연구논문

유한모집단에서 가중평균에 포함된 가중치의 효과*

Weighting Effect on the Weighted Mean in Finite Population

김규성**
Kyu-Seong Kim

표본조사에서 가중치는 설계 단계와 분석 단계에서 만들어지고 부여될 수 있다. 설계 단계의 가중치는 추출확률이나 응답률 등과 같은 표본 데이터 획득 지표에 관련되어 있고 분석 단계의 가중치는 모집단 수치나 다른 보조 변수정보 등과 같은 외적인 정보와 관련되어 있다. 그리고 최종가중치는 설계 단계의 가중치와 분석 단계의 가중치의 곱으로 만들어진다. 이 논문에서는 분석 단계에서 부여되는 가중치에 초점을 맞추어 가중평균으로 모평균을 추정할 때 가중평균에 포함된 가중치가 모평균 추론에 미치는 영향을 고찰하였다.

유한모집단에서 각 조사단위에 조사변수와 가중치가 쌍으로 있고 표본추출확률이 균등한 경우를 가정하였다. 이러한 조건에서 가중평균의 편향과 평균제곱오차를 구하여 가중평균은 모평균의 편향 추정량임을 보였고, 편향의 방향과 크기는 조사변수와 가중치의 상관관계로 설명할 수 있음을 보였다. 즉, 만일 가중치와 조사변수가 양의 상관관계가 있으면 가중평균은 모평균을 과대 추정하게 되고, 만일 음의 상관관계가 있으면 모평균을 과소 추정하게 된다. 그리고 두 변수의 상관계수가 크면 편향은 증가한다. 가중평균에 대한 이론적인 수식 유도와 함께 편향의 크기와 평균제곱오차의 크기를 수치적으로 검토하기 위하여 모의실험을 실시하였다. 모의실험에서는 상관계수가 -0.2과 0.6 사이에 있는 9개의 가중치를 생성하였고, 표본수는 100부터 400까지 고려하여 편향의 크기와 평균제곱오차의 크기를 수치적으로 구하였다. 하나의 결과로써 상관계수가 0.55이고 표본수가 400인 경우에 가중평균의 편향의 제곱이 평균제곱오차에서 차지하는 비율은 무려 82%에 이르는 것으로 나타났는데, 이는 가중평균의 편향이 어떤 경우에는 매우 심각할 수도 있음을 보여주는 것이다.

주제어: 래킹비 추정량, 사후층화추정량, 설계가중치, 최종가중치.

* 이 논문은 2005년도 서울시립대학교 학술연구조성비에 의하여 연구되었음.

** 교신저자(corresponding author): 서울시립대학교 통계학과 교수 김규성.
E-mail : kskim@uos.ac.kr

Weights can be made and imposed in both sample design stage and analysis stage in a sample survey. While in design stage weights are related with sample data acquisition quantities such as sample selection probability and response rate, in analysis stage weights are connected with external quantities, for instance population quantities and some auxiliary information. The final weight is the product of all weights in both stage. In the present paper, we focus on the weight in analysis stage and investigate the effect of such weights imposed on the weighted mean when estimating the population mean.

We consider a finite population with a pair of fixed survey value and weight in each unit, and suppose equal selection probability designs. Under the condition we derive the formulas of the bias as well as mean square error of the weighted mean and show that the weighted mean is biased and the direction and amount of the bias can be explained by the correlation between survey variate and weight: if the correlation coefficient is positive, then the weighted mean over-estimates the population mean, on the other hand, if negative, then under-estimates. Also the magnitude of bias is getting larger when the correlation coefficient is getting greater. In addition to theoretical derivation about the weighted mean, we conduct a simulation study to show quantities of the bias and mean square errors numerically. In the simulation, nine weights having correlation coefficient with survey variate from -0.2 to 0.6 are generated and four sample sizes from 100 to 400 are considered and then biases and mean square errors are calculated in each case. As a result, in the case of 400 sample size and 0.55 correlation coefficient, the amount of squared bias of the weighted mean occupies up to 82% among mean square error, which says the weighted mean might be biased very seriously in some cases.

key words : design weights, final weight, post-stratification estimator, raking ratio estimator.

I. 서론

표본조사에서 가중치를 부여한 뒤 추정치를 산출하는 것이 일반화되고 있다. 가중치를 부여하는 첫 번째 이유는 불균등한 추출확률이나 응답률을 보정

하여 표본의 대표성을 확보하기 위해서이다(예를 들면, Kish 1992; 류제복 1993). 작은 확률로 얻어진 데이터에 상대적으로 큰 가중치를 부여하여 표본 데이터의 대표성을 보정해 주는 것이다. 두 번째 이유는 보조변수의 일치성을 추구하면서 추론의 효율을 높이기 위하여 부여된다. 사후층화 후에 모집단 수치를 고려하여 만드는 사후층화추정량이나 래킹비 추정량이 이에 해당되고, 혹은 조사변수와 밀접한 연관이 있는 보조변수를 확보한 후에 추론에 반영하는 비추정량이나 회귀추정량도 분석 단계에서 가중치를 부여하는 대표적인 경우들이다. 이러한 추정량들은 가중치를 부여하지 않은 추정량보다 표준오차가 더 작아야 가중치를 부여한 효과가 나타난다. 또한 래킹비 추정량이나 이를 더 일반화한 일반회귀추정량에서 가중치를 부여하는 전제조건은 보조변수의 표본의 값과 모집단 값이 서로 일치하도록 하는 보조변수의 일치성이다(예를 들면, Sarndal et al. 1992; 김규성 2005).

최종가중치는 설계 단계에서 만들어진 가중치와 분석 단계에서 만들어진 가중치의 곱으로 나타난다. 표본추출 과정에서 생기는 불균등 추출확률을 가중치로 활용해야 하는가 문제는 관점에 따라 다른 답이 제시된다(예를 들면, Hansen et al. 1983; Royall 1970; Korn & Graubard 1995). 논쟁이 되는 설계가중치와는 달리 분석 단계에서 보조변수를 활용하여 가중치를 부여하는 것에 대해서는 별다른 이견은 없는 것 같다. 그 이유는 보조변수를 활용한 결과가 가중치로 표현되기 때문이다. 다만, 어떤 방법을 선택할 것인가에 대해서는 방법론적인 차이가 있을 수 있다. 이 과정에서 '가중치'라고 하는 용어를 사용하는 것이 적절한가에 대해서는 이견이 있을 수 있다. 왜냐하면 가중치 활용의 근본 취지는 설계 단계에 있고, 분석 단계에서 보조변수를 추론에 반영할 때에는 조정(adjustment), 혹은 보정(calibration)이란 용어도 흔히 사용하기 때문이다. 그러나 최종가중치는 설계 단계의 가중치와 분석 단계의 가중치의 곱이고, 추론에 사용되는 것이 최종가중치이므로 이를 분리하여 표현하는 것이 마땅하지 않다고 생각하여 이 논문에서는 단계에 관계없이 '가중치'라는 표현을 쓰기로 한다. 일반회귀추정량에서는 최종가중치를 'g-가중치'라고 표현하고 g-가중치는 설계가중치와 보조변수의 조정항목 혹은 보정항을 모두 포함한다(Sarndal et al. 1992; 김규성 2004).

최종가중치를 w_i라고 하고 y_i를 조사변수라고 할때 모평균 추정량으로서 다음과 같은 가중평균을 고려하자.

$$\bar{y}_w = \frac{\sum_{i \in s} w_i y_i}{\sum_{i \in s} w_i} \tag{1.1}$$

가중치 효과와 관련한 질문은 '가중평균으로 모평균을 추정할 때 가중치가 미치는 영향은 무엇인가?' 하는 것이다. 이 질문에 대해 잘 알려진 답은 두 가지이다. 첫째는 설계기반 관점의 답으로 만일 위의 가중치가 설계가중치이면 주어진 표본설계에서 가중평균은 일치추정량이며 점근적으로 비편향 추정량이라는 것이다. 따라서 표본의 수가 증가할수록 가중평균은 표본조사에서 활용하기에 바람직한 추정량이다. 두 번째 답은 분석의 관점에서 Kish(1992)가 내놓은 것이다. 만일 조사변수가 서로 독립이고 동일한 분포를 갖는다고 하면 위의 추정량은 모형 비편향이고 분산은 가중치를 부여하기 전보다 증가하며, 그 증가분은 상대표준오차의 제곱으로 나타난다는 것이다.

$$Var(\bar{y}_w) = \frac{\sigma^2}{n}(1 + cv_w^2) \tag{1.2}$$

여기에서 n은 표본크기이고, σ^2은 조사변수의 분산, 그리고 cv_w는 표본에 포함된 가중치의 상대표준오차이다. 따라서 가중치를 부여한 가중평균 (1.1)은 모형 비편향성이 유지되고 분산은 가중치를 부여하지 않은 추정량보다 $(1 + cv_w^2)$배 증가하는 것으로 나타난다.

Kish의 결과는 조사단위가 식별되지 않는 무한모집단에서 타당한 것이다. 무한모집단에서는 조사단위가 식별되지 않으므로 차례대로 추출된 n개의 조사단위가 표본이 되고, 가중치는 조사단위에 부여되는 것이 아니고 표본추출 순서에 부여된다. 따라서 무한모집단에서는 가중치와 조사변수가 서로 무관하다. 반면 유한모집단에서는 조사단위가 식별되고 표본은 유한모집단의 부분집합이다. 그리고 가중치는 조사단위에 부여되며 설계나 분석 단계에서 추출확률 및 보조변수를 이용하여 가중치를 만들기 때문에 가중치가 조사변수와 모두 무관하다고는 말하기 어렵다. 추론의 효율을 높이기 위하여 조사변수와

밀접한 연관이 있는 보조변수를 가중치 산출에 활용하기 때문이다. 따라서 유한모집단에서 조사변수와 가중치가 연관이 있는 경우에는 식 (1.2)의 분산식은 타당하지 않다. 이 논문에서는 이러한 점에 착안하여 식 (1.1)에 주어진 가중평균의 성질을 유한모집단에서 찾아보고자 한다. '유한모집단에서 가중평균은 비편향인가?' 그리고 '가중평균의 분산은 식 (1.2)와 같아지는가?'하는 것이 이 논문에서 밝히고자 하는 질문이다.

최종가중치는 설계가중치와 분석 단계의 보정항이 결합되어 만들어지므로 최종가중치의 효과를 완전하게 분석하기 위해서는 표본설계의 확률구조와 조사변수와 보조변수의 관련성을 동시에 구조적으로 파악해야 한다. 그러나 이 연구에서는 처음에 의도했던 일반적인 표본설계에서의 가중치 효과를 모두 다루지는 못하였다. 대신 연구 범위를 축소하여 균등추출확률을 갖는 표본설계 혹은 자체가중(self-weighting)이 되는 표본설계만을 연구 대상으로 하였다. 따라서 분석 단계에서 보조정보를 활용하여 가중치를 부여할 때 가중치가 모평균 추론에 미치는 영향만을 알아보고자 한다.

다음 절에서는 유한모집단에서 가중평균의 편향과 평균제곱오차를 구하고 가중평균의 성질을 고찰한다. 제3절에서는 모의실험을 통하여 2절에서 제시한 편향과 평균제곱오차가 수치적으로 어느 정도 크기인가를 확인해 본다. 마지막 4절에서는 간단한 요약과 함께 향후 연구 과제를 제시한다.

II. 유한모집단에서 가중치 부여 효과

크기가 N인 유한모집단에서 조사변수와 가중치를 쌍으로 고려하자.

$$(w_i, y_i),\ \ i=1,\ldots,N.$$

그리고 각 조사단위의 추출확률은 동일하다고 하자. 고찰의 대상이 되는 추정량은 식 (1.1)과 같은 가중평균이다. 몇 개의 예에서 가중치의 형태를 구체적으로 살펴보자. 사후층화추정량에서 가중치는 사후층 g의 모집단 크기 N_g와 표본크기 n_g로 만들어진 $w_i=N_g/n_g,\ (i\in g\text{층})$이 된다. 표본 단위의 추출확률은 동일하지만 사후층별로 부여되는 가중치는 서로 다

르다. 회귀추정량에서 보조변수를 x라고 하면 가중치 w_i는 다음과 같다. $w_i = N[1/n + (\overline{X} - \overline{x_s})(x_i - \overline{x_s})/\sum_s (x_i - \overline{x_s})^2]$. 보조변수가 하나이고 추출확률이 동일한 경우에 일반화회귀추정량을 고려하면 가중치 w_i는 다음과 같이 표현된다. $w_i = N[1/n + (\overline{X} - \overline{x_s})(x_i/v(x_i))/\sum_s x_i^2/v(x_i)]$, 여기서 $v(x_i)$는 i번째 단위와 관련한 분산항이다. 첫 번째와 두 번째 예에서는 가중치를 표본에 포함된 조사단위에 대하여 모두 더하면 모집단 크기가 된다. 즉, $\sum_s w_i = N$이다. 따라서 식 (2.1)에 나타난 추정량의 형태는 실질적으로 $\overline{y}_w = \sum_{i \in s} w_i y_i / N$가 된다. 이렇게 되는 이유는 처음부터 모집단 크기를 고려하여 추정가중치를 만들었기 때문이다. 사후층화에서 래킹비 추정량도 분모가 모집단 크기에 맞추어져 있기 때문에 이같은 성질이 유지된다. 그러나 세 번째 예에서는 이와 같은 규칙은 지켜지지 않는다. 분산항으로 인하여 가중평균 (2.1)의 분모는 모집단 크기가 되지 않기 때문이다.

이제 위의 예에서와 같이 분석 단계에서 만들어진 가중치를 부여하여 만든 가중평균의 성질을 규명하기 위하여 가중평균 $\overline{y}_w$의 편향과 분산을 구해본다. 이를 위하여 David & Sukhatme(1974)가 구한 결과를 응용하자. 편의상 $z_i = w_i y_i$, $i = 1, \ldots, N$라 하고, $\overline{Z} = \sum_{i=1}^{N} z_i/N$, $\overline{W} = \sum_{i=1}^{N} w_i/N$ 으로 나타내자. 그러면 가중평균 $\overline{y}_w$은 비추정량의 형태로 다음과 같이 표현 가능하다. $\overline{y}_w = \sum_s w_i y_i / \sum_s w_i = \overline{z}/\overline{w}$. 여기서 주의해야 할 점은 가중평균 $\overline{y}_w$가 추정하고자 하는 대상은 $\overline{Z}/\overline{W}$가 아니라 모평균 $\overline{Y}$라고 하는 사실이다. 이러한 차이점에 주의하면서 가중평균 $\overline{y}_w$의 편향을 구하면 다음과 같다.

$$B(\overline{y}_w) \approx \left(\frac{\overline{Z}}{\overline{W}} - \overline{Y}\right) + \frac{N-n}{(N-1)n}\left(\frac{\overline{Z}}{\overline{W}}\right) CV_w^2 \left(1 - \rho_{wz}\frac{CV_z}{CV_w}\right) \quad (2.1)$$

여기에서 CV_w, CV_z는 각각 w와 $z = wy$의 상대표준오차이고 ρ_{wz}는 두 변수의 상관계수이다.

$$CV_w^2=\frac{\sum_{i=1}^{N}(w_i-\bar{w})^2/N}{\bar{w}^2},\quad CV_z^2=\frac{\sum_{i=1}^{N}(z_i-\bar{z})^2/N}{\bar{z}^2} \tag{2.2}$$

$$\rho_{wz}=\frac{\sum_{i=1}^{N}(w_i-\bar{w})(z_i-\bar{z})/(N\bar{w}\,\bar{z})}{CV_wCV_z}$$

또한 가중평균의 평균제곱오차를 구하면 아래와 같다.

$$M(\bar{y}_w)\approx(\frac{\bar{Z}}{\bar{W}}-\bar{Y})^2+\frac{N-n}{(N-1)n}\left(\frac{\bar{Z}}{\bar{W}}\right)CV_w^2 \tag{2.3}$$

$$\times\left\{(\frac{\bar{Z}}{\bar{W}})\left((1-\rho_{wz}\frac{CV_z}{CV_w})^2+(1-\rho_{wz}^2)(\frac{CV_z}{CV_w})^2\right)+2(\frac{\bar{Z}}{\bar{W}}-\bar{Y})(1-\rho_{wz}\frac{CV_z}{CV_w})\right\}$$

식 (2.1)에서 편향 $B(\bar{y}_w)$의 우변 두 번째 항은 표본의 수가 증가하면 크기가 0이 되지만 우변 첫 번째 항은 표본의 크기하고는 무관하다. 다시 말하면 가중평균의 편향은 표본의 크기가 증가하더라도 사라지지 않는다. 가중평균은 비추정량 형태이므로 점근적으로 비편향성을 가질 것이라는 예상과는 달리 가중평균은 점근 비편향추정량이 아니다.

가중평균의 편향의 성질을 구체적으로 알아보기 위하여 다음의 식을 이용하자.

$$\bar{Z}=\frac{1}{N}\sum_{i=1}^{N}w_iy_i=Cov(w_iy_i)+\bar{W}\bar{Y}=\rho_{wy}\sigma_W\sigma_Y+\bar{W}\bar{Y} \tag{2.4}$$

그리고 식 (2.1)에 대입하면 가중평균의 근사 편향은 다음과 같음을 알 수 있다.

$$\frac{\bar{Z}}{\bar{W}}-\bar{Y}=\rho_{wy}CV_wCV_y+O(\frac{1}{n}) \tag{2.5}$$

따라서 표본의 크기가 적당히 커서 식 (2.5)의 뒷부분을 무시할 수 있다면 다음과 같은 성질을 유도할 수 있다.

성질: 표본의 수가 크면 가중평균 $\bar{y}_w$의 편향은 상관계수 ρ_{wy}에 의존한다. 즉,

(i) $\frac{\bar{Z}}{W}-\bar{Y}>0$ 일 필요충분조건은 $\rho_{wy}>0$ 이고,

(ii) $\frac{\bar{Z}}{W}-\bar{Y}<0$ 일 필요충분조건은 $\rho_{wy}<0$ 이며,

(iii) 근사 비편향일 필요충분조건은 $\rho_{wy}=0$ 이다.

바꿔 말하면, 가중치와 조사변수가 양의 상관관계가 있으면 가중평균은 모평균을 과대 추정하는 성질이 있고, 반대로 음의 상관관계가 있으면 과소 추정하는 성질이 있다고 할 수 있다. 그리고 가중평균이 모평균을 비편향 추정하는 경우는 두 변수의 상관계수가 0일 때이다.

또한 가중평균의 평균제곱오차는 식 (2.3)으로부터 다음과 같음을 알 수 있다.

$$M(\bar{y}_w)=(\frac{\bar{Z}}{W}-\bar{Y})^2+O(\frac{1}{n}) \qquad (2.6)$$

표본의 수가 증가하면 식 (2.6)의 우측 항 두 번째 부분은 0이 되고, 첫 번째 부분은 편향의 제곱이므로 결국 가중평균의 평균제곱오차는 표본의 수가 증가하면 편향의 제곱이 된다.

III. 모의실험

1. 유한모집단 생성

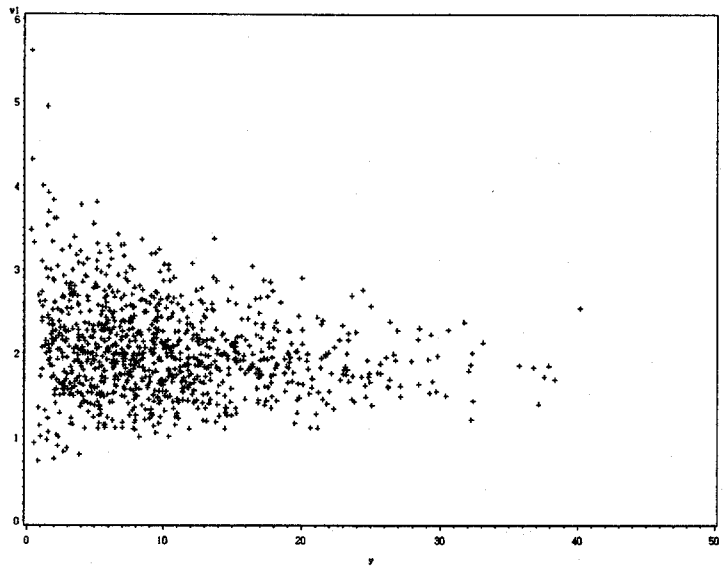
모의실험을 위하여 크기 $N=10,000$인 유한모집단을 생성하였다. 관심변수 y는 감마분포 $G(2,5)$에서 발생하였고, 가중치 w_i는 조사변수 y_i와의 상관계수 크기와 두 변수의 패턴을 고려하여 감마분포 $G(c_i, b_i)$에서 발생하였는데 이때 사용된 인자 b_i, c_i는 다음과 같다.

$$b_i=6.25\times y_i^{1.5\alpha}\times(8+5y_i^{\beta})^{-1},\quad c_i=0.03\times y_i^{-1.5\times\alpha}\times(8+5y_i^{\beta})^2.$$

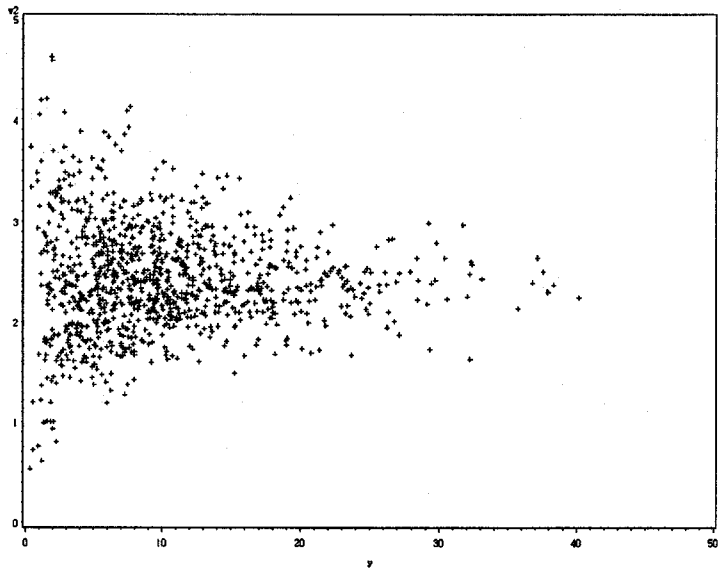
인자 α, β를 변화시켜 9개의 가중치를 만들었는데, 가중치 차례대로 이용한 (α, β)의 값은 다음과 같다: $(-0.5,-0.3)$, $(-0.5,0)$, $(-0.5,0.3)$, $(0,-0.3)$, $(0,0)$, $(0,0,3)$, $(0.2, -0,3)$, $(0.2,0)$, $(0.2,0.3)$. 결과적으로 두 변수의 상관계수 크기는 -0.21에서 0.55 사이가 되었는데, 첫 번째 가중치가 조사변수와 음의 상관계수가 가장 크고 두 번째 가중치는 0에 가까우며 세 번째 가중치가 가장 큰 상관계수를 갖는다. 나머지 6개 가중치는 처음 3개 가중치와 비슷한 패턴을 유지하면서 상관관계가 완화된 형태이다. 〈표 1〉에 관심변수와 9개 가중치의 기초통계량이 주어져 있고 〈그림 1〉 – 〈그림 3〉에 산점도가 주어져 있다.

〈표 1〉 관심변수와 9개 가중치의 기초통계량

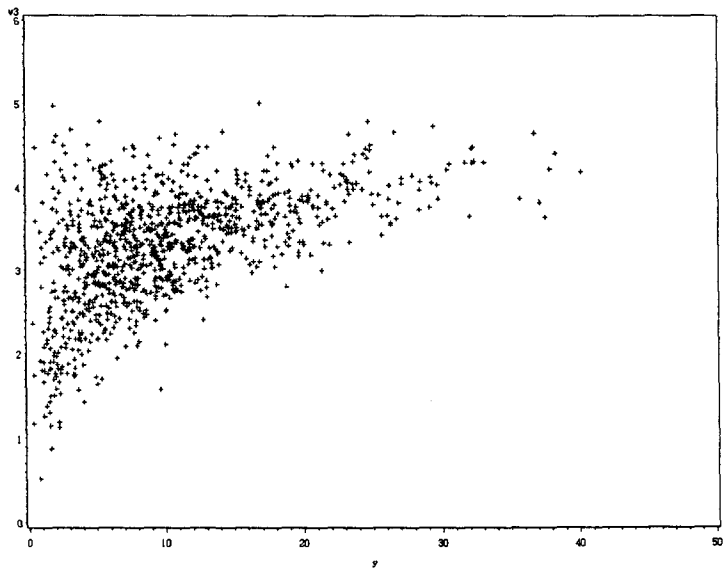
변수명	가중치 w의 분포			$z = w \times y$의 분포		
	평균	변동계수	관심변수와 상관계수	평균	변동계수	가중치와 상관계수
y	10.1116	69.9469	-	-	-	-
w_1	2.0209	28.1521	-0.2087	19.5946	69.2130	0.0580
w_2	2.4394	22.8339	0.0048	24.6852	72.4789	0.2019
w_3	3.2778	21.3126	0.5510	35.8663	83.2912	0.6250
w_4	2.0104	53.7375	-0.1137	19.4619	94.2105	0.5043
w_5	2.4376	44.8927	-0.0224	24.4742	87.6799	0.5004
w_6	3.2713	35.0443	0.3401	35.8362	89.8276	0.6184
w_7	2.0255	75.0877	-0.0751	19.6729	127.0858	0.6788
w_8	2.4513	60.4920	0.0121	24.9140	109.1789	0.6428
w_9	3.2653	47.8683	0.2413	35.6854	100.3170	0.6652



〈그림 1〉 w_1 vs. y
$\rho_{wy} = -0.2087$



〈그림 2〉 w_2 vs. y
$\rho_{wy} = 0.0048$



〈그림 3〉 w_3 vs. y
$\rho_{wy}=0.5510$

2. 가중치 부여 효과

가중치 부여 효과를 살펴보기 위하여 식 (2.1)에 있는 근사 편향을 표본수에 영향을 받지 않는 부분(B_1)과 표본수에 영향을 받는 부분(B_2)으로 나누어 크기를 비교해 보고, 또한 표본수를 증가시켜 가면서 근사 편향(B)의 크기를 구해 보았다.

$$B_1=\left(\frac{\bar{Z}}{\bar{W}}-\bar{Y}\right),\quad B_2=\left(\frac{\bar{Z}}{\bar{W}}\right)CV_w^2\left(1-\rho_{wz}\frac{CV_z}{CV_w}\right)$$

$$B=B_1+\frac{N-n}{(N-1)n}\times B_2$$

아래의 〈표 2〉에 결과가 있다. 〈표 2〉에서 보면, 상관계수가 크면 표본수에 영향을 받지 않는 부분인 B_1이 큰 값을 갖는다. 그리고 표본수가 1인 경

우에는 B_2의 크기가 작지 않지만 표본수가 증가하면 B_2의 크기는 줄어들어 편향(B)이 B_1에 근접함을 알 수 있고, 또한 표본수가 더 크게 증가한다 하더라고 편향의 크기는 0으로 가지 않음을 알 수 있다.

〈표 2〉 표본수에 따른 근사편향의 크기

가중치	상관계수 ρ_{wy}	B_1	B_2	근사편향(B)			
				n=100	n=200	n=300	n=400
w_1	-0.2087	-0.4156	0.6587	-0.4091	-0.4123	-0.4134	-0.4140
w_4	-0.1133	-0.4309	0.3238	-0.4277	-0.4294	-0.4299	-0.4302
w_7	-0.0751	-0.3989	-0.8155	-0.4070	-0.4029	-0.4016	-0.4009
w_5	-0.0224	-0.0713	0.0455	-0.0708	-0.0710	-0.0711	-0.0712
w_2	0.0048	0.0077	0.1894	0.0096	0.0087	0.0083	0.0082
w_8	0.0121	0.0519	-0.5961	0.0460	0.0490	0.0500	0.0505
w_9	0.2413	0.8170	-0.9868	0.8073	0.8122	0.8138	0.8147
w_6	0.3401	0.8431	-0.7871	0.8353	0.8392	0.8405	0.8412
w_3	0.5510	0.8305	-0.7170	0.8234	0.8270	0.8282	0.8288

평균제곱오차에서 가중치의 효과를 살펴보기 위하여 편향과 마찬가지로 평균제곱오차를 표본수에 영향을 받지 않는 부분(M_1)과 표본수에 영향을 받는 부분(M_2)으로 구분하여 크기를 구해 보고, 표본수가 커짐에 따라 근사 평균제곱오차(M)가 어떻게 변하는지를 살펴보았다.

$$M_1 = (\frac{\overline{Z}}{W} - \overline{Y})^2$$

$$M_2 = \left(\frac{\overline{Z}}{W}\right) CV_w^2 \times \left\{ (\frac{\overline{Z}}{W}) \left((1-\rho_{wz}\frac{CV_z}{CV_w})^2 + (1-\rho_{wz}^2)(\frac{CV_z}{CV_w})^2 \right) + 2(\frac{\overline{Z}}{W} - \overline{Y})(1-\rho_{wz}\frac{CV_z}{CV_w}) \right\}$$

$$M = M_1 + \frac{N-n}{(N-1)n} \times M_2$$

〈표 3〉에 평균제곱오차에 대한 결과가 있다. 〈표 3〉에서 보면, 상관계수가 큰 가중치의 M_1이 상관계수가 작은 가중치의 M_1보다 크게 나타난다. 또한 표본수가 증가하면 평균제곱오차가 줄어드는 경향이 있지만 0으로 수렴하는 것은 아니다. 편향으로 인한 양수의 M_1이 있기 때문이다.

〈표 3〉 표본수에 따른 평균제곱오차의 크기

가중치	상관계수 ρ_{wy}	M_1	M_2	근사 평균제곱오차(M)			
				n=100	n=200	n=300	n=400
w_1	-0.2087	0.1727	49.8117	0.6659	0.4168	0.3338	0.2923
w_4	-0.1133	0.1857	62.1051	0.8006	0.4901	0.3865	0.3348
w_7	-0.0751	0.1591	83.9786	0.9906	0.5707	0.4307	0.3607
w_5	-0.0224	0.0050	58.0903	0.5802	0.2897	0.1929	0.1445
w_2	0.0048	0.0000	52.2913	0.5178	0.2563	0.1691	0.1255
w_8	0.0121	0.0027	73.1516	0.7269	0.3611	0.2392	0.1782
w_9	0.2413	0.6676	69.6451	1.3571	1.0089	0.8928	0.8347
w_6	0.3401	0.7108	63.5205	1.3397	1.0221	0.9162	0.8633
w_3	0.5510	0.6898	60.7403	1.2912	0.9875	0.8862	0.8356

편향이 평균제곱오차에서 차지하는 비율과 표본수가 증가함에 따라 편향의 영향이 어떻게 변하는지를 알아보기 위하여 평균제곱오차에 대한 편향제곱의 비율(RB)을 다음과 같이 계산하였다.

$$RB(\%) = \frac{B^2}{M} \times 100$$

편향 대신 편향의 제곱을 사용한 것은 평균제곱오차 중에서 편향이 차지하는 비율을 보기 위해서이다. 아래의 〈표 4〉에 표본수에 따른 편향의 영향이 나타나 있다. 두 가지 사실을 발견할 수 있다. 첫째, 상관계수가 크면 편향의 영향이 증가한다는 사실이다. w_3의 경우, 표본수가 100이면 평균제곱오차 중 편향의 제곱이 차지하는 비율은 무려 52%에 이른다. 둘째, 표본수가 증가

하면 편향의 영향이 더 커진다. w_3의 경우 표본수가 400이면 편향제곱이 평균제곱오차에서 차지하는 비율은 82%에 이른다.

〈표 4〉 표본수에 따른 편향제곱의 비율

가중치	상관계수	편향제곱의 비율 RB (%)			
	ρ_{wy}	n=100	n=200	n=300	n=400
w_1	-0.2087	25.1324	40.7993	51.2187	58.6484
w_4	-0.1133	22.8561	37.6224	47.8170	55.2784
w_7	-0.0751	16.7261	28.4541	37.4468	44.5605
w_5	-0.0224	0.8654	1.7442	2.6252	3.5084
w_2	0.0048	0.0180	0.0295	0.0416	0.0539
w_8	0.0121	0.2921	0.6665	1.0474	1.4336
w_9	0.2413	48.0221	65.3908	74.1926	79.5118
w_6	0.3401	52.0818	68.9113	77.1148	81.9713
w_3	0.5510	52.5168	69.2686	77.4047	82.2121

IV. 결론

표본조사에서 가중치는 설계 단계와 분석 단계에서 만들어지고 부여될 수 있다. 추출확률과 연관된 설계가중치나 활용 가능한 보조변수를 이용하여 만드는 분석 단계의 가중치는 현실문제에서 그 활용 빈도가 꾸준히 많아지고 있다. 활용 빈도와는 별개의 가중치 부여 방법, 그리고 가중치 부여 효과 등은 표본조사 분야에서는 매우 중요한 문제의 하나로 지금까지도 논쟁 중에 있다. 이 논문에서는 균등한 표본추출확률을 갖는 표본조사에서 분석 단계에서 가중치를 부여하여 가중평균을 만든 후 모평균을 추정했을 때 가중치가 추론에 미치는 영향을 알아보았다.

표본조사에서는 모집단이 유한하고 각 조사단위에 그 조사단위의 특성을 반영한 가중치가 부여되기 때문에 가중치와 조사변수가 서로 무관하다고 보기 어렵다. 대신 조사변수와 밀접한 연관이 있는 보조변수를 사용하기 때문에 가

중치와 조사변수는 양의 관계이든 음의 관계이든 어느 정도의 관련이 있다고 보는 것이 현실적일 것이다. 이 논문에서는 이러한 점에 착안하여 유한모집단에서 가중치가 조사단위에 식별되어 부여될 때 부여된 가중치가 가중평균에서 어떤 효과를 나타내는가는 알아 보았다. 요약하면, 가중치가 조사변수와 양의 상관관계가 있으면 가중평균은 모평균을 과대 추정하고, 반대로 음의 상관관계가 있으면 모평균을 과소 추정하게 된다는 것이다. 그리고 두 변수가 서로 무관할 때만이 가중평균은 모평균을 비편향 추정하게 된다. 모의실험에서 예를 보였듯이 상관계수가 크고 표본의 수가 크면 평균제곱오차에서 편향이 차지하는 비율은 매우 높아질 수 있다.

이 논문에서 밝힌 결과는 분석단계에서 가중치 효과를 보인 Kish(1992)의 결과와는 매우 다르다. Kish는 가중평균이 모형 비편향이고 가중치의 표본상대표준오차의 제곱 크기만큼 분산이 증가한다고 하였다. 가중치를 부여하여 가중평균을 만들면 설계비편향성을 확보하는 이점이 있는 대신에 모형 분산이 증가하는 손실이 있다고 하는 것이다. 그러나 이 결과는 무한모집단을 전제로 하고 가중치와 조사변수가 무관하다는 전제에서 구한 것이므로 유한모집단에서 가중치와 조사변수의 상관성을 염두에 둔 본 연구의 결과와는 다르다.

이 논문에서는 계산의 어려움 때문에 처음 의도와는 달리 일반적인 표본설계를 다루지 못하고 균등한 추출확률을 갖는 표본설계만을 대상으로 하였다. 따라서 일반적인 표본설계로 대상을 넓혀서 가중평균의 편향과 평균제곱오차를 구할 수 있으면 그 결과는 이 논문의 결과를 포괄하는 더 훌륭한 결과가 될 것이다. 그리고 최종가중치를 포함하는 일반회귀추정량에 대해서도 유한모집단 성질을 밝힐 수 있을 것이다. 일반회귀추정량은 일반적인 표본설계에서 설계 일치성을 갖고 근사 모형비편향성도 갖는 우수한 추정량으로 알려져 있다 (Sarndal et al. 1992). 그러나 이러한 결과는 모두 점근적인 성질로서 모집단 및 표본의 크기가 동시에 증가한다는 확률구조를 염두에 둘 때 타당한 것이다. 하나의 고정된 유한모집단에서는 이러한 설명을 하지 못한다.

그렇다면 가중평균의 점근적 성질은 어떻게 될 것인가? 가중치가 조사변수와 무관하게 랜덤하게 부여된다면 가중평균은 점근적으로 비편향성을 가질 것

이다. 그러나 가중치가 모집단 크기가 커지더라도 조사변수와 관계가 조사단위의 식별성을 통해 유지된다면 가중평균의 편향은 그대로 유지될 것으로 전망된다. 이 부분은 모집단과 조사변수에 대한 확률 구조를 정립한 후에 밝혀져야 할 사항으로 추후 연구를 기대한다.

참고문헌

김규성. 2004. "표본조사에서 일반회귀추정량의 활용." ≪조사연구≫ 5(2): 49-70.

김규성. 2005. "표본의 대표성과 추정의 효율성." ≪조사연구≫ 6(1): 39-62.

류제복. 2003. "가중값 작성 개요. 표본조사에서 가중치 적용 및 활용." ≪2003년 통계의 날 기념워크숍 발표자료≫ 1-17.

David, I.P. and Sukhatme, B.V. 1974. "On the bias and mean square error of the ratio estimator." *Journal of the American Statistical Association* 69: 464-466.

Hansen, M. H., Madow, W. G. and Tepping, B. J. 1983. "An evaluation of model-dependent and probability-sampling inferences in sample surveys." *Journal of the American Statistical Association* 78: 776-807.

Kish, L. 1992. "Weighting for unequal P_i". *Journal of Official Statistics* 8: 183-200.

Korn, E. and Graunard, B. I. 1995. "Examples of differing weighted and unweighted estimates from a sample survey." *The American Statisticain* 49: 291-295.

Royall, R.M. 1970. "On the finite population sampling theory under certain linear regression models." *Biometrika*, 57: 377-387.

Samdal, C. E., Swensson, B. and Wretman, J. 1992. *Model Assisted Survey Sampling*. Springer-Verlag.