

*Journal of the Korean
Data & Information Science Society
1998, Vol. 9, No. 2, pp. 247 ~ 253*

변화시점이 있는 영과잉-포아송모형에서 돌출대립가설에 대한 우도비검정¹

김 경 무²

요약

영과잉-포아송모형에서 변화시점이 있는 경우, 돌출대립가설에 대한 우도비검정을 이용하여 변화시점의 유·무를 알아 보았다. 변화시점에 대한 추정은 최소제곱법을 이용하였고 이를 최우추정법을 이용하기 위한 초기치로 활용하였다. 또한 대립가설에 대한 몇 가지 흥미있는 모수들을 적률법을 이용하여 추정하였다. 모의실험을 통하여 이들 추정량을 비교하였고 결과 변화시점에 대한 추정은 최소제곱법보다는 최우추정법이 바람직하게 나타났고 흥미있는 몇 가지 모수들에 대해서는 최우추정량이 적률추정량보다 우수하게 나타났다.

주제어: 변화시점, 영과잉-포아송모형, 돌출대립가설, 우도비검정

1. 서 론

영과잉-포아송분포(Zero-Inflated Poisson: 이후 ZIP로 표기함)라 함은 이산형 확률분포에 있어서 정상적인 포아송 확률분포보다 영의 값이 과잉 관측되는 경우를 묘사하기 위하여 원래의 확률분포를 변형시킨 것을 말한다. 포아송분포는 생산공정단계에서 발생하는 불량품의 수에 관한 확률분포로서 지금까지 중요한 분포로 이용되어 왔다. 그러나 문명의 발달과 제품을 만들어 내는 기술의 고급화로 인하여 불량률은 현저하게 감소되어 가고 있다. 예를 들면, 반도체 분야에서 컴퓨터에 내장되어 있는 메모리 칩들은 생산공정의 정확도와 기술의 발달로 불량품으로 판정되는 경우가 드물다. 다시 말하면 여러번 실시한 표본검사 중 단위당 불량품의 수는 대부분 영이고 극히 적은 경우에 한하여 불량품이 발생한다고 볼 수 있다. 이러한 경우, 기존의 포아송분포에 적합시켜 통계적인 추정 및 검정을 한다면 이는 제 3종의 통계적 오류를 범하는 결과를 초래할 것이다.

¹이 논문은 1998학년도 대구대학교 학술연구비 지원에 의한 논문임.

²대구대학교 자연과학대학 통계학과 교수, (712-714) 경북 경산시 진량읍 내리리 15

생산공정과정에서 단위당 나타나는 불량품 수가 이러한 ZIP 분포를 따른다고 할 때, 불량품 수가 어떤 처음 시점에서 다음 변화시점까지 변화가 있다고 하자. 생산공정이 돌발적인 원인들로 인하여 변화가 되고 적당한 시간이 흐른 다음 다시 원상태로 환원되는 경우, 변화된 상태를 돌출상태(epidemic state)라 한다. 이때 미지의 두 시점을 변화시점(change-points)이라 한다. 변화시점에 관한 추정 및 검정 그리고 불량률의 변화가 어느 정도인지를 추정하는 것은 중요한 일일 것이다. 이러한 ZIP 분포는 처음으로 Singh(1963)에 의해 소개되었다. 그러나 이 분포는 수학적인 모형으로만 인식되어 응용분야가 다양하지 못하였다. 그 이후 Lambert(1992)는 ZIP 회귀모형을 통하여 공변량(covariate)들의 효과를 연구하고 이를 실제의 자료에 적용하였다. 한편 김경무(1998)는 변화시점이 한 개 있을 때 영과잉-포아송모형을 생각하고 변화시점에 관한 우도비 검정을 연구하였다. 돌출상태인 모형은 일반적으로 경제모형에서 많이 이용되는데 변화시점모형에 처음으로 적용한 이는 Broemeling-Tsurumi(1987)이다. 이들은 정규분포를 가정하고 여러 형태의 검정통계량의 성질을 연구하였다.

2절에서는 변화시점이 있는 ZIP분포와 돌출대립가설을 소개하고 3절에서는 돌출대립가설에 대한 변화시점의 유·무를 우도비검정을 통하여 알아본다. 다음 4절에서는 흥미있는 모수들에 대한 추정량을 적률 및 최우추정법에 의해 유도하고 이를 5절의 모의실험을 통하여 비교해 본다.

2. 돌출 ZIP모형

확률변수 Y 는 생산공정에서 일정 단위당 불량품이 나타나는 수로서 ZIP 분포를 따른다. ZIP는 포아송분포와 베르누이분포와의 혼합모형으로 볼 수 있다. 즉,

$$\begin{aligned} Y &\sim 0 & , & \text{ } p \text{의 확률로} \\ &\sim \text{Poisson}(\lambda) & , & \text{ } 1-p \text{의 확률}, \end{aligned}$$

여기에서 $0 \leq p \leq 1$ 는 불량품이 전혀 나타나지 않는 상태(perfect state)의 확률이며 $\lambda > 0$ 는 포아송분포의 평균이다. 이때 확률질량함수(pmf)는 아래와 같이 된다.

$$\begin{aligned} P(Y = k) &= p + (1-p)e^{-\lambda} & , & \text{ } k = 0 \\ &= (1-p)\lambda^k e^{-\lambda}/k! & , & \text{ } k = 1, 2, \dots \end{aligned}$$

앞으로 위 분포를 $ZIP(p, \lambda)$ 로 표기하기로 한다. 서로 독립인 확률변수 Y_1, Y_2, \dots, Y_n 가 시간의 흐름에 따라 연속적으로 얻을 수 있는 관측자료라 하자. 이때 생산공정과정 중 여러가지의 원인들로 인하여 돌출상태가 된다고 하자. 즉, 미지의 첫번째 시점 a 이후부터 두 번째 변화시점 b 까지 분포의 돌출변화가 있는 다음과 같은 모형을 생각할 수 있다.

$$\begin{aligned} Y_1, Y_2, \dots, Y_a, Y_{b+1}, Y_{b+2}, \dots, Y_n &\sim ZIP(p, \lambda) \\ Y_{a+1}, \dots, Y_b &\sim ZIP(p^*, \lambda^*), \end{aligned}$$

여기에서 양의 정수 $a, b (1 \leq a < b < n)$ 는 미지의 변화시점(changepoints)이고 p^* 와 λ^* 은 각각 돌출변화 이후 불량품이 전혀 나타나지 않는 상태의 확률과 불량품 수에 대한 평균을 의미한다. 만약 $p > p^*$ 혹은 $\lambda < \lambda^*$ 이 된다면, 돌출변화 이후 불량품수가 증가될 것이다. 우리는 위 모형을 돌출 ZIP모형, $ZIP(p, \lambda, a, b, p^*, \lambda^*)$ 이라 하겠고 6개의 모수를 포함하고 있다.

3. 우도비 검정

생산공정과정 중 돌출변화가 있는지를 검정하기 위하여, 변화가 없다는 귀무가설; $H_0 : p = p^*$, 그리고 $\lambda = \lambda^*$ 그리고 불량품의 수에 변화가 있다는 돌출대립가설; $H_1 : p \neq p^*$ 혹은 $\lambda \neq \lambda^*$ 을 설정할 수 있다. 위 경우 귀무가설에 대한 로그-우도함수(log-likelihood function)는

$$l(y ; p, \lambda) = \sum_{i=1}^n \ln[\{p + (1-p)e^{-\lambda}\}I(y_i = 0) + \{(1-p)\lambda^{y_i} e^{-\lambda}/y_i!\}I(y_i > 0)]$$

이 된다. 여기에서 $I(\cdot)$ 는 지시함수(indicator function)이다. 또한 돌출대립가설에 대한 로그-우도함수는 다음과 같다.

$$\begin{aligned} l(y ; p, \lambda, a, b, p^*, \lambda^*) &= \sum_{i=1}^n \ln \left[\{p + (1-p)e^{-\lambda}\} \{I(i \leq a, y_i = 0) + I(b < i \leq n, y_i = 0)\} \right. \\ &\quad + \{(1-p)\lambda^{y_i} e^{-\lambda}/y_i!\} \{I(i \leq a, y_i > 0) + I(b < i \leq n, y_i > 0)\} \\ &\quad \left. + \{p^* + (1-p^*)e^{-\lambda^*}\}I(a < i \leq b, y_i = 0) + \{(1-p^*)\lambda^{*y_i} e^{-\lambda^*}/y_i!\}I(a < i \leq b, y_i > 0) \right] \end{aligned}$$

위 돌출대립가설에 대한 로그-우도함수는 6개의 모수로 이루어진 복잡한 함수이다. 그러므로 p, λ, a, b, p^* 그리고 λ^* 의 최우추정량(MLE)을 각각 $\tilde{p}, \tilde{\lambda}, \tilde{a}, \tilde{b}, \tilde{p}^*$, 그리고 $\tilde{\lambda}^*$ 라 한다면, 이들 추정량은 간단한 형태로(closed form) 나타나지 않아 우도함수를 최대로하는 수치해석적 방법을 이용하여 구할 것이다. 연구자는 함수의 최대치를 구하는 한 가지 방법으로 Powell(1964) 방법을 이용하여 최우추정치를 구하였다. 이에 따른 우도비 검정통계량은

$$T_n = -2\{l(y ; \tilde{p}, \tilde{\lambda}) - l(y ; \tilde{p}, \tilde{\lambda}, \tilde{a}, \tilde{b}, \tilde{p}^*, \tilde{\lambda}^*)\}$$

이다. 위 통계량은 잘 알려져 있듯이 대표본에서 자유도 4인 카이제곱분포에 접근한다. 그러므로 T_n 의 값이 크면 귀무가설을 기각시킬 수 있다.

4. 추정

변화시점이 있는 ZIP모형에서 $p, \lambda, p^*, \lambda^*$ 그리고 두 변화시점 a 그리고 b 를 추정하는 일은 흥미있는 일일 것이다. 변화시점을 제외한 모든 모수들의 추정방법은 적률을 이용한 추정법(MME)과 최우추정법(MLE)을 이용할 것이다. MME는 비교적 알기 쉬운 형태로 유도될 수 있으며 이를 MLE를 수치해석적(반복법)으로 구하기 위한 초기값으로 이용할 수 있다. 먼저 변화시점 a, b 의 추정량은 돌출상태와 돌출이 아닌 상태의 관측치 제곱합이 최소가 되도록 구한다. 즉, (a, b) 의 최소제곱추정량 (\hat{a}, \hat{b}) 는 식 (1)을 최소로 하는 양의 정수 (c, d) , $1 \leq c < d < n$, 값을 변화시점에 대한 최소제곱추정량(LSE)으로 설정하였다.

$$\begin{aligned} & \sum_{i=c+1}^d \left(y_i - \frac{\sum_{i=c+1}^d y_i}{d-c} \right)^2 + \sum_{i=1}^c \left(y_i - \frac{\sum_{i=1}^c y_i + \sum_{i=d+1}^n y_i}{n-d+c} \right)^2 \\ & + \sum_{i=d+1}^n \left(y_i - \frac{\sum_{i=1}^c y_i + \sum_{i=d+1}^n y_i}{n-d+c} \right)^2 \end{aligned} \quad (1)$$

위 변화시점에 대한 추정량 역시 식 (1)에서 직접 유도될 수 없고 주어진 표본을 이용해서 구해야 할 것이다. 이제 변화시점의 추정량이 결정이 되면, 돌출 ZIP모형의 표본들 중 돌출상태의 표본과 돌출이 아닌 상태의 표본들을 대상으로 적률방법(method of moments)을 이용하면 식 (2)를 얻을 수 있고 식 (2)의 우변을 편의상 A,B,C 그리고 D라 두기로 하자. 식 (2)의 좌변은 돌출 ZIP모형에서 구한 것이고 우변은 표본에서 얻은 통계량이다. 순서대로 돌출상태가 아닐 때의 평균, 제곱평균 그리고 돌출상태에서의 평균, 제곱평균이다.

$$\begin{aligned} (1-p)\lambda &= \frac{\sum_{i=1}^{\hat{a}} y_i + \sum_{i=\hat{b}+1}^n y_i}{n-\hat{b}+\hat{a}} \equiv A, \\ (1-p)\lambda(1+\lambda) &= \frac{\sum_{i=1}^{\hat{a}} y_i^2 + \sum_{i=\hat{b}+1}^n y_i^2}{n-\hat{b}+\hat{a}} \equiv B, \\ (1-p^*)\lambda^* &= \sum_{i=\hat{a}+1}^{\hat{b}} \frac{y_i}{\hat{b}-\hat{a}} \equiv C, \\ (1-p^*)\lambda^*(1+\lambda^*) &= \sum_{i=\hat{a}+1}^{\hat{b}} \frac{y_i^2}{\hat{b}-\hat{a}} \equiv D. \end{aligned} \quad (2)$$

또한 식 (2)를 모수에 관하여 정리하면, 변화시점을 제외한 나머지 모수들의 MME는 식 (3)-(6)과 같다.

$$\hat{\lambda} = \frac{A}{B} - 1, \quad (3)$$

$$\hat{p} = 1 - \frac{A}{\hat{\lambda}}, \quad (4)$$

$$\hat{\lambda}^* = \frac{D}{C} - 1, \quad (5)$$

$$\hat{p}^* = 1 - \frac{C}{\hat{\lambda}^*}. \quad (6)$$

물론 위 모든 추정량들 양의 값을 지녀야 한다. 그리고 MME 추정치는 MLE를 구하기 위한 초기값으로 활용되어질 수 있다.

5. 모의실험

<표 5.1>은 영파잉-포아송 난수를 프로그램을 작성하여 추출한 표본이다. 실험조건은 돌출 영파잉-포아송분포, ZIP ($p, \lambda, a, b, p^*, \lambda^*$)에서 $p = 0.7, \lambda = 1.0, p^* = 0.2, \lambda^* = 2.5$ 표본 크기 $n = 50$ 이다. 또한 돌출상태의 시점 $a = 30$, 그리고 종점 $b = 40$ 으로 설정하였다. 이 자료를 보면 영이 66%나 차지하고 있어 기존의 포아송분포를 따른다고 볼 수 없다. 그리고 돌출변화 시점과 종점을 쉽게 찾기가 어렵다. 먼저 이 자료가 영파잉-포아송분포를 따른다고 했을 때, 돌출변화가 없다는 귀무가설과 돌출변화가 있다는 대립가설을 우도비 검정하여 보면 검정통계량 값은 $T_n = 94.878$ 이다. 이에 따른 자유도 4인 카이제곱분포의 유의확률은 0.000으로 귀무가설을 기각하게 되어 돌출변화가 있음을 알 수 있다. 이 결과는 모의 실험에서 영파잉-포아송분포에 돌출변화를 가한 사실과 같게 된다. 다음으로 돌출상태의 시점(a)와 종점(b)를 모른다고 생각하고 최소제곱법(LSE)과 최우추정법(MLE)을 이용하여 구하면 <표 5.2>와 같다. 실제의 변화시점 모두 $a = 30, b = 40$ 과 비교해 보면 LSE보다는 MLE가 비교적 잘 추정되어 있음을 알 수 있다. 또한 <표 5.2>의 팔호안의 값은 모의 실험 반복을 500번하여 얻은 평균제곱오차(MSE)를 나타낸다. 그 결과 역시 MLE가 LSE보다 우수하게 나타난다. 나머지 모두들을 적률법(MME)과 최우추정법을 이용하여 구한 결과가 <표 5.3>에 나타나 있다.

<표 5.3>의 결과를 살펴보면 MLE가 MME보다 참값에 가깝게 잘 추정되어 있음을 볼 수 있고 $\hat{a} = 35, \hat{b} = 39$ 사이의 표본들은 $p^* = 0.000$ 이므로 기존의 평균이 $\lambda^* = 1.751$ 인 포아송 분포를 따르게 됨을 알 수 있다.

<표 5.1> 돌출 영과잉-포아송 난수

0, 0, 0, 2, 1, 1, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 3, 0, 0, 0, 0, 0, 2, 2, 0, 0
0, 0, 0, 2, 0, 1, 0, 2, 0, 0, 1, 2, 1, 3, 0, 0, 0, 2, 0, 0, 2, 0, 0, 0, 2

<표 5.2> 변화시점 추정치

추정량	<i>a</i>	<i>b</i>
LSE	6(102.5)	13(123.7)
MLE	35(64.7)	39(85.4)

<표 5.3> 추정치 비교

모수	<i>p</i>	λ	p^*	λ^*
참값	0.700	1.000	0.200	2.500
MME	0.302	1.032	0.000	0.000
MLE	0.068	1.359	0.000	1.751

References

1. 김경무, (1998). 변화시점이 있는 영과잉-포아송모형, 통계이론방법연구, 9(1), 1-9.
2. Broemeling, L. D. & Tsurumi, H. (1987). *Econometrics and Structural Change*, New York: Marcel Dekker.
3. Lambert, D. (1992). Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing, *Technometrics*, 34, 1-14.
4. Powell, M. J. D. (1964). An Efficient Method for Finding the Minimum of a Function of Several Variables without Calculating Derivatives, *Computer Journal*, 7, 155-162.
5. Singh, S. N. (1963). A Note on Inflated Poisson Distribution, *Journal of the Indian Statistical Association*, 1, 140-144.
6. Yao, Q. (1993). Tests for change-points with epidemic alternatives, *Biometrika* , 80, 1, 179-191.

Likelihood Ratio Test for the Epidemic Alternatives on the Zero-Inflated Poisson Model³

Kyungmoo Kim⁴

Abstract

In case of the epidemic Zero-Inflated Poisson model, likelihood ratio test was used for testing epidemic alternatives. Epidemic changepoints were estimated by the method of least squares. It were used for starting points to estimate the maximum likelihood estimators. And several parameters were compared through the Monte Carlo simulations. As a result, maximum likelihood estimators for the epidemic changepoints and several parameters are better than the least squares and moment estimators.

³This paper was supported by Taegu University research fund, 1998.

⁴Professor, Department of Statistics, Taegu University, Kyungpook, 712-714 Korea.