Journal of the Korean
Statistical Society
Vol. 22, No. 2, 1993

포아송으로부터 부의 이항분포로의 이탈에 대한 검정통계량의 확장

이선호[1]

요 약

포아송분포로부터 부의 이항분포로의 이탈을 검색하는 통계량들이 자료의 형태에 따라 여러 가지 제시되었다. 그런데 대립가설인 부의 이항분포의 모수화 방법에 따라 분산과 평균의 구조가 변하고 국소 최적 검정 통계량도 달라진다는 것이 알려졌다. 본 논문에서는 대립가설을 일반적인 포아송 혼합분포로까지 확장시키고, 일반적인 형태의 분산과 평균의 구조에도 검정 가능한 새로운 통계량 L을 소개하고 있다. 또한 L 통계량은 포아송분포로부터 부의 이항분포로의 이탈을 다루는 기존의 여러 통계량들의 일반화된 형태임을 보였다. 점근적 상대효율과 모의실험을 통하여 L 통계량과 기존의 통계량들을 비교한 결과 분산과 평균사이의 구조에 상관없이 L 통계량이 우수한 것임을 입증하였다.

KEYWORDS: Poisson mixture models, overdispersion in Poisson models, mean-variance structures of negative binomial model, hyper Poisson variation.

1. 서론

포아송 모형은 가산자료의 회귀분석이나 범주형 자료를 분석하는데 폭 넓게 이용되고 있다. 그러나 포아송분포로써 가산자료를 모형화함에 있어 표본분산이 표본평균보다 큰 현상, 즉 초포아송 변이(hyper Poisson variation)의 문제가 야기되곤 한다.

Hausman, Hall과 Griliches(1984)는 기업에서 지출하는 연구개발 비용과 획득하는 특허 건수와의 관계를 규명하는데 과산포 포아송 모형을 사용하고 있다. 각 기업에서 연구개발에

[1]133-747, 서울특별시 성동구 군자동 98번지, 세종대학교 수학과

지출하는 비용이 확률변수이고 비용이 주어졌을 때 특허 건수의 조건분포는 포아송 분포를 이루게 된다. 따라서 특허 건수의 무조건분포는 포아송의 혼합분포를 이루게 되며, 이 경우 특허 건수 자료는 초 포아송 변이를 갖게 된다.

Margolin, Kaplan과 Zeiger(1981)는 살모넬라를 이용한 에임즈(Ames) 돌연변이 (mutagenecity) 검사에서 초 포아송 변이를 검색해 내었고, 이러한 초 포아송 변이를 무시했을 때의 문제점으로 비효율적인 추정, 모수 추정치에 대한 분산의 과소추정과 가설검정에서 오류를 범할 확률이 증가한다는 것 등을 지적하였다.

초 포아송 변이는 포아송 가정을 판단하는데 사용되는 자료들이 독립적이기는 하나 동일 분포를 갖지 못했기 때문에 발생한 것으로, 이러한 자료를 대상으로 포아송으로부터 포아송 혼합의 한 형태인 부의 이항분포로의 이탈을 다루는 문제는 많은 논문에서 다루어졌다.(Paul과 Plackett(1978), Collings와 Margolin(1985), Dean과 Lawless(1989), Kim과 Park(1992), Dean(1992) 등)

본 논문에서는 지금까지 유도된 기존의 통계량들과 이선호(1991)가 새로이 유도한 L 통계량을 비교하였다. L 통계량은 분산과 평균이 $\sigma^2 = \mu + c\mu^r$ 형태를 만족하는 일반적인 포아송의 혼합분포를 대립가설로 하였을 때 포아송분포로부터의 이탈을 검색하는 통계량으로 제시되었다. 2장에서는 포아송분포로부터의 이탈을 검정하기 위한 가정을 설정하였다. 3장에서는 초 포아송 변이를 검정하는 통계량들을 자료들의 형태에 따라 세가지로 분류하고 우수한 몇 가지 통계량들과 점근적으로 동등한 검정을 유도하였다. 또한 유효 스코어에 기초한 통계량인 L도 분류된 자료들의 형태에 맞게 L_B와 L_C로 나타내고 이를 점근적 상대효율과 모의실험을 통하여 부의 이항분포를 대립가설로 하고 분산과 평균 사이의 관계를 설정하고 구한 기존의 통계량들과 비교한 결과, L은 분산과 평균 사이의 구조에 상관없이 우수한 것임을 밝혔다.

2. 가정의 설정

일반적으로 어떤 분포를 추정하려 할 때 그에 속한 모수가 단일값을 가짐으로써 분포가 하나의 분포로 추정될 수도 있으나 모수가 확률변수일 경우에는 여러 분포의 가중평균이나 적분형태로 표현될 수 있고 이렇게 하여 새로운 분포가 만들어 질 수 있는데 이를 혼합분포라 한다.

부의 이항분포(negetive binomial distribution)는 포아송분포의 감마 혼합(a gamma mixture of Poisson)이다. 성공확률이 P, 실패확률이 $1-P$ 인 버눌리 시행을 독립적으로 반복 시행할 때 N번째 성공을 거둘 때까지 관찰해야 할 실패 횟수의 확률분포로 정의는 다음과 같다.

정의 1. 확률변수 X가 다음의 확률밀도함수를 가질 때 모수가 N, P인 부의 이항분포(negative binomial distribution)를 이룬다고 말하며, $X \sim NB(N, (1-P)/P)$로 표기한다.

$$Pr(X=x) = \binom{N+x-1}{N-1} P^N (1-P)^x, \qquad x = 0, 1, \cdots.$$

부의 이항분포에서는 모수화에 따라서 분산과 평균의 관계를 변화시킬 수 있다. 즉 $c > 0$, $m > 0$ 에 대해 $X \sim NB(1/c,\ cm)$ 이면 확률변수 X의 평균은 m, 분산은 $m(1+cm)$으로 분산은 평균의 이차함수 형태가 된다. 또한 $X \sim NB(m/c,\ c)$일 때는 평균은 m, 분산은 $m(c+1)$로 분산과 평균 사이가 선형관계가 된다. 분산과 평균 사이의 관계를 일반화시키는 모수화는 $X \sim NB(m^{2-r}/c,\ cm^{r-1})$로서 이때 평균은 m, 분산은 $m + cm^r$로 되어 분산과 평균 사이의 이차함수 및 선형관계를 모두 포함하게 된다.

X가 평균 λ 인 포아송분포를 이룰 때 $X \sim P(\lambda)$로 표기한다. 포아송분포에서는 분산과 평균이 같은 데 비해 위의 부의 이항분포의 경우는 $c \to 0$ 일 때 점근분포가 $X \sim P(m)$이 됨을 알 수 있다. 그러므로 귀무가설이 포아송분포이고 대립가설이 부의 분포일 때의 검정가설은 분산이 평균의 이차 함수일 경우 $X \sim NB(1/c,\ mc)$이며 $H_0 : c = 0$ 대 $H_a : c > 0$ 으로 놓을 수 있고, 분산과 평균이 선형 관계일 경우는 $X \sim NB(m/c,\ c)$ 이며 $H_0 : c = 0$ 대 $H_a : c > 0$ 으로 놓을 수 있다. 여기서 $c = 0$은 극한적 논리(limiting argument)로서 $c \to 0$을 의미한다.

포아송분포에서 과산포 문제는 1970년대부터 활발히 다루어져 부의 이항분포를 대립가설로 한 초 포아송 변이에 대한 많은 통계량들이 제시되었다. Collings와 Margolin(1985)은 귀무가설인 포아송분포를 만족하는 표본들의 형태를 다음과 같이 세 가지로 나누었고 본 논문에서도 이에 따라 포아송 가정의 적합성을 검정하기 위한 통계량을 아래의 분류에 따라 다루겠다.

경우 A: 확률표본 형태 (random sample)

모든 i에 대해 $E(Y_i) = m$을 만족하는 확률표본 $Y_1, Y_2, \cdots, Y_n$

경우 B: 원점을 통과하는 회귀 형태 (regression through the origin)

모든 i에 대해 $E(Y_i) = \beta_i m$ (단 β_i는 모두 알려진 양수)를 만족하는 확률변수 Y_1, Y_2, $\cdots$, Y_n

경우 C: 일원배열 형태 (one-way layout)

$j = 1, 2, \cdots, n_i$, $i = 1, 2, \cdots, k$ 일 때 $E(Y_{ij}) = m_i$ 를 만족하는 $\sum_{i=1}^{k} n_i$개의 확률변수들

3. 검정통계량의 비교

3-1. 확률표본 형태의 경우

포아송분포로부터 다른 분포로의 이탈을 검정하는 표준적 검정은 다음의 S_{A1}이 큰 값을 가질 때 H_0를 기각하는 분산검정(variance test)이 있다.

$$S_{A1} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y}_+)^2}{\overline{Y}_+}, \qquad \overline{Y}_+ = \frac{Y_+}{n} = \frac{\sum Y_i}{n}$$

Fisher *et al.*(1922)에 의해 제시된 이 검정통계량은 부의 이항분포를 대립가설로 하였을 때 모수 m의 분포함수인 감마분포의 분산값 σ^2에 대해 국소 최강력 불편검정 (locally most powerful unbiased test)임을 보였다. m의 값을 알 경우에는 분산검정이 동질성 검정에 대해 최적검정이 되지 못한다. Potthoff와 Whittinghill(1969)은 m의 값을 알 경우 $S_{A2} = \sum Y_i^2 - (2m+1)\sum Y_i$가 큰 값을 가질 때 H_0을 기각하는 것이 부의 이항분포를 대립가설로 하였을 때의 국소 최강력 검정(locally most powerful test)임을 보였다. Moran(1970)은 대립가설이 포아송분포의 일반적 혼합인 광범위한 분포의 집합일 때 S_{A1}에 기초한 검정은 $C(\alpha)$ 검정이 되며 우도비(likelihood ratio) 방법에 기초한 검정과 점근적으로 동일하다는 것을 보였다.

포아송 모형의 표본이 확률표본일 경우는 다음 절의 기대값이 원점을 통과하는 회귀 형태에서 모든 i의 값이 1인 특수한 경우이므로 다른 통계량과의 구체적인 비교는 다음 절에서 다루기로 한다.

3-2. 원점을 통과하는 회귀 형태의 경우

3-2-1. 검정통계량

원점을 통과하는 회귀 형태의 경우는 회귀분석에서 절편항이 없는 회귀선처럼 $E(Y_i) = \beta_i m (i = 1, 2, \cdots, n)$으로 표현되며 m은 경우에 따라 이미 알려진 수이기도 하고 미지수이기도 하지만 대개 β_i는 알려진 양수이다. 이 경우 포아송으로부터의 이탈을 검색하는 것도 분산검정을 적절하게 변형시킨 것을 사용할 수 있다. Rao(1952)는 다음의 S_B를 검정통계량으로 제안하였다.

$$S_B = \sum_{i=1}^{n} \frac{(Y_i - \beta_i \widehat{m})^2}{\beta_i \widehat{m}}, \qquad \widehat{m} = \frac{Y_+}{\beta_+} = \frac{\sum Y_i}{\sum \beta_i} \tag{1}$$

Potthoff와 Whittinghill(1969)은 부의 이항분포를 대립가설로 하고 m의 값을 알 때 $S_{B2} = \sum Y_i(Y_i - 1) - 2m \sum \beta_i Y_i$가 국소 최강력 검정임을 보였고 실제 자료를 이용하여 검정력을 비교하였다.

Collings와 Margolin(1985)은 분산이 평균의 이차 함수 형태일 때에 한하여 다음 Neyman(1959)의 $C(\alpha)$ 검정통계량이 국소 최강력 검정임을 유도해 보였다.

$$T_B = \sum_{i=1}^{n} \frac{(Y_i - \beta_i \widehat{m})^2}{\overline{Y}_+}, \qquad \widehat{m} = \frac{Y_+}{\beta_+} \tag{2}$$

Dean과 Lawless(1989)는 대립가설을 부의 이항분포에서부터 모수의 1차 적률 (moment)과 2차 적률이 유한한 포아송의 혼합분포의 경우로 확장하여 포아송분포로부터의 이탈을 검정하는 통계량 D_B를 식 (3)과 같이 구했는데, 이것은 m이 무한대로 갈 때 식(2)의 T_B와 점근

적으로 동등(asymptotically equivalent)한 검정임을 쉽게 알 수 있다.

$$D_B = \sum\{(Y_i - \beta_i\widehat{m})^2 - Y_i\}, \qquad \widehat{m} = \frac{Y_+}{\beta_+} \tag{3}$$

'점근적으로 동등한 검정'이라 함은 서로 다른 두 통계량에서 표본 크기나 평균이 극한값에 수렴하여 얻어지는 극한 형태가 서로 같음을 의미한다.

분산이 평균과 선형인 관계에 있을 때 Kim과 Park(1992)은 다음 식 (4)의 K_B에 기초한 검정이 국소 최강력 검정이며 m이 무한대로 감에 따라 통계량 K_B는 통계량 $S_B + n$와 점근적으로 동등함을 보였다.

$$K_B = \sum \frac{(Y_i - \beta_i\widehat{m})^2 - Y_i}{\beta_i\widehat{m}}, \qquad \widehat{m} = \frac{Y_+}{\beta_+} \tag{4}$$

이 선호(1991)는 평균과 분산의 관계가 임의의 $c > 0$, r에 대해 $\sigma^2 = \mu + c\mu^r$인 일반적인 포아송의 혼합분포를 대립가설로 하였을 때 포아송분포로부터의 이탈을 검색할 수 있는 식 (5)의 통계량 L_B를 유효스코어(efficient score)를 사용하여 유도하였다. 그러므로 포아송 혼합분포의 한 형태인 부의 이항분포를 대립가설로 하였을 때 L_B를 검정통계량으로 사용할 수 있다.

$$L_B = \sum \frac{P_i^{r-1}(Y_i - \beta_i\widehat{m})^2}{\beta_i\widehat{m}}, \qquad \widehat{m} = \frac{Y_+}{\beta_+}, \quad P_i = \frac{\beta_i}{\beta_+} \tag{5}$$

Cox와 Hinkley(1974,pp.113-121)는 유효스코어가 $c = 0$ 일 때 검정력 함수(power function)의 기울기를 극대화 시키기 때문에 유효스코어에 바탕을 둔 통계량들은 국소 최강력 불편 검정(locally most powerful unbiased test)이 된다고 하였다.

경우 B에 대해 지금까지 구하여진 검정 통계량은 크게 S_B,T_B와 L_B로 나눌 수 있다. 그 중 S_B와 T_B를 비교한 결과, Kim과 Park(1992)은 대립가설인 부의 이항분포 에서 평균과 분산의 구조가 달라지면 지역적 최강력 검정이 달라지는 것을 밝혔다. 또한 식 (5)의 통계량 L_B에 $r = 1$을 대입하면 식 (1)의 통계량 S_B와 동일하고, $r = 2$ 이면 식 (2)의 통계량 T_B와 동일해 진다. 그러므로 L_B는 S_B와 T_B를 동시에 일반화한 검정통계량임을 쉽게 알 수 있다.

3-2-2. 근사적 귀무가설 분포와 대립가설 분포

앞에서도 언급하였듯이 B의 경우에 $Y_1, Y_2, \cdots, Y_n$는 서로 독립이고 $Y_i \sim NB(\frac{1}{c}(\beta_i m)^{2-r}, c(\beta_i m)^{r-1})$을 이루어 $E(Y_i) = \beta_i m$ 이고 $Var(Y_i) = \beta_i m[1 + c(\beta_i m)^{r-1}]$ 가 성립한다($i = 1, 2, \cdots, n$). $\sigma^2 = \mu + c\mu^r$이 성립하는 이러한 모수화에서 $m \to \infty$ 이고 $c \to 0$ 이면서 $cm^{r-1} \to t$($t \geq 0$인 상수)일 때 Collings와 Margolin(1985, p.416)처럼 다음이 성립

함을 쉽게 알 수 있다.

$$Z_i = \frac{Y_i - \beta_i m}{\sqrt{\beta_i m\{1 + c(\beta_i m)^{r-1}\}}} \quad \dot{\sim} \quad N(0,1), \qquad i = 1, 2, \cdots, n$$

단 '$\dot{\sim}$'는 근사적인 분포를 이룸을 의미한다.

대수의 약법칙(weak law of large numbers)에 의해 m이 커짐에 따라 $Y_i/(\beta_i \widehat{m}) = 1 + o_p(1)$이므로 L_B는 다음의 정리 1과 같이 독립인 표준 정규 확률변수(independent standard normal r.v.)의 이차항 형태로 확률수렴(convergence in probability)한다.

정리 1.

$$L_B = \sum \frac{P_i^{r-1}(Y_i - \beta_i \widehat{m})^2}{\beta_i \widehat{m}} = \mathbf{Z}'\mathbf{U}\mathbf{Z} \dot{+} o_p(1),$$

단,

$$P_i = \frac{\beta_i}{\beta_+}$$

$$\sqrt{\underline{P}} = (\sqrt{P_1}, \sqrt{P_2}, \cdots, \sqrt{P_n})'$$

$$\mathbf{Z} = (Z_1, Z_2, \cdots, Z_n)'$$

$$\mathbf{U} = \mathbf{D}(\sqrt{W_i})(\mathbf{I} - \sqrt{\underline{P}}\sqrt{\underline{P}'})\mathbf{D}(P_i^{r-1})(\mathbf{I} - \sqrt{\underline{P}}\sqrt{\underline{P}'})\mathbf{D}(\sqrt{W_i})$$

$$\mathbf{D}(a_i) = \mathrm{diag}(a_1, \cdots, a_n)$$

$$W_i = 1 + t\beta_i^{r-1}$$

증명. $m \to \infty$에 따라 $\frac{\widehat{m}}{m} = 1 + o_p(1)$이므로 '$\xrightarrow{\mathcal{L}}$'가 분포수렴(convergence in distribution)을 의미할 때 슬룻스키(Slutsky)의 정리에 의해 다음이 성립한다.

$$\frac{P_i^{r-1}(Y_i - \beta_i \widehat{m})^2}{\beta_i \widehat{m}} \quad \xrightarrow{\mathcal{L}} \quad \frac{P_i^{r-1}(Y_i - \beta_i \widehat{m})^2}{\beta_i m}$$

또한 $Z_i^* = \sqrt{1 + t\beta_i^{r-1}} Z_i$라 놓으면 $m \to \infty$, $c \to 0$이고 $cm^{r-1} = t + o_p(1)$일 때 $Z_i^* = \sqrt{1 + c(\beta_i m)^{r-1}} Z_i + o_p(1)$이고 다음이 성립한다.

$$\begin{aligned}
\frac{\sqrt{P_i^{r-1}}(Y_i - \beta_i \widehat{m})}{\sqrt{\beta_i m}} &= \frac{\sqrt{P_i^{r-1}}(Y_i - \beta_i m)}{\sqrt{\beta_i m}} - \frac{\sqrt{P_i^{r-1}}(\beta_i \widehat{m} - \beta_i m)}{\sqrt{\beta_i m}} \\
&= \sqrt{P_i^{r-1}} Z_i^* - \frac{\sqrt{P_i^{r-1}}}{\sqrt{\beta_i}} \cdot \frac{\beta_i}{\beta_+} \cdot \frac{\beta_+ \widehat{m} - \beta_+ m}{\sqrt{m}} + o_p(1) \\
&= \sqrt{P_i^{r-1}}(Z_i^* - \sqrt{P_i} \sum_{k=1}^{n} Z_k^* \sqrt{P_k}) + o_p(1)
\end{aligned}$$

그런데 $\mathbf{Z}^* = \mathbf{D}(\sqrt{1+t\beta_i^{r-1}})\mathbf{Z}$이므로 다음의 결과를 얻을 수 있다.

$$\begin{aligned} L_B &= \sum \frac{P_i^{r-1}(Y_i-\beta_i\widehat{m})^2}{\beta_i m} + o_p(1) \\ &= \{\mathbf{D}(\sqrt{P_i^{r-1}})(\mathbf{I}-\sqrt{\underline{P}}\sqrt{\underline{P}'})\mathbf{Z}^*\}'\{\mathbf{D}(\sqrt{P_i^{r-1}})(\mathbf{I}-\sqrt{\underline{P}}\sqrt{\underline{P}'})\mathbf{Z}^*\} + o_p(1) \\ &= \mathbf{Z}'\mathbf{D}(\sqrt{W_i})(\mathbf{I}-\sqrt{\underline{P}}\sqrt{\underline{P}'})\mathbf{D}(P_i^{r-1})(\mathbf{I}-\sqrt{\underline{P}}\sqrt{\underline{P}'})\mathbf{D}(\sqrt{W_i})\mathbf{Z} + o_p(1) \\ &= \mathbf{Z}'\mathbf{U}\mathbf{Z} + o_p(1) \end{aligned}$$

Bishop, Feinberg와 Holland(1975, p.473)에 의해 정리 1의 통계량은 다음 식 (6)을 만족한다.

$$\mathbf{Z}'\mathbf{U}\mathbf{Z} \quad \xrightarrow{\mathcal{L}} \quad \sum_{i=1}^{n} \varphi_i X_i(1) \tag{6}$$

단, $\{\varphi\}_{i=1}^{n}$은 행렬 U의 고유근,
$Xi(1)$은 $i=1,\cdots,n$에 대해 독립이며 동일한 자유도 1인 카이 제곱 분포를 갖는 확률변수.

그러므로 $m \to \infty$, $c \to 0$이고 $cm^{r-1} = t + o_p(1)$일 때 다음이 성립한다.

$$L_B \quad \xrightarrow[H_a]{\mathcal{L}} \quad \sum_{i=1}^{n} \varphi_i X_i(1)$$

단, $\{\varphi_i\}_{i=1}^{n}$은 행렬 $\mathbf{U}$의 고유근,

$$L_B \quad \xrightarrow[H_0]{\mathcal{L}} \quad \sum_{i=1}^{n} \varphi_{0i} X_i(1)$$

단, $\{\varphi_{0i}\}_{i=1}^{n}$은 행렬 $\mathbf{U}|_{c=0}$의 고유근.

3-2-3. 점근적 상대효율(asymptotic relative efficiency)

통계량 B에 대한 통계량 A의 피트만 점근적 상대효율을 $e_{A|B}$라 할 때 앞 절의 근사분포를 이용하여 통계량 L_B, S_B와 T_B사이의 점근적 상대효율을 구하였다.

정리 2.

$$e_{S_B|L_B} = \frac{[\sum P_i^{r-1}(1-P_i)]^2/(n-1)}{[\sum P_i^{2r-2}(1-P_i)]^2/[\sum P_i^{2r-2} - 2\sum P_i^{2r-1} + (\sum P_i^r)^2]}$$

$$e_{T_B|L_B} = \frac{[\sum P_i^r(1-P_i)]^2/[(\sum P_I^2)^2 + \sum P_i^2 - 2\sum P_i^3]}{[\sum P_i^{2r-2}(1-P_i)]^2/[\sum P_i^{2r-2} - 2\sum P_i^{2r-1} + (\sum P_i^r)^2]}$$

$$e_{S_B|T_B} = \frac{[\sum P_i^{r-1}(1-P_i)]^2/(n-1)}{[\sum P_i^r(1-P_i)]^2/[(\sum P_i^2)^2 + \sum P_i^2 - 2\sum P_i^3]}$$

증명. $\{\varphi_i\}_{i=1}$과 $X_i(1)$가 앞 절의 조건들을 만족할 때 다음 두 식이 성립한다.

$$E(|\varphi_i X_i(1)|^4) = \varphi_i^4 E(|X_i(1)|^4) = 105\varphi_i^4 < \infty, \qquad i = 1, 2, \cdots, n.$$

$$\frac{\sum_{i=1}^n E(|\varphi_i X_i(1)|^4)}{Var^2(\sum_{i=1}^n \varphi_i X_i(1))} = \frac{\sum_{i=1}^n 105\varphi_i^4}{(\sum_{i=1}^n 2\varphi_i^2)^2} = o_p(1)$$

그러므로 L_B의 근사통계량인 $\sum \varphi_i X_i(1)$는 리아푸노프 조건(Lyapounov condition)을 만족하여 H_0과 H_a 아래에서 정규분포를 이룬다. 또한 $tr(A)$는 행렬 A의 대각선항의 합을 의미할 때 $\sum \varphi_i = tr(U)$, $\sum \varphi_i^2 = tr(U^2)$ 임을 이용하여 H_a 아래에서 L_B의 평균과 H_0 아래에서 L_B의 분산을 구하면 다음과 같다.

$$E_{H_a}(L_B) = \sum \{P_i^{r-1} - P_i^r\}\{1 + c(m\beta_i)^{r-1}\}$$
$$\frac{\partial}{\partial c} E_{H_a}(L_B) = \sum (m\beta_i)^{r-1}\{P_i^{r-1} - P_i^r\}$$
$$Var_{H_0}(L_B) = 2\{\sum P^{2r-2} - 2\sum P^{2r-1} + (\sum P^r)^2\}$$

같은 방법으로 S_B와 T_B 통계량의 평균과 분산을 구할 수 있고 피트만의 점근적 상대효율 공식(Kendall과 Stuart(1979, Vol.2 p.284))에 대입하여 위의 상대효율을 구하였다.

여기서 $r = 1$이면 $e_{S_B|L_B} = 1$ 이고 $r = 2$이면 $e_{T_B|L_B} = 1$ 임을 쉽게 알 수 있다. 또한 Kim과 Park(1992)은 $r = 0.5$일 때 $e_{T_B|S_B} \le 1$을 증명함으로써 $\sigma^2 = \mu + c\sqrt{\mu}$를 만족하는 부의 이항분포를 대립가설로 하였을 때의 검정통계량으로는 S_B가 T_B 보다 우수함을 보였다. 통계량 S_B와 L_B의 비교는 임의의 r에 대해 다음의 정리를 유도함으로써 L_B가 S_B보다 검정력이 뛰어남을 증명하였다.

정리 3. $e_{S_B|L_B} \le 1$, 단 $P_1 = P_2 = \cdots = P_n$일 때 등호 성립.

증명. 확률 변수 X에 대해 $E^2(X) \le E(X^2)$이므로

$$e_{S_B|L_B} = \frac{[\sum P_i^{r-1}(1-P_i)]^2/(n-1)}{[\sum P_i^{2r-2}(1-P_i)]^2/[\sum P_i^{2r-2} - 2\sum P_i^{2r-1} + (\sum P_i^r)^2]}$$

$$\leq \frac{1}{[\sum P_i^{2r-2}(1-P_i)]/[\sum P_i^{2r-2} - 2\sum P_i^{2r-1} + (\sum P_i^r)^2]}$$

$$= 1 - \frac{\sum P_i^{2r-1} - (\sum P_i^r)^2}{[\sum P_i^{2r-2}(1-P_i)]} \text{이다.}$$

확률변수 Y가 $Pr(Y = P_i^{r-1}) = P_i$(단 $i = 1, \cdots, n$)를 만족한다고 하자. 그러면 항상 $E^2(Y) \leq E(Y^2)$이므로 $(\sum P_i^r)^2 \leq \sum P_i^{2r-1}$ 이 성립한다. 그러므로 $e_{S_B|L_B} \leq 1$이다.

3-2-4. 모의실험(simulation)

앞에서 논한 세 통계량 L_B, S_B와 T_B의 점근적 상대효율은 점근 분포에 기초한 결과이다. 현실에서 이러한 점근 이론이 얼마나 적용될 수 있는지를 파악하기 위해서는 모의실험이 요구된다.

포트란서브루틴(Fortran Subroutine)인 IMSL(International Mathematical and Statistical Library)을 이용하여 $n = 10$인 경우의 자료집합을 생성하였으며 이를 1000번 되풀이하여 모의실험을 하였다.

$r = 0.5$, 1.0, 2.0, 3.0일 때의 모의실험 결과가 표 1-4 에 나타나 있다.

표 1. $r = 0.5$ 일 때의 모의 실험 결과표

C	검 정 력		
	L_B	S_B	T_B
0.0	0.048	0.058	0.045
0.5	0.091	0.073	0.059
1.0	0.122	0.116	0.076
2.0	0.166	0.140	0.099
3.0	0.219	0.205	0.125
4.0	0.295	0.279	0.170
5.5	0.393	0.390	0.225
7.0	0.483	0.477	0.271
10.0	0.586	0.612	0.359

$\beta = (0.1, 0.2, 0.4, 0.5, 0.5, 1.0, 1.5, 1.5, 2.0, 2.4)$, $m = 100$.

표 2. $r = 1.0$ 일 때의 모의 실험 결과표

C	검 정 력	
	$L_B = S_B$	T_B
0.0	0.043	0.038
0.1	0.094	0.080
0.2	0.122	0.095
0.3	0.181	0.137
0.5	0.240	0.203
0.7	0.381	0.271
1.0	0.463	0.335
1.3	0.602	0.471
1.7	0.728	0.581
2.2	0.804	0.688

$$\beta = (0.1, 0.2, 0.4, 0.5, 0.5, 1.0, 1.5, 1.5, 2.0, 2.4), \quad m = 100.$$

표 3. $r = 2.0$ 일 때의 모의 실험 결과표

C	검 정 력	
	$L_B = T_B$	S_B
0.000	0.051	0.052
0.001	0.095	0.088
0.002	0.145	0.113
0.003	0.187	0.166
0.005	0.290	0.248
0.008	0.414	0.390
0.011	0.501	0.484
0.015	0.642	0.601
0.019	0.697	0.691
0.023	0.775	0.759
0.028	0.818	0.823

$$\beta = (0.1, 0.2, 0.4, 0.5, 0.5, 1.0, 1.5, 1.5, 2.0, 2.4), \quad m = 100.$$

표 4. $r = 3.0$ 일 때의 모의 실험 결과표

C	검 정 력		
	L_B	S_B	T_B
0.000000	0.046	0.050	0.046
0.000005	0.084	0.069	0.082
0.000010	0.138	0.109	0.136
0.000015	0.178	0.126	0.175
0.000025	0.241	0.189	0.242
0.000035	0.340	0.268	0.339
0.000050	0.437	0.357	0.456
0.000065	0.495	0.431	0.497
0.000080	0.558	0.518	0.591
0.000100	0.630	0.589	0.653
0.000120	0.693	0.659	0.715
0.000140	0.726	0.702	0.753

$\beta = (0.1, 0.2, 0.4, 0.5, 0.5, 1.0, 1.5, 1.5, 2.0, 2.4), \quad m = 100.$

표 2의 결과는 $r = 1$일 때는 통계량 L_B가 T_B보다 더 우수하고 표 3은 $r = 2$ 일 때는 L_B가 S_B보다 우수한 통계량임을 보여 주고 있다. $r \geq 2$ 일 때는 T_B가 S_B보다 우수하고 L_B는 c가 0 가까이 있을 때 T_B보다 다소 높은 검정력을 보이고 있으며 c가 큰 값일 때는 T_B보다는 다소 검정력이 떨어지지만 S_B 보다는 우수한 결과를 보이고 있다. $r \leq 1$일 때는 통계량 L_B가 S_B와 T_B보다 우수함을 볼 수 있다. 그러므로 L_B는 분산과 평균의 구조에 상관없이 S_B 보다는 강력한 검정임을 유도하고, T_B에는 최소한 대안으로 사용할 수 있는 검정임을 알 수 있다.

새 통계량 L_B는 S_B와 T_B를 일반화시킨 통계량으로서 S_B와 T_B의 최적성이 적용되지 않는 $r \neq 1$과 $r \neq 2$의 경우와 그리고 r이 알려져 있지 않는 대부분의 현실적 상황에서도 적정한 검정 통계량이 된다고 결론 지을 수 있다.

3-3. 일원 배열 형태의 경우

3-3-1. 검정통계량

각각의 크기가 $n_1, \cdots, n_k$인 k개의 독립된 자료들의 일원배열형태에서 Gart(1983)는 $S_c = \sum_{i=1}^{k} \sum_{j=1}^{n_i} [(Y_{ij} - \overline{Y}_{i+})^2 / \overline{Y}_{i+}]$를 검정통계량으로 제시하였다. $\{Y_{ij}\}_{j=1}^{n_i}{}_{i=1}^{k}$가 독립적인 일원 배열 형태의 자료에서 $Y_{ij} \overset{ind}{\sim} NB(1/c,\ m_i c)$ $(j = 1, \cdots, n_i,\ i = 1, \cdots, k)$이고 검정가설이 $H_0 : c = 0$ 대 $H_a : c > 0$ 일 때 Collings와 Margolin(1985)은 다음 식 (7)이 포아송으로부터 부의 이항분포로의 이탈을 검정하는 국소 최강력 불편 검정임을 밝혔다.

$$T_C = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \widehat{m}_i)^2}{\widehat{m}}, \ \widehat{m} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}}{\sum n_i} = \overline{Y}_{++}, \ \widehat{m}_i = \frac{\sum_{j=1}^{n_i} Y_{ij}}{n_i} \tag{7}$$

Kim과 Park(1992)은 $Y_{ij} \overset{ind}{\sim} NB(m_i/c, \ c)$ $(j = 1, \cdots, n_i, \ i = 1, \cdots, k)$이고 검정가설이 $H_0 : c = 0$ 대 $H_a : c > 0$일 때 다음 식 (8) 의 값이 크면 H_0를 기각하는 것이 점근적으로 국소 최적 검정인 $C(\alpha)$ 검정임을 보였다.

$$K_C = \sum_{i=1}^{k}\sum_{j=1}^{n_i} \frac{(Y_{ij} - \widehat{m}_i)^2 - Y_{ij}}{\widehat{m}_i}, \quad \widehat{m}_i = \frac{\sum_{j=1}^{n_i} Y_{ij}}{n_i} = \overline{Y}_{i+} \tag{8}$$

그런데 $K_C = S_C - \sum_{i=1}^{k} n_i$이므로 K_C는 S_C와 점근적으로 동등하다는 것을 알 수 있다. 또한 원점을 통과하는 회귀 형태의 경우와 마찬가지로 대립가설인 부의 이항분포에서 분산과 평균 사이의 구조가 달라짐에 따라 지역적 최강력 검정이 달라진다.

이선호(1991)는 평균과 분산의 관계가 $\sigma^2 = \mu + c\mu^r$ (단 $c > 0$) 인 일반적인 포아송의 혼합분포를 대립가설로 하였을 때 포아송분포로부터의 이탈을 검색할 수 있는 식 (9)의 통계량 L_C를 유효스코어(efficient score)를 사용하여 유도하였다. 이 통계량을 부의 이항분포를 대립가설로 하였을 때의 검정통계량으로도 사용할 수 있다.

$$L_C = \sum\sum \frac{\widehat{m}_i^{r-2}(Y_{ij} - \widehat{m}_i)^2}{\widehat{m}^{r-1}}, \quad \widehat{m} = \frac{\sum n_i m_i}{\sum n_i} \tag{9}$$

이 통계량 L_C는 기존에 다루었던 통계량 S_C와 T_C의 일반형임을 쉽게 알 수 있다.

3-3-2. 근사적 귀무가설 분포와 대립가설 분포

$E(Y_{ij}) = m_i, (i = 1, \cdots, k, \ j = 1, \cdots, n_i)$이고, $\sigma^2 = \mu + c\mu^r$이라는 전제 아래에서 임의의 i에 대하여 $m_i \to \infty$이고 $c \to 0$ 이면서 $cm_i^{r-1} \to t_i$일 때 Collings와 Margolin(1985)에 의해 다음 식 (10)이 근사적으로 성립한다.

$$Z_{ij} = \frac{Y_{ij} - m_i}{\sqrt{m_i(1 + t_i)}} \quad \dot{\sim} \quad N(0,1) \tag{10}$$

$m_i \to \infty$임에 따라 $\widehat{m}_i/m_i = 1 + o_p(1)$이므로, L_C의 근사분포를 구하면 다음 정리 4와 같다.

정리 4.

$$\begin{aligned} L_C &= \sum\sum \frac{\widehat{m}_i^{r-2}(Y_{ij}-\widehat{m}_i)^2}{\widehat{m}^{r-1}} \\ &= \sum_i (\frac{\widehat{m}_i}{\widehat{m}})^{r-1}(1+t_i)\underline{Z}_i'(\mathbf{I}_i - \frac{1}{n_i}\underline{\mathbf{1}}_i\underline{\mathbf{1}}_i')\underline{Z}_i + o_p(1) \end{aligned}$$

단, $\underline{Z}_i = (Z_{i1}, Z_{i2}, \cdots, Z_{in_i})'$

$\mathbf{I}_i : n_i \times n_i$의 항등행렬

$\underline{\mathbf{1}}_i = (1, 1, \cdots, 1)'_{1\times n_i}$

증명. $m_i \to \infty$임에 따라 $\frac{\widehat{m}_i}{m_i} = 1 + o_p(1)$이므로 다음이 성립한다($i = 1, \cdots, k$).

$$L_C = \sum(\frac{\widehat{m}_i}{\widehat{m}})^{r-1}\left[\sum_{j=1}^{n_i}\frac{(Y_{ij}-\widehat{m}_i)^2}{\widehat{m}_i}\right] = \sum(\frac{\widehat{m}_i}{\widehat{m}})^{r-1}\left[\sum_{i=1}^{n_i}\frac{(Y_{ij}-\widehat{m}_i)^2}{m_i}\right] + o_p(1)$$

그런데, $\underline{Z}_i = (Z_{i1}, Z_{i2}, \cdots, Z_{in_i})'$일 때 정리 1의 증명과 같은 방법으로 다음의 결과를 얻을 수 있다.

$$\sum_{j=1}^{n_i}\frac{(Y_{ij}-\widehat{m}_i)^2}{m_i(1+t_i)} = \underline{Z}_i'(\mathbf{I}_i - \frac{1}{n_i}\underline{\mathbf{1}}_i\underline{\mathbf{1}}_i')\underline{Z}_i$$

또한 각각의 i가 서로 독립이므로

$$L_C = \sum(\frac{\widehat{m}_i}{\widehat{m}})^{r-1}(1+t_i)\underline{Z}_i'(\mathbf{I}_i - \frac{1}{n_i}\underline{\mathbf{1}}_i\underline{\mathbf{1}}_i')\underline{Z}_i + o_p(1)$$

이다.

Bishop, Fienberg와 Holand(1975,p.473)에 의해 다음 식 (11)을 얻을 수 있다.

$$\underline{Z}_i'(\mathbf{I}_i - \frac{1}{n_i}\underline{\mathbf{1}}_i\underline{\mathbf{1}}_i')\underline{Z}_i \quad \xrightarrow{\mathcal{L}} \quad \sum_{i=1}^{n_i}\eta_i X_i(1) \tag{11}$$

단 $\{\eta_i\}_{i=1}^{n_i}$는 $(\mathbf{I}_i - \frac{1}{n_i}\underline{\mathbf{1}}_i\underline{\mathbf{1}}_i')$의 고유근

그런데 $(\mathbf{I}_i - \frac{1}{n_i}\underline{\mathbf{1}}_i\underline{\mathbf{1}}_i')$는 대칭이며 멱등(idempotent)인 행렬이고 위수(order)는 $(n_i - 1)$이므로 식 (12)을 얻을 수 있다.

$$\underline{Z}_i'(\mathbf{I}_i - \frac{1}{n_i}\underline{\mathbf{1}}_i\underline{\mathbf{1}}_i')\underline{Z}_i \quad \xrightarrow{\mathcal{L}} \quad X(n_i - 1) \tag{12}$$

이를 이용하여 H_a과 H_0 아래에서의 근사분포를 구하였다.

$$L_C \quad \xrightarrow[H_a]{\mathcal{L}} \quad \sum(\frac{\widehat{m}_i}{\widehat{m}})^{r-1}(1+t_i)X(n_i-1) \tag{13}$$

$$\xrightarrow[H_0]{\mathcal{L}} \quad \sum(\frac{\widehat{m}_i}{\widehat{m}})^{r-1}X(n_i-1) \tag{14}$$

3-3-3. 점근적 상대효율(asymptotic relative efficiency)

대표본일 때 S_C, T_C와 L_C의 효율성을 비교하기 위해 다음과 같이 피트만의 점근적 상대효율을 구하였다.

정리 5.

$$\begin{aligned}
e_{S_C|L_C} &= \frac{\{\sum m_i^{r-1}(n_i-1)\}^2/(n-k)}{\sum m_i^{2r-2}(n_i-1)} \\
e_{T_C|L_C} &= \frac{\{\sum m_i^{r}(n_i-1)\}^2/\sum m_i^2(n_i-1)}{\sum m_i^{2r-2}(n_i-1)} \\
e_{S_C|T_C} &= \frac{\{\sum m_i^{r-1}(n_i-1)\}^2/(n-k)}{\{\sum m_i^{r}(n_i-1)\}^2/\sum m_i^2(n_i-1)}, \qquad n=\sum n_i
\end{aligned}$$

증명. $m_i \to \infty$ 일 때 $\widehat{m}_i/m_i = 1+o_p(1)$이므로 식 (13)과 (14)의 H_a과 H_0 아래에서의 L_C의 근사통계량은 정리 2와 마찬가지로 리아푸노프 조건(Lyapounov condition)을 만족하므로 정규분포를 가정할 수 있다. 또한 L_C의 근사통계량을 이용하여 H_a 아래에서 L_C의 평균과 H_0 아래에서 L_C의 분산을 구하면 다음과 같다.

$$\begin{aligned}
E_{H_a}(L_C) &= \sum(\frac{m_i}{m})^{r-1}(1+cm_i^{r-1})(n_i-1) \\
\frac{\partial}{\partial c}E_{H_a}(L_C) &= \sum \frac{m_i^{2r-2}}{m^{r-1}}(n_i-1) \\
Var_{H_0}(L_C) &= 2\sum(\frac{m_i}{m})^{2r-2}(n_i-1)
\end{aligned}$$

같은 방법으로 S_C와 T_C 통계량의 평균과 분산을 구할 수 있고 피트만의 점근적 상대효율 공식(Kendall과 Stuart(1979, Vol.2 p.284))에 대입하여 위의 상대효율을 구하였다.

L_C는 S_C와 T_C를 동시에 일반화한 검정통계량이기 때문에 $r=1$이면 $e_{S_C|L_C}=1$이고 $r=2$이면 $e_{T_C|L_C}=1$ 임을 쉽게 알 수 있다. 또한 Kim과 Park (1992)은 $r=0.5$일 때 $e_{T_C|S_C} \le 1$을 증명함으로써 $\sigma^2=\mu+c\sqrt{\mu}$ 를 만족하는 부의 이항분포를 대립가설로 하였을

때의 검정통계량으로는 S_C가 T_C 보다 더 우수함을 보였다. 통계량 S_C와 L_C의 비교는 임의의 r에 대해 다음의 정리를 유도함으로써 L_C가 S_C보다 검정력이 뛰어남을 증명하였다.

정리 6. $e_{S_C|L_C} \leq 1$, 단 $m_1 = m_2 = \cdots = m_n$일 때 등호 성립.

증명. 임의의 $i(i = 1, \cdots, k)$ 에 대해 $t_i = (n_i - 1)/(n-k)$ 라 하자. 그러면 $t_i \geq 0$이고 $\sum t_i = 0$이다. 이산확률변수 X가 $m_1, \cdots, m_k$의 값을 취할 수 있고 확률이 각각 $t_1, \cdots, t_k$일 때 $e_{S_C|L_C} = E^2(X^{r-1})/E(X^{2r-2})$이다. 그런데 임의의 n에 대해 $E^2(X^n) \leq E(X^{2n})$이 성립하므로 $e_{S_C|L_C} \leq 1$ 이다.

3-3-4. 모의실험

모의실험을 통하여 일원배열 형태의 소표본에서의 검정력을 비교하였다. 분산과 평균의 구조 $\sigma^2 = \mu + c\mu^r$ 에서 r = 0.5, 1.0, 2.0, 3.0 일 때에 대해 모의실험을 하고 다음 표 5-8을 얻었다.

표 5. $r = 0.5$ 일 때의 모의 실험 결과표

C	검 정 력		
	L_B	S_B	T_B
0.0	0.053	0.055	0.051
0.3	0.084	0.077	0.062
0.6	0.113	0.108	0.079
0.9	0.157	0.122	0.087
1.2	0.221	0.187	0.087
1.8	0.307	0.289	0.124
2.4	0.383	0.371	0.180
3.0	0.485	0.467	0.189
3.6	0.591	0.562	0.232
4.5	0.655	0.628	0.294
5.4	0.735	0.725	0.366

$n = (10, 10),\ m = (10, 100)$

표 6. $r = 1.0$ 일 때의 모의 실험 결과표

C	검 정 력	
	$L_C = S_C$	T_C
0.00	0.050	0.067
0.08	0.099	0.089
0.16	0.115	0.095
0.24	0.188	0.136
0.32	0.235	0.188
0.48	0.367	0.293
0.64	0.482	0.353
0.80	0.572	0.422
1.04	0.707	0.562
1.28	0.804	0.645

$n = (10, 10),\ m = (10, 100)$

표 7. $r = 2.0$ 일 때의 모의 실험 결과표

C	검 정 력	
	$L_C = T_C$	S_C
0.000	0.051	0.054
0.001	0.078	0.066
0.002	0.120	0.087
0.003	0.153	0.124
0.004	0.225	0.164
0.006	0.312	0.251
0.008	0.422	0.345
0.010	0.519	0.423
0.013	0.616	0.560
0.016	0.684	0.595
0.019	0.800	0.743

$n = (10, 10),\ m = (10, 100)$

표 8. $r = 3.0$ 일 때의 모의 실험 결과표

C	검 정 력		
	L_C	S_C	T_C
0.00000	0.052	0.048	0.056
0.00001	0.085	0.065	0.086
0.00002	0.098	0.082	0.097
0.00003	0.159	0.128	0.166
0.00004	0.201	0.141	0.206
0.00006	0.288	0.235	0.296
0.00008	0.421	0.315	0.422
0.00010	0.474	0.356	0.473
0.00012	0.578	0.455	0.583
0.00015	0.664	0.551	0.665
0.00018	0.723	0.621	0.717
0.00022	0.812	0.696	0.813

$n = (10, 10),\ m = (10, 100)$

일원 배열 형태일 때의 모의실험으로 앞에서 다루었던 원점을 통과하는 회귀 형태의 실험과 같은 결과를 얻었다. S_C, T_C 와 L_C 를 비교했을 때 통계량 S_C 나 T_C 는 통계량의 최적성이 적용되는 r 값에 따라 우수하기도 하고, 검정력이 떨어지기도 하였다. 그러나 통계량 L_C 는 r 의 값에 상관없이, 즉 평균과 분산 사이의 구조에 상관없이 검정력이 우수하였다.

4. 결론

통계량 L 은 분산과 평균이 $\sigma^2 = \mu + c\mu^r$ 을 만족하는 포아송의 혼합분포를 대립가설로 하였을 때 포아송분포로부터의 이탈을 검색할 수 있도록 유도된 것이다. 포아송 혼합분포의 한 형태인 부의 이항분포로의 이탈을 검색하는 통계량들과 비교한 결과 L 은 분산과 평균의 구조에서 r 값에 상관없이 우수한 것임을 밝혔다.

또한 피트만의 점근적 상대효율을 비교하여 $e_{S|L} \leq 1$ 임을 증명하여 L 통계량이 항상 S 통계량보다 우수함을 입증하였다.

L 통계량과 T 통계량을 피트만 점근적 상대효율을 통하여 비교하여 항상 L 이 T 보다 우수함을 증명하지는 못했지만 여러 가지 값을 대입하여 본 결과, $e_{T|L} \leq 1$ 의 결과를 얻었다. 그러므로 일반적으로 $e_{T|L} \leq 1$ 이 성립함을 증명하는 것이 추후 과제라 하겠다.

모의실험을 통하여서도 S와 T통계량 각각의 최적성이 적용되는 r값에 따라 검정력에 많은 차이를 보이나 L은 r이 어떤 값을 갖더라도 항상 S보다는 우수하고 T에는 최소한 대안으로 사용할 수 있는 우수한 통계량임을 밝혔다.

5. 감사의 글

본 논문은 저자의 박사학위 논문의 일부이며, 논문의 지도교수인 연세대학교의 김병수박사께 감사를 드린다. 아울러 익명의 두 분 심사위원의 비평과 논평이 본 논문 올 개선하는데 많은 도움이 되었음을 밝힌다.

참고문헌

(1) 이 선 호 (1991). 포아송분포에서의 과산포 검정, 박사학위 논문, 연세대학교 대학원.

(2) Bishop, Y. M., Fienberg, S. E. and Holland, P. W.(1975). *Discrete Multivariate Analysis*, MIT Press, Cambridge.

(3) Collings, Bruce J. and Margolin, Barry H.(1985). Testing goodness of fit for the Poisson assumption when observations are not identically distributed. *Journal of the American Statistical Association*, 80, 411-18.

(4) Cox, D. R. and Hinkley, D. V.(1974). *Theoratical Statistics.* Chapman and Hall, London.

(5) Dean, C.(1992). Testing for overdispersion in Poisson and binomial regression models. *Journal of the American Statistical Association*, 87, 451-457.

(6) Dean, C. and Lawless, J. F.(1989). Tests for detecting overdispersion in Poisson regression models. *Journal of the American Statistical Association*, 84, 467-72.

(7) Fisher, R. A., Thorton, H. G., and Mackenzie, W. A. (1922). The accuracy of the plating method of estimating the denstiy of bacterial populations. *Annals of Applied Biology*, 9, 325-359.

(8) Gart, J. J. (1983). The analysis of Poisson regression with an application in virology. *Biometrics*, 70, 269-74.

(9) Hausman, J., Hall, B. and Griliches, Z.(1984). Econometric models for count data with an application to the patents - R & D relationship. *Econometrica*, 52, 909-937.

(10) IMSL Co. (1989). *IMSL:Fortran Subroutines for Statistical analysis*, IMSL Co., Houston.

(11) Kendall, M. and Stuart, A. (1979). *The advanced theory of statistics* Vol.2, Charles Griffin & Company Limited, London.

(12) Kim, B. S. and Park, C.(1992). Some remarks on testing goodness of fit for the Poisson assumption. *Communications in Statistics*, 21, 979-96.

(13) Margolin, B. H., Kaplan, N. and Zeiger, E. (1981). Statistical analysis of the Ames salmonella / microsome test. *Proceedings of the National Academy of Sciences*, 78, 3779-83.

(14) Moran, P. A. P.(1970). On asymptotically optimal test of composite hypotheses. *Biometrika*, 57, 47-55.

(15) Neyman, J.(1959). Optimal asymptotic tests of composite hypotheses. *Probability and Statistics: The Herald Cramer Volume, ed. Ulf Grenander*, John Wiley and Sons, New York.

(16) Paul, S. R. and Plackett, R. L. (1978). Inference sensitivity for Poisson mixtures. *Biometrika*, 65, 591-602.

(17) Potthoff, R. F. and Whittinghill, M. (1969). Testing for homogeneity II: The Poisson distributions. *Biometrika*, 53, 183-90.

(18) Rao, C.R.(1952). *Advanced Statistical Methods in Biometric Research.* John Wiley, New York.

On the Extension of Test Statistics for Detecting Negative Binomial Departures from the Poisson Assumption

Sunho Lee[2]

ABSTRACT

Collings and Margolin(1985) developed a locally most powerful unbiased test for detecting a negative binomial departure from a Poisson model, when the variance was a quadratic function of the mean. Kim and Park(1992) obtained a locally optimal test, when the variance was a linear function of the mean. Kim and Park showed that different mean-variance structures of a negative binomial derived different optimal test.

These results are unified and extended by Lee(1991). Lee develops a locally most powerful unbiased test for detecting overdispersion in the Poisson model against a mixture of Poisson with the general mean-variance structure, $\sigma^2 = \mu + c\mu^r$.

Superiority of Lee's test is shown by the comparison of Pitman's asymptotic relative efficiencies and Monte Carlo simulation studies.

KEYWORDS: Poisson mixture models, overdispersion in Poisson models, mean-variance structures of negative binomial model, hyper Poisson variation.

[2]Department of Mathematics, Sejong University, Seoul, 133-747, Korea.