A simple and intuitive test for number-inflation or number-deflation

Paul Wilson[1], Jochen Einbeck[2]

[1] School of Mathematics and Computer Science/Statistical Cybermetrics Research Group, University of Wolverhampton, WV1 1LY, United Kingdom
[2] Department of Mathematical Sciences, Durham University, DH1 3LE, United Kingdom

E-mail for correspondence: `pauljwilson@wlv.ac.uk`

Abstract: We present a test of zero-modification which checks if the number of zeros is consistent with the hypothesized count distribution. This test is easily extended to test for inflation or deflation of *any non-negative values, and, by performing multiple tests of inflation/deflation of the counts present in observed data relative to* any given model, it is possible to assess the suitability of that model. Such multiple testing may be represented diagrammatically. The test for number-inflation/deflation is informally called the "Christmas Eve Test" as the original idea occurred to the main author on December 24th, 2014, and the diagrammatic method the "Durham Diagram" as it was developed during preparation for a talk at Durham University.

1 Problem and methodology

We are given random draws Y_i, $i = 1, \ldots, n$ from some count distribution, which is hypothesized to possess a specific parametric density function $f(y_i, \Theta_i)$. For instance, f may be the Poisson density, and Θ_i may correspond to a linear predictor $z_i^{\mathrm{T}}\beta$, with z_i a vector of covariates and β an unspecified parameter vector. We are interested in testing whether this distributional assumption is correct, or, in other words, whether the observed data are consistent with this specification.

We will consider initially the particular question of whether the observed number of zero's is consistent with this assumption (but generalize this idea to other values later on). Therefore, let $E(Y_i) = \mu_i$ and $p_i = P(Y_i = 0)$. Hence if X_i is a random variable which takes the value 1 if $Y_i = 0$ and 0 otherwise then X_i is a Bernoulli random variable with parameter p_i.

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

If $\mu_1 = \cdots = \mu_n = \mu$, i.e. μ does not depend on covariates, and hence all the p_i's are equal also, then the distribution of the number of zeros among $Y_1, \ldots Y_n$ is the sum of n independent Bernoulli random variables with parameter p, and hence is the binomial distribution $Bin(n, p)$, and thus has mean np and variance $np(1-p)$. Of more interest is when μ_i *does* depend on covariates, and hence the p_i's are not all equal. The sum, S_X, of n independent Bernoulli random variables $X_1, \ldots, X_n$ with parameters $p_1, \ldots, p_n$ respectively is known as a *Poisson–Binomial* distribution (Chen and Liu, 1997):

$$P(S_X = k) = \left\{ \prod_{i=1}^{n} (1 - p_i) \right\} \sum_{i_1 < \cdots < i_k} w_{i1} \cdots w_{ik}, \tag{1}$$

where $w_i = \frac{p_i}{1-p_i}$, $i = 1, \ldots, n$, and the summation is over all possible combinations of distinct $i_1, \ldots, i_k$ from $\{1, \ldots, n\}$. The R package *poibin* (Hong, 2013) implements both exact and approximate methods for computing the cdf of the Poisson-Binomial distribution.

2 The Christmas Eve test

We wish to determine whether data is zero-inflated or zero-deflated relative to a conditional count distribution (for instance, Poisson). The procedure for this is as follows:
(i) Fit the model according to the hypothesized count distribution; (ii) For each Y_i, estimate $P(Y_i = 0) = p_i$; (iii) Use *poibin* to determine a (say) 90% confidence interval.
If the observed number of zeros in the data exceeds the upper limit of the confidence interval then we have evidence of zero-inflation. If it is less than the lower limit we have evidence of zero–deflation. (Alternatively *poibin* will return a p value.)

2.1 Example

The $n = 100$ observations in the following table were simulated from a ZIP distribution with zero-inflation parameter 0.2 and a Poisson mean uniformly distributed on $[0.5, 1.5]$.

Y	0	1	2	3	4	5	6	7
Count	39	18	17	16	7	0	2	1

Representing the vector of Poisson means by `Z`, the three steps outlined in the previous subsection are implemented through the following R-code:

```
(i)    mod <- glm(Y ~ Z, family = poisson)
(ii)   mfv <- dpois(0, mod$fitted.values)
(iii)  qpoibin(c(0.05,0.95), pp = mfv)
```

Step (iii) returns the 90% confidence interval $[19, 35]$ for the expected number of zeros, hence as the observed number of zeros is greater than the upper limit of the confidence interval we may reject the null hypothesis that the observed number of zeros is consistent with a Poisson distribution.

2.2 Parameter estimation, power and type-one error rates

As visible from step (i) of the above example, the procedure requires estimation of the means μ_i under the hypothesis that the count distribution is correctly specified. However, these estimates may be poor if this hypothesis is wrong, rendering the distribution (1) incorrect too.
Indeed, we found in further investigation that the estimation of the Poisson parameter will be generally biased (but reasonably precise) if the data are in fact zero–inflated. However, by estimating the mean parameter from the truncated, positive data only, the estimates of the mean parameter became unbiased, but imprecise.
Simulations show that a combination of the two approaches is successful: excellent power and type-one error rates are achieved when the Poisson parameter is estimated as a $2:1$ weighted mean of the the two estimators. The type-one error rates and powers obtained when the Christmas Eve test is used as a test of zero-inflation for 100 observed data are shown in Figure 1. The corresponding rates for a score test and a likelihood ratio test are also illustrated. As is apparent the Christmas Eve test is the most powerful. Its type one error rate is comparable to that of the score test, but behaves better than that of the likelihood ratio test.

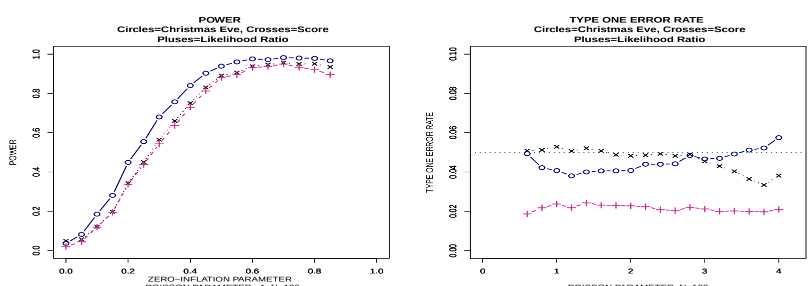

FIGURE 1. Power and Type One Error rates

3 Extending the test to positive values

The code of Section 2.1 may easily be modified to obtain confidence intervals for other values. For example a 90% confidence interval for the number of 1's under the Poisson model may be estimated to be $[20, 34]$, indicating that Y is "one-deflated" relative to the fitted Poisson model. Similarly it may be shown that $[23, 41]$, $[14, 30,]$, $[6, 18]$, $[1, 9]$, $[0, 5]$, $[0, 3]$ and $[0, 1]$ are 90% confidence intervals for the number of 2's, 3's, 4's, 5's, 6's and 7's under the Poisson model. This may be illustrated diagrammatically by a "Durham Diagram". In the left-hand diagram of Figure 2 the dotted lines represent the upper and lower limits of the confidence intervals for the counts under the Poisson model, and the dashed line the observed values. If the data is consistent with the reference model the dashed line should in general stay within the confidence bands, and departures from within the confidence bands indicate possible unsuitability of the reference model, and hence the left-hand diagram of Figure 2 indicates that a Poisson model is not suitable. The right-hand diagram of Figure 2 is constructed taking a zero-inflated Poisson distribution as the reference model; here we see that none of the observed counts exceed or fall short of the confidence intervals, indicating that a zero-inflated model may be suitable for the data.

4 Conclusion

The Christmas Eve Test is a highly intuitive test that when used to test zero-inflation has superior power to score and likelihood ratio tests, and an excellent type-one error rate. Whilst the Christmas Eve Test, including its extension to values other than zero, was originally developed with respect to zero-inflation it may be used to assess the suitability of any model for observed data.

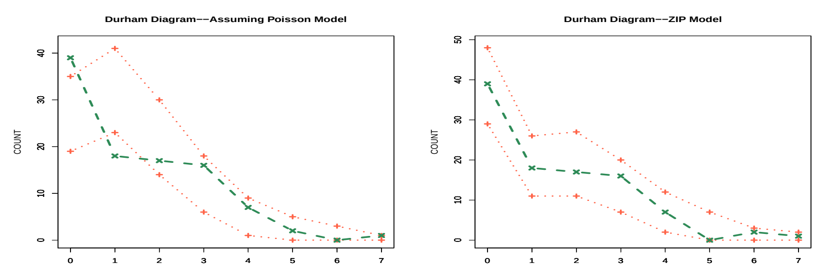

FIGURE 2. Durham Diagrams: Assuming Poisson and Zero-inflated Poisson

References

Chen, S.X. and Liu, J.S. (1997). Statistical applications of the Poisson-Binomial and conditional Bernoulli distributions. *Statistica Sinica*, **7**, 875 − 892.

Hong, Y. (2013). poibin: The Poisson Binomial distribution. url = http://CRAN.R-project.org/package=poibin.