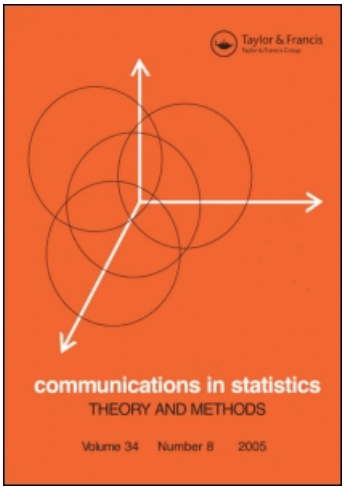

Communications in Statistics - Theory and Methods
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713597238

Full Bayesian Significance Test for Zero-Inflated Distributions

Josemar Rodrigues [a]
[a] UFSCar-DEs, São Carlos, SP, Brazil

To cite this Article Rodrigues, Josemar'Full Bayesian Significance Test for Zero-Inflated Distributions', Communications in Statistics - Theory and Methods, 35: 2, 299 — 307
To link to this Article: DOI: 10.1080/03610920500439984
URL: http://dx.doi.org/10.1080/03610920500439984

Communications in Statistics—Theory and Methods, 35: 299–307, 2006

ISSN: 0361-0926 print/1532-415X online
DOI: 10.1080/03610920500439984

Taylor & Francis
Taylor & Francis Group

Bayesian Inference

Full Bayesian Significance Test for Zero-Inflated Distributions

JOSEMAR RODRIGUES

UFSCar-DEs, São Carlos, SP-Brazil

In this article, we discuss an application of the Full Bayesian Significance Test (FBST) introduced by Pereira and Stern (1999) to compute the evidence of the Poisson distribution against the Zero-Inflated Poisson distribution (ZIP). The FBST is intuitive and easy to implement via Winbugs as an alternative to the classical tests formulated by Xie et al. (2001) in statistical process context. This evidence measure is based on the augmented data and used to test the fitting of the ZIP model for count data with excess of zeros in two illustrative examples in the statistical process control and the horticultural research.

Keywords Augmented data; Evidence; Gibbs sampling; Poisson distributions; Horticultural research; Mixture distributions; Statistical control process.

Mathematics Subject Classification 62F15.

1. The Zero-Inflated Distribution (ZID)

Usually the count data set with excess of zero can arise in many areas, particularly in quality studies in industry and horticultural research. This type of count data is modeled by the ZID and the classical fitting tests are based on asymptotic results (Xie et al., 2001). For example, the Poisson distribution has often been used for count data, however, this model does not provide a good fit to actual data when there is a frequent or excessive number of zero counts. In such a situation, the zero-inflated distribution (ZID) is more appropriate. In this article, this distribution is studied from the Bayesian point of view using the data augmentation algorithm. The zero-inflated Poisson distribution (ZIP) and illustrative examples, via MCMC algorithm implemented in Winbugs, are considered. In particular, the Full Bayesian Significance Test (FBST) is computed to verify a good fit of the Poisson distribution. The FBST is a measure of evidence of a precise hypothesis which is the posterior probability of the set of points having smaller density than the supremum over the

Received October 1, 2004; Accepted June 27, 2005
Address correspondence to Josemar Rodrigues, UFSCar-DEs-13565-905-São Carlos-SP-Brazil; E-mail: vjosemar@power.ufscar.br

Table 1
The discrete mixture distribution on θ

Θ	P
0	ω
θ	$1-\omega$

hypothesis. In the statistical process context this evidence measure can be used to check the validity of the Poisson distribution, since the ZIP model is complicated in order to construct the traditional c-chart. In this article, this measure is adapted to augmented data and easily computed via Winbugs. One interesting characteristic of the FBST based on the augmented data is not to protect the Poisson distribution against the inflated Poisson distribution when the parameter rate is large as do the classical tests.

There are many different ways to formulate the ZID model; in this article we consider the following way:

$$Pr[Y = y \mid \Theta = \theta] = \begin{cases} I_{\{\theta\}}(y), & \theta = 0 \\ p(y \mid \theta), & \theta > 0 \end{cases} \tag{1.1}$$

where $I_{\{0\}}(y)$ is a distribution which is degenerate at zero and $p(y \mid \theta)$ is a conditional probability function of Y given θ. The overdispersion discrete model is defined by

$$p(y) = Pr[Y = y] = \int_0^\infty Pr[Y = y \mid \theta] dP(\theta), \tag{1.2}$$

where P is a continuous or discrete measure on θ. We give a special attention to the following discrete measure shown in Table 1 for $0 \leq \omega < 1$. It is important to mention that this approach is restrictive since ω could be negative in some real cases. This kind of discrete measure P inflates the model $p(y \mid \theta)$ with zeros and gives the following ZID:

$$p(y \mid \theta, \omega) = \omega I_{\{0\}}(y) + (1-\omega) p(y \mid \theta). \tag{1.3}$$

For the ZID in (1.3) we have that

$$E_\theta[Y] = (1-\omega) E[Y \mid \theta] = \mu \quad \text{and} \quad Var_\theta[Y] = (1-\omega) Var[Y \mid \theta] + \frac{\omega}{1-\omega}\mu^2. \tag{1.4}$$

The second equation is a more general form to show overdispersion than the usual quadratic variance function. For example, for the zero-inflated Poisson we have

$$Var[Y] = \mu + \frac{\omega}{1-\omega}\mu^2.$$

2. The FBST Based on the Augmented Data

Suppose that $Y = (Y_1, \ldots, Y_n)$ is a vector of n independent random variables generated by the ZID model. Let $A = \{y_i : y_i = 0, i = 1, \ldots, n\}$ and $m = \#(A)$, then

the likelihood function is

$$L_Y[\theta, \omega] = [\omega + (1-\omega)p(0 \mid \theta)]^m (1-\omega)^{n-m} \prod_{y_i \notin A} p(y_i \mid \theta).$$

The elements of the set A come from either of two different groups, the degenerated distribution at zero or $p(0 \mid \theta)$. In this situation, it is natural to define an unobserved (or missing) latent allocation variable:

$$I_i = \begin{cases} 1 & \text{with probability } p(\theta, \omega), \\ 0 & \text{otherwise,} \end{cases}$$

for $i \in A$ and

$$p(\theta, \omega) = \frac{\omega}{\omega + (1-\omega)p(0 \mid \theta)}. \tag{2.1}$$

This latent variable indicates whether the ith element of A is drawn from the first component of (1.3) or not. So, the likelihood function based on the augmented data $D = \{Y, I\}$ (Tanner and Wong, 1987), where $I = (I_1, \ldots, I_m)$ is

$$\begin{aligned} L_D[\theta, \omega] &= L_Y[\omega, \theta] \prod_{i=1}^{m} p(\theta, \omega)^{I_i} (1 - p(\theta, \omega))^{1-I_i} \\ &= \underbrace{\omega^S (1-\omega)^{n-S}}_{\text{inflated zeros}} \underbrace{p(0 \mid \theta)^{m-S} \prod_{y_i \notin A} p(y_i \mid \theta)}_{\text{data from the model}} = L_1(\omega) L_2(\theta), \end{aligned} \tag{2.2}$$

where $S = \sum_{i=1}^{m} I_i \sim Bin[m, p(\theta, \omega)]$. Assuming a joint prior, $\pi(\theta, \omega)$, the joint posterior of (θ, ω), given D, is

$$\pi(\theta, \omega \mid D) \propto L_D[\theta, \omega] \pi(\theta, \omega).$$

2.1. *FBST via the Augmented Data*

The computation of the evidence measure for

$$H_o : \omega = 0 \quad \text{versus} \quad H_1 : \omega > 0 \tag{2.3}$$

is performed (see Pereira and Stern, 1999) in two steps, a numerical optimization step and a numerical integration step. These two steps are easily implemented via Winbugs and we only need the knowledge of the parameter space represented by the posterior distribution, avoiding the most important argument against Bayesian test called the Lindley's paradox.

Since the parameters w and θ are unrelated due to the factorization of the likelihood funtion, the FBST will be based on the marginal likelihood $L_2(\theta)$ as follows:

- Optimization step: Finding the mode, θ_o, of $\pi_o(\theta \mid Y)$ under $H_o : \theta \in \Theta_o$, where $\Theta_o = \{\theta : w = 0\}$, and the posterior density, $\pi_0(\theta \mid Y)$, is given by

$$\pi_o(\theta \mid Y) = \pi(\theta, 0) \prod_{i=1}^{n} p(y_i \mid \theta).$$

- Integration step:

$$Ev(H_o \mid D) = 1 - Pr[\theta \in T^*(D) \mid D] = 1 - \int_0^\infty \int_{T^*(D)}^\infty \pi(\theta, \omega \mid D) d\theta d\omega.$$

where

$$T^*(D) = \{\theta : \pi(\theta \mid D) \geq \pi(\theta_o \mid D)\}$$

Remarks. The FBST has the following properties (for more details we suggest refering to Pereira and Stern, 1999):

- We have that $Ev(H_o \mid D) = 1$ if and only if $\omega = 0$, so, if this evidence measure is "small" it means that the null set, $\{\theta_o\}$, is in a region of low posterior probability, and, consequently, the data D gives strong evidence in favour of $H_1 : \omega > 0$.
- This procedure is the FBST introduced by Pereira and Stern (1999) to test H_o, where T^* is the Highest Density Probability Set.
- Increasing the sample size the FBST converges to right 0/1 value (accept/reject decision).
- Considering only the observed sample allowing no adhoc artifice like a positive prior distribution on the precise hypothesis.
- For small values of ω and large values of θ it can be seen from (2.1) that the FSBT does not give protection (a small probability type one error) to the Poisson distribution as do the classical tests listed in Table 5 (Xie et al., 2001).

3. The Inflated Poisson Distribution

From now on, our detailed exposition is limited to the Poisson model

$$p(y \mid \theta) = \frac{\theta^y e^{-\theta}}{y!}.$$

However, the metodology is generic and can be applied to other discrete distributions. The likelihood function based on D is

$$L_D[\theta, \omega] \propto \omega^S (1-\omega)^{n-S} \theta^{\sum_{y_i \notin A} y_i} e^{-(n-S)\theta} = L_1(\omega) L_2(\theta),$$

where $S \sim Bin[m, p(\theta, \omega)]$ and

$$p(\theta, \omega) = \frac{\omega}{\omega + (1-\omega)e^{-\theta}}.$$

The likelihood function suggests the following independent priors:

$$\pi(\theta) \sim \Gamma[a, b] \quad \text{and} \quad \pi(\omega) \sim Beta[c, d]. \tag{3.1}$$

So, the joint posterior distribution for (θ, ω), given D, is

$$\pi(\theta, \omega \mid D) \propto \omega^{S+c-1} (1-\omega)^{n-S+d-1} \theta^{T+a-1} e^{-(n-S+b)\theta}, \tag{3.2}$$

where $T = \sum_{y_i \notin A} y_i$.

3.1. *Posterior Simulation Using MCMC Algorithm*

It is very easy to implement via Winbugs (see the Winbugs code in Sec. 4) and it consists of the following two steps:

- Step 1. Given $(\theta^{(j-1)}, \omega^{(j-1)})$ at the $(j-1)$-stage we draw $S^{(j)}$ from the $Bin[m, p(\theta^{(j-1)}, \omega^{(j-1)})]$.
- Step 2. Given $S^{(j)}$, we draw $(\theta^{(j)}, \omega^{(j)})$ from the densities

$$\begin{aligned} \omega^{(j)} &\sim Beta\left[S^{(j)} + c, n - S^{(j)} + d\right] \\ \theta^{(j)} &\sim \Gamma\left[\sum_{y_i \notin A} y_i + a, n - S^{(j)} + b\right] \end{aligned} \tag{3.3}$$

3.2. *FBST via the Augmented Data*

The computation of the evidence measure for

$$H_o : \omega = 0 \quad \text{versus} \quad H_1 : \omega > 0 \tag{3.4}$$

is performed in the following two steps:

- Optimization step. The mode, θ_o, of $\pi_o(\theta \mid Y)$ under $H_o : \theta \in \Theta_o$ is

$$\theta_o = \frac{T + a - 1}{n + b},$$

  where $\pi_o(\theta \mid Y)$ is the gamma density with parameters $T + a$ and $n + b$.
- Integration step.

$$Ev(H_o \mid D) = 1 - Pr[\theta \in T^{\star}(D) \mid D],$$

  where,

$$T^{\star}(D) = \{\theta : \pi(\theta \mid D) \geq \pi(\theta_o \mid D)\}.$$

Remark. The marginal density, $\pi(\theta \mid D)$, corresponds to a gamma distribution with the shape parameter $\sum_{i \notin A} y_i + a$ and the scale parameter $n - S + b$ which is maximized at $\frac{T+a-1}{n-S+b}$. Also, we remind that, under this posterior distribution, the FBST based on the augmented data is equivalent to the FBST formulated by Pereira and Stern (1999), to test $H_o : \theta = \frac{T+a-1}{n+b}$.

4. Some Illustrative Examples

Example 4.1 (Horticultural Research)**.** As an illustrative example of the ZIP model we consider the Poisson data (Table 3) provided by Marin and Jones (1993). The data are the number of roots produced by 270 micropropagated shoots of the columnar apple cultivar *Trajan*. During the rooting period, all shoots were maintained under identical conditions, but the shoots themselves were cultered on media containing different concentrations of the cytokimin BAP, in growth cabinets with an 8- or 16-hour photoperiod. The full experimental background is given by Marin and Jones (1993). As discussed by Rideout et al. (1998), the large numbers of zeros for 16-hour photoperiod are an obvious problem for the Poisson fit.

Table 2
Posterior quantiles and the FBST

	Photoperiod					
	8			16		
quantile	2.5	50	97.5	2.5	50	97.5
θ	6.72	7.15	7.57	4.84	5.36	5.95
ω	0.004	0.017	0.047	0.39	0.47	0.55
FBST		0.66			0.0	

The posterior summaries using the Bayesian procedure proposed in the previous sections are given in Table 2.

We have the following conclusions with respect to Example 4.1:

- It is clear from results of Table 3 that the Poisson distribution has serious problems of fitting for 16-hour photoperiod.
- Table 3 shows that the ZIP, with the non informative priors given by $c = d = 1$ and $a = b = 1.0E - 10$ is definitely better than the ordinary Poisson distribution.
- In Table 3, it interesting to pay attention for the nice fitting of the ZIP distribution at the point zero.
- Also, we have a reasonable overall fit as compared with the simple Poisson model. The same result was found by Rideout et al., 1998, using asymptotic results.
- These examples show that it is quite easy to implement this Bayesian procedure and the FBST in Winbugs.
- Also, in Table 2, the parameter ω shows in a simple way how the overdispersion is occurring at different levels of photoperiod. Although, we have a few zeros ($m = 2$) for the photoperiod level 8, the FBST gives a small evidence for H_o. It is reasonable because the parameter θ is quite large for this photoperiod level. This confirms our statement that the FSBT based on augmented data does not give protection to the Poisson distribution for large values of θ.

Example 4.2 (Statistical Process Control)**.** This example was discussed by Xie et al. (2001) with the purpose of comparing many different classical tests based on asymptotic likelihood theory. The data set in Table 4 is the read write errors discovered in a computer hard disk in a manufacturing process. In the statistical control process context (Xie et al., 2001) the variable Y is the number of non comformities in the unit after submitting a random shock which occurs with probability $1 - \omega$. It is assumed that Y follows a ZIP distribution. The upper control limit y_u for a control chart based on the number of non conformities can be obtained from the Bayesian point of view as the smallest integer solution of

$$P[Y \geq y_u \mid D] = \int p(y \mid \theta, \omega)\pi(\theta, \omega \mid D)d\theta\, d\omega \leq \alpha, \tag{4.1}$$

where α is the predetermined false alarm probability for the upper control limit y_u.

Table 3
Fitted frequency for the Poisson and ZIP models

	Photoperiod					
	8			16		
No. of roots	O	E_{ZIP}	E_P	O	E_{ZIP}	E_P
0	2	2.91	0.11	62	61.68	7.43
1	3	0.78	0.82	7	1.74	21.27
2	6	2.78	2.91	7	4.63	30.43
3	7	6.60	6.89	8	8.23	29.02
4	13	11.75	12.23	8	11.00	20.76
5	12	16.76	17.36	6	11.80	11.88
6	14	19.93	20.55	10	10.57	5.66
7	17	20.34	20.84	4	8.13	2.31
8	21	18.18	18.50	2	5.49	0.82
9	14	14.45	14.59	7	3.31	0.26
10	13	10.35	10.36	4	1.79	0.07
11	10	6.74	6.68	2	0.89	0.01
12	2	4.03	3.95	3	0.40	0.00
13	2	2.2	2.16			
14	3	1.14	1.09			
17	1	0.54	0.51			
No. of shoots	140			130		
	$\chi^2_{\text{ZIP}} = 21.22$, $\chi^2_P = 51.72$			$\chi^2_{\text{ZIP}} = 49.85$, $\chi^2_P = 2954$		
	$\chi^2_j = \sum_i \frac{(O_i - E_i)^2}{E_i}$, $j = P$, ZIP					
	Ev = 0.66			Ev = 0.0		

Fitted frequency: O = Total observed frequency; E_P = Fitted Poisson frequency; E_{ZIP} = Fitted ZIP frequency

For the data set in Table 4, (4.1) can be easily calculated via Winbugs given $P[Y \geq 14 \mid D] = 0.0062$ at the acceptable false rate $\alpha = 0.01$. This means that there should not be any alarm for values less or equal to 14 when the underlying model is the ZIP distribution. Using a non Bayesian approach, Xie et al. (2001) found the same value $y_u = 14$ at $\alpha = 0.01$.

The conclusions for this example are:

- Table 5 shows a perfect agreement between the classical tests and the FBST. Also, Table 6 confirms the MLE estimators obtained by Xie et al. (2001).
- The data set in Table 4 contains many units with zeros, that is, $m = 180$ and $n = 208$. It is clear from Table 5 that the ZIP model should be used instead of the Poisson model. The FBST gives a strong evidence in favour of the ZIP model.
- Winbugs code for Example 4.2:
  model;

$$\{S \sim dbin(pw, m)$$
$$pw \leftarrow w/(w + (1 - w) * exp(-theta))$$

$theta \sim dgamma(r, mu)$

$r \leftarrow y + a$

$y \leftarrow inprod(v1[\,], v2[\,])$

$mu \leftarrow n - S + b$

$w \sim dbeta(a1, mu1)$

$a1 \leftarrow S + c$

$mu1 \leftarrow n - S + d$

$for\ (i \text{ in } 1 : 10)\{O[i] \leftarrow n * (w * equals(v1[i], 0) + (1 - w) *$

$exp\ (-theta + v1[i] * log(theta) - logfact(v1[i]))\}$

$thetao \leftarrow (r - 1)/(n + b)$

$logLthetao \leftarrow (y + a - 1) * log(thetao) - (n - S + b) * thetao$

$logLtheta \leftarrow (y + a - 1) * log(theta) - (n - S + b) * theta$

$Ev \leftarrow 1 - step(logLtheta - logLthetao)\}$

$\%Data$

$list(v1 = c(0, 1, 2, 3, 4, 6, 9, 11, 15, 75), v2 = c(180, 11, 5, 2, 1, 2, 2, 1, 1, 2),$

$m = 180, c = 1, d = 1, a = 1.0E - 10, b = 1.0E - 10, n = 208)$

$\%Initial\ \ values$:

$list(w = 0.5, theta = 7)$

Table 4
A set of defect count data from a manufacturing process

0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	6	0	9
11	0	1	2	0	0	0	0	0	0	0	0	3	3	0	0	5	0	15	6
0	0	0	4	2	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
75	0	0	0	0	75	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0
0	2	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
0	0	1	0	0	0	0	0												

Table 5
Summary of the test statistics and the FBST

Test methods	Test statistic	Critical region	Accept/reject H_o
Score test, S_1	628.135	$S_1 > 6.6349$	Reject
Likelihood ratio test, S_2	806.243	$S_2 > 6.6349$	Reject
Chi-square test, S_3	350.197	$S_3 > 15.0863$	Reject
Confidence Interval test, S_4	0.1933	$S_4 < 1$	Reject
C test, S_5	25.0626	$\|S_5\| > 2.5758$	Reject
R test, S_6	25.0925	$\|S_6\| > 2.5758$	Reject
FBST			Ev = 0

Table 6
Posterior quantiles and the MLE

		Quantile			
	Mean	2.5	50	97.5	MLE
θ	8.454	7.427	8.435	9.574	8.6413
ω	0.8618	0.8125	0.863	0.905	0.8654

5. Conclusions

In this article, we consider the Full Bayesian Significance Test of Pereira and Stern (1999) based on the augmented data to ckeck if the data come from a Poisson or the Zero-Inflated Poisson model. This test is intuitive and easy to implement via Winbugs as an alternative to the classical tests or Bayes factors.

Acknowledgment

The author is very grateful to the anonymous referee for his helpful comments and valuable suggestions on previous version of this article.

References

Marin, J. O., Jones, W. (1993). Hadlow micropropagation of columnar apple trees. *J. Horticultural Sci.* 68:289–297.

Pereira, C. A. B., Stern, J. M. (1999). Evidence and credibility: full Bayesian significance test for precise hypothesis. *Entropy* 1:69–80.

Rideout, M., Demétrio, C. G. B., Hinde, J. (1998). Models for count data with many zeros. *Int. Biometric Conf.* Cape Town, South Africa.

Tanner, M. A., Wong, W. W. (1987). The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.* 82:528–540.

Xie, M., He, B., Goh, T. N. (2001). Zero-inflated Poisson model in statistical process control. *Computat. Statist. Data Anal.* 38:191–201.