A Note on Inconsistency of the Score Test

Sumathi K
Department of Statistics
Mangalore University, Mangalagangotri
Konaje, India
chaitra_udipi@yahoo.com

Aruna Rao K
Department of Statistics,
Mangalore University, Mangalagangotri
Konaje, India

Abstract

The score test proposed by Rao (1947) has been widely used in the recent years for data analysis and model building because of its simplicity. However, at the time of its computation, it has been found that the value of the score test statistic becomes negative. Freedman (2007) discussed some of the theoretical reasons for this inconsistency of the score test and observed that the test was inconsistent when the observed Fisher information matrix was used rather than the expected Fisher information matrix. The present paper is an attempt to demonstrate the inconsistency of the score test in terms of the power function. The paper further dispels the doubt regarding the use of score test.

Keywords: Model building, Consistency, Score test, Fisher information matrix, Power function.

1. Introduction

The score test proposed by Rao (1947) has been widely used in the recent years for data analysis and model building. This is because of the simplicity of the test in checking the adequacy of a model when nested models are used. However, at the time of computation, it has been found that the value of the score test statistic becomes negative. For a long time, the reason for this phenomenon was not known. A good discussion can be found in Morgan et al. (2007), Verbeke and Molenberghs (2007) and Freedman (2007).

Morgan et al. (2007) provided an example where the score test statistic attains a negative value. The example relates to the case of a zero-inflated Poisson distribution. The score test statistic became negative when the model was a poor fit to the data under the null hypothesis. Further investigations based on this were done by Verbeke and Molenberghs (2007) and Freedman (2007).

Freedman (2007) discussed some of the theoretical reasons for the inconsistency of the score test. He observed that the test was inconsistent when the observed Fisher information matrix was used rather than the expected Fisher information matrix and that the observed Fisher information matrix generated negative variance estimates at the maximum likelihood estimate (MLE) of the parameter of the distribution under the null hypothesis. He observed that the test

statistic can also become inconsistent when the expected likelihood equation has spurious roots.

Verbeke and Molenberghs (2007) mentioned that the problems associated with the score test can be considered along the following four dimensions.

1. Unconstrained versus constrained parameter spaces.
2. Use of observed versus expected Fisher information.
3. Rejection probabilities carried out under the null hypothesis, a correctly specified alternative, or a mis-specified alternative.
4. Asymptotic versus small sample behaviour.

The first and the second dimensions have been discussed by the authors.

When we were comparing the likelihood ratio, the Wald and the score tests for testing the inflation parameter $p=p_0$ of the inflated Poisson distribution, the graphs of the power functions of the score test and its perturbed versions exhibited fluctuations (decreasing or randomly decreasing and increasing after a certain stage) rather than increasing on either sides of the specified value of the inflate parameter. For details, see Sumathi and Rao (2010).

In the present paper, an attempt has been made to study the causes for the fluctuations in connection with the small sample power computation of the score test for a zero-inflated Poisson distribution. The numerical simulations show the inconsistency of the score test in small sample power computations in the form of fluctuations in the power curve. The inconsistency is because the test statistic becomes negative. As the frequency of the test statistic attaining a negative value increases, the power of the test decreases, which in turn causes fluctuations in the power function.

It has been found that the usual score test statistic is more consistent than its perturbed version obtained by using the unrestricted MLEs in the expected or the observed Fisher information matrix. This finding contradicts the conclusion of Freedman (2007) where the asymptotic properties were of major concern.

The remaining part of the paper is organized as follows. Section 2 describes the score tests for testing the inflate parameter $p=p_0$ of an inflated Poisson distribution. Section 3 discusses the small sample performance of the power functions. The paper concludes with a discussion in section 4.

2. Score tests

Rao (1947) proposed the score test. The advantage of this test is that the computation of the MLEs is not required when the hypothesis is simple. Details of the score test are available in Rao (1973), Cox and Hinkley (1974) and Severini (2000). In this section, the score test statistic for testing the inflate parameter $p=p_0$ of an inflated Poisson distribution (inflated at zero) has been described.

Consider a random variable Y which follows an inflated Poisson distribution (inflated at zero) with parameters p and λ. The probability mass function of Y is given by

$$P[Y=y]=\begin{cases} p+(1-p)\exp(-\lambda), \text{when } y=0 \\ (1-p)\exp(-\lambda)\dfrac{\lambda^y}{y!}, \text{when } y=1,2,... \end{cases} \quad (1)$$

Let $y_1, y_2, \ldots, y_n$ be a random sample of size n from the zero inflated Poisson distribution given by (1). The likelihood function based on the n observations is given by $L(p,\lambda;Y)=\left(p+(1-p)e^{-\lambda}\right)^{n_0}\prod_{y_i\neq 0}\dfrac{(1-p)e^{-\lambda}\lambda^{y_i}}{y_i!}$ (2)

where $Y = (y_1, y_2, \ldots, y_n)$ and n_0 is the number of observations that are zeroes.

The maximum likelihood (ML) equations for the estimation of p and λ after simplification are given by

$$n\left(1-e^{-\lambda}\right)p-\left(n_0-ne^{-\lambda}\right)=0 \quad (3)$$

$$\lambda-\left(\frac{\sum_{y_i\neq 0} y_i}{n-n_0}\right)\left(1-e^{-\lambda}\right)=0 \quad (4)$$

respectively. It may be noted that equation (4) is independent of the parameter p and is the ML equation for λ from a truncated Poisson distribution (truncated at zero) based on $n-n_0$ observations. Since there does not exist a closed form solution to equation (4), the ML estimate of λ can be obtained using a numerical method. In the current paper, the method used for simulation discussed in section 3, is the bisection method (Sastry (1994)).

Let the ML estimator of λ be denoted by $\hat{\lambda}$. The MLE of p is obtained by substituting the value of $\hat{\lambda}$ for λ in equation (3). For carrying out inference on p when $p=p_0$ (specified), the estimation of the restricted MLE of λ denoted by $\hat{\hat{\lambda}}$ is necessary. $\hat{\hat{\lambda}}$ is the solution of the restricted ML equation given by

$$\frac{1}{\lambda}\sum_{y_i\neq 0} y_i-(n-n_0)-\frac{n_0(1-p_0)e^{-\lambda}}{p_0+(1-p_0)e^{-\lambda}}=0 \quad (5)$$

The expected Fisher Information matrix for p and λ is given by

$$I = \begin{bmatrix} \frac{n\left(1 - e^{-\lambda}\right)}{(1-p)\left(p + (1-p)e^{-\lambda}\right)} & \frac{-ne^{-\lambda}}{p + (1-p)e^{-\lambda}} \\ \frac{-ne^{-\lambda}}{p + (1-p)e^{-\lambda}} & \frac{-n(1-p)\left(\lambda pe^{-\lambda} - p - e^{-\lambda} + pe^{-\lambda}\right)}{\lambda\left(p + (1-p)e^{-\lambda}\right)} \end{bmatrix} \quad (6)$$

For details, see Bhattacharya, et.al. (2008).

The score test statistic for testing the hypothesis H_0: p = p_0 is given by

$$W_S = U'I^{-1}U \quad (7)$$

where $U = \left[\frac{\partial}{\partial p}\log L; \frac{\partial}{\partial \lambda}\log L\right]$ (8)

is the score vector with $\frac{\partial}{\partial p}\log L = \frac{n_0\left(1-e^{-\lambda}\right)}{p+(1-p)e^{-\lambda}} - \frac{n-n_0}{1-p}$

$$\frac{\partial}{\partial \lambda}\log L = \frac{-n_0(1-p)e^{-\lambda}}{p+(1-p)e^{-\lambda}} - n - n_0 + \frac{\sum_{y_i \neq 0} y_i}{\lambda},$$

U' denotes the transpose of U and *I* is given by (6). The parameters p and λ in *U* and I are replaced by p_0 and $\hat{\hat{\lambda}}$.

Cox and Hinkley (1974) and Kale (1999) have suggested that the unrestricted MLEs can be used in place of the restricted MLEs in the Fisher information matrix. Details can also be found in Morgan et al. (2007), Verbeke and Molenberghs (2007) and Freedman (2007). Thus using the suggestions of Cox and Hinkley (1974) and Kale (1999), two perturbed versions of the score test have been developed. The test statistics have been denoted as W_{s_1} and W_{s_2} and have the same forms as W_S but the parameters p and λ in I are replaced by p_0 and $\hat{\lambda}$ for W_{s_1} and by $\hat{p}$ and $\hat{\lambda}$ for W_{s_2} respectively.

It has been noted by several authors (Cox and Snell (1989), Davison (2003), Morgan (2000), Pawitwan (2001), Morgan et al. (2007), Verbeke and Molenberghs (2007) and Freedman (2007)) that observed Fisher information matrix may be used in place of the expected Fisher information matrix in a score test statistic. For the zero inflated Poisson distribution considered in the present paper, the observed Fisher information matrix is

$$O = \begin{bmatrix} \frac{n_0\left(1-e^{-\lambda}\right)^2}{\left(p+(1-p)e^{-\lambda}\right)^2} + \frac{n-n_0}{(1-p)^2} & \frac{-n_0e^{-\lambda}}{p+(1-p)e^{-\lambda}}\left(1+\frac{\left(1-e^{-\lambda}\right)(1-p)}{\left(p+(1-p)e^{-\lambda}\right)}\right) \\ \frac{-n_0e^{-\lambda}}{p+(1-p)e^{-\lambda}}\left(1+\frac{\left(1-e^{-\lambda}\right)(1-p)}{\left(p+(1-p)e^{-\lambda}\right)}\right) & \frac{-n_0(1-p)e^{-\lambda}}{\left(p+(1-p)e^{-\lambda}\right)}\left(1-\frac{(1-p)e^{-\lambda}}{\left(p+(1-p)e^{-\lambda}\right)}\right) + \frac{\sum_{i\neq 0} y_i}{\lambda^2} \end{bmatrix} \quad (9)$$

3. Behaviour of the Power functions of the score tests

A simulation experiment was done to assess the behaviour of the power functions of the three test statistics viz., W_S, W_{s_1} and W_{s_2}. The simulation configurations used were as follows:

Level of significance α=0.05

Sample size *n* = 20, 40, 80, 100, 200, 400

p=0.1, 0.3, 0.5, 0.7, 0.9

λ =1, 3, 5, 7, 9

The number of simulations = 10000

The power functions of the score test and its perturbed versions were estimated using simulations. For details regarding simulations, refer Sumathi and Rao (2010). There were 25 configurations corresponding to each value of *n*.

From the simulations it was evident that for a sample of size n=20, the power functions of all the three test statistics showed fluctuations on the left of p_0 when testing for p_0=0.1, for all values of the mean parameter λ given above. When testing for p_0=0.3, the power function of the perturbed version of the score test statistic viz., W_{s_2}, where $\hat{p}$ and $\hat{\lambda}$ were used in Fisher information matrix, showed fluctuations on the left of p_0 for all values of λ, while the power functions of the score test statistic W_S and one of its perturbed versions viz., W_{s_1} exhibited fluctuations on the left of p_0 only when λ =1. The power function of all the three tests were normal when testing for p_0=0.5. When testing for p_0=0.7, the power function of the test statistic W_{s_2} exhibited fluctuations on either sides of p_0 for λ=1, 7 and 9, while the behaviour of the power functions of W_S and W_{s_1} was normal for all values of λ. Finally, when testing for p_0=0.9, the power functions of all the three test statistics showed fluctuations on the right of p_0 for all values of the mean parameter λ given above.

When the sample size was increased to 40, the behaviour of the power functions of the test statistics were the same as that when the sample size was 20. When a sample of size 80 was considered, the power function of the usual score test statistic W_S did not show any fluctuations, while the power functions of the two perturbed versions of the score test exhibited fluctuations on the left of p_0 when testing for p_0=0.1 and on the right of p_0 when testing for p_0=0.9. The same behaviour was observed when n=100, 200. Also, when a large sample of size 400 was considered, fluctuations were seen in the power functions of the two perturbed versions of the score test while testing for p_0=0.5, 0.7 and 0.9 corresponding to λ = 1 and 3. Similar type of fluctuations were seen in the power functions of the three score tests when the observed Fisher information matrix O given by (9) was used instead of the expected Fisher information matrix I.

Tables 1 and 2 indicate the tests whose power functions exhibit fluctuations for various combinations of p_0 and λ when the sample sizes n=20 and 40 were considered.

Table 1: The tests whose power functions showed fluctuations for combinations of p_0 and λ for sample sizes n=20 and 40 when the expected Fisher information matrix was used.

Sample size n	p_0	λ				
		1	3	5	7	9
20	0.1	W_S, W_{s_1}, W_{s_2}	W_S, W_{s_1}, W_{s_2}	W_S, W_{s_1}, W_{s_2}	W_S, W_{s_1}, W_{s_2}	W_S, W_{s_1}, W_{s_2}
	0.3	W_S, W_{s_1}, W_{s_2}	W_{s_2}	W_{s_2}	W_{s_2}	W_{s_2}
	0.5	-	-	-	-	-
	0.7	W_{s_2}	-	-	W_{s_2}	W_{s_2}
	0.9	W_S, W_{s_1}, W_{s_2}	W_S, W_{s_1}, W_{s_2}	W_S, W_{s_1}, W_{s_2}	W_S, W_{s_1}, W_{s_2}	W_S, W_{s_1}, W_{s_2}
40	0.1	W_S, W_{s_1}, W_{s_2}	W_S, W_{s_1}, W_{s_2}	W_S, W_{s_1}, W_{s_2}	W_S, W_{s_1}, W_{s_2}	W_S, W_{s_1}, W_{s_2}
	0.3	W_S, W_{s_1}, W_{s_2}	-	W_{s_2}	W_{s_2}	W_{s_2}
	0.5	-	W_{s_1}, W_{s_2}	W_{s_1}, W_{s_2}	W_{s_1}, W_{s_2}	W_{s_2}
	0.7	-	W_{s_2}	W_{s_2}	W_{s_2}	W_{s_2}
	0.9	W_S, W_{s_1}, W_{s_2}	W_S, W_{s_1}, W_{s_2}	W_S	W_S, W_{s_1}, W_{s_2}	W_S, W_{s_1}, W_{s_2}

Table 2: The tests whose power functions showed fluctuations for combinations of p_0 and λ for sample sizes n=20 and 40 when the observed Fisher information matrix was used.

Sample size n	p_0	λ				
		1	3	5	7	9
20	0.1	W_S, W_{s_1}, W_{s_2}	W_S, W_{s_1}, W_{s_2}	W_S, W_{s_1}, W_{s_2}	W_S, W_{s_1}, W_{s_2}	W_S, W_{s_1}, W_{s_2}
	0.3	W_S, W_{s_1}, W_{s_2}	W_{s_2}	W_{s_2}	W_{s_2}	W_{s_2}
	0.5	-	-	W_{s_2}	W_{s_2}	W_{s_2}
	0.7	W_S, W_{s_1}, W_{s_2}	W_{s_2}	W_{s_2}	W_{s_2}	W_{s_2}
	0.9	W_S, W_{s_1}, W_{s_2}	W_S, W_{s_1}, W_{s_2}	W_S, W_{s_1}, W_{s_2}	W_S, W_{s_1}, W_{s_2}	W_S, W_{s_1}, W_{s_2}
40	0.1	W_S, W_{s_1}, W_{s_2}	W_S, W_{s_1}, W_{s_2}	W_S, W_{s_1}, W_{s_2}	W_S, W_{s_1}, W_{s_2}	W_S, W_{s_1}, W_{s_2}
	0.3	W_{s_1}, W_{s_2}	-	W_{s_2}	W_{s_2}	W_{s_2}
	0.5	W_S, W_{s_1}, W_{s_2}	W_S	W_{s_2}	W_{s_2}	W_{s_2}
	0.7	-	W_{s_2}	W_{s_2}	W_{s_2}	W_{s_2}
	0.9	W_S, W_{s_1}, W_{s_2}	W_S, W_{s_1}, W_{s_2}	W_S, W_{s_1}, W_{s_2}	W_S, W_{s_1}, W_{s_2}	W_S, W_{s_1}, W_{s_2}

Below are the graphs showing the power functions of the score test statistics for a sample of size 40 while testing for p_0=0.1 and p_0=0.3 when λ=3 and 5. Figures 1, 2, 3 and 4 correspond to the case when the expected Fisher information matrix was used while figures 5, 6, 7 and 8 correspond to that when observed Fisher information matrix was used. The points on the x-axis of the graphs correspond to that in the neighborhood of p_0.

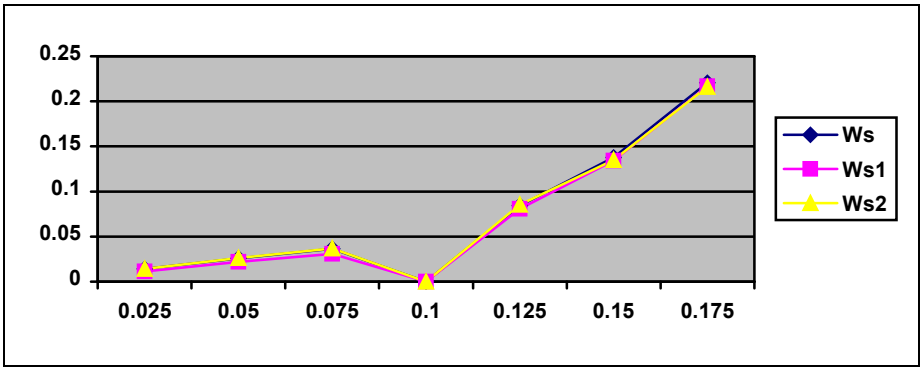

Figure 1: Power functions of the three tests using expected Fisher information when n=40, p_0=0.1, λ=3

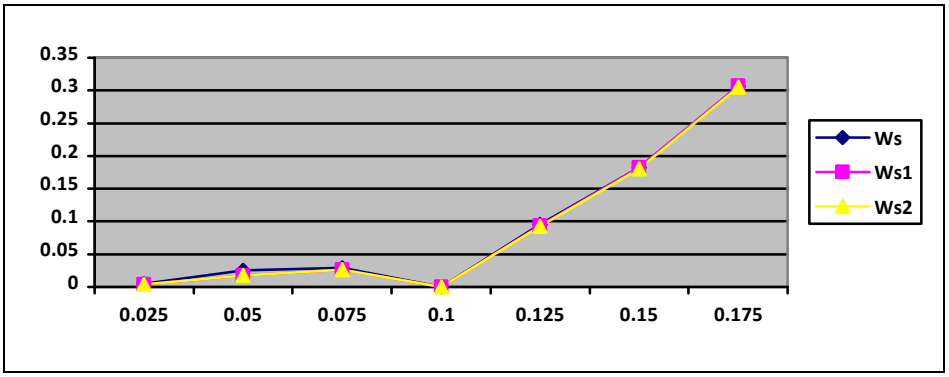

Figure 2: Power functions of the three tests using expected Fisher information when n=40, p_0=0.1, λ=5

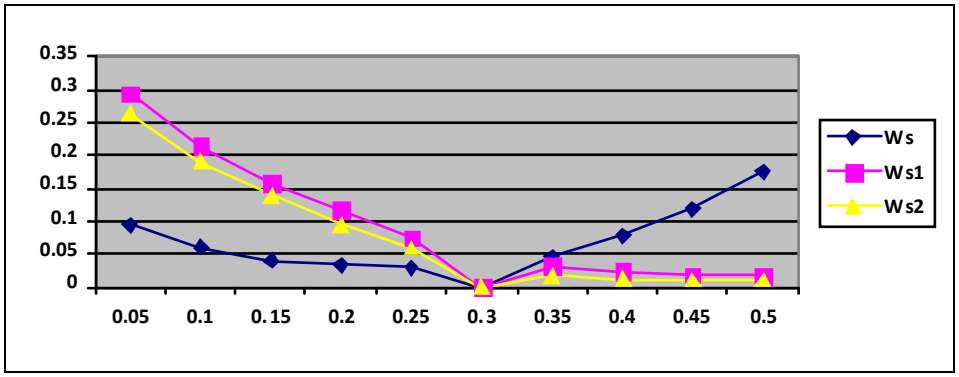

Figure 3: Power functions of the three tests using expected Fisher information when n=40, p_0=0.3, λ=3

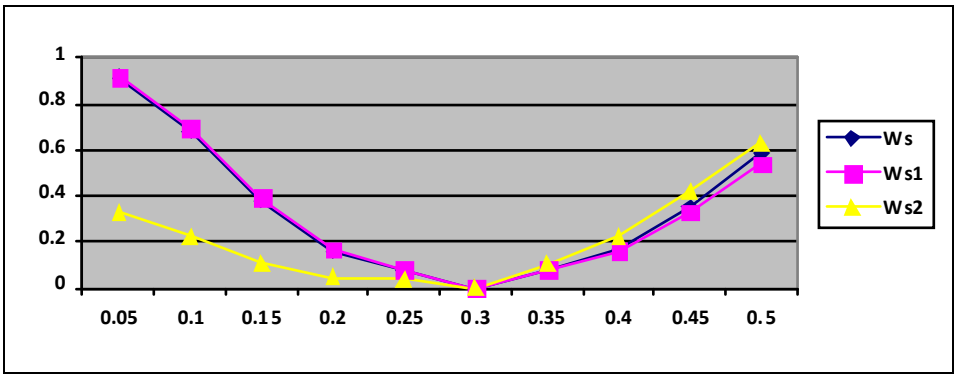

Figure 4: Power functions of the three tests using expected Fisher information when n=40, p_0=0.3, λ=5

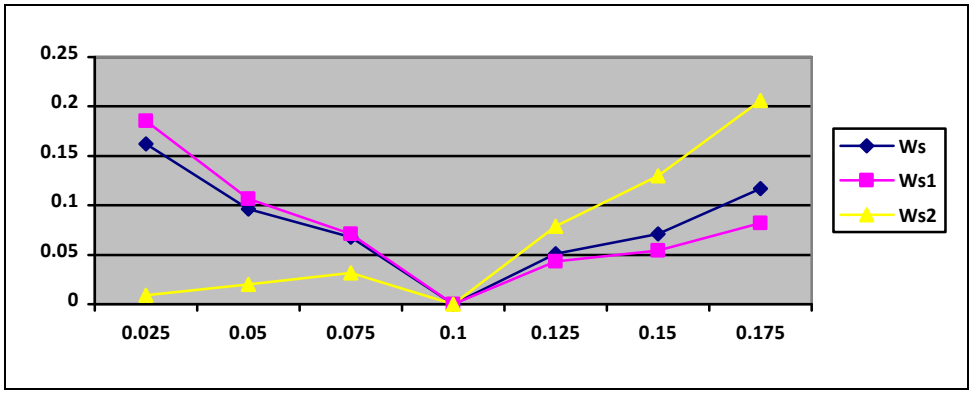

Figure 5: Power functions of the three tests using observed Fisher information when n=40, p_0=0.1, λ=3

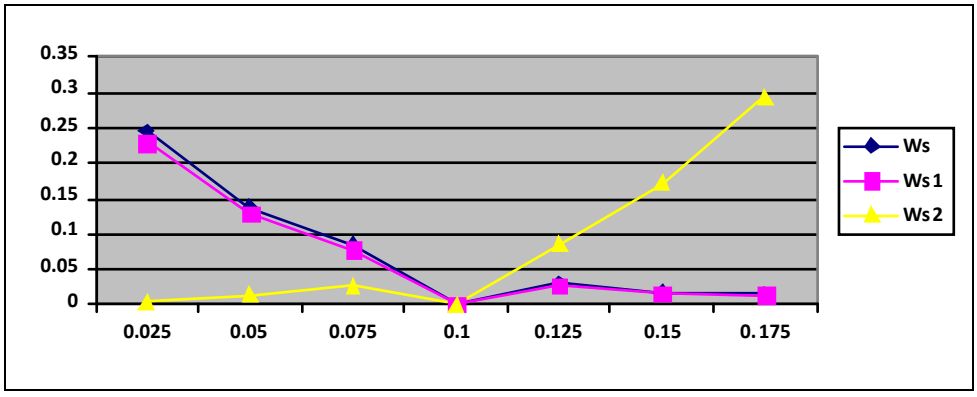

Figure 6: Power functions of the three tests using observed Fisher information when n=40, p_0=0.1, λ=5

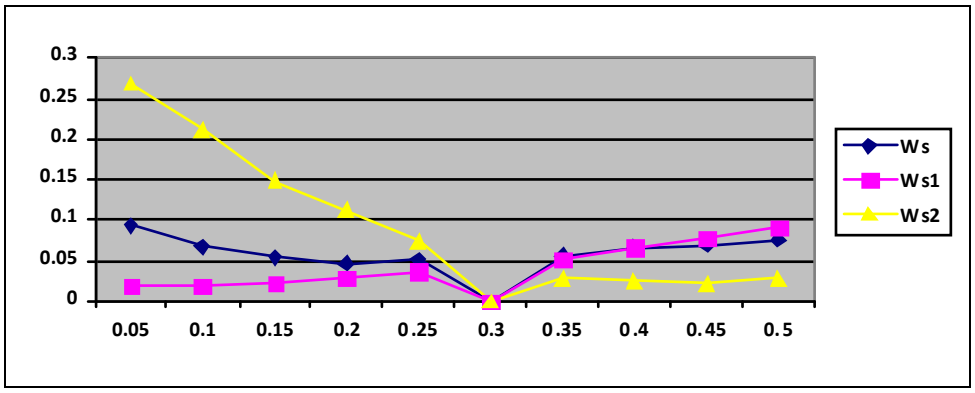

Figure 7: Power functions of the three tests using observed Fisher information when n=40, p_0=0.3, λ=3

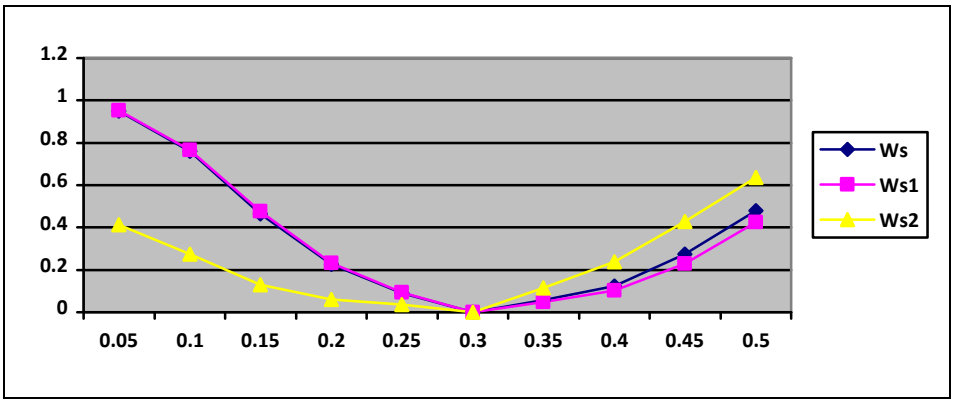

Figure 8: Power functions of the three tests using observed Fisher information when n=40, p_0=0.3, λ=5

The following table 3 summarises the number of times the three versions of the score test statistics exhibit fluctuations in the power functions for the twenty five configurations of the parameters p_0 and λ, for various sample sizes, when the expected and the observed Fisher information matrices were used.

Table 3

Sample size n	Information Matrix	Test Statistics		
		W_S	W_{s_1}	W_{s_2}
20	Expected	11	11	18
	Observed	12	12	23
40	Expected	11	13	21
	Observed	12	12	22
80	Expected	00	06	18
	Observed	02	08	18
100	Expected	01	09	19
	Observed	03	11	19
200	Expected	01	08	17
	Observed	02	07	18
400	Expected	01	04	07
	Observed	02	05	08

From the tables and graphs of the power functions of the test statistics, the following conclusions emerge.

1. The fluctuations either on the right or on the left of p_0 are seen in the powers of all the three test statistics for sample sizes n = 20 and n = 40 irrespective of whether the expected Fisher information matrix I or the observed Fisher information matrix O has been used.
2. The powers of the test statistics are very low (<0.1) in the neighbourhood of p_0. This is because the test statistics attain negative value at these points.
3. When the unrestricted ML estimators $\hat{p}$ and $\hat{\lambda}$ are used in the Fisher information matrix, may it be expected or observed, the power function of the perturbed version of the score test statistic viz., W_{s_2}, exhibits fluctuations even in situations where the usual score test statistic W_S and its other perturbed version W_{s_1} are normal.
4. The fluctuations generally occur for small samples when testing for small or large values of the inflate parameter p, especially when the power is estimated at the values of p nearer to the boundary of the parameter space for p viz., (0, 1).
5. The use of the unrestricted ML estimators $\hat{p}$ and $\hat{\lambda}$ in the observed Fisher information matrix does not improve the situation of inconsistency of the score test. This conclusion differs from that of Freedman (2007).
6. From table 3, it follows that the usual score test is the best compared to its perturbed versions, irrespective of whether the expected Fisher information matrix or the observed Fisher information matrix is used.
7. Also, it is evident from table 3, that the fluctuations are seen when either expected or observed Fisher information matrices are used, but it is more when observed information is used. Hence, the performance of the tests are better when expected Fisher information matrix is used.

4. Discussion and conclusion:

In this paper, the effect of inconsistency of the score test in estimating the power function of the test and its perturbed versions for small samples has been demonstrated. We observed that the power functions show fluctuations when the test statistics become negative and lead to the acceptance of the null hypothesis. This inconsistency occurs for small sample sizes and also when testing for small or large values of the inflate parameter p. However, if we look into the overall pattern of the power functions, the usual score test W_S and its perturbed version W_{s_1} do not exhibit fluctuations when the alternative hypothesis is at a moderate distance from the null hypothesis in either directions. This is an indication that one can safely use the score test for the analysis of count data. Also, it is advisable to use the expected Fisher information matrix rather than the observed

information matrix. Moreover, if one encounters a negative value of the score test statistic, then the score test should not be used for any inferential aspect. Instead, the user can think of either using the likelihood ratio test or the Wald test.

The usual score test does not have this inconsistent behaviour when the sample size is 80 or above except for p_0=0.1 and λ=1, while the two perturbed versions of the score test exhibit inconsistency even for large samples of size 200 and 400. Fluctuations are seen in the power functions of the test statistics when using either expected or observed Fisher information matrix. The inconsistency of the score test is a problem only when the null hypothesis is very much mis-specified and our analysis strengthens the conclusions of Morgan, et. al. (2007). Thus, from the present study, we recommend the conventional score test W_S for inference involving inflated distributions when the sample size n is at least 80.

Acknowledgement

The authors are grateful to the anonymous referees whose valuable comments led to a substantial improvement of the paper. The paper is dedicated to Professor C. R. Rao, the proposer of the score test.

References

1. Bhattacharya, A., Clarke, B.S., and Datta, G.S. (2008). A Bayesian test for excess zeros in a zero-inflated power series distribution. *IMS collections, Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K Sen*, 1, 89-104.
2. Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics*. Chapman and Hall: London.
3. Cox, D.R. and Snell, E. J. (1989). *Analysis of Binary Data*. Second Edition. Chapman and Hall: London.
4. Davison, A. C. (2003). *Statistical Models*. CUP: New York.
5. Freedman, D. A. (2007). How Can the Score Test Be Inconsistent? *The American Statistician*, Vol. 61, No. 4, 291-295.
6. Kale, B. K. (1999). *A First Course on Parametric Inference*. Narosa Publishing House, New Delhi.
7. Morgan, B. J. T. (2000). *Applied Statistical Modelling*. Arnold: London.
8. Morgan, B. J. T., Palmer, K. J. and Ridout, M. S. (2007). Negative Score Test Statistic. *The American Statistician*, Vol. 61, No. 4, 285-288.
9. Pawitwan, Y. (2001). *In All Likelihood: Statistical modeling and Inference using Likelihood.* Clarendon Press: Oxford.
10. Rao, C. R. (1973). *Linear Statistical Inference and its Applications*. Second Edition. John Wiley & Sons, New York, Inc.

11. Rao, C. R. (1947). Large Sample tests of Statistical Hypotheses concerning several parameters with applications to problems of estimation. *Proceedings of the Cambridge Philosophical Society*, 44, 50-57.

12. Sastry, S.S. (1994). Introductory Methods of Numerical Analysis. *Prentice Hall of India, New Delhi.*

13. Severini, T.A. (2000). *Likelihood Methods in Statistics*. Oxford University Press Inc, New York.

14. Sumathi, K. and Rao, A.K (2010). Tests for Cured Proportion for Recurrent Event Data. *Optimisation and Statistics.*

15. Verbeke, G and Molenberghs, G. (2007). What Can Go Wrong With the Score Test? *The American Statistician*, Vol. 61, No. 4, 289-290.