table is given in Table 5. The data were analyzed in (van den Broek 1995), who used a score test to detect if zero-inflation exists. Later, (Bhattacharya et al. 2008) and (Bayarri et al. 2008) applied Bayesian testing to test for zero-inflation. All

Table 5 The frequency table of the number of UTIs recorded in the patients admitted at the Department of Internal Medicine at the Utrecht University Hospital

X	0	1	2	3	Total
Frequency	81	9	7	1	98

of these analyses favor a ZIP distribution. Moreover, the use of a zero-inflated distribution is appropriate because the zeros are likely arising from two subgroups of men: one group that are otherwise healthy aside from having HIV (structural zeros), and one group that has a history of other issues with their urinary system (e.g., kidney stones) and, thus, could be at higher risk of eventually developing a UTI (random zeros).

The fiducial sample is set to be 10,000. The 95% GCI for the average number of UTIs that the patients have is (0.157, 0.434). The 95% SLRT confidence interval is (0.157, 0.426), which is close to the GCI, while the bootstrap confidence interval with 10,000 bootstrap samples is (0.143, 0.398). Even though one can easily calculate the sample mean from these data ($\bar{x} = 0.266$) and infer that, for example, the average number is less than 1, the approach we have presented now affords us with the additional insights that accompany confidence interval interpretations, such as the reliability of our estimate of the mean and how far the spread of that interval falls away from a particular value of interest. We also know from the coverage study in the previous section that the median width of the 95% GCI will be noticeably wider than the respective 95% bootstrap confidence interval. Practically speaking, this could have implications on the hospital's treatment plans for these UTIs. If treatment plans are benchmarked against the 95% bootstrap confidence intervals, then smaller and larger values beyond the respective limits of that interval will be omitted from such plans, whereas those values will be reflected via the 95% GCI.

6 Conclusion

In this article, generalized fiducial inference on ZIP and PH distributions was studied for the first time and applied to a healthcare dataset. The practical contribution of this method is that one can now easily calculate and report a confidence interval along with an estimate of the mean if using either a ZIP or PH model. The theoretical advantage of this method is that it achieves good small sample properties except for when the zero proportion π is large and the Poisson parameter λ is small. Also, it does not depend on the selection of priors like Bayesian inference, but it only relies on the data generation equation. A simulation study demonstrated that, for the confidence interval of the mean of ZIP and PH distributions, the generalized fiducial inference works very well for various scenarios. Since the Poisson distribution can be considered as a special case of ZIP or PH distribution, the simulation also shows our method for ZIP and PH distributions can accommodate constructing the confidence interval of the mean of a Poisson distribution. Thus, if the goal is only to construct a confidence interval for the mean of the count data, our approach can be applied directly since it will not be necessary to detect for zero-inflation or decide if the data are under-/over-dispersed. Furthermore, our fiducial approach performed equally well as the SLRT when testing the ZIP mean.

We note that there is some computational limitation of the proposed method since it involves finding the root of a sum of factorials. When the sample size is large or the Poisson mean parameter is large, the computational effort could become prohibitive. Uniformly valid approximation exists for Stirling numbers of the second kind (Temme 1993),

which can alleviate some of the computational burden, but this can translate to worse results for coverage probabilities. Such simulation results are not shown here. We also note that generalized fiducial inference can be very similar to Bayesian inference when a fiducial distribution is obtained. We highlighted earlier that (Bhattacharya et al. 2008) used Bayesian inference to test for the presence of zero-inflation. Future research will be focused on extending the use generalized fiducial inference for selecting the model between Poisson distribution and ZIP/PH models. Moreover, there are broader inference considerations when fitting ZIP/PH regression models, such as joint confidence intervals on the regression parameters and simultaneous confidence intervals over the values of the covariate space. The utility of such inference is underscored by the recent emphasis placed on marginalized ZI regression models, which focuses on modeling the mean response across the two states (Long et al. 2014; Todem et al. 2016; Martin and Hall 2017). These are further extensions worth considering in the generalized fiducial framework.

Abbreviations

GCI: Generalized confidence interval; GPQ: Generalized pivotal quantity; PH: Poisson hurdle; SLRT: Signed likelihood ratio test; UTI: Urinary tract infection; ZIP: Zero-inflated poisson

Acknowledgements

The authors acknowledge Professor Kalimuthu Krishnamoorthy from the Department of Mathematics at the University of Louisiana at Lafayette for early discussions on the topic treated in this paper. The authors are also thankful to two reviewers, whose comments helped improve the quality of this manuscript.

Authors' contributions

Zou contributed to the writing of the manuscript, development of methodology, and execution of simulation work. Hannig contributed to the writing of the manuscript, development of methodology, and development of the numerical routines. Young contributed to the organization and writing of the manuscript and designing the simulation studies. All authors read and approved the final manuscript.

Funding

This research was not funded by any institution.

Availability of data and materials

The UTI data are reported in the text. Simulation scripts for the results presented are available upon request from the authors.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Genentech, South San Francisco, California, USA. ²Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA. ³Dr. Bing Zhang Department of Statistics, University of Kentucky, Lexington, Kentucky, USA.

Received: 16 September 2020 Accepted: 5 February 2021

Published online: 06 March 2021

References

- Bayarri, M. J., Berger, J. O., Datta, G. S.: Objective Bayes Testing of Poisson Versus Inflated Poisson Models. In: Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh, B. Clarke and S. Ghosal (eds.) IMS Collections, vol. 3, pp. 105–121. Institute of Mathematical Statistics, Beachwood, (2008)
- Bhattacharya, A., Clarke, B. S., Datta, G. S.: A Bayesian test for excess zeros in a zero-inflated power series distribution. In: Balakrishnan, N., Peña, E. A., Silvapulle, M. J. (eds.) Beyond Parametrics in Interdisciplinary Research, pp. 89–104. Festschrift in Honor of Professor Pranab K. Sen, Institute of Mathematical Statistics, Beachwood, (2008)
- Cameron, A. C., Trivedi, P. K.: Regression-Based Tests for Overdispersion in the Poisson Model. *J. Econ.* **46**(3), 347–364 (1990)
- Cameron, A. C., Trivedi, P. K.: Regression Analysis of Count Data, 2nd edn. Cambridge, New York (2013)
- Cao, G., Hsu, W. W., Todem, D.: A Score-Type Test for Heterogeneity in Zero-Inflated Models in a Stratified Population. *Stat. Med.* **33**(12), 2103–2114 (2014)
- Deng, D., Paul, S. R.: Score Tests for Zero-Inflation and Over-Dispersion in Generalized Linear Models. *Stat. Sin.* **15**(1), 257–276 (2005)
- E, L., Hannig, J., Iyer, H.: Fiducial Intervals for Variance Components in an Unbalanced Two-Component Normal Mixed Linear Model. *J. Am. Stat. Assoc.* **103**(482), 854–865 (2008)
- Fisher, R. A.: The Fiducial Argument in Statistical Inference. *Ann. Eugenics.* **6**(4), 391–398 (1935)
- Gu, K., Ng, H. K., Tang, M. L., Schucany, W. R.: Testing the Ratio of Two Poisson Rates. *Biom. J.* **50**(2), 283–298 (2008)

- Hall, D. B., Berenhaut, K. S.: Score Tests for Heterogeneity and Overdispersion in Zero-Inflated Poisson and Binomial Regression Models. *Can. J. Stat.* **30**(3), 415–430 (2002)
- Hannig, J.: On Generalized Fiducial Inference. *Stat. Sin.* **19**(2), 491–544 (2009)
- Hannig, J., Iyer, H., Patterson, P.: Fiducial generalized confidence intervals. *J. Am. Stat. Assoc.* **101**(473), 254–269 (2006)
- Hannig, J., Iyer, H., Lai, R. C. S., Lee, T. C. M.: Generalized fiducial inference: A review and new results. *J. Am. Stat. Assoc.* **111**(515), 1346–1361 (2016)
- Itô, K.: *Poisson Point Processes and Their Applications to Markov Processes*. Springer, New York (2015)
- Janaskul, N., Hinde, J. P.: Score Tests for Zero-Inflated Poisson Models. *Comput. Stat. Data Anal.* **40**(1), 75–96 (2002)
- Janaskul, N., Hinde, J. P.: Score Tests for Extra-Zero Models in Zero-Inflated Negative Binomial Models. *Commun. Stat. Simul. Comput.* **38**(1), 92–108 (2008)
- Lambert, D.: Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. *Technometrics*. **34**(1), 1–14 (1992)
- Long, D. L., Preisser, J. S., Herring, A. H., Golin, C. E.: A Marginalized Zero-Inflated Poisson Regression Model with Overall Exposure Effects. *Stat. Med.* **33**(29), 5151–5165 (2014)
- Martin, J., Hall, D. B.: Marginal Zero-Inflated Regression Models for Count Regression. *J. Appl. Stat.* **44**(10), 1807–1826 (2017)
- Mathew, T., Young, D. S.: Fiducial-Based Tolerance Intervals for Some Discrete Distributions. *Comput. Stat. Data Anal.* **61**, 38–49 (2013)
- Mullahy, J.: Specification and Testing of Some Modified Count Data Models. *J. Econ.* **33**(3), 341–365 (1986)
- Ridout, M., Hinde, J., Demétrio CGB: A Score Test for Testing a Zero-Inflated Poisson Regression Model Against Zero-Inflated Negative Binomial Alternatives. *Biometrics*. **57**(1), 219–223 (2001)
- Springael, L., Van Nieuwenhuysse, I.: On the Sum of Independent Zero-Truncated Poisson Random Variables. Research paper 2006-011. June, 1–15 (2006)
- Temme, N. M.: Asymptotic Estimates of Stirling Numbers. *Stud. Appl. Math.* **89**(3), 233–243 (1993)
- Todem, D., Kim, K., Hsu, W. W.: Marginal Mean Models for Zero-Inflated Count Data. *Biometrics*. **72**(13), 986–994 (2016)
- Todem, D., Hsu, W. W., Fine, J. P.: A Quasi-Score Statistic for Homogeneity Testing Against Covariate-Varying Heterogeneity. *Scand. J. Stat.* **45**(3), 465–481 (2018)
- van den Broek, J.: A Score Test for Zero Inflation in a Poisson Distribution. *Biometrics*. **51**(2), 738–743 (1995)
- Waguespack, D., Krishnamoorthy, K., Lee, M.: Tests and Confidence Intervals for the Mean of a Zero-Inflated Poisson Distribution. *Am. J. Math. Manag. Sci.* **39**(4), 383–390 (2020)
- Weerahandi, S.: Generalized confidence intervals. In: *Exact statistical methods for data analysis*. Springer, New York, (1995)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)