유한모집단 평균에 대한 경험적 베이즈 추정

신 민 웅[1] 신 기 일[1]

요 약

초 모집단으로 부터 반복해 유한 모집단을 추출할 때, 이미 조사된 자료들을 이용하면 현재의 유한모집단 모수들을 더 효율적으로 추정할 수 있다. 이러한 문제에 대하여 Ericson(1969)이 유한 모집단 표본추출에서 베이지안 분석을 하였고, Ghosh와 Meeden(1986)은 정규 초 모집단을 가정하여 유한모집단 평균의 경험적 베이즈 추정을 하였다. Nandram과 Sedransk (1993)는 Ghosh와 Meeden(1986)의 유한모집단들의 분산이 모두 같다는 가정들을 완화하여 유한모집단 평균의 경험적 베이즈 추정을 하였다. 본 연구는 Nandram과 Sedransk의 결과를 층화표본추출의 경우로 일반화 하였다.

Key Words : 유한모집단, 초모집단, 경험적 베이즈 추정

1.서 론

이 논문은 초모집단(superpopulation)으로부터 생성된 유한모집단 (finite population) 표본추출에서 경험적 베이즈 추정(empirical Bayes estimation)문제를 다룬다. 이러한 추정의 예는 작은 지역에서의 추정문제로 인구추정, 실업률, 농산물 산출등의 추정이 있다. 그러한 경우에 어떤 특정 지역의 추정률의 효율성을 높이기 위해서는 유사한 인근지역으로 부터의 정보를 사용하여 더 개선된 추정치를 얻을 수 있다.

더우기, 큰 규모의 조사로 반복해 표본추출된 많은 유한모집단들은 시간의 경과에 따라 천천히 변한다. 결과적으로 이미 조사된 자료(data from earlier surveys)들을 이용하여 현재의 유한모집단 모수추정치(current estimates of finite population parameters)들을 더 개선하여 추정할 수 있다. 예를 들어 색맹같은 것은 그 출현수가 시간의 경과에 따라 매우 안정적으로 변하고 있다.

이러한 문제에 대한 연구로 Ericson(1969)이 유한모집단 표본추출에서의 베이지안 분석을 매우 성공적으로 수행하였다. Ghosh와 Meeden(1986)은 정규초모집단 (normal superpopulation model)을 가정하여 유한모집단 평균의 경험적 베이즈추정을 하였다. 경험적 베이즈 분석에서는 사전 모수들(prior parameters)이 미지이므로

1 한국외국어대학교 통계학과, 경기도 용인군 모현면 왕산리 산89

자료(표본)로 부터 추정되어야 한다.

계속해서 Ghosh와 Lahiri(1987)는 정규성의 가정을 완화하고, 사후선형성(posterior linearity)을 가정하였다. 그러나 그들은 표본분산들이 동일하다는 가정을 하여 실용성에 제한이 있었다. 이러한 가정들을 완화하여 Nandram과 Sedransk(1993)는 표본분산들이 동일하지 않은 경우로 일반화하였다.

본 논문은 2절에는 Nandram과 Sedransk(1993)의 결과를 층화표본추출의 경우로 일반화하여 유한모집단의 층별평균치를 추정한다. 3절에서는 층화표본추출을 m-단계 반복하였을 경우에 현재의 유한모집단 모수들을 추정한다.

2. 층화 유한모집단 추출

이 절에서는 층화표본추출을 할 경우를 생각한다. Y_{ij}를 i번째 층의 j번째 단위 ($i=1,\cdots,l$; $j=1,\cdots,N_i$)를 나타낸다고 하자. i번째 층으로 부터 표본의 크기 n_i인 표본을 $Y_{i1},\cdots,Y_{in_i}, i=1,\cdots,l$로 나타낸다.

우리의 목적은 유한모집단의 층별평균을 구하는데 있다. 즉 i번째 층의 평균

$$\gamma(Y_i)=\sum_{j=1}^{N_i} Y_{ij}/N_i \quad , \quad i=1,\cdots,l$$

을 추정하고자 한다. 단, $Y_i=(Y_{i1},\cdots,Y_{iNi})'$이다.

먼저 추론을 하는데 있어서, 우리는 다음과 같은 초 모집단 모형을 가정한다.

$$Y_{i1},\cdots,Y_{iN_i} \mid \mu_i,\sigma_i^2 \quad \sim^{iid} \quad N(\mu_i,\sigma_i^2) \tag{2.1}$$

로 서로 독립이다.

그리고 각 i에 대하여 서로 독립적으로

$$\mu_i \mid \sigma_i^2,\ \theta,\ \delta_i^2 \quad \sim \quad N(\theta,\delta_i^2) \tag{2.2}$$

라고 가정한다. 단, θ , σ_i^2은 고정되어 있고 미지이다. δ_i^2은 고정되어 있고, 기지라고 가정한다.

2.1 점 추정량(point estimator)

이제 θ , δ_i^2 은 고정되어 있고, $\delta_i>0$, $i=1,\cdots,l$은 기지라고 가정할 때에, $\gamma(Y_i)$의 사후평균(posterior mean)을 e_{Bi}라고 하면,

$$e_{Bi}=\overline{Y}_i-(1-f_i)w_i(\overline{Y}_i-\theta) \tag{2.3}$$

이다. 그리고 $\gamma(Y_i)$의 사후분산을 v_{Bi}^2이라 하면

$$v_{Bi}^2 = (1-f_i)\{f_i + (1-f_i)(1-w_i)\}\sigma_i^2/n_i \tag{2.4}$$

이다.

단, $w_i = (\sigma_i^2/n_i)\{\delta_i^2 + (\sigma_i^2/n_i)\}$이다. 그리고 $\overline{Y}_i = \sum_{j=1}^{n_i} Y_{ij}/n_i$, $f_i = n_i/N_i$, $i=1,\cdots,l$

이다.

σ_i^2을 추정하기 위하여, $s_i^2 = \sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_i)^2/(n_i-1)$ 을 사용한다.

σ_i^2과 δ_i^2이 기지일 때에, θ의 최대우도 추정량은

$$\hat{\theta}_* = \sum_{i=1}^{l}(1-w_i)\overline{Y}_i/\sum_{i=1}^{l}(1-w_i) \tag{2.5}$$

이다. 우리는 w_i의 추정치를 $\hat{w}_i$라 하면

$$\hat{\theta} = \sum_{i=1}^{l}(1-\hat{w}_i)\overline{Y}_i/\sum_{i=1}^{l}(1-\hat{w}_i),\ \delta_i^2>0,\ i=1,\cdots,l \tag{2.6}$$

단, $\hat{w}_i = (s_i^2/n_i)\{\delta_i^2 + (s_i^2/n_i)\}$

즉, i번째 층의 평균의 경험적 베이즈 추정치는

$$e_{EBi} = \overline{Y}_i - (1-f_i)\hat{w}_i(\overline{Y}_i - \hat{\theta}) \tag{2.7}$$

이다.

$\gamma(Y_i)$의 또 다른 추정치로 다음과 같은 추정치를 생각할 수 있다.

$$e_i = \overline{Y}_i$$

이다.

2.2 구간 추정량 (interval estimator)

$\theta, \delta^2, \sigma^2$ 그리고 $Y_{i1},\cdots,Y_{in_i}$가 주어져있을 때, $\gamma(Y_i)$의 $100(1-\alpha)\%$ Highest Posterior Density(HPD) 구간은

$$e_{Bi} \pm v_{Bi}\, z_{\alpha/2}$$

이다. 단, $z_{\alpha/2}$는 표준정규분포의 $100\{1-(\alpha/2)\}\%$ 점이다.

e_i을 중심으로한 신뢰구간은

$$e_i \pm \{(1-f_i)s_i^2/n_i\}^{1/2} z_{\alpha/2} \tag{2.8}$$

이다.

e_{EBi}에 대한 신뢰구간은

$$e_{EBi} \pm v_{EBi}\, z_{\alpha/2}$$

단, $v_{EBi}^2 = \hat{v}_{ui}^2 + \hat{v}_{\theta * i}^2$ 이다. 그리고 $\hat{v}_{ui}^2 = (1-f_i)\{f_i + (1-f_i)\hat{w}_i\}s_i^2/n_i$,

$\hat{v}_{\theta * i}^2 = (1-f_i)^2 \hat{w}_i^2 \left\{ \sum_{i=1}^{l} \hat{w}_i n_i / s_i^2 \right\}^{-1}$ 이다.

3. 층화유한모집단 추출을 m-단계 반복하였을 경우

이 절에서는 층화유한모집단들이 m단계 반복적으로 표본추출되었다고 하자. i-단계 유한 모집단 (i번째 추출된 유한 모집단)의 j번째 층의 원소들을 $X_1^{(j)}, \cdots, X_{N_i}^{(j)}$ ($i = 1, \cdots, m$ 와 $j = 1, \cdots, l$)로 표시한다. i-단계 유한모집단의 j번째 층으로부터 추출된 표본을 $X_1^{(j)}, \cdots, X_{n_i}^{(j)}$로 나타낸다. 즉 표본의 크기는 $n_i^{(j)}$이다.

우리의 목적은 m-단계 유한모집단의 층별평균

$$\gamma(X_m^{(j)}) = \sum_{k=1}^{N_m^{(j)}} X_{mk}^{(j)} / N_m^{(j)}, \quad j = 1, \cdots, l$$

을 추론하는데 있다. 즉 현재의 유한 모집단(m단계 유한모집단)의 층별 평균을 추론하는데 있다.

우리는 초모집단 모형을 다음과 같이 가정한다. 즉, i-단계 유한모집단의 j번째 층에 대하여

$$X_1^{(j)}, \cdots, X_{N_i}^{(j)} \mid \mu_i^{(j)}, \sigma_i^{2(j)} \sim^{iid} N(\mu_i^{(j)}, \sigma_i^{2(j)}) \tag{3.1}$$

$i = 1, \cdots, m$와 $j = 1, \cdots, l$ 이다.

또한, 모든 $i = 1, \cdots, m$ 와 $j = 1, \cdots, l$에 대하여

$$\mu_i^{(j)} \mid \theta, \delta_j^2 \sim N(\theta, \delta_j^2) \tag{3.2}$$

이라고 가정한다. 단, 모든 $i = 1, \cdots, m$ 와 $j = 1, \cdots, l$에 대하여 θ, δ_j^2, $\sigma_i^{2(j)}$은 고정되어 있고 미지이다.

3.1 점 추정량

이제 모든 $i=1,\cdots,m$ 와 $j=1,\cdots,l$에 대하여 θ, δ_j^2 그리고 $\sigma_i^{2(j)}$은 고정되어 있다고 가정한다. 현재 유한모집단(m단계)의 j번째 층의 $\gamma(X_m^{(j)})$의 사후평균과 $\gamma(X_m^{(j)})$의 사후분산을 각각 $e_{Bm}^{(j)}$ 그리고 $v_{Bm}^{2(j)}$라고 하자. 그러면

$$e_{Bm}^{(j)} = \overline{X}_m^{(j)} - (1-f_m^{(j)})w_m^{(j)}(\overline{X}_m^{(j)} - \theta) \tag{3.3}$$

이고

$$v_{Bm}^{2(j)} = (1-f_m^{(j)})\{f_m^{(j)} + (1-f_m^{(j)})(1-w_m^{(j)})\}\sigma_m^{2(j)}/n_m^{(j)} \tag{3.4}$$

이된다. 여기서 $w_m^{(j)} = (\sigma_m^{2(j)}/n_m^{(j)})/\{\delta_j^2 + (\sigma_m^{2(j)}/n_m^{(j)})\}$, $f_i^{(j)} = n_i^{(j)}/N_i^{(j)}$이다. 또한 $\overline{X}_i^{(j)} = \sum_{k=1}^{n_i^{(j)}} X_{ik}^{(j)}/n_i^{(j)}$, $i=1,\cdots,m$이고 $j=1,\cdots,l$이다.

우리는 $\sigma_i^{2(j)}$을 추정하기 위하여 $s_i^{2(j)}$을 사용한다. Nandram과 Sedransk(1993)와 유사하게 구한 $\delta_m^{2(j)}$의 추정치, $\hat{\delta}_j^2$은

$$\hat{\delta}_j^2 = \max(0, \hat{\delta}_{j*}^2) \tag{3.5}$$

이다.

단, $m \geq 4$ 일 때

$$\hat{\delta}_{j*}^2 = \frac{(m-1)(m-3)^{-1}\sum_{i=1}^{m} n_i^{(j)}\{\overline{Y}_i^{(j)} - n_{\cdot}^{-1}(\sum n_i^{(j)}\overline{Y}_i^{(j)})\}^2 - \sum_{i=1}^{m}(1-n_i n_{\cdot}^{-1})s_i^{2(j)}}{\sum_{i=1}^{m} n_i^{(j)}(1-n_i^{(j)} n_{\cdot}^{-1})}$$

$\sigma_i^{2(j)}$과 δ_j^2이 기지일 때는, θ의 최대우도 추정량은

$$\hat{\theta}_* = \frac{\sum_{j=1}^{l}\sum_{i=1}^{m}(1-w_i^{(j)})\overline{X}_i^{(j)}}{\sum_{j=1}^{l}\sum_{i=1}^{m}(1-w_i^{(j)})} \tag{3.6}$$

이다. 이 때 $w_i^{(j)}$의 추정치 $\hat{w}_i^{(j)}$를 사용하면, 우리는 θ의 추정치를 구할수 있다. 먼저 적어도 하나 이상의 $\hat{\delta}_j^2 > 0$ 일 때, θ의 추정치는 다음과 같다.

$$\hat{\theta} = \frac{\sum_{j=1}^{l}\sum_{i=1}^{m}(1-\hat{w}_i^{(j)})\overline{X}_i^{(j)}}{\sum_{j=1}^{l}\sum_{i=1}^{m}(1-\hat{w}_i^{(j)})} \tag{3.7}$$

또한 모든 $\delta_j^2 = 0$ 이면

$$\hat{\theta} = \frac{\sum_{j=1}^{l}\sum_{i=1}^{m} n_i^{(j)}\overline{Y}_i^{(j)}/s_i^{2(j)}}{\sum_{j=1}^{l}\sum_{i=1}^{m} n_i^{(j)}/s_i^{2(j)}} \tag{3.8}$$

이다. 단, $\hat{w}_i^{(j)} = (s_i^{2(j)}/n_i^{(j)})/\{\delta_j^2 + (s_i^{2(j)}/n_i^{(j)})\}$.

따라서, 경험적 베이즈 추정치는

$$e_{EBm}^{(j)} = \overline{X}_m^{(j)} - (1-f_m^{(j)})\hat{w}_m^{(j)}(\overline{X}_m^{(j)} - \hat{\theta}),\ j = 1, \cdots, l \tag{3.9}$$

이다.

우리는 다른 추정치로 $e_{sm}^{(j)} = \overline{X}_m^{(j)}$를 생각할 수 있다.

3.2 구간 추정량

$\delta_j^2,\ j = 1, \cdots, l$ 그리고 $X_1^{(j)}, \cdots, X_{N_i}^{(j)},\ i = 1, \cdots, m$ 이 주어졌을 때 $\gamma(X_m^{(j)})$의 베이즈 추정량의 $100(1-\alpha)\%$ *HPD* 구간은

$$e_{Bm}^{(j)} \pm v_{Bm}^{(j)} z_{\alpha/2} \tag{3.10}$$

이다.

$e_{sm}^{(j)} = \overline{X}_m^{(j)}$의 $100(1-\alpha)\%$ 신뢰 구간은 (2.8)과 유사하게 구할수 있다.

e_{EBm}을 중심으로 한 $100(1-\alpha)\%$ *HPD* 구간은

$$e_{EBm}^{(j)} \pm v_{EBm}^{2(j)} z_{\alpha/2} \tag{3.11}$$

이다. 단, $v_{EBm}^{2(j)} = \hat{v}_{um}^{2(j)} + \hat{v}_{\theta m*}^{2(j)}$이다. 그리고

$$\hat{v}_{um}^{2(j)} = (1-f_m^{(j)})\{f_m^{(j)} + (1-f_m^{(j)})\hat{w}_m^{(j)}\}s_m^{2(j)}/n_m^{(j)}$$

$$\hat{w}_m^{(j)} = (s_m^{2(j)}/n_m^{(j)})\{\delta_j^2 + (s_m^{2(j)}/n_m^{(j)})\}$$

$$\hat{v}_{\theta m*}^{2(j)} = (1-f_m^{(j)})^2 \hat{w}_m^{(j)2}\left\{\sum_{i=1}^{m}\hat{w}_i^{(j)} n_i^{(j)}/s_i^{2(j)}\right\}^{-1}$$

이다.

4. 결 론

우리의 연구는 Nandram과 Sedransk(1993)의 결과를 층화 표본추출의 경우로 확장하였다. 또한 층화평균의 정규가정을 완화하여 경험적 베이즈 추정을 하였다. 이와 같이 Nandram과 Sedransk(1993)의 결과는 여러가지의 복합표본설계(complex sample design)에 응용될 수 있다. 층화평균에 대한 정규가정중 분산이 작을수록 경험적 베이즈 추정의 효율성이 높아진다. 이는 다른 층 또는 다른 단계에서의 자료가 많은 정보를 제공하기 때문이다. 그러나 표본의 크기가 커질수록 경험적 베이즈 추정의 효율성이 상대적으로 단순평균에 비해 떨어진다.

참 고 문 헌

[1] Ericson, W. A. (1969), "Subjective Bayesian Models in Sampling Finite Populations" (with discussion) Journal of the Royal Statistical Society, Ser. B, 31,195-233

[2] Ghosh, M., and Lahiri, P. (1987), "Robust Empirical Bayes Estimation of means From Stratified Samples." Journal of the American Statistical Association, 82, 1153-1162.

[3] Ghosh, M., and Meeden, G. (1986). "Empirical Bayes Estimation in Finite Population Sampling", Journal of the American Statistical Association, 81, 1058-1062.

[4] Nandram, B., and Sedransk, J. (1993), "Empirical Bayes estimation for the Finite Population Mean on the Current Occasion", Journal of the American Statistical Association, 88, 994-1000.