Sample quantiles corresponding to mid p-values for zero–modification tests

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: Wilson and Einbeck (2015, 2016) propose a test for zero-modification relative to a stated model. The basis of the test is that the number of observed zeros follows a Poisson-binomial distribution. The decision to reject, or otherwise, the non zero–modified model is made by either (i) computing the mid p-value corresponding to the number of observed zeros, or (ii) comparing the number of observed zeros to the relevant "traditional" quantile of the appropriate Poisson–binomial distribution. In general either approach will result in the same decision, but occasionally discrepancies may occur. In this paper we investigate the use of mid-distribution quantiles in approach (ii) above, and show that this reduces the possibility of discrepancies.

Keywords: zero-modification, mid p-values, quantiles

1 Introduction

Wilson and Einbeck (2015) proposed a new and intuitive test for zero-modification that uses the observed number of zeros, n_0, in a given sample $y = y_1, y_2, \ldots, y_n$ from count variables Y_i and a set of covariates x_i, $i = 1, 2, \ldots, n$ to establish whether the distributional assumption $Y_i|x_i \sim G(y_i|\mu_i)$ where μ_i is a pre-specified parametric function of the x_i is consistent with N_0, the distribution of the number of zeros under G. This is achieved by referencing the value of n_0 to the appropriate Poisson-Binomial distribution (Chen and Liu, 1997).
To illustrate, consider the case where G is a Poisson model, and thus $p_i = p(0|\mu_i) = e^{-\mu_i}$ and let T_i be a random variable which takes the value 1

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

if $y_i = 0$ and 0 otherwise. Clearly T_i is a Bernoulli random variable with parameter p_i and thus N_0 may be formulated as the sum over independent Bernoulli experiments $T_1, T_2, \ldots, T_n$.
Based on this simple observation, consider the special case that there are no covariates, that is $\mu_1 = \mu_2 = \cdots = \mu_n = \mu$. In this case, the p_i's are equal also, and so the distribution of N_0 is a binomial distribution $Bin(n, p)$, where $p = e^{-\mu}$, and thus has mean np and variance $np(1-p)$. Based on this distribution, one can immediately compute quantiles corresponding to a given significance level, and use these as critical values for the test; alternatively one may determine the p-value corresponding to n_0, and reject or otherwise the Poisson model based upon this. If the μ_i do depend on covariates, N_0 is the sum of n independent Bernoulli random variables $T_1, T_2, \ldots, T_n$, and hence is a Poisson-Binomial distribution with parameters $p_1, p_2, \ldots, p_n$ and one proceeds by computing quantiles or p-values relative to this distribution, using, for example, the R package `poibin` (Hong, 2013).
Wilson and Einbeck (2016) proposed the use of *mid p-values*

$$\hat{\alpha}_{T,0.5}(t) = P_0[T > t] + 0.5P_0[T = t] = 0.5\left(P_0\left[T \geq t\right] + P_0[T \geq t+1]\right)$$

which Franck (1986) argues are more appropriate when the test statistic is discrete. Note that if T were continuous, then $P_0[T = t] = 0$ and the mid p-value is equivalent to the "traditional" p-value. It may be shown that the attainment rate of the proposed test when mid p-values are employed is superior to that when traditional p-values are used.
Wilson and Einbeck (2015) utilise the "traditional" quantile $Q(p) = \inf\{t \mid F(t) \geq p\}$ where $F(x) = P(X \leq x)$ is the cumulative distribution function of a random variable X. This may lead to discrepancies. An example, based upon the one-sided version of the test (i.e. we are testing for zero-inflation only), is the following:

1.1 Trajan Data

The data are the number of roots produced by $n = 270$ micropropagated shoots of the columnar apple cultivar Trajan. During the rooting period, all shoots were maintained under identical conditions, but the shoots themselves were cultured on media containing different concentrations of the cytokinin BAP, in growth cabinets with an 8 or 16 hour photoperiod. Full details of the experiment are to be found in Marin (1993). A striking feature of the data is that although almost all shoots produced under the 8 hour photoperiod rooted, only about half of those produced under the 16 hour photoperiod did. Overall $n_0 = 64$ shoots produced zero roots, of which only 2 were from the shorter photoperiod.
These data were analysed by Ridout and Demétrio (1992) and Ridout et al. (1998). If the model of the null hypothesis is a negative binomial (type–II) model, where both the mean and the size parameter are modelled by

photoperiod, then a (mid) p-value of 0.0871 for the test of Wilson and Einbeck (2015, 2016) is returned, indicating non-rejection of the negative binomial model at $\alpha = 0.05$. The traditional $5th$ and $95th$ quantiles of the distribution of N_0 are 47 and 66; the interval $[47, 66]$ is referred to as a 90% *fluctuation interval*. As $n_0 = 64$ is interior to this interval we conclude that n_0 *is* consistent with such a model (and inconsistent with the zero-inflated model) at a level of significance of $\alpha = 0.05$. An 80% fluctuation interval however is $[49, 64]$, and thus based upon this fluctuation interval we would fail to reject the negative-binomial model in favour of the strictly zero-inflated model at a level of significance of 0.10, but we would do so under the "p–value criterion".

2 Quantiles based on mid-distribution functions

Let X be a discrete random variable with distinct values $v_1 < v_2 < \cdots < v_d$, let $P(X = v_i) = p_i$. Ma et al. (2011) recommend the following quantile function for discrete distributions:

$$Q(p) = F^{-1}_{\text{mid}}(p) = \begin{cases} v_1 & \text{if } p < p_1/2 \\ v_k & \text{if } p = \pi_k,\ k = 1, \ldots, d \\ \lambda v_k + (1-\lambda) v_{k+1} & \text{if } p = \lambda \pi_k + (1-\lambda)\pi_{k+1} \\ & 0 < \lambda < 1,\ \ k = 1, \ldots, d-1 \\ v_d & \text{if } p > \pi_d \end{cases}$$

Where $\pi_k = \sum_{i=1}^{k-1} p_i + p_k/2$, that is, π_k is a lower-tailed mid-p-value.

2.1 Example: Mid Quantiles for a Binomial Distribution

Let $X \sim \text{Bin}(7, 0.35)$, and thus X has pmf and cdf:

x	0	1	2	3	4	5	6	7
$P(X = x)$	0.049	0.185	0.298	0.268	0.144	0.047	0.008	0.001
$P(X \leq x)$	0.049	0.234	0.532	0.800	0.944	0.991	0.999	1.000

and hence the "traditional" 90^{th} quantile of X is 4.
We determine the "mid–quantile" as follows:
$v_5 = 4, v_6 = 5$, $p_5 = 0.144$, $p_6 = 0.047$.
Hence $\pi_4 = 0.800 + 0.144/2 = 0.8720$, $\pi_5 = 0.944 + 0.047/2 = 0.9675$.
Note that $0.9 = 0.707\pi_4 + (1 - 0.707)\pi_5$, hence:

$$Q(0.9) = F^{-1}_{\text{mid}}(0.9) = 0.707 v_4 + (1 - 0.707) v_5 = 3.213$$

2.2 Example: Simulated Poisson Data

The 25 data of Table 1 are a random draw from a random variable W that is believed to follow a Poisson distribution. It is wished to test this belief.

TABLE 1.

0	1	2	3
16	4	4	1

It is estimated, using the adaptive mixture estimator of Wilson and Einbeck (2016), that the mean of W is $\mu = 1.171$, and hence under the null (Poisson) model $P(W = 0) = \exp(-1.171) = 0.310$. Hence the observed number of zeros in random samples of size 25 drawn from W will be $Bin(25, 0.310)$ distributed. The "traditional" 2.5^{th} and 97.5^{th} quantiles of such a distribution are 7 and 16 respectively, and hence a 95% fluctuation interval for the number of observed zeros under the Poisson distribution is $[7, 16]$ indicating non-rejection of the Poisson model at a level of significance of $\alpha = 0.05$, consistent with the traditional p-value of 0.064, but inconsistent with the mid p-value of 0.045. The 95% fluctuation interval based upon the mid quantiles is however $[6.52, 15.57]$, consistent with the mid p-value. These results are summarised in Table 2.2.

TABLE 2. $n = 25$, H_0:Poisson

$n_0 = 16$	p-value	$95\%FI$
traditional	0.064	$[7, 16]$
mid	0.045	$[6.52, 15.57]$

2.3 Example: Trajan Data Revisited

Here we re-compute the 80% fluctuation interval for the negative binomial model fitted to the Trajan data of Section 1.1 using the mid-distribution quantiles defined above. (Recall, here we are testing for strict zero-inflation, and thus the upper bound of the fluctuation interval serves as a test statistic for a one-sided test). We find that $\pi_{47} = 0.073$ and $\pi_{48} = 0.101$, thus $0.1 = 0.069\pi_{47} + (1 - 0.069)\pi_{48}$ and hence $Q(0.1) = (0.069 \times 47) + ((1 - 0.069) \times 48) = 47.931$. Similarly $\pi_{63} = 0.882$ and $\pi_{64} = 0.913$, thus $0.9 = 0.419\pi_{63} + (1 - 0.419)\pi_{64}$ and hence $Q(0.9) = (0.419 \times 63) + ((1 - 0.419) \times 64) = 63.581$. Thus, using the mid-distribution quantile we obtain a 80% fluctuation interval of $(47.931, 63.581)$, and hence $n_0 = 64$ is exterior to the confidence interval, and we reject the negative-binomial model in favour of the zero-inflated negative binomial model under both criteria. These results are summarised in Table 2.3.

TABLE 3.

$n_0 = 64$	p-value	90% FI	80% FI
traditional	0.1010	$[47, 66]$	$[49, 64]$
mid	0.0871	$[46.902, 65.679]$	$[47.931, 63.581]$

3 Conclusion

Decisions based upon mid-distribution quantiles as defined above will agree with those based upon mid p-values unless $p < p_1/2$ or $p > \pi_d$. With respect to the test proposed in Wilson and Einbeck (2015, 2016) these exceptions correspond to the observed data either containing no zeros, or consisting entirely of zeros, and hence the adoption of quantiles based upon mid-distribution functions results in fluctuation intervals that nearly entirely removes discrepancies that may sometimes occur between decisions based upon fluctuation intervals and mid p–values. Given that the power and attainment rates of the test when based upon mid p-values are excellent, such alignment is desirable. The adoption of such quantiles is straightforward. In this paper we only discuss the use of mid-distribution quantiles in relation to the test of Wilson and Einbeck (2015, 2016), but their application to other tests with discrete test statistics is worthy of investigation.

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. R package version 1.2. https://CRAN.R-project.org/package=poibin

Ma, Y., Genton M. and Parzen, E. (2011) Asymptotic properties of sample quantiles of discrete distributions. *Annals of the Institute of Statistical Mathematics* **63**, 227 – 243.

Ridout and Demétrio (1992) Generalized Linear Models for Positive Count Data. *Revista de matematica e estatistica* **10**, 139 – 148.

Ridout, M.S., Demétrio C.B. and Hinde, J. (1998) Models for count data with many zeros. In *Proceedings of the XIXth international biometric conference* **19**, 179 – 192).

Wilson, P. and Einbeck, J. (2015). A simple and intuitive test for number–inflation or number–deflation. In: Wagner, H. and Friedl, H. (Eds). Proc's of the 30th IWSM, Linz, Austria, Vol 2, pages 299–302.

Wilson, P. and Einbeck, J. (2016). On statistical testing and mean parameter estimation for zero–modification in count data regression. In: Dupuy, J. and Josse, J. (Eds). Proc's of the 31st IWSM, Rennes, France, Vol 1, pages 325 – 330.