응용통계연구 (2009)
22(2), 365–373

남아 출생률 자료에 대한 이질성 분석

임화경[1] · 송석헌[2] · 송주원[3]

[1]고려대학교 통계학과, [2]고려대학교 통계학과, [3]고려대학교 통계학과

(2008년 10월 접수, 2008년 11월 채택)

요 약

1990년대에 접어들면서 우리나라는 남아선호사상, 사회경제적 발전, 인구정책 등으로 인해 태아 성감별이나 불법 낙태가 사회적으로 확산되어 출생 성비의 불균형 현상이 초래되었다. 이러한 출생 성비의 불균형 현상을 모니터링 하기 위해서 시계열 분석 또는 지역 차이 분석 등의 많은 연구들이 있었지만 시간과 공간을 동시에 분석에 포함시키지 못했다는 단점이 있다. 본 연구는 시간과 지역성을 동시에 고려하여 출생 성비의 불균형 현상을 분석하였다. 행정구역상으로 구분된 234개의 구, 시, 군 지역에 대하여 2000년과 2001년도의 각 지역별 셋째 자녀 이상의 남녀 출생아 수 자료를 이질적 모집단을 가정하여 이항분포의 혼합모형에 적용시키는 방법을 제안한다. 혼합모형의 위치모수와 가중치, 상관계수 추정은 EM 알고리즘을 통해 수행되었고, ArcView GIS를 이용하여 지역의 이질성을 그림을 통해 표현하였다.

주요용어: 이항자료, 남아 출생률, 이질성, 혼합분포, EM 알고리즘.

1. 서론

생물학적 측면에 있어서 염색체 결합으로 인한 출생 성비는 같다고 보지만 자연 상태에서의 출생 성비는 여아 100명당 106명 정도로 남아 출생 수가 더 많다고 알려져 있다. 표 1.1은 1980년부터 2005년까지 5년 단위로 우리나라 여아 100명당 남아 출생 수를 전체 비율과 출산 순위별 비율로 나타낸 것이다. 우선 전체 출생아의 성비를 보면 1980년(105.3)부터 점점 높아지다 1990년(116.5)에 가장 높았고 그 이후는 성비가 조금씩 하락하여 자연 상태에서의 출생 성비인 106과 거의 비슷해졌다. 그리고 출산 순위별 성비로 보았을 때는 첫째 자녀는 1990년을 제외한 모든 연도에서, 둘째 자녀는 1990년과 1995년을 제외한 연도에서 대체로 자연 상태의 출생 성비와 유사한 반면 셋째 자녀와 넷째 자녀 이상을 보면 거의 모든 연도에 걸쳐 자연 상태의 출생 성비와 확연히 차이가 나는 것을 볼 수 있다. 그림 1.1은 지역별 출생 성비의 변화를 나타내는데 전국의 출생 성비를 나타내는 선 위에는 수치를 표시하였다. 경북, 경남, 부산, 대구 등 영남지역의 출생 성비는 대체적으로 전국 평균치보다 높으며, 대전, 충북, 제주도의 출생 성비는 비교적 높으나 일관성 있는 경향을 나타내지 않고 있고, 상대적으로 광주, 전남, 전북 등 호남지역과 서울, 경기, 인천 등 수도권의 출생 성비는 전국 평균을 밑돌고 있어 각 지역마다 출생 성비는 차이를 보인다. 이와 같이 출생아의 성비 불균형이 나타나는 이유는 여러 가지가 있을 수 있는데, 남존여비(男尊女卑)의 전근대적 가부장제의 잔재가 아직 남아있는데다가 생명윤리의 부재현상으로 인한 불법 태아 성감별이나 불법 낙태가 사회적으로 확산되어 결과적으로 출생 성비의 불균형 현상이 초래되었다는 보고가 있다 (조남훈과 서문희, 1994).

[3]교신저자: (136-701) 서울시 성북구 안암동 5-1, 고려대학교 통계학과, 부교수. E-mail: jsong@korea.ac.kr

표 **1.1.** 출산 순위별 출생 성비

(단위: 여아 100명당 남아 출생 수)

연도 \ 출생	전체	첫째 자녀	둘째 자녀	셋째 자녀	넷째 자녀 이상
1980	105.3	106.0	106.5	106.9	110.2
1985	109.4	106.0	107.8	129.2	146.8
1990	116.5	108.5	117.0	188.8	209.2
1995	113.2	105.8	111.7	177.2	203.9
2000	110.2	106.2	107.4	141.7	167.5
2005	107.7	104.8	106.4	127.7	132.6

(출처: 통계청, 인구동태 통계연보)

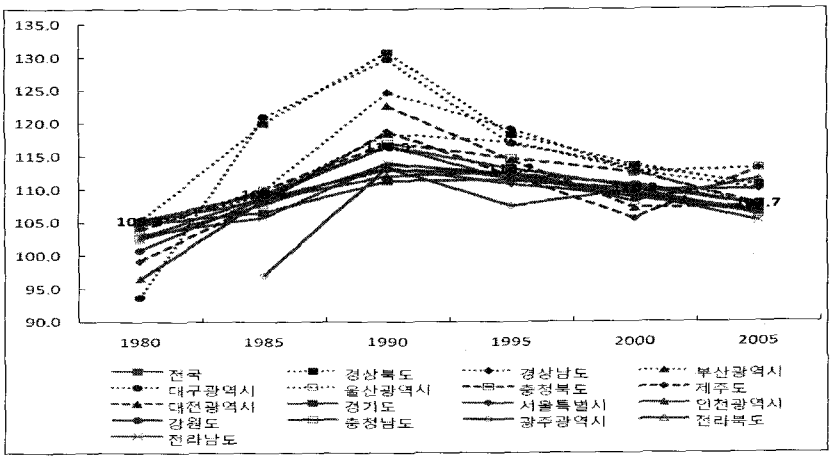


그림 **1.1.** 지역별 출생 성비 변화

이러한 출생 성비의 불균형 현상을 모니터링 하기 위해서 시계열 분석 또는 지역 차이 분석 등의 많은 연구들이 있었다 (Öberg, 1990; Böhning, 1999; 서문희, 1995; 김응식과 박상화, 1996). 하지만 위의 연구들은 고정된 한 시점에서의 지역적 이질성(heterogeneity)을 연구하거나 특정한 한 지역에서의 시간의 흐름에 따른 출생 성비의 변화를 분석하여, 시간과 공간을 동시에 분석에 포함시키지 못했다는 단점이 있다. 따라서 본 연구에서는 시간과 지역성을 동시에 고려하는 통계적 모형을 가지고 출생 성비의 불균형 현상을 분석하고자 한다.

출생아 수 n 중 남아, 여아의 출생 수를 각각 n^M, n^F라 하고($n = n^M + n^F$), 남아 출생 수 n^M은 모수 $\pi(= n^M/n)$를 갖는 이항분포를 따른다고 가정하자. 그림 1.2와 같이 모집단이 동질성(homogeneity)을 갖는 경우에는 하나의 모수 π를 가진 분포로서 모집단을 설명할 수 있으나, 알려지지 않은 이질성을 갖는 모집단의 경우에는 한 개의 모수를 가진 분포로 모집단 안의 K개의 부분 모집단(sub-population)의 다양한 특성의 변화를 설명하기에는 부족하여, 이 경우 서로 다른 K개의 모수를 갖는 이항분포의 혼합모형을 고려할 수 있다.

본 연구는 우리나라의 행정구역상으로 구분된 234개의 구, 시, 군 지역에 대하여 2000년과 2001년도의 각 지역별 셋째 자녀 이상의 남녀 출생아 수 자료를 가지고 이질성이 있는 모집단을 가정하여 이항분포의 혼합모형을 적용하여 분석하는 방법을 제안한다. 물론 90년대의 출생 성비의 차이가 더 심각하므로 이질성이 있는 모집단 자료로 더 적당하겠으나, 통계청에서 제공되는 자료의 형태가 구, 시, 군 지역으

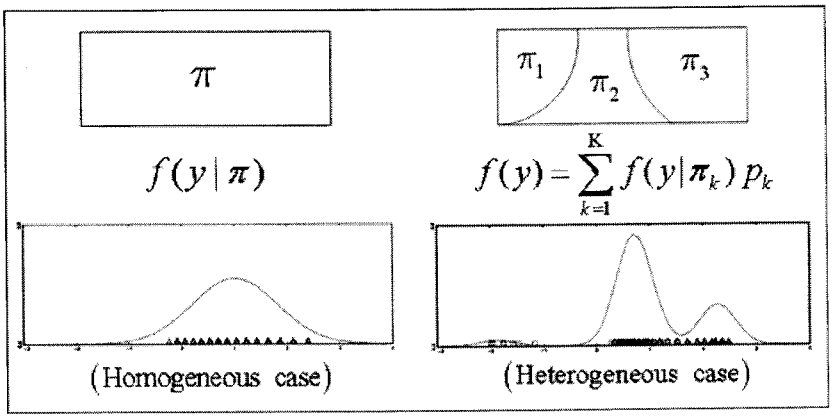


그림 **1.2.** 동질성(HOMOGENEITY)과 이질성(HETEROGENEITY)

로 구분된 출생아 수 자료의 경우에는 2000년부터 얻을 수 있어 본 연구는 2000년과 2001년도 자료를 가지고 분석하였다. 그리고 출생아 수 전체를 가지고 분석하기 보다는 셋째 자녀와 넷째 자녀 이상의 경우 자연 상태의 출생 성비와 차이가 많이 나므로 셋째 자녀 이상의 출생아 수 자료를 가지고 시간과 지역에 따른 출생 성비의 차이를 분석하였다.

2절에서는 동질적 모집단에서의 출생률에 대한 모형을 소개하고, 3절에서는 이질적 모집단의 경우에 적용할 수 있는 모형인 이항분포의 혼합모형에 대해 설명하고 모수 추정 방법인 EM 알고리즘을 소개한다. 4절에서는 출생 자료에 대한 진단 및 분석한 결과를 제시하고 5절에서 결론을 맺는다.

2. 동질적 모집단에서의 출생률에 대한 모형

자료가 하나의 모수를 갖는 분포로 설명될 수 있는 동질적 모집단 경우에 출생률에 대한 모형은 다음과 같다. i번째 지역의 j번째 시점에서 관측된 출생아 수 n_{ij} 중 남아, 여아의 출생 수를 각각 n_{ij}^M, n_{ij}^F라 하고($n_{ij} = n_{ij}^M + n_{ij}^F$), 남아 출생 수 $y_{ij} = n_{ij}^M$는 모수 π_{ij}를 갖는 이항분포를 따른다고 가정하자. y_{ij}의 확률밀도함수는

$$f(y_{ij}|\pi_{ij}) = \binom{n_{ij}}{y_{ij}} \pi_{ij}^{y_{ij}} (1-\pi_{ij})^{n_{ij}-y_{ij}}, \tag{2.1}$$

여기서 π_{ij}은 i번째 지역의 j번째 시점에서의 남아 출생률로서 정준연결함수를 사용한 일반화선형혼합모형(GLMM)은 다음과 같다.

$$\text{logit}(\pi_{ij}) = X_{ij}'\beta_j + Z_{ij}'b_{ij}, \quad i = 1,\ldots,n, \quad j = 1,\ldots,J, \tag{2.2}$$

여기서 $X_{ij} = (1, x_{ij1}, \ldots, x_{ijp})'$는 $(p+1)$차원의 고정효과요인이고, $\beta_j = (\beta_{j0}, \ldots, \beta_{jp})'$는 고정효과에 대한 $(p+1)$차원 모수벡터이다. $Z_{ij} = (z_{ij1}, \ldots, z_{ijq})'$는 임의효과에 대한 q차원 계획행렬이고, b_{ij}는 i번째 지역의 j번째 시점에서의 임의계수로서, b_{ij}의 분포는 $b_{ij} \sim MVN(0, \sum)$을 가정한다.

3. 이질적 모집단에서의 이항분포의 혼합모형

모집단이 여러 개의 부분 모집단으로 구성된 이질적 특성이 있다면 서로 다른 K개의 모수를 고려해야 하는데, 표준 가정은 다음과 같다.

• K는 고정된 상수이고 미지이다.

• 반응변수 $y_{ij},\ i=1,\ldots,n,\ j=1,\ldots,J$는 조건부 독립인 이항분포를 따른다.

$$y_{ij}|b_{ij} \sim \text{Bin}(n_{ij},\pi_{ij}), \qquad \text{여기서 } \text{logit}(\pi_{ij}) = X_{ij}{}'\beta_j + Z_{ij}{}'b_{ij}. \tag{3.1}$$

• 관측되지 않은 다변량 임의계수 벡터 $B_i = (b_{i1},\ldots,b_{iJ})$에 대한 분포 $G(B_i)$는 위치벡터 $B_k = (b_{k1},\ldots,b_{kJ}),\ k=1,\ldots,K$와 질량 $P(B_i = B_k) = p_k,\ \sum_{k=1}^{K} p_k = 1$을 갖는 이산확률 질량함수이다. 즉, $B_i \sim \sum_{k=1}^{K} p_k\ N(B_k,\Omega)$이고 각 부분 모집단은 동일한 공분산 Ω를 가진다고 가정한다.

우도함수는 다음과 같이 주어진다.

$$\prod_{i=1}^{n}\sum_{k=1}^{\text{K}}\left[\prod_{j=1}^{J} f\,(y_{ij}|B_k)\right] p_k. \tag{3.2}$$

EM 알고리즘을 이용하여 우도함수를 최대화시키면 혼합모형의 위치모수와 가중치, 회귀계수, 상관계수를 얻을 수 있는데 추정 과정은 다음과 같다.

단계 1. 추정할 모수에 대한 임의의 초기값을 설정한다.

$$\mathbf{b} = \begin{pmatrix} b_{11} & b_{21} & \cdots & b_{K1} \\ \vdots & \vdots & \ddots & \vdots \\ b_{1J} & b_{2J} & \cdots & b_{KJ} \end{pmatrix}, \qquad \mathbf{p} = (p_1 \ \ p_2 \ \ \cdots \ \ p_K)\,,$$

여기서 b_{kj}와 p_k는 k번째 부분 모집단의 j번째 시점에서의 위치모수와 가중치이다.

단계 2. E-단계

$$w_{ik} = \frac{p_k \prod_{j=1}^{J} f(y_{ij}|x_{ij},B_k)}{\sum_{k=1}^{K} p_k \prod_{j=1}^{J} f(y_{ij}|x_{ij},B_k)}.$$

단계 3. M-단계

$$\hat{p}_k = \sum\nolimits_{i=1}^{n} \frac{w_{ik}}{n}$$

$$\sum\nolimits_{i=1}^{n}\sum\nolimits_{k=1}^{\text{K}} w_{ik}\frac{\partial \log(f_{ik})}{\partial \delta} = 0 \ \ \rightarrow \ \ \hat{\delta} = \left(\hat{b}_{kj},\,\hat{\rho}\right),$$

$$\text{여기서} \quad f_{ik} = \prod_{j=1}^{J} f(y_{ij},\ b_{kj})$$

단계 4. 위의 E-단계와 M-단계를 반복 수행하여 관측된 로그우도함수 ℓ이 다음을 만족하면 실행을 중지한다.

$$\ell = \sum_{i=1}^{n} \log\left[\sum_{k=1}^{\text{K}-1} p_k f_{ik} + \left(1 - \sum_{k=1}^{\text{K}-1} p_k\right) f_{ik}\right], \qquad \frac{|\ell^{(r+1)} - \ell^{(r)}|}{|\ell^{(r)}|} < 10^{-6}.$$

자세한 내용은 임화경 등 (2008)을 참고하길 바란다.

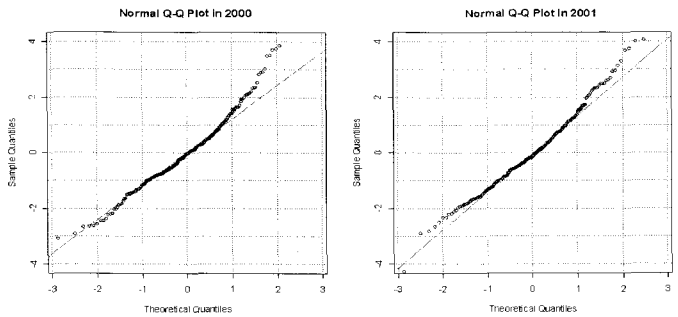


그림 **4.1.** 2000년, 2001년 셋째 자녀이상의 남아 출생률에 대한 Q-Q PLOT

표 **4.1.** 혼합 이항분포를 통해 구한 추정치들과 동질적 모집단 추정치의 비교

# of Mixing Components		Location parameter $\hat{\beta}_{kj}(\hat{\pi}_{kj})$ $J=1$	$J=2$	Weighting Prob.	Rho $\hat{\rho}$	Log-Likelihood $\bar{\ell}$
동질적(Homogeneous) 모집단						
$K=1$	Clust1	0.35(0.59)	0.34(0.58)	1.00	0.998	−1978.85
이질적(Heterogeneous) 모집단						
$K=2$	Clust1	−0.17(0.46)	−0.18(0.45)	0.85	0.997	−1914.04
	Clust2	0.16(0.54)	0.14(0.54)	0.15		
$K=3$	Clust1	−0.31(0.42)	−0.29(0.43)	0.13	0.998	−1910.52
	Clust2	−0.15(0.46)	−0.17(0.46)	0.73		
	Clust3	0.17(0.54)	0.15(0.54)	0.14		
$K=4$	Clust1	−0.31(0.42)	−0.29(0.43)	0.12	0.998	−1908.29
	Clust2	−0.15(0.46)	−0.17(0.46)	0.72		
	Clust3	0.13(0.53)	0.12(0.53)	0.14		
	Clust4	0.35(0.59)	0.28(0.57)	0.02		

4. 분석 결과

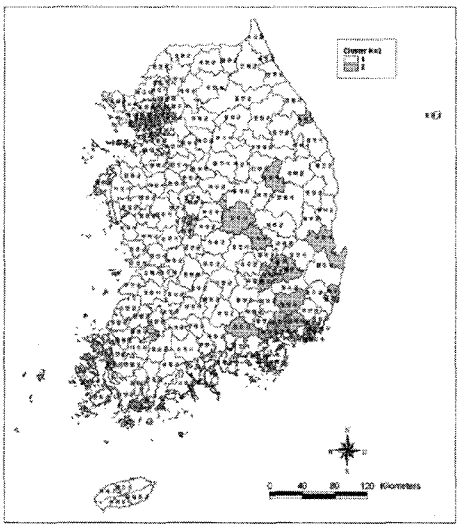
4.1. 출생 자료에 대한 진단

i번째 지역의 j번째 시점에서 남아 출생률의 추정값은 $\hat{\pi}_{ij} = n_{ij}^M / n_{ij}$이고, 표본의 개수 n_{ij}가 충분히 크면 정규근사를 이용하여 **Z**통계량의 값은 다음과 같이 표현할 수 있다.

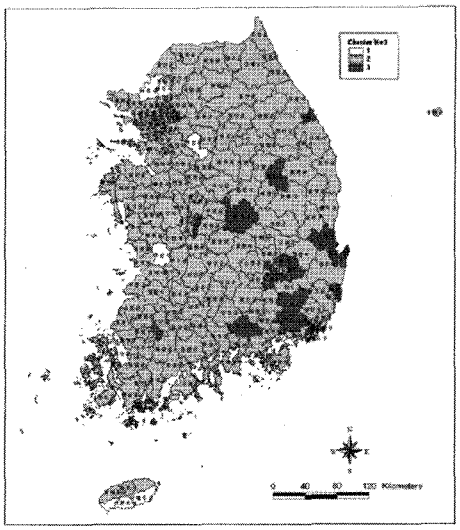
$$\mathbf{Z}_{ij} = \frac{\hat{\pi}_{ij} - \pi_{\cdot j}}{\sqrt{\pi_{\cdot j}(1-\pi_{\cdot j})/n_{ij}}}, \quad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, J, \tag{4.1}$$

여기서 $\pi_{\cdot j}$은 모집단의 j번째 시점에서의 남아 출생률을 의미한다. 동질적 모집단의 분포 하에서 **Z**통계량은 표준정규분포를 따르므로, 2000년과 2001년의 셋째 자녀 이상의 남아 출생률에 대한 자료가 정규 모집단의 표본인지를 알아보기 위해 Q-Q plot을 그려보았다 (그림 4.1 참조).

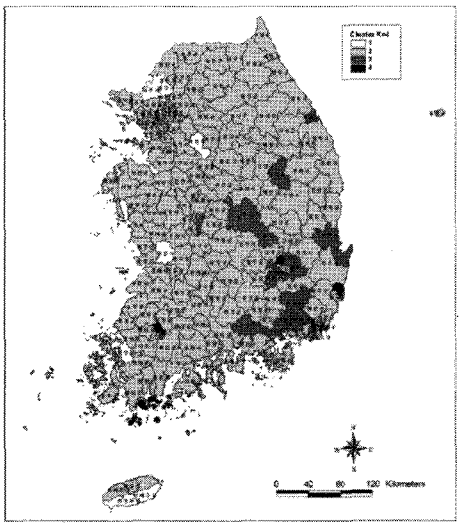
자료의 중간부분은 정규분포를 따르지만 상위, 하위의 꼬리부분은 정규분포와 다르게 나타나 모집단의 이질성을 의심할 수 있다.



(a) 군집 수 $K = 2$인 경우



(b) 군집 수 $K = 3$인 경우



(c) 군집 수 $K = 4$인 경우

그림 4.2. 남아 출생률에 대한 지역별 이질성

4.2. EM 알고리즘을 이용한 모수 추정 결과

2000년과 2001년의 셋째자녀 이상의 남아 출생률에 대하여 EM 알고리즘을 통해서 구한 혼합분포의 추정치는 표 4.1과 같다. 로그우도함수를 이용하여 적당한 부분 모집단의 개수를 선택하였는데 (Böhning, 1999), 이질적 모집단을 가정한 경우 적당한 부분 모집단의 개수는 로그우도값의 변화가 그리 크지 않게 되는 $K = 3$ 또는 4가 적당해 보인다. $K = 1$인 동질적 모집단의 경우와 비교해본다면 로그 우도함수값이 약 $68 \sim 71$정도 차이가 나므로 이질성이 있다고 가정한 경우가 모집단을 더 잘 설명해준다고 말할 수 있다. $K = 1$인 동질적 모집단에 비해 혼합분포의 수가 증가할수록 위치모수들은 다양한 값을 지

니며, 가중치를 보면 대부분의 지역들이 Cluster2에 속하는 것을 알 수 있다. 본 연구는 혼합분포의 수 K가 증가할수록 추정해야할 모수의 개수가 많아지므로, $K = 5$ 이상은 고려하지 않았다.

z_{ik}가 i번째 지역이 k번째 부분 모집단에 있을 경우($z_{ik} = 1$)를 나타낸 관측되지 않은 지시변수라고 하자. 그러면 i번째 지역이 k번째 부분 모집단에 속할 사후확률(posterior probabilities)은 다음과 같다.

$$P(z_{ik} = 1|y_{ij}, n_{ij}) = w_{ik} = \frac{\binom{n_{ij}}{y_{ij}} \pi_{kj}^{y_{ij}} (1 - \pi_{kj})^{n_{ij} - y_{ij}} \Pr(z_{ik} = 1)}{\sum_{k=1}^{K} \binom{n_{ij}}{y_{ij}} \pi_{kj}^{y_{ij}} (1 - \pi_{kj})^{n_{ij} - y_{ij}} \Pr(z_{ik} = 1)}. \tag{4.2}$$

모든 지역은 사후확률을 최대화 시키는 부분 모집단으로 분류되는데 ArcView GIS를 이용하여 그린 결과가 그림 4.2에 있다. 우리나라에서 남아선호사상이 심한 경상도 지역이 다른 지역들에 비해 남아 출생률이 높게 구분되어 지역별 남아 출생률의 이질성을 살펴볼 수 있는데, 그림 4.2(a) $K = 2$인 경우를 보면 대체로 경북, 경남, 부산, 대구 등 영남지역과 전남 완도군의 남아 출생률이 다른 지역에 비해 높게 구분됨을 볼 수 있고, 그림 4.2(b) $K = 3$으로 분류하면 그림 4.2(a) $K = 2$에서 남아 출생률이 다소 낮았던 Cluster1 지역들이 좀 더 세분화되어 서울(서대문), 경기도(파주시, 양주시, 부천시, 이천시, 광주군), 전북(익산시), 전남(목포시, 무안군), 제주도(서귀포시, 남제주군) 지역들은 남아 출생률이 가장 낮은 그룹으로 분류되었다. 그림 4.2(c) $K = 4$로 군집을 분류하면 그림 4.2(b) $K = 3$에서 남아 출생률이 가장 높게 구분된 Cluster3 지역들이 더 세분화 되어 부산(북구), 대구(동구, 북구), 전남(완도)지역은 가장 높은 남아 출생률을 가지고 있는 지역으로 분류되었다.

5. 결론

본 연구에서는 K개의 부분 모집단을 가진 이질성이 있는 모집단의 경우에 있어서 시간과 공간을 동시에 고려한 이항분포의 혼합모형을 제시하였다. 행정구역상으로 구분된 234개의 구, 시, 군 지역에 대하여 2000년과 2001년도의 각 지역별 셋째 자녀 이상의 남녀 출생아 수 자료를 가지고 분석한 결과 지역별 이질성이 존재함을 확인할 수 있었고, 동질성을 가정한 경우보다 모집단을 좀 더 잘 설명할 수 있음을 보였다. 이질성이 있는 지역들은 미래에 주의 깊게 모니터링 할 수 있기 때문에 인구정책상 좋은 정보로 이용될 수 있을 것이라 생각한다.

한편, 각 지역별 출생 성비에 영향을 미칠 수 있는 관련된 변수(직업, 소득수준, 종교, 출신지역(원적), 부인의 연령, 부인의 교육수준)들이 포함된 패널 형태의 자료가 있다면 좀 더 다양한 분석이 가능할 것이다.

감사의 말씀

본 연구에 대해 세심한 지도를 해주신 고려대학교 통계학과 허명회 교수님께 진심으로 감사의 말씀을 전합니다. 그리고 논문을 지도해 주시고 연구자의 자세를 가르쳐 주셨던 故 송석헌 교수님께 이 논문을 바칩니다.

참고문헌

조남훈, 서문희 (1994). 성비의 불균형 변동추이와 대응방안, <한국보건사회연구보고서>, 94-16.

서문희 (1995). 우리나라 출생성비 불균형의 지역차이에 관한 연구, <보건사회연구>, **15**, 143–173.

김응식, 박상화 (1996). 지역별 출생성비의 시계열적 추이에 관한 연구, <한국보건통계학회지>, **21**, 29–36.

임화경, 송석헌, 송주원, 전수영 (2008). 다변량 다수준 이항자료를 위한 일반화선형혼합모형, <응용통계연구>, **21**, 923–932.

통계청 (1980–2005). 인구동태통계연보.

Öberg, St. (1990). Spatial mapping of sex Ratios, *Popnet*, **18**, 5–10.

Böhning, D. (1999). *Computer Assisted Analysis of Mixtures and Applications: Meta-Analysis, Disease Mapping and Others*, Chapman & Hall/CRC, London.

Heterogeneity Analysis of the Male Birth Ratio Data

Hwa-Kyung Lim[1] · Seuck-Heun Song[2] · Juwon Song[3]

[1]Department of Statistics, Korea University; [2]Department of Statistics, Korea University; [3]Department of Statistics, Korea University

(Received October 2008; accepted November 2008)

Abstract

Since 1990, identifying the sex of fetus and illegal abortion has brought the sex ratio imbalance at birth in Korea due to a notion of preferring a son to a daughter, socio-economic development, population policy, and so forth. Although there have been many researches such as time series analysis and region difference analysis to monitor this sex ratio imbalance, they have a defect that time and space could not be included in the analysis simultaneously. This study analyzes the sex ratio imbalance at birth, taking into account time and region at the same time. The analysis considered the numbers of male and female babies, who were born as the third or latter in their families, in 2000 and 2001 at 234 Gu / Si / Goon administrative districts. Here, we suggest a mixture model of binomial distributions, assuming heterogeneous populations. The estimation of the location parameters, weights and correlation coefficient of the mixture model is conducted by the EM algorithm, and the heterogeneity of the regions is expressed as a picture using ArcView GIS.

Keywords: Binomial data, male birth ratio, heterogeneity, mixture model, EM-algorithm.

[3]Corresponding Author: Associate Professor, Department of Statistics, Korea University, Anam-Dong 5, Sungbuk-Gu, Seoul 136-701, Korea. E-mail: jsong@korea.ac.kr