Why Nematologists Should Use Zero-Inflated Models with Count Data And How To Use Them

Roberts, Lucas
rlucas7@vt.edu

Giese, W. Gill
giese@vt.edu

September 11, 2016

Abstract

Zero-inflated models are an essential yet seldom used tool by nematologists when modeling abundance data. Traditional sample means are inadequate estimators when excessive zeroes exist in the measurement counts. We show that using sample means when zero-inflation is present leads to a bias in the rate estimate causing scientists to underestimate the rate of nematodes in a soil sample. Moreover, we show that even if the nematologist collects a very large number of measurements, the error will persist. A test is provided which aids the scientist in determining the presence of zero-inflation. Additionally, an algorithm is given for estimation of the zero-inflation probability via a programmable spreadsheet. We provide a table of common statistical software with relevant information for the scientist to focus their software choice and how to use the chosen software.

1 Introduction

Raw nematodes counts, abundance measures, and rates within a unit volume of soil are all common measures of nematodes. The method of counting after elutriation is time consuming but conceptually straightforward for the researcher. In our own work [13] we found count models of root measures in vineyards to be inadequate statistical models for these observed data. The count data are often observed with a high prevalence of zeroes, indeed often beyond the frequency of zeroes suggested by the Poisson model itself. In cases with excess zeroes the Poisson model is inadequate to estimate the true, underlying rate of nematode abundance. Zero-inflated models offer an alternative that accounts for the excess zeroes observed as well as providing additional over dispersion flexibility. A distribution is called over-dispersed (relative to the Poisson) if the variance exceeds the mean, $\sigma^2 > \mu$.

In many cases nematologists, and scientists in general use the Poisson model not because it is appropriate for the observed data but because the scientist is familiar with the modeling approach of calculating sample means. When analyzing count data to obtain an estimate of the rate per unit, using a sample mean to estimate the rate is equivalent to assuming a Poisson count distribution. This article is meant to bring the zero-inflated models to the attention of scientists who model nematode abundance data and point out the pitfalls of using traditional methods in the presence of excess zero counts. The use of zero inflated models is common in other disciplines such as dentistry [4], road accidents [19], and software faults [16]. The zero-inflated Poisson model has been seen in the analysis of nematode count data by Murray et al. [17]. Murray et al. mentioned that the zero-inflation might come not from the lack of nematodes but from the symbiotic relation between certain species (southern root knot) and the root system, once the nematode embeds in the root system there is limited ability to measure the organism. The only time the organism may be measured is when recently-hatched larvae are detectable in the soil prior to invasion [17].

There are many readily available software applications for fitting zero-inflated models. We provide a table of common statistical software to fit zero-inflated models and compare relevant aspects of these software applications. Moreover, for the scientist who prefers to 'cook their own' statistics we also provide a

spreadsheet on the publisher's website so that the scientist may perform the calculations on simplified zero-inflated Poisson models themselves. Also the spreadsheet serves as an instructional aid. The scientist may study the formulae and calculations to ensure their understanding of the model and resulting calculations.

Beyond the statistical aspects of zero-inflated models for count data, there are very important scientific considerations for using these models. One salient consideration is that failing to use zero-inflated models when the data are in fact zero-inflated leads to a bias in the estimated rates of nematodes per unit volume of soil. The reported rate estimates determined under the traditional (sample average) models will under-estimate the rate parameter and may cause a grower to fail to act when in fact they should make efforts to control their nematode population either by rootstock selection, sanitation of field implements, and other mitigation methods like methyl bromide fumigation and beyond, as outlined by Zasada et al. [25]. Additionally, the scientist might assume that as long as they collect a sufficient number of samples the bias from misspecifying the Poisson model will be negligible. We show that the bias persists for any large number of measurements and a larger rate implies a larger error in estimation.

The remainder of this manuscript proceeds as follows. Section 2 discusses the theoretical challenges of rate estimation and presents the bias results for zero-inflated data. Section 3 gives a test statistic for the researcher to determine whether to use a zero-inflated or a Poisson model. Section 4 presents an EM algorithm to estimate the two parameters in a ZIP model. Also, we give a link to an excel spreadsheet where the algorithm is already programmed. Section 5 lists a table with comparisons of various software that may be used to fit a zero-inflated Poisson regression model. Finally, Section 6 concludes with a discussion and summary of the manuscript.

2   Estimation challenges

The bias of an estimator is defined as the difference between the expected value of the estimator and the parameter to be estimated. In the case of count data often the researcher wants to know the rate parameter which indicates an average number of nematodes per unit of soil volume. Let the rate parameter be denoted λ and let the zero inflation probability be denoted ϕ. If the researchers use the average of the counts, denoted $\bar{y}$, and the data are in fact distributed according to a ZIP distribution then the expected value of the sample mean is

$$\mathbb{E}(\bar{y}) = (1 - \phi)\lambda. \tag{1}$$

Thus the bias of the estimate is

$$\underbrace{\mathbb{E}(\bar{y}) - \lambda}_{=bias} = -\phi\lambda, \tag{2}$$

so that the larger the zero-inflation or the rate parameter, the larger in real terms, the bias of the estimate.

One common approach to analyzing excess zeroes in count data is to add 1 to each observed count. If the counts are indeed Poisson distributed, the resulting random variable is a size-biased Poisson [3, 2]. Regardless of the underlying distribution if you use the counts with 1 added to each, then the rate estimate will be biased by 1. To remove this bias you may safely subtract 1 from the average. Note that this does not solve the problem of excess zeroes nor does the procedure alleviate the bias of the estimator detailed above. Another common method especially when analyzing abundance data (e.g. [15]) is to condition on the count being greater than zero which is equivalent to ignoring the zero counts. Under a ZIP distribution one can show that the distribution $\Pr(Y = y|y > 0)$ has a zero-truncated Poisson distribution. For more on the statistical properties of this distribution see Cohen [8] and Singh [20]. Unfortunately, the sample mean is a biased estimate of the rate parameter λ in this zero-truncated case as well. Moreover, some authors prefer to use a transformed version of the count with a 1 added. Some examples are $log(y + 1)$ [14, 6] or $\sqrt{2(y + 1)}$ [1, 24] and similar variations to these two transforms. Both of these transforms lead to biased estimators of the underlying rate. The second transform which uses a squared root, has origins in the variance stabilization literature [12] and would be reasonable without the addition of a 1 *only if the underlying counts are Poisson*

distributed. In the applied papers where these transforms were applied, no likelihood ratio test to check the Poisson assumption was applied.

Of course the researcher might reasonably assume that with enough measurements, the bias is not a problem. This assumption is known as asymptotic unbiasedness in the statistics literature. We now detail that this assumption, although reasonable, is erroneous.

Perhaps the most common measure of the closeness of an estimator to the true parameter is the mean squared error (MSE). The MSE is defined as the expected value of the squared difference between the estimator and the true value of the parameter. Formally written,

$$MSE(\bar{y}, \lambda) = \mathbb{E}(\bar{y} - \lambda)^2, \tag{3}$$

We use the well-known bias variance decomposition of the MSE equation to calculate MSE in this case and we use the conditional variance formula for calculating the variance of the ZIP random variable. The bias has already been given above in Equation 1. The MSE for the average as an estimator of the rate is

$$MSE(\bar{y}, \lambda) = \phi^2 \lambda^2 + \frac{\lambda(1 - \phi) + \lambda^2(1 - \phi)\phi}{n}. \tag{4}$$

This last equation tells us that even with an infinite number of samples ($n \to \infty$) unless the zero-inflation is non-existent, there will be a bias proportional to the square of the (true) rate and the traditional methods will be suboptimal to use in place of estimators of zero-inflation. This section has discussed the statistical estimation of rates for data which are truly zero-inflated. In practice we never know for certain whether the data are truly zero-inflated or not and we must rely on statistical tests to make this determination. The next section discusses a method to make the Poisson or ZIP determination as well as a statistical nuance of the procedure.

3   Model determination: Poisson or ZIP?

There are two commonly used methods to determine whether the count data you are modeling is from a zero-inflated distribution or not: the likelihood ratio (LR) test and the Vuong test [22]. There is a fair amount of disagreement within the scholarly community about when to use the Vuong test or the LR test [23] in zero-inflated models. Notwithstanding these disputes, the Vuong test is actually a class of tests that are appropriate when comparing two zero-inflated regression models that may or may not be nested. We will not detail the Vuong test for reasons of scope but the interested reader is referred to the original paper by Vuong as well as two econometrics textbooks [5, 10]. The likelihood ratio test method is appropriate when comparing a Poisson model against a zero-inflated model. We will not detail the likelihood ratio test but rather give an asymptotically equivalent test known as the score test. We detail the score test because the form of the test statistic is simple, intuitive and easily calculated. The formula for the score statistic is

$$S_n = \frac{(n_0 - np_0)^2}{np_0(1 - p_0) - n\bar{y}p_0^2}, \tag{5}$$

where n_0 denotes the number of observations recorded 0, n is the total number of observations including the zeroes, p_0 is an estimate for the probability of observing a zero in a Poisson distribution, and $\bar{y}$ is the sample mean of all counts including the observed zeroes [21]. A common estimate for p_0 would be to use $\exp(-\bar{y})$, although the alternative estimator $(1 - 1/n)^{n\bar{y}}$ is also reasonable. Both estimates give very similar answers in most cases. If the observed counts are truly Poisson distributed then the test statistic S_n will be distributed according to a chi-squared distribution with 1 degree of freedom. We provide a table with common critical values for the chi-squared with 1 degree of freedom in Table 1.

The score test statistic is constructed by taking the derivative of the log-likelihood with respect to the parameter(s) assuming the null hypothesis is true. In our case the likelihood is a ZIP distribution. We parametrize in terms of the odds $\theta = \phi/(1 - \phi)$. Then the score function for θ is

$$\sum_{i=1}^{n} \left(\frac{-1}{1+\theta} + \mathbb{I}(y_i = 0)\frac{1}{(\theta + \exp[-\lambda_i])} \right). \tag{6}$$

One can readily check that the expected value of the score equation above is 0 when summed over all observations by using $\mathbb{E}(\mathbb{I}(y_i = 0)) = (\theta + e^{-\lambda_i})/(1+\theta)$. Note here the notation $\mathbb{I}(A)$ is an indicator function which takes the value 1 when the event A occurs and 0 otherwise. Using standard results in statistics, this score statistic will be asymptotically distributed with a chi-squared distribution with 1 degree of freedom. The score test *will* converge to the chi-squared distribution because the zero-inflation parameter (ϕ) does not have a boundary at the value 0 as is commonly assumed but at the value $-e^{\lambda}/(1-e^{\lambda})$. For cases when ϕ is less than 0 the term zero-inflated may not accurately describe reality because we are reducing the probability of the 0 values to a smaller probability than would be assumed with a Poisson distribution. The case when the likelihood ratio statistic has a parameter on the boundary of the parameter space is a well studied and understood problem within statistics that occurs in other models as well, such as a mixed effect model when you want to test whether one of the variance components is 0 or not [9]. In boundary cases the results state that the distribution of the test statistic has probability one-half at the zero point and then other one-half probability is spread over a chi-squared distribution with a single degree of freedom [7, 11, 18]. Table 1 gives the critical values of the chi-squared distribution for common type 1 error rates.
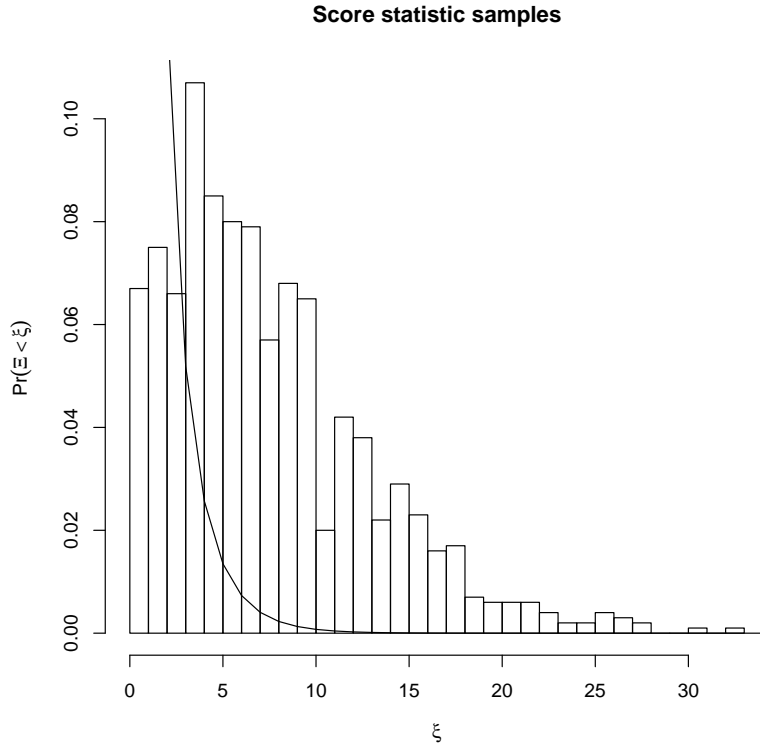

Figure 1: Simulated and theoretical values of the score statistic with Poisson random variables.

Score test values	
P-value	Critical value ξ
0.001	0.682447
0.01	0.680257
0.025	0.676564
0.05	0.670281
0.1	0.657218
0.25	0.613524

Table 1: The critical values of the score test.

4 An algorithm for estimating ϕ and λ with no covariates

In this section we detail an algorithm for estimating the two parameters of a ZIP model that is readily programmable within a spreadsheet with basic arithmetic calculations. Although the algorithm contains a while loop that is potentially infinite, in our experience the number of iterations until the algorithm converges is typically no more than 20. Moreover, the calculations required in each iteration are readily programmed so that the researcher may apply the technique to various data via copy and paste to mimic the iterations of the while loop. The algorithm follows:

Algorithm 1 ZIP EM algorithm

1: **procedure** ZIP EM$(\phi^0, \lambda^0, \varepsilon, \vec{y}, \vec{w})$
2:     **for** $i \leftarrow 1, n$ **do**
3:         **if** $y_i = 0$ **then**
4:             $w_i \leftarrow (1 + \exp[-\lambda_i^0 - \log(\phi_i^0/(1 - \phi_i^0))])^{-1}$
5:         **else**
6:             $w_i \leftarrow 0$
7:     $t \leftarrow 1$
8:     **while** $\varepsilon > |\phi^t - \phi^{t-1}|$ and $\varepsilon > |\lambda^t - \lambda^{t-1}|$ **do**
9:         $\phi^t \leftarrow \sum_{i=1}^{n} w_i/n$
10:
11:         $\lambda^t \leftarrow \sum_{i=1}^{n}(1 - w_i)y_i / \left(\sum_{i=1}^{n}(1 - w_i)\right)$
12:
13:         **for** $i \leftarrow 1, n$ **do**
14:             **if** $y_i = 0$ **then**
15:                 $w_i \leftarrow (1 + \exp[-\lambda_i^t - \log(\phi_i^t/(1 - \phi_i^t))])^{-1}$
16:             **else**
17:                 $w_i \leftarrow 0$
18:         $t \leftarrow t + 1$
19:
20:     **return** ϕ^t, λ^t $\triangleright$ The EM algorithm (MLE) estimates
21:

The algorithmic description is a formal way of stating the procedure to get the estimates which in essence involves iterating the calculations on lines 9 and 11 and updating the weights (w_i) in lines 15 and 17 after a first initialization of the relevant values. For initial values of ϕ^0, λ^0, and $\vec{z}$, we suggest random uniform numbers between 0 and 1 for all three. Note that final values output by the algorithm for w_i and ϕ^t will be between 0 and 1 while λ^t may be any positive value. Although a while loop with a stopping condition on the successive estimates of the rate and zero-inflation parameters is formally used in the algorithm, often replacing the stopping conditions with $t < 10$ or another sufficiently large whole number of iterations is

	Species				
	stunt	spiral	ring	dagger	rootknot
n	144	144	144	144	144
n_0	54	12	135	59	121
$\sum_{i=1}^{n} y_i$	12360	38120	420	4620	1920
$\bar{y}$	85.93	264.72	2.92	32.08	12.33
S	3.83×10^{38}	9.28×10^{114}	2.64×10^3	2.07×10^7	6.28×10^7

Table 2: Summary statistics for example zero-inflated Poisson analysis.

sufficient for many reasonable settings of numerical accuracy (ε).

A spreadsheet to accompany the manuscript is available to illustrate the necessary calculations as well as provide a working template for the nematologist. The spreadsheet is titled `zip_calcs.xlsx` and is available in the supplementary material. In the spreadsheet we use a two significant digit stopping rule and we also calculate the score statistic. The data provided in the spreadsheet are simulated Poisson variates so the score statistic has a large pvalue indicating evidence in favor of the null hypothesis that the data are distributed Poisson. Nonetheless, for illustrative purposes we proceed with the zero-inflated Poisson algorithm calculations so that the reader may study the calculation procedure. Also this example provides evidence that the algorithm provides a reasonable answer in the case of Poisson data, a small ($\phi \approx 0.03$) zero inflation estimate.

4.1 Numerical Example

In this section we present the calculation of the score statistic to determine zero-inflation status and then use the test outcome to determine the method of estimating the rate and-if applicable-the zero-inflation probability.

First consider the summary data in Table 2 although some of the calculated statistics are redundant we include them in the table for completeness. The sample mean row may be calculated using the formula $\sum_{i=1}^{n} y_i/n$.

First let us consider the stunt species. Recalling the formula for the score test statistic,

$$S_n = \frac{(n_0 - np_0)^2}{np_0(1-p_0) - n\bar{y}p_0^2} \tag{7}$$

$$= \frac{(54 - 144 \times e^{-85.94})^2}{144 \times e^{-85.94} \times (1 - e^{-85.94}) - 12360 \times e^{-2 \times 85.94}} \tag{8}$$

$$= \frac{2.92 \times 10^3}{7.61 \times 10^{-36}} \tag{9}$$

$$\approx 3.83 \times 10^{38} \tag{10}$$

Depending on how many significant digits are kept in the intermediate steps of the calculation the test statistic may be calculated slightly differently. However, in the stunt case the data are so overwhelmingly zero-inflated that the difference will not change the conclusion of the test; the data are zero-inflated.

Let us now turn to the ring species of nematode. The score statistic for these count data is:

$$S_n = \frac{(n_0 - np_0)^2}{np_0(1-p_0) - n\bar{y}p_0^2} \tag{11}$$

$$= \frac{(135 - 144 \times e^{-2.92})^2}{144 \times e^{-2.92} \times (1 - e^{-2.92}) - 420 \times e^{-2 \times 2.92}} \tag{12}$$

$$= \frac{1.62 \times 10^{24}}{6.14} \tag{13}$$

$$\approx 2.64 \times 10^3 \tag{14}$$

zero-inflated regression software comparison						
software	Poisson	ZIP	ZINB	LR test	Vuong test	cost
JMP	Y	Y				$
Minitab	Y*	Y	Y	Y		$
R	Y	Y	Y	Y	Y	OS
SAS	Y	Y	Y	Y	Y	$
Stata	Y					$
Statistica	Y					$
SPSS	Y					$

Table 3: A comparison of common statistical software for zero-inflated Poisson analysis.

This tells us that this specie is likely distributed zero-inflated as well. Using sample means to estimate rates in both of these cases would cause us to underestimate the true rates as predicted by the bias given above in Equation 1.

Now let us calculate the estimated rates under the zero-inflated Poisson model as well as under the Poisson model. Under the Poisson model the estimated rates are given by the sample means shown in Table 2 above. Under the ZIP model the rates for the stunt and for the ring are 86.791 and 5.98898 respectively. The difference between these estimates and the Poisson rate estimates are,

$$85.83 - 86.791 = -0.95766 \tag{15}$$

for stunt nematodes and

$$2.916667 - 5.98898 = -3.07232 \tag{16}$$

for ring nematodes.

Note that the differences here under the Poisson and the Zero-inflated Poisson models for ring species is substantial whereas for stunt nematodes the empirical bias is subtle. This matches the intuition that the larger the number of zeroes in the count data, the larger the bias. In the cases of stunt nematodes the increase in the amount rate parameter (λ) will impact the bias was more than offset by the decrease in the proportion of zeroes (ϕ).

5   Comparison of ZIP regression software

There are several commercially available and freely available versions of software routines to provide coefficient estimates in the case that you have covariate information on your nematode count measurements. One common example might be the type of ground cover or cover crop present on the soil sample that is elutriated to get the nematode counts. These covariate measurements might commonly occur in vineyard experimental designs as well as lab studies or from typical viticultural production practices. Table 5 displays the various common software packages used for statistical analysis and whether these software contain routine(s) for fitting zero-inflated regression models.

6   Discussion

In this brief research note we discussed the zero-inflated distributions and several statistical and scientific implications of modeling nematode abundance using standard statistical techniques with count data. We argue that nematologists should be familiar with zero-inflated models and use these methods in place of the sample mean statistic to estimate rates of nematodes in soil volume data where appropriate.

We provided a score test statistic for the researcher to decide between a Poisson or a ZIP model for a collection of observations. The score test may be used both whether covariates are available or not. When

deciding between two competing ZIP regression models with non-nested covariate structure the Vuong test should be used. Moreover, we stressed both the theoretical and computational aspects of these methods as well as statistical nuances of the methods. Finally, we provided an algorithm to estimate the rate of nematodes and the zero-inflation probability component under a zero-inflated Poisson model. Additionally, we provided a spreadsheet with calculations ready programmed for the researcher to modify as needed for their nematode abundance data analysis.

References

[1] Francis J Anscombe. The transformation of Poisson, binomial and negative-binomial data. *Biometrika*, 35(3/4):246–254, 1948.

[2] Richard Arratia and Larry Goldstein. Size bias, sampling, the waiting time paradox, and infinite divisibility: when is the increment independent? *arXiv preprint arXiv:1007.3910*, 2010.

[3] Richard Arratia, Larry Goldstein, and Fred Kochman. Size bias for one and all. *arXiv preprint arXiv:1308.2729*, 2013.

[4] Dankmar Böhning, Ekkehart Dietz, Peter Schlattmann, Lisette Mendonca, and Ursula Kirchner. The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 162(2):195–209, 1999.

[5] A Colin Cameron and Pravin K Trivedi. *Microeconometrics: methods and applications*. Cambridge university press, 2005.

[6] M Centinari, JE Vanden Heuvel, M Goebel, MS Smith, and TL Bauerle. Root-zone management practices impact above and below ground growth in Cabernet Franc grapevines. *Australian Journal of Grape and Wine Research*, 22(1):137–148, 2016.

[7] Herman Chernoff. On the distribution of the likelihood ratio. *The Annals of Mathematical Statistics*, pages 573–578, 1954.

[8] A Clifford Cohen. Estimating the parameter in a conditional Poisson distribution. *Biometrics*, 16(2):203–211, 1960.

[9] Ciprian M Crainiceanu and David Ruppert. Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):165–185, 2004.

[10] Russell Davidson and James G MacKinnon. *Estimation and inference in econometrics*. Cambridge Univ Press, 1995.

[11] Ziding Feng and Charles E McCulloch. Statistical inference using maximum likelihood estimation and the generalized likelihood ratio when the true parameter is on the boundary of the parameter space. *Statistics & Probability Letters*, 13(4):325–332, 1992.

[12] Murray F Freeman and John W Tukey. Transformations related to the angular and the square root. *The Annals of Mathematical Statistics*, pages 607–611, 1950.

[13] W Gill Giese, Tony K Wolf, Ciro Velasco-Cruz, and Lucas Roberts. Cover crop and root pruning effects on the rooting pattern of so4 rootstock grafted to Cabernet Sauvignon. *American Journal of Enology and Viticulture*, 67(1):105–115, 2016.

[14] Amanda D Howland, R Paul Schreiner, and Inga A Zasada. Spatial distribution of plant-parasitic nematodes in semi-arid Vitis vinifera vineyards in Washington. *Journal of nematology*, 46(4):321, 2014.

[15] Ganpati B Jagdale, Ted Holladay, PM Brannen, WO Cline, Paula Agudelo, AP Nyczepir, and JP Noe. Incidence and pathogenicity of plant-parasitic nematodes associated with blueberry (Vaccinium spp.) replant disease in Georgia and North Carolina. *Journal of nematology*, 45(2):92, 2013.

[16] Taghi M Khoshgoftaar, Kehan Gao, and Robert M Szabo. An application of zero-inflated Poisson regression for software fault prediction. In *Software Reliability Engineering, 2001. ISSRE 2001. Proceedings. 12th International Symposium on*, pages 66–73. IEEE, 2001.

[17] Leigh Murray, Stephen H Thomas, Jill Schroeder, Scott Kreider, Zhining Ou, JM Trojan, and C Fiore. Modeling the root-knot nematode/nutsedge pest complex: Perspectives from weed science, nematology and statistics. In *23rd Annual conference on applied statistics in agriculture*, pages 128–151. Kansas State University, 2011.

[18] Steven G Self and Kung-Yee Liang. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398):605–610, 1987.

[19] Venky Shankar, John Milton, and F Mannering. Modeling accident frequencies as zero-altered probability processes: an empirical inquiry. *Accident Analysis & Prevention*, 29(6):829–837, 1997.

[20] Jagbir Singh. A characterization of positive Poisson distribution and its statistical application. *SIAM Journal on Applied Mathematics*, 34(3):545–548, 1978.

[21] Jan Van den Broek. A score test for zero inflation in a Poisson distribution. *Biometrics*, pages 738–743, 1995.

[22] Quang H Vuong. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*, pages 307–333, 1989.

[23] Paul Wilson. The misuse of the Vuong test for non-nested models to test for zero-inflation. *Economics Letters*, 127:51–53, 2015.

[24] Guan Yu. Variance stabilizing transformations of Poisson, binomial and negative binomial distributions. *Statistics & Probability Letters*, 79(14):1621–1629, 2009.

[25] Inga A Zasada, John M Halbrendt, Nancy Kokalis-Burelle, James LaMondia, Michael V McKenry, and Joe W Noling. Managing nematodes without methyl bromide. *Annual review of phytopathology*, 48:311–328, 2010.