Proceedings of Joint Conference of The Korean Data Analysis Society and Korean Data And Information Science Society, October 17-18, 2003, pp. 1-11

A Study on the Estimation of Confidence Intervals for Discrete Distribution

김대학[1], 오광식[2], 이상복[3]

요 약

일반적으로 모수에 대한 신뢰구간 추정량이 점 추정량보다 훨씬 더 선호되고 있으며 많이 알려져 있다. 그러나 이산형 분포의 경우에는 주로 대 표본 근사 이론에 입각한 근사 신뢰구간이 많이 사용되고 있다. 본 논문에서는 여러 가지 이산형 분포 가운데에서 가장 많이 활용되고 있는 이항분포와 포아송 분포의 모수에 대한 다양한 신뢰구간 추정량들을 소개하고 대 표본 근사 이론에 의한 신뢰구간뿐만 아니라 소 표본의 경우에도 유용하게 이용될 수 있는 신뢰구간 등을 살펴보고 이들 신뢰구간들을 비교하였다.

Keywords : 신뢰구간, 소 표본, 이산형분포, 이항분포, 포아송분포, 대표본 근사

1. 서론

대부분의 통계학 교재에서 이항분포와 포아송분포의 중요성과 응용성을 소개하고 있다. 이들은 우리의 생활가운데 자주 사용되는 대표적인 이산형 분포로서 여러 가지 성질들이 잘 알려져 있다. 포아송분포는 발생빈도가 드문 희귀한 사건의 횟수의 관찰 등 다양한 현상에서 발견되며 생물학 등 제 학문분야에 응용성이 높은 이산형 분포이고 관찰 값이 성공과 실패로 구분되는 이항인 경우 이항분포의 성공의 모비율 p는 모집단에서 어떤 특정한 속성을 갖는 개체의 비율을 나타내는 모수로서 이의 추정과 관련하여 많은 응용이 이루어 지고 있다.

그러나 각 분포의 모수에 대한 신뢰구간 추정량의 차원에서 보면 그다지 상세히 다루어지지 않고 대부분의 경우 대표본 근사이론을 이용한 근사 신뢰구간만을 소개하고 있다. 본 논문에서는 이항분포와 포아송분포의 여러 신뢰구간추정량들을 각각 소개하고 2장에서는 이항분포의 신뢰구간, 3장에서는 포아송분포의 신뢰구간들을 설명하였으며 4장에서는 정확신뢰구간, 스코어 신뢰구간 그리고 연속성을 수정한 신뢰구간 등을 이용하여 소표본의 경우에 신뢰구간을 구하고 근사 신뢰구간과 비교하였다.

[1]712-702, 경북 경산시 하양읍 금락리, 대구가톨릭대학교 환경정보학부 정보통계학전공 교수
E-Mail : dhkim@cu.ac.kr
[2]712-702, 경북 경산시 하양읍 금락리, 대구가톨릭대학교 환경정보학부 정보통계학전공 교수
[3]712-702, 경북 경산시 하양읍 금락리, 대구가톨릭대학교 환경정보학부 정보통계학전공 교수

2. 이항분포의 모비율 신뢰구간 추정

크기가 n인 성공과 실패로 구분될 수 있는 랜덤표본 $X_1, X_2, \cdots, X_n$에서 특정한 속성을 갖는 개수를 X라 두고 $\hat{p}=X/n$을 표본비율이라 놓자. 모비율 p의 신뢰구간 구축과 관련된 연구는 최근 Chen(1990), Leemis와 Trivedi(1996) 등에 의해 여러 성질들이 계속 연구되고 있다. Vollset(1993)은 모비율 p의 신뢰구간을 구축하는 다양한 방법들을 비교하였으며, Leemis와 Trivedi(1996)은 모비율의 신뢰구간을 구축함에 있어 정규근사 방법과 포아송근사 방법을 비교하여 표본비율이 낮은 경우 포아송 근사 신뢰구간을 활용하는 것이 좋음을 보인바 있다. 특히 가장 최근의 연구로 Agresti와 Coull(1998)을 들 수 있는데, 그들은 Clopper와 Pearson(1934)의 "정확"신뢰구간보다 근사이론을 이용한 신뢰구간이 더 나을 수 있다는 것을 모의실험으로 보인바 있다.

p의 $100(1-\alpha)\%$ 정확신뢰구간을 구하기 위해서는 관찰 값 x에 대해 유의수준 $\alpha/2$에서 귀무가설 $H_0: p=p_0$을 기각하지 않는 모든 p_0를 계산하면 된다. 이때 얻어지는 모든 p_0 중 최소값이 신뢰구간의 하한, 최대값이 신뢰구간의 상한이 된다. 즉, 최소값 p_L과 최대값 p_U는

$$P(X\geq x|p=p_L)=\sum_{k=x}^{n}\binom{n}{k}p_L^k(1-p_L)^{n-k}\leq \alpha/2,$$

$$P(X\leq x|p=p_U)=\sum_{k=0}^{x}\binom{n}{k}p_U^k(1-p_U)^{n-k}\leq \alpha/2$$

로 계산된다. 정확신뢰구간을 구하기 위해서는 상당한 계산이 요구되지만 Blyth(1986)나 Hald(1952)의 쉬운 계산방법이 존재한다. 정확신뢰구간(EX)에 대한 닫힌 형태는 다음과 같다.

$$\text{EX}: \left[1+\frac{n-x+1}{xF_{2x,2(n-x+1),1-\alpha/2}}\right]^{-1} < p < \left[1+\frac{n-x}{(x+1)F_{2(x+1),2(n-x),\alpha/2}}\right]^{-1} \tag{2.1}$$

여기서 $F_{a,b,c}$는 자유도 a, b를 따르는 F분포의 $100(1-c)\%$ 분위점이다. 이 신뢰구간을 Clopper-Pearson(1934)의 정확신뢰구간이라 한다.

이제 표본비율 $\hat{p}$의 분산을 $V(p)=\sqrt{p(1-p)/n}$로 놓으면, 확률변수 $(p-\hat{p})^2$는 정규근사에 의해 $(p-\hat{p})^2=cV(\tilde{p})$로 표현할 수 있다. 여기서, c는 $\chi^2(1)$분포의 $100(1-\alpha)\%$ 분위점이며, $\tilde{p}$는 p에 대한 표현식이다. 이제 $\tilde{p}$에 적당한 값을 대입하여 위의 이차방정식을 풀면 두 값 $[p_L, p_U]$를 얻게된다. 잘 알려진 바대로 $\tilde{p}=\hat{p}$을 대입하여 방정식을 풀면, 모비율 p에 대한 $100(1-\alpha)\%$ 신뢰구간은 다음과 같이 주어진다.

$$\text{WA}: \hat{p}\pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} \tag{2.2}$$

여기서 z_α 는 표준정규분포의 $100(1-\alpha)\%$ 분위점이다. 이 신뢰구간을 Wald 신뢰구간이라고 하

며, 대부분의 통계학 책에서 소개하고 있다. 또한 $(p-\hat{p})^2 = cV(\hat{p})$에서 $\tilde{p}=p$를 대입하여 이차방정식을 풀면 다음과 같은 Score 신뢰구간이 주어진다.

$$\text{SC}: \frac{X+z^2_{\alpha/2}/2 \pm z_{\alpha/2}\sqrt{X-X^2/n+z^2_{\alpha/2}/4}}{n+z^2_{\alpha/2}} \tag{2.3}$$

정규근사의 또 다른 방법으로 arcsin 변환을 이용하는 것이 있다. 이는 $\hat{p}$의 분산이 모비율 p의 함수이므로, 분산이 p에 의존하지 않는 함수를 유도하는 과정에서 arcsin 변환을 취한 신뢰구간은 다음과 같이 계산된다.

$$\text{AS}: \sin^2[\arcsin(\hat{p}^{\frac{1}{2}}) \pm \frac{Z_{\alpha/2}}{2\sqrt{n}}] \tag{2.4}$$

Chen(1990)은 arcsin 변환을 통하여, 정규분포로의 수렴속도가 빨라짐을 보인 바 있다. 정규근사 방법은 이산형 분포를 따르는 통계량을 연속확률분포인 정규분포로 근사함으로서 약간의 오차가 존재하리라고 예상할 수 있다. 그러므로 수정된 값을 대입시키는 연속성 수정(continuity correction)으로 오차를 보정하는 방법을 생각할 수 있다. 즉, 이항분포 X에 대해 $P(X=a)$의 정규근사는 $P(a-0.5 \le Y \le a+0.5)$이다. 여기서, Y는 X와 평균과 분산이 같은 정규분포이다. 이에 따라, 식 (2.2)에 연속성 수정을 하면 다음과 같은 신뢰구간이 주어진다.

$$\text{WA2}: \frac{X}{n} \pm \left\{ \frac{z_{\alpha/2}}{\sqrt{n}} \sqrt{\left(\frac{X}{n}\right)\left(1-\frac{x}{n}\right)} + \frac{1}{2n} \right\} \tag{2.5}$$

또한, Score 신뢰구간에 연속성 수정을 적용하면

$$\text{SC2}: \frac{(X \pm 0.5)+z^2_{\alpha/2}/2 \pm z_{\alpha/2}\sqrt{(X \pm 0.5)-(X^2 \pm 0.5)^2/n+z^2_{\alpha/2}/4}}{n+z^2_{\alpha/2}} \tag{2.6}$$

이다. 그런데, 평균이 np이고 분산이 $np(1-p)$인 정규분포 Y에 대해 비교적 큰 n에 대해 $P(|X-np| \le z_\alpha \sqrt{np(1-p)}) \approx 1-\alpha$ 가 성립하고 또한 $(X/n)(1-X/n)$는 $p(1-p)$로 확률수렴 하므로 $P(|X-np| \le z_\alpha \sqrt{n\hat{p}(1-\hat{p})}) \approx 1-\alpha$ 가 된다. 그러나 $\sqrt{\hat{p}(1-\hat{p})}$는 $\sqrt{p(1-p)}$를 상당히 하향추정하는 경향이 있다(Blyth와 Still, 1983). 그러므로 정규근사값 $z_{\alpha/2}$를 적당한 $\gamma z_{\alpha/2}$로 변환한 신뢰구간을 생각할 수 있다. 이에 대해 T-분포를 고려하여

$$\text{BS}: \frac{X}{n} \pm \left\{ \frac{t(n-1,\alpha/2)}{\sqrt{n}} \sqrt{\left(\frac{X}{n}\right)\left(1-\frac{X}{n}\right)} + \frac{1}{2n} \right\} \tag{2.7}$$

로 수정할 수 있다. Blyth와 Still(1983)은 $\gamma > 1$ 을 만족하는 최적의 식을 다음과 같이 제시하였다.

$$\gamma^2=\frac{n}{n-z_{\alpha/2}^2-2z_{\alpha/2}/\sqrt{n}-1/n+O1/n^3\sqrt{n)}},\ r\approx\sqrt{\frac{n}{n-z_{\alpha/2}^2-2z_{\alpha/2}/\sqrt{n}-1/n}}$$

이에 따라 Blyth와 Still(1983)의 조정된 신뢰구간은 다음과 같이 주어진다.

$$\text{BS2}:\frac{X}{n}\pm\left\{\frac{z_{\alpha/2}}{\sqrt{n-z_{\alpha/2}^2-2z_{\alpha/2}/\sqrt{n}-1/n}}\sqrt{\left(\frac{X}{n}\right)\left(1-\frac{X}{n}\right)+\frac{1}{2n}}\right\} \tag{2.8}$$

한편, 포아송분포의 모수에 대한 정확신뢰구간으로부터 모비율 p의 신뢰구간을 유도하는 방법을 살펴보자. 포아송분포의 모수를 μ라 했을 때, μ의 정확신뢰구간의 상한과 하한은 각각 다음을 만족하는 μ_U와 μ_L이다.

$$P(X\geq x\mid\mu=\mu_L)=\sum_{k=x}^{n}\frac{(e^{-\mu_L}\mu_L^k)}{k!}\leq\alpha/2,\ P(X\leq x\mid\mu=\mu_U)=\sum_{k=0}^{x}\frac{(e^{-\mu_u}\mu_U^k)}{k!}\leq\alpha/2$$

여기서 포아송분포의 누적확률인 $P(X\leq x\mid\mu)$는 자유도가 $v=2(1+x)$인 카이제곱분포 형태로 다음과 같이 표현된다.

$$\sum_{k=0}^{x}\frac{(e^{-\mu}\mu^k)}{k!}=P\left(\chi_v^2>2\mu\right)$$

위의 사실과 이항분포의 포아송근사를 활용하여 $np\rightarrow\mu$로 놓으면, 다음과 같이 포아송분포의 정확확률로 모비율의 신뢰구간을 구할 수 있게 된다. 즉

$$P(X\geq x\mid\mu=np_L)=\sum_{k=x}^{\infty}\frac{(e^{-np_L}(np_L)^k)}{k!}=1-\sum_{k=0}^{x-1}\frac{(e^{-np_L}(np_L)^k)}{k!}\leq\alpha/2$$

로 되고 포아송분포의 누적확률은 카이제곱분포 $P\chi_{2x}^2\geq 2np_L=1-\alpha/2$가 되어 신뢰구간의 하한 p_L을 얻을 수 있으며, 비슷한 방법으로 상한 p_U도 다음과 같이 계산된다.

$$\text{PO}:p_L=\frac{1}{2n}\chi_{2x,\alpha/2}^2\ ,\ p_U=\frac{1}{2n}\chi_{2(x+1),1-\alpha/2}^2 \tag{2.9}$$

포아송 신뢰구간의 성질에 대해서는 Leemis와 Trivedi(1996)에 의해 자세히 언급되었다. 특히 그들의 연구에서 표본의 크기와 표본비율이 주어질 때 정규근사를 사용하여야 하는지 포아송근사를 사용하여야 하는지의 경계점을 제시하기도 하였다.

Score 신뢰구간의 중간 추정값은 95%에서 $\tilde{p}=(X+z^2/2)/(n+z^2)\approx(X+2)/(n+4)$로 주어진다. 이는 모비율을 2개의 성공과 2개의 실패를 더한 값으로 이동(shift)하여 추정한다는 의미를 지닌다. Agresti와 Coull(1998)은 $(X+2)/(n+4)$를 이용한 신뢰구간이 Score 신뢰구간과 매우 유사

한 결과를 유도함을 모의실험으로 보였다. 그런데 비슷한 이동추정으로 베이지안 추정을 생각할 수 있다. 사전분포로 Beta (β, γ)를 이용하면, 모비율 p의 베이지안 추정량은 $(X+\beta)/(n+\beta+\gamma)$이다. 일반성을 잃지 않고 $\beta=\gamma$로 두면 $(X+\beta)/(n+2\beta)$이므로 베이지안 추정량을 이용한 신뢰구간은 다음과 같다.

$$\left(\frac{X+\beta}{n+2\beta}\right)\pm\frac{z_{\alpha/2}}{\sqrt{n}}\left[\left(\frac{X+\beta}{n+2\beta}\right)\left(1-\frac{X+\beta}{n+2\beta}\right)\right]$$

Chen(1990)은 $\beta=kz_{\alpha/2}$에 대해서 최적의 k를 제시하였는데 $\beta=z_{\alpha/2}^2/2$를 추천하고 있다. 그러므로 Chen(1990)의 결과를 이용한 신뢰구간은 다음과 같이 주어진다.

$$\text{CH}:\left(\frac{X+z_{\alpha/2}^2/2}{n+z_{\alpha/2}^2}\right)\pm\frac{z_{\alpha/2}}{\sqrt{n}}\left[\left(\frac{X+z_{\alpha/2}^2/2}{n+z_{\alpha/2}^2}\right)\left(1-\frac{X+z_{\alpha/2}^2/2}{n+z_{\alpha/2}^2}\right)\right] \tag{2.10}$$

이 신뢰구간은 Score 신뢰구간의 중간값을 이용한 신뢰구간과 매우 유사함을 알 수 있다.

3. 포아송분포의 평균 μ의 신뢰구간 추정량

이제 평균이 μ인 포아송분포를 따르는 n개의 랜덤표본 $X_1, X_2, \cdots, X_n$이 주어졌을 때, 기본적인 통계적 추론 중의 하나는 모수 μ에 대한 신뢰구간을 추정하는 것이다. 이에 대해 대부분의 통계학 책에서는 표본평균 $\overline{X}$를 이용한

$$\overline{X}\pm z_{1-\alpha/2}\sqrt{\overline{X}/n}$$

과 같은 신뢰구간을 소개하고 있는데 이를 Wald 유형의 신뢰구간이라 부른다. 여기서 $z_{1-\alpha}$는 표준정규분포의 $100(1-\alpha)\%$ 분위수이다. Wald 유형의 신뢰구간은 포아송분포가 지니는 이산성 때문에 주어진 명목수준을 하향 추정하는 경향이 있다. 이와는 달리 대조적인 신뢰구간으로 "정확"(exact)신뢰구간을 들 수 있다. Agresti와 Coull(1998)이 이항분포에서 실험한 결과를 살펴보면 대부분의 정확신뢰구간은 주어진 명목수준을 상당부분 상향 추정하는 경향이 있다는 것이다.

랜덤표본 $X_1, X_2, \cdots, X_n$의 합 X는 평균이 $\nu=n\mu$인 포아송분포를 따른다. ν의 $100(1-\alpha)\%$ 정확신뢰구간을 구하려면, 관찰값들의 합 x에 대해 유의수준 $\alpha/2$에서 귀무가설 $H_0: \nu=\nu_0$을 기각하지 않는 모든 ν_0를 계산해야 된다. 이때 얻어지는 모든 ν_0중 최소값을 정확신뢰구간의 하한으로 하고, 최대값을 정확신뢰구간의 상한으로 놓는다. 즉, 최소값 ν_L과 최대값 ν_U는 다음과 같다.

$$P(X\geq x|\nu=\nu_L)=\sum_{k=x}^{\infty}e^{-\nu_L}\nu_L^k/k!\leq\alpha/2,\ P(X\leq x|\nu=\nu_U)=\sum_{k=0}^{x}e^{-\nu_U}\nu_U^k/k!\leq\alpha/2$$

이와 같은 신뢰구간을 구하기 위해서 상당한 계산이 요구된다. 그러나 포아송분포의 누적확률 $P(X\le x|\nu)$는 자유도가 $v=2(1+x)$ 인 카이제곱분포 형태로 표현됨을 이용하여 정확신뢰구간을 쉽게 유도할 수 있다. 간단한 표현을 위하여 신뢰구간의 하한을 L 신뢰구간의 상한을 U라 표기하자. μ에 대한 정확신뢰구간의 상한과 하한은 ν에 대해 표본의 크기 n으로 나누어 구한다. 이러한 정확신뢰구간(EX)은 다음과 같다.

$$\text{EX: } \mu_L=\frac{1}{2n}\chi^2_{2(x+1),\alpha/2},\ \text{EX: } \mu_U=\frac{1}{2n}\chi^2_{2(x+1),1-\alpha/2} \tag{3.1}$$

정확신뢰구간을 구할 때는 카이제곱분포를 이용할 수 있으나, 다음과 같은 카이제곱분포의 정규근사에 의하여 정규분포로부터 신뢰구간을 구할 수 있다. 즉,

$$\chi^2_{\nu,\alpha}=\nu\left(1-\frac{2}{9\nu}+z_\alpha\sqrt{\frac{2}{9\nu}}\right)^3$$

위의 근사관계식에 의해 다음과 같은 카이제곱근사(CH) 신뢰구간을 얻을 수 있다.

$$\text{CH: } \mu_L=\bar{x}\left[1-\frac{1}{9x}-\frac{z_{1-\alpha/2}}{3\sqrt{x}}\right]^3,\ \text{CH(U): } \mu_U=\left(\bar{x}+\frac{1}{n}\right)\left[1-\frac{1}{9(x+1)}+\frac{z_{1-\alpha/2}}{3\sqrt{x+1}}\right]^3 \tag{3.2}$$

위의 신뢰구간은 근본적으로 정확신뢰구간(EX)에서 카이제곱분포를 정규근사 한 것에 불과하므로 정확신뢰구간과 큰 차이를 보이지 않을 것으로 예상된다. Wald 유형 신뢰구간(WA)은 모수 μ에 대한 최대우도추정량 $\bar{X}$ 와 정규근사를 이용하여 쉽게 얻을 수 있다. 즉,

$$\text{WA}:\ \bar{X}\pm z_{1-\alpha/2}\sqrt{\bar{X}/n} \tag{3.3}$$

포아송분포는 평균과 분산이 일치하는 성질을 이용하면 분산을 1차 적률로 추정한 신뢰구간을 구축할 수 있으나, 분산을 2차 적률로 추정하면 또 다른 신뢰구간을 얻을 수 있다. 즉 표본분산 S^2을 이용하여 다음과 같은 Wald 신뢰구간(W2)을 구축할 수 있다.

$$\text{W2}:\ \bar{X}\pm z_{1-\alpha/2}S/\sqrt{n} \tag{3.4}$$

그런데, Wald 신뢰구간(WA)처럼 분산을 추정하는 것이 아니라, Wilson(1927)이 지적한 바처럼, 분산을 그대로 모수 μ로 추정하여, 2차 방정식을 풀면 신뢰구간을 구할 수 있다. 이를 Score 신뢰구간(SC)이라 한다. Score 신뢰구간(SC)은 다음과 같다.

$$\text{SC}:\ \bar{x}\left[1+\frac{1}{2x}z^2_{1-\alpha/2}\left[1\pm\left(1+\frac{4x}{z^2_{1-\alpha/2}}\right)^{1/2}\right]\right] \tag{3.5}$$

한편, 위의 Score 신뢰구간에 연속성 수정을 가하여 다른 신뢰구간을 구축할 수 있다. 비교적 큰

모수 ν에 대해 포아송 확률을 다음의 정규근사로 표현할 수 있다.

$$\sum_{k=0}^{x}(e^{-\nu}\nu^{k})/k! = P\left(Z < \frac{x+1/2-\nu}{\sqrt{\nu}}\right)$$

위의 관계로부터 아래와 같은 다른 형태의 Score 신뢰구간(S2)을 얻을 수 있다. 즉,

$$\begin{aligned}
\text{S2(L)} &: \bar{x} - \frac{1}{2n} + \frac{1}{2n}z^2_{1-\alpha/2} - z_{1-\alpha/2}\sqrt{\frac{\bar{x}}{n} - \frac{1}{2n^2} + \frac{z^2_{1-\alpha/2}}{4n^2}} \\
\text{S2(U)} &: \bar{x} + \frac{1}{2n} + \frac{1}{2n}z^2_{1-\alpha/2} + z_{1-\alpha/2}\sqrt{\frac{\bar{x}}{n} + \frac{1}{2n^2} + \frac{z^2_{1-\alpha/2}}{4n^2}}
\end{aligned} \tag{3.6}$$

또한 연속성 수정의 근사보다 좀 더 정확한 근사식은 Sahai와 Khurshid(1993)의 Molenaar 근사이다. 이에 의하면 Sahai와 Khurshid(1993)의 Molenaar 신뢰구간(MO)은 다음과 같다.

$$\text{MO: } \bar{x} + \frac{2z^2_{\alpha/2}+1}{6n} \pm \left[\frac{1}{2n} + \frac{1}{n}\sqrt{z^2_{\alpha/2}\left(x \pm \frac{1}{2} + \frac{z^2_{\alpha/2}+2}{18}\right)}\right] \tag{3.7}$$

Molenaar 근사식에 의해 신뢰구간을 구축하면 연속성수정보다 포아송분포의 누적확률에 더 유사하므로 정확신뢰구간과 매우 비슷한 신뢰구간을 구축하게 될 것이다. 신뢰구간을 구하는 또 다른 방법으로 테일러 전개와 분산안정화 방법을 활용하는 것이다. 이에 대해 본 연구에서는 Sahai와 Khurshid(1993)의 Bartlett 방법(BA), Anscombe 방법(AN), Freeman과 Tukey 방법(FT), Hald 방법(HA)을 고려하였다. 각 방법의 신뢰구간은 다음과 같이 표기할 수 있다.

$$\text{BA: } \frac{1}{n}\left(\sqrt{x} \pm \frac{1}{2}z_{\alpha/2}\right)^2 \tag{3.8}$$

$$\text{AN: } \frac{1}{n}\left(\sqrt{x+\frac{3}{8}} \pm \frac{1}{2}z_{\alpha/2}\right)^2 - \frac{3}{8n} \tag{3.9}$$

$$\text{FT: } \frac{1}{4n}\left[\sqrt{x}+\sqrt{x+1} \pm z_{\alpha/2})^2 - 1\right] \tag{3.10}$$

$$\text{HA: } \frac{1}{n}\left[\sqrt{x-\frac{1}{2}} \pm \frac{z_{\alpha/2}}{2}\right]^2 + \frac{1}{2n} \tag{3.11}$$

4. 소 표본 신뢰구간의 계산과 비교

본 절에서는 2절과 3절에서 소개한 10가지 이항분포의 모비율에 대한 신뢰구간과 11가지 포아송분포의 모수에 대한 신뢰구간 추정방법을 이용하여 소 표본의 경우 각 신뢰구간을 계산하고자 한다. 표본의 크기가 커짐에 따라 정규근사 신뢰구간과 기타의 신뢰구간들의 추정치 사이가 크지 않으리라는 예상을 확인하기 위하여 비교적 대표본이라 볼 수 있는 30과 50의 경우도 포함시켰다.

신뢰구간의 명목수준은 가장 많이 사용되는 $1-\alpha = 0.95$를 상정하였다.

표현의 편리를 위하여 각 표에서의 각 신뢰구간 추정방법들은 가능한 범위 내에서 한글로 표현하였고 괄호안에 약자를 표기하였다. 또한 각 신뢰구간의 해당되는 식 [식 (2.1)~식 (2.10)]과 [식 (3.1)~(3.11)]의 번호를 삽입하여 알아보기 쉽게 하였다.

[표 1]은 표본의 수가 아주 작은 n=5일 때 이항분포의 여러 표본비율의 경우에 각 방법들의 신뢰구간의 상한과 하한을 나타내었다. [표 1]을 살펴보면 정규근사 방법의 결과는 정확 방법과는 상당한 차이를 보이고 있으며 정규근사의 연속성 수정을 가한 방법들 간에는 차이가 거의 나타나지 않음을 알 수 있다. 포아송 근사의 경우는 상한이 1의 값을 가짐으로서 포아송 근사 방법의 큰 의미를 발견하기가 어렵다. 이는 대표본이 아닌 소표본의 경우 근사이론을 적용하기 어려운 이론적 배경에 기인하는 것으로 판단된다.

[표 1]. 이항분포에서 $\hat{p}$의 변화에 따른 신뢰구간(n=5)

표본비율 $\hat{p}$		$\hat{p}$=0.3		$\hat{p}$=0.4		$\hat{p}$=0.5		$\hat{p}$=0.6		$\hat{p}$=0.7	
방법\신뢰구간	식번호	하한	상한	하한	상한	하한	상한	하한	상한	하한	상한
정규근사(WA)	(2.2)	.0000	.7016	.0000	.8294	.0617	.9382	.1705	1.000	.2983	1.000
정확방법(EX)	(2.1)	.0225	.7905	.0527	.8533	.0943	.9056	.1466	.9472	.2094	.9774
스코어(SC)	(2.3)	.0725	.7012	.1176	.7692	.1704	.8295	.2307	.8823	.2987	.9274
스코어수정(SC2)	(2.6)	.0362	.7375	.0725	.8143	.1176	.8823	.1704	.9426	.2307	.9954
정규근사수정1(WA2)	(2.5)	.0000	.8016	.0000	.9294	.0000	1.000	.0705	1.000	.1983	1.000
정규근사수정2(BS)	(2.7)	.0000	.9690	.0000	1.000	.0000	1.000	.0000	1.000	.0309	1.000
정규근사수정3(BS2)	(2.8)	.0000	.8016	.0000	.9294	.0000	1.000	.0705	1.000	.1983	1.000
chen(CH)	(2.10)	.0000	.8137	.0079	.8788	.0617	.9382	.1211	.9920	.1862	1.000
아크사인(AS)	(2.4)	.0198	.7242	.0595	.8125	.1157	.8842	.1874	.9404	.2757	.9801
포아송근사(PO)	(2.9)	.0215	1.000	.0484	1.000	.0831	1.000	.1237	1.000	.1689	1.000

[표 2]. 표본의 크기 변화에 따른 신뢰구간(표본비율 $\hat{p}$=0.8)

표본수 n		n=5		n=10		n=20		n=30		n=50	
방법\신뢰구간	식번호	하한	상한	하한	상한	하한	상한	하한	상한	하한	상한
정규근사(WA)	(2.2)	.4493	1.000	.5520	1.000	.6246	.9735	.6586	.9431	.6891	.9108
정확방법(EX)	(2.1)	.2835	.9949	.4439	.9747	.5633	.9426	.6143	.9228	.6628	.8996
스코어(SC)	(2.3)	.3755	.9637	.4901	.9433	.5839	.9193	.6269	.9049	.6696	.8875
스코어수정(SC2)	(2.6)	.2987	1.000	.4421	.9912	.5573	.9460	.6086	.9231	.6585	.8986
정규근사수정1(WA2)	(2.5)	.3493	1.000	.5020	1.000	.5996	1.000	.6401	.9598	.6791	.9208
정규근사수정2(BS)	(2.7)	.2033	1.000	.4638	1.000	.5877	1.000	.6339	.9660	.6763	.9236
정규근사수정3(BS2)	(2.8)	.3493	1.000	.5020	1.000	.5741	1.000	.6278	.9721	.6738	.9261
chen(CH)	(2.10)	.2573	1.000	.4374	.9960	.5623	.9410	.6144	.9174	.6635	.8936
아크사인(AS)	(2.4)	.3845	.9993	.5118	.9765	.6019	.9413	.6408	.9210	.6790	.8980
포아송근사(PO)	(2.9)	.2179	1.000	.3453	1.000	.4572	1.000	.5125	1.000	.5714	1.000

포아송 근사에 의한 신뢰구간은 정확신뢰구간에 비해서 매우 크게 과대 추정되고 있다. 포아송 근사방법이 과대추정 되는 경향은 어느 정도 예상된 일인데 이에 대해 Leemis와 Trivedi(1996)는 $n \geq 20, p \leq 0.05$ 이거나, $n \geq 100, np \leq 10$ 일 때 포아송 근사를 활용하는 것이 유리하다고 제시 한

바 있다.

[표 2]는 표본비율이 0.8일 때 다섯 가지 표본의 크기에 대한 각 방법의 신뢰구간의 하한과 상한이 나타나 있다. 포아송 근사방법의 경우는 신뢰구간의 상한이 1로 나타나 표본비율이 0.8 정도로 큰 경우에 좋은 방법은 아닌 것처럼 보인다.

포아송분포의 모수 μ의 11가지 신뢰구간 추정량들을 계산한 결과가 [표 3]에 나타나 있다. 아주 작은 크기인 $n=5$일 때 평균의 변화에 따른 각 방법의 신뢰구간이 [표 3]에 나타나 있다. 이 경우에는 주어진 표본의 크기만큼 한 번 생성된 랜덤표본으로부터 계산된 신뢰구간들로서 각 방법에 따라 약간씩의 차이를 보이고 있다. [표 3]을 자세히 살펴보면 정규근사의 방법들은 정확방법보다 항상 하한의 값이 작게 나타나고 있음을 알수 있다.

[표 3]. 포아송분포의 모수의 변화에 따른 신뢰구간($n=5$)

μ		μ=1		μ=2		μ=3		μ=4		μ=5	
방법\신뢰구간	식번호	하한	상한	하한	상한	하한	상한	하한	상한	하한	상한
정규근사1(WA)	(3.3)	0.1234	1.8765	1.0420	3.7579	2.4036	5.9963	4.5143	9.0856	3.6890	7.9109
정규근사2(W2)	(3.4)	0.2395	2.2395	0.4796	4.3203	2.4036	5.9963	4.8012	8.7987	2.5556	9.0443
정확방법(EX)	(3.1)	0.3246	2.3336	1.2401	4.1923	2.5998	6.4201	4.7084	9.5026	3.8843	8.3297
스코어방법(SC)	(3.5)	0.4271	2.3411	1.3729	4.1953	2.7471	6.4210	4.8663	9.5018	4.0385	8.3297
스코어-수정(S2)	(3.6)	0.3681	2.4804	1.3004	4.3223	2.6682	6.5418	4.7830	9.6184	3.9565	8.4475
Molenaar(MO)	(3.7)	0.3284	2.3354	1.2414	4.1932	2.6006	6.4206	4.7096	9.5026	3.8848	8.3301
카이제곱근사(CH)	(3.2)	0.3222	2.3336	1.2387	4.1925	2.5988	6.4204	4.7084	9.5026	3.8835	8.3300
Battlet(BA)	(3.8)	0.3155	2.0685	1.2341	3.9499	2.5957	6.1884	4.7063	9.2777	3.8811	8.1030
Anscombe(AN)	(3.9)	0.2832	2.1008	1.2131	3.9710	2.5797	6.2043	4.6938	9.2903	3.8675	8.1166
Freedman(FT)	(3.10)	0.3214	2.1581	1.2554	4.0266	2.6240	6.2589	4.7393	9.3440	3.9126	8.1706
Hald(HA)	(3.11)	0.3605	2.0236	1.2627	3.9213	2.6172	6.1668	4.7232	9.2608	3.8994	8.0847

[표 4]. 포아송분포의 랜덤표본

대상표본	μ	n	랜덤표본
[표본 1]	4.17	5	1 6 4 7 6
[표본 2]	8.34	10	7 6 6 11 15 5 7 4 6 6
[표본 3]	16.68	20	27 23 17 12 26 23 21 17 18 16 14 13 14 6 17 14 15 18 14 20
[표본 4]	25.02	30	43 30 28 22 27 25 23 21 22 37 23 25 20 24 27 21 30 27 16 27 27 25 24 27 22 26 35 30 27 26
[표본 5]	41.69	50	42 39 38 36 37 57 41 51 39 43 31 46 52 42 43 39 43 45 43 40 39 41 36 37 39 46 43 42 38 36 40 44 28 52 52 44 47 52 25 43 42 49 40 38 32 41 45 40 48 52

표본의 크기에 따른 신뢰구간의 변화를 보기 위하여 표본의 크기를 5부터 50 까지 변화시키면서 신뢰구간을 계산한 결과는 [표 5]에 나타내었다. 표본의 크기가 큰 경우에는 11가지 방법의 신뢰구간이 아주 비슷하게 나타나나 표본의 크기가 작을 때에는 하한과 상한의 차이가 상당해 보인다. 이들 신뢰구간들은 해당되는 평균을 가지는 포아송분포로부터 생성된 랜덤표본에 기초한 것으로 랜덤표본들은 [표 4]에 나타나 있다.

[표 5]. 11가지 방법의 포아송 모수의 신뢰구간 추정결과

표본수 n		n=5		n=10		n=20		n=30		n=50	
대상표본		표본1		표본2		표본3		표본4		표본5	
방법\신뢰구간	식번호	하한	상한	하한	상한	하한	상한	하한	상한	하한	상한
정규근사1(WA)	(3.3)	2.8796	6.7203	5.6254	8.9745	16.4726	20.2273	24.4005	28.0661	40.1645	43.7554
정규근사2(W2)	(3.4)	2.7073	6.8926	5.2747	9.3252	16.3127	20.3872	24.3405	28.1261	40.2087	43.7112
정확방법(EX)	(3.1)	3.0754	7.1420	5.7218	9.1787	16.5203	20.3268	24.4322	28.1320	40.4835	43.7947
스코어방법(SC)	(3.5)	3.2257	7.1425	5.8065	9.1776	16.5662	20.3258	24.4634	28.1312	40.2025	43.7943
스코어-수정(S2)	(3.6)	3.1454	7.2607	5.7622	9.2333	16.5424	20.3521	24.4473	28.1485	40.1927	43.8045
Molenaar(MO)	(3.7)	3.0761	7.1425	5.7221	9.1787	16.5204	20.3268	24.4323	28.1319	40.1835	43.7947
카이제곱근사(CH)	(3.2)	3.0745	7.1423	5.7218	9.1787	16.5203	20.3268	24.4322	28.1320	40.1835	43.7947
Battlet(BA)	(3.8)	3.0717	6.9124	5.7214	9.0706	16.5206	20.2753	24.4325	28.0981	40.1837	43.7746
Anscombe(AN)	(3.9)	3.0567	6.9237	5.7171	9.0749	16.5196	20.2763	24.4321	28.0985	40.1835	43.7748
Freedman(FT)	(3.10)	3.1013	6.9817	5.7406	9.1012	16.5318	20.2891	24.4403	28.1070	40.1885	43.7799
Hald(HA)	(3.11)	3.0918	6.8923	5.7271	9.0648	16.5219	20.2741	24.4331	28.0975	40.1839	43.7744

5. 결론과 토의

이항분포에서 Vollset(1993), Agresti와 Coull(1998)등에 의해 Wald 방법이 지니는 문제점이 많이 지적되어 왔으며 그 대안으로 Wilson(1927)에 의한 Score 방법들이 주로 추천되어 왔다. 그러나 포아송분포에 있어서 Agresti와 Coull(1998)의 이항분포와 비슷한 결과를 유도하리라는 언급을 제외하고 구체적인 포아송 신뢰구간에 대한 연구가 많지 않은 실정이다. 본 논문에서는 이항분포와 포아송분포의 여러 가지 신뢰구간을 살펴보았고 사용되는 신뢰구간 추정량에 따라 차이가 발생함을 알 수 있었다. 이산형 자료에서 표본의 크기가 작은 경우 정확한 추론은 매우 중요한 역할을 한다는 점을 간과할 수 없다.

참고문헌

[1] Agresti, A. and Coull, B.(1998) Approximate is Better than "Exact" for Interval Estimation of Binomial Proportions. *The American Statistician*, Vol. 52, 119-126.

[2] Blyth, C.R.(1986) Approximate binomial confidence limits, Journal of the American Statistical Association, Vol. 81, 843-855;Corrigenda(1989), Vol. 84, 636.

[3] Blyth, C.R. and Still, H.A.(1983) Binomial confidence intervals, *Journal of the American Statistical Association*, Vol. 78, 108-116.

[4] Chen, H.(1990) the accuracy of approximate intervals for a binomial parameter, *Journal of the American Statistical Association*, Vol. 85, 514-518.

[5] Clopper, C.J. and Pearson, E.S.(1934) The use of confidence or fiducial limits illustrated in the case of the binomial, *Biometrika*, Vol. 26, 404-413.

[6] Hald, A.(1952) *Statistical Theory with Engineering Applications,* John Wiley, New York.

[7] Leemis, L.M. and Trivedi, K.S.(1996) A comparison of approximate interval estimators for the bernoulli parameter, *The American Statistician,* Vol. 50, 63-68.

[8] Sahai, H. and Khurshid, A.(1993) Confidence Intervals for the Mean of a Poisson Distribution: A Review. *Biometrical Journal* vol 7. 857-67.

[9] Vollset, S.E.(1993) Confidence intervals for a binomial proportion, *Statistics in Medicine,* Vol. 12, 809-824.

[10] Wilson, E.B.(1927) Probable inference, the law of succession, and statistical inference, *Journal of the American Statistical Association,* Vol. 22, 209-212.