⊠연구논문

층화 표본에서 단위 무응답에 대한 가중치 조정 방법

염준근 · 손창균
동국대학교 통계학과

The Weighting Adjustment for Unit Nonresponse in the Stratified Sampling

JoonKeun Yum · ChangKyoon Son
Dept. of Statistics, Dongguk University

Abstract

In sampling survey the nonresponse reduces the precision of the estimator because of the nonresponse bias of the estimator.

Deville, et al.(1993) considered the generalized raking procedure with the auxiliary information under five distance measures for reducing the nonresponse bias of the estimator.

This paper extends the classical weighting adjustment of Deville, et al.(1993) to the stratified sampling case with three among five measures.

1. 서론

단위무응답이 발생하는 경우 가중치 조정기법의 하나인 raking비 수정은 1940년 미국의 인구센서스에서 완전한 인구수와 표본 수간의 일치성을 보장하기 위해 Deming과 Stephan 에 의해 처음으로 제안되었다. Raking비 수정 과정은 2차원 분류표에서 표본 가중치를 반복적으로 사용하여 행과 열의 수정값을 이전 단계의 값으로 이용하게 된다. Oh와 Scheuren(1983)은 단순임의 추출 하에서 무응답이 존재하는 경우 추정량들의 조건부 평균과 분산에 대해 언급하였다. Brackstone과 Rao(1979)은 반복회수를 4회까지 전개하여 raking 비 추정량의 근사분산과 편의를 구하였다.

Bankier(1986)는 유한 근사분산을 순환적인 방법으로 계산하고, Causey(1972)는 최소 분류 정보 이론(minimum discrimination information theory)에 의거하여 수정인자

의 근사분산을 도출하였다. 또한 Causey(1983,1984)는 2차원 분류표에서 모집단의 칸 비율을 추정하기위해 표본의 주변 비율을 이용하여 여러 가지 거리측도에 근거한 모집단 칸 비율 추정량들에 대한 효율성을 비교하였다. Binder(1983)는 관심모수에 대한 추정량의 근사분산을 계산하는 일반적인 방법을 제시하였다. 그리고 Oh와 Scheuren (1987)은 기존의 비추정과 raking비 추정량을 혼합하여 사용할 수 있는 수정된 raking 비 추정량을 제안하였고, Binder와 Theberge(1988)은 raking비 추정량의 수렴된 해에 대한 점근적인 성질과 무응답이 특정한 구조를 가질 때 추정량이 불편임을 보였고, 또한 이 추정량의 점근분산을 제시하였다.

최근에 raking비 추정의 일반적인 형태로서 Deville과 Särndal(1992)는 Horvitz-Thompson 추정량에 기초하여 보조 변수를 이용한 유한 모집단의 총합을 추정함에 있어서 주변 총합이나 칸 수에 대한 보정 추정(calibration estimation) 과정을 적용하여, 모집단 총합에 대한 보정 추정량과 분산을 제안하였고, Deville외(1993)은 이들의 연구를 더욱 일반적인 경우로 확장하였다.

따라서 본 논문에서는 모집단 총합을 추정함에 있어서 Deville 등(1993)의 이론을 보다 실제적인 경우로 확장하여, 단위 무응답이 발생한 경우 보조정보를 이용하여 무응답으로 인한 편의를 줄이고, 정도를 높힐 수 있는 방법으로서 층화 추출의 경우 최소제곱 거리 측도, 최소 분류 거리 측도 그리고 최소 엔트로피 측도등 3가지 측도를 이용하여 각 층의 모집단 총합과 그에 따르는 분산을 비교하였고, 이 측도들 간의 추정량의 정확도를 비교하여 보았다.

2. 층화 표본

확률추출의 이론은 많은 조사통계의 분야에 기여한 것이 사실이다. 그러나 단순임의 추출하에서 통제할 수 없는 여러 가지 사실들이 있는데, 이를 보완하기 위해 층화추출이나 집락추출을 이용한다. 특히 층화추출은 모집단의 추정값의 정도를 높히고, 따라서 모집단의 표준오차를 크게 감소시킬 수 있는 장점을 지니고 있다. 또한 집락추출의 경우에는 단순임의추출 보다 표본의 정도가 낮아지는 문제가 있기는 하지만, 집락을 추출단위로 하기 때문에 추출작업이 편리하고, 단순임의 추출보다 비용이 적게든다는 장점을 지니고 있다.

따라서 본문에서는 실제 조사에서 많이 사용하는 층화추출의 경우, 일반화 raking 과정을 이용하여 관심변수 Y와 연관된 보조변수 $\boldsymbol{X}$를 층화변수로 이용하여 모집단을 층화한 다음, 각각의 모집단의 각 층으로부터 확률비례 추출에 의해 표본을 추출한 후, 추출된 표본을 조사하는 과정에서 발생하는 조사단위의 무응답의 가중치를 수정하는 과정을 살펴보았다. 이 장의 1절에서 층화 추출의 경우 일반화 raking 추정과정을 이용하여 일반화 회귀 추정량을 도출하였고, 2절에서는 보조변수 벡터를 이용하여 기지의 칸수와 주변수에 대해 보정 추정과정을 전개하였고, 3절에서는 모집단 총합에 대한 추정량의 분산을 구하였다.

2.1 회귀 추정량의 도출

유한모집단 $U=\{1,2,\cdots,N\}$을 고려하고, 이들이 각각 L개의 층으로 구성되고 층의 크기가 N_h인 h층의 모집단에서 n_h개의 표본을 확률비례 비복원으로 추출하였다고 하자. 층화 확률비례 비복원추출을 실시하는 경우 Horvitz-Thompson 추정량을 사용하며, 이때 원래의 추출가중치 d_{hk}와의 차이를 최소로 하는 새로운 가중치를 구하기 위해 보조변수 $\boldsymbol{X}$에 대한 보정방정식이 필요하다. 포함 확률 $\pi_{hk}=\Pr(k\in s_h)$와 $\pi_{hkl}=\Pr(k\cap l\in s_h)$인 추출설계에 의해 추출한다고 하자. y_{hk}를 모집단 단위 중 h층의 k번째 관심변수의 값이라 하고, 이와 관련하여 $j=1,\cdots,J$개의 보조변수 벡터 $\boldsymbol{x}_{hk}=(x_{hk1},x_{hk2},\cdots,x_{hkj},\cdots,x_{hkJ})'$를 고려하자. 여기서, x_{hkj}는 h층의 k번째 단위의 j번째 보조 변수를 의미한다.

그러면 모집단에 있는 단위 k가 h층의 표본에 포함되는 경우에 $k\in s_h(\subseteq U_h)$인 단위에 대해 $(y_{hk},\ \boldsymbol{x}_{hk})$를 관찰하게 되며, 보조변수 $\boldsymbol{X}$는 h층에 대한 모집단 총합 $X_h=\sum_{U_h}\boldsymbol{x}_{hk}$으로 정확히 알고 있다고 가정하자. 여기서 s_h는 h층에 대해 모집단 U_h로 부터 추출한 표본을 의미한다. 목적은 관심변수 Y의 h층에 대한 모집단 총합 $Y_h=\sum_{k\in U_h}y_{hk}=\sum_{U_h}y_{hk}$을 추정하는데 있다. 보조변수 $\boldsymbol{X}$에 대해 층별 총합에 대한 추정량은 다음과 같다.

$$\hat{Y}_h=\sum^{n_h}d_{hk}y_{hk}=\hat{Y}_{HT},\quad \hat{X}_h=\sum^{n_h}d_{hk}\,\boldsymbol{x}_{hk}=\hat{X}_{HT} \tag{2.1.1}$$

여기서 $d_{hk}=\pi_{hk}^{-1}$로서 추출가중치 이며, h층에 대한 총합은 Horvitz-Thompson추정량으로 표현된다.

각각의 h층에 대한 표본 단위들의 새로운 추출 가중치 w_{hk}를 구하기 위해, 보조변수 $\boldsymbol{X}$를 이용한 보정방정식은 다음과 같이 형태가 될 것이다.

$$\sum^{n_h}w_{hk}\,\boldsymbol{x}_{hk}=\sum^{N_h}\boldsymbol{x}_{hk}=X_h \tag{2.1.2}$$

Deville과 Särndal(1992)이 언급한 것과 같이 최적화 문제는 다음과 같다.

$$Min_{\,w_{hk}}\ \{\sum d_{hk}G(w_{hk}/d_{hk})-\ \boldsymbol{\lambda}_h'(\sum_{s_h}w_{hk}\,\boldsymbol{x}_{hk}-\sum_{U_h}\boldsymbol{x}_{hk})\} \tag{2.1.3}$$

여기서, 거리측도 $G(\cdot)$는 $G(w_{hk}/d_{hk})=\sum(w_{hk}-d_{hk})^2/2d_{hk}$로서 최소 제곱 거리측도를 의미하며, $\boldsymbol{\lambda}_h=(\lambda_{h1},\cdots,\lambda_{hj},\cdots,\lambda_{hJ})'$는 h층에 대해 $J\times1$벡터인 라그란쥐 승수 이다.

따라서 식(2.1.3)에서 정의된 거리 측도를 최소로하는 w_{hk}는 라그란쥐 승수법을 이용하여 다음과 같은 결과를 얻게 된다.

$$\begin{aligned} w_{hk} &= d_{hk}F(q_{hk}\ \boldsymbol{x}_{hk}'\ \boldsymbol{\lambda}_h) \\ &= d_{hk}(1+q_{hk}\ \boldsymbol{x}_{hk}'\ \boldsymbol{\lambda}_h) \end{aligned} \tag{2.1.4}$$

여기서, 보조변수 $\boldsymbol{X}$의 h층에 대한 모집단 총합에 대한 추정량 $\hat{X}_h = \sum^{n_h} d_{hk}x_{hk}$ $= \hat{X}_{HT}$으로, h층에 대한 라그란쥐 승수 $\boldsymbol{\lambda}_h$ 는 다음과 같이 결정된다.

$$\boldsymbol{\lambda}_h = \hat{\boldsymbol{T}}_h^{-1}(X_h - \hat{X}_h) \tag{2.1.5}$$

식(2.1.4)로부터 q_{hk}는 표본 h층의 k번째 단위에 대한 가중치로서 거리측도를 일반적으로 표현하기 위해 사용된 기지의 가중치이며, 전개를 간단히 하기 위해 $q_{hk}=1$인 균일 가중치로 언급하였다. 이는 거리 함수의 1차 미분에 대한 역함수로서 $q_{hk}=F_{hk}(0)'$ 인 성질을 가진다.

층화추출에 있어서 분리 회귀추정량은 각 층별 회귀계수 $\boldsymbol{B}_h$의 추정량 $\hat{\boldsymbol{B}}_h$를 구하여 다음과 같이 층별총합을 얻게 된다.

$$\hat{Y}_{wh} = \sum^{n_h} w_{hk}y_{hk} = \hat{Y}_h + (X_h - \hat{X}_h)'\ \hat{\boldsymbol{B}}_h = \hat{Y}_{regh} \tag{2.1.6}$$

여기서, $\hat{\boldsymbol{B}}_h = \hat{\boldsymbol{T}}_h^{-1}\sum^{n_h} d_{hk}q_{hk}\ \boldsymbol{x}_{hk}y_{hk} = \hat{\boldsymbol{T}}_h^{-1}\ \hat{\boldsymbol{t}}_h$ 이다.

[정리 1] 다음의 $\hat{\boldsymbol{B}}_h$를 이용하여 h층에 대한 보정추정량의 식(2.1.6)은 근사적으로 불편추정량이다.

$$\hat{\boldsymbol{B}}_h = \hat{\boldsymbol{T}}_h^{-1}\ \hat{\boldsymbol{t}}_h \tag{2.1.7}$$

증 명) 식(2.1.6)으로부터 h층에 대한 보정추정량은 다음과 같이 표현할 수 있다.

$$\begin{aligned} \hat{Y}_{wh} = \sum w_{hk}y_{hk} &= \hat{Y}_h + (X_h - \hat{X}_h)'\ \hat{\boldsymbol{B}}_h \\ &= \hat{Y}_h + (X_h - \hat{X}_h)'\ \hat{\boldsymbol{T}}_h^{-1}\ \hat{\boldsymbol{t}}_h \\ &= f(\hat{Y}_h,\ \hat{X}_h,\ \hat{\boldsymbol{T}}_h,\ \hat{\boldsymbol{t}}_h) \end{aligned}$$

보정추정량 $\hat{Y}_{wh}$은 함수 $f(\hat{Y}_h, \hat{X}_h, \hat{\boldsymbol{T}}_h, \hat{\boldsymbol{t}}_h)$에 대해 비선형이므로, 이 식을 $\hat{Y}_h$, $\hat{X}_h$, $\hat{\boldsymbol{T}}_h$, $\hat{\boldsymbol{t}}_h$에 대해 각각 편미분하면 다음과 같다.

여기서 $\hat{\boldsymbol{B}}_h = \hat{\boldsymbol{T}}_h^{-1}\hat{\boldsymbol{t}}_h$으로 표현하면, $\hat{\boldsymbol{T}}_h = \sum_{s_h} d_{hk}q_{hk}\boldsymbol{x}_{hk}\boldsymbol{x}_{hk}'$이고, 이것은 $\hat{t}_{jj'h} = \sum_{s_h} d_{hk}q_{hk}x_{jhk}x_{j'hk} = \hat{t}_{j'jh}$인 원소를 가진 행렬이며, $\hat{\boldsymbol{t}}_h = \sum_{s_h} d_{hk}q_{hk}\boldsymbol{x}_{hk}y_{hk}$는 $\hat{t}_{jh0} = \sum_{s_h} d_{hk}q_{hk}x_{jhk}y_{hk}$인 원소를 가진 벡터이다.

$$\frac{\partial f}{\partial \hat{Y}_h} = 1, \quad \frac{\partial f}{\partial \hat{X}_{jh}} = -\hat{B}_{jh}, \quad j = 1, \cdots, J$$

$$\frac{\partial f}{\partial \hat{t}_{jj'h}} = (X_h - \hat{X}_h)'(-\hat{\boldsymbol{T}}_h^{-1}\Lambda_{jj'h}\hat{\boldsymbol{T}}_h^{-1})\hat{\boldsymbol{t}}_h, \quad j \le j' = 1, \cdots, J$$

$$\frac{\partial f}{\partial \hat{t}_{jh0}} = (X_h - \hat{X}_h)'\hat{\boldsymbol{T}}_h^{-1}\lambda_{jh}, \quad j = 1, \cdots, J,$$

여기서 λ_{jh}는 $J \times 1$ 벡터로서 h층에 대해 j번째 원소가 1이고 나머지는 0인 값을 가지며, $\Lambda_{jj'h}$는 $J \times J$인 행렬에서 (j, j')와 (j', j)인 위치에서 1인 값을 가지며, 그 외는 0의 값을 가진다. 이를 테일러 전개한 후 정리하면 다음과 같다.

$$\begin{aligned} \hat{Y}_{wh} \doteq \hat{Y}_{h0} &= Y_h + (\hat{Y}_h - Y_h) - \sum_{j=1}^{J} B_{hj}(\hat{X}_{jh} - X_{jh}) \\ &= \hat{Y}_h + (X_h - \hat{X}_h)'\boldsymbol{B}_h \\ &= \sum_{U_h} y_{hk}^o + \sum_{s_h} d_{hk}E_{hk} \end{aligned}$$

양변에 각각 기대값을 취하면,

$$E(\hat{Y}_{wh}) \doteq E(\hat{Y}_{h0}) = \sum_{U_h} y_{hk}^o + E(\sum_{s_h} d_{hk}E_{hk}) = Y_h$$

이고, 여기서 $y_{hk}^0 = \boldsymbol{x}_{hk}'\boldsymbol{B}_h$ 로서 모회귀 계수 $\boldsymbol{B}_h$를 이용한 h층의 k번째 관심변수에 대한 예측치이며, $E_{hk} = y_{hk} - x_{hk}'\boldsymbol{B}_h$으로 잔차이다. ■

[따름정리 1] 층별 총합에 대한 점근 분산의 식은 위의 [정리1]을 이용하면 다음과 같다.

$$AV(\hat{Y}_{wh}) = \sum\sum_{U}(\pi_{hkl} - \pi_{hk}\pi_{hl})(d_{hk}E_{hk})(d_{hl}E_{hl}) \tag{2.1.8}$$

증 명) [정리 1]을 이용하여 다음과 같이 간단히 증명할 수 있다.

$$\begin{aligned} V(\hat{Y}_{wh}) &\doteq V(\hat{Y}_{h0}) \\ &= V(\sum_{S_h} d_{hk}E_{hk}) \\ &= \sum\sum_{U_h}(\pi_{hkl} - \pi_{hk}\pi_{hl})d_{hk}d_{hl}E_{hk}E_{hl} \end{aligned}$$ ∎

위의 분산은 보정 추정량이 점근적으로 불편추정량이므로, 분산을 AV로 표현하였다.

[따름정리 2] 식(2.1.7)의 모회귀 계수 $\boldsymbol{B}_h$의 추정량 $\hat{\boldsymbol{B}}_h$는 점근분산 식(2.1.8)을 최소하는 최적(optimal)의 추정량이다.

증 명) Särndal 등(1992) 참조 ∎

따라서, h층의 모회귀 계수 $\boldsymbol{B}_h$가 미지이므로 표본으로부터 추정을 해야하며, $\boldsymbol{B}_h$의 가중 추정량 $\hat{\boldsymbol{B}}_{wh}$는 [따름정리2]로 부터 다음의 제곱식을 최소로 한다.

$$SS_{U_h} = \sum_{U_h} q_{hk}(y_{hk} - \boldsymbol{x}_{hk}'\boldsymbol{B}_h)^2 = \sum_{U_h} q_{hk}E_{hk}^2 \tag{2.1.9}$$

위의 점근 분산을 추정하기 위해 $\boldsymbol{B}_h$가 미지이므로 잔차 $E_{hk} = y_{hk} - \boldsymbol{x}_{hk}'\boldsymbol{B}_h$를 이용할 수 없고, $\boldsymbol{B}_h$의 가중추정량 $\hat{\boldsymbol{B}}_{wh}$는 식(2.1.9)의 SS_{U_h}를 미지의 모집단 총합으로 고려하여 구할 수 있으며, SS_{U_h}의 보정된 가중추정량은 $SS_{wh} = \sum_{S_h} w_{hk}q_{hk}E_{hk}^2$ 이고, 이는 표본으로부터의 정규방정식인 $(\sum_{S_h} w_{hk}q_{hk}\boldsymbol{x}_{hk}\boldsymbol{x}_{hk}')\hat{\boldsymbol{B}}_{wh} = \sum_{S_h} w_{hk}q_{hk}\boldsymbol{x}_{hk}y_{hk}$을 만족하는 추정량 $\hat{\boldsymbol{B}}_{wh}$에 의해 최소가 된다. 따라서 표본에 근거한 잔차는 $e_{hk} = y_{hk} - \boldsymbol{x}_{hk}'\hat{\boldsymbol{B}}_{wh}$에 의해 계산할 수 있으며, 분산추정량은 다음과 같다.

$$\hat{V}(\hat{Y}_{wh}) = \sum\sum_{S_h}(\pi_{hkl} - \pi_{hk}\pi_{hl})/\pi_{hkl}(w_{hk}e_{hk})(w_{hl}e_{hl}) \tag{2.1.10}$$

2.2 가중치 조정과정

Deville 등(1993)에 의해 전체 관심 모집단 중 h층의 칸 (i, j)에 대해 응답한 단위들의 칸 수 N_{hij}가 기지이고, 보조변수 $\boldsymbol{X}$를 이용하여 가중치에 대한 보정을 실시하게 되며, 이러한 경우 벡터 $\boldsymbol{x}_{hk}$는 모집단에서 h층의 단위 k가 칸에 포함됨을 나타내는 $(RC-1)$개의 0과 하나의 1로 구성된다. 따라서 보조변수 벡터 $\boldsymbol{x}_{hk}$는 h층에 대해 완전하게 관찰되게 된다. 그러면 $X_h = \sum_{U_h} \boldsymbol{x}_{hk}$는 $R \times C$개의 기지 모집단 칸 수 N_{hij}로 구성된다. 칸 (i, j)에서 h층의 모든 k에 대해 $\boldsymbol{x}_{hk}{}' \boldsymbol{\lambda}_h$는 상수이며, $\boldsymbol{x}_{hk}{}' \boldsymbol{\lambda}_h = \lambda_{hij}$라 하자. 그러면 식(2.1.4)로 부터 $F(\boldsymbol{x}_{hk}{}' \boldsymbol{\lambda}_h) = F(\lambda_{hij}) = N_{hij}/m_{hij}$이고, $m_{hij} = \sum_{s_h} 1/\pi_{hk} = \sum_{s_h} d_{hk}$이며, s_{hij}는 표본 s_h 중 h층에 있는 단위들의 칸 수를 나타낸다. 식(2.1.4)로 부터 함수 F에 관계없이 칸 (i, j)에 있는 h층의 모든 단위 k에 대해 보정된 가중치는 $w_{hk} = d_{hk} N_{hij}/m_{hij}$가 된다. 여기서 $m_h = \sum_i \sum_j m_{hij}$ 이다. 따라서, 식(2.1.4)에 의해 다음과 같은 추정량을 얻을 수 있다.

$$\hat{Y}_{ph} = \sum_i \sum_j N_{hij} \bar{y}_{s_{hij}} = N_h \bar{y}_{s_h} \tag{2.2.1}$$

여기서 $\bar{y}_{s_{hij}} = (\sum_{s_{hij}} y_{hk}/\pi_{hk})/m_{hij}$는 표본 칸에 h층에서 응답한 단위의 π^{-1}로 가중된 y의 평균이다.

다음으로 기지의 주변수에 대한 보정 과정을 살펴보면, 우선 주변수 상에서 2차원 분류표에 대해 가중치를 보정한다고 하자. h층에 대해 보조변수 벡터 $\boldsymbol{x}_{hk}$가 정의되어야 하며, 따라서 $\sum_{U_h} \boldsymbol{x}_{hk}$는 보정을 위해 이용할 정보를 요약하게 되다. 이는 모집단 $R \times C$ 표에서 주변수를 나타내며, 다음과 같이 정의할 수 있다.

$$\boldsymbol{x}_{hk} = (\delta_{h1\cdot k}, \delta_{h2\cdot k}, \cdots, \delta_{hR\cdot k}, \delta_{h\cdot 1k}, \delta_{h\cdot 2k}, \cdots, \delta_{h\cdot Ck})'$$

여기서, $\delta_{hi\cdot k} = 1$, 만일 모집단에서 h층에 있는 단위 k가 i행에 있으면,
$= 0$, 그 외,
$\delta_{h\cdot jk} = 1$, 만일 모집단에서 h층에 있는 단위 k가 j열에 있으면,
$= 0$, 그 외.

따라서 $\sum_{U_h} \boldsymbol{x}_{hk}$는 h층에 대해 다음과 같이 R개의 행과 C개의 열에 대한 단위들의 주변수들로 나타낼 수 있다.

$$\sum_{U_h} \boldsymbol{x}_{hk} = (N_{h1+}, N_{h2+}, \cdots, N_{hR+}, N_{h+1}, N_{h+2}, \cdots, N_{h+C})' \tag{2.2.2}$$

여기서, $N_{hi+} = \sum_j^C N_{hij}$, $N_{h+j} = \sum_i^R N_{hij}$ 이다.

라그란쥐 승수 $\boldsymbol{\lambda}_h = (\alpha_{h1}, \alpha_{h2}, \cdots, \alpha_{hR}, \beta_{h1}, \beta_{h2}, \cdots, \beta_{hC})'$ 라 하면, h층의 단위 k가 항상 칸 (i, j) 에 포함될 때면, $\boldsymbol{x}_{hk}' \boldsymbol{\lambda}_h = \alpha_{hi} + \beta_{hj}$가 된다. F는 앞에서 언급한 조건을 따르는 거리함수 G에 해당하는 함수라 하자.

그러면, $F(\boldsymbol{x}_{hk}' \boldsymbol{\lambda}_h) = F(\alpha_{hi} + \beta_{hj})$이다. h층에서 칸 (i, j) 에 응답한 단위들의 수 $m_{hij} = \sum_{s_{hij}} d_{hk}$로서, 보정방정식은 다음과 같다.

$$\sum_j^C m_{hij} F(\alpha_{hi} + \beta_{hj}) = N_{hi+} \tag{2.2.3}$$

$$\sum_i^R m_{hij} F(\alpha_{hi} + \beta_{hj}) = N_{h+j} \tag{2.2.4}$$

h층의 행과 열의 효과인 α_{hi}와 β_{hj}가 결정되면, 보정된 칸 수에 대한 추정량으로 다음과 같다.

$$\widetilde{m}_{hij} = m_{hij} F(\alpha_{hi} + \beta_{hj}) \tag{2.2.5}$$

여기서, $\widetilde{m}_h = \sum_i \sum_j \widetilde{m}_{hij}$ 이다.

또한 보정된 가중치는 다음과 같이 얻어질 것이다.

$$w_{hk} = d_{hk} \widetilde{m}_{hij} / m_{hij} \tag{2.2.6}$$

결국 보정된 가중치를 이용하여 구하고자 하는 주변수에 대한 보정 추정량은 다음과 같다.

$$\hat{Y}_{wh} = \sum_{s_h} w_{hk} y_{hk} = \sum_i \sum_j \widetilde{m}_{hij} \overline{y}_{s_{hij}} = \hat{Y}_{mh} \tag{2.2.7}$$

여기서, $\overline{y}_{s_{hij}} = (\sum_{s_{hij}} y_{hk} / \pi_{hk}) / m_{hij}$ 는 h층에 있는 표본단위의 π^{-1}로 가중된 y의 평균이다.

식(2.2.5)와 (2.2.6)으로부터 칸의 인자 $F(\alpha_{hi}+\beta_{hj})$는 칸 (i, j)에 있는 h층의 임의의 단위 k에 대해 무응답으로 발생하는 가중치의 변화를 측정한다.

응답자들의 주변 정보를 이용한 칸 수 추정량의 벡터를 다음과 같이 정의하자.

$$\boldsymbol{m}_h = (m_{h11}, m_{h12}, \cdots, m_{hij}, \cdots, m_{hRC})' \tag{2.2.8}$$

여기서, $m_{hij} = \sum_{S_{hij}} d_{hk}$ 이고, $m_h = \sum_i \sum_j m_{hij}$이다.

우선 주변수에 대한 추정량 $\hat{Y}_{mh}$와 사후 층화 추정량 $\hat{Y}_{ph}$를 조건부 분산과 편의를 계산하기 위해 적절한 형태로 표현하자. 이는 h층의 모집단 칸 U_{hij}에 있는 단위 k에 대해 $y_{hk} = \alpha_{hi} + \beta_{hj} + E_{hk}$인 2원 가법효과 모형을 이용하여 유한모집단을 모수화하는 것이다. 여기서 주어진 h층에 대해 α_{hi}는 행의 효과이며, β_{hj}는 열의 효과, E_{hk}는 오차항이다. 정규방정식으로 정의된 α_{hi}, β_{hj}는 고정된 미지의 값이므로 추정해야 한다. 관심 변수 Y의 h층의 k번째 단위의 값을 $y_{hk} = \boldsymbol{x}_{hk}' \boldsymbol{B}_h + E_{hk}$으로 표현할 수 있으며, 여기서 $\boldsymbol{B}_h = (\alpha_{h1}, \cdots, \alpha_{hR}, \beta_{h1}, \cdots, \beta_{hC})'$이며, $\boldsymbol{x}_{hk}$는 이미 앞에서 정의되었고, α_{hi}와 β_{hj} 를 구하기 위한 정규방정식은 다음과 같다.

$$\sum_j^C N_{hij}(\alpha_{hi} + \beta_{hj}) = N_{hi+} \bar{y}_{hi+} \tag{2.2.9}$$

$$\sum_i^R N_{hij}(\alpha_{hi} + \beta_{hj}) = N_{h+j} \bar{y}_{h+j} \tag{2.2.10}$$

$y_{hk} = \alpha_{hi} + \beta_{hj} + E_{hk}$에서 k번째 오차항을 분해하여, $E_{hk} = L_{hij} + R_{hk}$라 하면, $L_{hij} = \bar{y}_{hij} - (\alpha_{hi} + \beta_{hj})$는 가법모형의 적합결여를 나타내며, $R_{hk} = y_{hk} - \bar{y}_{hij}$는 칸 평균의 편차이고, $\bar{y}_{hij} = \sum_{U_{hij}} y_{hk}/N_{hij}$ 이다. 또한 h층의 평균은 $\sum_i \sum_j \bar{y}_{hij} = \bar{y}_h$ 가 된다. 다음과 같이 L_{hij}에 대해 조건을 부여하자.

$$\sum^R N_{hij} L_{hij} = \sum^C N_{hij} L_{hij} = 0 \tag{2.2.11}$$

관찰치 y_{hk}는 모든 $k \in U_{hij}$ 에 대해, 결국 다음과 같이 가법모형 예측과 가법모형의 적합결여, 그리고 칸 평균의 편차항의 결합으로 표현된다.

$$y_{hk} = (\alpha_{hi} + \beta_{hj}) + L_{hij} + R_{hk} \tag{2.2.12}$$

식(2.2.7)로 부터 다음을 얻게 된다.

$$\hat{Y}_{mh} = \sum\sum \tilde{m}_{hij}(\alpha_{hi} + \beta_{hj} + L_{hij} + \overline{R}_{s_{hij}}) \tag{2.2.13}$$

여기서 $\overline{R}_{s_{hij}} = \sum_{S_{hij}} d_{hk} R_{hk} / m_{hij}$ 이다.

h층의 총합 Y_h에 대해 대응된 수식이 필요하기 때문에 다음 식을 이용하자.

$$\sum\sum \tilde{m}_{hij}\alpha_{hi} = \sum^{R} N_{hi+}\alpha_{hi} \tag{2.2.14}$$

$$\sum\sum \tilde{m}_{hij}\beta_{hj} = \sum^{C} N_{h+j}\beta_{hj} \tag{2.2.15}$$

위의 방정식과 식(2.2.11)을 이용하여 다음을 구할 수 있다.

$$Y_h = \sum_{U_h} y_{hk} = \sum_i\sum_j N_{hij}(\alpha_{hi} + \beta_{hj}) = \sum_i\sum_j \tilde{m}_{hij}(\alpha_{hi} + \beta_{hj}) \tag{2.2.16}$$

식(2.2.12), (2.2.14), (2.2.15), (2.2.16)을 이용하여 $\hat{Y}_{mh}$의 오차를 다음과 같이 표현할 수 있다.

$$\hat{Y}_{mh} - Y_h = \sum_i\sum_j(\tilde{m}_{hij} - N_{hij})L_{hij} + \sum_i\sum_j \tilde{m}_{hij}\overline{R}_{s_{hij}} \tag{2.2.17}$$

사후 층화 추정량(2.2.1)의 오차에 대한 수식은 다음과 같이 간단히 도출할 수 있다.

$$\hat{Y}_{ph} - Y_h = \sum_i\sum_j N_{hij}\overline{R}_{s_{hij}} \tag{2.2.18}$$

2.3 분산 추정량

전통적인 raking비 추정량의 분산은 점근적인 경우라 하더라도 구하기가 그다지 쉽지 않다. 단순임의 비복원 추출하에서 Brackstone과 Rao(1979), Konijn(1981), Choudhrij와 Lee(1987)은 반복횟 수가 1회 또는 아주 적은 수의 반복 후에 그 과정이 수렴한다는 가정 하에서 반복 분산식을 도출하였다. Binder와 Theberge(1988)에 의해 분산추정량이 제안되었지만 수식이 복잡하다는 단점이 있다. 여기서는 층화 확률비례 비복원 추출인 경우 일반적인 분산추정량의 식(2.1.10)을 이용하고, 기지의 주변수에 대한 추정량 $\hat{Y}_{mh}$가 h층의 회귀 추정량인 $\hat{Y}_{regh}$에 근사하다는 조건을 만족하는 거리측도 G에 대한 성질을 이용하여 $\hat{Y}_{mh}$의 분산추정량을 도출하였다. $\hat{Y}_{mh}$의 대표본

분산은 오차항을 $E_{hk}=y_{hk}-(\alpha_{hi}+\beta_{hj})$라 놓고 식(2.1.9)에 의해 제시된다. 분산추정량은 식(2.1.10)로 부터 다음의 표본에 근거한 잔차식을 대입하여 얻을 수 있다.

$$e_{hk}=y_{hk}-(\hat{\alpha}_{hi}+\hat{\beta}_{hj})=\hat{L}_{hij}+\hat{R}_{hk} \tag{2.3.1}$$

여기서 $\hat{R}_{hk}=y_{hk}-\bar{y}_{s_{hij}}$이고, $\hat{L}_{hij}=\bar{y}_{s_{hij}}-(\hat{\alpha}_{hi}+\hat{\beta}_{hj})$이다.

$\hat{\alpha}_{hi}$, $\hat{\beta}_{hj}$는 식(2.2.9)와 식(2.2.10)에 상대적으로 표본에 근거한 정규방정식인 다음의 해이다.

$$\sum^{C}\tilde{m}_{hij}(\hat{\alpha}_{hi}+\hat{\beta}_{hj})=\sum_{S_{hi}}w_{hk}y_{hk}=\sum^{C}\tilde{m}_{hij}\bar{y}_{s_{hij}} \tag{2.3.2}$$

$$\sum^{R}\tilde{m}_{hij}(\hat{\alpha}_{hi}+\hat{\beta}_{hj})=\sum_{S_{hij}}w_{hk}y_{hk}=\sum^{R}\tilde{m}_{hij}\bar{y}_{s_{hij}} \tag{2.3.3}$$

결국 구하고자 하는 분산 추정량은 다음과 같다.

$$\hat{V}(\hat{Y}_{mh})=\sum\sum_{S_h}(\pi_{hkl}-\pi_{hk}\pi_{hl})/\pi_{hkl}(w_{hk}w_{hl})(\hat{L}_{hij}+\hat{R}_{hk})(\hat{L}_{hij}+\hat{R}_{hl}) \tag{2.3.4}$$

만일 표본을 층화 임의추출 설계인 경우로 가정하면 식(2.3.4)에서 응답표본 s_{hij}에 있는 모든 k에 대해 가중치는 $w_{hk}=\tilde{m}_{hij}/n_{hij}$ 이며, 이를 식(2.3.4)에 대입하면 다음과 같다.

$$\hat{V}(\hat{Y}_{mh})=\frac{n_h}{(n_h-1)}(1-f_h)\sum_{S_h}w_{hk}^2e_{hk}^2=\hat{V}_{h1}+\hat{V}_{h2} \tag{2.3.5}$$

여기서

$$\hat{V}_{h1}=\frac{n_h}{n_h-1}(1-f_h)\sum_i\sum_j(\tilde{m}_{hij})^2(\hat{L}_{hij})^2/n_{hij},$$

$$\hat{V}_{h2}=\frac{n_h}{n_h-1}(1-f_h)\sum_i\sum_j(\tilde{m}_{hij})^2 s_{s_{hij}}^{\ 2}/n_{hij}$$

이고, $s_{s_{hij}}^{\ 2}=\sum_i\sum_j(y_{hk}-\bar{y}_{s_{hij}})^2/n_{hij}$는 h층의 분산으로서 칸 (i,j)에 있는 y의 표본 분산이다.

3. 수치적 예제

1) 최소제곱 거리 측도

우선 보정 가중치를 구하기 위해 행과 열의 효과인 α_{hi}, β_{hj}를 추정하기 해야 하며, 이를 위해 Causey(1984)가 제안한 반복적인 방법을 이용하자. 거리함수 G는 식(2.1.3)의 경우로 제한하면, 그에 따르는 함수 F는 다음과 같다.

$$F(\ \boldsymbol{x}_{hk}\ \boldsymbol{\lambda}_h) = 1 + \alpha_{hi} + \beta_{hj} \tag{3.1}$$

따라서 식(2.2.3)과 식(2.2.4)는 다음과 같이 다시 표현할 수 있다.

$$\sum_j^C m_{hij}(1 + \alpha_{hi} + \beta_{hj}) = N_{hi+} \tag{3.2}$$

$$\sum_i^R m_{hij}(1 + \alpha_{hi} + \beta_{hj}) = N_{h+j} \tag{3.3}$$

초기치로 $\beta_{hC}^{(0)} = 0$ 라 하면, 식(3.2)로부터 다음과 같은 식을 유도할 수 있다.

$$\alpha_{hi}^{(1)} = (N_{hi+} - m_{hi+} - \sum_j^C m_{hij}\beta_{hj}^{(0)})/m_{hi+} \tag{3.4}$$

식(3.4)에서 구한 행의 효과 $\alpha_{hi}^{(1)}$의 값을 이용하여, 다시 식(3.3)에 대입하면 다음과 같은 식을 얻게 된다.

$$\beta_{hj}^{(1)} = (N_{h+j} - m_{h+j} - \sum_i^R m_{hij}\alpha_{hi}^{(1)})/m_{h+j} \tag{3.5}$$

식(3.4)와 식(3.5)를 반복적으로 사용하여 $\alpha_{hi}^{(k)}$와 $\beta_{hj}^{(k)}$의 값이 안정화 될 때, 이 값을 추정값인 $\hat{\alpha}_{hi}$와 $\hat{\beta}_{hj}$으로 사용한다.

Deville 등(1993)이 제시한 보정추정량에 대한 분산 추정량을 층화 확률비례 표본에 적용하여 층화 표본에 대한 보정추정량과 그에 따르는 분산을 각각 구하여 보았다.

우선 전개를 간단히 하기 위해 층의 수는 $L=2$라 하고, 각 층의 크기는 같다고 가정하자. 또한 행과 열이 2×2인 표를 이용한다고 하자. 그러면 $h=1,2$이고 $R=C=2$로써 행과 열의 효과를 반복식을 구하기 위한 식(3.2)와 식(3.3)은 다음과 같이 표현될 것이다.

$$\sum_{j=1}^{2} m_{hij}(1+\alpha_{hi}+\beta_{hj}) = N_{hi+} \quad , \quad h=1,\ 2,\ i=1,2 \tag{3.6}$$

$$\sum_{i=1}^{2} m_{hij}(1+\alpha_{hi}+\beta_{hj}) = N_{h+j} \quad , \quad h=1,\ 2,\ j=1,2 \tag{3.7}$$

이로부터 층 $h=1$ 일 때, 초기치로 $\beta_{11}^{(0)} = \beta_{12}^{(0)} = 0$ 이라 놓고 식(3.4)와 식(3.5)을 1회 반복하면 다음과 같은 행과 열의 값을 얻게 된다.

$$\alpha_{11}^{(1)} = (N_{11+} - m_{11+} - (m_{111}\beta_{11}^{(0)} + m_{112}\beta_{12}^{(0)}))/m_{11+}, \quad i=1,\ j=1,2$$

$$\alpha_{12}^{(1)} = (N_{12+} - m_{12+} - (m_{121}\beta_{11}^{(0)} + m_{122}\beta_{12}^{(0)}))/m_{12+}, \quad i=2,\ j=1,2$$

$$\beta_{11}^{(1)} = (N_{1+1} - m_{1+1} - (m_{111}\alpha_{11}^{(1)} + m_{121}\alpha_{12}^{(1)}))/m_{1+1} \quad j=1,\ i=1,2$$

$$\beta_{12}^{(1)} = (N_{1+2} - m_{1+2} - (m_{112}\alpha_{11}^{(1)} + m_{122}\alpha_{12}^{(1)}))/m_{1+2} \quad j=2,\ i=1,2$$

같은 방법으로 $h=2$ 인 층에 대해 행과 열의 효과를 구하기 위해 초기치로 $\beta_{21}^{(0)} = \beta_{22}^{(0)} = 0$ 이라 놓고 식(3.4)와 식(3.5)를 1회 반복하여 구한다.

이러한 방법으로 행과 열의 효과를 추정한 후, 이 값을 식(2.2.5)와 식(2.2.6)에 대입하면 보정된 칸 수 추정량과 보정가중치를 구할 수 있다. 또한 이 값들을 식(2.2.7)에 대입하여 각층의 모집단 총합을 추정할 수 있으며, 모집단 개체가 표본에 포함될 확률이 기지라고 가정하여, 분산 추정량을 구할 수 있을 것이다.

2) 최소 분류 거리 측도

Deville 등(1993)에 의해 제시된 최소 분류 거리 측도에 따르는 함수 F는 다음과 같다.

$$F(\ \boldsymbol{x}_{hk}\ \boldsymbol{\lambda}_h) = \exp(\alpha_{hi}+\beta_{hj}) \tag{3.8}$$

따라서, 식(2.2.3)과 식(2.2.4)는 다음과 같이 다시 표현할 수 있다.

$$\sum_{j}^{C} m_{hij}\exp(\alpha_{hi}+\beta_{hj}) = N_{hi+} \tag{3.9}$$

$$\sum_{i}^{R} m_{hij}\exp(\alpha_{hi}+\beta_{hj}) = N_{h+j} \tag{3.10}$$

앞의 과정과 같이 초기치로 $\beta_{hC}^{(0)}=0$ 라 하면, 층 1에 대해 다음과 같은 식을 유도할 수 있다.

$$\hat{\alpha}_{11} = \log[N_{11+}/(m_{111}\exp(\hat{\beta}_{11})+m_{112}\exp(\hat{\beta}_{12}))],\ \ i=1,\ \ j=1,2$$

$$\hat{\alpha}_{12} = \log[N_{12+}/(m_{121}\exp(\hat{\beta}_{11})+m_{122}\exp(\hat{\beta}_{12}))],\ \ i=2,\ \ j=1,2$$

$$\hat{\beta}_{11} = \log[N_{1+1}/(m_{111}\exp(\hat{\alpha}_{11})+m_{121}\exp(\hat{\alpha}_{21}))],\ \ j=1,\ \ i=1,2$$

$$\hat{\beta}_{12} = \log[N_{1+2}/(m_{112}\exp(\hat{\alpha}_{11})+m_{122}\exp(\hat{\alpha}_{12}))],\ \ j=2,\ \ i=1,2$$

층 2에 대해서도 동일한 과정을 반복하여 각각의 행과 열의 효과 $\hat{\alpha}_{21}$, $\hat{\alpha}_{22}$, $\hat{\beta}_{21}$, $\hat{\beta}_{22}$ 를 추정할 수 있다.

3) 최소 엔트로피 측도

최소 엔트로피 측도에 따르는 함수 F는 다음과 같다.

$$F(\boldsymbol{x}_{hk}\ \boldsymbol{\lambda}_h) = (1-\alpha_{hi}-\beta_{hj})^{-1} \tag{3.11}$$

따라서 식(2.2.3)과 식(2.2.4)는 다음과 같이 다시 표현할 수 있다.

$$\sum_j^C m_{hij}(1-\alpha_{hi}-\beta_{hj})^{-1} = N_{hi+} \tag{3.12}$$

$$\sum_i^R m_{hij}(1-\alpha_{hi}-\beta_{hj})^{-1} = N_{h+j} \tag{3.13}$$

앞의 과정과 같이 초기치로 $\beta_{hC}^{(0)}=0$ 라 하여, 층 1에 대해 $\hat{\alpha}_{11}$, $\hat{\alpha}_{12}$ 에 대한 식을 유도할 수 있다. 각각의 추정량에 대한 식은 복잡하여 생략하기로 한다. 이를 반복적인 방법으로 $\hat{\beta}_{11}$, $\hat{\beta}_{12}$ 에 대한 식으로 풀어 각각의 값을 추정할 수 있다. 또한 층 2에 대해서도 동일한 과정을 반복하여 각각의 행과 열의 효과인 $\hat{\alpha}_{21}$, $\hat{\alpha}_{22}$, $\hat{\beta}_{21}$, $\hat{\beta}_{22}$ 를 추정할 수 있다.

다음은 Little과 Wu(1991)의 2차 국민건강 영양진단조사 자료를 인용하였으며, 이 자료를 이용하여 각 층의 미지의 모집단 칸수를 추정한 결과이다.

< 표 3.1 > 모집단 (단위:명)

층구분	소득 \ 지역	지 방	도 시	합 계
층 1	저 소 득	5,226,550	2,853,838	8,080,388
	고 소 득	12,582,716	4,659,327	17,242,043
	합 계	17,809,266	7,513,165	25,322,431
층 2	저 소 득	8,369,063	4,950,535	13,319,598
	고 소 득	8,103,178	3,899,655	12,002,833
	합 계	16,472,241	8,850,190	25,322,431

(단, 모집단의 칸수는 미지라고 가정함)

< 표 3.2 > 응답표본수 (단위:명)

층구분	소득 \ 지역	지 방	도 시	합 계
층 1	저 소 득	583	3,769	4,352
	고 소 득	2,308	1,614	3,922
	합 계	2,891	5,383	8,274
층 2	저 소 득	1,043	3,310	4,353
	고 소 득	1,849	2,072	3,921
	합 계	2,892	5,382	8,274

< 표 3.3 > 행과 열의 추정값

층 \ 추정량	$\hat{\alpha}_{11}$	$\hat{\alpha}_{12}$	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$
h=1	1.0118[1]	484.766	5772.032	1248.940
	7.4693[2]	7.8308	0.9666	-0.07136
	0.9997[3]	0.999743	0.000102	-0.000429
층 \ 추정량	$\hat{\alpha}_{21}$	$\hat{\alpha}_{22}$	$\hat{\beta}_{21}$	$\hat{\beta}_{22}$
h=2	2.4251[1]	994.67	6329.87	2025.51
	7.4760[2]	7.2396	1.31599	0.36373
	0.9997[3]	0.99963	0.000153	-0.000293

([1]최소제곱 거리 측도, [2]최소분류 거리측도, [3]최소엔트로피측도)

또한 <표 3.1>의 모집단 주변수와 <표 3.2>의 표본 칸수를 이용한 추정된 모집단의 칸 수는 각각 최소거리 측도와 최소 분류 거리 측도에 대해 다음과 같이 나타났다.

< 표 3.4.1 > 최소 제곱 거리측도를 이용한 모집단 칸 수의 추정량 (단위:명)

층 구 분	소득 \ 지역	지 방	도 시	합 계
층 1	저 소 득	3,366,268	4,714,830	8,081,098
	고 소 득	14,442,998	2,799,813	17,242,811
	합 계	17,809,266	7,514,643	25,323,909
층 2	저 소 득	6,605,622	6,713,577	13,319,199
	고 소 득	9,866,619	2,137,963	12,004,582
	합 계	16,472,241	8,851,540	25,323,781

<표 3.4.1>로 부터 표본과 모집단의 주변수를 고려하여 초기 가충치를 $d_{hk}=$ 25,322,431/8,274 로 하여, 각 층에 대해 행과 열의 추정된 효과를 이용하여 모집단 총합에 대한 분산 추정량은 층 1에 대해 $\widehat{V}(\ \widehat{Y}_{m1})=2.674E21$ 이고, 층 2에 대해 $\widehat{V}(\ \widehat{Y}_{m2})=2.1094E21$ 으로 나타났다.

< 표 3.4.2 > 최소 분류 거리측도를 이용한 모집단 칸 수의 추정량 (단위:명)

층 구 분	소득 \ 지역	지 방	도 시	합 계
층 1	저 소 득	2,687,685	6,153,490	8,841,175
	고 소 득	15,121,581	3,744,989	18,866,570
	합 계	17,809,266	9,898,479	27,707,745
층 2	저 소 득	6,864,878	8,406,455	15,271,333
	고 소 득	9,607,363	4,154,258	13,761,621
	합 계	16,472,241	12,560,713	29,032,954

<표 3.4.2>로 부터 각 층에 대해 행과 열의 추정된 효과를 이용하여 모집단 총합에 대한 분산추정량은 층 1에 대해 $\widehat{V}(\ \widehat{Y}_{m1})=2.4547E21$ 이고, 층 2에 대해 $\widehat{V}(\ \widehat{Y}_{m2})$ $=1.6717E21$ 으로 나타났다.

< 표 3.4.3 > 최소 엔트로피 측도를 이용한 모집단 칸 수의 추정량 (단위:명)

층 구 분	소득 \ 지역	지 방	도 시	합 계
층 1	저 소 득	2,920,465	5,159,842	8,080,307
	고 소 득	14,888,801	2,353,323	17,242,124
	합 계	17,809,266	7,513,165	25,322,431
층 2	저 소 득	7,633,911	5,685,608	13,319,519
	고 소 득	8,838,330	3,164,582	12,002,912
	합 계	16,472,241	8,850,190	25,322,431

<표 3.4.3>로부터 각 층에 대해 행과 열의 추정된 효과를 이용하여 모집단 총합에 대한 분산추정량은 층 1에 대해 $\hat{V}(\hat{Y}_{m1}) = 1.422E21$ 이고, 층 2에 대해 $\hat{V}(\hat{Y}_{m2}) = 1.3824E21$ 로 나타났다.

4. 결론

인구 총 조사나 사업체 총 조사와 같이 비용이 많이 소요되는 조사에서 이전의 모집단 자료를 이용하여 관심 모집단으로부터 특정한 표본추출 설계를 이용하여 얻은 정보를 토대로 모집단 총합을 추정하는 과정에 대해 살펴보았다. 특히 본 논문에서 이용한 표본추출설계는 층화 확률비례 추출인 경우에 대해 단순임의 추출의 경우 보다 실제적인 문제를 다루어 보았다.

본 연구에서 특별히 지적할 사항은 행의 효과가 감소함에 따라 열의 효과는 증가하지만, 추정된 칸의 수가 안정화 될 때 거의 변화가 없음을 알 수 있었다. 또한 각 층별 모집단 총합에 대해 추정량을 비교해 보면 최소 제곱 거리측도를 이용한 결과가 상대적으로 최소 분류 거리 측도를 이용한 결과 보다 실제 모집단 총합에 가까운 값임을 알 수 있다. 또한 최소 엔트로피 측도를 이용한 모집단 총합에 대한 추정값은 다른 측도에 비해 거의 실제 모집단 총합과 일치함을 알 수 있으며, 이때 분산 추정량 역시 최소 엔트로피 측도의 경우가 가장 작은 분산을 갖는 것으로 나타났다.

결과적으로 무응답 편의를 줄이기 위한 방법으로 표본 가중치 수정 방법을 이용하는 경우에 있어서 3가지 측도 중 분산 추정량이 가장 적은 최소 엔트로피 측도를 이용하는 것이 적합하며, 무응답이 발생한 경우 각 층의 칸수를 추정함에 있어서 추정량의 정확도 면에서도 최소 엔트로피 측도가 다른 두 가지 측도에 비해 우월한 것으로 나타났다.

앞으로 연구할 과제로는 논문에서 다루어 보았던 거리측도를 이용하여 집락 표본에 대해 적용하고 이들에 대한 효율성에 대해 연구하는 것이다.

참고문헌

[1] Bankier, M.D.(1986), "Estimators based on several stratified samples with applications to multiple frame surveys," *Journal of American Statistical Association*, 81, pp. 1074-1079.

[2] Binder, D.A.(1983), "On the variances of asymptotically normal estimators from complex surveys," *International Statistical Review*, 51, pp. 279-292.

[3] Binder, D.A. and Theberge, A.(1988), "Estimating the variance of raking-ratio estimators," *The Canadian Journal of Statistics*, 16(supplement), pp. 45-55.

[4] Brackstone, D.A. and Rao, J.N.K.(1979), "An investigation of raking ratio estimators," *Sankhy*ä., Ser. C, 41, Part 2, pp. 97-114.

[5] Causey, B.D.(1972), "Sensitivity of Raked contingency table totals to changes in problem conditions," *Annals of Mathematical Statistics,* 43, pp. 656-658.

[6] Causey, B.D.(1983), "Estimation of Proportions for Multinomial Contingency Tables Subject to Marginal Constraints," *Communication in Statistics Theory and Methodology*, 13, pp. 2581-2587.

[7] Causey, B.D.(1984), "Estimation under Generalized Sampling of Cell Proportions for Contingency Tables Subject to Marginal Constraints," *Communication in Statistics Theory and Methodology*, 13(20), pp. 2487-2494.

[8] Choudhry, G.H. and Lee, H.(1987), "Variance Estimation for the Canadian Labour Force Survey," *Survey Methodology,* Vol. 13, pp. 147-161.

[9] Deming, W.E. and Stephan, F.F.(1940), "On a least square adjustment of the sampled frequency table when the expected marginal totals are known," *Annals of Mathematical Statistics,* 11. pp. 427-444.

[10] Deville, J.C. and Särndal, C.E.(1992), "Calibration Estimators in Survey Sampling," *Journal of American Statistical Association*, 87, pp. 376-382.

[11] Deville, J.C., Särndal, C., and Sautory, O.(1993), "Generalized Raking Procedures in Survey Sampling," *Journal of American Statistical Association*, 88, pp. 1013-1020.

[12] Konjin, H.S.(1981), "Biases, Variances, and Covariances of Raking Ratio Estimators for Marginal and Cell Totals and Averages of Observed Characteristics," Metrika, 28, pp. 109-121.

[13] Little, R.J.A. and Wu, M.(1991), "Models for Contingency Tables With Known Margins When Target and Sampled Populations Differ," *Journal of American Statistical Association*, 86, pp. 87-95.

[14] Oh and Scheuren(1983), "Weighting adjustment for unit nonresponse," In *Incomplete Data in Sample Survey Vol. 2, Theory and Bibliographies*, (Eds.W.G. Madow, I.Olkin and D.B. Rubin), New York: Academic Press pp. 143-184.

[15] Oh and Scheuren(1987), "Modified Raking Ratio Estimation," *Survey Methodology, Vol 13*, pp. 209-219.

[16] Särndal, C., Swensson, B., and Wretman, J.H.(1992), *Model Assisted Survey Sampling*, New York ,Springer-Verlag.