The Korean Communications
in Statistics Vol. 12 No. 2, 2005
pp. 365-379

Variance Estimation for Imputed Survey Data using Balanced Repeated Replication Method

Junsuk Lee[1], Taekyong Hong[2], and Pyong Namkung[3]

Abstract

Balanced Repeated Replication(BRR) is widely used to estimate the variance of linear or nonlinear estimators from complex sampling surveys. Most of survey data sets include imputed missing values and treat the imputed values as observed data. But applying the standard BRR variance estimation formula for imputed data does not produce valid variance estimators. Shao, Chen and Chen(1998) proposed an adjusted BRR method by adjusting the imputed data to produce more accurate variance estimators. In this paper, another adjusted BRR method is proposed with examples of real data.

Keywords : balanced repeated replication, jackknife, hot-deck, imputation

1. 서 론

모집단 특성값 추정을 위한 표본조사 설계는 모집단 구조나 조사 방법의 다양화로 인해 단순 표본조사 설계 하에서는 보다 정확한 추정이 어렵다. 따라서 층화추출, 집락추출, 다단계 층화추출 등 다양한 표본추출 방법들을 혼용하는 복합 표본추출 설계가 많이 이용되고 있다. 이러한 복합 표본추출 설계 하에서의 분산추정 방법은 대단히 복잡하기 때문에 최근 많이 이용되고 있는 분산 추정 방법으로는 테일러 선형화 방법(Taylor linearization method)과 잭나이프 방법(Jackknife method), 균형 순환 반복법(Balanced repeated replication) 그리고 붓스트랩 방법(Bootstrap) 등과 같은 표본 재사용 방법(sample reuse method)들이 있다. 이들 방법 중 균형 이분표본(Balanced half sample) 혹은 균형 반복법(Balanced Repeated Replication method : BRR)은 다단계 층화추출에서 비선형 추정량의 분산추정 방법으로 McCarthy(1969)에 의해 제안되어 널리 사용되어 왔다.

복합 표본조사에서 분산추정을 위한 방법들이 제안되었지만 현실적으로 표본조사를 실시함에 있어 근본적인 문제점은 필연적으로 발생되는 무응답으로 인한 조사 자료의 불완전성에 있다. 따라서 조사자들은 모집단 총합이나 분산 등과 같은 관심있는 모수들의 보다 정확한 추정을 위해

1) Lecturer, Department of Statistics, Sungkyunkwan University, Seoul, 110-745, Korea
E-mail : jslee@skku.edu
2) Lecturer, Department of Statistics, Sungkyunkwan University, Seoul, 110-745, Korea
3) Professor, Department of Statistics, Sungkyunkwan University, Seoul, 110-745, Korea

대체(imputation) 방법을 이용하여 완전자료를 구성하려 노력한다. 그러나 대체된 자료는 실제 자료가 아니므로 분산추정에 있어서 구조적 편향(structural bias)을 가질 수 있고 추정량의 분산이 상당히 과소 추정되는 경향이 있다. 그러므로 대체된 자료를 이용한 분산추정 문제는 최근에 많은 연구가 이루어지고 있으며, 복합 표본조사에서 보다 좋은 분산추정량을 구하는 방법들이 제안되고 있다. 즉, Rubin(1978)의 다중대체(multiple imputation), Rao와 Shao(1992)의 조정된 jackknife 방법, Shao와 Sitter(1996)는 bootstrap 방법 등 통계량의 분포와 관계없이 적용될 수 있는 방법들이 제안되었지만 jackknife 방법에 비해 계산량이 많다는 단점이 있다.

분산추정 시 계산 상 jackknife 방법과 비슷한 계산량과 사분위수와 같은 비평활 통계량의 일치 분산추정량을 제공하는 방법이 BRR 방법이다. 그러나 다른 방법들과 마찬가지로 일반적인 BRR 방법도 대체로 인한 분산의 증가를 고려하지 않기 때문에 일치 추정량을 제공하지 못한다. 이러한 문제의 해결을 위해 BRR방법의 적절한 조정이 필요하여 Shao, Chen, Chen(1998)은 대체된 자료의 일치 분산추정량을 제공할 수 있는 조정된 BRR방법을 제안하였다.

본 논문에서는 분산추정 시 대체된 자료의 새로운 조정방법을 제안하고, 실제 자료를 사용한 모의실험을 통해 새로운 조정방법과 표준 BRR방법, 조정된 BRR방법 그리고 jackknife 방법들과의 상대 편향을 비교해 보고자 한다.

2. 결측값이 발생한 경우 대체 방법

표본조사에서 얻어진 자료는 조사 항목에 빈 칸이 없는 완전한 장방형의 형태로 구해지지 못하는 것이 보통이다. 즉, 무응답이라 불리는 결측값이 생기게 된다. 무응답을 적절한 다른 값으로 대체한 완전자료로부터 만들어진 추정량의 분산은 표본추출 변동과 대체변동을 모두 포함하고 있기 때문에 무응답이 없는 경우의 분산보다 일반적으로 크게 나타난다. 따라서 무응답의 대체값을 실제 관측값으로 간주하고 일반적인 분산추정을 하는 경우 대체분산이 고려되지 않기 때문에 실제 분산이 과소 추정되어 통계적 추론의 과오를 범하게 된다.

2.1 단위 무응답의 경우

단위 무응답의 경우 크기 N의 모집단으로부터 크기 n의 표본을 비복원 단순임의추출하여 모집단 총합의 분산을 추정하는 문제를 고려할 때, 단위 무응답의 발생으로 크기 $n^{(r)}$의 표본만이 구성되었다면 $n^{(r)}$은 분명히 n보다 작을 것이다.

따라서 모집단 총합 추정량 $\hat{Y}$의 분산은

$$V_{srs}(\hat{Y}) = N^2(1-n/N)S^2/n$$

이므로 표본을 응답자 층과 무응답자 층으로 나누어 각각의 크기를 N_1, N_2 그리고 각각의 층 가중값을 $W_1 = N_1/N$, $W_2 = N_2/N$로 하면

$$E(\hat{Y}) - Y = N\bar{Y}_1 - (N_1\bar{Y}_1 + N_2\bar{Y}_2) = N_2(\bar{Y}_1 - \bar{Y}_2) = Bias(\hat{Y})$$

만큼의 편향(bias)이 발생하게 된다. 이런 단위 무응답에 의한 편향은 재가중 방법이나 사후층화 방법 그리고 재조사 등을 통해 조정할 수 있다.

2.2 항목무응답의 경우

2.2.1 평균 대체 방법

이 방법은 무응답이 발생한 칸에 동일 문항의 응답자들의 평균값을 대체하는 방법이다. 응답자들의 평균값을 $\bar{y}_r=\frac{1}{r}\sum_{k\in A_r} y_k$라 하면 평균대체 방법을 이용한 완전자료는 다음과 같다.

$$y_k^*=\begin{cases} y_k\,, & k\in A_r \quad ; \text{응답값} \\ \bar{y}_r\,, & k\in A_{s-r} \; ; \text{평균대체값} \end{cases}$$

여기서 A_s는 크기 n인 표본, A_r은 크기 r인 응답자 집합을 의미한다.
평균대체의 경우 대체 후 표본평균은 응답자의 평균값과 동일하게 나타난다.

$$\bar{y}_I=\frac{1}{n}\left(\sum_{k\in A_r} y_k+\sum_{k\in A_{s-r}}\bar{y}_r\right)=\bar{y}_r \tag{2.1}$$

2.2.2 비 대체 방법

비 대체 방법은 관심변수 y와 상관관계가 높은 보조변수 x가 있을 때 유용한 방법으로 무응답 항목 k에 비추정값 $\left(\frac{\bar{y}_r}{\bar{x}_r}\right)x_k$를 대체하는 방법이다.

$$y_k^*=\begin{cases} y_k\,, & k\in A_r \quad ; \text{응답값} \\ \left(\frac{\bar{y}_r}{\bar{x}_r}\right)x_k\,, & k\in A_{s-r} \; ; \text{평균대체값} \end{cases}$$

비 대체 후 표본평균은 비추정량이 된다. 즉,

$$\bar{y}_I=\frac{1}{n}\left\{\sum_{k\in A_r} y_k+\sum_{k\in A_{s-r}}\left(\frac{\bar{y}_r}{\bar{x}_r}\right)x_k\right\}. \tag{2.2}$$

와 같이 된다.

2.2.3 핫덱 대체 방법

핫덱(hot deck)대체 방법은 무응답이 발생한 항목에 응답자 중의 한 응답값을 랜덤하게 선정하여 대체하는 방법이다. 즉,

$$y_k^* = \begin{cases} y_k, & k \in A_r \\ y_j,\ A_r\text{에서 랜덤하게 선정된 값}, & k \in A_{s-r} \end{cases}$$

이 되며 핫덱 대체 후 표본평균은 식(2.3)과 같다.

$$\bar{y}_I = \frac{1}{n}\left\{ \sum_{i \in A_r} y_i + \sum_{i \in A_{s-r}} y_i^* \right\} = \frac{1}{n}\left(r\bar{y}_r + (n-r)\,\bar{y}_m^*\right). \tag{2.3}$$

여기서 $\bar{y}_m^*$은 대체된 값들의 평균이다.

2.2.4 다중 대체

다중 대체(multiple imputation)방법은 단일대체 방법들이 대체 후 분산을 과소추정하게 되는 문제를 개선하기 위해 제안된 방법이다. 이 방법은 각 결측값에 대한 대체값을 2개 이상의 가능한 벡터(vector)로 대체하는 방법으로 한 개의 결측값에 대한 대체값으로 가능한 여러 가지 경우에 따라 만들어지는 대체값들을 대상으로 결측값을 구하는 방법이다.

Rubin(1978)은 대체로 인한 분산의 변동을 고려한 다중 대체를 제안했으며 이 과정은 동일한 대체 절차를 통한 M개의 대체값으로 각각의 결측값을 대체한 $M(\geq 2)$개의 완전한 자료 셋을 필요로 한다. $\bar{y}_{I1}, \cdots, \bar{y}_{IM}$을 단순임의추출 하에서 모집단 평균 $\bar{Y}$의 M개의 대체 추정량이라 하면 $\bar{Y}$의 최종 대체 추정량은 $\bar{y}_{I.} = \sum_{l=1}^{M} \bar{y}_{Il}/M$이 되고, 분산추정량은

$$Var(\bar{y}_{I.}) = \frac{1}{M}\left(\frac{1}{n} - \frac{1}{N}\right)S_{I1}^2 + \frac{M+1}{M}\left\{\frac{1}{M-1}\sum_{l=1}^{M}(\bar{y}_{Il} - \bar{y}_{I.})^2\right\} \tag{2.4}$$

이다. 여기서 s_{I1}^2은 l번째 자료의 표본분산이다. 또한 우변의 첫 번째 항인 $\frac{1}{M}\left(\frac{1}{n} - \frac{1}{N}\right)S_{I1}^2$은 추정값과 결합된 대체 내 분산이고, 두 번째 항의 $\frac{1}{M-1}\sum_{l=1}^{M}(\bar{y}_{Il} - \bar{y}_{I.})^2$는 대체 간 분산이며 $\frac{M+1}{M}$은 많은 대체값들 중에서 유한개만을 사용함으로써 발생하는 오차를 줄이기 위한 조정값이다.

3. 복합 표본조사에서 분산추정 방법

최근 사용되고 있는 복합 표본조사에서의 분산 추정방법들은 아래와 같다.

3.1 테일러 선형화 방법

이 방법은 테일러 급수 전개(Taylor series expansion)에 의하여 선형추정량(linear estimator)으로 근사한 후 분산을 계산하는 방법이다.

$t=(t_1, t_2, \cdots, t_k)$를 추정량들의 집합이라 하고 그 기댓값들의 집합을 $T=(T_1, T_2, \cdots, T_k)$라 하자. 만일 추정될 함수 $\theta=F(T)$가 $\hat{\theta}=F(t)$에 의해 추정된다고 하면 테일러 급수 전개식의 첫 번째 항만을 이용하면 근사적으로 아래의 식이 성립한다.

$$\hat{\theta}=F(t)-F(T)=\sum_{i=1}^{k}(t_i-T_i)\frac{\partial F}{\partial T_i}$$

그리고 $\hat{\theta}$의 분산은 근사적으로 선형함수 $\sum_{i=1}^{k}(t_i-T_i)\frac{\partial F}{\partial T_i}$의 분산에 의해 다음과 같이 표현된다.

$$V(\hat{\theta})=\sum_{i=1}^{k}\left(\frac{\partial F}{\partial T_i}\right)^2 V(t_i)+\sum_{i\neq j}\frac{\partial F}{\partial T_i}\frac{\partial F}{\partial T_j}Cov(t_i, t_j) \tag{3.1}$$

3.2 균형 순환 반복법

균형 순환 반복법은 McCarthy(1969)에 의해 최초로 제안된 방법으로 각 층에서 두 개의 1차 단위를 선택하는 층화추출에서의 분산추정을 위해 개발되었다. 즉, 각 층에서 두 개의 1차 단위들은 동일한 확률로 선택되며 이분표본은 각 층으로부터 한 개의 1차 단위가 랜덤하게 선택됨으로써 얻어진다.

$\hat{\theta}_H$ $\hat{\theta}_c$ 그리고 $\hat{\theta}_s$를 각각 선택된 이분표본에서 나머지 이분표본에서 그리고 전체표본에서 계산된 추정량이라 하자. 단순한 선형추정량으로

$$(\hat{\theta}_H-\hat{\theta}_s)^2=(\hat{\theta}_c-\hat{\theta}_s)^2=\frac{1}{4}(\hat{\theta}_H-\hat{\theta}_c)^2$$

는 유한모집단 수정계수(finite population correction : fpc)가 무시된다면 모두 $V(\hat{\theta})$의 불편추정량이 된다. 여기서 오직 두 개의 이분표본만을 사용할 경우 분산추정량은 신뢰할 수 없지만 반복수를 늘려가면서 정도(precision)를 높일 수 있다.

3.3 Jackknife 방법

단순임의추출 하에서 $\hat{\theta}_{(j)}$를 $\hat{\theta}$와 동일한 형태의 추정량이라 하자. 여기서, $\hat{\theta}_{(j)}$는 j번째 관측값이 제외된 상태에서의 추정량이다. 따라서 만일 $\hat{\theta}=\bar{y}$라면 $\hat{\theta}_{(j)}=\bar{y}_{(j)}=\sum_{i\neq j}y_i/(n-1)$이 된다. 그러면 jackknife 분산추정량은 다음과 같은 형태가 된다.

$$\hat{V}_J(\hat{\theta})=\frac{n-1}{n}\sum_{j=1}^{n}(\hat{\theta}_{(j)}-\hat{\theta})^2$$

층화 다단 집락추출의 경우 H개의 층이 있고, 층 h에서 n_h개의 1차 추출단위가 선택된다고 가정하자. Jackknife 방법의 적용을 위해 한 번에 한 개의 1차 추출단위를 제거한다. $\hat{\theta}_{(hj)}$를 층 h내의 j번째 1차 추출단위가 제거된 상태에서의 추정량이라 하면 분산추정량은 식 (3.2)와 같다.

$$\hat{V}_J(\hat{\theta}) = \sum_{h=1}^{H} \frac{n_h - 1}{n_h} \sum_{j=1}^{n_h} (\hat{\theta}_{(hj)} - \hat{\theta})^2 \tag{3.2}$$

3.4 Bootstrap 방법

알려져 있지 않은 확률분포 F로부터 독립적으로 n개의 관측값 $y_1, y_2, \cdots, y_n$을 뽑는 경우 각각의 값 y_i ($i=1, 2, \cdots, n$)들은 동일한 가중값 $1/n$을 갖는다. 이 추정된 분포로부터 표본크기 $n_1(<n)$의 m개의 bootstrap 표본을 뽑는다. 그러면 전체 표본에 기초한 $\hat{\theta}$의 분산추정량은 식 (3.3)과 같다.

$$\hat{V}(\hat{\theta}) = \sum_{k=1}^{m} (\hat{\theta}_k - \hat{\theta})^2/(m-1) \tag{3.3}$$

여기서 $\hat{\theta}_k$는 k번째 bootstrap 표본으로부터 계산된 추정값이다.

Boostrap 방법의 장점은 일반적인 표본추출 설계에서 사분위수 등과 같은 비평활 함수에의 적용이 가능하다.

4. 새로운 조정 방법 제안

본 논문에서는 대체된 값들이 있는 경우 BRR 방법과 조정된 BRR 방법을 이용하여 분산추정을 하는 방법과 새로운 대체값의 조정 방법을 제안하고자 한다.

4.1 결정적 대체

결정적 대체(deterministic imputation)는 대체되는 값들이 어떤 특정한 값으로 결정되어져 있다는 의미로 대체되는 값들이 랜덤하게 선택되는 랜덤 대체와 상반되는 개념이다.

4.1.1 평균 대체

평균 대체에 있어서 조정된 BRR방법은 대체된 값들의 조정을 통해 대체분산을 유도할 수 있다.

$$\tilde{y}(\alpha_t)_{hij} = \begin{cases} \tilde{y}_{hij} + E(\alpha_t)_{A_k}(\tilde{y}_{hij}) - E_{A_k}(\tilde{y}_{hij}) & ; \ y_{hij}\text{가 대체된값인 경우} \\ y_{hij} & ; \ y_{hij}\text{가관측값인 경우} \end{cases}$$

즉, 원래의 대체된 값에 t이분표본 내의 자료만을 이용한 기대값과 t 반복 표본 내의 자료를 모두 이용한 기대값의 차이를 가중함으로써 대체값을 조정하고 있다.

본 논문에서는 대체된 값이 있는 경우 다음과 같은 재조정된 방법을 제안하려 한다.

$$\tilde{y}(\alpha_t)_{hij}=\begin{cases}\tilde{y}_{hij}+\{E(\alpha_t)_{A_k}(\tilde{y}_{hij})+E(\alpha_t)^c{}_{A_k}(\tilde{y}_{hij})\}-2E_{A_k}(\tilde{y}_{hij}) & ; y_{hij}\text{가 대체된값인 경우}\\ y_{hij} & ; y_{hij}\text{가관측값인 경우}\end{cases} \tag{4.1}$$

여기서 $E(\alpha_t)^c$는 선택된 이분표본을 제외한 나머지 여이분표본 내의 자료를 이용한 기대값이다. 이러한 시도는 Risto와 Erkki(1994)이 제안한 BRR에서의 7가지 분산추정식 중에서 이분표본과 여이분표본을 이용하여 각각의 분산을 구한 뒤 두 분산을 더하여 둘로 나눈 추정식이 상당히 정확한 추정을 함을 보였다.

4.1.2 중위수 대체

결정적 대체에 있어서 또 하나의 방법은 특이값들을 포함한 자료의 경우에 중위수의 평균을 이용한 조정방법으로 조정식은 다음과 같다.

$$\tilde{y}(\alpha_t)_{hij}=\begin{cases}\tilde{y}_{hij}+M(\alpha_t)_{A_k}(\tilde{y}_{hij})-M_{A_k}(\tilde{y}_{hij}) & :y_{hij}\text{가 대체된값일때}\\ y_{hij} & :y_{hij}\text{가관측값일때}\end{cases} \tag{4.2}$$

여기서 $M(\alpha_t)_{A_k}$는 이분표본 내의 자료들의 중위수의 평균이며 M_{A_k}는 전체 자료들의 중위수들의 평균이다.

4.2 랜덤 대체

랜덤 대체는 대체값들이 정해진 특정한 값으로 대체되는 것과 달리 결측이 있는 대체층 내에서 랜덤한 값으로 선택되어 대체되는 경우를 의미한다. 대표적인 랜덤 대체로 랜덤 비 대체와 가중 핫덱 대체의 경우를 살펴보자.

4.2.1 랜덤 비 대체

랜덤 비 대체에 있어서 Shao와 Chen 그리고 Chen(1998)은 다음과 같은 조정식을 제안하였다.

$$\tilde{y}(\alpha)_{hij}=\begin{cases}\tilde{y}_{hij}+\left(\dfrac{\sum_{A_k}w(\alpha_t)_{hij}y_{hij}}{\sum_{A_k}w(\alpha_t)_{hij}x_{hij}}-\dfrac{\sum_{A_k}w_{hij}y_{hij}}{\sum_{A_k}w_{hij}x_{hij}}\right)x_{hij} & :y_{hij}\text{가 대체된 값일 때}\\ y_{hij} & :y_{hij}\text{가 관측값일 때}\end{cases} \tag{4.3}$$

즉, 이분표본 내의 값들의 가중된 비에서 전체표본의 가중된 비의 차이로 조정을 하고 있다. 이러한 랜덤 비 대체의 경우 본 논문에서는 다음과 같은 조정식을 사용하였다. 새로운 조정식의 기본 개념은 평균대체나 중위수 대체의 경우와 동일하다.

$$\tilde{y}(\alpha_t)_{hij}=\begin{cases}\tilde{y}_{hij}+\left\{\left(\dfrac{\sum_{A_k} w(\alpha_t)_{hij} y_{hij}}{\sum_{A_k} w(\alpha_t)_{hij} x_{hij}}+\dfrac{\sum_{A_k} w(\alpha_t)^{c}_{hij} y_{hij}}{\sum_{A_k} w(\alpha_t)^{c}_{hij} x_{hij}}\right)-2\left(\dfrac{\sum_{A_k} w_{hij} y_{hij}}{\sum_{A_k} w_{hij} x_{hij}}\right)\right\} x_{hij} \\ \quad : y_{hij} \text{ 가 대체된 값인 경우} \\ y_{hij} : y_{hij} \text{ 가 관측값인 경우}\end{cases} \tag{4.4}$$

4.2.2 가중 핫덱 대체

가중 핫덱(weighted hot-deck)대체 방법은 무응답이 발생한 층에서 대체값을 응답값 중에서 층 가중값을 고려하여 랜덤하게 선택하는 방법으로 조정식은 다음과 같다(Shao와 Chen 그리고 Chen(1998)).

$$\tilde{y}(\alpha_t)_{hij}=\begin{cases}\tilde{y}_{hij}+\dfrac{\sum_{A_k} w(\alpha_t)_{hij} y_{hij}}{\sum_{A_k} w(\alpha_t)_{hij}}-\dfrac{\sum_{A_k} w_{hij} y_{hij}}{\sum_{A_k} w_{hij}} & : y_{hij} \text{ 가대체된 값인 경우} \\ y_{hij} & : y_{hij} \text{ 가 관측값인 경우}\end{cases} \tag{4.4}$$

본 논문에서는 다음과 같은 조정식을 사용하였다.

$$\tilde{y}(\alpha_t)_{hij}=\begin{cases}\tilde{y}_{hij}+\left(\dfrac{\sum_{A_k} w(\alpha_t)_{hij} y_{hij}}{\sum_{A_k} w(\alpha_t)_{hij}}+\dfrac{\sum_{A_k} w(\alpha_t)^{c}_{hij} y_{hij}}{\sum_{A_k} w(\alpha_t)^{c}_{hij}}\right)-2\dfrac{\sum_{A_k} w_{hij} y_{hij}}{\sum_{A_k} w_{hij}} \\ y_{hij}\end{cases} \tag{4.5}$$

5. 실제 자료를 이용한 분산추정 방법들 간의 비교

대체된 값들이 있는 경우 새로운 대체값의 조정 방법을 이용한 모의실험을 통해 그 결과를 비교해 보고자 한다. 모의실험을 위한 자료로는 통계청 2001 농업 총조사 자료 중 일부를 발췌하여 사용한 실제자료이며, 관심변수로는 영농 기간을 보조변수로는 영농 기간과 상관관계가 높을 것으로 기대되는 영농자 나이를 선택하였다. 전체 자료 중에서 랜덤하게 2,000명을 선택하여 6개의 층으로 구분하였다. 각 층은 영농 기간에 따라 동일한 크기로 구성하여 동일한 층가중값을 주었고 이분표본에서 각 층마다 두 개씩의 1차추출단위를 선택하였다. 실제 두 변수의 상관계수는 0.799

이다. 이 자료는 무응답이 없는 완전자료이기 때문에 BRR의 매 반복에서 인위적으로 랜덤하게 결측값을 생성시켜 각 대체 방법을 이용하여 대체하였다. 또한 각 추정 분산의 상대편향(relative bias)을 구하여 무응답율의 증가에 따른 각 조정 방법들을 비교하였으며, 각 방법들의 상대효율(relative efficiency)을 보기 위해 평균제곱오차(MSE)의 비로 비교하였다.

5.1 평균대체

모의실험은 자료에 결측이 없을 경우와 각각 5%, 10%, 20%, 30%, 50%의 결측값들이 있을 경우 일반적인 BRR 방법, 조정된 BRR 방법 그리고 새로 제안한 재조정된 BRR 방법을 이용하여 실시하였다. 각각의 경우에 있어서 분산을 추정한 결과들이 <표 5.1>에 나타나 있다.

무응답이 없는 완전자료의 경우 분산의 참값은 234.7885이다.

<표 5. 1> 평균대체에서 조정 후 분산추정 결과

무응답률	평균대체	표준BRR	조정된BRR	재조정된BRR
5%	220.9892	232.2717	226.9484	226.9937
10%	212.5352	230.1309	224.1099	224.0416
20%	187.1234	222.3651	209.1848	209.3825
30%	164.0715	199.8041	191.3785	191.4473
50%	113.9678	170.5872	150.7599	150.5349

<표 5. 1>의 평균대체에서 조정후 분산추정 결과를 보면 평균대체의 경우는 예상대로 응답률이 저하될수록 무응답이 없는 경우에 비해 분산의 과소 추정 정도가 심해짐을 볼 수 있고, 표준 BRR의 경우와 조정된 BRR 그리고 재조정된 BRR의 경우 모두 무응답률이 증가할수록 미세하게 분산을 과소 추정 하다가 무응답률 50%에서는 과소추정의 정도가 급격하게 심해졌다.

또한 각 조정 방법들에 대한 상대편향은 <표 5. 2>와 같다.

<표 5. 2> 각 무응답률에 따른 조정 방법들의 상대편향

무응답률	평균대체	표준BRR	조정된BRR	재조정된BRR
5%	-0.0588	-0.0588	-0.0107	-0.0344
10%	-0.0948	-0.0948	-0.0198	-0.0455
20%	-0.2030	-0.2030	-0.0529	-0.1090
30%	-0.3012	-0.3012	-0.1490	-0.1849
50%	-0.5146	-0.5146	-0.2734	-0.3579

상대편향을 통해 보면 조정된 BRR의 경우와 재조정된 BRR의 경우 거의 차이를 보이지 않고 있으며 조정을 거치지 않은 표준 BRR의 경우에는 무응답률이 높아질수록 상대편향이 급격히 커짐을 볼 수 있다.

5.2 중위수 대체

같은 자료를 이용한 모의실험 결과가 <표 5. 3>에 나타나 있다.

<표 5. 3> 중위수 대체에서 조정 후 분산추정 결과

무응답률	중위수대체	표준BRR	조정된BRR	재조정된BRR
5%	221.2163	232.3071	226.9620	226.9132
10%	212.9716	229.8546	224.5630	224.0467
20%	187.9868	221.8971	209.4756	209.1364
30%	165.0395	199.1550	192.0500	192.0509
50%	115.0499	170.2395	150.7863	151.1558

중위수 대체의 경우도 평균대체의 경우와 마찬가지로 분산이 과소 추정되고 있음을 볼 수 있는데 표준 BRR, 조정된 BRR 그리고 재조정된 BRR 세 방법 모두 유사한 수준으로 분산의 과소추정이 나타났다.

각 조정 방법들에 대한 상대편향은 <표 5. 4>와 같다.

<표 5. 4> 각 무응답률에 따른 조정 방법들의 상대편향

무응답률	중위수대체	표준BRR	조정된BRR	재조정된BRR
5%	-0.0578	-0.0578	-0.0106	-0.0333
10%	-0.0929	-0.0929	-0.0210	-0.0436
20%	-0.1993	-0.1993	-0.0549	-0.1078
30%	-0.2971	-0.2971	-0.1518	-0.1820
50%	-0.5100	-0.5100	-0.2749	-0.3578

5.3 랜덤 비 대체

랜덤 비 대체의 경우 조정 후 분산을 추정한 결과는 <표 5. 5>과 같다.

<표 5. 5> 비 대체에서 조정 후 분산추정 결과

무응답률	비 대체	표준BRR	조정된BRR	재조정된BRR
5%	222.7983	243.9652	227.5318	227.8151
10%	216.0543	267.7275	225.4126	225.8049
20%	194.4969	283.3811	212.8106	212.6932
30%	175.1411	310.0743	197.7344	197.4461
50%	133.0885	339.1769	164.0948	163.7697

<표 5. 5>의 결과를 보면 랜덤 비 대체의 경우에는 대체값을 조정하지 않은 경우가 분산의 과소추정 정도가 심하고 표준 BRR 방법은 대체로 인한 분산을 과대추정하는 경향이 있으며 마지막 열의 재조정된 BRR의 경우에서는 조정된 BRR 방법과 유사하게 모분산을 과소추정 해주고 있다..

각 방법들의 상대편향은 <표 5. 6>에 나타나 있다. 상대편향을 비교해 보아도 비 대체의 경우 무응답률 증가에 따라 심한 과소편의를 나타내고 있고 반대로 표준 BRR의 경우 상당한 과대추정의 결과를 보이고 있다. 재조정된 BRR의 경우 다른 방법에 비해 가장 작은 상대편향을 보이고 있는데 무응답률 50%에서 상대편향이 갑자기 증가함을 볼 수 있다.

<표 5. 6> 각 무응답률에 따른 조정 방법들의 상대편향

무응답률	비 대체	표준BRR	조정된BRR	재조정된BRR
5%	-0.0511	-0.0511	0.0391	-0.0309
10%	-0.0798	-0.0798	0.1403	-0.0399
20%	-0.1716	-0.1716	0.2070	-0.0936
30%	-0.2540	-0.2540	0.3207	-0.1578
50%	-0.4332	-0.4332	0.4446	-0.3011

5.4 가중 핫덱 대체

같은 자료로 모의실험을 실시한 결과가 <표 5. 7>에 나타나 있다.

<표 5. 7> 가중 핫덱 대체에서 조정 후 분산추정 결과

무응답률	가중핫덱대체	표준BRR	조정된BRR	재조정된BRR
5%	235.1997	232.5841	232.7120	232.6635
10%	236.3216	233.2881	233.9654	234.3414
20%	232.2752	230.8934	229.2962	228.7992
30%	236.0329	231.5407	232.5779	232.0004
50%	230.7539	216.4858	219.7897	219.7000

이 경우는 특이하게도 아무런 조정이나 다른 방법을 사용하지 않은 경우가 오히려 분산을 잘 추정해 주는 것을 볼 수 있다. 이는 사용된 자료가 대체방법과 잘 맞고 있어 조정이 오히려 분산 추정의 효율을 떨어뜨릴 수도 있음을 보여준다.

각 무응답률에 따른 조정 방법들의 상대편향은 <표 5. 8>과 같다.

<표 5. 8> 각 무응답률에 따른 조정 방법들의 상대편향

무응답률	가중핫덱대체	표준BRR	조정된BRR	재조정된BRR
5%	0.0018	-0.0285	-0.0094	-0.0088
10%	0.0065	-0.0441	-0.0064	-0.0035
20%	-0.0107	-0.1069	-0.0166	-0.0234
30%	0.0053	-0.1479	-0.0138	-0.0094
50%	-0.0172	-0.2659	-0.0780	-0.0639

상대편향을 보면 대체 조정을 하지 않은 경우 안정적인 편의를 보이고 있고 조정된 BRR 방법과 재조정된 BRR 방법 모두 음의 편의를 보이고 있으나 그 정도는 극히 미미하다.

5.5 다중 대체

다중 대체된 자료의 경우는 분산 추정을 위해 서로 다른 대체값으로 대체된 $M \geq 2$개의 자료 셋을 필요로 한다. 모의실험으로 통계청 자료를 가지고 각각 평균대체, 중위수 대체, 비율 대체 그리고 가중 핫덱 대체된 $M=4$개의 자료 셋으로부터 각각의 분산을 추정한 후 다중대체 분산을 추정하게 된다. 모의실험 결과가 <표 5. 9>에 나타나있는데, 이 결과를 보면 조정된 BRR과 재조

정된 BRR이 거의 유사하게 분산을 과소추정 하고 있고 표준 BRR 방법이 더 좋은 결과를 보임을 알 수 있다.

<표 5. 9> 다중대체 I 에서 조정 후 분산추정 결과

무응답률	다중대체 I	표준BRR	조정된BRR	재조정된BRR
5%	232.6263	235.2820	228.5385	228.5964
10%	236.1633	240.2503	227.0127	227.0586
20%	233.9335	239.6342	215.1918	215.0028
30%	234.4382	235.1435	203.4352	203.2362
50%	228.0496	224.1224	171.3577	171.2901

각 무응답률에 따른 조정 방법들의 상대편향은 <표 5. 10>과 같다.

<표 5. 10> 각 무응답률에 따른 조정 방법들의 상대편향

무응답률	다중대체 I	표준BRR	조정된BRR	재조정된BRR
5%	-0.0490	0.0021	-0.0266	-0.0415
10%	-0.0779	0.0233	-0.0331	-0.0652
20%	-0.1702	0.0206	-0.0835	-0.1462
30%	-0.2501	0.0015	-0.1335	-0.2118
50%	-0.4309	-0.0454	-0.2702	-0.3687

다중 대체의 경우 사용된 대체 방법들 중 평균대체와 중위수 대체의 영향으로 분산이 과소 추정됨을 알 수 있는데 상대적으로 분산 추정의 정도가 좋은 비 대체 방법과 가중 핫덱 대체 방법만을 이용한 다중대체에서의 모의실험 결과는 <표 5. 11>과 같다

<표 5. 11> 다중대체 II 에서 조정 후 분산추정 결과

무응답률	다중대체 II	표준BRR	조정된BRR	재조정된BRR
5%	232.6263	238.2746	230.1219	230.2393
10%	236.1633	250.5078	229.6890	230.0731
20%	233.9335	257.1373	221.0534	220.7462
30%	234.4382	270.8075	215.1562	214.7232
50%	228.0496	277.8314	191.9423	191.7348

이 경우를 살펴보면 두 가지 대체 방법만을 이용한 결과 분산의 과소 추정 정도가 약간 줄어들기는 하지만 표준 BRR의 경우는 비 대체 방법의 경우와 마찬가지로 무응답률이 증가할수록 분산을 상당히 과대 추정하고 있다. 조정된 BRR 방법과 재조정된 BRR의 경우는 별다른 영향을 받지 않음을 볼 수 있다. 위의 결과는 모의실험에 사용된 자료가 비 대체와 가중 핫덱 대체를 이용한 다중대체 방법 자체에 잘 적합하고 있어 대체된 값에 대한 조정이 오히려 좋지 않은 추정을 하게 됨을 알 수 있다. 각 무응답률에 따른 조정 방법들의 상대편향은 <표 5. 12>와 같다.

상대편향을 살펴보면 다중대체 방법들의 선택에 따라서 조정된 BRR의 경우는 무응답률의 증가에 따라 분산의 과대 추정 정도가 심해지는데 비해 표준 BRR 방법과 재조정된 BRR 방법은 다중대체 방법들의 선택에 관계없이 비교적 안정된 분산 추정량을 제공함을 알 수 있다.

<표 5. 12> 각 무응답률에 따른 조정 방법들의 상대편향

무응답률	다중대체Ⅱ	표준BRR	조정된BRR	재조정된BRR
5%	-0.0398	0.0148	-0.0247	-0.0199
10%	-0.0620	0.0670	-0.0366	-0.0217
20%	-0.1392	0.0952	-0.0912	-0.0585
30%	-0.2010	0.1534	-0.1244	-0.0836
50%	-0.3495	0.1833	-0.2252	-0.1825

반대로 조정을 하지 않은 일반 분산식에 의한 상대편향이 상대적으로 작은 편향을 보이고 있는데 이는 다중대체 방법 자체가 대체로 인한 분산의 변동을 고려하여 제시된 것이기 때문이여 다중대체된 상태에서 조정이 오히려 좋지 않은 결과를 나타낸 것으로 생각할 수 있다.

<표 5. 13> 각 대체 방법들과 대체된 값들의 조정 방법에 따른 상대효율

대체방법	무응답률	Jackknife	조정된 BRR (MSEa)	재조정된 BRR (MSEr)	MSE_i/MSE_r	MSE_a/MSE_r
평균대체 MSE_{M1}	5%	411.4096	288.4153	287.7523	1.4297	1.0023
	10%	707.7458	338.1425	339.5378	2.0844	0.9959
	20%	2459.0887	864.7318	854.8482	2.8766	1.0116
	30%	5164.9610	2075.8020	2069.9061	2.4953	1.0028
	50%	14711.6118	7211.5582	7249.2039	2.0294	0.9948
중위수대체 MSE_{M2}	5%	405.4219	288.2163	288.9334	1.4032	0.9975
	10%	688.9466	329.1233	339.4336	2.0297	0.9696
	20%	2378.3827	850.2172	867.1682	2.7427	0.9805
	30%	5029.9598	2018.6286	2018.5565	2.4919	1.0000
	50%	14452.3749	7207.1577	7145.5916	2.0226	1.0086
랜덤비율대체 MSE_R	5%	366.5629	280.1909	276.4427	1.3260	1.0136
	10%	567.0249	313.3206	306.5095	1.8499	1.0222
	20%	1817.9131	695.8375	700.8967	2.5937	0.9928
	30%	3732.9577	1570.7402	1591.9005	2.3450	0.9867
	50%	10475.9796	5161.6904	5207.4418	2.0117	0.9912
가중핫덱대체 MSE_H	5%	235.3688	237.0239	237.1791	0.9924	0.9993
	10%	238.6721	234.6429	234.5413	1.0176	1.0004
	20%	238.5917	259.4615	264.6709	0.9015	0.9803
	30%	237.5814	237.4645	239.7741	0.9909	0.9904
	50%	247.0317	444.7536	447.3632	0.5522	0.9942
다중대체Ⅰ MSE_{M3}	5%	237.3013	267.6004	266.9385	0.8890	1.0025
	10%	238.0534	287.4753	286.8094	0.8300	1.0023
	20%	234.6645	599.2214	606.4765	0.3869	0.9880
	30%	234.5609	1186.4628	1198.7864	0.1957	0.9897
	50%	273.4617	4194.8252	4203.3392	0.0651	0.9980
다중대체Ⅱ MSE_{M4}	5%	237.3013	251.8990	250.9343	0.9457	1.0038
	10%	238.0534	255.6938	252.3077	0.9435	1.0134
	20%	234.6645	409.7059	417.9327	0.5615	0.9803
	30%	234.5609	600.5843	617.3382	0.3800	0.9729
	50%	273.4617	2027.7421	2045.3529	0.1337	0.9914

이상의 결과들을 토대로 jackknife 방법과 조정된 BRR 그리고 재조정된 BRR들과의 상대효율을 보기 위하여 각 대체방법들의 경우에서 평균제곱오차(MSE)를 구하여 각각의 상대효율을 비교하여 보았다. 그 결과들을 종합하여 <표 5. 13>에 요약하였다.

<표 5. 13>의 결과를 보면 모든 경우에서 재조정된 BRR이 jackknife 방법과 비교하여 월등한 상대효율을 보이고 있으며, 조정된 BRR과의 비교에도 재조정된 BRR 방법이 무응답률이 50%인 경우를 제외하고 더 높은 상대효율을 나타내고 있다.

6. 결 론

복합 표본조사에 있어서 분산 추정 방법들을 살펴보고, 무응답 혹은 여러 가지 다른 이유로 결측값이 발생했을 때 완전자료를 구성하기 위한 여러 가지 대체 방법들을 알아보았다. 대체방법은 그 자체로 완전자료를 구성하여 장방형의 자료를 이용한 일반적 통계 소프트웨어를 이용한 분석을 가능하게 해주고, 불편성과 일치성을 갖는 좋은 모수추정량을 만들 수 있다는 장점이 있다. 반면 실제 분산 추정시의 문제는 대체된 자료를 관측된 실제 자료로 간주하고 분석을 함으로써 발생하는 추정의 편의이다. 즉, 대체로 인한 분산의 변동을 고려하지 않음으로써 생기는 분산의 과소 추정 문제가 발생하는 것이다.

이 문제를 해결하는 방안으로 다중대체 방법, 조정된 jackknife 방법, 조정된 BRR 방법 그리고 조정된 bootstrap 방법 등이 제안되었는데 본 논문에서는 조정된 BRR 방법에서 제안되었던 대체된 자료의 조정 방법의 대안으로 여이분표본 자료를 동시에 이용한 조정 방법을 제안하였다.

실제 자료인 2001년 통계청 농업 총조사 자료를 이용하여 모의실험을 실시하여 기존의 표준 BRR 방법과 Shao와 Chen 그리고 Chen(1998)에 의해 제안되었던 대체된 자료의 조정 방법과의 결과를 비교하였다.

그 결과를 살펴보면 <표 5. 13>에서와 같이 결정적 대체 방법인 평균 대체와 중위수 대체의 경우 새로 제안한 조정 방법은 조정된 BRR 방법과 거의 유사한 결과를 보였다.

이상의 결과를 토대로 무응답으로 인한 결측값의 대체 방법은 자료의 구조나 성질에 따라 달리 선택되어야 만이 좋은 결과를 기대할 수 있고, 또한 대체된 자료의 조정이 오히려 분산추정의 정도를 떨어뜨릴 수도 있으므로 결측값의 대체나 대체된 값들의 조정에는 세심한 주의가 필요하다 하겠다.

참고문헌

[1] Bickel, P. J. and Freedman, D. A.(1984). Asymptotic Normality and the Bootstrap in Stratified Sampling, *The Annals of Statistics*, 12, 470-482.

[2] Efron, B. and Tibshirani,R.J.(1993). *An Introduction to the Bootstrap*, Chapman & Hall

[3] Krewski, D. and Rao, J. N. K. (1981). Inference from Stratified Samples : Properties of the Linearization, Jackknife and Balanced Repeated Replication Methods, *The Annals of Statistics*, 9, 1010-1019.

[4] Lee, H., Rancourt, E. and Sarndal, C. E. (1994). Experiments with Variance Estimation

from Survey Data with Imputed Values. *Journal of Official Statistics*, 10, 231-243.

[5] Lehtonen, R. and Pahkinen, E. J. (1994). *Practical Methods for Design and Analysis of Complex Surveys*, John Wiley & Sons.

[6] Rao, J. N. K. and Shao, J. (1992). Jackknife Variance Estimation with Survey Data under Hot-Deck Imputation, *Biometrika,* 79, 811-822.

[7] Rao, J. N. K. and Shao, J.(1996). On Balanced Half-Sample Variance Estimation in Stratified Random Sampling, *Journal of the American Statistical Association*, 91, 343-348.

[8] Risto, Lehtonen. and Erkki, J. P. (1994). *Practical Methods for Design and Analysis of Complex Surveys*, John Wiley & Sons.

[9] Rubin, D. B. (1978). Multiple Imputation for Nonresponse in Surveys, Wiley, New York.

[10] Shao, J. Chen, Y. Chen.(1998). Balanced Repeated Replication for Stratified Multistage Data Under Imputation. *Journal of the American Statistical Association,* 93, 819-831.

[11] Shao, J. and Sitter, R. R. (1996). Bootstrap for Imputed Survey Data. *Journal of the American Statistical Association,* 91, 1278-1288.

[2004년 10월 접수. 2005년 5월 채택]