제2장

농업통계의 이중추출 추정방법 연구[1)]

진 영

제1절 머리말

1. 연구배경 및 필요성

특정 집단의 어떤 특성을 알고 싶을 때, 전체에 대한 조사를 할 수도 있지만 시간, 비용, 신속성 등을 고려하여 특정 집단을 잘 대표할 수 있는 표본을 추출하여 표본조사를 실시하는 경우가 더욱 많다.

매 5년마다 실시되는 농림어업총조사를 목표모집단으로 다양한 농어업통계 표본조사를 실시한다. 농림어업총조사 조사시기와 각종 농어업통계 표본설계 실시시기는 약 2~3년 정도 차이가 있고, 농촌환경의 급변화로 표본설계 당시 농가인 가구가 실제 조사시점에서는 농가가 아닌 일반가구로 변경되어 표본대체가 많이 이루어진다. 또한, 표본설계 당시 농가의 주요 특성인 영농형태・전겸업 등이 조사시점에서는 다른 경우가 다수 발생하여 농업통계의 단층현상이 심해질 수 있으며, 결과분석에 따른 농촌동향 분석이 실제와 다르게 파악되거나 분석되어 괴리현상이 나타날 수 있다.

그러므로, 최근 농촌 환경변화를 반영한 목표모집단을 확보하고, 각각의 농어업조사 목적에 적합하게 목표모집단을 분류하고 허용오차에 맞는 표본규모를 결정하여 대표성 높은 표본농가를 추출하는 표본설계 및 추정방법 연구가 필요하다.

---

1) 2012년 농업통계 표본개편과 관련된 「복합 이중추출방법에 따른 추정량 연구」에 대한 통합보고서에서 첫 번째로 진행되었던 연구로서 2011년 하반기(7~12월)에 연구했던 내용을 정리한 연구보고서임

2012년 농업통계의 표본설계 개편방향은 농업조사 자료를 주표본(master sample)으로 하여 4종 관련 농업조사(농가경제조사, 농축산물생산비조사, 가축동향조사, 양곡소비량조사) 표본을 이중추출방법으로 추출하여 통계조사를 실시할 예정이다.

이중추출방법은 중복추출의 한 방법으로서 모집단의 사전정보가 부족하거나 정확한 표본틀이 없을 때 실시하는 표본설계 방법이다. 2012년부터 개편될 농업통계 표본설계는 표본개편에 따른 시계열 단층을 최소화하기 위해 이중추출법(double sampling)을 적용할 계획이므로 이중추출 표본설계와 추정방법에 대해 단계별로 체계적인 정리를 겸한 연구를 수행할 필요가 있다.

2. 연구목표 및 범위

본 연구의 궁극적인 목표는 2012년부터 점차적으로 개편될 농업통계 표본설계 기본방향인 이중추출방법의 표본설계 효과를 미리 분석하는 것이다. 그중에서도 농업통계의 핵심인 농가경제조사의 이중추출방법에 대한 표본설계 효과를 연구할 계획이다.

과거에는 0·5년도 농업총조사 자료를 모집단으로 관련 농업통계의 표본설계를 하였지만, 이번에는 매년 12월에 실시하는 농업조사를 주표본(마스터모집단)으로 농가경제조사, 농축산물생산비조사, 가축동향조사, 양곡소비량조사를 위한 표본을 추출하는 이중추출 표본설계를 할 계획이다. 여기서 주표본으로 이용되는 농업조사는 2010년 농업총조사 자료를 모집단으로 2011년 하반기에 표본설계 되었으며, 2011년 12월에 통계조사를 실시할 예정이다.

이중추출방법은 통계청에서 농업통계에 처음 적용하는 표본설계 방법이므로 차근차근 단계별로 체계적인 정리를 겸한 연구를 하는 것이 중요하다고 외부 전문위원들이 권고하였다. 이런 권고사항을 고려하여 이중추출 표본설계 효과 분석 연구는 크게 3단계로 이루어질 것이다.

- 1단계: 이중추출방법의 선행연구로 기초문헌연구와 국내외 연구동향을 파악하는 것이다. 이중추출방법이란 어떤 방법인지를 구체적으로 설명하고, 주요 특성과 추정방법 등을 요약하고 정리한다. 또한, 국내・외의 주요 연구동향과 적용사례를 검토한다.

- 2단계: 농가경제조사 표본개편(안)을 참조한 층화집락-층화를 위한 복합 이중추출 추정량을 검토하여 1차적으로 제시할 것이다. 이중추출 표본설계 효과를 분석하기 위해서 농가경제조사 주요 변수들의 모집단이 필요하지만 현실적으로 구할 수 없으므로 유사한 표본모집단을 구축하여 추정량을 산출하여야 한다. 이중추출방법은 1차 표본인 농업조사 표본에서 2차 표본인 농가경제조사 표본농가를 추출하는 것으로서 2차 표본은 1차 표본의 집합표본이다. 하지만, 현재 사용가능한 농업조사와 농가경제조사 자료는 2005년 농업총조사를 모집단으로 각각 표본설계하여 중복부분이 없으므로 이용가능한 농업조사 자료를 활용하여 복합 이중추출을 적용할 농가경제조사 유사 모집단 자료를 생성해야 하며, 이에 따른 연구가 필요하다.

- 3단계: 구축된 농가경제조사 모집단을 대상으로 이중추출 모의 표본설계를 하여 2차 표본추출 시 나타날 수 있는 중점 사항들을 점검하고 표본설계 효과를 측정한다. 조건부 결합추정과 반복 분산추정 등의 효율성을 비교하여 농가경제조사 표본설계에 적합한 추정방법을 검토하고 제시할 것이다.

본 연구의 주요 목적은 농업통계 이중추출방법에 따른 표본설계 효과를 사전에 검토하는 것이다. 따라서, 이번 연구에서는 다음과 같이 이중추출방법과 최근 연구사례 등에 대해서 연구하였다. 제2절에서 많은 조사방법론에서 언급한 이중추출 표본설계에 대해 정리하여 기술하였고, 제3절에서는 이중추출관련 국내외 연구동향과 실제 적용사례를 살펴보았으며, 향후 연구방향에 대해서 검토하였다. 제4절에서는 연구결과를 요약하고, 향후 진행될 연구사항과 연구방향에 대해 다시 언급하면서 이번 연구를 마무리하였다.

본 연구의 범위와 관련해서 간단히 언급해 둘 것이 있다. 이중추출방법은 두 단계에 걸쳐 동일표본에 대해 간략조사와 심층조사를 하여 2차 표본들의 결과를 추정 시 1차의 간략조사 정보를 보조정보로 이용할 수 있어 비추정 또는 회귀추정에 매우 유익하다. 하지만, 이번 연구에서는 이중추출 표본설계의 추정에 대한 효율성에 연구포인트를 맞춰서 연구하였다. 보조정보를 이용한 추정방법 연구는 본 연구를 마친 후에 실제 조사자료를 이용하여 다른 연구과제로 연구를 진행하여야 될 것이다.

제2절 이중추출 표본설계

1. 서론

가. 개요

이중추출은 중복추출의 한 방법으로 층화추출이나 집락추출과 같이 비용을 절감하고 표본의 정도를 높이기 위한 표본추출방법이다. 이중추출(二重抽出)은 이상추출(二相抽出), 중복추출(重複抽出)이라고도 하며, 영어로는 double-sampling, double-phase sampling, two-phase sampling이라고 한다. 2차 이상의 부차표본에서 관심변수 값이 얻어지면 다상추출(多相抽出, multi-phase sampling)이라 한다.

층화를 위한 목적으로 1차 표본을 추출하기 때문에 이중추출은 초기에 층화를 위한 모집단의 사전정보가 부족하거나 정확한 표본틀이 없는 경우에 유용하다. 전통적으로 산림통계, 어종통계 등의 자원통계에 자주 사용하는 방법이다.

이중추출방법은 각각의 표본들이 정보를 공유한다는 가정 하에 두 단계에 걸쳐 표본을 추출한다. 모집단에서 1차 표본(the first-phase sample)을 추출하여 2차 표본추출을 위한 정보를 얻고, 1차 표본에서 다시 2차 부차표본(the second-phase sample)을 추출하여 관심변수의 정보를 수집한다.

이와 같이, 1차에서 보다 최근 정보를 수집할 수 있고, 2차 표본설계 시 1차에서 획득한 정보를 활용할 수 있으며, 추정 시에 1차에서 수집한 정보를 이용하여 정도 높은 추정을 할 수 있는 장점이 있다.

1차 표본 자료의 유용성을 살펴보자.

- 2차 표본의 층화를 위해 정보를 제공한다.
  (To stratify the second-phase sample)
- 각 추정치(차이추정, 비추정, 회귀추정 등)의 정도 향상에 기여한다.
  (To improve the estimation by using a difference, ratio or regression estimator)
- 무응답을 위한 부표본을 추출할 때 이용한다.
  (To draw a subsample of non-respondent units)

나. 이중추출과 이단추출의 구분

단계별 추출단위가 같은가 다른가에 따라 이중추출(二重抽出, two-phase sampling)과 이단추출(二段抽出, two-stage sampling)로 구분할 수 있다.

첫째, 이중추출은 1차 추출단위와 2차 추출단위가 같고, 이단추출은 1차 추출단위와 2차 추출단위가 다르다.

둘째, 이중추출에서는 1차 표본에 대해서도 조사를 실시하고 여기서 얻어진 정보는 층화추출을 위해 사용될 수 있고, 1차 표본 정보를 2차 표본에서 얻어지는 여러 특성치의 추정에 활용하여 추정치의 정도를 높인다. 반면에, 이단추출에서는 1차 추출단위는 2차 추출단위의 집합이다. 2차 추출단위는 1차 추출단위의 부분집합으로 1차 추출단위에 대해서 조사를 하지 않는다.

예를 들면, 가구조사의 이중추출에서 1차 추출단위(psu)는 가구가 되고, 2차 추출단위(ssu)도 가구가 된다. 또한, 행정자료나 통계조사를 통해 1차 추출단위에서 얻어진 정보를 2차 표본설계와 특성치의 추정에 활용할 수 있다. 하지만 이단추출에서는 1차 추출단위(psu)는 행정단위인 동이고, 2차 추출단위(ssu)는 그 동을 구성하는 가구로 1차 추출단위는 조사하지 않으며, 2차 표본에 대해서만 조사를 한다.

다. 이중추출의 역사

이중추출은 매우 효과적이고 효율적인 비용의 표본추출방법으로서 Neyman(1938)[2]이 가장 먼저 제안하였다. Rao(1973)[3]는 이중추출에서의 층화추출방법과 무응답 문제에 대해 연구하였고, Cochran(1977)은 이중추출의 기본 추정치에 대해 표본추출론 저서에 수록하였다. Yates(1981)는 1953년 초판 발행에서 센서스를 비롯한 실제조사에 적용할 수 있는 유용한 표본추출의 한 방법으로 이중추출법을 언급하였다.

1990년 이후 연구동향을 살펴보면, 보조자료를 이용한 비추정 또는 회귀추정 연구가 많으며, 반복분산에 대한 연구도 활발하다. 또한 실제로 조사에 적용사례도 찾아볼 수 있었다. 제2절에서 국내외 관련 선행연구에 대해 기술하였다.

라. Horvitz-Thompson 추정[4]

2차 추출에서는 최종 표본추출확률이 1차 추출의 관측값에 의존하지 않는 불변성(invariance)을 기본 전제로 하는데, 이중추출은 이러한 불변성을 가정하지 않는 보다 일반적인 추출이다. 2차 표본추출확률이 1차 표본의 관측치에 의해 결정된다는 것은 결국 2차 표본추출확률값이 고정값이 아닌 확률변수 값이 된다는 것을 의미하고, 따라서 이 경우 기존의 HT(Horvitz-Thompson) 추정이론을 적용할 수 없을 것이다.

불변성(invariance)은 2차 표본추출확률이 1차 표본에 의존하지 않는 특성인데, 이중

2) Neyman, J. (1938), "Contribution to the theory of sampling human populations," JSSA, 33, pp.101-116.
3) Rao, J. N. K. (1973), On double sampling for stratification and analytic surveys. Biometrika, 60, pp.125-133.
4) 김재광 (2008), 표본조사론, 자유아카데미

추출은 2차 표본추출확률이 1차 표본에 의해 결정되는 확률변수의 특성을 가지므로 HT 추정이론을 적용할 수 없을 것이다. 김재광(2008)은 이중추출에서의 HT추정 이론을 적용하기 위해 조건부 결합추정방법을 소개하고 있다.

1차 추출에서 크기 n인 표본(s_1)을 단순임의 추출하여 y_i 관측 후, 2차 표본(s_2) 추출에서 y_i에 비례하는 확률비례추출(PPS)을 하는 경우에 2차 표본추출확률은 (2.1)와 같고, 모집단의 추출단위 i가 1차 표본에 포함될 확률은 (2.2)가 된다.

$$\pi_{i|s1}^{(2)} = \Pr(i \in s_2 | s_1) = \frac{r y_i}{\Sigma_{k \in s_1} y_k} \tag{2.1}$$

$$\pi_i = \sum_{i \in s1} \pi_{i|s_1}^{(2)} P_1(s_1) = E_1(\pi_{i|s_1}^{(2)}) \tag{2.2}$$

(2.1)에서 어떤 1차 표본(s_1)이 추출되느냐에 따라 2차 표본추출 확률값 $\pi_{i|s_1}^{(2)}$가 달라지므로 $\pi_{i|s_1}^{(2)}$는 확률변수가 된다. 따라서, 불변성을 만족하려면 1차 추출된 표본에서 2차 표본추출하는 조건부 2차 표본추출 확률과 모집단에서 2차 표본을 추출할 확률이 동일하면 된다. 즉, $\pi_{i|s_1}^{(2)} = \pi_i^{(2)}$를 만족하면 (2.2)에서 (2.3)이 성립하고, HT 추정량 구현이 가능하게 된다.

$$\pi_i = \sum_{k \in s_1} P_1(s_1) \pi_i^{(2)} = \pi_i^{(1)} \pi_i^{(2)} \tag{2.3}$$

이를 바탕으로 이중추출 추정량을 구해보자. 이중추출의 1차 추출에서 y_i를 정보를 획득하였을 때, 모집단 총계 추정량은 (2.4)과 같고, 1차 표본에서 2차 표본을 추출하는 모집단 총계의 조건부 비편향 추정량(conditional unbiased estimator)은 (2.5)와 같다.

$$\hat{Y}_1 = \sum_{i \in s_1} \frac{y_i}{\pi_i^{(1)}} \tag{2.4}$$

$$\hat{Y}^* = \sum_{i \in s_2} \frac{y_i}{\pi_i^{(1)} \pi_{i|s_1}^{(2)}} = \sum_{i \in s_2} \frac{y_i}{\pi_i^*} \tag{2.5}$$

(2.5) 추정량은 π_i^*-추정량이 되며, 이 π_i^*-추정량은 모집단 총계에 대해 비편향추정량을 구현한다. 또한, 이중추출의 분산은 조건부 분산 (2.6)을 만족하므로 (2.7)과 같은 식이 도출된다.

$$V(\hat{Y}^*) = V[E(\hat{Y}^* | s_1)] + E[V(\hat{Y}^* | s_1)] \tag{2.6}$$

$$V(\hat{Y}^*) = V\left(\sum_{i\in s_1}\frac{y_i}{\pi_i^{(1)}}\right) + E\left(\sum_{i\in s_1}\sum_{j\in s_1}(\pi_{ij|s_1}^{(2)} - \pi_{i|s_1}^{(2)}\pi_{j|s_1}^{(2)})\frac{y_i}{\pi_i^*}\frac{y_j}{\pi_j^*}\right) \quad (2.7)$$

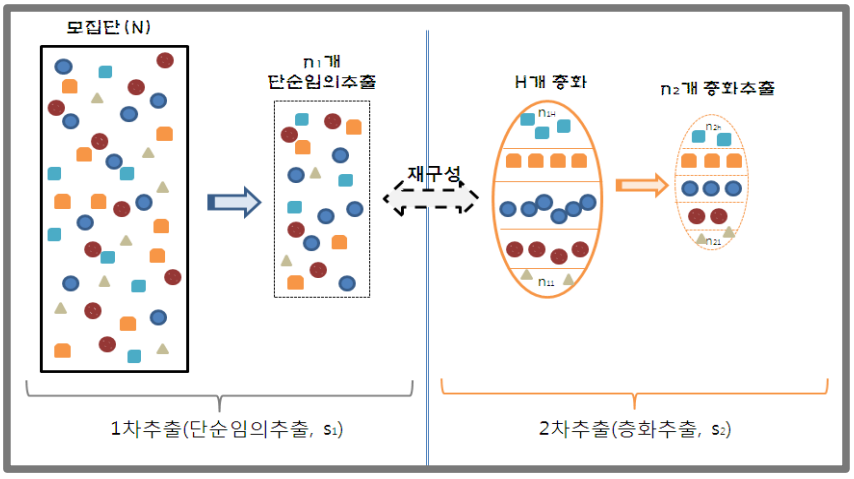
(2.7)에서 $\pi_{ij|s_1}^{(2)} = \Pr(i\in s_2, j\in s_2|s_1)$으로 1차 표본($s_1$)의 추출단위 i,j가 2차 추출에 포함될 조건부 결합표본을 포함할 확률을 의미한다. (2.7)에서 첫 번째 항은 1차 표본추출로 인한 분산항이고 두 번째 항은 2차 표본추출로 인한 분산항을 나타낸다.

2. 단순임의-층화를 위한 이중추출방법

가. 개요

층화를 위한 이중추출방법은 1차 표본을 단순임의 추출하여 정보를 수집하여 층화하고, 층화된 1차 표본에서 2차 표본을 추출하는 방법으로 가장 기본적인 이중추출방법으로 전체적인 표본설계 형태는 층화 추출형태이다.

N개의 유한모집단(U)에서 n_1개의 1차 표본(1st-phase sample, s_1)을 추출하며, s_1이 추출될 확률은 $\pi_1 = \pi_{1i} = \sum_{i\in s_1} p_1(s_1)$ 이다. 2차 표본을 추출하기 위해서 1차 표본을 H개 층화하고 $n_{1h}(\in s_{1h})$로 표시한다. 층화된 1차 표본에서 n_2개의 2차 표본(s_2)을 추출한다. h층에서 $n_{2h}(< n_{1h}, \in s_{2h})$를 만족하며, $n_2 = \sum_h^H n_{2h}$ 이다. 추출된 1차 표본에서 2차 표본이 추출될 확률은 조건부 확률인 $\pi_{n_2|n_1} = \pi_{n_{2i}|n_{1i}} = \sum_{a\in s_2} p(s_2|s_1)$이며, 각 층의 추출률은 n_{2h}/n_{1h}이고, 관심변수는 y_i이다. 단순임의-층화를 위한 이중추출 표본설계 과정은 [그림 2-1]을 참조한다.



[그림 2-1] 이중추출(단순임의-층화) 표본설계 과정

나. 이중추출 추정식

층화를 위한 이중추출에 있어서 1차 표본의 목적은 층별 가중값을 추정하는 것이고, 2차 표본의 목적은 각 층의 평균 $\overline{Y}_h$를 추정하는데 있다. 층화를 위한 이중추출에 있어서 모평균 $\overline{Y}$는 다음과 같으며 W_h는 모집단에서 h층의 상대크기를 나타낸다.

$$\overline{Y} = \sum W_h \overline{Y}_h \tag{2.8}$$

$$\begin{cases} W_h = \dfrac{N_h}{N} \\ \overline{Y}_h = \dfrac{1}{N_h}\sum_{i}^{N_h} y_i \end{cases}$$

모평균의 비편향추정량 $\overline{y}_{d.st}$는 다음과 같이 표시하며, w_h는 표본에서 h층의 상대크기이다.

$$\overline{y}_{d.st} = \sum_{h}^{H} w_h \overline{y}_{2h} = \sum_{h}^{H} \sum_{i \in s_{2h}} \frac{n_{1h}}{n_1} \frac{y_i}{n_{2h}} \qquad (w_h = \frac{n_{1h}}{n_1}) \tag{2.9}$$

$\overline{y}_{d.st}$분산은 조건부 분산을 적용하여 구할 수 있다.

$$V(\overline{y}) = V_1(\overline{y}_{d.st}) + V_2(\overline{y}_{d.st}) = V_1\{E_2(\overline{y}_{d.st})\} + E_1\{V_2(\overline{y}_{d.st})\} \tag{2.10}$$

이중추출에서 2차 표본 n_{2h}의 y_i값을 이용하여 산출하지만, h층의 1차 표본에 속하는 n_{1h}개 단위의 층별 y_i값을 구한다고 가정하면 (2.11)이 성립한다.

$$E_2(\overline{y}_{d.st}) = \sum_{h} w_h \overline{y}_{1h} = \frac{1}{n_1} \sum_{h} \sum_{i \in s_{1h}} y_i = \overline{y}_1 \tag{2.11}$$

(2.11)를 참조하여 추정량 $\overline{y}_{d.st}$는 (2.12)와 같이 도출할 수 있다.

$$\overline{y}_{d.st} = \sum w_h \overline{y}_{2h} = \sum w_h \overline{y}_{1h} + \sum w_h (\overline{y}_{2h} - \overline{y}_{1h}) \tag{2.12}$$

(2.11)는 크기 n_1인 임의표본 평균 형태이므로 분산은 (2.13)과 같다.

$$V_1\{E_2(\bar{y}_{d.st})\} = V_1(\sum_h w_h \bar{y}_{1h}) = V(\bar{y}_1) = \frac{N-n_1}{N}\frac{S^2}{n_1} = \left(\frac{1}{n_1}-\frac{1}{N}\right)S^2 \quad (2.13)$$

$$\begin{cases} \bar{y}_1 = \frac{1}{n_1}\sum_h^H \sum_{i\in s_{1h}} y_i \\ S^2 = \frac{1}{N-1}\sum_i^N (y_i - \bar{y}_N)^2 \end{cases}$$

(2.11)에 (2.12)를 대입하여 $E_1\{V_2(\bar{y}_{d.st})\}$를 구하면 (2.14)과 같다.

$$\begin{aligned} E_1\{V_2(\bar{y}_{d.st})\} &= E_1 V_2(\sum w_h(\bar{y}_{2h}-\bar{y}_{1h}) \\ &= E_1\{\sum w_h^2 V(\bar{y}_{2h}-\bar{y}_{1h})^*\} \\ &= E_1\left\{\sum w_h^2 s_{1h}^2(\frac{1}{n_{2h}}-\frac{1}{n_{1h}})\right\} \\ &= \sum(\frac{1}{n_{2h}}-\frac{1}{n_{1h}})w_h^2 s_{1h}^2 \end{aligned} \quad (2.14)$$

$$\begin{cases} s_{1h}^2 = \frac{1}{n_{1h}-1}\sum_{i\in s_{1h}}(y_i-\bar{y}_{1h})^2 \\ \bar{y}_{1h} = \frac{1}{n_{1h}}\sum_{i\in s_{1h}} y_i \end{cases}$$

$$* V(\bar{y}_{2h}-\bar{y}_{1h}) = V(\bar{y}_{2h}) - 2Cov(\bar{y}_{2h},\bar{y}_{1h}) + V(\bar{y}_{1h}) = V(\bar{y}_{2h}) - V(\bar{y}_{1h})$$

$$\left(\because Cov(\bar{y}_{2h},\bar{y}_{1h}) = E_1E_2(\bar{y}_{2h}-E\bar{y}_2)(\bar{y}_{1h}-E\bar{y}_1) = E_1(\bar{y}_{1h}-\bar{Y}_2)(\bar{y}_{1h}-\bar{Y}_2) = V(\bar{y}_{1h})\right)$$

(2.13), (2.14)를 결합하여 $\bar{y}_{d.st}$의 조건부 분산을 구하면 (2.15)과 같다.

$$V(\bar{y}_{d.st}) = \left(\frac{1}{n_1}-\frac{1}{N}\right)S^2 + \sum_h^H(\frac{1}{n_{2h}}-\frac{1}{n_{1h}})w_h^2 s_{1h}^2 \quad (2.15)$$

S^2은 (2.16)와 같으므로 (2.15)에 대입하여 (2.17), (2.18)과 같이 층간분산과 층내분산을 도출할 수 있다.

$$S^2 \doteq \sum_h^H w_h[(\bar{y}_{1h} - \bar{y}_1)^2 + s_{1h}^2] \tag{2.16}$$

$$V(\bar{y}_{d.st}) \doteq \frac{1}{n_1}(1-f_1)\sum_h^H w_h(\bar{y}_{1h} - \bar{y}_1)^2 + \sum_h^H(\frac{1}{n_{2h}} - \frac{1}{n_{1h}}f_1)w_h^2 s_{1h}^2 \tag{2.17}$$

$$= \frac{1}{n_1}(1-f_1)\, s_b^2 + \sum_h^H(\frac{1}{n_{2h}} - \frac{1}{n_{1h}}f_1)\, s_w^2 \qquad (f_1 = n_1/N) \tag{2.18}$$

(2.18)에서 이중추출은 층간분산(between stratum variance)과 층내분산(within stratum variance)로 이루어졌다는 것을 알 수 있다.

(2.17)에서 $\bar{y}_{1h}, s_{1h}^2$를 $\bar{y}_{2h}, s_{2h}^2$로 각각 대체하면 회귀분산추정치(Linearized Variance estimator) (2.19)을 구할 수 있다.

$$V(\bar{y}_d) \doteq \frac{1}{n_1}(1-f_1)\sum_h^H w_h(\bar{y}_{2h} - \bar{y}_2)^2 + \sum_h^H(\frac{1}{n_{2h}} - \frac{1}{n_{1h}}f_1)w_h^2 s_{2h}^2 \tag{2.19}$$

$$\begin{cases} \bar{y}_2 = \sum_h^H w_h \bar{y}_{2h} \\ \bar{y}_{2h} = \frac{1}{n_{2h}} \sum_{i \in s_2} y_i \\ s_{2h}^2 = \frac{1}{n_{2h}-1} \sum_{i \in s_2} (y_i - \bar{y}_{2h})^2 \end{cases}$$

3. 복합 이중추출방법

Kish(1995)는 복합이중추출(Complex Two-phase Selections)에 대해서 간략하게 소개하였다. 모집단을 층화하여 1단계 표본을 추출하고 2단계 표본을 집락추출하였을 때 또는 1단계는 집락표본을 추출하고 2단계 표본은 층화집락 추출한 경우 등을 언급하였다. 비용을 절감할 수 있지만 복합분산으로서 분산이 증가할 수 있다고 하였으며, 자동차 등록명부, 어업허가명부, 사회보험명부 등에서 복합 표본추출을 할 수 있다고 하였다.

모집단 N을 G개로 층화하여 층별로 주요 변수로 정렬하고, 두 단계 모두 층내는 random이라고 가정 시, 평균과 분산은 이중층화추출방법에 따른다고 하였다.

1단계 표본(n_L)은 모집단을 G개 층화하여 표본추출하며, 추출률은 다음과 같다.

$$[\frac{n_g}{N_g}\ ,\ (n_L = \Sigma\, n_g)\ ]$$

만약 substrata(n_{gLh})에서 정렬하여 층화한 경우(n_L), n_{gLh}은 g번째 층의 h번째 substratum이며, substrata에서 정렬하여 보조정보를 얻은 n_L의 추출률은 n_{gh}/n_{gLh}이다.

평균과 분산은 다음과 같다.

$$\sum_g^G W_g \bar{y}_g = \sum_g^G W_g \sum_h^{H_g} w_{gh}\, \bar{y}_{gh} \tag{2.20}$$

$$\sum_g^G W_g^2\ Var(\bar{y}_g) = \sum_g^G W_g^2\ [\ \sum_h^{H_g} w_{gh}^2\ \frac{s_{gh}}{n_{gh}} + \frac{1}{n_g}\sum_h^{H_g} w_{gh}(\bar{y}_{gh} - \bar{y}_g)^2] \tag{2.21}$$

또한, 1단계 집락표본으로 비용을 절감하고, 2단계 표본은 층화집락 추출한 경우에는 복합분산으로서 분산이 증가할 수 있다고 하였다.

복합 이중추출은 층화를 위한 이중추출보다 가정과 표본설계과정이 복잡하며, 분산 추정량도 단순하지 않다. 따라서, 복합 이중추출의 경우를 설명하고 간략하게 기술하였다.

가. 층화-층화 이중추출법

1차 표본과 2차 표본을 둘다 층화 추출한 경우이다.

유한모집단(U)은 H 층화되어 있으며, h층은 N_h개 추출단위로 구성되어 있다. 전체 규모 N은 층화 표본설계 형태로서 $N = \sum_h^H N_h = N_H$로 표시할 수 있다.

모집단에서 $n_1 (= \sum_h n_{1h} = n_h)$개의 1차 표본(1st-phase sample, s_1)을 층화추출한다. s_1이 추출될 확률은 $\pi_1 = \pi_{1i} = \sum_{i \in s_1} p_1(s_1)$이며, h층 표본규모는 $n_{1h} = n_{1h} (\in s_{1h})$이다.

2차 표본을 추출하기 위해서 층화추출한 1차 표본에서 주요 특성별로 G개로 층화하여 $n_{1g} = N_g (\in s_1 = s_{1g})$로 표시하고, $\sum_g^G n_{1g} (= n_1)$이 성립한다. G개 층화된 1차 표본들에서 n_2개의 2차 표본(s_2)을 층화추출한다. 2차 표본은 $n_{2g} (< n_{1g,}\ s_{2g})$를 만족하며, $n_2 = \sum_g^G n_{2g}$ 이다. 층화추출된 1차 표본에서 2차 층화표본이 추출될 확률은 조건부 확률인 $\pi_{n_2|n_1} = \pi_{n_{2i}|n_{1i}} = \sum_{i \in s_2} p(s_2|s_1)$이며, g층의 추출률은 n_{2g}/n_{1g}이고, 관심변수는 y_j이다.[5)]

5) 관련 논문은 Binder, Babyak, Brodeur, Hidiroglou, Jocelyn(2000) 등이 연구한 'Variance Estimation for Two-Phase Stratified Sampling'이다.

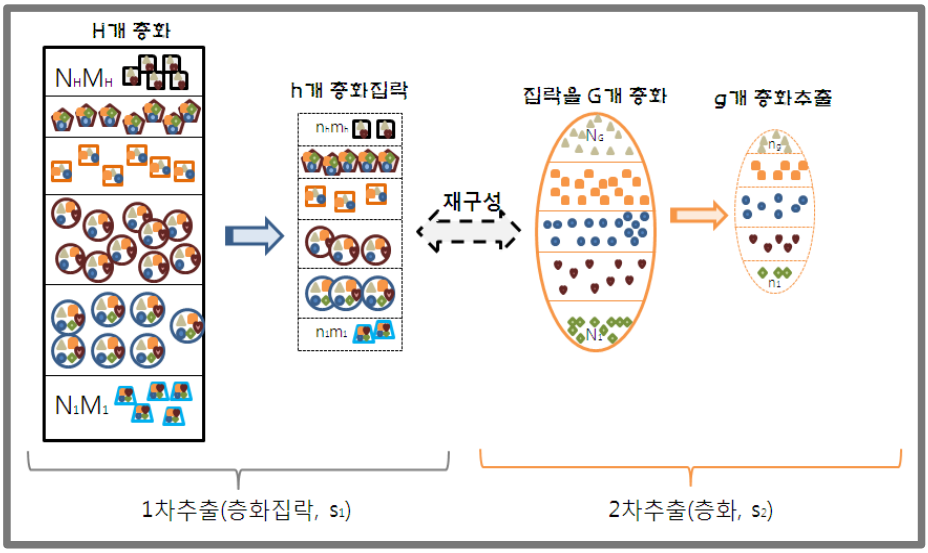
나. 층화집락-층화 이중추출법

1차 표본은 층화집락 추출하고, 2차 표본은 층화한 경우로서 앞으로 연구할 농가경제조사의 복합 이중추출과 유사한 경우이다.

유한모집단(U)은 H개 층화되어 있으며, h층은 N_h개 집락의 M_h개 추출단위로 구성되어 있다. 전체규모 N은 층화집락 표본설계 형태로서 $N=\sum_h^H N_h M_h = N_H M_H$로 표시할 수 있다.

모집단에서 $n_1(=\sum_h n_{1h} m_{1h} = n_h m_h)$개의 1차표본(1st-phase sample, s_1)을 단순임의추출한다. s_1이 추출될 확률은 $\pi_1 = \pi_{1i} = \sum_{i \in s_1} p_1(s_1)$이며, h층의 표본규모는 $n_{1h} = n_{1h} m_{1h} (\in s_{1hm})$이다.

2차 표본을 추출하기 위해서 층화집락 추출한 1차 표본에서 집락을 풀어서 주요 특성별로 G개로 층화하여 $n_{1g} = N_g (\in s_1 = s_{1g})$로 표시하고, $\sum_g^G n_{1g} = n_1 (= n_h m_h)$이 성립한다. G개 층화된 1차 표본틀에서 n_2개의 2차 표본(s_2)을 층화추출한다. 2차 표본은 $n_{2g} (< n_{1g}, s_{2g})$를 만족하며, $n_2 = \sum_g^G n_{2g}$ 이다. 층화집락 추출된 1차 표본에서 2차 층화표본이 추출될 확률은 조건부 확률인 $\pi_{n_2|n_1} = \pi_{n_{2i}|n_{1i}} = \sum_{i \in s_2} p(s_2|s_1)$이며, 각 층의 추출률은 n_{2g}/n_{1g}이고, 관심변수는 y_j이다.



[그림 2-2] 이중추출(층화집락-층화) 표본설계 과정

위와 같은 경우를 농업통계에 적용한 복합 이중추출 추정량의 효율성에 대해서 앞으로 연구할 것이다.

4. 이중추출의 효율성

이중추출의 단계별 표본추출 분산에 의한 상대정도를 이용하여 이중추출의 효율성에 대해 정리해 보았다(Yates, 1981). 상대정도(Relative Precision)는 동일한 추출단위에 근거한 두 가지 상이한 추출방법으로 같은 규모의 추출단위를 표본으로 선택할 경우, 두 가지 방법에 의한 추정치의 추출분산의 백분비로 정의할 수 있다.

이중추출 차별로 추출된 표본에 대해 다음과 같이 정의한다.

A = 1차 표본의 분산으로 1차 표본으로 추출된 모든 x에서 산출한다.
n_1 : A의 표본크기, s_1^2 : A의 분산
B = 1차에서 추출된 2차 표본의 분산으로 오차가 없다고 간주한다.
n_2: B의 표본크기, s_2^2: B의 분산

여기에서는 이해를 돕기 위해 단상표본(single-phase sample), 이상표본(two-phase sample)으로 구분한다. 한 번에 n_1 개 또는 n_2개의 단상표본(single-phase sample)을 추출하지 않고, 단계별로 n_1, n_2개 이상표본(two-phase sample)을 추출하는 경우($n_2 \in n_1$) 분산을 (2.22)와 같이 고려할 수 있다.

$$A = \frac{1}{n_1} s_2^2, \qquad A' = \frac{1}{n_2} s_1^2, \qquad B = \frac{1}{n_2}(1 - \frac{n_2}{n_1}) s_2^2 \tag{2.22}$$

분산 (2.22)와 표본추출률 $\lambda\,(= n_2 / n_1)$ 와 표준편차의 비 $k\,(= s_2 / s_1)$ 를 이용하여 상대정도를 살펴보면, 다음과 같다.

■ 2차 표본에 1차 표본이 포함됨으로써 발생하는 이득의 상대 정도
(단상표본으로 n_2을 추출했을 때, 1차 표본정보를 포함함으로써 발생하는 이익)

$$\frac{A'}{A+B} = \frac{1}{(1-\lambda)k^2 + \lambda} \tag{2.23}$$

■ 2차 표본으로 1차 모든 표본을 모두 사용하지 않음으로써 생기는 손실의 상대 정도
(단상표본으로 n_2을 추출했을 때, 1차 표본에서 추출되지 않은 $(n_1 - n_2)$의 정보가 부족하여 발생하는 손실)

$$\frac{A}{A+B}=\frac{\lambda}{(1-\lambda)k^2+\lambda} \tag{2.24}$$

<표 2-1>은 이상표본과 단상표본의 표본추출률 λ와 표준편차의 비 k를 이용하여 상대정도를 정리한 표이다. 효율적인 2차 표본의 표준오차(standard error)는 1차 표준편차를 절반으로 나눈 것인 $k^2=1/4$이라고 가정하면, <표 2-1>에서 a, b의미는 다음과 같다.

a는 1차 표본 정보를 얻기 위해 표본규모를 4배 증가하면, 2차 표본의 정보는 2.29배로 증가함을 의미하고, b는 2차 표본으로 1차 표본의 $1/4$만을 사용한다면, 2차 표본의 손실된 정보는 0.57임을 의미한다.

〈표 2-1〉 이상표본과 단상표본의 상대 정도 비교

이상표본(Two-phase sample) 단상표본(Single-phase sample)	n_1과 n_2 n_2			n_1과 n_2 n_1		
λ \ k^2	1/2	1/4	1/8	1/2	1/4	1/8
$\lambda=1/2$	1.33	1.6	1.78	0.67	0.8	0.89
$\lambda=1/4$	1.6	2.29[a]	2.91	0.4	0.57[b]	0.73
$\lambda=1/8$	1.78	2.91	4.27	0.22	0.36	0.53

5. 주표본 추출(Master Sampling)

박재수의 조사방법론(1996)에서는 미국 농업조사의 주표본 추출(master sampling)에 대해 소개하고 있다. 주표본 추출은 1차 표본으로 집락을 추출한 이중추출이지만, 일반적인 이중추출과 다른 점은 1차 표본은 일단 추출되면 고정되어 여러 가지 조사에 이용된다는 점이다. 여기서 1차 표본은 주표본(master sample)이라 하고, 2차 표본을 뽑는 방법을 주표본 추출(master sampling)이라고 한다.

1943년 미국 농업부문 통계조사에서 농장군이 집락(cluster)인 집락추출인데, 농업에 관계되는 각종 통계조사를 이 표본에서 다시 뽑아 실시할 수 있게끔 계획하였다. 미국에서는 농장추출 또는 농업에서의 지역추출의 유효적절한 방법이 없었을 뿐만 아니라 그때까지 사용된 추출방법에도 문제가 있어서 이런 방법을 이용하였다.

주표본의 장·단점은 다음과 같다.

- 같은 분석단위 또는 조사단위로 실시할 수 있는 각종 조사를 동일한 주표본에서 뽑은 표본을 사용함으로써, 각 조사 간의 유기적 연락을 도모하고 분석 시에 표본오차를 감소시킬 수 있다.
- 상이한 조사에 대하여 별도의 표본설계를 할 필요없이 쉽게 표본을 뽑을 수 있으며, 주표본에 대한 충분한 지식이 축적되어 이용할 수 있다. 특히, 소규모조사나 갑자기 실시되는 1회 조사에도 사용할 수 있다.
- 표본이 주표본에 한정되므로 조사의 성격에 따라 모집단에서 직접 뽑은 경우에 비하여 정도가 좋지 못하다.
- 장기간 사용하면 대표성을 상실하게 된다.

제3절 최근 연구사례

제3절에서는 이중추출 관련 국내외 연구동향과 적용 실례에 대해 살펴볼 것이다. 전반적으로 단순임의-층화를 위한 이중추출방법과 보조자료를 이용한 비추정 또는 회귀추정에 대한 연구가 주요 내용이었으며, 2000년 이후에는 이중추출의 반복분산 추정과 회귀추정에 대한 연구도 활발하였다. 실제 조사에서는 단순임의-층화를 위한 이중추출법과 회귀추정에 대한 적용사례가 많았으며, 드물게 집락표본을 위한 이중추출법 연구사례를 찾을 수 있었으며, 향후 연구내용과 방향을 정하는데 도움이 되었다.

1. 연구동향

국외에 비해 국내에서는 이중추출에 대한 연구와 적용사례가 많지는 않았지만 표본배분, 추정법, 무응답 등의 주제로 연구되었으며, 장애인경제활동실태조사나 복지실태조사 등을 이중추출법으로 표본설계하여 조사하였다.

우리나라에서는 지은숙이 이중추출의 회귀추정에 대해 연구하였고(1987), 김익찬은 Bayes 및 최대최소 과정을 이중추출에서의 최적 표본설계에 도입하여 비추정량과 회귀추정량에 적용하였다(1988). 이현과 김영원(1995)은 집락표본을 이용한 이중추출법과 비추정에 대해 연구하였고, 김규성 외 2명(2001)이 이중추출에서 모평균 추정방법에 고찰하였다. 비추정량, 회귀추정량과 결합비 추정량을 제안하고 최적 표본수와 최소분산을 유도하였다. 염준근 외 2명(2002)는 단위 무응답이 발생했을 때 이중추출방법을 이용한 가중조정방법과 분산추정량에 대해 연구하였다. 김호일(2002)은 비례배분에 의한 이중추출법에서의 분산의 추이를 알아보고, 분산감소를 위한 표본설계방법을 연구하였다.

최근 국외의 이중추출 연구는 1차 표본의 보조정보를 이용한 비추정 또는 회귀추정 연구가 활발하였으며, 반복분산 추정방법의 하나인 잭나이프 분산추정을 비추정 또는 회귀추정에 더불어 연구하였다.

실례 적용 연구로는 미국, 그리스, 칠레 등에서 자원통계에 적용한 연구와 캐나다에서 사업체와 고용조사에 적용한 사례가 있었다. 대부분 단순임의-층화 이중추출의 실례였고, 집락표본이나 두 개의 표본틀을 적용한 경우가 있었다.

1990년 이후 연구를 살펴보면, 1993년 브렛과 풀러(Breidt and Fuller)[6]는 보조 자료를 이용하여 효율적인 삼상추출방법(three-phase sampling)과 추정치를 연구하였고, 1994년에는 차우드후리와 로이(Chaudhuri and Roy)[7]가 이중추출에서 최적의 단순회귀 추정치 특성에 중점을 두고 연구하였다. 듀폰트(Dupont)[8]는 1995년에 부가적 회귀추정치에 대해 연구하였으며, 1998년에는 히디롤루와 산달(Hidiroglou and Sarndal)[9]이 이중추출에서 보조정보의 유용성을 위한 자료보정(calibration)에 의한 회귀추정치에 대해 연구하였다.

2000년 이후, 보조정보를 활용한 이중추출의 비추정에 중점을 두어 두 명의 싱과 김(Singh, H.P., Singh, S. and Kim, J.M., 2006)과 싱·차우한과 사완·스마렌다체(Singh, Chauhan and Sawan, Smarandache, 2007), 뷰산· 판데·카타라(Bhushan, Pandey and Katara, 2008) 등은 두 개의 보조변수를 이용한 비추정에 대해 연구하였고, 하니프·하크와 샤바즈(Hanif, Hag and Shahbaz, 2010)는 다수의 보조변수를 이용한 비추정 연구를 하였다

Sven Berg(1972)는 집락(cluster) 또는 이단표본추출(two-stage sampling)을 위한 이중추출에 대해 기술하였다. 1차는 임의단순표본추출, 2차는 집락 또는 이단표본추출한 경우이다. 1차 표본은 2차 표본추출에서 집락별 확률비례표본추출을 할 수 있도록 크기를 충분히 고려해야 한다고 하였다.

자료 간의 정보 공유여부에 따른 이중추출의 추정방법과 표본조사는 별개로 취급되었다. 다른 자료와의 정보 공유는 물론 동일 표본틀에서 추출된다는 가정이 필요 없는 이중추출을 '비공유 이중추출(non-nested double sampling)'이라고 한다. Des Raj(1968)[10]와 Cochran(1977)이 표본추출 방법론에 간략하게 논의하였으며, 이런 가정을 통계조사기관에서 적용하였다. 대표적인 예로 캐나다 통계국의 고용 및 급료와 고용시간 조사(Canadian Survey of Employment, Payrolls and Hours, SEPH)이다.

---

6) Breidt, J., and Fuller, W. A. (1993). Regression weighting for multiphase samples. Sankhya, 55, pp.297-309

7) Chaudhuri, A., and Roy, D. (1994). Model assisted survey sampling strategy in two phases. Metrika, 41, pp.355-362

8) Dupont, F.(1995). Redressement alternatifs en presence de plusieurs niveaux d'information auxillaire. Internal report from INSEE, Paris, France.

9) Hidiroglou, M. A., and Sarndal, C. E.(1998). Use of auxiliary information for two-phase sampling. Survey Methodology, 24, pp.11-20

10) Des Raj(1968). Sampling Theory. TMH edition

다른 예로는 Deville(1999)[11]이 연구한 INSEE에서 실시된 가구조사와 Chojnacky (1998)가 미국 농무부의 산림자원조사, Bazigos와 Kavadas(2007)의 그리스 어종통계 등에 적용한 사례가 있다. 집락표본을 위한 이중추출로는 칠레의 Robotham외 2명(2008)의 황새치(swordfish) 연구사례가 있다.

2. 이중추출 모집단 관련 연구

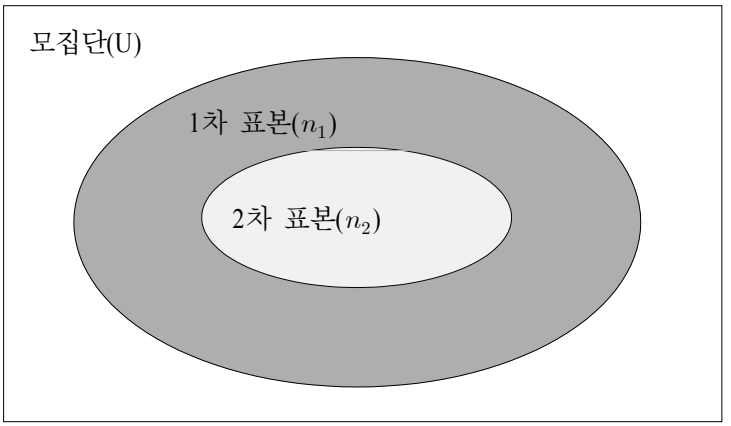
Hidiroglou(2001)는 이중추출에서의 모집단에 대해 언급하면서 모집단과 각 단계별 표본의 관계를 설명하였는데 이를 정리하면 다음과 같다.

이중추출에서는 두 단계에 걸쳐 표본을 추출하는데, 각 단계에서 추출된 표본이 서로 종속될 수도 있고, 독립성을 가질 수도 있다. 또한 다른 표본추출방법과 달리 1차와 2차 표본 모집단이 동일할 수도 있고, 다를 수도 있다.

여기에서는 이중추출에서 단계별 표본이 종속된 경우에 대해 연구하였으며, 독립된 경우에 대해서는 간단하게 기술하였다.

가. 단계별 표본이 종속관계인 경우(Nested Case)

1차 표본과 2차 표본이 종속관계인 경우이다. 동일 모집단에서 1차 표본을 추출하고, 추출된 1차 표본에서 2차 표본을 추출하므로 두 단계의 표본이 종속되며, 2차 표본은 1차 표본에 속하게 된다.



[그림 2-3] 단계별 표본이 종속관계인 경우

11) Deville, J. C. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. Survey methodmology, 25, pp.193-204.

모집단을 $U=\{1,...,k,...,N\}$로 가정하면, 모집단 U에서 추출한 1차 표본은 $n_1(n_1 \subseteq U)$이고, k가 n_1에 추출될 확률은 $\pi_{1k}= P(k \in n_1)$이다. n_1에서 추출한 2차 표본은 $n_2(n_2 \subseteq n_1 \subseteq U)$이다. k가 n에 추출될 확률은 우선 n_1에 추출되어야 하므로 $\pi_{2k|n_1}= P(k \in n_2 | n_1)$ 조건부 확률로 표시할 수 있다.

$\pi_{1k} > 0\,(k \in U)$이면, $\pi_{2k|n_1} > 0 (k \in n_1)$ 이 성립된다. 이때 표본 k의 1차 표본 가중치는 $w_{1k}= 1/\pi_{1k}$가 되고, 2차 표본 가중치는 $w_{2k}= 1/\pi_{2k|n_1}$으로 표시할 수 있다. 그러므로 $k \in n_2$일 때, 표본 k의 전체 가중치는 $w_k^* = w_{1k} \times w_{2k}$로 표시할 수 있다. 1차에서 구한 k의 보조 자료는 x_k이다. 총계 추정치를 간단히 정의하면 다음과 같다.

모집단 $\quad X_1 = \sum^{U} x_{1k}$

1차 표본 $\quad \widehat{X}_1= \sum^{n_1} w_{1k} \times x_{1k} \qquad \widehat{X}= \sum^{n_1} w_{1k} \times x_k$

2차 표본 $\quad \widehat{\widehat{X}}_1= \sum^{n_2} w_k^* \times x_{1k} \qquad \widehat{\widehat{X}}= \sum^{n_2} w_k^* \times x_k$

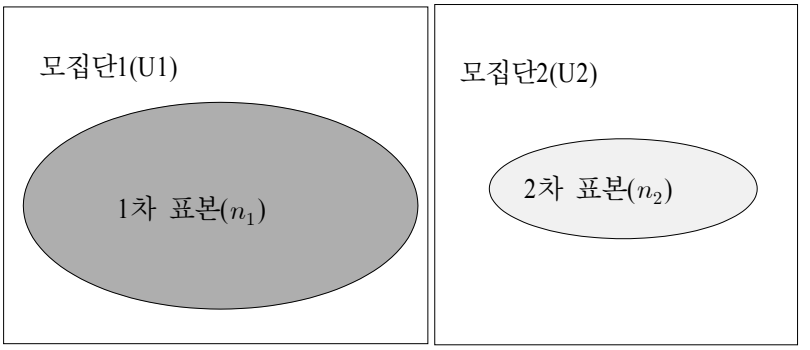
$\qquad\qquad \widehat{Y} = \sum^{n_2} w_k^* \times y_k$

나. 단계별 표본이 독립관계인 경우(Non-Nested Case)

1차 표본과 2차 표본이 독립적인 관계인 경우는 2가지이다. 다른 모집단에서 1차, 2차 표본을 각각 추출하는 경우와 동일 모집단에서 1차와 2차 표본을 독립적으로 추출하는 경우이다.

1) 다른 모집단에서 단계별 표본을 독립적으로 추출

다른 두 모집단을 $U_1=\{1,...,k,...,N_1\}$, $U_2=\{1,...,k,...,N_2\}$로 가정하고, 두 모집단에서 추출된 각 표본은 $n_1(n_1 \subseteq U_1)$, $n_2(n_2 \subseteq U_2)$이며, 두 표본은 독립적이다. k가 각 표본에 추출될 확률은 $\pi_{1k}^{(1)}= P(k \in n_1) > 0$, $\pi_{2k}^{(2)}= P(k \in n_2) > 0$이다.

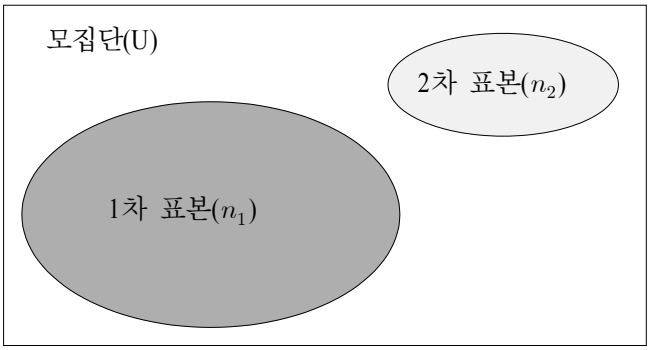


[그림 2-4] 다른 모집단에서 단계별 표본을 독립적으로 추출

1차 표본인 n_1에 추출될 k의 가중치는 $w_{1k}^{(1)} = 1/\pi_{1k}^{(1)}$ 이고, k가 2차 표본인 n에 추출될 가중치는 $w_{2k}^{(2)} = 1/\pi_{2k}^{(2)}$ 이다.

2) 동일 모집단에서 단계별 표본을 독립적으로 추출

동일 모집단에서 1차 표본과 2차 표본을 독립적으로 추출하는 경우는 매우 드물다. 이런 경우, 1차 표본 n_1과 2차 표본은 n_2의 결합 확률(joint inclusion probability)이 매우 복잡할 수 있다. 만약 두 표본을 단순임의추출법으로 추출한다면 간단하게 계산할 수 있는데, 결합추출확률(joint selection probabilities)을 이용하여 분산추정한 탐(Tam)[12]의 연구를 참고한다.



[그림 2-5] 동일 모집단에서 단계별 표본을 독립적으로 추출

12) Tam, S. M. (1984). On covariances from nested samples. The American Statistician, 38, pp.288-289

3. 회귀추정을 위한 반복 분산추정

표본조사에서 얻어지는 분산 추정치에 의해 표본조사 결과가 얼마나 정확한지, 신뢰성 있는 추정결과인지를 평가하게 된다. 좋은 분산 추정량은 분산 추정량이 작아서 안정적이어야 하며, 비편향이거나 편향이어도 매우 작아야 되며, 양수로 계산이 복잡하지 않아야 된다.

반복분산 추정방법(replication variance estimation)은 자료에서 여러번 가상의 표본추출을 수행하고, 수행된 결과를 기반으로 분산를 추정하여 결과의 정확성과 신뢰성을 평가하는 방법이다. 반복분산추정은 다목적 조사 또는 복합조사 시 매우 효율적이며, 테일러 전개의 편도함수(partial derivative) 계산을 포함하지 않아 보다 쉽게 분산추정치를 산출할 수 있다.

대표적인 반복분산 방법으로는 붓스트랩(Bootstrap), 몬테카를로(MonteCarlo), 잭나이프 방법(Jackknife technique) 등이 있는데, 실무 적용에 있어 잭나이프 반복분산 방법이 효율적이어서 이중추출에서도 관련 연구가 활발하다.

잭나이프 방법은 처음에는 Quenouille(1949)에 의해 비 편향(ratio bias)의 크기를 줄여줄 수 있는 기법으로 소개되었다가 Tukey(1958)에 의해 분산추정 기법으로 사용될 수 있음이 밝혀져 지금은 실제 조사에서 분산 추정기법으로 많이 사용되는 방법이다.

잭나이프 방법을 이용하면 계산법은 약간 복잡하나 여러 가지로 유용하고 효율적인 경우가 많다. 잭나이프 방법의 특징은 결합 추정량이 편향(bias)를 갖는 경우에도 잭나이프 추정량은 그에 대한 영향을 덜 받으면서, 결합 추정량보다 더욱 정규분포에 근사한 분포가 된다. 잭나이프 추정은 일반 추정보다 약간 복잡하다는 단점이 있으나 신뢰성 문제뿐만 아니라 기타 측정문제에도 응용될 수 있어 여러 통계조사 추정 시 유용하게 이용될 수 있다.

중복추출의 응용으로 주표본(master sample)을 만들어 놓고, 필요에 따라 여기에서 부표본을 뽑아 사용할 수 있음을 설명한 바 있다. 이와 같은 생각으로 조사계획 시에 미리 여러 조의 반복표본을 구성해 놓으면 불확실한 조사예산하에서도 계획을 수립할 수 있고, 확정된 예산에 맞는 표본크기에 따라 반복표본을 선택하여 사용할 수 있다. 이것은 반복부표본의 근본적인 장점의 하나가 각 부표본이 추정하고자 하는 모수에 대하여 독립적이고도 동등한 추정치를 제공한다는 점을 응용한 것이다.

이중추출의 장점은 효율적인 추정이라고 언급한 바 있다. 모집단과 관련된 변수 또는 1단계의 조사변수를 보조변수를 이용하여 비 추정, 회귀추정 등 다양한 추정을 할 수 있는데, 이런 장점을 이용한 이중추출에서의 잭나이프 연구가 활발하다.

이중추출에서의 잭나이프의 연구는 Rao and Shao(1992)[13]가 Hot-deck imputation 처리시

REE를 위한 잭나이프 분산추정치를 제안하였고, Yung and Rao(2000)[14]는 사후층화로 확대하였다. Fuller(1998)[15]는 이중추출에서 회귀추정를 위한 반복분산 방법을 제안하였다. 2000년 이후에는 김재광이 Fuller, Navarro, Sitter, Yu 등과 공동연구 형태로 잭나이프 방법에 대해 연구하고 있다. 여기서는 Kim, J. K. and Yu, C. L. (2011)에서 층화를 위한 이중추출의 회귀추정에서 잭나이프 추정의 기본개념에 대해 요약하였다

제2절의 층화를 위한 이중추출의 (2.9)에서 (2.25)를 도출할 수 있다.

$$\bar{y}_d^{(k)} = \frac{1}{N}\sum_h^H \hat{N}_{1h}^{(k)}\,\bar{y}_{2h}^{(k)} \tag{2.25}$$

여기서 k는 잭나이프 반복에서 제거되는 반복표본(index)를 의미한다. (2.25)에서 구성요소를 자세히 기술하면 (2.26), (2.27)와 같다.

$$\begin{aligned}\frac{1}{N}\hat{N}_{1h}^{(k)} &= \Sigma_{i\in s_1} w_i^{(k)} \\ &= \begin{cases} \dfrac{1}{(n_1-1)}(n_{1h}-1) & (k\in s_1) \\ \dfrac{1}{(n_1-1)} & (k\notin s_1) \end{cases}\end{aligned} \tag{2.26}$$

$$\begin{aligned}\bar{y}_{2h}^{(k)} &= \frac{\Sigma_{i\in s_2} w_i^{(k)} y_i}{\Sigma_{i\in s_2} w_i^{(k)}} \\ &= \begin{cases} \dfrac{1}{(n_{2h-1})}(n_{2h}\bar{y}_{2h}-y_k) & (k\in s_2) \\ \bar{y}_{2h} & (k\notin s_2) \end{cases}\end{aligned} \tag{2.27}$$

(2.9)와 (2.25)를 이용하여 층화를 위한 이중추출에서의 잭나이프 분산 추정량를 도출할 수 있다.

$$\hat{V}_J = \sum_{k\in s_1}\frac{n-1}{n}(1-f_1)(\bar{y}_d^{(k)}-\bar{y}_d)^2 \qquad (f_1=\frac{n_1}{N}) \tag{2.28}$$

$$\hat{V}_J \doteq \frac{1}{n_1}(1-f_1)\sum_h^H w_h(\bar{y}_{1h}-\bar{y}_1)^2 + (1-f_1)\sum_h^H \frac{1}{n_{2h}} w_h^2 s_{1h}^2 \tag{2.29}$$

---

13) Rao, J. N. K., and Shao, J. (1992), “Jackknife Variance EstimationWith Survey Data Under Hot Deck Imputation,” Biometrika, 79, 811–822.

14) Yung, W., and Rao, J. N. K. (2000), “ Jackknife Variance Estimation Under Imputation for Estimators Using Postratification Information,” JSSA, 95, 903–915.

15) Fuller, W. A. (1998), "Replication Variance Estimation for Two-Phase Samples," Statistica Sinica, 8, 1153-1164.

(2.17)과 (2.29)를 비교하여 잭나이프 분산 추정량의 편의(bias)를 구하면 다음과 같다.

$$Bias(\hat{V}_J) \doteq -E\left\{f_1 \sum_h^H (\frac{1}{n_{2h}} - \frac{1}{n_{1h}}) s_{1h}^2\right\} \tag{2.30}$$

(2.30)에서 1차 표본 추출률이 $f_1 \doteq 0$ 이라면, 편의는 거의 없어서 $Bias \doteq 0$ 이 성립한다. 하지만, 이외의 경우에는 잭나이프 분산추정량은 분산을 과소추정하게 되므로 이런 현상을 해결하기 위해 (2.31)을 고려할 수 있다.

$$\bar{y}_{2h}^{(k)} = \begin{cases} \frac{1}{(n_{2h} - \delta_{2h})}(n_{2h}\bar{y}_{2h} - \delta_{2h} y_k) & (k \in s_2) \\ \bar{y}_{2h} & (k \notin s_2) \end{cases} \tag{2.31}$$

(2.31)에서 $\delta_{2h} = 1$이면 $\bar{y}_{2h}^{(k)}$는 (2.27)와 같다. (2.31)를 이용하여 잭나이프 분산추정량 (2.32)와 편의(2.33)을 도출할 수 있다.

$$\hat{V}_J \doteq \frac{1}{n_1}(1-f_1)\sum_h^H w_h(\bar{y}_{1h} - \bar{y}_1)^2 + (1-f_1)\sum_h^H \frac{(n_{2h}-1)\delta_{2h}^2}{(n_{1h} - \delta_{2h})^2} w_h^2 s_{1h}^2 \tag{2.32}$$

$$Bias(\hat{V}_J) \doteq E\left[\sum_h^H \left\{(1-f_1)\frac{(n_{2h}-1)\delta_{2h}^2}{(n_{2h}-\delta_{2h})^2} - \frac{1}{n_{2h}}\left(1 - f_1 \frac{n_{2h}}{n_{1h}}\right)\right\} w_h^2 s_{1h}^2\right] \tag{2.33}$$

(2.33)에서 δ_{2h}가 (2.34)와 같으면 편의는 '0' 이 된다.

$$\delta_{2h} = \frac{n_{2h}}{1 + \sqrt{n_{2h}(n_{2h}-1)}/d_{2h}} \tag{2.34}$$

$$\left(d_{1h} = \sqrt{(1 - f_1 n_{2h}\frac{1}{n_{1h}})/(1-f_1)}\right)$$

(2.34)를 참조하여 δ_{2h}를 결정해주면, 잭나이프 분산 추정치는 $f_1 \doteq 0$ 이 아니어도 불편추정량이 된다.

4. 이중추출방법 적용실례

이중추출방법을 실제로 적용한 실례를 간략하게 소개하였는데, 대체로 자원통계에서 이중추출을 적용한 것을 알 수 있다.

가. 미국

미국 농림부(USDA,United States Department of Agriculture)의 산림자원조사(Forest Inventory)에서 Chojnacky는 단순임의-층화 이중추출을 적용하여 표본설계하였다. 산림지도에서 표본단위를 1km 격자(grid)구분 후 계통추출하였는데, grid와 모집단 간의 강한 상관관계가 없으면 단순임의추출 분산추정 적용이 가능하다고 하였다.

Rao, Cochran, Bickford 등은 분산 CV 차이를 비교하였으며, 비 추정치를 산출하였다.

나. 캐나다

캐나다 통계청의 Hidiroglou(2001) 연구에서는 이중추출의 모집단과 단계별 표본과의 관계를 알 수 있으며, QRCS(Quartely Retail Commodity Survey, 소매사업체 조사), SEPH (Canadian Survey of Employment, Payrolls and Hours, 고용 및 급료와 고용시간 조사)에 이중추출을 적용하여 회귀추정하였다.

SEPH(Rancourt and Hidiroglou)[16]에서는 두 개의 독립표본들을 동일 모집단을 대표하는 각기 다른 표본틀에서 추출하였다. 보조변수(x)인 고용인 수와 총 급료는 캐나다 국세청(Canada Customs and Revenue Agency)의 행정파일에서 추출된 표본에서 정보를 수집하였고, 관심항목(y_i) 고용인 총 노동시간과 총 소득은 캐나다 사업체 등록 통계(Statistics Canada Business Register)에서 추출된 표본에서 수집하였다.

다. 그리스

Bazigos와 Kavadas(2007)는 그리스 어종통계의 어종연령 추정 시 단순임의-층화를 위한 이중추출방법을 적용하였다. 1차 표본으로 추출된 어종의 길이의 분포를 조사하였고, 길이로 층화한 그룹에서 2차 표본을 추출하여 연령을 추정하였다.

라. 칠레

칠레의 Robotham, Young and Saavedra-Nievas(2008)는 어종통계에 대한 연구는 집락

16) Rancourt, E., Hidiroglou, M. A. (1998). Use of administrative records in the Canadian survey of employment, payrolls and hours. Proceedings of the Survey Methods Section, pp.39-47.

표본을 위한 이중추출의 실례이다. 어종통계 연구에서 황새치 길이별 집락에서 1차는 집락추출, 2차는 임의추출하여 이중추출하였다. 그리고, 복합추정에 사용하는 잭나이프 분산추정과 단순임의-층화 이중추출 분산추정을 비교하여 효율성을 비교하였다.

잭나이프 분산과 단순임의-층화를 위한 이중추출 분산의 상대효율(CV) 비교 결과, 단순임의-층화 이중추출 분산의 효율성이 잭나이프 분산보다 좋았으나 단순임의-층화 이중추출 분산의 상대편향은 음수로 나타나 모집단에 대한 설명력이 약했으며, 과소추정된 것으로 판명되었다.

집락모집단에서 단순임의추출하여 집락 내 종속성을 제거하는 경우에는 정도가 과소추정되어 설명력이 낮아지므로, 잭나이프 분산추정량이 의미가 있다고 할 수 있다.

이와 같이 집락표본의 특징을 제거하여 분산추정하는 경우는 효율성이 낮아질 수 있으므로 표본의 특징을 충분히 반영하여 효율성을 높일 수 있는 분산 추정식 연구가 필요하다.

위에서 살펴본 것과 같이 실제로 이중추출이 적용된 예를 보면, 이중추출은 명확한 표본추출틀을 구축할 수 없는 자원통계에서 많이 사용되고 있다. 이중에서 집락표본에 대한 이중추출인 칠레의 어종통계에 대한 결과가 주목할 만하다. 우리나라에서 이번에 개편할 2010년 기준 농업통계는 기본적으로 1차 표본이 집락형태이기 때문이다. 집락표본의 특성을 반영한 분산추정을 하여야 표본설계의 효율성을 높일 수 있다고 했으므로 향후 연구 시 집락표본의 특성을 반영한 추정량 연구를 주요 연구내용으로 하여 연구를 진행할 계획이다.

제4절 맺음말

1. 연구결과 요약

사회·경제가 복잡하고 다양한 형태로 급격히 변화하고 있는데, 도시화·시장개방·기후변화 등의 여파로 농업환경은 더욱 빠르게 변하고 있다. 그러다보니 전수조사인 농림어업총조사 조사시기와 농업조사 표본개편 시기 차이가 2~3년임에도 불구하고, 추출된 표본농가의 특성이 변화한 경우가 다수 발생하였다.

일반적으로 표본개편 시 추출된 새로운 표본 특성과 과거 표본 특성이 완전히 일치되지 않아 시계열 단층 현상이 발생할 수 있는데, 최근 국내외 환경의 급변화로 시계열 단층 현상이 심해져 농촌동향 분석이 실제와 다른 괴리현상이 나타날 수 있으므로 농촌의 최근 정보를 최대한 반영할 수 있는 표본설계 및 추정기법의 연구가 필요하다.

2012년부터 점진적으로 개편할 농업통계조사의 표본설계 개편방향은 농업조사 자료를

주표본(master sample)으로 관련 농업조사 표본을 이중추출방법으로 추출하여 통계조사를 실시할 예정이다. 이중추출방법이 통계청에서는 농업통계에 처음 적용하는 표본설계 방법이어서 크게 3단계로 선행연구, 모집단구축, 모의 표본설계로 이중추출 표본설계 효과분석이 이루어질 것이다. 이번 연구는 첫 단계로 기초문헌을 연구하여 이중추출 방법이란 어떤 방법인지 살펴보고, 국내외 연구동향을 파악하여 정리하였다.

이중추출방법은 동일한 표본을 대상으로 간략조사와 심층조사를 두 단계에 걸쳐 실시하는 표본설계로 1차 표본을 부모집단으로 간주하여 1단계 조사에서 획득한 표본의 최근 정보를 이용하여 2차 표본설계를 실시한다. 그럼으로써 최근 정보를 이용한 표본설계를 할 수 있어 표본의 대표성을 높여 통계의 정도를 향상시킬 수 있는 장점이 있다.

이중추출 관련 국내외 연구동향은 전반적으로 단순임의-층화를 위한 이중추출방법과 보조자료를 이용한 비추정 또는 회귀추정에 대한 연구가 주요 내용이었으며, 2000년 이후에는 이중추출의 반복분산 추정과 회귀추정에 대한 연구도 활발하였다.

실제 자원조사에서 단순임의-층화를 위한 이중추출법과 회귀추정에 대한 적용사례가 있었으며, 드물게 집락표본을 위한 이중추출법 적용사례도 있었다. 집락표본 적용사례인 칠레의 어종통계 결과를 참고하여 집락표본의 특성을 반영한 집락표본의 특성을 반영한 분산 추정량 연구로 표본설계 효율성을 검토하는 연구방향으로 연구를 진행할 계획이다.

2. 향후 중점 연구사항

본 연구를 기초연구로 후속 연구는 이중추출 표본설계 효과를 분석하기 위하여 농가경제조사 이중추출 유사 모집단을 구축하는 것이다. 현재 활용이 가능한 자료는 2005년 기준 자료로서 이중추출을 하기 위해서는 2차 표본인 농가경제조사를 1차 표본인 농업조사의 부분집합 형태로 연계하여야 하는데, 더욱 세밀하게 연계하기 위해 매년 초에 실시하는 농가명부조사 자료를 이용할 것이다. 이러한 연계과정에서 모집단 분포, 공통변수와 주요변수 특성분석 등이 이루어져야 하며, 다양한 구성요인으로 층 세분화 방법, 연계자료의 정확성 등을 분석하여야 한다.

3개 자료를 연계하여 구축된 농가경제조사 표본틀을 이용하여 여러 개의 표본을 추출하고 제시된 추정치의 타당성과 안정성을 측정하여 추정치의 효율성을 검증한다.

또한, 이중추출 모의 표본설계를 하여 2차 표본추출 시 나타날 수 있는 중점 사항들을 점검하고 본 연구의 궁극적인 목표인 표본설계 효과를 측정하는 것이다. 또한, 조건부 결합추정식과 잭나이프 추정식의 효율성을 검토하여 농가경제조사 표본설계에 적합한 추정방법을 검토·제시할 것이다.

참고문헌

김규성·김진석·이선순 (2001). 이중추출에서의 모평균 추정. 응용통계연구 제14권 1호 pp.13-24.

김익찬 (1988). 이중 표본추출에서의 BAYES 및 최대최소 과정. 성균관대학교 통계학 박사 논문.

김재광 (2008). 표본조사론. 자유아카데미.

김호일 (2002). 이중추출법의 비례배분 경우에 분산의 변화에 관한 소고. Journal of Natural Science Vol., 9.

박재수 (1996). 표본조사론. 박영사.

박홍래 (1997). 통계조사론. 영지문화사.

염준근, 손창균, 정영미 (2002). 이중 추출 방법을 이용한 단위 무응답의 가중치 조정 방법에 관한 연구. 춘계 한국통계학회

이경·김영원 (1995). 집락추출법을 이용한 이중추출법에 관한 연구. 숙명여대 자연과학 논문집 제6호 135~145.

지은숙(1987). A study on optimum allocation and regression estimators of double sampling. The J. of KWU, Vol. 16.

Bazigos, G. and Kavadas(2007). Optimal sampling designs for large-scale fishery sample surveys in Greece. Mediterranean Marine Science Vol. 8/2, pp.65-82.

Binder, D. A., Babyak, C., Brodeur, M., Hidiroglou, M., and Jocelyn, W. (2000). Variance Estimation for Two-Phase Stratified Sampling. Canadian Journal of Statistics, Vol. 28, No. 3, pp.751 – 764.

Chaudhuri, A., and Roy, D. (1994). Model assisted survey sampling strategy in two phases. Metrika, 41, pp.355-362.

Chojnacky, D. C. (1998). Double Sampling for Stratification: A forest inventory application in the interior west. United States Department of Agriculture, Research Paper RMRS-RP-7, June 1998.

Cochran, W. G. (1977). Sampling Techniques (3rd edition). New York: John Wiley & Sons, Inc.

Des Raj(1968). Sampling Theory. TMH edition

Deville, J. C. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. Survey methodmology, 25, pp.193-204.

Dupont, F.(1995). Redressement alternatifs en presence de plusieurs niveaux d'information auxillaire. Internal report from INSEE, Paris, France.

Fuller, W. A. (1998). Replication variance estimation for two-phase samples. Statistica Sinica 8, pp.1153-1164.

Hidiroglou, M. A. (2001). Double sampling. Survey Methodology, Vol. 27, No.2, pp.143-154.

Hidiroglou, M. A., and Sarndal, C. E.(1998). Use of auxiliary information for two-phase sampling. Survey Methodology, 24, pp.11-20.

Kim, J. K., Navarro, A. and Fuller, W. A. (2006). Replication Variance Estimation for Two-Phase Stratified Sampling. JSSA, Vol. 101, No. 473

Kim, J. K. and Yu, C. L. (2011), Replication variance estimation under two-phase sampling. Survey Methodology, Vol. 37, No. 1, pp.67-74, Statistics Canada.

Kish, L. (1965). Survey Sampling. New York: John Wiley & Sons, Inc.

Neyman, J. (1938). Contribution to the theory of sampling human populations. Journal of the American Statistical Association, 33, pp.101-116.

Rao, J. N. K., and Shao, J. (1992), "Jackknife Variance EstimationWith Survey Data Under Hot Deck Imputation," Biometrika, 79, 811–822.

Robotham, H. Young, Z. I. Saavedra-Nievas, J. C. (2008). Jackknife method for estimating the variance of the age composition using two-phase sampling with an application to commercial catches of swordfish(Xiphias gladius). Fisheries research93, pp.135-139.

S¨arndal, C. E., Swensson, B. and Wretman, J. (1992). Model Assisted Survey Sampling. Springer-Verlag, New York.

Sven Berg (1972). Double Sampling for Cluster- or Two-stage Sampling. Int. Stat. Rev. Vol. 40, No. 1.

Wolter, K. M. (2003). Introduction to Variance Estimation. Springer Verlag.

Yates, F. (1981). Sampling Methods for Censuses and Surveys (4th edition). London: Griffin and Co.

Yung, W., and Rao, J. N. K. (2000), " Jackknife Variance Estimation Under Imputation for Estimators Using Postratification Information," JSSA, 95, 903–915.