Journal of the Korean Data & Information Science Society
2015, **26**(2), 313–322

http://dx.doi.org/10.7465/jkdi.2015.26.2.313
한국데이터정보과학회지

이항-퇴화 혼합분포의 최우추정법[†]

황선영[1] · 손승혜[2] · 오창혁[3]

[123]영남대학교 통계학과

접수 2015년 1월 17일, 수정 2015년 2월 9일, 게재확정 2015년 3월 18일

요 약

본 연구에서는 하나의 균일분포 또는 퇴화분포와 두 개의 이항분포의 혼합분포 모형에 대하여 최우추정법을 소개하며, 제시된 모형에 대하여 시뮬레이션을 통해 최우추정량의 성질을 밝히며, 실험을 통해 얻은 강의 평가 자료에 대하여 퇴화분포를 가지는 혼합분포에 대하여 적용하여 보았다. 특히 퇴화분포는 한국의 문화 특성상 가운데 값을 선호하는 현상을 모형화하는데 유용하게 사용될 수 있음을 보였다.

주요용어: 우도함수, 이산균일분포, 최우추정법, 퇴화분포, 혼합분포.

1. 서론

혼합분포는 분포의 이질성을 나타내는 유용한 방법이며 자료가 얻어지는 모집단이 두 개 이상의 이질적 집단으로 구성되어 있는 경우에 여러 분야에서 폭넓게 사용되고 있다 (McLachlan과 Peel, 2001). 혼합분포는 몇 개의 성분분포로 이루어지며, 성분분포는 연속형 또는 이산형이 될 수 있다. 성분분포가 이산형인 경우는 이항분포, 포아송분포, 이산균일분포 등이 흔히 사용된다. 그중에서 이항분포를 성분으로 가지는 혼합분포의 이론과 적용에 대한 많은 연구가 이루어져 왔다 (Blischke, 1964; Johnson 등, 2005; Liu 등, 2006). 한편, Oh (2014)는 이동 이항분포의 혼합분포의 최우추정치를 찾는 방법을 제안하였고, Bonnini 등 (2012)은 이항분포와 이산균일분포의 혼합분포에서, Domenico (2003)는 이산균일분포와 이동 이항분포의 혼합분포에서, Lee와 Oh (2006)와 Oh (2006)는 이동 포아송분포의 혼합분포에서 모수의 추정과 적용문제를 다루었다.

특히 이산균일분포와 이항분포의 혼합분포는 상품선호도와 같은 고객 만족도 조사에서 상품을 사용해 본 적이 있는 집단과 그렇지 않는 집단을 모형화하는 데 적용되었다 (Piccolo와 D'Elia, 2008; Iannario와 Piccolo, 2011; Iannario 등, 2012; Kenett와 Salini, 2011; Iannario, 2012b). 또한, D'Elia와 Piccolo (2005)은 설문지 응답자가 두 개의 서로 다른 형태의 균질 하위 그룹으로 나누어진다는 가정 하에서 서수형 자료에 대하여 혼합분포를 사용하였다. Cicia 등 (2010)은 커피소비자의 상품선호도를 측정하는 모형으로 이항분포와 이산균일분포의 혼합분포를 사용하여 불확실한 집단과 선호집단으로 구분하였다. Piccolo (2003)는 심리학 분야의 응용문제에 적용하기 위해 이산균일분포와 이항분포의 혼합분포를 고려하였으며, 두 집단으로부터 얻은 자료를 적합하는데 유용하다.

[†] 이 연구는 2011년도 영남대학교 연구비 지원에 의해 수행되었음.
[1] (712-749) 경북 경산시 대학로 280, 영남대학교 통계학과, 석사과정.
[2] (712-749) 경북 경산시 대학로 280, 영남대학교 통계학과, 석사학위.
[3] 교신저자: (712-749) 경북 경산시 대학로 280, 영남대학교 통계학과, 교수. E–mail: choh@yu.ac.kr

Iannario와 Piccolo (2009)는 혼합분포의 추정문제를 다루었다. Iannario (2010)는 이동 이항분포와 이산균일분포의 혼합분포에서 성분확인성 (identifiability)를 조사하였다. Iannario (2012a)는 최우추정법에서 EM 알고리듬의 계산속도를 향상시키기 위하여 이산균일분포와 이항분포의 사전추정량을 제시하였다. 더욱이 Corduas (2011)는 서수형 자료에 대한 이항분포의 혼합분포를 사용하여 평가자 집단을 군집화하는 절차를 제안하였다.

강의에 대한 학생 평가에서 평가자의 집단에 대한 동질성의 가정이 적절하지 않다는 연구 결과가 제시되었다. Beran과 Violato (2005)은 평점에 대한 기대와 학생들의 강의 평가가 유의할만한 수준의 관련이 있다는 사실을 371,131명의 캐나다 대학생 강의평가 설문지 (the Universal Student Ratings of Instruction instrument in Canada)로부터 얻었다. 따라서 좋은 성적을 기대하는 학생들과 그렇지 않은 학생들은 강의 평가에서 서로 다른 집단으로 간주될 수 있다.

한편, 한국에는 '중용'이라는 문화가 있어 설문 조사시 응답자가 그 질문에 대하여 잘 모르거나 혹은 대답하기 싫을 때 응답항 중에서 가운데 항을 선택하는 경우가 있다. 이는 이러한 집단이 무작위로 응답한다고 하는 가정이 성립하지 않으며 오히려 가운데 문항에 고착하여 응답한다는 가정이 보다 타당한 것으로 보인다. 이 경우에는 무작위 응답에 대한 균일분포의 대응대신에 '무관심' 응답에 대한 퇴화분포의 적용이 더 적절한 것으로 생각된다. 따라서 본 연구에서는 하나의 이산균일분포 (또는 퇴화분포)와 두 개의 이항분포의 혼합분포를 모형으로 제시하고 이 모형에 대한 모수의 추정치 방법을 찾아본다. 제시된 방법은 Dempster 등 (1977)의 EM 알고리듬을 사용하여 최우추정량을 찾는 것이며, EM 알고리듬과 적용 방법은 McLachlan과 Krishnan (2008)에 자세히 소개되어 있다. 최우추정치를 구하는 EM 알고리듬의 개별 단계를 제시하며, 추정치의 성질을 파악하기 위하여 시뮬레이션 실험을 하였으며, 실험을 통하여 얻은 대학생 강의 평가 자료에 대하여 이 모형을 적합시켜서 보았다.

다음 절에서는 이산 균일분포와 두 이항분포의 혼합분포의 추정을 위한 EM 알고리듬을 서술한다. 제 3절에서는 시뮬레이션 결과를 제시하고 설명한다. 제4절에서는 강의평가자료에 대한 제시된 모형을 적합하여 본다. 마지막 절은 논의과 결론을 제시한다.

2. 혼합분포와 EM 알고리듬

범위 $\{0, 1, 2, \cdots, m\}$ 상에서 확률함수

$$f_U(x) = \frac{1}{m+1}, \tag{2.1}$$

$$f_B(x; m, \theta) = \binom{m}{x} \theta^x (1-\theta)^{m-x}, \tag{2.2}$$

를 가지는 이항분포와 이산균일분포를 생각하자. 단, θ는 0과 1 사이의 실수이며, m은 일반성을 잃지 않고 주어진 양의 짝수이라고 가정한다. 하나의 이산균일분포와 두 개의 이항분포의 혼합분포는 확률함수

$$f(x; \Phi) = \pi_0 f_U(x) + \sum_{i=1}^{2} \pi_i f_B(x; \theta_i) \tag{2.3}$$

를 가진다고 한다. 단, π_i, $i = 0, 1, 2$는 성분 i에 대한 가중치로써 $0 \le \pi_i \le 1$와 $\pi_0 + \pi_1 + \pi_2 = 1$를 만족한다. 여기서 모수에 대하여 기호 $\Phi = (\pi_0, \pi_1, \pi_2; \theta_1, \theta_2)$를 사용한다. 식 (2.3)의 확률함수를 가지는 확률변수 X의 평균과 분산은 다음과 같이 얻어진다:

$$E(X) = \pi_0 \frac{m}{2} + \pi_1 m\theta_1 + \pi_2 m\theta_2 \tag{2.4}$$

$$
\begin{aligned}
V(X) = {} & \pi_0(1-\pi_0)\left(\frac{m}{2}\right)^2 + \pi_0\frac{m^2}{12} + \pi_1(1-\pi_1)(m\theta_1)^2 + \pi_1 m\theta_1(1-\theta_1) \\
& + \pi_2(1-\pi_2)(m\theta_2)^2 + \pi_2 m\theta_2(1-\theta_2) \\
& - m^2\left\{\pi_1\theta_1(\pi_0+\pi_2\theta_2) + \pi_2\theta_2(\pi_0+\pi_1\theta_1)\right\}
\end{aligned} \tag{2.5}
$$

확률함수 (2.3)은 확률함수의 가중합으로 표현되기 때문에 이와 관련된 최우추정치를 구하는 것은 어려운 문제이다. 따라서 EM 알고리듬과 같은 반복적 방법을 적용하는 것이 보편적이다.

n개의 관측치 $\boldsymbol{x} = (x_1, \cdots, x_n)$가 식 (2.3)을 따르는 혼합분포에서 얻어졌다고 가정하자. 여기서 자료는 성분에 관한 정보가 없는 불완전자료라고 가정한다. 이런 구조에서는 확률함수 $f(x;\Phi)$는 불완전자료에 대한 확률함수로 간주된다. 모수 Φ에 대한 우도함수는 다음으로 주어진다:

$$
L_{\boldsymbol{x}}(\Phi) = \sum_{j=1}^{n} \log f(x_j;\Phi). \tag{2.6}
$$

우도함수 $L_{\boldsymbol{x}}(\Phi)$를 최대화시킴으로 최우추정치 $\widehat{\Phi}_{\boldsymbol{x}}$ of Φ를 얻을 수 있다. 최우추정치를 얻기 위하여 EM 알고리듬을 적용하며 이를 위해 먼저 자료가 어느 집단에서 얻어졌는지를 파악하는 성분벡터를 추정한다. 관측값 x_j에 대한 성분벡터 $z_j = (z_{j0}, z_{j1}, z_{j2})$는 x_j의 성분에 관한 정보를 제시하며 원소 z_{ji}는 x_j의 값이 성분 j에 속하면 1 아니면 0이 된다.

자료 $(x_1, z_1), \cdots, (x_n, z_n)$는 $\boldsymbol{x}$에 대하여 EM 알고리듬을 위한 완전자료로 간주되며 완전자료에 대한 우도함수는 다음과 같이 주어진다:

$$
L_{\boldsymbol{x},\boldsymbol{z}}(\Phi) = \sum_{j=1}^{n} z_{j0}\left\{\log\pi_0 + \log f_U(x_j)\right\} + \sum_{i=1}^{2}\sum_{j=1}^{n} z_{ji}\left\{\log\pi_i + \log f_B(x_j;\theta_i)\right\}. \tag{2.7}
$$

완전자료에 대한 최우추정치는 (2.7)로부터 쉽게 얻어지며 다음과 같다:

$$
\widehat{\pi}_i = \frac{\sum_{j=1}^{n} z_{ji}}{n},\ i = 0,1,2; \quad \hat{\theta}_i = \frac{\sum_{j=1}^{n} z_{ji}x_j}{m\sum_{j=1}^{n} z_{ji}},\ i = 1,2. \tag{2.8}
$$

분모의 값이 0이 되는 것을 피하기 위해 적어도 하나의 관측치가 각 성분으로부터 얻어졌다고 가정한다. 성분값 z_{ji}의 추정치를 $\widehat{z}_{ji}$로, 이에 대응되는 z_j의 추정치를 $\widehat{z}_j$으로 나타내기로 하자. 그러면 추정된 완전자료 $(x_1, \widehat{z_1}), ..., (x_n, \widehat{z_n})$를 이용하여 (2.8)로 모수를 추정할 수 있다.

만약 p번째 반복에서 추정된 모수를 $\Phi^{(p)}$라고 하면 $p+1$번째 반복에서 z_{ji}는 기대값으로 추정된다:

$$
\begin{aligned}
z_{j0}^{(p+1)} &= \frac{\pi_0^{(p)} f_U(x_j)}{\pi_0^{(p)} f_U(x_j) + \sum_{h=1}^{2}\pi_h^{(p)} f_B(x_j;\theta_h^{(p)})}, \\
z_{ji}^{(p+1)} &= \frac{\pi_i^{(p)} f_i(x_j;\theta_i^{(p)})}{\pi_0^{(p)} f_U(x_j) + \sum_{h=1}^{2}\pi_h^{(p)} f_B(x_j;\theta_h^{(p)})},\ i = 1,2.
\end{aligned} \tag{2.9}
$$

$p+1$번째 반복에서 추정된 완전자료 $(x_1, z_1^{(p+1)}), \cdots, (x_n, z_n^{(p+1)})$에 대하여 식 (2.8)을 사용하여 추정치를 얻는다. 모수 추정을 위한 EM 알고리듬의 단계를 다음과 같이 나타낼 수 있다:

EM 알고리듬

단계 1: $p \leftarrow 1$이라고 둔다.

모수 Φ의 초기값을 $\Phi^{(p)} = (\pi_0^{(p)}, \pi_1^{(p)}, \pi_2^{(p)}; \theta_1^{(p)}, \theta_2^{(2)})$ 라고 둔다.

단계 2: (E-단계) 주어진 모수 $\Phi^{(p)}$에 대하여, (2.9)를 사용하여 추정치 $z_1^{(p+1)}, \cdots, z_n^{(p+1)}$를 얻는다.

단계 3: (M-단계) 추정된 완전자료를 $(x_1,\ z_1^{(p+1)}), \cdots, (x_n, z_n^{(p+1)})$ 식 (2.8)에 사용하여 추정치 $\Phi^{(p+1)}$를 얻는다.

단계 4: 만약 모수의 추정치가 수렴조건을 만족하지 않으면, $p \leftarrow p+1$와 $\Phi^{(p)} \leftarrow \Phi^{(p+1)}$로 두고 단계 2로 간다. 만족한다면 최종추정치를 $\widehat{\Phi}$로 둔다.

EM 알고리듬의 단계 4에서는 다음의 수렴 조건을 사용한다: 주어진 상한 $\delta > 0$에 대하여

$$\left| L_{\boldsymbol{x},\boldsymbol{z}}(\Phi^{(p+1)}) - L_{\boldsymbol{x},\boldsymbol{z}}(\Phi^{(p)}) \right| < \delta.$$

한편, 식 (2.1)의 균일분포는 무작위로 대답하는 응답자을 모형화하기 위하여 사용되었으나, 한국에서는 질문에 대하여 잘모르거나 혹은 대답하는 것 자체를 꺼리는 등의 경우에 응답항 중에서 가운데 값을 선택하는 문화가 있다. 이러한 경우에는 가운데 값에 퇴화하는 분포를 사용하는 모형이 적당한 것으로 보인다. 이 경우에는 식 (2.3) 에서 $f_U(x)$ 대신에 퇴화확률함수 $f_D(x) = I_{\{m/2\}}(x)$ 를 사용하는 혼합분포

$$f(x;\Phi) = \pi_0 f_D(x) + \sum_{i=1}^{2} \pi_i f_B(x;\theta_i) \tag{2.10}$$

를 가정하는 것이 타당할 것이다. 이 혼합분포의 평균은 균일분포를 이용한 혼합분포의 평균 (2.4)와 같으며 분산은 식 (2.5)에서 $\pi_1 m^2/12$를 뺀 값으로 주어진다.

퇴화확률함수에 대한 혼합분포 (2.10)를 가정했을 때 모수 Φ의 추정을 위한 EM 알고리듬은 혼합분포 (2.3)에 대한 EM 알고리듬과 같다. 다만, 추정된 완전자료는 식 (2.9)에서 $f_U(\cdot)$ 대신에 $f_D(\cdot)$를 사용하여 얻어진다.

3. 몬테카를로 실험

하나의 이산균일분포 (또는 퇴화분포)와 두 개의 이항분포의 혼합분포에 대한 EM 알고리듬의 성능을 평가하기 위하여 시뮬레이션을 실행하였다. 각 시뮬레이션에서 반복회수는 5000번으로 하고 각 반복에서 표본크기는 n=100으로 하였다. 혼합분포의 성분 가중치는 (π_0, π_1, π_2)=(.2, .4, .4)로 고정하고 (θ_1, θ_2)의 값을 (.1, .9), (.2, .8), (.3, .7), 그리고 (.4, .6)로 변화시켰다. θ와 π의 값의 조합에 대하여, m을 5에서부터 10까지 변화시켰다. Table 3.1과 3.2에 주어진 값은각각 균일분포와 퇴화분포에 대응되는 혼합분포에 대하여 5000번의 시뮬레이션에서 얻은 모수의 추정값의 표본평균과 표본표준오차 (괄호안)이다.

각 시뮬레이션에서 혼합분포를 따르는 $n = 100$개의 값을 생성한 후, EM 알고리듬을 적용하여 추정치를 얻었다. EM 알고리듬을 위한 π와 θ의 초기 추정값은 단순하게 정했다. 즉, $\pi_0 = \pi_1 = \pi_2 = 1/3$, 그리고 $\theta_1 = 0.25$과 $\theta_2 = 0.75$로 하였다. 한편, (π_1, π_2)의 각 값에 대하여, 표본평균은 참값에 가까와 지고 있다. (θ_1, θ_2)의 각 쌍에 대하여서, $\hat{\theta}$의 표준오차는 m의 값이 증가함에 따라 감소한다. 그리고 m이 증가함에 따라 $\hat{\pi}$의 표준오차는 감소하는 경향을 보이고 있다. 두 이항분포의 모수 θ_1과 θ_2가 거리가 멀어질수록, 혼합분포의 두 개의 이항분포 성분은 더 잘 추정되는 것으로 나타나고 있고 이는 당연한 결과라고 할 수 있다. $\hat{\pi}_0$의 추정치의 평균은 참값 0.2보다 작아지는 경향을 보이고 있다. 한편 π_1과 π_2의 추정치의 평균은 참값 0.4보다 커지는 경향을 보이고 있다. 그러나 표본평균과 참값의 차이는 그다지 크지 않다. $\hat{\theta}$의 추정치의 평균은 참값과 상당히 가깝다 특히 θ_1과 θ_2가 상당히 떨어져 있을 때 특

히 그러하다. 각 경우에 π_1과 π_2의 표준오차는 같다. 왜냐하면 이들이 합이 1이고 표본평균의 분산은 각각 $n\pi_1\pi_2$이기 때문이다.

한편 퇴화분포에 의한 혼합분포가 균일분포에 의한 혼합분포에 비하여 표준오차가 전반적으로 작음을 볼 수 있다.

Table 3.1 Sample means and standard errors of 5000 simulations for a mixture of a uniform and two binomial distributions with g=3, π_0= 0.20, π_1= 0.40, and π_2= 0.40. In each simulation the sample size is n=100

θ_1	θ_2	m	π_0	π_1	π_2	θ_1	θ_2
0.100	0.900	5	0.185	0.407	0.408	0.098	0.902
			(0.168)	(0.099)	(0.097)	(0.036)	(0.034)
		6	0.178	0.410	0.411	0.100	0.900
			(0.144)	(0.087)	(0.088)	(0.030)	(0.030)
		7	0.179	0.411	0.410	0.101	0.899
			(0.126)	(0.080)	(0.080)	(0.026)	(0.026)
		8	0.184	0.407	0.408	0.100	0.899
			(0.112)	(0.076)	(0.075)	(0.023)	(0.024)
		9	0.184	0.409	0.407	0.101	0.899
			(0.102)	(0.070)	(0.071)	(0.021)	(0.022)
		10	0.188	0.406	0.406	0.101	0.899
			(0.095)	(0.069)	(0.067)	(0.020)	(0.020)
0.200	0.800	5	0.231	0.382	0.387	0.192	0.807
			(0.276)	(0.151)	(0.154)	(0.073)	(0.074)
		6	0.215	0.392	0.392	0.195	0.804
			(0.243)	(0.135)	(0.135)	(0.056)	(0.054)
		7	0.202	0.399	0.399	0.197	0.803
			(0.214)	(0.120)	(0.121)	(0.043)	(0.043)
		8	0.195	0.403	0.403	0.198	0.801
			(0.192)	(0.111)	(0.109)	(0.036)	(0.035)
		9	0.186	0.407	0.407	0.199	0.801
			(0.175)	(0.102)	(0.102)	(0.032)	(0.031)
		10	0.187	0.407	0.406	0.199	0.801
			(0.162)	(0.096)	(0.096)	(0.028)	(0.029)
0.300	0.700	5	0.179	0.411	0.411	0.299	0.702
			(0.209)	(0.172)	(0.172)	(0.113)	(0.113)
		6	0.170	0.416	0.415	0.297	0.705
			(0.188)	(0.151)	(0.150)	(0.088)	(0.088)
		7	0.175	0.412	0.412	0.296	0.703
			(0.172)	(0.133)	(0.134)	(0.073)	(0.073)
		8	0.167	0.418	0.415	0.298	0.704
			(0.157)	(0.118)	(0.120)	(0.059)	(0.059)
		9	0.174	0.414	0.412	0.298	0.704
			(0.147)	(0.109)	(0.109)	(0.051)	(0.051)
		10	0.179	0.411	0.410	0.298	0.701
			(0.140)	(0.100)	(0.102)	(0.045)	(0.045)
0.400	0.600	5	0.133	0.435	0.433	0.383	0.617
			(0.142)	(0.215)	(0.214)	(0.133)	(0.133)
		6	0.141	0.425	0.434	0.384	0.616
			(0.133)	(0.210)	(0.210)	(0.124)	(0.121)
		7	0.148	0.428	0.424	0.383	0.618
			(0.124)	(0.211)	(0.211)	(0.116)	(0.117)
		8	0.149	0.422	0.429	0.376	0.620
			(0.116)	(0.215)	(0.214)	(0.112)	(0.111)
		9	0.158	0.423	0.419	0.382	0.621
			(0.112)	(0.211)	(0.210)	(0.104)	(0.104)
		10	0.159	0.417	0.424	0.378	0.618
			(0.106)	(0.209)	(0.208)	(0.100)	(0.098)

Table 3.2 Sample means and standard errors of 500 simulations for a mixture of a degenerated and two binomial distribtuions with g=3, π_0= 0.20, π_1= 0.40, and π_2= 0.40. In each simulation the sample size is n=100

θ_1	θ_2	m	π_0	π_1	π_2	θ_1	θ_2
0.100	0.900	5	0.198	0.402	0.400	0.100	0.901
			(0.048)	(0.052)	(0.049)	(0.031)	(0.025)
		6	0.199	0.401	0.400	0.101	0.900
			(0.040)	(0.050)	(0.050)	(0.022)	(0.022)
		7	0.202	0.399	0.399	0.100	0.901
			(0.044)	(0.050)	(0.050)	(0.020)	(0.019)
		8	0.201	0.400	0.399	0.101	0.898
			(0.042)	(0.048)	(0.049)	(0.017)	(0.018)
		9	0.201	0.403	0.396	0.100	0.899
			(0.042)	(0.051)	(0.051)	(0.017)	(0.016)
		10	0.202	0.397	0.401	0.101	0.900
			(0.042)	(0.050)	(0.050)	(0.015)	(0.015)
0.200	0.800	5	0.196	0.410	0.395	0.201	0.804
			(0.065)	(0.090)	(0.065)	(0.057)	(0.045)
		6	0.198	0.398	0.404	0.198	0.799
			(0.053)	(0.057)	(0.060)	(0.036)	(0.039)
		7	0.197	0.403	0.401	0.201	0.801
			(0.054)	(0.069)	(0.057)	(0.040)	(0.030)
		8	0.198	0.404	0.398	0.200	0.803
			(0.046)	(0.053)	(0.054)	(0.028)	(0.027)
		9	0.198	0.402	0.399	0.203	0.801
			(0.046)	(0.055)	(0.050)	(0.028)	(0.024)
		10	0.195	0.403	0.402	0.201	0.799
			(0.043)	(0.052)	(0.052)	(0.023)	(0.023)
0.300	0.700	5	0.189	0.421	0.389	0.293	0.718
			(0.073)	(0.174)	(0.147)	(0.089)	(0.091)
		6	0.187	0.424	0.389	0.299	0.717
			(0.068)	(0.149)	(0.144)	(0.076)	(0.079)
		7	0.192	0.420	0.388	0.303	0.712
			(0.064)	(0.134)	(0.116)	(0.065)	(0.063)
		8	0.191	0.402	0.407	0.297	0.700
			(0.061)	(0.101)	(0.102)	(0.053)	(0.053)
		9	0.192	0.416	0.392	0.303	0.703
			(0.056)	(0.092)	(0.078)	(0.044)	(0.041)
		10	0.194	0.402	0.404	0.298	0.700
			(0.055)	(0.077)	(0.075)	(0.038)	(0.037)
0.400	0.600	5	0.202	0.380	0.418	0.345	0.642
			(0.083)	(0.264)	(0.251)	(0.168)	(0.164)
		6	0.193	0.397	0.411	0.355	0.636
			(0.072)	(0.252)	(0.247)	(0.154)	(0.143)
		7	0.196	0.390	0.414	0.358	0.629
			(0.073)	(0.245)	(0.241)	(0.133)	(0.129)
		8	0.194	0.408	0.398	0.366	0.637
			(0.072)	(0.247)	(0.246)	(0.128)	(0.130)
		9	0.193	0.412	0.395	0.374	0.627
			(0.066)	(0.235)	(0.226)	(0.112)	(0.113)
		10	0.191	0.438	0.371	0.385	0.636
			(0.068)	(0.230)	(0.226)	(0.098)	(0.105)

4. 혼합분포의 적용 예제

최근에 들어 한국의 대부분의 대학교에서는 강의평가를 실시하고 있다. 강의 평가는 학기의 마지막

주 근처에 실시되며 이때는 대부분의 학생들이 자신의 중간고사 성적이나 과제 성적, 자신의 출석, 학습 이해도 등을 바탕으로 해당 과목의 성적을 예측할 수 있다. 학생들 사이에서는 일반적으로 '좋은' 평점은 4점대 (A+ 혹은 A), '나쁜' 평점은 2점대 (C+ 이하)로 인지되고 있다. 좋은 평점을 기대하는 학생은 나쁜 평점을 기대하는 학생들 보다 강의평가에서 강의자를 더 좋게 평가하는 것으로 알려져있다. 한편으로는 강의평가에 대하여 무관심하지만 강의평가에 참가하여야 자신의 성적을 미리 보고 성적 이의신청을 할 수 있는 제도가 있는 경우에 강의평가에서 무관심적으로 즉, 가운데 응답항을 무조건적으로 선택하는 학생이 존재하는 것도 현실이다. 따라서 강의평가에서 자신의 예상 평점에 따라 응답의 형태가 다른 집단이 존재한다는 모형을 가정할 수가 있고, 여기서는 서로 다른 세 개의 집단의 존재를 가정한다.

이를 알아보기 위하여 2014학년도 2학기에 저자 중 한 사람이 가르치는 강의 한 과목의 중간 고사를 끝낸 후 학생들에게 실험용 강의 평가를 수업 중에 실시하였다. 먼저 무작위로 생성된 예상 평점을 학생들에게 나누어 주었다. 예상 평점은 4점 이상과 2점 이하의 두 종류이었다. 그런 다음 예상점수가 마음에 드는지에 대해 '예'와 '아니오'로 답하는 문항과 "강의가 충실하게 진행되었다"고 느끼는지를 묻는 7점 척도의 문항이 제시되었다. 여기서 1점은 '아주 아니다', 4점은 '보통이다', 그리고 7점은 '아주 그렇다'의 순으로 서수적으로 대응된다. 강의평가에 참가한 학생 수는 모두 71명이었고, 그 중에서 예상점수가 마음에 들지 않는다는 설문지 37매에 대한 도수분포는 Table 4.1과 같았다. '보통'에 해당하는 4번 답항의 돗수가 그 주위의 돗수보다 월등히 높음을 볼 수가 있고, 이는 응답자 중에 좋지 않은 평점이 기대된다는 평가에 대하여 불만 등을 표출하는 것이라고 짐작할 수 있다. 따라서 값 4에서의 퇴화분포를 포함하는 모형 (2.10)을 고려한다.

$m = 6$인 경우에 식 (2.10)의 퇴화분포와 두 개의 이항분포의 혼합분포를 가정했을 때의 추정된 모수는 $\hat{\pi}_0 = 0.088$, $\hat{\pi}_1 = 0.491$, $\hat{\pi}_2 = 0.422$, $\hat{\theta}_1 = 0.466$과 $\hat{\theta}_2 = 0.913$이었다. 한편 모형 적합을 위한 χ^2-검정통계량의 값은 0.944이며 이에 대한 유의확률은 대략 0.98이다. Table 4.1의 도수분포와 적합된 도수분포를 나타낸 것이 Figure 4.1이며, 좁은 청색 막대는 추정된 혼합분포의 확률에 관한 것이며 넓은 분홍색 막대는 관측 상대도수에 대한 것이다.

Table 4.1 Frequencies of lecture evaluation data

Response	1	2	3	4	5	6	7	Sum
Frequency	0	2	6	9	4	7	9	37

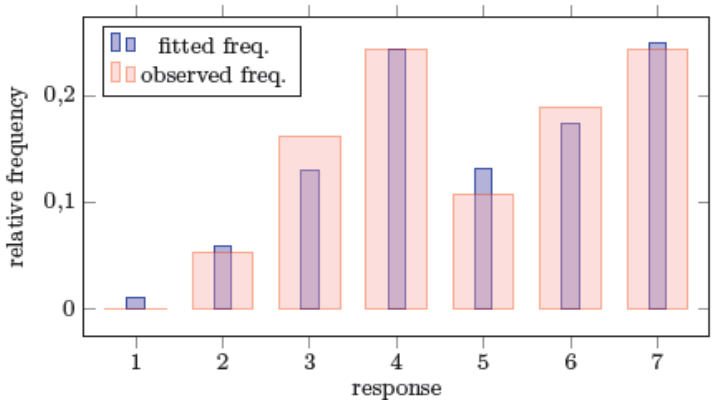

Figure 4.1 Histogram of lecture evaluation data overlayed with fitted mixture distribution.

5. 토의 및 결론

본 연구에서는 하나의 이산균일분포 혹은 퇴화분포와 두 개의 이항분포의 혼합분포 모형을 제안하였고 이를 따르는 확률변수의 평균과 분산을 구하였다. 또한 주어진 혼합분포에 대하여 주어진 자료를 불완전자료로 간주하고 성분 정보를 추정하여 완전자료를 추정하는 방법과 추정된 완전자료로부터 우도함수를 최대화하는 방법을 구하여 EM 알고리듬의 절차를 제시하였다. 또한 몬테카를로 실험을 통하여 제시된 최우추정법이 하나의 이산균일분포 혹은 퇴화분포와 두 개의 이항분포의 혼합분포에서 모수의 추정치들의 평균은 모수의 참값과 가까와짐을 보였다.

제시된 모형을 실제 자료에 적용하기 위하여 표본 조사를 통하여 얻은 강의 평가 자료를 적합하는 데 사용하였으며, 적합의 결과는 만족스러운 것으로 평가할 수 있다.

그러나 강의 평가자료에서 보듯이 강의평가를 극단적으로 잘 하는 집단 즉, 혼합분포의 이항분포 성분 중 $\theta > 0.9$인 경우에 대하는 적합에 다소 문제가 있음을 지적할 수 있다. 이 경우에는 이항분포 대신에 음이항분포 또는 절단 음이항분포를 사용하는 고려해 볼 수 있으며 이는 후속 연구에서 진행될 수 있을 것이다.

본 연구에서 제시한 혼합분포의 모형은 무작위 응답 혹은 무관심 응답자가 존재하는 두 이질적 집단에서의 표본 자료의 분석에서 적절히 사용될 수 있을 것이다.

References

Beran, T. and Violato, C. (2005). Ratings of university teacher instruction: how much do student and course characteristics really matter? *Assessment & Evaluation in Higher Education*, **30**, 593-601.

Blischke, W. (1964). Estimating the parameters of mixture of binomial distributions. *Journal of the American Statistical Association*, **59**, 510-528.

Bonnini, S., Piccolo, D., Salmaso, L. and Solmi, F. (2012). Permutation inference for a class of mixture models. *Communications in Statistics-Theory and Methods*, **41**, 2879-2895.

Cicia, G., Corduas, M., Giudice, T. D. and Piccolo, D. (2010). Valuing consumer preferences with the CUB model: A case study of fair trade coffee. *International Journal on Food System Dynamics*, **1**, 82-93.

Corduas, M. (2011). A study on university students' opinions about teaching quality: a model based approach for clustering ordinal data. In M. Attanasio & V. Capursi Jackson (Eds.), *Statistical Methods for the Evaluation of University Systems*. Heidelberg: Springer.

Čekanavičius, V., Peköz, E. A., Röllin, A. and Shwartz, M. (2009). A three-parameter binomial approximation. Available from `http://arxiv.org/abs/0906.2855`.

Copsey, K. and Webb, A. (2003). Bayesian gamma mixture model approach to radar target recognition. *IEEE Transactions on Aerospace and Electronic Systems*, **39**, 1201-1217.

D'Elia, A. and Piccolo, D. (2005). A mixture model for preference data analysis. *Computational Statistics and Data Analysis*, **49**, 917-934.

Dempster, A. P., Laird, N. M. and Rubin, D. R. (1977). Maximum likelihood from incomplete data. *Journal of the Royal Statistical Society B*, **39**, 1-38.

Domenico, P. (2003). On the moments of a mixture of uniform and shifted binomial random variables. *Quaderni di Statistica*, **5**, 1-20.

Greenwald, A. G. (2002). Constructs in student ratings of instructors. In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement*. New York: Erlbaum.

Iannario M. (2010). On the identifiability of a mixture model for ordinal data. *METRON*, **LXVIII**, 87-94.

Iannario M. (2012a). Preliminary estimators for a mixture model of ordinal data. *Adv Data Anal Classif*, **5**, 163–184.

Iannario M. (2012b). Modelling shelter choices in a class of mixture models for ordinal responses. *Stat Methods Appl*, **21**, 1–22.

Iannario, M., Manisera, M., Piccolo, D. and Zuccolotto, P. (2012). Sensory analysis in the food industry as a tool for marketing decisions. *Adv Data Anal Classif*, **6**, 303–321.

Iannario M. and Piccolo D. (2011). CUB Models: Statistical Methods and Empirical Evidence. *Modern Analysis of Customer Satisfaction Surveys*, Kenett R. S. and Salini S. (Eds). John Wiley and Sons: Chichester: UK.

Johnson, N. L., Kemp, A. W. and Kotz, S. (2005). *Univariate discrete distributions*, 3rd ed., Wiley-Interscience, New York.

Kenett, R. S. and Salini, S. (2011). Modern analysis of customer satisfaction surveys: comparison of models and integrated analysis. *Applied Stochastic Models in Business and Industry*, **27**, 465-475.

Lee, H. J. and Oh, C. (2006). Estimation in mixture of shifted Poisson distributions with known shift parameters. *Journal of the Korean Data & Information Science Society*, **17**, 785-794.

Liu, Z., Almhana, J., Choulakian, V. and McGorman, R. (2006). Online EM algorithm for mixture with application to internet traffic modeling. *Computational Statistics & Data Analysis*, **50**, 1052–1071.

McLachlan, G. J. and Krishnan, T. (2008). *The EM algorithm and extensions*, 2nd ed., Wiley, Hoboken, NJ.

McLachlan, G. J. and Peel, D. (2001). *Finite mixture models*, John Wiley & Sons, Inc., New York.

Oh, C. (2006). Estimation in mixture of shifted Poisson distributions. *Journal of the Korean Data & Information Science Society*, **17**, 1209-1217.

Oh, C. (2014). A maximum likelihood estimation method for a mixture of shifted binomial distributions. *Journal of the Korean Data & Information Science Society*, **25**, 255-261.

Piccolo D. (2003). On the moments of a mixture of uniform and shifted binomial random variables. *Quaderni di Statistica*, **5**, 85-104.

Piccolo D. and D'Elia A. (2008). A new approach for modeling consumers' preferences. *Food Quality Preference*, **19**, 247-259.

Skipper, M. (2012). A Pólya approximation to the Poisson–binomial law. *Journal of Apply Probability*, **49**, 745-757.

Journal of the Korean Data & Information Science Society
2015, **26**(2), 313–322

http://dx.doi.org/10.7465/jkdi.2015.26.2.313
한국데이터정보과학회지

Maximum likelihood estimation for a mixture distribution[†]

Seonyeong Hwang[1] · Seung Hye Sohn[2] · Changhyuck Oh[3]

[123]Department of Statistics, Yeungnam University

Received 17 January 2015, revised 9 February 2015, accepted 18 March 2015

Abstract

A mixture distribution of a discrete uniform or degenerated distribution and two binomial distribution is proposed and a method of obtaining the maximum likelihood estimates of the parameters is provided. For the proposed model simulation studies were conducted to see performance of the maximum likelihood estimates and a mixture of a degenerated distribution and two binomial distributions was applied to fit a lecture evaluation data to show a good result.

Keywords: Binomial distribution, degenerated distribution, likelihood, maximum likelihood, mixture distribution.

[†] This research was supported by the 2011 Yeungnam University Research Fund.
[1] Master program, Department of Statistics, Yeungnam University, Gyeongsan 712-749, Korea.
[2] Master degree, Department of Statistics, Yeungnam University, Gyeongsan 712-749, Korea.
[3] Corresponding author: Professor, Department of Statistics, Yeungnam University, Gyeongsan 712-749, Korea. E–mail: choh@yu.ac.kr