Malaysian Journal of Science **26** (2): 127 – 133 (2007)

A Test for the Mean in the Zero-Inflated Poisson Distribution

Nik Ahmad Kamal* and **Pooi Ah Hin**

Institute of Mathematical Sciences, Faculty of Science, University of Malaya 50603 Kuala Lumpur, Malaysia
*mymailbox@um.edu.my (corresponding author)

Received 18th October 2005, accepted 20th June 2007

ABSTRACT In analyzing Poisson-count data, sometimes a lot of zeros are observed. When there are too many zeros, a zero-inflated Poisson distribution can be a more suitable model to use. A test for the mean θ in the imperfect state can be obtained by using the conditional maximum likelihood estimator $\tilde{\theta}$ of θ and the asymptotic variance of $\tilde{\theta}$. For moderately small sample size, the probability of the rejection region under the null hypothesis is found to have only a small variation around the targeted value as the value of the nuisance parameter varies.

ABSTRAK Dalam menganalisis data bilangan Poisson, sering kali banyak sifar dicerap. Apabila terdapat banyak sifar, taburan Poisson inflasi sifar boleh dijadikan model yang lebih sesuai digunakan. Dalam kes ini, ujian bagi min θ boleh di dapati dengan menggunakan penganggar kebolehjadian maksimum bersyarat $\tilde{\theta}$ bagi θ bersama varians asimptot bagi $\tilde{\theta}$. Bagi saiz sampel yang agak kecil kebarangkalian kawasan penolakan di bawah hipotesis nol didapati mempunyai variasi yang kecil disekeliling nilai sebenar, apabila nilai parameter kacauganggu diubah.

(Zero-inflated Poisson distribution, conditional likelihood function, asymptotic variance, reject region)

INTRODUCTION

In many manufacturing processes, when the process is in its perfect state, defects are rarely observed. But, when there is an equipment or process problem, defects may occur according to a discrete distribution. When this distribution is Poisson (θ), the number of defects X_i at the *i*th inspection of the process is given by

$$\Pr(X_i = x_i) = \begin{cases} \phi + (1-\phi)e^{-\theta} & \text{for } x_i = 0 \\ (1-\phi)\dfrac{\theta^{x_i} e^{-\theta}}{x_i!} & \text{for } x_i = 1, 2, 3, \ldots \end{cases}$$

where $\theta>0$ and $0\leq\phi<1$ with $(1-\phi)$ as probability that the system is in an imperfect state. The distribution is called the zero-inflated Poisson distribution (ZIP), (Cohen [1], Johnson and Kotz [7]). Yip [9] used ZIP to model the number of insects per leaf. Heilbron [5] proposed zero-altered Poisson and negative binomial regression models and applied them to study high-risk human behavior. Lambert [8] used ZIP regression model to analyze data on soldering defects on printed wiring boards. Gupta, Gupta and Tripathi [3, 4] proposed zero-adjusted discrete distributions.

It is interesting to note that the conditional density of X_i given X_i is non zero, is independent of ϕ,

$$\Pr(X_i = x_i \mid A_i = a_i) = \left\{ \frac{\theta^{x_i} e^{-\theta}}{x_i!} / (1 - e^{-\theta}) \right\}^{a_i} \tag{1}$$

where $a_i = \begin{cases} 0 & \text{if } x_i = 0 \\ 1 & \text{if } x_i > 0 \end{cases}$

and the density of the number $A = \sum_{i=1}^{n} A_i$ of non zero X_i has the following binomial distribution

$$A \overset{d}{=} Bin\ [n, (1-\phi)(1-e^{-\theta})] \tag{2}$$

where n is the number of inspections.

The maximum likelihood estimator $(\tilde{\theta}, \tilde{\phi})$ of (θ, ϕ) based on the conditional density (1) is then given by

$$\tilde{x} = \frac{\tilde{\theta}}{1-e^{-\tilde{\theta}}} \text{ and } (1-\tilde{\phi})(1-e^{-\tilde{\theta}}) = \frac{a}{n}$$

provided that $(1-\frac{a}{n}) > e^{-\tilde{\theta}}$, where

$a = \sum_{i=1}^{n} a_i$ and $\tilde{x} = \sum_{i=1}^{n} a_i x_i / a$ (see David and Johnson [2], Irwin and Goen [6] and Yip [9]).

The asymptotic variance of $\tilde{\theta}$ is given by

$$Var(\tilde{\theta}) = \left\{ \frac{a}{1-e^{-\theta}} \left(\frac{1}{\theta} - \frac{e^{-\theta}}{1-e^{-\theta}} \right) \right\}^{-1} \tag{3}$$

Yip [9] selected the sample size n=500 and performed 100 simulation runs to investigate the performance of $\tilde{\theta}$ and $\tilde{\phi}$, and the estimators $\hat{\theta}$ and $\hat{\phi}$ based on unconditional likelihood function, by examining the bias and standard error of the series of estimates obtained. The simulation results show that when n is as large as 500, the performance of $(\tilde{\theta}, \tilde{\phi})$ is about the same as that of $(\hat{\theta}, \hat{\phi})$. But it is much easier to compute $(\tilde{\theta}, \tilde{\phi})$.

TESTING THE HYPOTHESIS REGARDING θ

When the system is in an imperfect state, the parameter θ determines the mean as well as the entire distribution of the number of defects. The parameter θ is therefore of great interest to us.

In this paper, we consider the problem of testing the mean θ when the process is in the imperfect state in the presence of the nuisance parameter ϕ. We consider the plausibility of using the test statistics

$$t = \frac{\tilde{\theta} - \theta_o}{\sqrt{\hat{V}ar\ (\tilde{\theta})}} \tag{4}$$

for testing the null hypothesis H_o: $\theta=\theta_o$ against the alternative hypothesis H_A: $\theta\neq\theta_o$. In (4), $\hat{Var}(\tilde{\theta})$ is the estimated asymptotic variance obtained by replacing θ in (3) with $\tilde{\theta}$.

We may choose a rejection region given by

$$R = \{\underline{x} : |t| > 1.96\} \tag{5}$$

for testing H_o: $\theta=\theta_o$.

When ϕ is small and n is large, t will have approximately a normal distribution and the level of the test based on R will be about 0.05. But when ϕ is large or n is small, the probability of R under H_o would deviate from 0.05 and the extent of deviation would depend on the value of ϕ.

Presently for given values of θ_o and n, we find the probability $Pr(|t|>1.96|\ \theta=\theta_o,\phi)$ of the rejection region under H_o. We may next try to find a value ϕ^* of ϕ such that the probability of the rejection region under H_o will not deviate from 0.05 by more than a small value δ (for example, δ=0.01, 0.02 or 0.03) for $\phi\in[0, \phi^*]$. The interpretation of ϕ^* is that in the situation in which we would not expect ϕ to be as large as ϕ^* the test would have a significance level which will not deviate from 0.05 by more than δ. If the value of ϕ could be larger than ϕ^* , then we should not use the test as the type I error then could deviate from 0.05 by more than δ.

PROBABILITY OF THE REJECTION REGION UNDER H_O

In this section, the probability $\pi(\theta_o,\phi) = \mathbf{Pr}(R|\theta = \theta_o,\phi)$ of the rejection region under the null hypothesis shall be derived. Following (2) we note that the number A of the X_i which are nonzero has the following binomial distribution:

$$f(a;\theta,\phi) = \binom{n}{a}(1-\phi)^a(1-e^{-\theta})^a[\phi + e^{-\theta}(1-\phi)]^{n-a}.$$

For $a \geq 1$, the moment generating function of the total number X of defects conditional on $A=a$ is:

$$M_{X|A=a}(t) = (e^{\theta}-1)^{-a}\left(\sum_{r=1}^{\infty}\frac{(\theta e^t)^r}{r!}\right)^a$$

$$= (e^{\theta}-1)^{-a}\left\{\theta^a e^{at} + \sum_{x=a+1}^{\infty}\left[\frac{1}{x!}\sum_{r=1}^{a}\binom{a}{r}(-1)^{a+r}(r\theta)^x\right]e^{xt}\right\}$$

Therefore

$$\mathbf{Pr}(X = a|A = a) = (e^{\theta}-1)^{-a}\theta^a \qquad (6), \qquad \text{and}$$

$$\mathbf{Pr}(X = x|A = a) = (e^{\theta}-1)^{-a}\frac{1}{x!}\sum_{r=1}^{a}\binom{a}{r}(-1)^{a+r}(r\theta)^a \qquad (7)$$

for $x = a+1, a+2, \ldots$

The joint probability function of (A, X) is then given by

$$\mathbf{Pr}(A = 0, X = 0) = [1-(1-\phi)(1-e^{-\theta})]^n \quad \mathbf{Pr}(A = a, X = x) = \mathbf{Pr}(X = x|A = a)\mathbf{Pr}(A = a)$$

$a \geq 1, x \geq a$, where $\mathbf{Pr}(X = x|A = a)$ is given by (6) and (7) and

$$\mathbf{Pr}(A = a) = \binom{n}{a}[(1-\phi)(1-e^{-\theta})]^a[1-(1-\phi)(1-e^{-\theta})]^{n-a}$$

The probability of the rejection region is then given by

$$\pi(\theta_o,\phi) = \sum_{\substack{(a,x)\text{ such} \\ \text{that }|t|>1.96}}\sum \mathbf{Pr}(A = a, X = x)$$

NUMERICAL EXAMPLES

When the values of $\boldsymbol{n}$ and θ_0 have been chosen, we can plot the probability of the rejection region under $\boldsymbol{H_o}$ (i.e. $\pi(\theta_o,\phi)$) against ϕ. In Figure 1, the probability $\pi(\boldsymbol{i},\phi)$ has been plotted against ϕ for $\theta_0 = 1, 2, 3, 4$ and $\boldsymbol{n} = 10, 15, 20, 25, 50$. The common characteristic of the curves obtained is that the curve initially is stable around a certain constant $\boldsymbol{c}$ when ϕ is small, and it increases quite sharply when ϕ approaches 1. Furthermore the value of $\boldsymbol{c}$ approaches 0.05 when $\boldsymbol{n}$ is large.

For each curve, it would be useful to note the value of ϕ^* such that the curve is stable around a certain constant when $\phi \in [0, \phi^*]$. For example, when $\theta_0 = 1$ and $\boldsymbol{n} = 50$, the value of $|\pi(1,\phi) - 0.05|$ is less than 0.03 when $0 \leq \phi \leq \phi^* = 0.53$. This means if we are quite sure that the actual value of ϕ is less than 0.53, then the test may be taken to have a significance level given by 0.05 ± 0.03. Table $\boldsymbol{i}$ gives the value of ϕ^* such that $|\pi(\theta_0,\phi) - 0.05| \leq (0.01)\boldsymbol{i}$, $\boldsymbol{i} = 1,2,3$.

Table 1 to Table 3 may be used to determine the size of $\boldsymbol{n}$ such that the test would have approximately a level of 0.05. The procedure is as follows:

First we choose the level of tolerance δ (δ= 0.01, 0.02 or 0.03) for the deviation of the significance level from the target value 0.05. We next note the values of θ_0 and $\tilde{\phi}$ and refer to one of Table 1 to Table 3. If the cell which corresponds to θ_0 and a given value $\boldsymbol{n}^*$ of $\boldsymbol{n}$ is larger than $\tilde{\phi}$, then $\hat{\boldsymbol{n}} = \boldsymbol{n}^*$ would be adequate.

For example if δ = 0.03, $\theta_0 = 3$ and $\tilde{\phi} = 0.6$, then Table 3 should be referred. The cell which corresponds to $\theta_0 = 3$ and $\boldsymbol{n}^* = 15$ gives the value 0.71 which is larger than 0.6. Thus the value of $\boldsymbol{n}$ = 15 would be adequate.

A dash in Table 1 to Table 3 means that for the indicated values of θ_0 and $\boldsymbol{n}$, the deviation of the significance level of the proposed test of mean from 0.05 is more than the indicate level of tolerance for all values of ϕ.

Table 1. Value of ϕ^* within which the level of the test is 0.05 ± 0.01

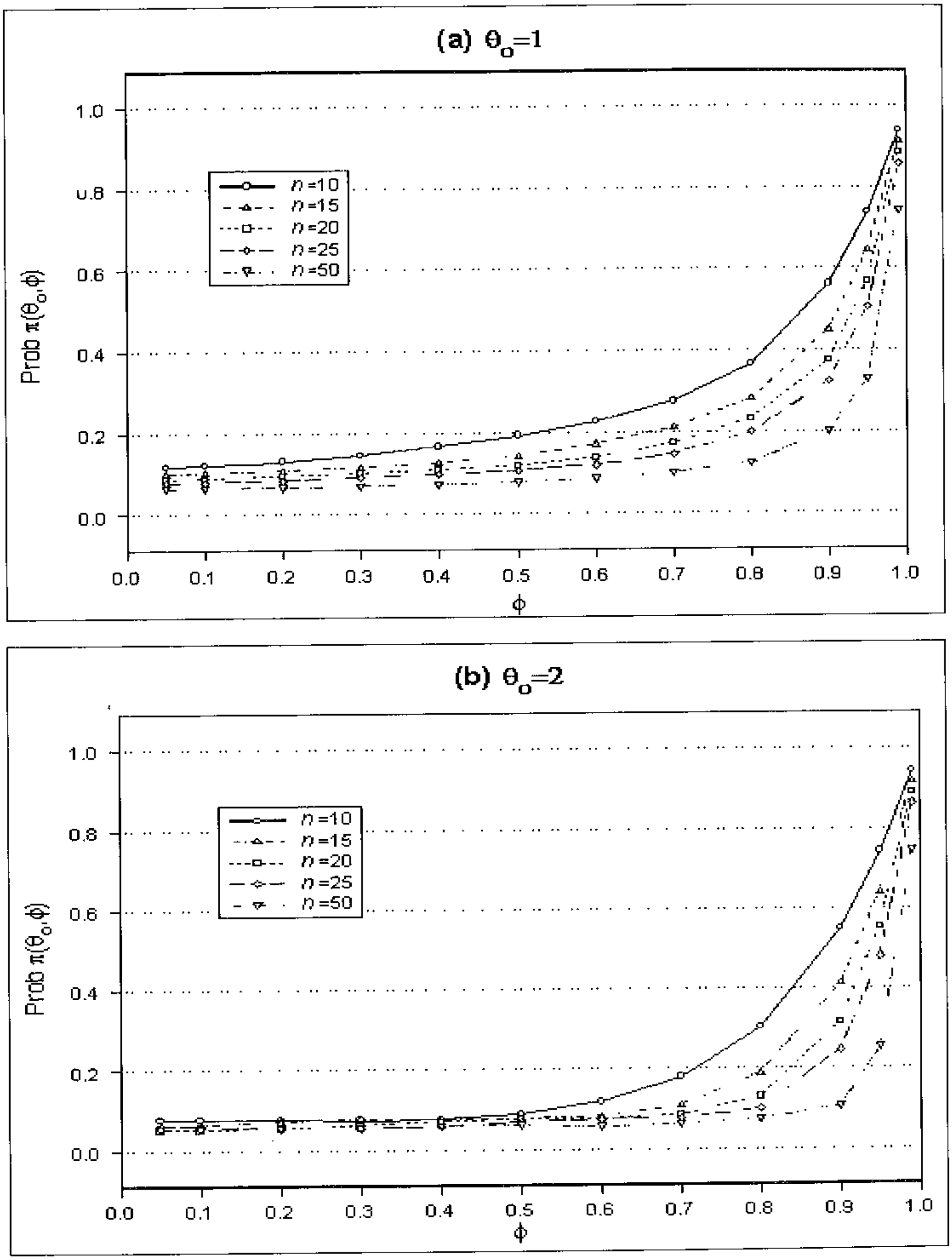
θ_0 \ $\boldsymbol{n}$	**10**	**15**	**20**	**25**	**50**
1	-	-	-	-	-
2	-	0.07	0.28	0.42	0.70
3	0.33	0.52	0.63	0.70	0.82
4	0.37	0.53	0.64	0.71	0.82
5	0.34	0.53	0.64	0.71	0.83

Table 2. Value of ϕ^* within which the level of the test is 0.05 ± 0.02

θ_0 \ $\boldsymbol{n}$	**10**	**15**	**20**	**25**	**50**
1	-	-	-	-	0.34
2	-	0.19	0.40	0.54	0.77
3	0.52	0.65	0.72	0.78	0.87
4	0.56	0.70	0.64	0.80	0.90
5	0.53	0.67	0.73	0.79	0.89

Table 3. Value of ϕ^* within which the level of the test is 0.05 ± 0.03

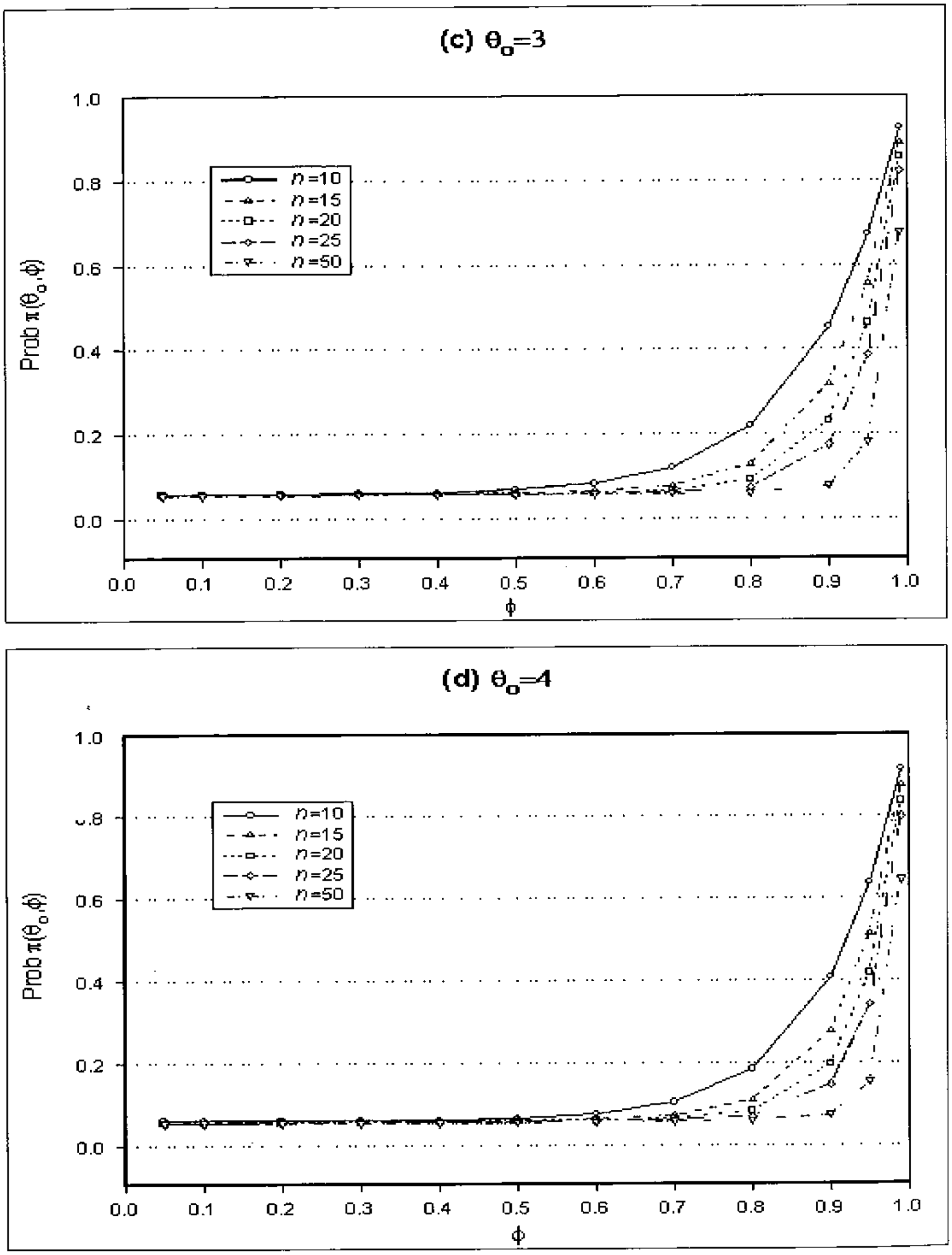
θ_0 \ $\boldsymbol{n}$	**10**	**15**	**20**	**25**	**50**
1	-	-	-	0.11	0.54
2	0.42	0.57	0.65	0.71	0.82
3	0.59	0.71	0.76	0.81	0.90
4	0.62	0.72	0.80	0.81	0.91
5	0.61	0.72	0.79	0.81	0.90



(a) θ_o=1

(b) θ_o=2

Figure 1. The probability $\pi(\theta_o,\phi)$ against ϕ for n=10, 15, 20, 25, 50



(c) θ_o=3

(d) θ_o=4

Figure 1. The probability $\pi(\theta_o,\phi)$ against ϕ for n = 10, 15, 20, 25, 50 (continued)

CONCLUDING REMARKS

There are a number of problems which may be considered for future research.

For example, when the deviation of the significance level of the proposed test of mean from the target value 0.05 is large, we may consider the problem of adjusting the critical value 1.96 to another value such that the resulting deviation would be small. We may also attempt to find the sample size n such that the test has a specified minimum power when the value of θ is fairly different from θ_o.

REFERENCES

1. Cohen, A. C. (1963). Estimation in mixtures of discrete distributions. *Proceedings of the International Symposium on Discrete Distributions, Montreal*. pp. 373 - 378.
2. David, F. N. and Johnson, N. L. (1952). The Truncated Poisson. *Biometrics* **8**: 275 - 285.
3. Gupta, P.L., Gupta, R.C. and Tripathi, R.C. (1995). Inflated modified power series with applications. *Communications in Statistics* **24** (9): 2355 - 2379.
4. Gupta, P.L., Gupta, R.C. and Tripathi, R.C. (1996). Analysis of zero-adjusted count data. *Computational Statistics and Data Analysis* **23**: 207 - 218.
5. Heilborn, D.C. (1989). *Generalized Linear Models for Altered Zero Probabilities and Over Dispersion in Count Data.* Unpublished Technical Report, University of California, San Francisco, Department of Epidemeology and Biostatistics.
6. Irwin, R. F. and Goen, R. L. (1958). Minimum variance unbiased estimators estimation from the truncated Poisson distribution. *Ann. Math. Statist.* **29**: 755 - 765.
7. Johnson, N.L. and Kotz, S. (1969). *Distributions in Statistics: Discrete Distributions*. Boston: Houghton Miffin.
8. Lambert, D. (1992). Zero inflated Poisson regression with an application to defects in manufacturing. *Technometrics* **34**: 1 - 14.
9. Yip, P. (1988). Inference about the mean of a Poisson distribution in the presence of a nuisance parameter. *Austral. J. Statist.* **30** (3): 299 - 306.