응용통계연구 제 8권 2호
1995년 9월 pp. 147-161.

가산자료(count data)의 과산포 검색: 일반화 과정

김 병 수[1], 오 경 주[2], 박 철 용[3]

요 약

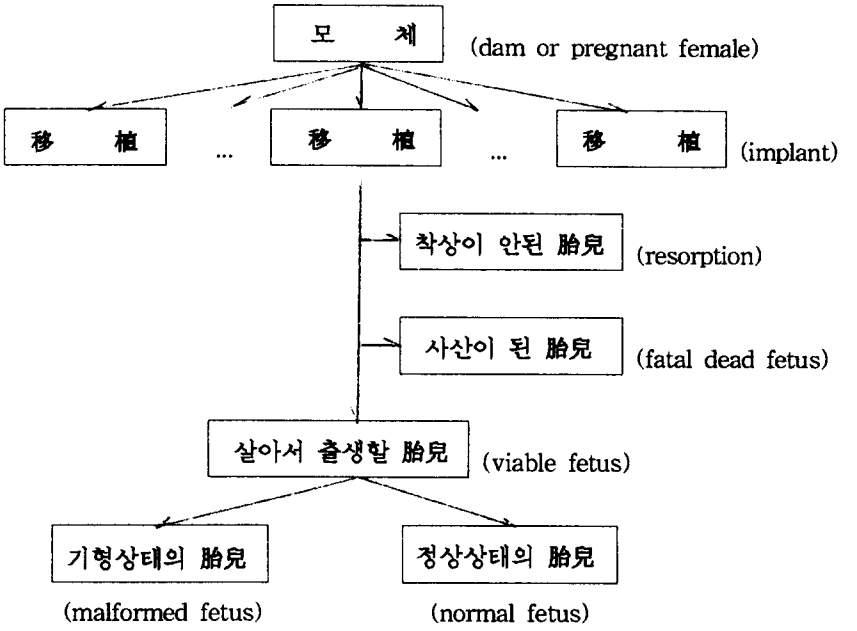
생검실험에서는 다산을 통해 번식하는 쥐와 같은 설치류동물들을 실험대상으로 하여 이항분포나 포아송분포 하에서 가산자료(count data)를 많이 생성한다. 다산을 통해 태어난 동물들을 독립적인 실험대상으로 간주하여 자료분석을 하면, 同腹仔효과로 인해 기존의 평균과 분산사이의 관계를 벗어나는 과산포현상이 종종 나타난다. 이러한 현상을 무시했을 때 모수추정치에 대한 분산을 과소추정하고, 이로 인하여 가설검정에서 낮은 검정력을 갖게된다. 이러한 문제점을 해결하기 위하여 최근 10여년간 과산포현상을 검색하는 통계량들과 과산포를 반영하는 모형들이 제시되었는데, 이를 개관하고 이러한 절차들의 일반화 과정을 자료 유형별로 비교분석한다.

1. 서 론

생검실험(bioassay)에서는 쥐와 같은 설치류 동물을 대상으로 약물이나 독성물질의 반응을 측정한다. 주로 실험대상으로 쓰는 쥐와 같은 동물은 하나의 모체가 胎兒를 낳을 때 多産을 하게 된다. 하나의 모체에서 나온 胎兒들은 다른 모체에서 나온 胎兒들과 비교해서 비슷한 속성이나 성질을 갖게되고, 이로 인하여 생검실험에서 유사한 반응을 나타낸다. 이때 모체(pregnant female)를 댐(dam)이라고 하며, 한 어미에서 나온 胎兒들을 통털어 同腹仔(litter)라 하고, 同腹仔끼리 나타내는 비슷한 반응을 同腹仔효과(litter effect)라 한다. 同腹仔효과를 일으키는 원인에는 여러가지가 있는데 특히 실험에 쓰이는 동물을 다루는 방법의 차이, 유전적인 요소에서의 차이, 각각의 어미가 자라난 환경의 차이와 나이의 차이등을 들 수 있다. 同腹仔효과를 무시하고 각각의 胎兒들의 반응이 독립이라고 간주하여 생검실험을 하게 되면 실험에서 얻은 자료가 過散布현상을 보인다(Haseman and Kupper, 1979).

Ryan(1992)은 同腹仔효과가 다음의 <그림 1>과 같은 位階構造를 가짐을 보이고, 하나의 정상 胎兒가 나오기 위해서는 착상이 된 후 살아서 출생해야 하고 기형이 되지 말아야 한다는 점에서 착상이 안된 胎兒(resorption), 사산이 된 胎兒(fatal dead fetus), 기형이 된 胎兒(malformed fetus), 정상상태의 胎兒(normal fetus)에 위계적 관계(hierarchical relationship)를 구성하였고, 이러한 구조를 염두에 두고 통계모형을 설정해야 한다고 주장했다.

1) (120-749) 서울특별시 서대문구 신촌동 134번지, 연세대학교 응용통계학과.
2) (110-450) 서울특별시 종로구 원남동 66-21 보령빌딩, 금강기획 마케팅전략연구소.
3) (704-701) 대구직할시 달서구 신당동 1000번지, 계명대학교 통계학과.



<그림 1> Ryan(1993)이 제시한 同腹仔효과의 구조

2. 과산포 검색에 관한 여러가지 모형

과산포현상을 개선하기 위한 방법의 하나는 과산포를 나타내는 모수를 포함하는 통계모형을 개발하고, 모수들의 최대우도추정값을 구한 후, 스코어 검정통계량을 유도하는 것이다(Liang and McCullagh, 1993). 생검실험에서 생성되는 자료의 모형으로 자주 쓰이는 분포는 이항분포와 포아송분포이다. 만일 과산포가 존재할 때, 이항분포나 포아송분포를 선택하여 분석을 하게 되면 서론에서 언급한 것처럼 추론상의 여러가지 문제를 야기한다. 이때에는 분포를 가정하지 않은 비모수적 방법을 이용하여 실험에서 얻은 자료를 분석하는 것도 한가지 방법이다.

표기의 편의를 위해 본고에서는 다음과 같은 정의들을 사용한다. 확률변수 X가 시행회수와 성공확률이 각각 n, p인 이항분포를 따를 때 $X \sim B(n, p)$로, 평균이 λ인 포아송분포를 따를 때 $X \sim$ Poisson(λ)로, 자유도가 ν인 카이제곱분포를 따를 때 $X \sim \chi^2(\nu)$로 표시한다. 또한 X가 다음의 식(2-1)과 같은 확률질량함수를 가질 때 $X \sim NB(N, (1-p)/p)$로 표기한다.

$$P(X=x)=\binom{N+x-1}{N-1} p^N(1-p)^x, \quad x=0,1,2,\ldots \tag{2-1}$$

$V=v$로 주어졌을 때 U의 조건분포가 F를 따르면 $(U|v) \sim F$로 표기한다.

2.1. 초이항변이에 관한 모형

同腹仔중 죽은 胎兒수를 관측치로 삼는 포아송분포와는 다르게 이항분포는 한 모체내 同腹仔수에 대한 죽은 胎兒수의 비율을 관측치로 삼는다. 따라서 이항분포는 모체내 同腹仔수와 죽은 胎兒수의 연관성(association)을 함께 고려한다는 점에서 포아송분포와 구별된다. 경우에 따라서는 同腹仔의 수와 죽은 胎兒의 수는 약간의 양의 상관관계가 있는 것으로 알려졌다(Haseman and Kupper, 1979).

다음부터 소개되는 이항분포의 혼합분포들은 반응변수가 二進자료(binary data)일 때이다. 즉 죽지않은 胎兒는 0으로, 죽은 胎兒는 1로 표시한다. 또한 X_{ij}는 $i(i=1,2,\ldots,M)$번째 모체의 $j(j=1,2,\ldots,n_i)$번째 태아의 생사를 나타내는 베르누이 확률변수이고, $X_i=\sum_{j=1}^{n_i} X_{ij}$는 i번째 모체에서 죽은 태아의 수를 나타내는 확률변수이다.

2.1.1. 베타이항분포(Beta-Binomial Distribution)

Williams(1975)는 한 모체내에서 반응을 보이는 同腹仔의 수는 이항분포를 따르며 모체의 胎兒가 죽을 확률은 베타분포를 따라 변한다고 가정하였다. 이에 근거하여 X_i의 주변분포(marginal distribution)로서 다음 식(2-2)와 같은 베타이항분포(beta-binomial distribution)를 유도하였다.

$$P(X_i = x) = \binom{n_i}{x} B(\alpha, n_i+\beta-x) / B(\alpha, \beta) \qquad (2\text{-}2)$$

단, $B(a, b) = \int_0^1 u^{a-1}(1-u)^{b-1}\,du$, $\quad a>0,\ b>0$이다.

모수 α, β를 추정하기 위해서 다음과 같이 모수변환을 한다.

$$p = \alpha(\alpha+\beta)^{-1}, \quad \phi = (\alpha+\beta)^{-1}$$

변환된 모수 p, ϕ는 초이항변이를 모수화함에 있어 α, β보다 의미있는 모수가 된다. 변환된 모수 p는 베타분포의 기대값이 되며, ϕ는 각 모체내 同腹仔간 변이(inter-litter variation)를 나타내고, 베타분포의 분산인 $p(1-p)\phi(1+\phi)^{-1}$의 구성인자로서 척도모수(scale parameter)를 결정하는 역할을 한다. 따라서 식(2-2)를 만족할 경우 $X_i \sim BB(n_i, p, \phi)$로 표기하기로 하며 이 때 X_i의 기대값과 분산은 각각 n_ip, $n_ip(1-p)(n_i\phi+1)/(1+\phi)$이며 $\phi \to 0$일 때 $X_i \sim B(n_i, p)$가 됨을 알 수 있다.

위의 베타이항분포의 경우 과산포측도(a measure of excess dispersion)를 $\tau^2 = 1/(1+\alpha+\beta)$로 놓았을 때 이항분포에 대한 산포계수(dispersion factor)가 $(n_i\phi+1)/(1+\phi) = 1+(n_i-1)\tau^2$이 되어 同腹仔크기(litter size)에 따라 변동한다. 하지만 일반화 선형모형(generalized linear models)의 경우 고정된 산포계수가 가정된다. Liang and McCullagh(1993)는 다섯개의 실제자료군이 위의 두가지 산포계수를 가지는 모형 중에 어느 쪽에 더 잘 적합되는지 비교, 분석하

였다.

2.1.2. 상관이 있는 이항분포(Correlated Binomial Distribution)

Kupper and Haseman(1978)은 모체내의 胎兒가 죽을 확률은 그것의 모체가 속한 그룹에 따라 결정되고, 각 모체내의 胎兒들은 상관관계를 가진다는 가정을 하고 식(2-3)과 같은 상관이 있는 이항분포(correlated binomial distribution)를 제안하였다.

$$P(X_i = x_i) = \binom{n_i}{x_i} p^{x_i}(1-p)^{n_i - x_i} f(x_{i1}, \cdots, x_{in_i}), \quad x_i = 0, 1, \ldots, n_i \tag{2-3}$$

여기서 $f(x_{i1}, \cdots, x_{in_i})$는 상호독립이 아닌 한 모체내 n_i개 胎兒들의 同腹仔효과를 반영하는 수정인자(correction factor)로 3차 이상의 상관관계를 무시하면 식(2-4)와 같은 확률질량함수를 얻을 수 있다.

$$P(X_i = x_i) = \binom{n_i}{x_i} p^{x_i}(1-p)^{n_i - x_i} \cdot \left\{1 + \frac{\theta}{2p^2(1-p)^2} + \left[(x_i - n_i p)^2 + x_i(2p-1) - n_i p^2 \right] \right\} \tag{2-4}$$

단, $\theta = \mathrm{Cov}(X_{ij}, X_{ij'})$, $j \neq j'$ 이다.

식(2-4)의 관계를 만족할 때 $X_i \sim CB(n_i, p, \theta)$로 표기하며 이 때 X_i의 기대값과 분산이 각각 $n_i p$, $n_i p(1-p) + n_i(n_i - 1)\theta$가 되며 $\theta = 0$일 때 $X_i \sim B(n_i, p)$가 됨을 알 수 있다.

Kupper and Haseman(1978)은 베타이항분포와 상관이 있는 이항분포를 실제자료에 근거하여 비교하고 자신들이 주장한 상관이 있는 이항분포가 베타이항분포에 비해 더 큰 우도비값을 가짐을 보였다.

2.1.3. 상관이 있는 베타이항분포(Beta-Correlated Binomial Distribution)

Paul(1987)은 상관이 있는 이항분포와 베타이항분포를 결합하여 상관이 있는 베타이항분포(beta-correlated binomial distribution)를 제안하였다. 이는 상관이 있는 이항분포의 베타혼합(beta mixture of correlated binomial)으로 세모수(three-parameter) 분포가 된다. X_i가 상관이 있는 베타이항분포를 따를 때 기대값과 분산은 $n_i p$, $[\, n_i p(1-p)(n_i \phi + 1) / (\phi + 1)] + n_i(n_1 - 1)\theta$가 되며, $\theta = 0$이면 $X_i \sim BB(n_i, p, \phi)$, $\phi \to 0$이면 $X_i \sim CB(n_i, p, \theta)$, $\theta = 0$ 과 $\phi \to 0$이면 $X_i \sim B(n_i, p)$가 된다.

2.2. 초포아송 변이에 관한 모형

모체내 죽은 胎兒수를 관측치로 삼는 포아송분포는 모체내 同腹仔크기(litter size)를 고려하

지 않는다. 이항분포인 경우 반응변수는 同腹仔수에 대한 죽은 胎兒수의 비율이 되므로 죽은 胎兒수가 0에 가깝게 되면 반응변수가 0에 가깝게 되어 분석을 위한 계산시 어려움이 따른다. 그러나, 포아송분포인 경우 죽은 胎兒수가 0이 되어서는 안되는 특별한 이론적 제약이 없다 (Haseman and Soares, 1976).

2.2.1. 음이항분포(Negative Binomial Distribution)

포아송분포에서 과산포문제는 활발히 다루어져 음이항분포를 대립가설로 놓고 초포아송변이에 대한 많은 통계량이 제시되었다. 음이항분포는 포아송분포의 감마혼합(gamma mixture of Poisson)으로, 모수가 변함에 따라 평균과 분산의 관계가 변한다. $X \sim NB(1/c, cm)$일 때 평균은 m, 분산은 $m(1+cm)$으로 분산은 평균의 이차함수 형태가 되며 $X \sim NB(m/c, c)$일 때 평균은 m, 분산은 $m(1+c)$로 분산과 평균이 선형관계가 된다. 또한 $X \sim NB(m^{2-r}/c, cm^{r-1})$일 때 평균은 m, 분산은 $m(1+cm^{r-1})$으로 분산과 평균이 r차관계인 일반적인 관계가 된다. 음이항분포는 $c>0$의 범위를 가지며, 위의 모든 경우에서 $c \to 0$일 때 $X \sim$ Poisson(m)이 된다.

Potthoff and Whittinghill(1966)은 포아송분포의 동질성 검정을 위한 통계량을 비동질성을 대립가설로 하여 유도하였으며, 대립가설이 음이항분포일 때 그 검정은 국소최강력검정(locally most powerful test)임을 보였다. Collings and Margolin(1985)은 분산이 평균의 이차함수 형태일 때 포아송분포로부터 음이항분포로의 이탈을 검색하는 국소최적검정(locally optimal test)을 찾았다. Kim and Park(1992)은 분산이 평균의 일차함수일 때 국소최적검정을 유도하고 콜링스와 마골린의 결과와 비교하여 초포아송변이의 검색의 문제에서 음이항분포의 모수화에 따라 국소최적검정이 바뀌어질 수 있다는 것을 보이고 있다. 이선호(1993)는 분산과 평균이 r차관계인 일반적인 경우에 국소최적검정을 유도하였다. Lee, Park and Kim(1994)은 음이항분포와 이항분포를 포함하는 캇츠분포군을 대립가설로 놓고 분산과 평균이 r차관계인 일반적인 경우 포아송분포로부터 이탈을 검색하는 국소최적검정을 유도하였다.

2.2.2. 포아송 혼합분포(Mixed Poisson Distribution)

Cox(1983)와 Chesher(1984)는 모수의 사전분포(prior distribution)가 일반적인 경우에도 과산포가 없는 귀무가설분포(null distribution)에서 모수의 추정량을 구하여 스코어검정을 행할 수 있음을 보였다. 이들의 논법을 이용하여 Dean and Lawless(1989)는 포아송회귀모형에서의 이탈을 검색하는 검정통계량을 유도하였다.

포아송혼합분포는 다음 식(2-5)와 같은 포아송분포로 부터 유도되었다.

$$(Y_i \mid x_i) \sim \text{Poisson}(\mu_i(x_i;\beta)), \quad i = 1, 2, \cdots, n \tag{2-5}$$

여기서 Y_i는 독립적인 반응변수이며 x_i는 $p\times 1$연관된 공변수벡터(associated covariate vector)이며, β는 $p\times 1$ 회귀계수벡터이다. 다시 말해, 위의 식(2-5)에 독립적이고 동일한 분포를 갖는 양의 확률변수 $\nu_1, \nu_2, \cdots, \nu_n$를 도입하면 식(2-6)과 같은 포아송 혼합모형이 된다.

$$(Y_i \mid x_i, \nu_i) \sim \text{Poisson}(\nu_i \mu_i(x_i; \beta)) \tag{2-6}$$

이때, ν_i는 유한한 분산을 가지며 일반성을 잃지 않고 $E(\nu_i) = 1$, $\text{Var}(\nu_i) = \tau$가 성립한다고 가정할 수 있다. $\tau = 0$ 이면 포아송분포가 되며 $\tau > 0$ 이면 일반적 포아송 혼합분포가 된다.

Dean(1992)은 한단계 더 나아가 이항분포와 포아송분포를 포함하는 지수계에서의 이탈을 검색하는 검정통계량을 유도하였다. 본고에서는 이항분포의 혼합분포와 포아송분포의 혼합분포를 포함하는 일반적인 혼합분포를 다루며 3.3에서 다루고 있다.

2.3. 일반화 선형모형(Generalized Linear Model)에서의 이탈을 다루는 모형

Ganio and Schaffer(1992)는 일반화 선형모형에서의 이탈을 검색하는 우도비검정과 스코어검정을 유도하였다. 서로 독립인 종속변수 $Y_1, Y_2, \ldots Y_n$에 대한 모형은 일반화 선형모형에서 산포모수(dispersion parameter)가 설명변수에 의존적인 다음의 식(2-7)과 (2-8)로 나타낼 수 있다.

$$E(Y_i) = \mu_i = h(\eta_i); \quad \eta_i = x_i^t \beta \tag{2-7}$$

$$\text{var}(Y_i) = V(\mu_i)/(a_i \phi_i); \ \phi_i = g(\gamma_i); \ \gamma_i = \lambda + z_i^t \alpha \tag{2-8}$$

여기서 x_i, β는 각각 $p \times 1$ 설명변수와 회귀계수이며, h는 이차미분이 가능한 알려진 단조함수(monotone function)이며, V는 알려진 양함수(positive function)이며, a_i는 알려진 상수이며, g는 이차미분이 가능한 양함수이며, λ는 절편모수이며, z_i, α는 각각 $q \times 1$ 설명변수와 회귀계수이다.

(2-7)과 (2-8)을 결합하는 방법으로 Efron(1986)이 제시한 이중지수계(double exponential family)가 널리 사용되며 다음의 식(2-9)의 확장된 유사우도함수(extended quasilikelihood function)나 다음의 식 (2-10)의 의사우도함수(psudolikelihood function)에 근거하여 검정통계량을 유도할 수 있다.

$$l(\mu, \phi; y) = \sum (1/2)[\ \ln(\phi_i) - \phi_i D(y_i, \mu_i)] \tag{2-9}$$

$$l_p(\mu, \phi; y) = \sum (1/2)[\ \ln(\phi_i) - \phi_i R(y_i, \mu_i)] \tag{2-10}$$

여기서 $D(y_i, \mu_i)$는 일모수 지수계(one-parameter exponential family)에서 정의된 편차(deviance)이며, $R(y_i, \mu_i) = (y_i - \mu_i)^2/[\ V(\mu_i)/a_i]$이다.

3. 과산포검색을 위한 검정통계량

3.1. 이항분포에서의 이탈에 관한 검정통계량

Tarone(1979)은 과산포현상이 일어날 때, 이항분포의 적합성을 검정하기 위해 상관이 있는 이항분포(correlated binomial distribution), 베타이항분포(beta-binomial distribution)를 각각 대

립가설로 놓고 스코어 검정통계량을 유도하였다.

3.1.1. 베타이항분포로의 이탈에 관한 검정통계량

2.1.1에서 밝힌 것처럼 대립가설이 베타이항분포일 때 검정할 가설은 $H_0:\phi=0$ 대 $H_1:\phi>0$가 된다. $\phi=0$ 이면 자료들은 이항분포를 따름을 의미하며, 베타이항분포 하에서 로그우도함수를 구한 후, 스코어통계량을 유도하면 다음 식(3-1)과 같다.

$$Z = \left\{ \sum_{i=1}^{M} \frac{(x_i - n_i\hat{p})^2}{(\hat{p}\hat{q})} - \sum_{i=1}^{M} n_i \right\} \left\{ 2\sum_{i=1}^{M} n_i(n_i-1) \right\}^{-1/2} \underset{H_0}{\dot{\sim}} N(0,1) \qquad (3\text{-}1)$$

여기서 $\hat{p}$는 각 모체내 同腹仔가 죽을 확률 p의 추정치, $\hat{q}=1-\hat{p}$이며 '$\underset{H_0}{\dot{\sim}}$'은 '$\sim$'을 중심으로 좌측의 통계량이 우측의 분포를 귀무가설 하에서 점근적으로(asymptotically) 접근함을 의미한다. 위의 식(3-1)은 대립가설이 베타이항분포일 때 각 모체내 胎兒수가 동일하든 동일하지 않든, 점근적으로 최적검정통계량(asymptotically optimal test statistic)이 된다(Tarone, 1979).

3.1.2. 상관이 있는 이항분포로의 이탈에 관한 검정통계량

2.1.2에서 밝힌 것처럼 대립가설이 상관이 있는 이항분포일 때 검정할 가설은 $H_0:\theta=0$ 대 $H_1:\theta>0$ 이 된다. $\theta=0$ 이면 자료들은 이항분포를 따름을 의미하며, 상관이 있는 이항분포하에서 로그우도함수를 구한 후, 스코어통계량을 유도하면 다음 식(3-2)와 같다.

$$X_C^2 = \left\{ \sum_{i=1}^{M} (x_i - n_i\hat{p})^2 / (\hat{p}\hat{q}) - \sum_{i=1}^{M} n_i \right\}^2 \left\{ 2\sum_{i=1}^{M} n_i(n_i-1) \right\}^{-1} \underset{H_0}{\dot{\sim}} \chi(1) \qquad (3\text{-}2)$$

여기서 $\hat{p}$는 각 모체내 同腹仔가 죽을 확률 p의 추정치, $\hat{q}=1-\hat{p}$이다. 위의 식(3-2)는 대립가설이 상관이 있는 이항분포일 때 점근적으로 최적검정통계량(asymptotically optimal test statistic)이 된다(Tarone, 1979).

3.2. 포아송분포에서의 이탈에 관한 검정통계량

3.2.1. 음이항분포로의 이탈에 관한 검정통계량

2.2.1에서 밝힌 것처럼 과산포를 검정하는 가설은 $H_0:c=0$ 대 $H_a:c>0$이 된다. 이때, $c=0$는 $c\to 0$로 해석한다.

Collings and Margolin(1985)의 분류를 따라 표본들의 기대값이 확률표본형태일 때, 하나의 원점을 통과하는 회귀형태일 때, 일원배열형태일 때로 나누어 검정통계량을 제시하고자 한다.

(경우 1) 확률표본형태(random sample)

$$E(Y_i) = m \quad , \quad i = 1, \cdots , n$$

포아송분포로부터의 이탈을 다룬 표준적 검정으로 다음 식(3-3)과 같은 통계량을 이용하는 피셔의 분산검정(Fisher's variance test)이 있다.

$$S_A = \sum_{i=1}^{n}(Y_i - \overline{Y}_+)^2 / \overline{Y}_+ \underset{H_0}{\sim} \chi^2(n-1) \ , \quad \overline{Y}_+ = \sum_{i=1}^{n} Y_i / n \qquad (3\text{-}3)$$

Potthoff and Whittinghill(1966)은 위의 식(3-3)의 검정통계량에 입각한 검정은 국소최강력검정(locally most powerful test)임을 보였다.

(경우 2) 원점을 통과하는 회귀형태(regression through the origin)

$$E(Y_i) = \beta_i m \quad , \quad \beta_i = \text{알려진양수}, \quad i = 1, \cdots , n$$

포아송분포로부터의 이탈을 검색하기 위해 Rao(1952)는 피셔의 분산검정통계량을 적절히 변형시켜 다음 식(3-4)를 제안하였다.

$$S_B = \sum_{i=1}^{n}(Y_i - \beta_i \hat{m})^2 / \beta_i \hat{m} \underset{H_0}{\sim} \chi^2(n-1) \quad , \quad \hat{m} = \sum_{j=1}^{n} Y_j / \sum_{t=1}^{n} \beta_t \qquad (3\text{-}4)$$

Collings and Margolin(1985)은 분산이 평균의 이차함수 형태일 때 다음 식(3-5)의 T_B에 근거한 검정이 국소최적검정(locally optimal test)이며 S_B에 근거한 검정보다 점근적 상대효율(asymptotic relative efficiency) 측면에서 더 우수한 검정임을 보였다.

$$T_B = \sum_{i=1}^{n}(Y_i - \beta_i \hat{m})^2 / \overline{Y}_+ \qquad (3\text{-}5)$$

Kim and Park(1992)은 분산과 평균이 선형관계에 있을 때 다음 식(3-6)의 K_B에 근거한 검정은 국소최강력불편검정(locally most powerful unbiased test)이 되며 T_B에 근거한 검정보다 점근적으로 우수한 검정임을 보였다. (K_B와 S_B는 점근적으로 동등한 통계량이기 때문에 S_B에 근거한 검정이 T_B에 근거한 검정보다 우수한 통계량이 됨을 알 수 있다.)

$$K_B = \sum_{i=1}^{n}\{(Y_i - \beta_i \hat{m})^2 - Y_i\} / \beta_i \hat{m} \qquad (3\text{-}6)$$

이선호(1993)는 분산과 평균이 r차관계, 즉 $\sigma^2 = m + cm^r$일 때 다음 식(3-7)의 L_B에 근거한 검정이 국소최적검정임을 보였고 Lee, Park and Kim(1995)은 이 검정이 S_B, T_B에 근거한 검정들보다 점근적으로 우수함을 보였다.

$$L_B = \sum_{i=1}^{n} p_i^{r-1} \frac{(Y_i - \beta_i \hat{m})^2}{\beta_i \hat{m}} \quad , \quad p_i = \beta_i / \sum_{j=1}^{n} \beta_j \qquad (3\text{-}7)$$

(경우 3) 일원배열형태(one-way layout)

$$E(Y_{ij}) = m_i, \quad i = 1, \cdots, k, \; j = 1, \cdots, n_i, \; n = \sum_{i=1}^{k} n_i$$

자료가 일원배열형태일 때 Gart(1964)는 피셔의 분산검정에서 이용하는 검정통계량 S_A를 변형하여 다음 식(3-8)과 같은 검정통계량을 제안하였다.

$$S_C = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i+})^2 / \overline{Y}_{i+} \underset{H_0}{\sim} \chi^2(n-k), \; \overline{Y}_{i+} = \sum_{j=1}^{n_i} Y_{ij} / n_i \quad (3\text{-}8)$$

Collings and Margolin(1985)은 분산이 평균의 이차함수일 때 다음 식(3-9)의 T_C에 근거한 검정이 국소최강력불편검정이며 S_C에 근거한 검정보다 점근적으로 우수한 검정임을 보였다.

$$T_C = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i+})^2 / \overline{Y}_{++}, \quad \overline{Y}_{++} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij} / n \quad (3\text{-}9)$$

Kim and Park(1992)은 분산과 평균이 선형관계일 때 S_C에 근거한 검정이 국소최적검정이며 T_C에 근거한 검정보다 점근적으로 우수한 검정임을 보였다. 또한 이선호(1993)는 분산과 평균이 r차관계일 때 다음의 식(3-10)의 L_C에 근거한 검정이 국소최적검정임을 보였고 Lee, Park and Kim(1995)에 의해 이 검정이 S_C, T_C에 근거한 검정보다 점근적으로 우수한 검정임이 밝혀졌다.

$$L_C = \sum_{i=1}^{k}\sum_{j=1}^{n_i} \frac{\overline{Y}_{i+}^{\,r-2}(Y_{ij} - \overline{Y}_{i+})^2}{\overline{Y}_{++}^{\,r-1}} \quad (3\text{-}10)$$

3.2.2. 포아송 혼합분포로의 이탈에 관한 검정통계량

Dean and Lawless(1989)는 포아송회귀모형일 때 대립가설을 음이항분포에서 식(2-6)과 같은 일반적인 포아송 혼합분포로 확장하여 초포아송변이를 검색하는 통계량을 제시함으로써 Collings and Margolin(1985)의 검정통계량을 일반화하였다.

2.2.2에서 밝혔듯이 귀무가설이 포아송분포이고 대립가설이 포아송 혼합분포일 때의 가설은 $H_0: \tau = 0$ 대 $H_a: \tau > 0$이 된다. $\tau = 0$ 을 검정하는 통계량은 다음의 식(3-11)이 된다 (Lee, 1986; Cameron and Trivedi, 1986; Dean and Lawless, 1989).

$$T = \frac{1}{2}\sum_{i=1}^{n}\{(Y_i - \widehat{\mu_i})^2 - Y_i\} \quad (3\text{-}11)$$

단, $\widehat{\mu_i} = \mu_i(x_i; \hat{\beta})$이며, $\hat{\beta}$는 β의 귀무가설 하의 최우추정량이다.

μ_i가 고정되고 $n \to \infty$일 때 T와 점근적으로 동등한 검정통계량은 다음 식(3-12)가 된다.

$$T_1 = \frac{T}{(\frac{1}{2}\sum_{i=1}^{n}\hat{\mu}_i^2)^{1/2}} = \frac{\sum_{i=1}^{n}\{(Y_i - \hat{\mu}_i)^2 - Y_i\}}{(2\sum_{i=1}^{n}\hat{\mu}_i^2)^{1/2}} \underset{H_0}{\sim} N(0,1) \quad (3\text{-}12)$$

또한, n 이 고정되고 $\mu_i \to \infty$일 때 T와 점근적으로 동등한 검정통계량은 다음 식(3-13)이 된다.

$$T_2 = \sum_{i=1}^{n}(Y_i - \hat{\mu}_i)^2 / \overline{Y}_+ \underset{H_0}{\sim} \sum_{i=1}^{n}\lambda_i \chi_i^2(1) \quad (3\text{-}13)$$

단, $\chi_i^2(1)$, $i=1,\cdots,n$은 독립적인 자유도 1의 카이제곱확률변수이고, λ_i는 $V = \mu_+^{-1}W^{1/2}(I-H)W^{1/2}$의 i번째 고유치(eigenvalue), $H = W^{1/2}X(X'WX)^{-1}X'W^{1/2}$, X는 ij번째 원소가 $\mu_i^{-1}(\partial\mu_i/\partial\beta_j)$인 $n\times p$행렬, $W = \mathrm{diag}(\mu_1,\cdots,\mu_n)$, $\mu_+ = \sum\mu_i$이다. Dean and Lawless(1989)는 T_1, T_2가 소표본 하에서 정규분포에 보다 가깝도록 수정하여 각각 T_a, T_b를 제시하고 모의실험을 통하여 검정통계량 T_a와 T_b는 비슷한 검정력을 가지며, 적합도검정에 사용하는 피어슨의 카이제곱통계량보다 높은 검정력을 가짐을 보였다.

3.3. 지수계에 관한 과산포 검정통계량

Dean(1992)은 포아송분포와 이항분포를 포함하는 지수계(exponential family)에 대하여 과산포현상에 관한 검정통계량 S를 유도하고, S통계량은 Tarone(1979), Collings and Margolin(1985), Dean and Lawless(1989), Kim and Park(1992)등이 제시한 검정통계량을 일반화한 통계량임을 보였다.

n개의 독립적 관찰치 Y_i가 다음 식(3-14)의 확률밀도함수에 따른다고 하자.

$$f(Y_i;\ \theta_i) = \exp\{a(\theta_i)Y_i - g(\theta_i) + c(Y_i)\} \quad (3\text{-}14)$$

단, $\theta_i = \theta_i(x_i;\beta)$, x_i는 $p\times 1$ 공변수벡터, β 는 $p\times 1$ 회귀계수벡터이다.

위의 식(3-14)의 지수계 확률밀도함수에 과산포현상을 포함할 수 있는, 연속이고 독립인 변수 θ_i^*를 새로이 도입하면 다음 식(3-15)가 된다.

$$f(Y_i \mid \theta_i^*) = \exp\{a(\theta_i^*)Y_i - g(\theta_i^*) + c(Y_i)\} \quad (3\text{-}15)$$

여기서 θ_i^*는 평균과 분산이 식(3-16)과 같으며 식(3-17)이 가정되는 변수이다.

$$E(\theta_i^*) = \theta_i(x_i;\beta), \quad V(\theta_i^*) = \tau b_i(\theta_i) > 0 \quad (3\text{-}16)$$

$$E\{(\theta_i^* - \theta_i)^r\} = \alpha_r \quad ; \ \alpha_r = o_p(\tau) \quad , r\geq 3 \quad (3\text{-}17)$$

$\tau \to 0$일 때 $f(Y_i;\theta_i^*)$는 $f(Y_i;\theta_i)$가 된다. 따라서, 지수계에서 과산포현상을 검정하는 가설

은 $H_0:\tau=0$ 대 $H_1:\tau>0$이 된다.

이 때 $\tau=0$을 검정하는 스코어 검정통계량은 $\sum_{i=1}^{n} T_i(\hat{\theta}_i)$이 되는데 여기서 $\hat{\theta}_i$는 귀무가설 하의 θ_i의 최대우도추정치이고 $T_i(\theta_i) = b_i(a_i')^2\{(Y_i-\mu_i)^2 - (a_i')^{-2}(g_i''-a_i''Y_i)\}/2$이다. 또한 표준화된 스코어 검정통계량을 구하면 다음 식(3-18)의 S가 된다.

$$S = \sum_{i=1}^{n} T_i(\hat{\theta}_i) \;/\; \hat{V} \;\overset{\cdot}{\underset{H_0}{\sim}}\; N(0,1) \tag{3-18}$$

여기서 $\hat{V} = V(\hat{\theta}_1, \cdots, \hat{\theta}_n)$는 $\hat{\theta}_1, \ldots, \hat{\theta}_n$에서 계산된 $\sum_{i=1}^{n} T_i(\hat{\theta}_i)$의 점근적 표준오차의 추정량이다.

Dean(1992)은 귀무가설이 지수계인 분포이고 대립가설이 과산포현상을 보이며 지수계로부터 이탈된 분포일때, 통계량 S는 귀무가설이 포아송분포와 이항분포일때 유도된 검정통계량을 일반화시킨 검정통계량이 됨을 보였다. 또한, 통계량 S를 이용하면 대립가설이 될 수 있을 것으로 여겨지는 분포를 선택함에 따라 검정통계량을 결정할 수도 있고, 대립가설의 평균과 분산사이의 구조를 선택함에 따라 검정통계량을 결정할 수도 있음을 보이고 있다.

3.4. 일반화 선형모형(Generalized Linear Model)에서의 이탈에 관한 검정통계량

식(2-9)에서 만약 $\alpha=0$ 이면 일반화 선형모형이 되며 따라서 검정하려는 가설은 $H_0:\alpha=0$이 된다. 식(2-9)의 확장된 유사우도함수(extended quasilikelihood function)와 식(2-10)의 의사우도함수(pseudolikelihood function)에 근거한 우도비검정은 다음의 식(3-19)와 식(3-20)의 검정통계량이 된다 (Ganio and Schaffer(1992)).

$$DLR = 2[\; l(\hat{\mu},\ \hat{\phi}; y) - l(\ \hat{\mu}_0,\ \hat{\phi}_0; y)] \tag{3-19}$$

$$PLR = 2[\; l_{\mathcal{K}}\hat{\mu},\ \hat{\phi}; y) - l_{\mathcal{K}}\ \hat{\mu}_0,\ \hat{\phi}_0; y)] \tag{3-20}$$

여기서 $\hat{\mu}$, $\hat{\phi}$는 전모형(full model) 하에서의 최우추정량이며 $\hat{\mu}_0$, $\hat{\phi}_0$는 귀무가설 하에서의 최우추정량이다. 또한 확장된 유사우도함수와 의사우도함수에 근거한 스코어검정통계량은 다음의 식(3-21)과 식(3-22)와 같아진다 (Ganio and Schaffer(1992)).

$$DS = (2\ \overline{D}^2)^{-1}\sum \hat{D}_i z_i(\sum z_i z_i^t)^{-1}\sum \hat{D}_i z_i^t \tag{3-21}$$

$$PS = (2\ \overline{R}^2)^{-1}\sum \hat{R}_i z_i(\sum z_i z_i^t)^{-1}\sum \hat{R}_i z_i^t \tag{3-22}$$

여기서 $\hat{D}_i = D(y_i,\ \hat{\mu}_i)$는 귀무가설 하에서 적합된 i번째 편차통계량(deviance statistic)이며 $\overline{D} = \sum \hat{D}_i/n$이며, $\hat{R}_i = R(y_i,\ \hat{\mu}_i)$는 귀무가설 하에서 적합된 i번째 피어슨 적합도검정통계량(Pearson goodness of fit statistic)이며 $\overline{R} = \sum \hat{R}_i/n$이다.

4. 추후 연구 과제

초이항변이에 관한 이항분포의 혼합분포에서 지금까지는 반응변수를 이진자료(binary data)로, 즉 죽은 胎兒(dead fetus)는 1로, 사는 胎兒(viable fetus)는 0으로 놓고 관측치를 얻었지만, Ryan(1992)이 제시한 사는 胎兒를 정상상태의 胎兒(normal fetus)와 기형상태의 胎兒(malformed fetus)로 나누어서 반응변수를 삼분화하여 관측치를 얻는 방법도 고려해 볼 수 있다. 이경우 McCullagh and Nelder(1989)의 5장에 소개된 다항자료에 대한 모형을 이용하여 산포모수(dispersion parameter)를 첨가하면 분석이 가능할 것이라 생각된다.

과산포 검색을 위해 본문에서 제시한 검정통계량들은 귀무가설에 기초한 스코어통계량들이었다. 최근에 Boos(1993)는 초이항변이가 일어나는 용량-반응자료(dose- response data)의 과산포 검색을 위해, 점근적 검정중의 하나이며 대립가설에 기초해서 검정통계량을 유도하는 왈드검정(Wald test)을 제시하고, 왈드검정과 스코어통계량을 이용한 검정은 서로 비슷한 결과를 가짐을 보여주었다. 따라서, 앞으로 왈드검정에 관한 연구를 함으로써, 스코어통계량을 이용해서 얻은 검정결과를 확인시켜 줄 수 있는 검정통계량을 개발할 수 있다고 생각한다.

참 고 문 헌

[1] 이선호 (1993). 포아송분포로부터 음이항분포로의 이탈에 대한 검정통계량의 확장, 「통계학연구」, 제22권, 171-189.

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

[3] Boos, D. D. (1993). Analysis of dose-response data in the presence of extrabinomial variation, *Applied Statistics*, Vol. 42, 173-183.

[4] Cameron, A. C. and Trivedi, P. K. (1986). Econometric models based on count data: comparisons and applications of some estimators and tests, *Journal of Applied Econometrics*, Vol. 1, 29-53.

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

[6] 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, 411-418.

[7] Cox, D. R. (1983). Some remarks on overdispersion, *Biomerika*, Vol. 70, 269-74.

[8] Dean, C. and Lawless, J. F. (1989). Tests for detecting overdispersion in Poisson regression models, *Journal of the American Statistical Association,* Vol. 84, 467-472.

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

[10] Efron, B. (1986). Double Exponential Families and Their Use in Generalized Linear Regression, *Journal of the American Statistical Association*, Vol. 81, 709-721.

[11] Gart, J. J. (1964). The analysis of Poisson regression with an application in virology, *Biometrika*, Vol. 51, 517-521.

[12] Ganio, L.M., and Schaffer, D.W. (1992). Diagnostics for Overdispersion, *Journal of the American Statistical Association*, Vol. 87, 795-804.

[13] Haseman, J. K. and Soares, E. R. (1976). The distribution of fetal death in control mice and its implications of statistical tests for dominant lethal effects, *Mutation Research*, Vol. 41, 277-288.

[14] Haseman, J. K. and Kupper, L. L. (1979). Analysis of dichotomous response data from certain toxicological experiments, *Biometrics*, Vol. 35, 281-293.

[15] Kim, B. S. (1988). Some remarks on locally most powerful unbiased tests for the detection of the mixtures of Poisson or binomial distributions, *Communications in Statistics*, Vol. A17, 3733-3741.

[16] Kim, B. S. and Park, C. (1992). Some remarks on testing goodness of fit for the Poisson assumption, *Communications in Statistics*, Vol. A21, 979-995.

[17] Kim, B. S. and Margolin, B. H. (1992). Testing goodness of fit of a multinomial model against overdispersed alternatives, *Biometrics*, Vol. 48, 711-719.

[18] Kupper, L. L. and Haseman, J. K. (1978). The use of a correlated binomial model for the analysis of certain toxicological experiments, *Biometrics*, Vol. 34, 69-76.

[19] Lee, L. F. (1986). Specification test for Poisson regression models, *International Economic Review*, Vol. 27, 689-706.

[20] Lee, S., Park, C., and Kim, B. S. (1995). Tests for detecting overdispersion in Poisson models, To appear in *Communications in Statistics, A*.

[21] Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika*, Vol. 73, 13-22.

[22] Liang, K. Y. and McCullagh, P. (1993). Case studies in binary dispersion, *Biometrics*, Vol. 49, 623-630.

[23] Lockhart, A. C., Piegosch, W. W., and Bishop, J. B. (1992). Assessing overdispersion and dose-response in the male dominant lethal essay, *Mutation Research*, Vol. 272, 35-58.

[24] McCullagh, P., and Nelder, J.A. (1989). *Generalized Linear Models*, second edition, Chapman and Hall, New York:.

[25] 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, 3779-3783.

[26] Paul, S. R. (1987). On the beta-correlated binomial distribution - A three parameter generalization of the binomial distribution, *Communications in Statistics*, Vol. A16, 1473-1478.

[27] Potthoff, R. F. and Whittinghill, M. (1966). Testing for homogeneity - II. The Poisson distribution, *Biometrika*, Vol. 53, 183-190.

[28] Rao, C. R. (1952). *Advanced statistical methods in biometric research*, John Wiley, New York.

[29] Ryan, L. (1993). Potency measures for developmental toxicity, *Environmetrics*, Vol. 4, 507-518.

[30] Tarone, R. E. (1979). Testing the goodness of fit of the binomial distribution *Biometrika*, Vol. 66, 585-590.

[31] Williams, D. A. (1975). The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity, *Biometrics*, Vol. 31, 949-952.

Overdispersion in Count Data - a Review

Kim, Byung Soo[4], Oh, Kyong Joo[5], and Park, Cheolyong[6]

Abstract

The primary objective of this paper is to review parametric models and test statistics related to overdispersion of count data. Poisson or binomial assumption often fails to explain overdispersion. We reviewed real examples of overdispersion in count data that occurred in toxicological or teratological experiments. We also reviewed several models that were suggested for implementing the extra-binomial variation or hyper-Poisson variability, and we noted how these models were generalized and further developed. The approaches that have been suggested for the overdispersion fall into two broad categories. The one is to develop a parametric model for it, and the other is to assume a particular relationship between the variance and the mean of the response variable and to derive a score test statistic for detecting the overdispersion. Recently, Dean(1992) derived a general score test statistic for detecting overdispersion from the exponential family.

4) Department of Applied Statistics, Yonsei University, Seoul, 120-749, KOREA.

5) Institute of Marketing Strategy, Diamond Ad. Ltd., Boryung Bldg., 66-21, Wonnam-Dong, Chongro-Ku, Seoul, 110-450, KOREA.

6) Department of Statistics, Keimyung University, Taegu, 704-701, KOREA.