The Korean Communications
in Statistics Vol. 6, No. 1, 1999
pp. 131-141

Variance Estimation Using Poststratified Complex Sample

Kyu Seong Kim[1)]

Abstract

Estimators for domains and approximate estimators of their variance are derived using post-stratified complex sample. Furthermore, we propose an adjusted variance estimator of a domain mean in case of considering the post-stratified complex sample as simple random sample.

A simulation study based on the data of Farm Household Economy Survey is presented to compare variance estimators numerically. From the study, we showed that our adjusted variance estimator compensate for the under-estimation problem considerably.

1. 서 론

통계조사는 주로 유한모집단(finite population)을 대상으로 한다. 이러한 모집단은 그 자체가 지리적, 환경적 속성을 가지고 있기 때문에 통상적인 랜덤표본(random sample)을 선정하기 어려운 경우가 많다. 따라서 모집단의 구조를 능률적으로 대표하는 표본으로 층화(stratification)와 집락화(clustering)를 통하여 선출되는 복합표본(complex sample)을 많이 이용하게 된다. 유한모집단에서 복합표본의 불가피성과 이에 근거한 추론은 Kish와 Frankel(1974)에 의해 명료하게 설명되었다.

관심있는 영역(domain of interest)의 모평균이나 모총계를 추정하고자 할 때 이 영역이 층화에 고려되지 않았다면, 사후층화(Post-stratification)를 통하여 관심 영역의 모수를 추정하게 된다. 사후층화는 표본을 선출하여 관측치를 조사한 후, 이를 근거로 표본을 사후적으로 분류하는 방법이다. 예를 들어 연령별, 직업별, 학력별 특성치가 필요한 조사에서는 사후층화에 의한 추론이 요구된다. 왜냐하면, 조사 전에 연령별, 직업별 응답자를 파악하는 것은 거의 불가능하거나, 조사비용이 많이 들어 사후층화를 하는 것이 불가피하기 때문이다.

실제 많은 경우에 사후층화에 의한 추론이 이루어지고 있긴 하지만, 이에 대한 관심부족으로 많은 오류를 내포하고 있는 것이 사실이다. (Holt와 Smith(1979), Little(1993)). 흔히 나타나는 오류의 근본적인 출발점은 표본이 추출된 확률분포를 무시하고 랜덤표본으로 간주하여 추론을 하는 것인데, 그 이유는 두 가지로 설명된다. 하나는 보통의 복합표본은 표본추출과정이 복잡하기 때문에 표본이 추출된 확률분포에 근거하여 추론을 할 경우, 추론이 대단히 어려워지기 때문에 표본추출확률분포(sampling distribution)를 이용하는 대신 랜덤확률분포를 이용하는 것이고, 다른 하나는 표본추출확률분포에 근거한 추론을 하기 위해서는 표본추출확률(selection probability)과 집락크기(cluster size)등 보조정보가 유지, 보전되어야 하는데 많은 경우에 그렇지 못하다. 따라서 정보부

1) Assistant Professor, Department of Computer Science and Statistics, the University of Seoul, Seoul, 130-743, Korea.

족으로 인하여 랜덤표본으로 간주하여 추론을 하게 되는 것이다.

본 연구에서는 복합표본에 기초한 관심영역에 대한 추론 방법을 제안하고자 한다. 먼저 층화2단집락표본에 근거한 관심영역의 모총계 및 모평균의 추정량을 구하고, 이 추정량들의 불편분산추정량(unbiased variance estimator) 및 근사불편분산추정량(approximate unbiased variance estimator)를 구한다. 표본추출확률분포에 근거한 분산추정량은 복합표본에 근거한 정확한 추론의 이론적 바탕을 제공할 것이다. 다음으로 층화2단집락표본을 랜덤표본으로 간주하고 추론을 할 경우에 발생하는 오류를 밝히고, 이러한 오류를 수정할 수 있는 수정된 분산추정량을 제안한다. 결과적으로 이 경우 발생하는 오류는 분산의 과소추정(under-estimation)인데, 통계조사에서 분산의 과소추정은 심각한 추론의 문제를 발생시킬 소지가 있다. 즉, 실제 보다 분산이 더 작다고 판단함으로써 통계조사 결과의 신뢰도를 과대평가하는 것이다. 본 연구에서는 분산의 과소추정을 방지하는 수정된 분산추정량을 유도하여, 불가피하게 복합표본을 랜덤표본으로 간주한다 하더라도 오류를 줄일 수 있는 방법을 제안하고자 한다.

제2절에서는 복합표본에 근거한 관심영역의 모총계 및 모평균의 분산추정량을 유도하고, 제3절에서는 복합표본을 랜덤표본으로 간주한 경우에 발생하는 오류를 줄일 수 있는 수정된 분산추정량을 제안한다. 제4절에서는 농가경제조사 표본을 이용하여 제안된 분산추정량의 효율을 수치적으로 비교하고, 마지막으로 제5절에서는 결론을 언급한다.

2. 복합표본에 근거한 분산추정

2.1 관심영역의 모총계 추정

층화2단집락표본추출(stratified two-stage cluster sampling)을 고려하자. 모집단은 L개의 층으로 구분되고 각 층은 N_h ($h=1,\ldots,L$)개의 1차추출단위(Primary sampling unit ; PSU)로 구성되어 있다. 또한 각 PSU는 M_{hi} ($i=1,\ldots,N_h$, $M_h=\sum_{i=1}^{N_h} M_{hi}$)개의 2차추출단위(Secondary sampling unit; SSU)로 이루어진다. 각 층에서 n_h개의 PSU를 확률비례추출하고 추출된 PSU에서 m_{hi} ($m_h=\sum_{i\in s_h} m_{hi}$),개의 SSU를 랜덤추출한다. PSU의 표본을 s_h라 하고 SSU의 표본을 s_{hi}라고 했을 때 모총계, $Y=\sum_{h=1}^{L}\sum_{i=1}^{N_h}\sum_{j=1}^{M_{hi}} y_{hij}$, 의 불편추정량은 다음과 같다.

$$\hat{Y}=\sum_{h=1}^{L}\sum_{i\in s_h}\frac{1}{\pi_{hi}}\sum_{j\in s_{hi}}\frac{y_{hij}}{\pi_{j|hi}}. \tag{2.1}$$

여기서 π_{hi}는 (hi)번째 PSU의 포함확률(inclusion probability)이며, $\pi_{j|hi}$는 (hi)번째 PSU에서 j번째 SSU가 선출될 포함확률이다. 이 두 포함확률은 모두 표본추출확률분포(sampling distribution)에 근거한다.

관심의 대상이 되는 영역을 D로 나타내고, D의 분포는 사후층화에 의하여 파악이 가능하다고

하자. 그리고 원래 표본 중에서 D에 속하는 표본추출단위(sampling unit)들의 모임을 s_h^*, s_{hi}^*라고 하자. 그러면 관심영역 D의 모총계, $Y^* = \sum_{h=1}^{L}\sum_{i=1}^{N_h}\sum_{j=1}^{M_{hi}} y_{hij}$, 의 추정량은 다음과 같이 표현된다.

$$\hat{Y}^* = \sum_{h=1}^{L}\sum_{i\in s_h^*}\frac{1}{\pi_{hi}}\sum_{j\in s_{hi}}\frac{y_{hij}}{\pi_{j|hi}}. \tag{2.2}$$

그런데 s_h^*와 s_{hi}^*는 원래 표본인 s_h와 s_{hi}를 사후적으로 분류하여 만든 표본이므로 표본수가 랜덤이다. 표본수가 주어졌다고 가정하면 D 위에서의 표본추출확률분포를 생각할 수 있고, 이러한 확률분포는 모집단이 관심영역 D로 제한된 축소된확률분포(reduced sampling distribution)가 된다. 축소된 확률분포에서 표본추출단위들의 포함확률을 다음과 같다.

$$\pi_{hi}^* = n_h^*\frac{M_{hi}^*}{M_h^{**}} \quad : \ D\text{에서 } (hi)\text{번째 PSU의 포함확률} \tag{2.3}$$

$$\pi_{j|hi}^* = \frac{m_{hi}^*}{M_{hi}^*} \quad : \ D\text{의 } (hi)\text{번째 PSU에서 } j\text{번째 SSU의 포함확률}$$

여기서 $M_h^{**} = \sum_{i=1}^{N_h} M_{hi}^*$이며, '*'가 붙은 항은 D에서 '*'가 붙지 않은 항에 대응되는 값들이다. 위의 식 (2.2)에서 제시된 추정량은 다음과 같이 표현될 수 있다.

$$\hat{Y}^* = \sum_{h=1}^{L}\sum_{i\in s_h^*}\frac{\pi_{hi}^*}{\pi_{hi}}\frac{1}{\pi_{hi}^*}\sum_{j\in s_{hi}}\frac{\pi_{j|hi}^*}{\pi_{j|hi}}\frac{y_{hij}}{\pi_{j|hi}^*}. \tag{2.4}$$

명백하게 관심영역에서 모총계 추정량 $\hat{Y}^*$은 조건부 편의추정량(conditionally biased estimator)이 된다. 물론 전체영역으로 추론의 범위를 넓히면 $\hat{Y}^*$은 불편추정량이 된다.

관심영역 D에서 모총계 추정량 $\hat{Y}^*$의 일반적인 조건부 분산은 유도가 가능하지만, 현실적인 유용성을 감안하여 PSU를 확률비례복원추출하는 경우로 국한하자. 우선 다음의 기호들을 약속한다.

$$\sigma_{hi}^2 = Var\{\bar{y}_{hi}^*|hi\} = (1-\frac{m_{hi}^*}{M_{hi}^*})\frac{S_{hi}^{*2}}{m_{hi}^*}, \quad S_{hi}^{2*} = \frac{1}{M_{hi}^*-1}\sum_{i=1}^{M_{hi}}(y_{hij}-\bar{Y}_{hi}^*)^2 \tag{2.5}$$

$$A_h = \frac{n_h^*}{n_h}\frac{M_h}{M_h^{**}}, \quad B_{hi} = \frac{m_{hi}^*}{m_{hi}}\frac{M_{hi}}{M_{hi}^*},$$

$\mathbf{m}^* = \{m_{hi}^* | \ h=1,\ldots,L, \ i=1,\ldots,N_h^*\}$: D 에서의 표본수

정리 2.1 PSU를 집락크기에 비례하는 확률비례복원추출하고, SSU를 랜덤추출하는 층화2단집락 표본추출을 가정하자. 그러면 관심영역의 모총계의 불편추정량은 다음과 같다.

$$\hat{Y}^* = \sum_{h=1}^{L} \frac{M_h}{n_h} \sum_{i \in s_h^*} \frac{m_{hi}^*}{m_{hi}} \bar{y}_{hi}^* \tag{2.6}$$

여기서 $\bar{y}_{hi}^* = \sum_{j \in s_{hi}^*} y_{hij}/m_{hi}^*$이다. 그리고 모총계 추정량 $\hat{Y}^*$의 조건부 분산은 아래와 같다.

$$Var\{\hat{Y}^* \mid \mathbf{m}^*\} = \sum_{h=1}^{L} A_h^2 [\frac{1}{2n_h^*} \sum_{i=1}^{N_h^*} \sum_{j \neq i}^{N_h^*} M_{hi} M_{hj} (\frac{m_{hi}^*}{m_{hi}} \overline{Y_{hi}^*} - \frac{m_{hj}^*}{m_{hj}} \overline{Y_{hj}^*})^2 + \sum_{i=1}^{N_h^*} \frac{B_{hi}^2 M_{hi}^{*2} \sigma_{hi}^2}{\pi_{hi}^*}] \tag{2.7}$$

또한 조건부 분산은 아래의 식으로 불편 추정할 수 있다.

$$v(\hat{Y}^* \mid \mathbf{m}^*) = \sum_{h=1}^{L} \frac{M_h^2}{2n_h^2(n_h^*-1)} \sum_{i \in s_h^*} \sum_{j \in s_h^*} (\frac{m_{hi}^*}{m_{hi}} \bar{y}_{hi}^* - \frac{m_{hj}^*}{m_{hj}} \bar{y}_{hj}^*)^2 \quad \blacksquare \tag{2.8}$$

증명. 조건부 분산을 2단계로 나누어 구한다.

$$\begin{aligned} Var\{\hat{Y}^* \mid \mathbf{m}^*\} &= \sum_{h=1}^{L} A_h^2 [\, V_1\{ \sum_{i \in s_h^*} \frac{B_{hi} M_{hi}^* E_2\{\bar{y}_{hi}^*\}}{\pi_{hi}^*} \mid \mathbf{m}^* \} \\ &\qquad + E_1\{ \sum_{i \in s_h^*} \frac{B_{hi}^2 M_{hi}^{*2} V_2\{\bar{y}_{hi}^*\}}{\pi_{hi}^{*2}} \mid \mathbf{m}^* \}] \\ &= \sum_{h=1}^{L} A_h^2 [\, V_1\{ \sum_{i \in s_h^*} \frac{B_{hi} Y_{hi}^*}{\pi_{hi}^*} \mid \mathbf{m}^* \} + E_1\{ \sum_{i \in s_h^*} \frac{B_{hi}^2 M_{hi}^{*2} \sigma_{hi}^2}{\pi_{hi}^{*2}} \mid \mathbf{m}^* \}] \end{aligned}$$

여기에 다음의 두 항을 대입하여 조건부 분산을 구할 수 있다.

$$V_1\{ \sum_{i \in s_h^*} \frac{B_{hi} Y_{hi}^*}{n_h^* p_{hi}^*} \mid \mathbf{m}^* \} = \frac{1}{2n_h^*} \sum_{i=1}^{N_h^*} \sum_{j \neq i}^{N_h^*} M_{hi} M_{hj} (\frac{m_{hi}^*}{m_{hi}} \overline{Y_{hi}^*} - \frac{m_{hj}^*}{m_{hj}} \overline{Y_{hj}^*})^2,$$

$$E_1\{ \sum_{i \in s_h^*} \frac{B_{hi}^2 M_{hi}^{*2} \sigma_{hi}^2}{n_h^{*2} p_{hi}^{*2}} \mid \mathbf{m}^* \} = \sum_{i=1}^{N_h^*} \frac{B_{hi}^2 M_{hi}^{*2} \sigma_{hi}^2}{\pi_{hi}^*}.$$

또한, 2단계 확률비례복원추출의 불편분산추정량이 다음과 같음을 이용하고,

$$v(\hat{Y}^* \mid \mathbf{m}^*) = \sum_{h=1}^{L} A_h^2 \frac{1}{2n_h^{*2}(n_h^*-1)} \sum_{i \in s_h^*} \sum_{j \in s_h^*} (\frac{B_{hi} \overline{y_{hi}^*} M_{hi}^*}{p_{hi}^*} - \frac{B_{hj} \overline{y_{hj}^*} M_{hj}^*}{p_{hj}^*})^2,$$

여기에 $A_h = n_h^* M_h / n_h M_h^{**}$, $B_{hi} = m_{hi}^* M_{hi} / m_{hi} M_{hi}^*$ 그리고 $p_{hi}^* = M_{hi}/M_h^{**}$을 대입하면 식(2.8)의 조건부 분산의 불편추정량을 얻을 수 있다. ■

표본 s_{hi}들의 크기와 s_{hi}^*들의 크기가 비슷한 경우에 식 (2.8)에서 주어진 불편분산추정량은 근사(approximation)를 통하여 간단한 형태로 표현될 수 있다. s_{hi}들의 크기와 s_{hi}^*들 크기가 비슷하다고 가정하면 $m_h = \sum_{i \in s_h} m_{hi} \approx n_h m_{hi}$, $m_h^* = \sum_{i \in s_h} m_{hi}^* \approx n_h^* m_{hi}^*$, 그리고 $m_{hi}^*/m_{hi} \approx m_{hj}^*/m_{hj}$이 된다. 이 결과를 식(2.8)에 대입하면 불편분산추정량은 아래의 식으로 간단하게 표현된다.

$$
\begin{aligned}
v(\hat{Y}^* \mid \mathbf{m}^*) &\approx \sum_{h=1}^{L} M_h^2 \left(\frac{m_{hi}^*}{m_{hi}}\right)^2 \frac{n_h^*}{n_h^2} \frac{1}{2n_h^*(n_h^*-1)} \sum_{i \in s_h} \sum_{j \in s_h} (\overline{y_{hi}^*} - \overline{y_{hj}^*})^2 \\
&= \sum_{h=1}^{L} M_h^2 \left(\frac{m_{hi}^*}{m_{hi}}\right)^2 \frac{n_h^*}{n_h^2} s_h^{2*} \frac{1}{m_{hi}^*} \{(\overline{m_h^*}-1)\tau_h^* + 1\} \\
&\approx \sum_{h=1}^{L} M_h^2 \left(\frac{m_h^*}{m_h}\right)^2 \frac{s_h^{2*}}{m_h^*} \{(\overline{m_h^*}-1)\tau_h^* + 1\}
\end{aligned}
\tag{2.9}
$$

여기에서 τ_h^*는 관심영역에서 h층의 급내상관계수(intraclass correlation coefficient)이고, $\overline{m_h^*}$는 D의 h층에서 m_{hi}^*들의 평균값이다. 위 근사식은 분산추정량 $v(\hat{Y}^* \mid \mathbf{m}^*)$이 층내 표본분산($s_h^{2*}$)과 급내상관계수($\tau_h^*$)를 포함하는 식으로 표현되어 집락의 효과가 급내상관계수의 형태로 나타남을 보여준다. 여기에서 보여진 집락의 효과는 다음절에서 제안할 수정된 분산추정량을 만드는데 반영된다.

2.2 관심영역의 모평균 추정

관심영역의 모평균 추정량은 모총계와 관심영역의 크기를 추정한 후 비추정량(ratio estimator)의 형태로 주어질 수 있다. 그리고 비추정량은 선형화를 통해 근사가 가능하다.

$$
\hat{\bar{Y}}^* = \frac{\hat{Y}^*}{\hat{M}^*} \approx \bar{Y}^* + \frac{1}{M^*}(\hat{Y}^* - \bar{Y}^* \hat{M}^*). \tag{2.10}
$$

여기에서 M^*은 관심영역의 크기이고 $\hat{Y}^*$은 (2.6)에 주어진 모총계 추정량이며, $\hat{M}^*$은 관심영역 크기의 불편추정량으로 (2.6)의 식과 유사하게 주어진다.

$$
\hat{M}^* = \sum_{h=1}^{L} \frac{M_h}{n_h} \sum_{i \in s_h} \frac{m_{hi}^*}{m_{hi}} \tag{2.11}
$$

식 (2.10)에 주어진 관심영역의 모평균 추정량은 편의추정량이지만 근사적으로 불편추정량이다. 그리고 식 (2.10)의 우변을 이용하면 모평균 추정량의 근사분산을 유도할 수 있다.

정리 2.2 정리 2.1과 동일한 표본추출확률분포에서, 식 (2.10)에 주어진 관심영역의 모평균 추정량, $\hat{\bar{Y}}^*$, 의 조건부 근사분산은 아래와 같다.

$$Var\{\hat{\bar{Y}}^* \mid \mathbf{m}^*\} \approx \frac{1}{M^{*2}} \sum_{h=1}^{L} A_h^2 \left[\frac{1}{2n_h^*} \sum_{i=1}^{N_h^*} \sum_{j \neq i}^{N_h^*} M_{hi} M_{hj} (Z_{hi}^* - Z_{hj}^*)^2 + \sum_{i=1}^{N_h^*} \frac{B_{hi}^2 M_{hi}^{*2} \sigma_{hi}^2}{\pi_{hi}^*} \right] \tag{2.12}$$

여기에서 $Z_{hi}^* = (m_{hi}^*/m_{hi})(\overline{Y_{hi}^*} - \overline{Y_h^*})$이다. ■

증명. 식 (2.10)을 이용하면 모평균 추정량의 분산은 다음과 같이 근사된다.

$$\begin{aligned} Var\{\hat{\bar{Y}}^* \mid \mathbf{m}^*\} &\approx \frac{1}{M^{*2}} Var\{\hat{Y}^* - \overline{Y}^* \hat{M}^* \mid m_T^*\} \\ &= \frac{1}{M^{*2}} [\, Var\{\hat{Y}^* \mid m_T^*\} + \overline{Y}^{*2} Var\{\hat{M}^* \mid m_T^*\} - 2\overline{Y}^* Cov\{\hat{Y}^*, \hat{M}^* \mid m_T^*\}] \end{aligned} \tag{2.13}$$

그리고 주어진 표본추출확률분포에서 $\hat{Y}^*$와 $\hat{M}^*$의 분산 및 공분산은 다음과 같다.

$$\begin{aligned} Var\{\hat{Y}^* \mid \mathbf{m}^*\} &= \sum_{h=1}^{L} A_h^2 \left[\frac{1}{2n_h^*} \sum_{i=1}^{N_h^*} \sum_{j \neq i}^{N_h^*} M_{hi} M_{hj} \left(\frac{m_{hi}^*}{m_{hi}} \overline{Y_{hi}^*} - \frac{m_{hj}^*}{m_{hj}} \overline{Y_{hj}^*} \right)^2 + \sum_{i=1}^{N_h^*} \frac{B_{hi}^2 M_{hi}^{*2} \sigma_{hi}^2}{n_h^* b_{hi}^*} \right] \\ Var\{\hat{M}^* \mid \mathbf{m}^*\} &= \sum_{h=1}^{L} A_h^2 \left[\frac{1}{2n_h^*} \sum_{i=1}^{N_h^*} \sum_{j \neq i}^{N_h^*} M_{hi} M_{hj} \left(\frac{m_{hi}^*}{m_{hi}} - \frac{m_{hj}^*}{m_{hj}} \right)^2 \right] \\ Cov\{\hat{Y}^*, \hat{M}^* \mid \mathbf{m}^*\} &= \sum_{h=1}^{L} A_h^2 \left[\frac{1}{2n_h^*} \sum_{i=1}^{N_h^*} \sum_{j \neq i}^{N_h^*} M_{hi} M_{hj} \left(\frac{m_{hi}^*}{m_{hi}} \overline{Y_{hi}^*} - \frac{m_{hj}^*}{m_{hj}} \overline{Y_{hj}^*} \right) \left(\frac{m_{hi}^*}{m_{hi}} - \frac{m_{hj}^*}{m_{hj}} \right) \right] \end{aligned} \tag{2.14}$$

따라서 (2.14)의 세 식을 (2.13)에 대입하면 (2.12)의 결과를 얻을 수 있다. ■

모평균 추정량의 분산추정량은 식 (2.13)에서 각각의 항에 불편추정량을 대입하여 구할 수 있다. 각각의 항에 불편추정량을 대입한 후 얻어진 분산추정량은 다음과 같이 간단하게 표현된다.

$$v(\hat{\bar{Y}}^* \mid \mathbf{m}^*) = \frac{1}{\hat{M}^{*2}} \sum_{h=1}^{L} \frac{M_h^2}{2n_h^2(n_h^* - 1)} \sum_{i \in s_h^*} \sum_{j \in s_h^*} (z_{hi}^* - z_{hj}^*)^2 \tag{2.15}$$

여기에서 $z_{hi}^* = (m_{hi}^*/m_{hi})(\bar{y}_{hi}^* - \hat{\bar{Y}}^*)$이다. 그리고 m_{hi}^*가 비슷한 경우에 분산추정량은 다음의 식으로 근사된다.

$$v(\widehat{Y^*}\mid \mathbf{m}^*) \approx \frac{1}{\widehat{M^{*2}}} \sum_{h=1}^{L} M_h^2 \left(\frac{m_h^*}{m_h}\right)^2 \frac{s_h^{2*}}{m_h^*}[(\overline{m_h^*}-1)\tau_h^*+1] \tag{2.16}$$

3. 랜덤표본에 근거한 모평균 추정량의 분산추정

관심영역의 모평균 추정에서 사후층화된 복합표본을 랜덤표본으로 간주하는 경우는 두 가지이다. 다항목 표본조사에서는 여러 항목을 조사하고 추론에 이용되는 변수들이 많기 때문에 층내에서 SSU의 표본추출확률을 동일하게 하는 경우가 많다. 만일 s_{hi}의 크기를 m으로 동일하게 하고 PSU를 집락크기에 비례하도록 표본을 선정하면, SSU의 포함확률은 층내에서 모두 동일해진다.

$$\pi_{hij} = n_h \frac{M_{hi}}{M_h} \times \frac{m}{M_{hi}} = \frac{n_h m}{M_h} \tag{3.1}$$

이렇게 선정된 복합표본은 랜덤표본으로 간주되어 다양한 자료분석에 이용된다. 또 다른 하나는 표본추출과정에 대한 보조정보, 즉 표본추출확률, 집락 크기 등, 을 가지고 있지 못하여 랜덤표본으로 간주할 수 밖에 없는 경우이다. 어느 경우든 복합표본을 랜덤표본으로 간주하고 분석을 하는 경우는 흔히 볼 수 있는 현상이다.

랜덤표본으로 간주하는 경우 관심영역에서의 모평균 추정량은 자기가중(self-weighting) 평균이 된다.

$$\bar{y}^* = \sum_{h=1}^{L} \left(\frac{m_h^*}{m_T^*}\right) \frac{1}{m_h^*} \sum_{i \in s_h} \sum_{j \in s_{hi}} y_{hij}, \tag{3.2}$$

여기에서 m_T^*는 관심영역에서 선정된 SSU의 총수이다. 그리고 랜덤표본에 기초한 가중평균의 조건부 불편분산추정량은 다음과 같이 된다.

$$v(\bar{y}^*\mid \mathbf{m}^*) = \sum_{h=1}^{L} \left(\frac{m_h^*}{m_T^*}\right)^2 \frac{s_h^{2*}}{m_h^*}. \tag{3.3}$$

집락화를 통하여 선정된 복합표본은 랜덤표본에 비해서 고르게 분포하지 않는다. 만일 관심영역의 급내상관계수가 클 경우에는 복합표본에 근거한 모평균 추정량은 랜덤표본에 기초한 그것보다 분산이 크다. 따라서 복합표본을 랜덤표본으로 간주할 경우 급내상관계수가 큰 값을 가지면 (3.3)에 제시된 분산추정량은 실제 분산을 과소 추정하게 된다. 즉, (3.3)에 제시된 분산추정량은 사용이 간편한 반면 급내상관계수가 큰 모집단에서는 과소추정을 하는 오류를 범하게 된다. 이러한 단점은 급내상관계수를 분산추정량에 고려해 줌으로써 수정할 수 있는데, 식 (2.9)로부터 수정된 분산추정량의 형태를 유추할 수 있다.

본 연구에서 제안하는 관심영역의 모평균 추정량에 대한 수정된 분산추정량은 아래와 같다.

$$v(\bar{y}^* \mid \mathbf{m}^*) = \sum_{h=1}^{L} \left(\frac{m_h^*}{m_T^*}\right)^2 \frac{s_h^{2*}}{m_h^*} \left((\overline{m_h^*} - 1)\tau_h^* + 1\right) \tag{3.4}$$

제안된 분산추정량은 모집단의 구조를 반영하는 급내상관계수를 포함함으로써 집락화로 인한 추론의 오류를 방지하는 장점이 있다. 다음절에서는 예제를 통하여 제안된 분산추정량의 효율을 수치적으로 보인다.

4. 예 제

복합표본과 랜덤표본에 기초한 분산추정량들과 수정된 분산추정량의 효율을 수치적으로 비교해 보기 위하여 1996년 농가경제조사의 표본자료 중 일부를 이용한다. 농가경제조사 표본은 우리나라의 9개 도를 층으로 하고, 부락을 PSU로, 농가를 SSU로 설정한 후 층화2단집락추출에 의해 전국에서 선정된 3,140 농가로 구성되었으며, 본 연구의 분석에 이용된 표본농가의 수는 3,085이다. 그리고 여러 조사변수 중 본 연구에 이용된 변수는 주요변수 8개이다. (y_1:농가소득, y_2:농업소득, y_3:농업조수입, y_4:농업경영비, y_5:농업 외 소득, y_6:농가 잉여 및 손실, y_7:농가자산, y_8:농가부채). 조사된 표본농가는 5개의 사후 층으로 분류되었는데, 농업조수입 중 최대작물수입에 따라 분류되었다. 아래의 <표 4.1>에 사후층의 기준 및 표본수가 나타나 있다. 각각의 사후 층은 모두 관심영역이 되며, 표본조사 후에 층을 만들었다.

<표 4.1> 사후층 기준 및 표본수

사후층 번호	표본농가수	층 분류 기준
1	1,576	농업조수입 중 벼 수입이 최대인 농가
2	333	농업조수입 중 과수 수입이 최대인 농가
3	641	농업조수입 중 채소 수입이 최대인 농가
4	237	농업조수입 중 축산 수입이 최대인 농가
5	298	기타

효율성 비교에 이용된 분산추정량은 아래의 6개이다.

$$v_1 = \frac{1}{\widehat{M}^{*2}} \sum_{h=1}^{L} \frac{M_h^2}{2n_h^2(n_h^* - 1)} \sum_{i \in s_h^*} \sum_{j \in s_h^*} (z_{hi}^* - z_{hj}^*)^2,$$

$$v_2 = \frac{1}{\widehat{M}^{*2}} \sum_{h=1}^{L} M_h^2 \left(\frac{m_h^*}{m_h}\right)^2 \frac{s_h^{2*}}{m_h^*} [(\overline{m^*} - 1)\tau^* + 1],$$

$$v_3 = \frac{s^{2*}}{m^*}, \qquad v_4 = \frac{s^{2*}}{m^*}[(\overline{m^*} - 1)\tau^* + 1],$$

$$v_5 = \sum_{h=1}^{L} \left(\frac{m_h^*}{m^*}\right)^2 \frac{s_h^{2*}}{m_h^*}, \qquad v_6 = \sum_{h=1}^{L} \left(\frac{m_h^*}{m^*}\right)^2 \frac{s_h^{2*}}{m_h^*}[(\overline{m^*} - 1)\tau^* + 1].$$

여기서 v_3과 v_4는 관심영역의 표본을 층을 구분하지 않고 랜덤표본으로 간주한 경우의 분산 추정량이고, v_2, v_4, v_6에 포함된 급내상관계수 τ^*는 관심영역의 표본전체에서 구한 값인데, τ_h^*대신 τ^*을 이용한 이유는 관심영역내의 각 층에서 구한 τ_h^*는 표본수가 적어서 그 값의 변동이 심하기 때문이다. 각 분산추정량의 효율성은 v_1을 기준으로 분산의 비, $v_k/v_1, (k=2,\ldots,6)$,를 구하여 비교하였다. <표 4.2>는 비교의 결과를 보여준다.

5개 분산추정량($v_2,\ldots,v_6$)은 모두 대체적으로 과소추정의 경향을 보이고 있다. 그 중 v_1 가장 근접한 값을 보이는 것은 v_2로서 약 88% 정도이며, 다음이 v_4, v_6로 약 87%, 82% 정도이다. 랜덤표본으로 간주한 경우의 분산추정량인 v_3, v_5은 64%, 61%로서 과소추정의 정도가 심하여 심각한 추론의 오류를 내포한다. v_2에서 12%의 과소추정은 사후층화된 복합표본의 크기를 동일하다고 가정하여 생긴 손실이며, v_3, v_5에서 36%, 39%의 과소추정은 복합표본을 랜덤표본으로 간주하여 발생한 손실이다. 본 연구에서 제안한 수정된 분산추정량의 경우 13%, 18%의 과소추정을 보여 v_3, v_5에 비해 상당부분 오류를 방지함을 보여준다.

5. 결 론

대부분의 표본조사는 랜덤표본보다는 복합표본을 이용한다. 따라서 표본조사 자료를 이용한 통계적 추론은 복합표본에 근거한 추론이 되어야 한다. 많은 경우에 복합표본을 랜덤표본으로 간주하고 통계 추론을 하는데, 만일 복합표본을 랜덤표본으로 간주하여 발생하는 손실이 크면 추론에 심각한 문제가 발생하게 된다.

본 연구에서는 관심영역의 모총계 및 모평균의 추론 문제를 다루었다. 이를 위하여 층화2단집락추출을 하여 선정된 표본을 다시 관심영역으로 사후층화하여 만든 복합표본을 추론에 이용하였다. 복합표본에 근거한 관심영역의 모평균 및 모총계 추정량을 구하였고, 더불어 분산, 분산추정량 및 분산추정량의 근사식을 유도하였다.

또한 복합표본을 랜덤표본으로 간주한 경우, 관심영역의 모평균 추정량의 분산추정량을 구하여 분산의 과소추정은 집락들의 급내상관계수와 관계가 있음을 보였고, 분산추정량에 급내상관계수를 반영한 수정된 분산추정량을 제안함으로써 분산의 과소추정 문제를 방지할 수 있는 추론의 방법을 제시하였다.

<표 4.2> 분산추정량의 효율성 비교

사후 층 번호	조 사 변 수							
	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8
	v_2/v_1							
1	0.9640	0.7765	0.8818	0.7809	0.8336	0.9627	0.9544	0.7720
2	1.3640	0.7214	0.8285	0.7756	0.8052	1.3631	0.9354	1.6718
3	0.8313	0.6908	0.8168	0.6673	0.6804	1.0422	0.6531	0.9018
4	0.9948	1.0265	0.6446	0.8601	0.7734	0.9805	0.7738	0.7911
5	0.6861	0.5916	1.0208	0.5172	0.5509	1.2324	0.7962	1.1848
	v_3/v_1							
1	0.5264	0.3764	0.4512	0.3796	0.4691	0.6997	0.5125	0.4699
2	0.9158	0.4799	0.5916	0.5317	0.6877	1.2661	0.6465	2.0364
3	0.5042	0.3528	0.5776	0.2941	0.2898	0.7470	0.5283	0.5782
4	0.8809	0.9680	0.3750	0.7794	0.6619	0.9536	0.8257	0.6803
5	0.4246	0.3452	0.7190	0.2594	0.2740	0.9849	0.7584	0.9549
	v_4/v_1							
1	1.0264	0.7559	0.9114	0.7666	0.8420	0.9209	1.2429	0.6760
2	1.4090	0.6990	0.8470	0.7998	1.0525	1.2547	1.1678	1.7778
3	0.8343	0.6813	0.8249	0.6277	0.5947	0.9337	0.9403	0.7155
4	0.9239	1.0042	0.4920	0.8383	0.7472	1.1239	1.1987	0.8215
5	0.5146	0.4487	0.7492	0.4357	0.4602	0.8892	0.9902	0.9398
	v_5/v_1							
1	0.5019	0.3671	0.4297	0.3677	0.4554	0.6894	0.4350	0.4659
2	0.8699	0.4674	0.5682	0.5024	0.6259	1.2657	0.4993	1.9889
3	0.4767	0.3379	0.5455	0.2831	0.2839	0.7223	0.4270	0.5761
4	0.8542	0.9324	0.3442	0.7611	0.6496	0.9209	0.6705	0.6805
5	0.4149	0.3400	0.7084	0.2377	0.2415	1.0088	0.5992	0.9180
	v_6/v_1							
1	0.9787	0.7372	0.8680	0.7426	0.8171	0.9073	1.0550	0.6701
2	1.3384	0.6808	0.8136	0.7557	0.9579	1.2543	0.9019	1.7363
3	0.7887	0.6526	0.7790	0.6043	0.5826	0.9028	0.7601	0.7130
4	0.8959	0.9673	0.4516	0.8187	0.7333	1.0853	0.9735	0.8218
5	0.5028	0.4418	0.7382	0.3993	0.4057	0.9108	0.7823	0.9035

참 고 문 헌

[1] Casady, R.J. and Valliant, R. (1993). Conditional properties of post-stratified estimators under normal theory, *Survey Methodology*, Vol. 19, 183-192.

[2] Cochran, W.G. (1977). *Sampling Techniques* (3rd ed.). New York, John Wiley.

[3] Djerf, K. (1997). Effects of post-stratification on the estimates of the Finnish labor force survey. *Journal of Official Statistics*, Vol. 13, 29-39.

[4] Eltinge, J.L. and Jang, D.S. (1996). Stability measures for variance component estimators under a stratified multistage design. *Survey Methodology*, Vol. 22, 157-165.

[5] Gelman, A. and Little, T.C. (1997). Poststratification into many categories using hierarchical logistic regression. *Survey Methodology*, Vol. 23, 127-135.

[6] Holt, D. and Smith, T.M.F. (1979). Post stratification. *Journal of Royall Statistical Society.* Ser. A, Vol. 142, 33-46.

[7] Jager, P. (1986). Post-stratification against bias in sampling. *International Statistical Review*, Vol. 54, 159-167.

[8] Kish, L. and Frankel, M.R. (1974). Inference from complex samples. *Journal of Royall Statistical Society*, Vol. 36, 1-37.

[9] Little, R.J.A. (1993). Post-stratification : A Modelers perspective. *Journal of the American Statistical Association*, Vol. 88, 1001-1012.

[10] Sarnsal, C.E., Swensson, B. and Wretman, J. (1992). *Model assisted survey sampling.* Springer-Verlag, New York.

[11] Valliant, R. (1987). Conditional properties of some estimators in stratified sampling. *Journal of the American Statistical Association*, Vol. 82, 509-519.

[12] Valliant, R. (1993). Poststratification and conditional variance estimation. *Journal of the American Statistical Association*, Vol. 88, 89-96.