Journal of the Korean Data & Information Science Society
2017, **28**(3), 493–504

http://dx.doi.org/10.7465/jkdi.2017.28.3.493
한국데이터정보과학회지

영 변환 모형 산포형태모수와 두 적합도 검정통계량 사이의 유사성 비교[†]

윤유정[1] · 김홍기[2]

[1]아태인구연구원 · [2]충남대학교 정보통계학과

접수 2017년 4월 8일, 수정 2017년 5월 15일, 게재확정 2017년 5월 18일

요 약

통계청 인구총조사의 출생아 수 자료는 우리가 쉽게 접할 수 있는 가산 자료이며 국가경쟁력 제고를 위한 정부의 출산정책 결정 및 그 기대효과 분석의 기반이 되는 자료이다. 출생아 수 자료 분석에 있어서 포아송 모형 등 가산 모형이 우월하다는 선행 연구결과에 의하여 가산 모형을 통한 자료 분석방법이 활용되고 있다. 이 때 가산 모형에서 가장 많이 사용하는 포아송 모형은 균등산포라는 제한적인 가정을 토대로 하기 때문에 출생아 수 자료 분석에 이 포아송 모형을 그대로 적용한다면 정보의 손실과 편향추정을 피할 수 없게 된다. 이러한 한계를 극복하기 위해 Ghosh 와 Kim (2007)은 영 과잉과 부족으로 인한 과대산포와 과소산포를 동시에 설명할 수 있는 영 변환 모형 (zero-altered model)을 제안하였다. 본 논문에서는 Ghosh 와 Kim (2007)의 영 변환 모형을 적용하여 실제 출생아수분포에서 영 변환 모형의 산포형태모수 δ를 도출하고 그 역할에 대하여 분석한다. 그리고 관측분포에서의 산포형태모수 δ와 이론적분포와의 차이를 비교하기 위한 적합도 검정통계량과의 유사성을 확인한다.

주요용어: 과대산포, 과소산포, 영 과잉, 영 부족, 영 변환 모형, 콜모고로프 검정.

1. 서론

가산 자료 (count data)는 종속변수인 관찰 값이 비음정수 (non-negative integer value)인 계수형 자료이며 평균과 분산이 같다는 균등분포 (uniform distribution)의 성질을 가지고 있는 포아송분포 (Poisson distribution)등이 가산 자료 분석에 널리 활용되고 있다.

분석에 앞서 모집단의 분포가 특정 분포를 따른다는 모수 모형 (parametric model)의 가정 하에서 관측분포에 적합한 이론적 모형을 적용하게 되는데 이론적 분포의 분산보다 관측분포의 분산이 큰 과대산포 (over-dispersion), 관측분포의 분산이 작은 과소산포 (under-dispersion)가 발생하기도 한다. 이 때 정보가 될 수 있는 영 (zero)의 과잉 및 부족이 과대산포와 과소산포의 다양한 원인 중 하나가 된다. 이 경우 균등분포를 가정한 이론적 분포를 그대로 적용한다면 정보의 손실과 편향 추정 (biased estimation)을 피할 수 없게 된다.

[†] 이 연구는 2015년도 충남대학교 학술 연구비에 의해 지원되었음. 이 논문은 윤유정의 석사 논문의 발췌 논문이다.

[1] (35209) 대전광역시 서구 청사로 148, 아태인구연구원 (APPI), 연구원.
[2] 교신저자: (34134) 대전광역시 유성구 대학로 99, 충남대학교 정보통계학과, 교수. E-mail: hong-giekim@cnu.ac.kr

영의 과잉으로 인한 과대산포는 실제 가산 자료에서 빈번히 관측되고 있기에 과대산포 문제를 해결하기 위한 이론적 모형의 대안으로 수정된 모형의 필요성이 부각되면서 Lambert (1992), Van den Broeck (1995), Hall (2000), Ridout 등 (1998; 2001) 등에 의해 영 과잉 모형이 연구되었다. 한편, Xiang 등 (2006), Xie 등 (2001)과 Zhao (2006) 등은 이러한 분포들을 이용한 포아송 회귀분석에 대한 연구를 진행하였다.

나아가 Heilbron (1994), Gupta 등 (1996), Castillo 와 Perez-Casany (2005), Ghosh 와 Kim (2007) 등의 연구에 의해 영 과잉 과대산포 모형에만 국한되지 않고 영 부족 과소산포 모형까지 설명할 수 있는 모형이 제시되었다. 특히 Ghosh 와 Kim (2007)이 제시한 영 변환 모형은 Castillo and Perez-Casany (2005) 등의 모수적 (parametric) 일반화 모형의 한계를 넘어 모수적 요소와 비모수적 요소를 포함하는 준모수적 (semi-parametric) 모형을 제시함으로써 가산자료에서의 보다 다양한 분석이 가능하도록 하였다.

본 논문에서는 영 과잉 모형과 영 부족 모형을 모두 설명할 수 있는 Ghosh 와 Kim (2007) 영 변환 모형을 적용하여 실제 관측분포에서 영 변환 모형의 산포형태모수 δ의 역할을 파악하였다. 그리고 관측분포에서의 산포형태모수 δ와 적합도검정의 검정통계량과의 연관성을 확인하였다. 한편, 적합도 검정 통계량에 관한 최근 연구로는 Kang 등 (2014)의 연구가 있다.

2. 영 변환 모형

Ghosh 와 Kim (2007)의 영 변환 모형은 포아송 분포와 같이 균등산포를 가정한 모형에서 영의 과잉으로 인한 과대산포와 영의 부족으로 인한 과소산포에 적용하기 위한 모형이다. 포아송회귀 모형을 이용한 최근 연구로 설계사들의 이직 요인을 분석한 연구가 있다 (Chun, 2016).

$N = \{0, 1, 2, \cdots\}$ 값을 가지는 이산확률변수 U의 확률질량함수를 $f_0(u)$라고 할 때, 모든 u에서 $f_0(u) < 1$로 가정한다. 여기서 산포형태모수 δ를 사용하여 $u = 0$에서의 확률을 조정함으로써 다음과 같은 새로운 확률변수 X의 확률질량함수 $f_\delta(x)$를 얻을 수 있다. 산포형태모수 δ의 범위는 $\delta \in (-1, 1)$이다.

$$f_\delta(x;\delta) = P(X = x) = \begin{cases} \delta_+^2 + (1-\delta^2) f_0(0), & x = 0, \\ \left\{1 - \delta_+^2 + \delta_-^2 \left(\frac{f_0(0)}{1-f_0(0)}\right)\right\} f_0(x), & x = 1, 2, \cdots. \end{cases} \tag{2.1}$$

위 식에서 $\delta_+ = \max(\delta, 0)$이고 $\delta_- = \max(-\delta, 0)$이며 $\delta = 0$에서 미분이 가능하도록 δ를 δ^2으로 표현했을 때 $\delta^2 = \delta_+^2 + \delta_-^2$ 관계가 성립한다. 확률질량함수 $f_\delta(X)$는 δ^*와 f_0^*가 존재하여 $f_\delta(x) = f_{\delta^*}(x)$, $f_0(0) = f_0^*(0)$가 되고 이는 $\delta = \delta^*$, $f_0(x) = f_0^*(x)$이므로 유일한 표현식으로 과대산포와 과소산포를 모두 설명할 수 있다 (Ghosh 와 Kim, 2007).

확률변수 U의 평균을 μ_0, 분산을 σ_0^2이라 할 때, 새로운 확률변수인 X의 평균과 분산을 μ, σ^2라고 하면 다음의 식이 도출된다.

$$\mu = h(\delta)\mu_0, \quad \sigma^2 = h(\delta)\sigma_0^2 + \frac{1-h(\delta)}{h(\delta)}\mu^2, \tag{2.2}$$

여기서 $h(\delta) = \left(1 - \delta_+^2 + \delta_-^2 \frac{f_0(0)}{1-f_0(0)}\right)$.

분산 식을 정리하면 다음과 같이 $\sigma^2 - \mu$는 산포형태모수 δ의 부호와 같음을 설명할 수 있다 (Ghosh 와 Kim, 2007).

$$\frac{1-h(\delta)}{h(\delta)} = \begin{cases} \frac{\delta^2}{1-\delta^2}, & \text{여기서 } \delta > 0, \\ \frac{-\delta^2 f_0(0)}{1-(1-\delta^2)f_0(0)}, & \text{여기서 } \delta \le 0. \end{cases} \tag{2.3}$$

이를 통해 산포의 유형과 산포형태모수 δ의 관계를 살펴보면 다음과 같다.

$$\delta < 0\text{인 경우, } \mu > \sigma^2\text{으로 과소산포,}$$
$$\delta = 0\text{인 경우, } \mu = \sigma^2\text{으로 균등산포,}$$
$$\delta > 0\text{인 경우, } \mu < \sigma^2\text{으로 과대산포.}$$

산포형태모수 δ는 위의 식의 평균과 분산 그리고 $g(0)$의 함수로서 다음과 같이 주어진다. 여기서 $g(0)$은 $x = 0$에서의 확률을 의미한다 (Ghosh 와 Kim, 2007).

$$\delta = \begin{cases} \sqrt{\frac{\sigma^2-\mu}{\sigma^2-\mu+\mu^2}}, & \sigma^2 \geq \mu \text{ 일 때,} \\ -\sqrt{\frac{\mu-\sigma^2}{\mu-\sigma^2+\frac{g(0)}{1-g(0)}\mu^2}}, & \sigma^2 < \mu \text{ 일 때.} \end{cases} \tag{2.4}$$

산포형태모수 δ값의 변화에 따른 확률질량함수를 통해 다음과 같이 영 과잉 모형과 영 부족 모형을 상세하게 설명할 수 있다.

① 이론적 모형
◎ $\delta = 0$인 경우 $\delta_+ = \delta_- = 0$ 이므로

$$f_\delta(x) = \begin{cases} f_0(0), & x = 0, \\ f_0(x), & x = 1, 2, \cdots . \end{cases}$$

영 과잉이나 영 부족을 나타내지 않는 이론적 분포의 확률질량함수와 같다.

② 영 과잉 모형
◎ $0 < \delta < 1$인 경우 $\delta_+ = \delta, \delta_- = 0$이므로

$$f_\delta(x) = \begin{cases} f_0(0) + \delta^2(1 - f_0(0)), & x = 0, \\ (1-\delta)^2 f_0(x), & x = 1, 2, \cdots . \end{cases}$$

영의 값에서는 발생확률이 증가하고 영 이외의 값에서는 발생확률이 감소한다.
◎ $\delta = 1$인 경우 $\delta_+ = 1, \delta_- = 0$이므로

$$f_\delta(x) = \begin{cases} 1, & x = 0, \\ 0, & x = 1, 2, \cdots . \end{cases}$$

영의 값에서 발생확률이 극대화되고 영 이외의 값에서는 발생확률이 없다.

③ 영 부족 모형
◎ $-1 < \delta < 0$인 경우 $\delta_+ = 0, \delta_- = -\delta$ 이므로

$$f_\delta(x) = \begin{cases} (1-\delta^2) f_0(0), & x = 0, \\ \left(1 + \delta^2 \frac{f_0(0)}{1-f_0(0)}\right), & x = 1, 2, \cdots . \end{cases}$$

영의 값에서 발생확률이 감소하고 영 이외의 값에서는 발생확률이 증가한다.

◎$\delta = -1$인 경우 $\delta_+ = 0, \delta_- = 1$ 이므로

$$f_\delta(x) = \begin{cases} 0, & x = 0, \\ \frac{1}{1-f_0(0)} f_0(x), & x = 1, 2, \cdots . \end{cases}$$

영의 값에서 발생확률이 없고 영 이외의 값에서 발생확률이 극대화된다. 위에서 보듯이 산포형태모수 δ의 범위는 $-1 \le \delta \le 1$이다.

표본 데이터에서 δ의 추정치 $\hat{\delta}$를 구하기 위해서는 μ와 σ^2 그리고 $g(0)$ 대신 표본평균 $\bar{X}$, 표본분산 S^2, 표본에서 영이 차지하는 비율 P_0가 필요하게 된다. 다음 식에서 $I_0(x)$는 x가 0일 때, 함수 값이 1인 지시함수 (indicator function)이다.

$$\overline{X} = \sum_{i=1}^{n} X_i/n, \quad S^2 = \sum_{i=1}^{n} (X - \overline{X})^2/(n-1), \quad P_0 = \sum_{i=1}^{n} I_0(X_i)/n. \tag{2.5}$$

이를 통해 산포형태모수 δ의 추정치 $\hat{\delta}$를 구하면 다음과 같다. (Ghosh 와 Kim, 2007)

$$\hat{\delta} = \begin{cases} \sqrt{\frac{\frac{n-1}{n} S^2 - \overline{X}}{\frac{n-1}{n} S^2 - \overline{X} + (\overline{X})^2}}, & \frac{n-1}{n} S^2 \ge \overline{X} \text{ 일 때}, \\ -\sqrt{\frac{\overline{X} - \frac{n-1}{n} S^2}{\overline{X} - \frac{n-1}{n} S^2 + \frac{P_0}{1-P_0} (\overline{X})^2}}, & \frac{n-1}{n} S^2 < \overline{X} \text{ 일 때}. \end{cases}$$

3. 대한민국 기혼여성 출생아수 산포형태모수 추정치 $\hat{\delta}$

본 논문에 활용된 대한민국 기혼여성 출생아 수 자료는 1980년, 1990년, 2000년, 2010년 데이터로서 인구총조사가 시행되는 각 연도에서 5% ∼ 10%내에서 조사된 표본자료이다. 이 출생아수 분포에 대한 산포형태모수의 추정치 $\hat{\delta}$는 다음의 Table 3.1과 같으며, 이 산포형태모수 δ를 추정하기 위한 평균, 분산, 영 확률에 대한 분석은 Ra (2012), Kim (2013)의 논문에서도 상세히 소개되어있다.

Table 3.1 Estimates of dispersion type parameter δ $\left(\hat{\delta}\right)$ of number of births in Korean married women

Year \ Age	15-19	20-24	25-29	30-34	35-39	40-44	45-49
1980	-0.486	-0.590	-0.813	-0.928	-0.928	-0.919	-0.875
1990	-0.407	-0.605	-0.823	-0.947	-0.961	-0.962	-0.952
2000	-0.296	-0.569	-0.732	-0.919	-0.957	-0.962	-0.961
2010	-0.598	-0.722	-0.734	-0.874	-0.950	-0.969	-0.973
Year \ Age	50-54	55-59	60-64	65-69	70-74	75+	Total
1980	-0.632	0.123	0.177	0.198	0.200	0.200	0.317
1990	-0.935	-0.888	-0.581	0.123	0.159	0.137	0.304
2000	-0.954	-0.929	-0.922	-0.852	-0.567	0.134	0.080
2010	-0.973	-0.968	-0.964	-0.946	-0.919	-0.821	-0.792

본 데이터에서 얻어진 산포형태모수 $\hat{\delta}$은 2000년까지 양수이며 2010년에는 음수로 나타났고 전체 연령에서 1980년의 0.317이 최대치, 2010년의 -0.792가 최소치이다. 연령별로 산포형태모수 $\hat{\delta}$을 살펴보면 15세 이상부터 50세 미만의 연령에서는 모두 음수로 영 부족 모형임을 확인할 수 있다. 반면 50세 이상의 연령부터는 연도별로 산포의 유형이 차이가 있는 것으로 나타났는데 1980년에는 50세 이상의 연령에서 모두 양수로 영 과잉 상태이나 1990년 이후부터는 영 과잉 현상이 감소하고 영 부족 상황이 증가하는 것을 Figure 3.1에서 확인할 수 있다.

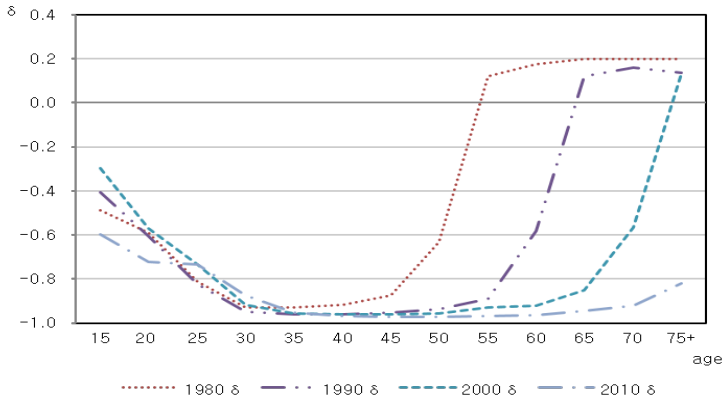

Figure 3.1 Dispersion type parameter of number of births in years

4. 산포형태모수 추정치 과 적합도검정 검정통계량과의 연관성 분석

4.1. χ^2 적합도 검정통계량

카이제곱 적합도 검정은 가설로 주어진 이론적 확률분포에서의 기댓값인 기대도수와 실제 관측도수와의 차이를 비교함으로써 관측분포의 모집단이 특정한 이론적 확률분포를 따르는지에 대한 검정 방법이다. 확률표본의 관측값이 범주 j $(j = 1, \cdots, c)$ 로 분류되어 있을 때 카이제곱 검정통계량은 다음과 같다.

$$O_j = \text{각 범주의 분류된 관측도수},$$
$$E_j = H_0\text{하에서 범주 에 대한 기대도수},$$
$$\chi^2 = \sum_{j=1}^{c} \frac{(O_j - E_j)^2}{E_j}.$$

출생아수의 관측 자료의 모집단은 포아송 분포를 따른다는 귀무가설 H_0 하에서 검정통계량 χ^2은 근사적으로 자유도가 $c-1$인 카이제곱분포를 따른다. 다음의 Table 4.1은 포아송 분포를 검정하는 카이제곱 통계량을 나타낸 표이다.

Table 4.1 Chi-square statistics (χ^2) of number of births in Korean married women

Year \ Age	15-19	20-24	25-29	30-34	35-39	40-44	45-49
1980	1014.7	27035.2	198936.0	295893.7	151759.1	89902.4	54026.1
1990	143.1	25156.6	389277.7	1262125.5	708536.0	329065.4	155181.9
2000	54.1	9212.4	171617.8	1168936.0	1753670.1	1445740.4	677407.5
2010	408.4	10430.1	68879.5	368334.8	1097083.7	1506601.6	1637891.7
Year \ Age	50-54	55-59	60-64	65-69	70-74	75+	Total
1980	43280.5	66414.2	60096.7	45048.3	28508.1	24929.9	893617.6
1990	82339.2	47442.6	38831.8	42304.4	33886.5	26104.4	1254988.4
2000	297848.8	124447.5	80397.6	53183.1	43451.6	55210.0	2271349.8
2010	1336280.6	614131.9	281449.4	129038.9	70649.9	49697.0	3403438.5

카이제곱 검정통계량의 크기를 통해 관측분포와 모집단의 분포와의 동일성에 대한 차이의 정도를 확

인 할 수 있으며 이는 영 변환 모형 산포형태모수 δ의 크기와 비교해 볼 수 있다. 다음의 Figure 4.1은 출생아수 분포의 카이제곱 검정통계량과 산포형태모수 $\hat{\delta}$을 비교한 그래프이다.

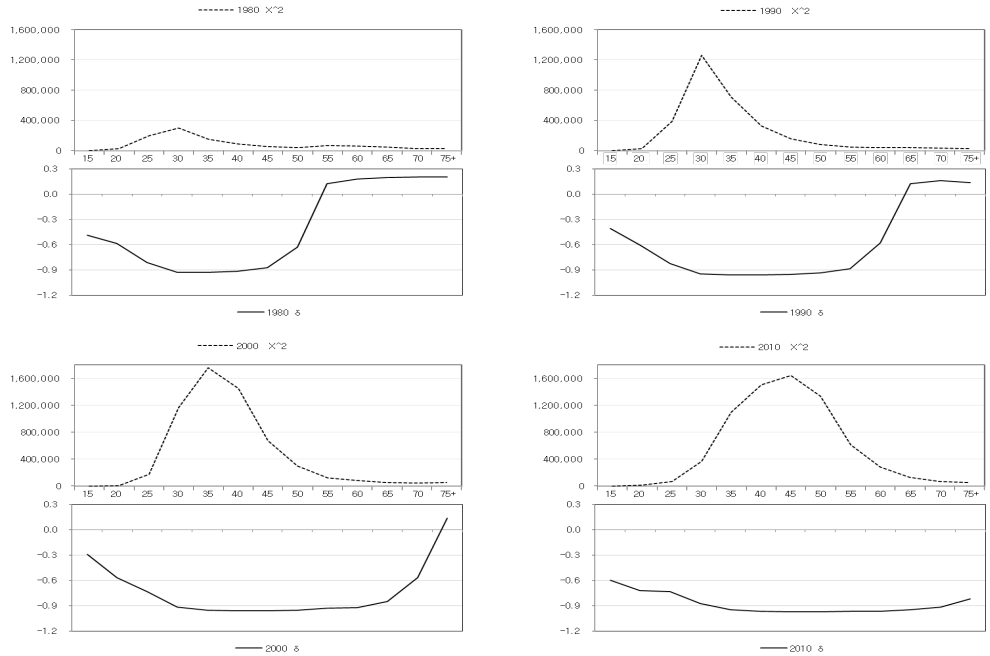

Figure 4.1 Chi-square statistics (above) and dispersion type parameter δ (below) of number of births in years (from top-left, years 1980,1990, 2000, and 2010)

앞서 영 변환 모형의 설명에서 균등산포일 때 $\delta = 0$이며 영 과잉 과대산포 모형일수록 $\delta = 1$에 가까워지고 영 부족 과소산포 모형일수록 $\delta = -1$에 가까워진다고 설명하였다. Figure 4.1을 보면 카이제곱 검정통계량의 크기와 산포형태모수 $\hat{\delta}$의 절대값의 크기가 관련성이 있음을 확인할 수 있다. 즉 산포형태모수 $\hat{\delta}$에서 균등산포를 벗어나는 정도가 커짐에 따라 카이제곱 검정통계량도 영향을 받는다는 것이다. 하지만 카이제곱 적합도 검정통계량만으로는 산포형태모수 $\hat{\delta}$에 따른 영 과잉 또는 영 부족 모형의 구분을 설명할 수는 없다는 한계가 있다. 이는 카이제곱 통계량이 방향성이 없다는 사실 (단측검정 불가능)에 기인한다.

4.2. 콜모고로프 (Kolmogorov) 적합도 검정통계량 D

콜모고로프 (Kolmogorov) 적합도 검정은 관측분포의 누적확률분포함수 $F_n(x)$와 이론적분포의 누적확률분포함수 $F_0(x)$를 비교하여 서로 같은 분포를 따르는지에 대한 적합도 검정 방법이다.

크기가 n인 확률표본으로부터 얻어지는 콜모고로프 검정통계량 D, D^+, D^-은 다음과 같다.

$$F_n(x) = \text{확률표본 } X_1, \cdots, X_n \text{이 갖는 누적확률표본분포함수},$$
$$F_0(x) = \text{특정 확률분포의 누적확률분포함수},$$
$$(\text{특정 확률분포를 포아송 분포로 가정하면} F_0(x) = \sum_{y=0}^{[x]} \frac{e^{-\lambda}\lambda^y}{y!}),$$
$$D^+ = \sup_x \{F_0(x) - F_n(x)\},$$
$$D^- = \sup_x \{F_0(x) - F_n(x)\},$$
$$D = \sup_x \{|F_0(x) - F_n(x)|\} = \max\{D^+, D^-\}.$$

출생아수의 관측 분포의 모집단은 포아송 분포를 따른다는 귀무가설 H_0하에서 콜모고로프 적합도 검정통계량 D를 계산한 결과는 다음의 Table 4.2와 같다.

Table 4.2 Kolmogorov goodness-of-fit statistic D of number of births in Korean married women

Year \ Age	15-19	20-24	25-29	30-34	35-39	40-44	45-49
1980	0.036	0.057	0.093	0.146	0.099	0.078	0.063
1990	0.019	0.067	0.116	0.196	0.202	0.149	0.096
2000	0.012	0.057	0.084	0.179	0.238	0.227	0.202
2010	0.066	0.112	0.107	0.130	0.170	0.220	0.231
Year \ Age	50-54	55-59	60-64	65-69	70-74	75+	Total
1980	0.050	0.048	0.045	0.052	0.052	0.053	0.053
1990	0.076	0.061	0.051	0.043	0.042	0.037	0.080
2000	0.148	0.086	0.076	0.068	0.058	0.044	0.093
2010	0.218	0.193	0.145	0.088	0.075	0.045	0.100

콜모고로프 검정통계량의 크기를 통해 관측분포와 모집단의 분포와의 동일성에 대한 차이의 정도를 확인할 수 있으며 이는 영 변환 모형 산포형태모수 $\hat{\delta}$의 크기와 비교해 볼 수 있다. 다음의 Figure 4.2는 출생아수 분포의 콜모고로프 검정통계량 D와 산포형태모수 $\hat{\delta}$을 비교한 그래프이다.

Figure 4.2에서 보면 콜모고로프 검정통계량 D는 산포형태모수 $\hat{\delta}$에서 균등산포를 벗어나는 정도가 커짐에 따라 영향을 받으므로 검정통계량 D의 크기가 산포형태모수 $\hat{\delta}$의 절대값의 크기와 연관성이 있음을 알 수 있다. 그러나 콜모고로프 검정통계량 D는 카이제곱 적합도 검정통계량과 마찬가지로 영 과잉 또는 영 부족 모형은 설명하지 못함을 확인할 수 있다. 따라서 이 절에서 계산한 콜모고로프 검정통계량 D는 검정통계량 D^+, D^-의 값 중에서의 최대치를 나타낸 것이므로 검정통계량 D^+와 검정통계량 D^-를 모두 나타내는 그래프에서는 어떠한 결과가 도출되는지 확인할 필요성이 있다.

4.3. 콜모고로프 검정통계량 D^+, D^-와 산포형태모수 $\hat{\delta}$의 연관성 분석

검정통계량 D^+, D^-를 이용하면 관측분포와 모집단 분포와의 동일성 여부에 대한 검정을 넘어 관측분포와 모집단 분포의 차이가 갖는 방향성에 대한 검정이 가능하다. 출생아수 관측분포의 모집단이 포아송 분포를 따른다는 귀무가설 H_0 하에서 콜모고로프 적합도 검정통계량 D^+, D^-를 계산한 결과는 다음의 Table 4.3, Table 4.4와 같다.

콜모고로프 적합도 검정통계량 D는 검정통계량 D^+, D^-의 값 중에서의 최대치를 검정통계량으로 나타내었다. 이 때 콜모고로프 적합도 검정통계량 D^+, D^-는 관측분포와 모집단의 분포와의 차이를 나타

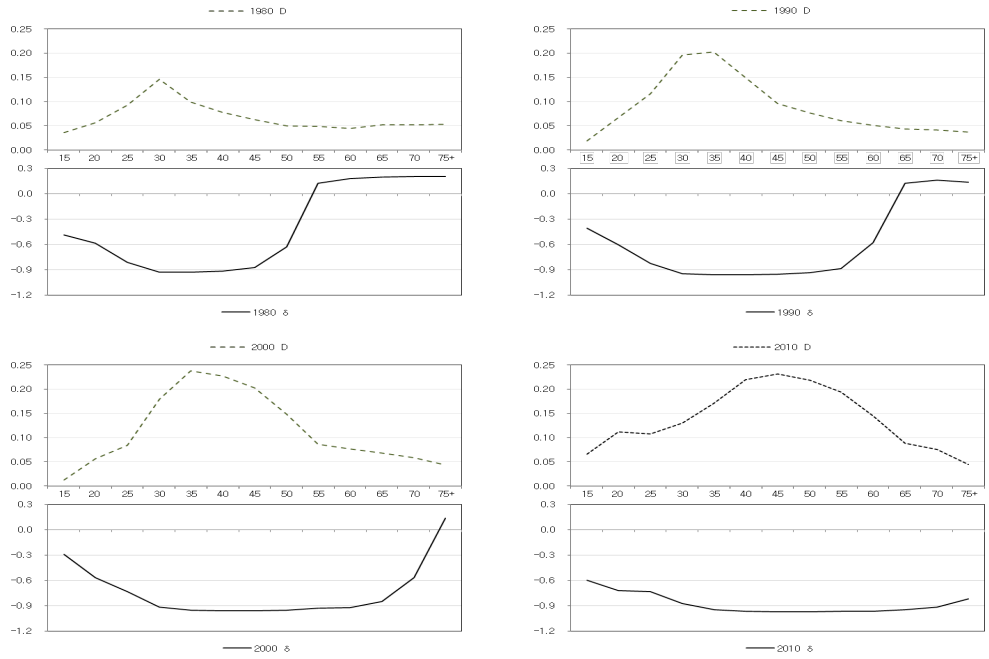

Figure 4.2 Kolmogorov goodness-of-fit statistic D (above) and dispersion type parameter δ (below) of number of births in years (from top-left, years 1980,1990, 2000, and 2010)

Table 4.3 Kolmogorov goodness-of-fit statistic D^+ of number of births in Korean married women

Year \ Age	15-19	20-24	25-29	30-34	35-39	40-44	45-49
1980	0.036	0.057	0.093	0.146	0.099	0.078	0.063
1990	0.019	0.067	0.116	0.196	0.202	0.149	0.096
2000	0.012	0.057	0.084	0.179	0.238	0.227	0.202
2010	0.066	0.112	0.107	0.130	0.170	0.220	0.231
Year \ Age	50-54	55-59	60-64	65-69	70-74	75+	Total
1980	0.050	0.048	0.044	0.040	0.036	0.032	0.043
1990	0.076	0.061	0.051	0.043	0.036	0.025	0.037
2000	0.148	0.086	0.076	0.068	0.058	0.043	0.067
2010	0.218	0.193	0.145	0.088	0.075	0.045	0.095

내므로 영 변환 모형 산포형태모수 $\hat{\delta}$의 크기와 비교할 수 있으며 이는 3장, Table 3.1의 산포형태모수 $\hat{\delta}$의 영 과잉 및 영 부족 모형과의 연관성이 있음을 확인할 수 있다. Figure 4.3은 출생아수 분포의 콜모고로프 적합도 검정통계량 D^+, D^-와 산포형태모수 $\hat{\delta}$을 비교한 그래프이다.

Figure 4.3에서 콜모고로프 적합도 검정통계량 D^+, D^-는 산포형태모수 $\hat{\delta}$의 크기뿐만 아니라 영 과잉 과대산포와 영 부족 과소산포를 구분하는 산포형태모수 δ의 역할도 설명할 수 있음을 알 수 있다.

대체적으로 산포형태모수 $\hat{\delta}$의 부호가 -인 영 부족 모형의 경우 콜모고로프 적합도 검정에서는 검정통계량 D^+가 검정통계량 D로 채택되었고 산포형태모수 $\hat{\delta}$의 부호가 +인 영 과잉 모형의 경우 검정통계량 D^-가 검정통계량 D로 채택되었다. 즉, 영 부족 모형에서 영 과잉 모형으로 전환되는 구간에서는 검정통계량 D^+의 그래프와 검정통계량 D^-가 교차되는 현상을 보이고 있다.

Table 4.4 Kolmogorov goodness-of-fit statistic D^- of number of births in Korean married women

Year \ Age	15-19	20-24	25-29	30-34	35-39	40-44	45-49
1980	0.027	0.032	0.019	0.083	0.062	0.049	0.036
1990	0.015	0.035	0.103	0.149	0.119	0.089	0.064
2000	0.009	0.027	0.076	0.162	0.172	0.148	0.117
2010	0.055	0.057	0.053	0.112	0.155	0.158	0.162
Year \ Age	50-54	55-59	60-64	65-69	70-74	75+	Total
1980	0.025	0.039	0.045	0.052	0.052	0.053	0.053
1990	0.049	0.039	0.027	0.037	0.042	0.037	0.080
2000	0.088	0.062	0.053	0.041	0.032	0.044	0.093
2010	0.144	0.115	0.085	0.064	0.052	0.025	0.100

5. 결론

영 과잉으로 인한 과대산포, 영 부족으로 인한 과소산포가 나타나는 가산자료에서의 분석을 위해 고안된 Ghosh 와 Kim (2007) 영 변환 모형은 영 과잉 모형과 영 부족 모형을 모두 설명할 수 있으며 그 바탕에는 영 변환 모형의 산포형태모수 δ의 역할을 이해하는 것이 필요하다. 본 논문에서는 평균과 분산 그리고 영 확률에 근거한 식으로 구성된 Ghosh 와 Kim (2007) 영 변환 모형의 산포형태모수 δ의 원리를 이해하고 실제 자료에 적용하였다.

대한민국 기혼여성 출생아수 자료에서 추정한 산포형태모수 δ에 대한 분석에서 보면 1980년에는 영 과잉 과대산포가 나타나는 연령대와 영 부족 과소산포가 나타나는 연령대가 혼재되어 있었으나 조사연도가 거듭될수록 영 과잉 과대산포가 나타나는 연령대가 감소하기 시작하여 2010년에는 전 연령대에서 영 부족 현상이 나타나는 결과를 확인하였다.

이와 같이 조사연도와 연령대에 따라 달라지는 산포형태모수 δ의 크기에 따라 관측분포와 이론적 포아송 분포와의 동일성에 대한 차이도 설명할 수 있는지 관심을 가지고 관측분포에서 적합도 검정통계량과 콜모고로프 적합도 검정통계량을 계산하여 산포형태모수 δ와 비교하였다.

카이제곱 적합도 검정통계량과 콜모고로프 적합도 검정통계량 D는 산포형태모수 δ와 비교했을 때는 적합도 검정통계량의 크기와 산포형태모수 δ의 절대값의 크기가 어느 정도 비례하는 사실은 확인할 수 있었지만 과대산포와 과소산포를 구분하는 산포형태모수 δ의 방향성에 대해서는 설명하지 못하였다. 반면 콜모고로프 적합도 검정통계량 D^+, D^-는 적합도 검정통계량과 산포형태모수 의 절대값의 크기뿐만 아니라 산포형태모수 δ의 방향성도 설명할 수 있는 것으로 나타났다.

대한민국 기혼여성 출생아수 자료의 산포형태모수 와 콜모고로프 적합도 검정통계량에 대한 분석에서 보면 1975년에 영 부족 과소산포가 나타나는 연령대에서는 검정통계량 D^-보다 큰 검정통계량 D^+가 검정통계량으로 채택되었으나 영 과잉 과대산포가 나타나는 연령대에서는 대체적으로 검정통계량 D^+보다 큰 검정통계량 D^-가 검정통계량으로 채택된 것을 알 수 있다. 또한 조사연도가 거듭될수록 산포형태모수 δ의 부호가 -인 수치, 즉 영 부족 과소산포 모형으로 수렴하면서 2010년도에는 모든 연령대에서 검정통계량 D^+가 검정통계량으로 채택된 것을 확인할 수 있다.

이와 같이 본 논문을 통해 콜모고로프 적합도 검정통계량 D^+, D^-는 단순한 분포의 동일성에 대한 검정뿐만 아니라 산포의 형태를 구분하는 산포형태모수의 역할도 어느 정도 설명할 수 있음을 알 수 있다.

References

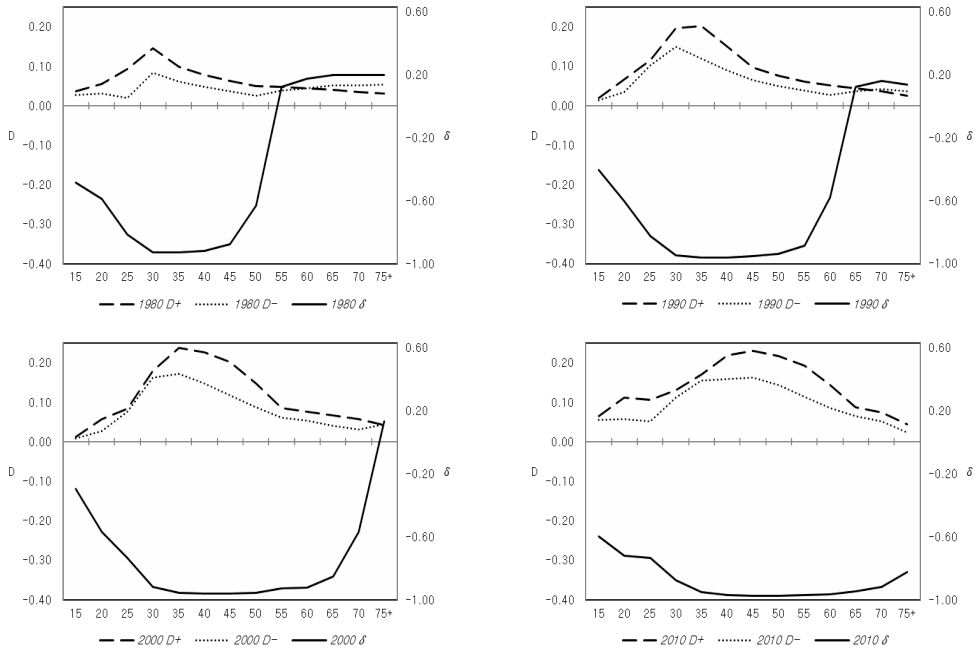

Figure 4.3 Kolmogorov goodness-of-fit statistics D, D^+ and D^- of number of births in Korean married women (from top-left, years 1980,1990, 2000, and 2010)

Castillo, J. D. and Perez-Casany, M. (2005). Overdispersed and underdispersed Poisson generalizations. *Journal of Statistical Planning and Inference*, **134**, 486-500.

Chun, H. (2016). The factors of insurance solicitor's turnovers of life insurance using Poisson regression. *Journal of the Korean Data & Information Science Society*, **27**, 1337-1347.

Ghosh, S. K. and Kim, H. (2007). Semiparametric inference based on a class of zero-altered distributions. *Statistical Methodology*, **4**, 371-378.

Gupta, P. L., Gupta, R. C. and Tripathi, R. C. (1996). Analysis of zero-adjusted count data. *Computational Statistics and Data Analysis*, **23**, 207-218.

Hall, D. B. (2000). Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics* **56**, 1030-1039.

Heilbron, D. C. (1994). Zero-altered and other regression models for count data with added zeroes. *Biometrical Journal*, **36**, 531-547.

Kang, S. Han, J. Seo, Y. and Jeong, J. (2014). Goodness-of-fit tests for the inverse Weibull or extreme value distribution based on multiply type-II censored samples. *Journal of the Korean Data & Information Science Society*, **25**, 903-914.

Kim, K. (2013). *A study on the role of the dispersion parameter in Ghosh and Kim's zero-altered model*, Master's thesis, Graduate school of Chungnam national university, Daejeon, Republic of Korea.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**, 1-14.

Ra, Y. (2012). *Application of zero-altered model to the distribution of number of children of Korean married women*, Master's thesis, Graduate school of Chungnam national university, Daejeon, Republic of Korea.

Ridout, M., Demetrio, C. G. B. and Hinde, J. (1998). Models for count data with many zeros. *International Biometric Conference*.

Ridout, M., Hinde, J. and Demerio, C. G. B. (2001). A score test for testing a zero-infalted poisson regression model against zero-inflated negative binomial alternatives. *Biometrics*, **57**, 219-223.

Van den Broeck, J. (1995). A Score test for zero inflation in poisson distribution. *Biometrics*, **51**, 738-743.

Xiang, L., Lee, A. H., Yau, K. K. W. and McLachlan, G. J. (2006). A score test for over-dispersion in

zero-inflated poisson mixed regression model. *Inter Science*, **26**, 1608-1622.

Xie, M., He, B. and Goh, T. N. (2001). Zero-inflated Poisson model in statistical process control. *Computational Statistics and Data Analysis*, **38**, 191-201.

Zhao, Y. (2006). Score test for generalization and zero-inflation in countdata modeling. *ProQuest*, **131**.

Journal of the Korean Data & Information Science Society
2017, **28**(3), 493–504

http://dx.doi.org/10.7465/jkdi.2017.28.3.493
한국데이터정보과학회지

Similarity between the dispersion parameter in zero-altered model and the two goodness-of-fit statistics[†]

Yujeong Yun[1] · Honggie Kim[2]

[1]Research Division, Asia Pacific Population Institute
[2]Department of Information and Statistics, Chungnam National University

Received 8 April 2017, revised 15 May 2017, accepted 18 May 2017

Abstract

We often observe count data that exhibit over-dispersion, originating from too many zeros, and under-dispersion, originating from too few zeros. To handle this types of problems, the zero-altered distribution model is designed by Ghosh and Kim in 2007. Their model can control both over-dispersion and under-dispersion with a single parameter, which had been impossible ever. The dispersion type depends on the sign of the parameter δ in zero-altered distribution. In this study, we demonstrate the role of the dispersion type parameter δ through the data of the number of births in Korea. Employing both the chi-square statistic and the Kolmogorov statistic for goodness-of-fit, we also explained any difference between the theoretical distribution and the observed one that exhibits either over-dispersion or under-dispersion. Finally this study shows whether the test statistics for goodness-of-fit show any similarity with the role of the dispersion type parameter δ or not.

Keywords: Kolmogorov test, over-dispersion, under-dispersion, zero-altered model, zero deflation, zero inflation.

[†] This research is fully supported by 2015 CNU research fund. This paper is based on part of Yujeong Yun's Master thesis.

[1] Research division, Asia Pacific Population Institute (APPI), 148, Cheongsa-ro, Seo-gu, Daejeon, Republic of Korea.

[2] Corresponding author: Professor, Department of Information and Statistics, CNU, 99, Daehak-ro, Yuseong-gu, Daejeon, Republic of Korea.
E-mail: honggiekim@cnu.ac.kr