응용통계연구 제 10권 1호
1997년 3월 pp. 105-119

일원배열 가산자료에서의 처리효과 비교[1)]

이 선 호[2)]

요 약

일원배열형태의 가산 자료집합에서 각 군의 평균을 이용하여 처리효과를 비교할 수 있다. Barnwal과 Paul(1988)은 각 군의 산포모수가 같다는 가정 아래에서 처리에 따른 차이를 검정하는 우도검정통계량과 $C(\alpha)$ 통계량을 유도하였는데 본 연구에서는 이러한 가정이 만족되지 않아도 검정할 수 있도록 통계량을 일반화하였다. 또한 음이항분포 대신 Efron(1986)의 이중지수계 포아송모형을 도입하여 새로운 통계량을 제시하였다. 모의실험을 통해 이중지수계 포아송모형으로부터 유도된 $C(\alpha)$ 통계량이 어느 경우에나 적합함을 밝혔다.

1. 서론

의학 연구의 많은 실험결과가 가산자료(count data)의 형태로 얻어지고 있고 각 군의 평균을 비교하여 처리효과를 분석하기도 한다. 쥐를 이용한 생검실험(bioassay)에서 모체내 태아 사망수를 관측하여 비교함으로써 약물이나 독성 물질의 반응을 비교하는 검사(dominant lethal test)(Haseman과 Soares, 1976; McCaughran과 Arnold, 1976; Lockart et al., 1992)를 예로 들 수 있다. 또한 살모넬라를 이용한 돌연변이 검사 자료(Margolin et al., 1981), 유전학에서 노르웨이산 양의 동복자 수(litter size)의 조사(Olesen et al., 1994)등 여러 방면에서 가산자료들을 접할 수 있고 이러한 자료의 분석 모형으로 포아송분포(Poisson distribution)와 이의 과산포 형태인 음이항분포(negative binomial distribution)가 많이 쓰인다(Margolin et al., 1981; Foulley et al., 1987; Tempelman과 Gianola, 1996).

약물이나 식품 첨가제 등의 효과를 알아보기 위한 모체내 태아 사망수 비교실험에서 McCaughran과 Arnold(1976)는 처리군들의 태아 사망수는 음이항분포를 따른다는 것을 보였고, 각 군의 평균이 처리효과를 알아보기 위한 좋은 척도임을 밝혔다. 이를 바탕으로 Barnwal과 Paul(1988)은 각 처리군의 산포모수(dispersion parameter)가 같을 때 군의 평균을 비교함으로써 처리효과를 분석할 수 있는 우도비 검정법에 따른 통계량과 Neyman(1959)의 $C(\alpha)$ 통계량을 유도하였다.

우도비 검정통계량은 모수의 추정이 귀무가설과 대립가설에서 모두 필요하다는 단점이 있다.

1) 이 연구는 1995년도 세종대학교 대양학술연구비에 의하여 연구되었슴.
2) (143-747) 서울시 광진구 군자동 98, 세종대학교 수학과 조교수.

반면에 $C(\alpha)$통계량은 여러 가지 장점이 있는데 첫째는 통계량이 미리 정해진 유의수준을 대체적으로 잘 지킨다는 것이고, 둘째는 귀무가설 아래에서의 모수 추정만을 필요로 한다는 것이다. 또한 대개의 경우 통계량이 간단하다는 장점이 있다.

본 논문의 주목적은 각 군의 산포모수가 같다는 가정이 성립하지 않아도 처리효과를 분석하는 것이다. 실질적으로 분포의 평균μ, 분산 σ^2과 산포모수 c사이의 관계를 $\sigma^2=\mu+c\mu^2$로 가정하면 c값의 조그만 변동도 분산에는 큰 영향을 미칠 수 있으므로 공통의 산포모수라는 제약 대신 각 군의 산포모수를 각각 추정하여 처리 효과를 비교하는 통계량을 유도하였다. 또다른 목적은 포아송분포의 모수가 반드시 감마분포(gamma distribution)를 따라야 하는 음이항분포 대신 포아송분포를 포함하고 과산포(overdispersion)뿐만 아니라 적산포(underdispersion)도 나타낼 수 있는 Efron(1986)의 이중지수계 포아송분포(double exponential Poisson family)를 도입하여 모형을 좀 더 일반화하는 것이다.

제2절에서는 음이항분포 모형에서의 모수의 추정과 각 군의 평균을 비교하는 통계량을 유도하였다. 또한 제3절에서는 이중지수계 포아송모형의 가정 아래에서 모수를 추정하고 통계량을 유도하였다. 이렇게 서로 다른 모형 아래에서 유도한 통계량을 모의실험을 통하여 제4절에서 비교하고 제5절에 예제를 제시하였다.

2. 음이항분포 모형

McCaughran과 Arnold(1976)는 태아 사망수의 모형 설정에서 모체 내의 태아 사망수는 포아송분포를 따르지만 각 모체의 사망률은 감마분포를 따른다고 가정하여 포아송의 감마혼합인 음이항분포(negative binomial distribution) 사용을 제시하였다.

음이항분포는 나타낼 수 있는 방법이 많은데 Collings와 Margolin(1985)처럼 확률변수 X의 평균 μ와 분산 σ^2 사이에 $\sigma^2=\mu+c\mu^2$의 이차함수를 만족하는 아래의 식 (1)을 X의 확률밀도함수로 정의하고 이를 $X\sim NB(\mu, c)$로 표기하였다.

$$\Pr(X=x)=\frac{\Gamma(X+c^{-1})}{x!\,\Gamma(c^{-1})}\left(\frac{c\mu}{1+c\mu}\right)^{x}\left(\frac{1}{1+c\mu}\right)^{1/c} \tag{1}$$
$$x=0,1,\cdots,\infty,\quad \mu>0,\quad c>0$$

식 (1)은 c→0일 때 점근분포가 평균이 μ인 포아송분포로 수렴함을 알 수 있다.

각 크기가 $n_1,\cdots,n_k$인 k개 군의 일원배열형태의 독립된 자료들에서 확률변수 X_{ij} $(j=1,\cdots,n_i,\ i=1,\cdots,k)$는 $X_{ij}\sim NB(\mu_i, c_i)$로 나타낼 수 있다. 또한 K는 모수와 관계없는 관찰값들만의 상수라 할 때 로그우도함수 l_{NB}은 다음과 같다.

$$\begin{aligned} &l_{NB}(\ \mu=(\mu_1,\cdots,\mu_k),\ c=(c_1,\cdots,\ c_k)) \\ &=K+\sum_{i=1}^{k}\sum_{j=1}^{n_i}[X_{ij}\log\mu_i-(X_{ij}+\ c_i^{-1})\log(1+c_i\mu_i)+\sum_{l=1}^{X_{ij}}\log(1+c_i(l-1))] \end{aligned} \tag{2}$$

k개 군의 처리효과를 비교하기 위해서는 $\mu_1=\cdots=\mu_k=\mu$의 가설을 다룰 수 있지만 $\mu_i=\mu+\phi_i(i=1,\cdots,k,\ \phi_k=0)$이고 $\boldsymbol{\phi}=(\phi_1,\cdots,\phi_{k-1})'$라 하면 본 논문에서 다룰 가설은 H_0: $\boldsymbol{\phi}=\mathbf{0}$와 H_1: $\boldsymbol{\phi}\neq\mathbf{0}$로 간단히 표현할 수 있다.

2.1 모수의 추정

$i=1,\cdots,k$일 때 H_1의 가정 아래에서 μ_i의 최대우도추정량은 $\hat{\mu}=\mu_{1i}=\sum_{j=1}^{n_i}X_{ij}/n_i=\overline{X}_{i.}$이고, c_i의 최대우도추정량, c_{1i}는 아래의 식을 만족하는 근 c_i로서 반드시 양수여야 하겠다.

$$\log(1+c_i\,\overline{X}_{i.})-\sum_{j=1}^{n_i}\sum_{l=1}^{X_{ij}}\frac{c_i}{1+c_i(l-1)}\ =\ 0$$

Barnwal과 Paul(1988)은 산포모수가 공통일 때, 즉 $c_1=\cdots=c_k=c$일 때 c의 최대우도추정량 c_1은 다음을 만족하는 양의 근, c임을 보였다.

$$\sum_{i=1}^{k}n_i\left[\log(1+c\,\overline{X}_{i.})-\sum_{j=1}^{n_i}\sum_{l=1}^{X_{ij}}\frac{c}{1+c(l-1)}\right]\ =\ 0$$

H_0이 사실일 때 μ의 최대우도추정량 μ_0은 $\mu_0=\overline{X}$를 만족하고 c_i와 공통 산포모수 c의 추정량, c_{0i}, c_0은 위와 같은 방법으로 구할 수 있다.

Barnwal과 Paul은 음이항분포에서 산포모수의 최대우도추정량을 구하기가 쉽지 않음을 언급하고 효율성은 조금 떨어지지만 반복법을 사용하지 않고도 구할 수 있는 적률추정량 사용을 제시하였는데 다음과 같이 H_0과 H_1에서의 산포모수의 적률추정량을 구할 수 있다.

$$c_1=\frac{\sum n_i(S_i^2-\overline{X}_{i.})}{\sum n_i\,\overline{X}_{i.}^{\,2}},\qquad c_{1i}=\frac{(S_i^2-\overline{X}_{i.})}{\overline{X}_{i.}^{\,2}}$$

$$c_0=\frac{(S^2-\overline{X})}{\overline{X}^2},\qquad c_{0i}=\frac{S_i^2-\overline{X}}{\overline{X}^2}$$

$$\text{단,}\quad S^2=\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\overline{X})^2}{\sum_{i=1}^{k}n_i-1},\quad S_i^2=\sum_{j=1}^{n_i}\frac{(X_{ij}-\overline{X}_i)^2}{n_i-1}$$

2.2 우도비 검정법(likelihood ratio test)

식 (2)의 로그우도함수에 μ와 c의 H_0과 H_1에서의 추정량, μ_0와 c_0, μ_1과 c_1를 이용하여 유도한 우도비 검정은 다음의 검정통계량이 되며 이는 자유도가 $k-1$인 카이제곱분포로 수렴한다.

$$NLR = 2[l_{NB}(\mu_1, c_1) - l_{NB}(\mu_o, c_0)]$$
$$= 2\sum_{i=1}^{k}\left[\frac{n_i}{c_{0i}}\log(1+c_{0i}\mu_0) - \frac{n_i}{c_{1i}}\log(1+c_{1i}\mu_{1i}) + \sum_{j=1}^{n_i}\left(X_{ij}\log\frac{\mu_{1i}(1+c_{0i}\mu_0)}{\mu_0(1+c_{1i}\mu_{1i})} + \sum_{l=1}^{X_{ij}}\log\frac{1+c_{1i}(l-1)}{1+c_{0i}(l-1)}\right)\right]$$

각 군 사이에 공통의 산포모수를 가정할 때 통계량 NLR에 공통 산포모수 c의 H_0과 H_1에서의 추정량 c_0, c_1을 대입하면 Barnwal과 Paul의 통계량과 같게 된다.

2.3 $C(\alpha)$ 검정법

$X_{ij} \sim N(\mu_i, c_i)$ $(j=1,\cdots,n_i, i=1,\cdots,k)$에서 $\Phi=0$의 검정을 위해 Tarone(1985)과 같은 방법으로 장애모수(nuiance parameter)들을 벡터 $\Lambda=(\lambda_1,\cdots,\lambda_{k+1})'=(\mu, c_1,\cdots,c_k)'$로 놓고 $\psi_i=\left[\frac{dl_{NB}}{d\phi_i}\right]_{\Phi=0}$ $(i=1,\cdots,k-1)$, $\gamma_j=\left[\frac{dl_{NB}}{d\lambda_j}\right]_{\Phi=0}$ $(j=1,\cdots,k+1)$이라 정의하자. $\hat{\Lambda}$가 H_0에서 Λ의 $\sqrt{n}$ 일치추정량($\sqrt{n}$ consistent estimator)이라면 β_{ij}가 γ_j에서의 ψ_i의 부분회귀계수일 때 $C(\alpha)$통계량은 $S_i(\hat{\Lambda})=\psi_i(\hat{\Lambda})-\sum_{j=1}^{k+1}\beta_{ij}\gamma_j(\hat{\Lambda})$에 기초한다(Barnwal과 Paul(1988)).

$S(\Lambda)=(S_1(\Lambda),\cdots,S_{k-1}(\Lambda))'$의 분산-공분산벡터 $Cov(S(\Lambda))$는 아래의 요인을 갖는 벡터 D, A, B가 존재할 때 $D-AB^{-1}A'$의 값을 갖는다.

$$D_{it}=E\left[-\frac{d^2l}{d\phi_i d\phi_t}\right]_{\Phi=0} \qquad i,t=1,\cdots,k-1$$
$$A_{ip}=E\left[-\frac{d^2l}{d\phi_i d\lambda_p}\right]_{\Phi=0} \qquad i=1,\cdots,k-1 \qquad p=1,\cdots,k+1$$
$$B_{ps}=E\left[-\frac{d^2l}{d\lambda_p d\lambda_s}\right]_{\Phi=0} \qquad p,s=1,\cdots,k+1$$

S, A, B, D에 $\hat{\Lambda}$을 대입하여 $C(\alpha)$통계량 $S'(D-AB^{-1}A')^{-1}S$을 얻게 되며 이는 자유도가 $k-1$인 카이제곱분포로 수렴한다(Neyman, 1959; Moran, 1970).

$\hat{\Lambda}$의 최대우도추정량 대신 역시 $\sqrt{n}$ 일치추정량인 적률추정량을 사용하면 $S_i(\hat{\Lambda})$는 $S_i(\hat{\Lambda})=\psi_i(\hat{\Lambda})$로 축소되어 스코어(score) 통계량과 같아진다(Paul과 Islam(1995)). 이로부터 다음의 $C(\alpha)$검정통계량 NCA를 유도하였다.

$$NCA = \sum_{i=1}^{k} \frac{n_i(\overline{X}_{i.} - \overline{X})^2}{\overline{X}(1+c_{0i}\overline{X})}$$

여기서 각 군의 산포모수의 추정량 $c_{01}, \cdots, c_{0k}$ 대신 공통의 산포모수의 추정량 c_0를 대입하면 Barnwal과 Paul(1988)이 유도했던 통계량과 같아진다.

3. 이중지수계 포아송모형

일모수지수계(one parameter exponential family)에 속하는 분포들 중 이항분포나 포아송분포의 분산은 평균과 종속적인 관계에 있다. 분산과 평균사이에 독립성을 부여하기 위해 Efron(1986)은 지수계에 새로운 산포모수 θ를 삽입하는 이중지수계(double exponential family)를 개발하였다. 평균이 μ이고 분산이 μ/θ인 포아송분포의 이중지수계 모형은 아래와 같고, X~DEP(μ, θ)로 표현하겠다.

$$f(\mu, \theta; x) = \theta^{\frac{1}{2}} e^{-\theta\mu}\left(\frac{e^{-x}x^x}{x!}\right)\left(\frac{e\mu}{x}\right)^{\theta x}, \quad x=0,1,2,\cdots, \quad \mu>0, \quad \theta>0$$

X~DEP(μ, θ)에서 $\theta=1$일 때 확률변수 X는 포아송분포를 따르며 $0<\theta<1$일 때는 과산포, $\theta>1$일 때는 적산포 현상을 보인다.

모체내 태아 사망수에 대한 분석에서 각 모체의 사망률이 감마분포를 따른다고 가정하고 음이항분포를 많이 쓰지만 실질적으로 사망률이 그 가정을 만족하는 지에 대한 아무런 이론적 뒷받침이 없다. 그러므로 이러한 자료집합을 위해 포아송모형의 과산포와 적산포를 모두 포함하고 최대우도추정량을 구하기 쉬운 이중지수계 포아송모형을 사용하게 되었다.

각 크기가 $n_1, \cdots, n_k$인 k군의 독립된 자료들의 일원배열형태에서 평균이 $\mu_i = \mu + \phi_i$, 분산이 μ_i/θ_i의 이중 지수계 포아송분포의 확률변수 $X_{ij}(j=1,\cdots,n_i, i=1,\cdots,k)$는 $X_{ij} \sim DEP(\mu_i, \theta_i)$로 나타낼 수 있고, $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_k)'$이며 K는 모수와 무관한 상수일 때 다음의 로그우도함수를 갖는다.

$$l_D(\boldsymbol{\mu}, \boldsymbol{\theta}) = K + \sum_{i=1}^{k}\left[\frac{n_i}{2}\log\theta_i - n_i\theta_i\mu_i + \sum_{j=1}^{n_i}\theta_i X_{ij}(1+\log\frac{\mu_i}{X_{ij}})\right] \qquad (3)$$

k개 군의 처리효과를 비교하기 위한 가설은 $H_0: \boldsymbol{\phi} = \mathbf{0}$와 $H_1: \boldsymbol{\phi} \neq \mathbf{0}$로 앞 절과 같다.

3.1 모수의 추정

이중지수계 포아송분포의 H_0에서의 모수의 최대우도추정량은 아래의 식 (4)와 (5)를 만족하는 $\boldsymbol{\mu}_0 = (\mu_0, \cdots, \mu_0)$, $\boldsymbol{\theta}_0 = (\theta_{01}, \cdots, \theta_{0k})$로 반복적인 방법을 써야만 얻을 수 있지만 음이항분포의 경우에 비해 쉽게 수렴값을 구할 수 있다.

$$\sum_i \sum_j (\theta_{0i} - \frac{\theta_{0i} x_{ij}}{\mu_0}) = 0 \tag{4}$$

$$\sum_j (\frac{1}{2\theta_{0i}} - \mu_0 + x_{ij}(1 + \log \frac{\mu_0}{x_{ij}})) = 0 \tag{5}$$

H_1에서의 모수의 추정량 $\mu_1 = (\mu_{11}, \cdots, \mu_{1k})$, $\theta_1 = (\theta_{11}, \cdots, \theta_{1k})$는 다음과 같이 직접 대입하여 구할 수 있다.

$$\mu_{1i} = \overline{X}_{i.}$$

$$\theta_{1i} = \frac{n_i}{2\sum_{j=1}^{n_i} X_{ij} \log \frac{X_{ij}}{\mu_{1i}}} \qquad i = 1, \cdots, k$$

또한 각 군의 산포모수가 모두 같을 때 H_0과 H_1에서의 최대우도추정량 μ_0, $\theta_0 = (\theta_0, \cdots, \theta_0)$와 μ_1, $\theta_1 = (\theta_1, \cdots, \theta_1)$은 다음을 만족한다.

$$\mu_0 = \overline{X}$$

$$\theta_0 = \frac{\sum_{i=1}^{k} n_i}{2\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij} \log \frac{X_{ij}}{\mu_0}}$$

$$\mu_{1i} = \overline{X}_{i.} \qquad i = 1, \cdots, k$$

$$\theta_1 = \frac{\sum_{i=1}^{k} n_i}{2\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij} \log \frac{X_{ij}}{\mu_{1i}}}$$

3.2 우도비 검정법

각 군의 평균간에 차이가 있는지에 대한 우도비 검정은 제2.2절과 같은 방법으로 식 (3)의 로그우도함수 l_D에 H_0과 H_1에서의 μ, θ의 추정량을 대입하여 유도한 아래의 통계량을 이용할 수 있다.

$$\begin{aligned} DLR &= 2[l_D(\mu_1, \theta_1) - l_D(\mu_0, \theta_0)] \\ &= \sum_{i=1}^{k} \Big[n_i (\log \frac{\theta_{1i}}{\theta_{0i}} - 2\theta_{1i}\mu_{1i} + 2\theta_{0i}\mu_0) \\ &\qquad + 2\sum_{j=1}^{n_i} X_{ij} (\theta_{1i}(1 + \log \frac{\mu_{1i}}{X_{ij}}) - \theta_{0i}(1 + \log \frac{\mu_0}{X_{ij}})) \Big] \end{aligned}$$

각 군의 산포모수가 모두 같은 경우 검정통계량이 다음과 같이 간단해 짐을 쉽게 보일 수 있다.

$$DLR=\Sigma n_i \log \frac{\theta_1}{\theta_0}$$

3.3 C(α)검정법

모수벡터 $\boldsymbol{\phi}=(\phi_1,\phi_2,\cdots,\phi_{k-1})'$, 장애모수벡터 $\boldsymbol{\Lambda}=(\lambda_1,\cdots,\lambda_{k+1})'=(\mu,\theta_1,\cdots,\theta_k)'$로부터 출발하여 이중지수계 포아송분포의 로그우도함수 l_D에 대하여 $\psi_i=\left[\frac{dl_D}{d\phi_i}\right]_{\boldsymbol{\phi}=0}$ $(i=1,\cdots,k-1)$, $\gamma_j=\left[\frac{dl_D}{d\lambda_j}\right]_{\boldsymbol{\phi}=0}$ $(j=1,\cdots,k+1)$를 정의하고 제2.3절의 전개방법을 l_D에 적용하여 벡터 D, A, B를 얻을 수 있다. 또한 $C(\alpha)$ 통계량의 기초가 되는 $S_i(\Lambda)=\psi_i(\Lambda)-\sum_{j=1}^{k+1}\beta_{ij}\gamma_j(\Lambda)$의 γ_j에서의 ψ_i의 부분회귀계수 β_{ij}의 추정량은 아래와 같이 구할 수 있고 이로부터 $S_i(\Lambda)$를 계산하였다.

$$\boldsymbol{\beta}=\begin{pmatrix}\beta_{11} & \beta_{12} & \cdots & \beta_{1(k+1)}\\ \vdots & & & \\ \beta_{(k-1)1} & \beta_{(k-1)2} & \cdots & \beta_{(k-1)(k+1)}\end{pmatrix}=AB^{-1}=\begin{pmatrix}\frac{n_1\theta_1}{\sum n_i\theta_i} & 0 & \cdots & 0\\ & \vdots & & \\ \frac{n_{k-1}\theta_{k-1}}{\sum n_i\theta_i} & 0 & \cdots & 0\end{pmatrix}$$

$$\begin{aligned}S_i(\Lambda) &= \psi_i(\Lambda)-\frac{n_i\theta_i}{\sum n_t\theta_t}\gamma_j(\Lambda)\\ &= \frac{1}{\mu}\left(\theta_i\sum X_{ij}-\frac{n_i\theta_i}{\sum_t n_t\theta_t}\sum_i\left(\theta_i\sum_j X_{ij}\right)\right)\end{aligned}$$

행렬을 이용한 계산이 좀 복잡하기는 하지만 벡터 D, A, B에 $\hat{\boldsymbol{\Lambda}}$을 대입하고 $S(\hat{\boldsymbol{\Lambda}})=(S_1(\hat{\boldsymbol{\Lambda}}),\cdots,S_{k-1}(\hat{\boldsymbol{\Lambda}}))'$과 함께 $\boldsymbol{\Phi}=\mathbf{0}$을 검정하는 아래의 $C(\alpha)$ 통계량 $DCA=S'Cov(S)^{-1}S=S'(D-AB^{-1}A')^{-1}S$을 얻게 되며 이는 자유도가 $k-1$인 카이제곱분포로 수렴한다(Neyman, 1959; Moran, 1970).

$$DCA=\sum_{i=1}^{k}n_i\frac{(\overline{X_{i.}}-\mu_0)^2}{\mu_0/\theta_{0i}}=\sum_{i=1}^{k}n_i\frac{(\overline{X}_{i.}-\overline{X})^2}{\overline{X}/\theta_{0i}}$$

여기서도 H_0이 사실이고 산포모수가 공통일 때 각 군의 추정량 $\theta_{01},\cdots,\theta_{0k}$대신 공통추정량 θ_0를 대입함으로써 처리효과를 비교할 수 있는 통계량을 얻을 수 있다.

4. 모의실험

주어진 자료집합이 음이항분포 모형을 따른다고 가정한 경우와 이중지수계 포아송 모형을 따른다고 가정한 경우, 산포모수가 공통인 경우와 그렇지 않은 경우로 나누어 각각의 경우에서 우도비 검정과 $C(\alpha)$ 검정에 의해 유도한 각 군의 처리효과를 비교할 수 있는 통계량들을 모의실험을 통하여 유의수준 α=0.05에서의 제1종 오류와 검정력을 비교하였다. 이 중에서 음이항분포 모형으로부터 유도한 통계량들의 산포모수 추정량은 최대우도추정량 대신 검정력이 약간 떨어지기는 하나(0.02-0.05) 구하기 쉬운 적률추정량을 사용하였다.

치료군의 갯수가 k=2, 3인 경우, 각 군의 크기는 동일하게 n=5, 10, 30, 50 인 경우의 모든 조합에 대하여 앞에서 유도한 네 가지 통계량들을 공통의 산포모수가 존재하는 경우와 그렇지 않은 경우로 나누어 IMSL(International Mathematical and Statistical Library)을 이용하여 생성한 음이항분포모형, 포아송의 균등혼합분포 모형(uniform mixture of Poisson model)과 이산균등분포 모형(discrete uniform model)의 세 가지 자료집합에 대하여 2000번 반복을 거쳐 모의실험을 하였다. 이 때 각 자료집합은 평균이 μ=10과 30이고 분산은 주어진 산포모수에 대해 $\mu+c\mu^2=\mu/\theta$ 를 만족한다.

여러 조합으로 얻어진 실험결과중 치료군의 갯수와 각 군의 크기에 따른 통계량간의 별다른 차이가 없었고, 각 군 평균의 크고 작음이 검정력에 영향을 미치지 않았으므로 본 논문에서는 k=3, n=10, 30이며 군의 평균 μ=10인 경우로 제한하여 표 1-3를 만들었다.

표 1. 유의수준 α=0.05일 때 음이항분포 모형에서의 제1종 오류와 검정력
표 2. 유의수준 α=0.05일 때 포아송의 균등혼합분포 모형에서의 제1종 오류와 검정력
표 3. 유의수준 α=0.05일 때 이산균등분포 모형에서의 제1종 오류와 검정력

표 1의 음이항분포 자료집합을 이용한 모의실험 결과에서 우도 검정법이 $C(\alpha)$ 검정법보다 검정력이 월등히 뛰어남을 볼 수 있다. 그러나 우도 검정은 n=10일 때 제1종 오류가 유의수준의 95% 신뢰구간인 (0.05-0.0096, 0.05+0.0096)의 상한에서 크게 벗어나 검정통계량으로 부적합하고 표본의 크기가 커짐에 따라 이 현상은 완화되지만 우도검정법을 사용하기 위해서는 공통의 산포모수라는 가정이 반드시 수반되어야 한다. $C(\alpha)$ 검정통계량 중에서도 이중지수계 포아송분포로부터 유도한 통계량이 음이항분포로부터 유도한 통계량보다 각 군의 크기가 작을 때 처리효과를 비교하는 데 더 적합하다. 또한 각 군의 산포모수를 각각 추정한 경우의 $C(\alpha)$ 통계량이 실제 자료집합의 산포모수가 서로 공통이건 아니건 상관없이 항상 검정력이 큼을 볼 수 있다. 그러므로 각 군의 처리효과를 비교하기 위해서는 DCA가 가장 적합한 통계량이고 각 군의 산포모수를 따로 추정함으로써 검정력이 더욱 좋아졌다. 군의 크기가 클 경우에는 이외에도 NCA와 공통의 산포모수라는 가정 아래에서 NLR, DLR을 쓸 수도 있다.

표 2의 자료집합은 포아송분포의 모수가 균일분포를 따르는 경우로 음이항분포와는 다른 형태의 포아송혼합모형에서 생성되었지만 표 1의 결과와 같은 양상을 띄었다. 이는 모수가 따르는 분포들의 로그우도의 국소적 움직임이 비슷함(Chesher, 1984)에 따른 당연한 결과라 하겠다.

표 1과 2의 자료집합은 모두 포아송의 과산포 형태로 $c \geq 0$인 경우만 해당되지만 표 3에서는 모수 μ를 중심으로 분산이 $\mu + c\mu^2$인 이산균등분포를 이용하여 $c \geq 0$ 뿐만아니라 $c < 0$인 경우도 다룰 수 있는 자료집합을 생성하였다. 그러므로 $c < 0$일 경우에는 $c \geq 0$ 을 가정하는 음이항분포로부터 유도된 통계량의 제1종 오류가 유의수준의 95%신뢰구간을 이탈하는 정도가 더욱 심화됨을 알 수 있다. 그러나 DCA 통계량은 군의 크기와 c값에 상관없이 항상 유의수준을 잘 지킬 뿐만이 아니라 각 군의 산포모수가 같지 않을 경우에도 검정력이 우수한 통계량임을 확인하였다. 앞의 두 자료집합에서의 결과와는 달리 NLR 통계량은 군의 크기가 클 경우에도 부적합함이 밝혀졌다.

5. 예제

McCaughran과 Arnold(1976)는 태아의 사망(embryonic death)에 대한 모형으로 음이항분포가 매우 적합함을 보였다. 또한 표 4의 약물 효과 비교에서 분산 안정화 과정을 거쳐 자료를 변형(variance stabilizing transformation)하여 세 군 사이에 차이가 없음을 보였다.

표 4의 자료를 제2절과 제3절에서 유도한 통계량들에 대입한 결과 표 5를 얻었는데 모든 통계량들의 검정결과가 McCaughran과 Arnold의 결론과 일치함을 볼 수 있다. 각 군의 크기가 작으므로 앞의 모의실험에서 보았듯이 각각의 다른 산포모수를 인정하는 경우의 통계량 NLR의 유의확률(p-value)이 다른 유의확률과는 차이가 있었다.

표 4. 대조군과 두 치료군 사이의 태아 사망 개체수 비교

	10개 모체의 관찰된 태아 사망 개체수	$\bar{X}$	$\hat{\theta}$
대조군	0 2 1 0 0 0 0 1 0 0	0.4	0.7256
약물 치료군 1	0 0 1 0 0 1 3 0 1 1	0.7	0.7413
약물 치료군 2	0 0 1 1 2 2 0 2 0 4	1.2	0.6163

Barnwal과 Paul(1988)도 이 자료를 이용하여 약물의 효과가 없음을 보였지만 이들이 유도한 NLR와 NCA 통계량을 사용하기 위해서는 각 군의 평균 비교 이전에 산포모수들이 모두 같은지에 대한 검정이 선행되어야 한다. 그러나 각 군의 산포모수를 인정하는 경우의 통계량은 이러한 번거로움없이 사용할 수 있고 DCA통계량은 특히 검정력이 높음을 볼 수 있다.

표 5. 표 4의 처리효과 분석 결과

	공통의 산포모수를 가정				각 군의 산포모수를 인정			
검정통계량	NLR	NCA	DLR	DCA	NLR	NCA	DLR	DCA
유의확률	0.200	0.221	0.206	0.210	0.062	0.210	0.103	0.233

6. 결론

일원배열형 이산자료의 처리 효과 비교에서 우도검정법이 $C(\alpha)$ 검정법보다 검정력이 높으나 제1종 오류가 유의수준을 훨씬 넘는 경우가 많았으므로 좋은 검정법이 아님이 모의실험 결과 나타났다. 그리고 여러 가지 자료집합에 대한 $C(\alpha)$ 검정에서는 음이항분포 모형과 이중지수계 포아송분포의 모형으로부터 유도된 통계량간에 검정력이 비슷하였고 자료집합에 따라 후자가 약간 우월한 경우도 있었다.

Barnwal과 Paul(1988)은 각 군의 산포모수가 모두 같은 경우로 제한하여 처리효과비교를 하였는데 이를 각 군의 산포모수가 서로 다를 경우에도 가능하도록 확장하였고 이렇게 함으로써 검정력이 향상됨을 보았다.

본 논문에서 유도한 통계량 DCA는 이중지수계 포아송모형으로부터 유도되어 최대우도추정량을 구하기 쉽고 $C(\alpha)$ 통계량임으로 계산이 간단하고 대립가설에서의 모수의 추정이 필요없다는 장점이 있다. 또한 모의실험 결과 어느 경우에서나 처리 효과 비교에 적합함을 보였으며 군의 크기가 작을 수록 다른 통계량들보다 월등히 우수함을 보였다.

참 고 문 헌

[1] Barnwal, P.K. and Paul, S.R. (1988). Analysis of one-way layout of count data with negative binomial variation, *Biometrika*, Vol. 75, 215-222.

[2] Chesher, A. (1984). Testing for neglected heterogeneity, *Econometrika*, Vol. 52, 865-72.

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

[4] Efron, B. (1986). Double exponential families and their use in generalized linear regression, *Journal of the American Statistical Association*, Vol. 81, 709-21.

[5] Foulley, J.L., Gianola, D., and Im, S. (1987). Genetic evaluation of traits distributed as Poisson binomial with reference to reproductive characters, *Theoretical and Applied Genetics*, Vol.73, 870-877.

[6] Haseman, J.K. and Soares, E.R. (1976). The distribution of fetal death in control mice and its implications on statistical tests for dominant lethal effects, *Mutation Research*, Vol. 41, 277-88.

[7] Lockhart, A., Piegorsch, W. and Bishop, J. (1992). Assessing overdispersion and dose-response in the male dominant lethal assay, *Mutation Research*, Vol. 272, 35-58.

[8] MacCaughran, D.A. and Arnold, D.W. (1976). Statistical models for numbers of implantation sites and embryonic deaths in mice, *Toxicology and Applied Pharmacology*, Vol. 38, 325-333.

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

[10] Moran, P.A.P. (1970). On asymptotically optimal tests of composite hypothesis, *Biometrika*, Vol. 57, 45-75.

[11] Neyman, J. (1959). Optimal asymptotic tests of composite hypothesis, *In Probability and Statistics: The Harold Cramer Volume*, U. Grenander (ed). New York: Wiley.

[12] Olesen, I., Perez_Enciso, M., Gianola, D. and Thomas, D. (1994). A comparison of normal and nonnormal models for number of lambs born in Norwegian sheep, *Journal of Animal Science*, Vol. 72, 1166-73.

[13] Paul S.R. and Islam A. (1995). Analysis of Proportions in the Presence of Over-/Under-dispersion, *Biometrics*, Vol. 51, 1400-1410.

[14] Tarone, R. E.(1985). On heterogeneity tests based on efficient scores, *Biometrika*, Vol. 72, 91-95.

[15] Tempelman, R. and Gianola, D.(1996). A mixed effects model for overdispersed count data in animal breeding, *Biometrics*, Vol. 52, 265-79.

표 1. 유의수준 α=0.05일 때 음이항분포모형에서의 제1종 오류와 검정력

(a) $n_1 = n_2 = n_3 = 10$

μ_1	μ_2	μ_3	c_1	c_2	c_3	공통의 산포모수 추정				각 군의 산포모수 추정			
						NLR	NCA	DLR	DCA	NLR	NCA	DLR	DCA
10	10	10	0.00	0.00	0.00	0.065	0.045	0.068	0.049	0.101	0.041	0.097	0.054
			0.00	0.00	0.10	0.065	0.043	0.068	0.047	0.114	0.045	0.097	0.058
			0.05	0.10	0.15	0.057	0.035	0.057	0.042	0.112	0.036	0.094	0.048
			0.15	0.15	0.15	0.062	0.036	0.064	0.041	0.133	0.040	0.101	0.046
10	11	11	0.00	0.00	0.00	0.108	0.080	0.113	0.083	0.168	0.069	0.154	0.092
			0.00	0.00	0.10	0.097	0.068	0.104	0.076	0.176	0.069	0.154	0.086
			0.05	0.10	0.15	0.093	0.058	0.100	0.072	0.166	0.069	0.142	0.084
			0.15	0.15	0.15	0.089	0.056	0.093	0.065	0.172	0.058	0.138	0.069
10	11	12	0.00	0.00	0.00	0.219	0.175	0.221	0.181	0.290	0.148	0.269	0.192
			0.00	0.00	0.10	0.167	0.123	0.174	0.141	0.257	0.133	0.235	0.159
			0.05	0.10	0.15	0.132	0.098	0.137	0.110	0.236	0.106	0.203	0.123
			0.15	0.15	0.15	0.138	0.093	0.141	0.108	0.219	0.092	0.178	0.104

(b) $n_1 = n_2 = n_3 = 30$

μ_1	μ_2	μ_3	c_1	c_2	c_3	공통의 산포모수 추정				각 군의 산포모수 추정			
						NLR	NCA	DLR	DCA	NLR	NCA	DLR	DCA
10	10	10	0.00	0.00	0.00	0.052	0.047	0.052	0.045	0.064	0.043	0.061	0.047
			0.00	0.00	0.10	0.054	0.046	0.052	0.045	0.089	0.054	0.075	0.056
			0.05	0.10	0.15	0.056	0.048	0.056	0.050	0.108	0.054	0.083	0.056
			0.15	0.15	0.15	0.050	0.045	0.051	0.047	0.101	0.052	0.071	0.052
10	11	11	0.00	0.00	0.00	0.204	0.186	0.198	0.183	0.229	0.184	0.222	0.186
			0.00	0.00	0.10	0.163	0.148	0.161	0.149	0.257	0.200	0.242	0.202
			0.05	0.10	0.15	0.112	0.096	0.112	0.105	0.197	0.135	0.169	0.129
			0.15	0.15	0.15	0.112	0.093	0.113	0.098	0.193	0.109	0.151	0.105
10	11	12	0.00	0.00	0.00	0.535	0.511	0.526	0.505	0.572	0.505	0.560	0.512
			0.00	0.00	0.10	0.412	0.377	0.419	0.394	0.504	0.418	0.478	0.419
			0.05	0.10	0.15	0.283	0.255	0.285	0.270	0.403	0.300	0.362	0.290
			0.15	0.15	0.15	0.233	0.225	0.238	0.228	0.344	0.240	0.293	0.234

표 2. 유의수준 α=0.05일 때 포아송의 균등혼합분포모형에서의 제1종 오류와 검정력

(a) $n_1 = n_2 = n_3 = 10$

μ_1	μ_2	μ_3	c_1	c_2	c_3	공통의 산포모수 추정				각 군의 산포모수 추정			
						NLR	NCA	DLR	DCA	NLR	NCA	DLR	DCA
10	10	10	0.05	0.05	0.05	0.068	0.045	0.068	0.048	0.109	0.042	0.098	0.053
			0.05	0.05	0.10	0.068	0.051	0.068	0.053	0.112	0.044	0.099	0.055
			0.05	0.10	0.15	0.068	0.047	0.067	0.050	0.115	0.045	0.096	0.056
			0.15	0.15	0.15	0.057	0.045	0.058	0.045	0.105	0.040	0.083	0.047
10	11	11	0.05	0.05	0.05	0.105	0.078	0.109	0.079	0.156	0.065	0.141	0.078
			0.05	0.05	0.10	0.108	0.076	0.108	0.081	0.163	0.067	0.145	0.082
			0.05	0.10	0.15	0.086	0.058	0.086	0.066	0.150	0.061	0.132	0.073
			0.15	0.15	0.15	0.089	0.065	0.092	0.064	0.149	0.062	0.119	0.068
10	11	12	0.05	0.05	0.05	0.162	0.124	0.163	0.131	0.226	0.107	0.202	0.135
			0.05	0.05	0.10	0.149	0.107	0.153	0.122	0.225	0.101	0.194	0.121
			0.05	0.10	0.15	0.128	0.094	0.130	0.102	0.213	0.097	0.189	0.112
			0.15	0.15	0.15	0.099	0.077	0.102	0.073	0.172	0.063	0.140	0.079

(b) $n_1 = n_2 = n_3 = 30$

μ_1	μ_2	μ_3	c_1	c_2	c_3	공통의 산포모수 추정				각 군의 산포모수 추정			
						NLR	NCA	DLR	DCA	NLR	NCA	DLR	DCA
10	10	10	0.05	0.05	0.05	0.055	0.050	0.054	0.049	0.079	0.049	0.067	0.049
			0.05	0.05	0.10	0.052	0.046	0.049	0.044	0.082	0.049	0.069	0.047
			0.05	0.10	0.15	0.049	0.046	0.048	0.044	0.082	0.049	0.067	0.046
			0.15	0.15	0.15	0.052	0.050	0.053	0.046	0.097	0.051	0.070	0.047
10	11	11	0.05	0.05	0.05	0.159	0.148	0.156	0.146	0.204	0.147	0.186	0.148
			0.05	0.05	0.10	0.135	0.124	0.134	0.122	0.194	0.139	0.171	0.135
			0.05	0.10	0.15	0.122	0.115	0.122	0.112	0.196	0.141	0.173	0.135
			0.15	0.15	0.15	0.099	0.098	0.100	0.090	0.166	0.102	0.131	0.095
10	11	12	0.05	0.05	0.05	0.356	0.338	0.352	0.338	0.424	0.337	0.391	0.341
			0.05	0.05	0.10	0.327	0.307	0.327	0.309	0.397	0.320	0.369	0.311
			0.05	0.10	0.15	0.278	0.263	0.280	0.263	0.381	0.301	0.353	0.290
			0.15	0.15	0.15	0.213	0.213	0.217	0.203	0.321	0.208	0.265	0.201

표 3. 유의수준 α=0.05일 때 이산균등분포모형에서의 제1종 오류와 검정력

(a) $n_1 = n_2 = n_3 = 10$

μ_1	μ_2	μ_3	c_1	c_2	c_3	공통의 산포모수 추정				각 군의 산포모수 추정			
						NLR	NCA	DLR	DCA	NLR	NCA	DLR	DCA
10	10	10	-0.01	-0.01	-0.01	0.147	0.052	0.064	0.053	0.285	0.042	0.089	0.057
			-0.01	0.00	0.02	0.126	0.040	0.060	0.042	0.213	0.032	0.080	0.052
			0.00	0.01	0.03	0.090	0.042	0.058	0.043	0.141	0.036	0.079	0.047
			0.03	0.03	0.03	0.140	0.041	0.058	0.041	0.265	0.032	0.080	0.043
10	11	11	-0.01	-0.01	-0.01	0.250	0.101	0.132	0.103	0.381	0.083	0.149	0.100
			-0.01	0.00	0.02	0.191	0.076	0.109	0.079	0.286	0.068	0.135	0.087
			0.00	0.01	0.03	0.156	0.085	0.106	0.084	0.255	0.065	0.132	0.083
			0.03	0.03	0.03	0.152	0.083	0.109	0.080	0.173	0.070	0.132	0.086
10	11	12	-0.01	-0.01	-0.01	0.377	0.204	0.264	0.211	0.420	0.164	0.271	0.201
			-0.01	0.00	0.02	0.273	0.172	0.221	0.176	0.338	0.128	0.241	0.161
			0.00	0.01	0.03	0.303	0.152	0.192	0.156	0.481	0.119	0.213	0.144
			0.03	0.03	0.03	0.312	0.157	0.202	0.160	0.448	0.124	0.215	0.153

(b) $n_1 = n_2 = n_3 = 30$

μ_1	μ_2	μ_3	c_1	c_2	c_3	공통의 산포모수 추정				각 군의 산포모수 추정			
						NLR	NCA	DLR	DCA	NLR	NCA	DLR	DCA
10	10	10	-0.01	-0.01	-0.01	0.132	0.055	0.056	0.052	0.229	0.052	0.063	0.054
			-0.01	0.00	0.02	0.107	0.046	0.045	0.043	0.181	0.045	0.053	0.045
			0.00	0.01	0.03	0.083	0.049	0.051	0.045	0.138	0.045	0.054	0.047
			0.03	0.03	0.03	0.128	0.047	0.048	0.043	0.216	0.045	0.055	0.047
10	11	11	-0.01	-0.01	-0.01	0.433	0.243	0.256	0.237	0.549	0.230	0.269	0.230
			-0.01	0.00	0.02	0.340	0.204	0.220	0.199	0.449	0.211	0.241	0.207
			0.00	0.01	0.03	0.269	0.190	0.198	0.186	0.384	0.191	0.221	0.186
			0.03	0.03	0.03	0.235	0.167	0.171	0.154	0.190	0.166	0.193	0.161
10	11	12	-0.01	-0.01	-0.01	0.752	0.627	0.645	0.623	0.718	0.596	0.643	0.599
			-0.01	0.00	0.02	0.618	0.525	0.541	0.522	0.636	0.517	0.557	0.513
			0.00	0.01	0.03	0.615	0.435	0.444	0.429	0.817	0.423	0.463	0.416
			0.03	0.03	0.03	0.588	0.419	0.425	0.405	0.727	0.403	0.454	0.394

Analysis of Counts in the One-way Layout[3)]

Sunho Lee[4)]

Abstract

Barnwal and Paul(1988) derived the likelihood ratio statistic and $C(\alpha)$ statistic for testing the equality of the means of several groups of count data in the presence of a common dispersion parameter. These tests are generalized to be applicable without the restriction of a common dispersion parameter. And the assumed model of data is also extended from negative binomial to double exponential Poisson model. Monte Carlo simulations show the superiority of $C(\alpha)$ statistic based on the double exponential Poisson family which has a very simple form and requires estimates of the parameters only under the null hypothesis.

3) This research was supported by Daeyang Research Foundation of Sejong University.

4) Assistant Professor, Dept. of Mathematics, Sejong University, Seoul, 143-747, Korea.