연구논문

층화 다단계 샘플링에서 설계 기반 분산추정*

Design-based Variance Estimation under Stratified Multi-stage Sampling

김 규 성**
Kim, Kyuseong

층화 다단계 샘플링에서 모총계 추정을 위하여 동질선형추정량을 고려하고, 이 추정량의 설계 기반 분산추정법을 고찰하였다. 한 방법은 분산을 일단계 분산과 이단계 분산으로 구분하여 각 층에서 각각을 비편향 추정하는 방법이고, 또 다른 방법은 이단계 표본에서 선정한 부차표본을 이용하여 일단계 분산만을 추정하여 전체분산을 비편향 추정하는 방법이다. 전자는 이단계 분산이 추정 가능할 때 이용하기 좋으며 후자는 이단계 분산을 추정할 수 없을 때 용이하게 쓸 수 있다. 각각의 추정법에 대하여 동질선형추정량에 대한 비음 비편향 분산 추정량의 형태를 제안하였다. 향 후 실제 조사에서 본 논문에서 제안한 분산추정법이 효과적으로 사용될 수 있기를 기대한다.

We investigate design-based variance estimation methods of homogeneous linear estimator for population total under stratified multi-stage sampling. One method is unbiasedly estimating the first stage variance and the second stage variance separately in each stratum. And another is sub-sampling method that estimating the first stage variance only by using sub-sample selected from the second stage sample so that resulting estimator is unbiased for the total variance. The first is useful when the second stage unbiased estimator is available and the second is when the second stage variance is not estimable. For each case, we proposed a form of non-negative unbiased variance estimator. We expect the proposed variance estimation methods can be effectively used for many practical surveys.

* 이 논문은 1999년도 서울시립대학교 학술연구조성비에 의하여 연구되었음.

** 서울시립대학교 컴퓨터 · 통계학과(E-mail: kskim@uoscc.uos.ac.kr)

Ⅰ. 서 론

사회 · 경제 조사의 결과로 생산되는 모집단 특성치의 추정치, 예컨대 모평균, 모총계, 모비율, 상관계수 추정치 등에 대한 신뢰도를 표현하는 측도로 통상적으로 추정치의 표준오차, 변이계수 혹은 신뢰구간 등이 종종 사용된다. 그런데 이러한 측도는 모두 추정치의 분산을 통해서 구현되기 때문에 이러한 측도를 실제 조사에서 이용하기 위해서는 조사 데이터를 이용한 분산의 추정이 먼저 이루어져야 한다. 유한 모집단을 대상으로 하는 표본이론에서는 모수 특성치에 대한 추론으로 설계기반 추론(design based inference)을 주로 하여왔다. 설계기반 추론은 표본이 추출된 설계에 기초한 추론을 말하며, 표본추출 방법과 이에 연관된 추정량 선택이 중요한 결정사항이다. 그리고 선택된 설계 및 추정량의 성과는 설계 기반 분산을 구해서 평가하게 된다.

통상적으로 사회 · 경제 조사에서는 층화, 집락화, 다단계 추출 등의 절차를 거친 복합 표본(complex sample)을 이용하게 된다. 그리고 모수 추정량은 이러한 복합 표본에서 조사된 관측값으로 만들어진다. 특히 전국 혹은 일부 지역을 대상으로 하는 대규모 조사에서는 지역을 기준으로 하는 층화(stratification)를 하고, 세부 지역 혹은 조사 특성에 맞는 조사구를 형성하여 집락화(clustering)를 하며 각 단계별로 독립적으로 표본 추출을 실시하는 경우가 흔하다. 이때 층화와 집락화는 여러 단계에 걸쳐 실시될 수 있다. 예로써 우리나라의 농가의 소득, 지출, 부채 등 농촌 경제의 지표를 생산하는 농가경제조사의 경우 부락을 집락으로 하고 영농형태와 지대를 층으로 하는 층화를 하였다. 그리고 각 층에서 부락을 부락크기에 비례하는 확률비례추출을 하였으며, 추출된 부락에서 농가는 소득순으로 정렬한 후 계통추출을 하였다(Kim, 1998). 농가경제조사의 설계는 층화 다단계 설계의 전형적인 예이다. 또 다른 예로서 미국의 노동력 인구 조사(Current population survey, CPS)에서는 각 주를 설계단위로 하여 도시와 카운티를 일차 추출단위로 삼아 집락화를 한 후, 노동력 지표가 비슷한 집락끼리 묶는 층화를 하였다. 그리고 센서스 블록을 형성하여 2단계 집락화를 하였다. 표본추출은 일차 집락은 크기 비례 확률추

출을 하였으며, 2차 집락은 여러 가지 사회 · 경제 속성을 고려하여 집락을 정렬한 후 계통추출을 하였다(Census Bureau, 1997). CPS 설계 역시 층화 다단계 샘플링의 전형적인 예이다.

조사 결과의 평가를 위해서는 모수 특성치의 추정치에 대한 분산을 추정하는 과정이 뒤따라야 한다. 그런데 설계 기반 추론에서는 표본추출방법과 추정량에 따라 분산 추정법도 다르게 나타난다. 이러한 상황에서 분산추정법에 대한 탐구는 두 가지 방향으로 진행될 수 있는데, 하나는 개별 설계를 중요시하여 개개의 설계에 충실한 추정법을 찾는 것이며 다른 하나는 여러 가지 설계를 하나의 집합으로 묶은 후 그 집합에 공통적으로 적용할 수 있는 추정법을 찾는 것이다. 전자의 경우는 개별 설계에 충실한 최적의 추정법을 구현할 수 있는 반면 설계에 따라 결과가 달라지기 때문에 그 종류가 많아지는 단점이 있다. 후자의 경우는 여러 설계에 동시에 적용할 수 있는 방법이므로 구체적인 공식보다는 일반적인 공식이 얻어지며 개별설계에 대해서는 효과적이지 않을 수도 있다. 그러나 컴퓨터 프로그램을 통하여 결과를 생산하는 경우 계산 알고리즘을 일반화하여 입력하면 여러 설계에 대하여 동시에 이용이 가능하므로 실제적인 유용성이 전자보다 후자가 더 크다고 할 수 있다. 특히 많은 통계를 생산하는 국가 기관이나 조사 전문회사에서는 개별 조사마다 분산 추정법을 구하여 계산하는 것 보다는 일반적인 계산 알고리즘에 의하여 결과를 생산하는 것이 더 효과적일 수 있다.

표본 조사 이론에 등장하는 대부분의 모총계 추정량은 아래의 식으로 표현되는 동질선형추정량(homogeneous linear estimator)의 일종이다.

$$t = \sum_{i \in s} w_i(s)\, y_i \tag{1.1}$$

여기에서 관측값 y_i에 부여되는 계수 $w_i(s)$는 표본 s에 영향을 받을 수 있으며 또한 조사 단위 i에도 영향을 받을 수 있다.

예를 들어, 조사 단위에만 영향을 받는 경우 계수의 형태는 $w_i(s) = w_i$가 되며 이러한 계수를 취하는 모평균의 비편향 추정량은 Horvitz-Thompson (HT) 추정량으로 표현될 수 있다.

$$t_1 = \sum_{i \in s} \frac{y_i}{\pi_i} \tag{1.2}$$

여기서 π_i 는 i번째 조사단위가 표본에 포함될 포함확률이다. HT 추정량에 연관되는 표본추출법은 포함확률비례 추출법으로서, 이 설계에서는 조사단위가 표본에 포함될 확률이 단위의 초기 추출확률과 비례하게 된다. 두 번째 예로서 계수가 표본에만 영향을 받는 경우로는, 즉 $w_i(s) = w(s)$, 비추정량을 고려할 수 있다.

$$t_2 = \left(\frac{\bar{y}_s}{\bar{x}_s} \right) X \tag{1.3}$$

여기서 $\bar{y}_s$ 와 $\bar{x}_s$는 표본 s에서 계산되는 표본 평균들이고 X는 보조 변수 x의 모총계이다. 비 추정량과 연관이 깊은 표본 추출법으로는 단순임의추출과 누적크기 비례 추출법을 생각할 수 있다. 단순임의표본을 이용하면 잘 알려진 바와 같이 비추정량은 편향 추정량이 되며, 점근적으로 표본의 수가 클 때 비편향성을 부여받을 수 있다. 반면에 누적크기 비례 표본을 이용하면 비추정량은 비편향 추정량이 되어 편향이 사라진다.

세 번째 예로서 표본과 조사 단위에 동시에 의존하는 추정량으로는 Murthy의 추정량을 생각할 수 있다.

$$t_3 = \sum_{i \in s} \frac{p(s|i)}{p(s)} y_i \tag{1.4}$$

여기에서 $p(s)$는 표본 s가 추출될 확률이며, $p(s|i)$는 조사 단위 i가 추출된 조건에서 표본 s가 추출될 확률이다. Murthy의 추정량과 연관된 표본추출법은 비복원 확률비례추출법이다. 이 추출법에서는 추출단계마다 추출확률이 바뀌므로 이를 반영한 추정량 t_3는 조사 단위와 표본에 모두 의존하는 형태를 띠게 된다.

설계 기반 분산추정법 탐구에서 개별설계를 중요시하면 예로 든 세 가지 추정량에 대한 추정법을 각각 구해야 할 것이다. 그러나 세 가지 설계가 추출법은 모두 다르지만 세 설계 모두 표본 크기가 일정하다고 하면 고정 표본 크기 설계에 포함되므로 고정 표본 크기 설계에 대한 일

반적인 분산 추정법을 구현하면 위의 세 설계는 모두 동일한 계산 알고리즘에 의하여 분산 추정을 실시할 수 있다.

본 논문에서는 후자의 방법을 취하여 층화 다단계 샘플링에서 분산추정법을 고찰한다. 모총계 추정량은 개별 설계의 추정량을 대부분 포함하는 일반적인 동질선형추정량을 층화 다단계 설계에서 고려하며, 이 추정량에 대한 분산 추정법을 알아본다. 제 2장에서는 다단계 샘플링에서 동질선형추정량에 대한 기존의 분산 추정법에 대한 연구 결과를 정리하고 제 3장에서는 층화 다단계 샘플링에서 동질선형 추정량에 대한 분산 추정법을 고찰한다. 마지막으로 제 4장에서는 설계 기반 분산 추정법에 대한 토의와 더불어 향후 연구 과제를 언급한다.

Ⅱ. 다단계 샘플링에서 분산추정

다단계 샘플링에서 동질선형추정량의 분산은 1단계의 분산과 하위 단계의 분산으로 구분되어 표현된다. 따라서 다단계 샘플링에서 설계 기반 분산 추정법은 일단계의 분산과 하위단계의 분산을 분리하여 추정하는 방법이 널리 연구되었다. 이때 분산 추정량이 가져야할 속성으로는 보통 비편향성과 비음성이 강조되어 왔다.

다단계 샘플링에서 분산 추정법은 Durbin(1953)에 의하여 연구가 시작되었으며, 이때 연구 대상 추정량은 HT 추정량이었다. 동질선형추정량에 대한 분산 추정법은 Raj(1966)에 의하여 개발되기 시작하였고 이후 Rao (1975)가 좀더 일반적인 형태로 이론을 발전시켰다. 이들의 방법은 동질선형추정량의 일단계 분산과 이단계 분산을 구한 후, 각각을 비편향 추정하여 전체 분산의 비편향 추정량을 만드는 것이다. 이들의 방법은 다단계 샘플링에서 동질선형추정량의 비편향 분산추정량을 체계적으로 구하는 방법을 제공한 점이 평가할 만 하다. 반면, 이 방법은 비편향성만 고려가 되었기 때문에 분산추정량임에도 불구하고 경우에 따라서는 음수의 분산추정값을 제공하는 위험이 있다. 그리고 기본적으로 이 방법은 2단계 분산이 비편향 추정 가능할 때 이용할 수 있는 방법이기 때문에

2단계 분산을 추정하지 못하는 설계에서는 이용하지 못하는 제약이 있다.

일단계 샘플링에서 비음 비편향 분산추정법에 대한 연구는 Vijayan (1975), Rao 와 Vijayan(1977)에 의해서 체계가 잡혔으며, Vijayan 외 2인 (1995)은 일반적인 비음 정치 행렬(non-negative definite matrix)에 비음 비편향 추정법을 확장하였다. 이들의 생각은 분산추정량을 이차 형식으로 바꾸어 대칭인 형태로 표현한 후 비음 추정량의 필요조건을 구하는 것이다. 이 형태는 HT 추정량의 분산추정량으로 Yate와 Grundy(1953)이 제안한 분산추정량의 형태와 유사하게 나타난다. 이 때 분산추정량을 이차형식으로 바꾸어 대칭인 형태로 표현하기 위해서는 관측치가 취하는 값 중에 분산을 0으로 하는 관측치가 포함되어 있어야 한다는 조건이 전제되어 있다. 보통의 조사에서는 이 조건이 대부분 만족되지만 일부 조사에서 관측치의 범위가 제한이 되는 경우에는 이들의 결과는 이용이 불가능하다. 이같은 관측치의 범위 문제를 일반적으로 해결한 연구 결과는 Padmawar(1998)에 의하여 발표되었다. 일단계 샘플링에서 분산추정량의 비음성에 대한 연구 결과는 다단계 샘플링으로 확대 적용할 수 있다. 이러한 연구는 Rao(1977) 과 Chaudhuri 외 2인(2000) 등이 있다.

이제까지 설명한 방법들은 다단계 샘플링에서 일단계 분산과 하위단계의 분산을 추정 가능할 때 이용할 수 있는 방법들이다. 그런데 경우에 따라서는 하위단계의 분산 추정이 어려울 수도 있다. 예컨대, 실제 조사에서는 표본의 집중을 방지하기 위하여 보조변수를 이용하여 조사단위들을 정렬한 후 계통추출을 사용하는 경우가 많은데, 계통추출을 사용하게 되면 설계 기반 추론에서는 분산추정이 불가능하다. 즉 만일 2단계에서 계통추출로 표본을 선정하는 경우 2단계 분산 추정은 불가능해지며 따라서 위에서 설명한 방법으로는 분산추정을 구현할 수가 없다.

이러한 문제에 착안하여, 이단계 분산을 직접 구하지 않고 일단계 분산만을 추정하여 전체 분산을 추정하는 방법이 Srinath 와 Hidiroglou(1980)에 의하여 개발되었다. 이들은 동질선형추정량이 아닌 HT 추정량을 대상으로 하여 이론을 전개하였으며, 동질선형추정량에 대한 이론의 확장은 Arnab(1988)에 의해서 이루어졌다. 이들의 방법은 이단계 표본 중에서

일부 표본을 다시 추출하는 부차 표집(sub-sampling)을 하는 것인데, 이들은 다음과 같은 점에 착안하였다. 즉, 이단계에서 사용된 표본 중에서 일부 표본을 부차 추출하여 일단계 분산 추정량을 만들면 표본수의 부족으로 일단계 분산의 과대 추정이 발생한다. 그런데 만일 과대 추정되는 양만큼이 이단계 분산과 일치한다면 전체적으로 일단계 분산 추정만으로 전체 분산을 비편향 추정하는 효과를 보게 되는 것이다. 즉, 이단계 표본 중 적절한 수의 부차 표본를 선정하여 만든 일단계 분산 추정량만으로 전체 분산을 비편향 추정하는 방식이며, 당연히 적정한 부차 표본수를 결정하는 문제가 중요한 관건이 된다.

부차 표집법은 이단계 이하 단계에서 분산추정이 용이하지 않을 때 이용할 수 있는 유용한 방법이다. 그리고 분산추정량의 형태가 일단계 분산 추정량의 형태만을 포함하므로 추정 공식이 간단하여 이용이 쉽다는 장점이 있다. 반면에 부차 표본수를 정확하게 계산해야 하는 일이 추가되며 경우에 따라서는 부차 표본수가 자연수로 정확하게 떨어지지 않을 수도 있다. 자연수로 떨어지지 않으면 확률화를 이용하며 표본수를 재차 결정해야 한다. 또 다른 단점은 조사에 이용된 전체 표본이 아닌 일부 표본만을 분산추정에 이용했기 때문에 비록 분산추정량이 분산을 비편향 추정한다 하더라도 분산추정량의 안정성(stability)은 상당히 떨어질 수 밖에 없다. 전체 분산 중에서 이단계 분산의 비중이 클수록 이를 보정하기 위해서 부차 표본수는 줄어들 것이므로 분산 추정량은 더욱 불안정해지게 된다. 따라서 이 방법은 이단계 분산의 비중이 전체 분산에 비하여 크지 않을 때 이용하는 것이 효과적일 것이다.

Ⅲ. 층화 다단계 샘플링에서 분산추정

앞 절의 분산 추정법의 논리는 동일하게 층화 다단계 샘플링에 적용할 수 있다. 본 절에서는 층화 다단계 샘플링에서 모총계에 대한 동질선형 추정량의 분산추정법을 고찰해 본다.

표본 추출은 각 층별로 독립적으로 이루어지는 것으로 가정하고, 각

층에서 일단계 표본은 고정 크기 표본 설계(fixed size design)에 의하여 추출한다고 하자. 그리고 하위 단계에서 표본 추출에 대한 제약은 없으며 단지 집락의 모총계의 비편향 추정량과 비편향 분산추정량을 제공할 수 있는 표본추출법을 이용한다고 가정하자. 이러한 표본추출법을 전제로 층화 다단계 샘플링에서 모총계에 대한 동질선형추정량은 다음과 같이 정의할 수 있다.

$$t = \sum_{h=1}^{L} \sum_{i \in s_h} w_i(s_h)\, t_{hi} \tag{3.1}$$

여기서 L은 층의 수이며, t_{hi}는 h층의 i번째 집락 총계 Y_{hi}에 대한 비편향 추정량이고, $w_i(s_h)$는 조사단위 i와 일단계 표본 s_h에 의존하는 계수이다.

이단계 분산 계산법을 이용하여 t의 분산을 구하고, Vijayan 외 2인 (1995)의 결과를 이용하여 이차형식을 대칭인 형태로 변환하면 t의 분산을 다음과 같이 표현할 수 있다.

$$\begin{aligned} Var(t) &= V_1 + V_2 \\ &= \sum_{h=1}^{L} \left\{ -\frac{1}{2} \sum_{i=1}^{N_h} \sum_{j=1}^{N_h} d_{hij}\, w_{hi}\, w_{hj} \left(\frac{Y_{hi}}{w_{hi}} - \frac{Y_{hj}}{w_{hj}} \right)^2 + \sum_{i=1}^{N_h} b_{hii}\, \sigma_{hi}^2 \right\} \end{aligned} \tag{3.2}$$

여기에서 N_h는 h 층의 크기이며, w_{hi}는 Y_{hi}가 취하는 값 중에서 t의 분산을 0으로 만드는 값이다. 또한 $d_{hij} = b_{hij} - 1$, $b_{hij} = E\{w_i(s_h) w_j(s_h)\}$이다. 그리고 σ_{hi}^2는 h 층의 i번째 집락의 총계 추정치에 대한 이단계 분산이다.

위의 식 (3.2)에서 보듯이 전체 분산 $Var(t)$은 일단계 분산 V_1과 이단계 분산 V_2의 합으로 구성된다. Raj 등의 분산 추정 절차에 따라 각각을 비편향 추정하여 전체 분산 $Var(t)$의 비음 비편향 추정량의 형태를 구해본다. 만일 이단계 분산의 비편향 추정량 $\hat{\sigma}_{hi}^2$을 구할 수 있으면, 분산 $Var(t)$의 비음 비편향 추정량은 다음과 같은 형태로 표현될 수 있음을 보일 수 있다(Kim, 2001).

$$v_1(t) = \sum_{h=1}^{L} \left\{ -\frac{1}{2} \sum_{i \in s_h} \sum_{j \in s_h} c_{ij}(s_h)\, w_{hi}\, w_{hj} \left(\frac{t_{hi}}{w_{hi}} - \frac{t_{hj}}{w_{hj}} \right)^2 + \sum_{i \in s_h} e_i(s_h)\, \hat{\sigma}_{hi}^2 \right\} \tag{3.3}$$

여기서 $E\{c_{ij}(s_h)\} = d_{hij}$, $E\{e_i(s_h)\} = 1$이다.

위의 식 (3.3)에서 계수 $c_{ij}(s_h)$와 $e_i(s_h)$는 각 층에서 사용된 일단계 표본추출 방법에 따라 다른 형태를 갖게 되며, $\hat{\sigma}^2_{hi}$의 형태는 하위단계의 표본 추출방법에 따라 정해진다. 예를 들어 포함확률비례 추출을 하면 다음의 값을 구하여 이용하면 된다.

$$c_{ij}^{(1)}(s_h) = \frac{d_{hij}}{\pi_{hij}},\ e_i^{(1)}(s_h) = \frac{1}{\pi_{hi}} \tag{3.4}$$

만일 비복원 확률비례추출을 했을 때의 계수는

$$c_{ij}^{(2)}(s_h) = \frac{d_{hij}p(s_h|i,j)}{p(s_h)},\ e_i^{(2)}(s_h) = \frac{p(s_h|i)}{p(s_h)} \tag{3.5}$$

이다. 여기서 $p(s|i,j)$는 집락 (i,j)가 뽑힌 후, 표본 s가 뽑힐 확률이다. 또 다른 예로 누적 크기 확률비례추출을 하면 계수는 다음의 값을 이용할 수 있을 것이다.

$$c_{ij}^{(3)}(s_h) = \frac{d_{hij}}{M_2\, p(s_h)},\ e_i^{(3)}(s_h) = \frac{1}{M_1\, p(s_h)} \tag{3.6}$$

여기서 $M_r = \binom{N-r}{n-r}$, $r=1,2$이다. 만일 하위단계에서 분산추정이 현실적으로 어렵다면 부차 표집법을 이용한 분산추정법을 대신 이용할 수 있다 (Srinath and Hidiroglou, 1980; Arnab, 1988).

$$v_2(t) = \sum_{h=1}^{L} \left\{ -\frac{1}{2} \sum_{i\in s_h} \sum_{j\in s_h} c_{ij}(s_h)\, w_{hi}\, w_{hj} \left(\frac{t_{hi}'}{w_{hi}} - \frac{t_{hj}'}{w_{hj}} \right)^2 \right\}. \tag{3.7}$$

여기서 t_{hi}'는 부차표본으로 구성한 h층의 i번째 집락 모총계의 비편향 추정량이다. 분산 추정량 v_2가 전체 분산의 비편향 추정량이 되도록 하는 부차 표본수는 다음의 관계식으로부터 구할 수 있다(Kim, 2001).

$$Var_2(t_{hi}') = \sigma_{hi}^{2\prime} = \frac{b_{hii}}{b_{hii}-1}\sigma_{hi}^2 \tag{3.8}$$

부차 표본수는 표본 추출방법에 의존하기 때문에 일반적으로 그 비율을 말하기는 어렵다. 하나의 예로서, 부차 표본수를 구해본다. 각 층에서 일단계 추출은 포함확률 비례 추출을 하고, 이단계에서는 단순임의추출을 했다고 하자. 그러면 $b_{hii} = 1/\pi_{hi}$ 이고 $\sigma_{hi}^2 \approx \sigma^2 / m_{hi}$ 이므로 다음이 성립한다.

$$m_{hi}' = (1 - \pi_{hi})\, m_{hi} = (1 - n_h\, p_{hi})\, m_{hi} \tag{3.9}$$

여기서 m_{hi}', m_{hi}는 h층의 i번째 집락에서 이단계 부차 표본수와 원 표본수이다. 가장 간단한 경우로 일단계 추출이 단순임의추출이면, 즉 $p_{hi} = 1/N_h$, 이면 $f_h = n_h/N_h$ 라고 할 때

$$m_{hi}' = (1 - f_h)\, m_{hi} \tag{3.10}$$

가 되어 부차 표본에서 줄어드는 표본의 비율은 일단계 추출의 표본비율 f_h와 같게 된다. 즉 일단계 표본추출비율 f_h이 크면 일단계 분산은 줄어드는 대신 상대적으로 이단계 분산이 커지고, 이단계 분산을 보정하기 위해서는 더 적은 수의 이단계 부차 표본을 선정해야 하는 것이다.

Ⅳ. 토 의

본 논문에서는 층화 다단계 샘플링에서 설계 기반 접근법을 이용한 모총계 추정량의 분산 추정법을 고찰하였다. 본 논문에서 고찰한 분산 추정식 v_1과 v_2는 개별설계에 대한 식이 아닌 고정 크기 설계를 모두 포함하는 일반적인 식이므로 그 활용범위가 넓을 것으로 생각된다. 컴퓨터 프로그램을 통하여 계산 알고리즘을 구현한 후 계수 c_{ij}와 e_i만 표본추출에 맞게 계산하여 입력하면 비음 비편향 분산 추정량을 체계적으로 얻을 수 있다.

보통의 사회 · 경제 조사는 다목적 조사이므로 모총계 이외에 모비율,

상관계수, 회귀계수 등 다양한 모집단 특성치를 산출하게 된다. 그런데 앞에서 설명한 분산 추정법은 모총계에 관한 것이기 때문에 모총계 이외의 모수 추정에는 직접적으로 활용할 수는 없다. 그리고 일반적으로 다양한 특성치에 대한 설계 기반 추론은 너무 복잡해지기 때문에 모비율, 상관계수, 회귀계수 추정에 설계기반 추정을 직접 이용하는 것은 바람직하지 않다. 대신 선형화 방법이나 재표집(resampling) 방법으로서 잭나이프(Jackknife) 방법, 균등이분표본(balanced half sample) 방법, 붓스트랩(bootstrap) 방법 등을 이용하여 비모수적으로 분산을 추정하는 것이 현실적으로 간편하고 효과적이다. 아직까지 층화 다단계 샘플링에서 동질선형추정량에 대한 함수의 분산추정법으로 재표집 방법을 이용한 연구결과는 발표되지 않고 있으나 향후 이에 대한 연구는 어렵지 않게 수행될 수 있을 것으로 보인다.

표본이론에서 조사의 결과로서 추정치의 분산을 추정하는 방법은 설계 기반 접근법과 모형 기반 접근법(model-based approach) 그리고 양자를 절충한 모형 보조(model assisted approach) 방법이 있다. 모형기반 추정법은 유한모집단을 하나의 표본으로 보고, 유한모집단의 생성 이전단계의 초 모집단(super-population)을 고려하여 초 모집단에 대한 모형을 가정한 후 이 모형에 기초하여 추론을 하는 방법이다. 모형기반 추정법은 가정된 모형의 타당성에 추론의 성과가 달려있다. 즉, 가정된 모형이 잘 맞으면 추론의 효율은 높은 반면, 모형이 틀리면 추론은 오류를 범하게 된다. 이와는 반대로 설계 기반 추론에서는 모형을 가정하지 않으므로 추론의 오류를 범할 위험은 적으나 추론의 효율은 상대적으로 모형기반 추론에 비해 낮게 된다. 즉 추론의 효율성과 타당성이라는 측면에서 두 방법은 각각 장 · 단점을 가지고 있는 것이다. 모형 보조 접근법은 보조변수를 이용하여 두 방법을 절충하려는 방법론이다.

실제 조사에서는 어느 한 추정법을 고수하는 것보다는 개별 조사에 적용이 쉽고, 계산이 용이하며 설명이 잘되는 방법을 선택하여 사용하는 것이 현실적이다. 많은 연구 결과가 보여주듯이 설계 기반 추정이 잘 적용되는 조사가 있는가 하면 모형기반 추정이 더 효율적인 조사도 있다. 향 후 많은 조사에서 본 논문에서 제안한 분산 추정법이 적용되기를 기

대한다.

참고문헌

Arnab, R. 1988. "Variance estimation in multi-stage sampling." *The Australian Journal of Statistics* 30: 107-110.

Census Bureau. 1997. *CPS sample design.* CPS technical paper 63. U.S.A.

Chaudhuri, A., Adhikary, A.K. and Dihidar, S. 2000. "Mean square error estimation in multi-stage sampling." *Metrika* 52: 115-131.

Durbin, J. 1953. "Some results in sampling theory when one units are selected with unequal probabilities." *Journal of Royal Statistical Society* Series B 15: 262-269.

Kim, Kyuseong 1998. "The current status and the improvable directions of the farm household economy survey." *The Korean Journal of Applied Statistics* 11: 29-39.

Kim, Kyuseong 2001. "Non-negative unbiased MSE estimation under stratified multi-stage sampling." *Journal of Korean Statistical Society.* (submitted).

Padmawar, V.R. 1998. "On estimating nonnegative definite quadratic forms." *Metrika* 48: 231-244.

Raj, D. 1966. "Some remarks on a simple procedure of sampling without replacement." *Journal of the American Statistical Association* 61: 391-396.

Rao, J.N.K. 1975. "Unbiased variance estimation for multi-stage designs." *Sankhyā* Series C 37: 133-139.

Rao, J.N.K. and Vijayan, K. 1977. "On estimating the variance in sampling with probability proportional to aggregate size." *Journal of the American Statistical Association* 72: 579-584.

Rao, J.N.K. and Wu, C.F.J. 1988. "Resampling inference with complex survey

data." *Journal of the American Statistical Association* 83: 231-241.

Srinath, K.P. and Hidiroglou, M.A. 1980. "Estimation of variance in multi-stage sampling." *Metrika* 27: 121-125.

Vijayan, K. 1975. "On estimating the variance in unequal probability sampling." *Journal of the American Statistical Association* 70: 713-716.

Vijayan, K., Mukhopadhyay, P. and Bhattachayya, S. 1995. "On non-negative unbiased estimation of quadratic forms in finite population sampling." *The Australian Journal of Statistics* 37: 169-178.

Yates, F. and Grundy, P.M. 1953. "Selection without replacement from within strata with probability proportional to size." *Journal of Royal Statistical Society* Serise B 15: 253-261.

Zou, G. 2000. "Variance estimation for unequal probability sampling." *Metrika* 50: 71-82.