전화조사 응답자의 표집방법으로서 CNU방법과 비례확률 표집방법의 비교연구

趙 盛 謙

차 례

전화조사에서 가구내 응답자를 추출하는 방법은 매우 중요하다. 최종 응답자인 가구내 응답자의 표집이 비확률적으로 이루어질 경우, 그 표본의 대표성을 객관적으로 판단할 아무런 기준이 없기 때문이다. 전화조사에서의 표본추출은 2단계로 나누어 실시되는데, 1단계에서는 전화번호를 그리고 2단계에서는 추출된 전화번호 즉 가구내에서 구체적인 조사대상자를 추출하게 된다. 한국의 경우 전화번호는 확률표집방법에 의해 추출되고 있지만 가구내 응답자는 할당표집 방법으로 표집되는 경우가 많다. 이처럼 가구내 응답자의 추출이 확률표집 방법에 의거하지 않는 경우, 표집오차의 계산이 가능하지가 않게 된다. 따라서 일단 전화번호를 확률적으로 추출한 다음에는 가구내 응답자를 어떻게 확률적으로 추출하는가 하는 문제가 중요하게 된다.

이러한 맥락에서 최근 양승목 등이 개발하여 제시한 CNU방법은 전화조사 상황에서 가구내 응답자의 무작위 추출방법을 제시했다는 점에서 그 의미가 크다.[1] 그러나 CNU방법은 가구 크기에 따라 그 구성원에게 각기 다른 선정확률을 부여하기 때문에 전집치 추정에 편파(bias)가 야기된다. 본 연구의 목적은 이러한 CNU방법의 한계점을 극복하는 대안적 방법으로서 비례확률 추출방법을 제시하고 이의 실용성을 검토해 보고자 하는 것이다.

1) 양승목·김현주·조성겸, "전화조사에서 가구내 응답자의 무선확률표집에 관한 연구," 「신문학보」, 제26호(1991), 189-214.

1. CNU방법의 개요와 한계점

(1) CNU방법의 개요

CNU방법은 양승목 등이 전화조사에서 가구내 응답자를 표집하는 과정에 적용하기 위해 개발한 것으로서 무작위로 추출된 전화번호(즉 각 기구)에서 응답자를 추출하는 절차를 제시한 것이다. 구체적으로 CNU방법은 각 가구내 성인 수와 성인 남자 수를 기준으로 가구를 모두 14가지로 구분한 다음, 각 가구 유형별로 선정대상자를 다음 〈표 1〉과 같이 정하였다.

〈표 1〉 CNU 선정표에서의 가구 유형별 선정대상자

		성인수			
		1명	2명	3명	4명 이상
남자수	0명	여자	나이 많은 여자 나이 적은 여자	제일 나이 많은 여자 둘째로 나이 많은 여자 제일 나이 적은 여자	제일 나이 많은 여자 둘째로 나이 많은 여자 세째로 나이 많은 여자 제일 나이 적은 여자
	1명	남자	남자 여자	남자 나이 많은 여자 나이 적은 여자	남자 제일 나이 많은 여자 둘째로 나이 많은 여자 제일 나이 적은 여자
	2명		나이 많은 남자 나이 적은 남자	나이 많은 남자 나이 적은 남자 여자	나이 많은 남자 나이 적은 남자 제일 나이 많은 여자 제일 나이 적은 여자
	3명			제일 나이 많은 남자 둘째로 나이 많은 남자 제일 나이 적은 남자	제일 나이 많은 남자 둘째로 나이 많은 남자 제일 나이 적은 남자 제일 나이 적은 여자
	4명 이상				제일 나이 많은 남자 둘째로 나이 많은 남자 세째로 나이 많은 남자 제일 나이 적은 남자

그 다음 이들 가구 유형별 선정대상자들이 동일한 선정확률을 갖도록 작성된 12가지 선정표를 이용하여 최종 응답자를 추출하도록 되어 있는데, 그 중의 하나를 예로 들면 〈표 2〉와 같다. 따라서 이 방법을 이용할 경우에는 이와 같은 12가지 선정표를 무작위로 각 설문지에 부착한 다음, 해당 가구의 성인 수와 성인 남자 수에 따라 주어진 선정표 상에서 응

〈표 2〉 CNU 선정표(12유형 중의 하나)

		성	인	수	
		1명	2명	3명	4명 이상
남자수	0명	여자	나이 많은 여자	제일 나이 많은 여자	제일 나이 많은 여자
	1명	남자	남자	나이 많은 여자	제일 나이 많은 여자
	2명		나이 많은 남자	여자	제일 나이 많은 여자
	3명			제일 나이 많은 남자	제일 나이 적은 여자
	4명 이상				제일 나이 많은 남자

답자를 고르면 되는 것이다. 이와 같이 함으로써 CNU방법은 동일 가구내의 구성원들에게는 동일한 선정확률을 부여하고 있다.

CNU방법은 실제 적용결과 그 타당성과 실용성이 높은 것으로 나타났는데, CNU방법을 이용하여 추출한 표본의 경우 할당표집 방법에 의해 추출된 표본보다 전집분포에 보다 근접하는 학력분포를 보였으며, 소요 시간 및 경비 등 실용성의 측면에서도 할당표집 방법보다 우월한 것으로 나타났다.[2)]

(2) CNU방법의 한계점

CNU방법은 각 가구에서 확률적으로 응답자를 추출하는 방법을 제시함으로써 확률표집이 가능토록 했다는 점에 그 의의가 있다. 그러나 이 방법에도 한계점이 있는데 무엇보다도 가구 크기가 다를 경우 응답자들의 추출확률이 달라진다는 점이다. 즉 가구 i의 선정확률을 K_i, 구성원의 수를 N_i라고 할때, 가구 i의 구성원 j가 표본에 포함될 확률 P_{ij}는 K_i/N_i가 된다. 그런데 가구는 동일한 확률로 선정되기 때문에 P_{ij}는 N_i에 반비례하게 된다. 즉 가구크기에 따라 그 구성원의 선정확률이 달라진다는 것이다.

이와 같이 최종표집단위의 선정확률이 각기 다르게 부여된 표집의 경우, 그 선정 확률상의 차이를 무시한 표본의 단순 통계치는 전집치의 편파 추정치(biased estimator)가 된다. 예컨대 가구 수가 M, 조사 대상자 수가 N 그리고 가구 i의 구성원 수가 N_i인 전집에서 m 가구를 표본추출한 다음 각 가구에서 무작위로 1명씩의 응답자를 추출한 경우를 살펴보면, 이때, 변인 X의 전집평균을 표본의 단순 평균치인

$$\bar{x}=\frac{1}{m}\Sigma X_{ij}$$

로 추정하는 경우 이 추정치의 기대치는

$$E(\bar{x})=E_iE_j\left(\frac{1}{m}\sum^{m}X_{ij}\right)$$

$$=\frac{1}{M}\sum^{M}\bar{X}_i$$

2) 양승목 등, 앞의 논문.

가 되므로 변인 X의 전집 평균치인

$$\bar{\bar{X}}=\frac{1}{N}\sum^{M}N_i\bar{\bar{X}}_i$$

와 일치하지 않게 된다. 즉 CNU방법을 이용하여 추출한 표본의 단순평균치로 전집치를 추정할 경우 그 추정치의 기대치와 실제 전집치간의 차 즉 편파는

$$\text{편파(bias)}=\frac{1}{M}\sum^{M}\bar{\bar{X}}_i-\frac{1}{N}\sum^{M}N_i\bar{\bar{X}}_i$$

와 같이 된다.

이처럼 CNU방법은 가구 크기에 따라 그 구성원에 대한 선정확률을 다르게 부여하기 때문에 모수치 추정에 편파가 있게 되는 한계가 있는 것이다.

(3) CNU방법에서의 편파의 정도

그러면 CNU방법을 이용하여 추출한 표본으로 부터 모수치를 추정하는 경우에 발생하는 이러한 편파의 크기는 실제로 어느 정도인가?

앞서 살펴본 추정치의 편파는[3]

$$\text{편파}=\frac{1}{M}\sum^{M}\bar{\bar{X}}_i-\frac{1}{N}\sum^{M}N_i\bar{\bar{X}}_i$$

$$=\frac{1}{\bar{N}}C_{ov}(N_i,\bar{\bar{X}}_i)$$

가 된다. 즉 추정치 편파의 크기는 추정하고자 하는 변인의 가구내 평균치와 가구크기간의 공변량에 비례하는 것이다. 따라서 가구크기와 상관정도가 낮은 변인의 경우에는 가구크기

3) $\text{편파(bias)}=E(\bar{x})-\bar{\bar{X}}$

$$=\frac{1}{M}\Sigma\bar{\bar{X}}_i-\bar{\bar{X}}$$

$$=\frac{1}{M}[\Sigma\bar{\bar{X}}_i-M\bar{\bar{X}}]$$

그런데

$$M\bar{\bar{X}}=M\frac{\Sigma N_i\bar{\bar{X}}_i}{N}$$

$$=\frac{\Sigma N_i\bar{\bar{X}}_i}{\bar{N}}$$

따라서

$$\text{편파}=\frac{1}{M}\left[\Sigma\bar{\bar{X}}_i-\frac{\Sigma N_i\bar{\bar{X}}_i}{\bar{N}}\right]$$

$$=\frac{1}{M\bar{N}}[\Sigma\bar{N}\bar{\bar{X}}_i-\sum N_i\bar{\bar{X}}_i]$$

$$=\frac{1}{M\bar{N}}[\Sigma(\bar{N}-N_i)\bar{\bar{X}}_i]$$

$$=\frac{1}{M\bar{N}}[\Sigma(\bar{N}-N_i)(\bar{\bar{X}}_i-\bar{\bar{X}})+\Sigma\bar{\bar{X}}(\bar{N}-N_i)]$$

$$=\frac{1}{\bar{N}}Cov(N_i,\bar{\bar{X}}_i)$$

에 따라 구성원들의 선정확률이 달라지는 CNU방법의 한계점은 무시될 수 있을 것이다. 문제는 한국의 경우 주요 변인들의 값이 과연 가구크기와 어느 정도의 상관정도를 보이느냐에 있다고 하겠다.

전화조사에서 관심이 되는 변인들과 가구크기간의 관계를 직접적으로 검증한 자료는 없지만 가구통계자료들을 재분석해 본 결과, 가구구성원의 수는 그 구성원들의 연령과 밀접한 관계를 보이는 것으로 나타났다. 한국인구보건연구원에서 1974년에 실시한 '한국출산력조사'의 결과를 재분석해 본 결과, 다음 〈표 3〉에서 보듯이 가족 형태에 따라 구성원들의 연령분포 및 가구내 성인 수가 달라지는 것으로 나타났는데, 가구내 성인 수가 적은 가족형태의 경우에는 대체로 20~30대가 많았고 반면에 성인 수가 많은 가족 형태의 경우에는 모든 연령층이 고르게 분포되어 있었다. 즉 전체 가구의 24.9%를 차지하는 '1세대 젊은 핵가족'과 '2세대 젊은 핵가족'의 경우 가구내 성인 수가 평균 2명인데 이들의 연령분포의 4분편차가 대략 22세에서 32세도 나타났다. 그리고 연령이 주로 30~40대에 분포되어 있는 '2세대 팽창 핵가족'의 경우는 가구당 성인수가 2.2명이었는데 이러한 가구가 20.1%에 달하고 있다. 반면에 전체 가구의 18.9%에 이르는 '3세대 가족'의 경우 가구내 성인 수가 평균 각각 3.5명과 5.6명으로 나타났으며 이들의 연령은 전연령층에 고르게 분포되어 있었다. 즉 성인 수가 적은 가구일 수록 연령이 낮은 것이다.

〈표 3〉 가족형태별 가구내 20세 이상 성인 수 평균 및 연령분포*

가 족 형 태	가 구 수	평 균 성 인 수	연령 4분편차**
1세대 젊은 핵가족	322호	2.0명	22~32세
2세대 젊은 핵가족	2,104	2.0	27~37
2세대 팽창 핵가족	1,953	2.2	37~47
2세대 늙은 핵가족	2,245	2.3	27~52
2세대 복합 가족	126	3.7	27~57
3세대 직계 가족	37	5.6	27~57
3세대 복합 가핵	1,805	3.5	32~62
1세대 늙은 핵가족	224	2.0	52~67
기 타	916	0.9	22~62
계	9,732	2.3	

* 한국인구보건연구원, 『한국의 가족형태와 가족주기에 관한 연구』에 제시된 자료를 재분석한 것임.
** 원 자료에 연령이 5세 급간으로 되어 있기 때문에 그 급간의 중간값을 취해서 계산했음.

따라서 최소한 연령 또는 연령과 관계가 높은 변인의 경우에는 가구크기에 따라 선정확률이 달라지는 CNU방식에 의해 표본을 추출할 경우 전집치 추정에 어느 정도의 편파가 있을 것으로 예측된다. 한국의 경우, 연령에 따라 학력 및 소득정도가 달라진다는 점에서 그리고 이러한 편파는 표본의 크기를 늘려도 일정하게 나타난다는 점에서 이러한 편파가 없는 불편 추정치를 얻을 수 있는 방법이 요청된다고 하겠다.

(4) 불편 추정치의 획득 방법

앞서 살펴보았듯이 한국의 경우 가구크기와 연령이 밀접한 관계가 있다는 점에서 CNU방법에 의한 추정치가 편파(bias)를 보일 가능성이 어느 정도 있다고 보겠다. 그러면 불편 추정치를 얻기 위한 대안적 방법에는 어떤 것들이 있는가?

CNU방법의 한계점을 극복하는 방법에는 크게 두가지 방법이 있다고 보겠다. 첫째는 통계적으로 보완하는 가중평균치를 이용하여 전집치를 추정하는 방법이다. 즉 가구 및 가구내 응답자는 CNU방법을 그대로 이용하여 추출하되, 다만 가구크기에 비례하는 가중치 w_i를 각 가구에 부여한 가중평균치

$$\bar{x}=\frac{\sum^{m} w_i X_{ij}}{\sum^{m} w_i}$$

로 전집치를 추정하는 방법이다. 이러한 가중 평균치는 일종의 비율추정치(ratio estimator)로서 전집치의 불편추정치라고는 할 수 없으나 그 편파는 무시될 수 있는 수준이다.

두번째 방법은 모든 선정대상자에게 동일 선정확률이 부여되도록 가구 선정확률 K_i를 가구크기에 비례하여 부여하는 방법이다. 즉 K_i/N_i가 상수가 되도록 K_i를 달리하는 일종의 비례확률 표집 방법을 이용하여 가구를 추출하는 방법이다. 이 경우, 그 전집 추정치는 표본의 단순평균치인

$$\bar{x}=\frac{1}{m}\sum^{m} X_{ij}$$

가 되는데 이것의 기대값은

$$E(\bar{x})=E_i E_j\left(\frac{1}{m}\sum^{m} x_{ij}\right)$$

$$=E_i\left(\frac{1}{m}\sum^{m} \bar{X}_i\right)$$

이다. 그런데 1차 표집 단위인 가구의 선정확률은 N_i/N에 비례하므로

$$E(\bar{x})=\frac{1}{N}(\sum \bar{X}_i N_i)$$

가 되어 전집치의 불편추정치가 된다.

이와 같이 불편추정치를 얻는 방법에는 두 가지가 있는데[4] 본 논문에서는 먼저 비례확률 표집방법을 제시한 다음 이 비례확률 표집방법과 CNU방법 및 가중치 부여 방법 등을 비교해 보도록 하겠다.

4) 물론 이 외에도 가구 구분을 무시하고 개인들에게 동등 확률을 부여하여 추출하는 방법도 생각해 볼 수 있다. 예컨대 전화번호는 동일확률을 부여하여 추출한 다음, 각 가구의 구성원을 일정한 순으로 배열하면서 매 5번째, 또는 매 3번째 구성원을 추출하는 방법을 들 수 있다. 그러나 이 방법은 가구 구성원을 배열하는데 어려움이 있고, 또 동일 가구에서 두사람을 추출해야 하는 경우가 발생하는 등 문제점이 있기 때문에 본 논문에서는 고려하지 않았다.

2. CNU방법에 대한 대안적 방법으로서 비례확률 표집방법

본 연구에서 제시하고자 하는 비례확률 표집방법은 CNU방법에 대한 대안적 방법으로서 가구크기와 관계없이 모든 조사 대상자에게 동일한 확률을 부여하는 표집방법이다. 이 표집방법에서의 선정표는 모두 4가지로 작성되었는데, CNU방법의 선정표와는 달리 가구크기에 비례하여 가구선정확률이 부여되도록 작성되었다. 그러나 CNU방법과 마찬가지로 4인 가구까지는 모든 구성원을 표본에 포함시키되, 가구내 성인 수가 5인 이상인 경우에는 4인까지만 조사대상자에 포함시키도록 하였는데, 이와 같이 한 가구에서 4인까지만 조사대상자로 포함시킨 것은 양승목 등이 지적한 바와 같이 5인 이상 가구에서 일부 구성원을 누락시키는 것이 그다지 큰 편차를 가져오지 않을 것으로 판단되었기 때문이다. 다음 〈표 4〉는 이러한 비례확률 선정표를 요약한 것이다(각각의 선정표는 〈부록〉 참조).

〈표 4〉 비례확률 표집방법의 선정표

		선정표	성	인	수	
		유 형	1명	2명	3명	4명 이상
남자수	0명	1	여 자	나이 많은 여자	제일 나이 많은 여자	제일 나이 많은 여자
		2	조사안함	조사안함	둘째로 나이 많은 여자	둘째로 나이 많은 여자
		3	조사안함	나이 적은 여자	제일 나이 적은 여자	세째로 나이 많은 여자
		4	조사안함	조사안함	조사안함	제일 나이 적은 여자
	1명	1	조사안함	조사안함	남 자	남 자
		2	남 자	남 자	조사안함	제일 나이 많은 여자
		3	조사안함	조사안함	나이 많은 여자	둘째로 나이 많은 여자
		4	조사안함	여 자	나이 적은 여자	제일 나이 적은 여자
	2명	1		나이 많은 남자	나이 많은 남자	나이 많은 남자
		2		조사안함	나이 적은 남자	나이 적은 남자
		3		나이 적은 남자	조사안함	제일 나이 많은 여자
		4		조사안함	여 자	제일 나이 적은 여자
	3명	1			제일 나이 많은 남자	제일 나이 많은 남자
		2			둘째로 나이 많은 남자	둘째로 나이 많은 남자
		3			제일 나이 적은 남자	제일 나이 적은 남자
		4			조사안함	제일 나이 적은 여자
	4명 이상	1				제일 나이 많은 남자
		2				둘째로 나이 많은 남자
		3				세째로 나이 많은 남자
		4				제일 나이 적은 남자

위의 표에서 '조사안함'이란 해당 가구에서 응답자를 추출하지 않고 건너뛴다는 것을 의미한다. 사용방법은 CNU선정표의 경우와 같다. 즉 먼저 4개 선정표를 무작위로 각 설문지에 부착한 다음 해당가구의 성인 수와 성인 남자 수에 따라 주어진 설문지에 부착된 선정표에서 면접대상자를 선정하면 된다. 그 대상자가 '남자'라고 되어 있으면 남자를 선정하고, '조사안함'이라고 되어 있으면 그 가구에서는 응답자를 추출하지 않고 다음 전화번호에서 응답자를 추출한다는 것이다. 이처럼 위의 선정표에 따라 응답자를 추출하게 되면 매 4개 1인 가구중 어느 한가구에서만 응답자를 추출하게 되고, 매 4개 2인 가구중 2개 가구, 매 4개 3인 가구중 3개 가구, 그리고 4인 이상 가구 모두에서 응답자를 추출하게 된다. 따라서 각 가구는 그크기에 따라 표본에 포함될 확률이 부여된 셈이다.

3. CNU방법과 비례확률 표집방법의 비교

이상에서 보았듯이 CNU방법의 한계점을 극복하는 방안에는 두가지가 있다. 하나는 CNU방법을 그대로 사용하면서 다만 가중치를 이용하여 통계적으로 보완하는 방법이고 또 하나는 본 논문에서 제시한 비례확률 표집방법이다. 그렇다면 이들 두 방법 중 어느 방법이 보다 효율적이며 실용적인가?

(1) 비교 방법

표집 방법은 그 추정치가 불편 추정치여야 함은 물론 그 정확성(accuracy)이 높아야 한다. 즉 표집오차와 편파가 적어야 할 것이다. 또한 실제 적용과정에서의 실용성도 있어야 할 것이다. 때문에 여기서는 정확성과 실용성의 두 가지 기준을 가지고 비례확률 표집방법과 CNU방법을 비교해 보았다. 이러한 표집 방법의 비교에서 본 연구에서는 이론적으로 도출된 평균오차 자승화(Mean Square Error)뿐만 아니라 실제 자료에서 산출한 평균오차 자승화도 아울러 비교해 보았다.

비교를 위한 자료는 실험전집을 이용하여 구했는데, 한국의 가족형태와 유사하도록 실험전집을 추출한 다음, 이 실험전집으로 부터 각기 다른 방법으로 표집하여 연령 및 성별을 추정하였을 때, 그 평균오차 자승화가 어떻게 달라지는가를 비교해 보았다. 실험전집 자료는 유치원, 국민학교, 중학교, 고등학교 및 대학생을 대상으로 504명의 가구를 추출하여 수집하였는데, 이처럼 추출된 504가구에서 20세 이상 성인 남녀 1,345명의 성별과 연령을 측정하였다. 이 때, 중학교 이하에서는 학적부 기록을 토대로 그리고 고등학교 이상에서는 본인에게 미리 준비한 양식에 따라 기재토록 하여 수집하였다.

이렇게 수집된 실험전집에서의 가구내 20세 이상 성인수 분포를 1985년도 전국 인구/주택센서스 자료와 비교해 보면, 본 연구의 실험전집의 성인 1인 가구비율이 1.8%로 센서스

〈표 5〉 실험전집과 센서스 자료의 가구내 20세 이상 성인 수 분포

가구내 20세 이상 성인 수	본 연구의 실험전집분포	85년도 센서스 분포
0명	0.0%	1.0%
1	1.8	11.8
2	57.7	54.0
3	20.0	18.8
4	14.1	9.4
5이상	6.4	5.0
	100.0	100.0

** 센서스 분포는 경제기획원 인구통계국에 의뢰하여 1985년도 센서스 자료중 20만 가구를 단순무선표집하여 산출한 것임.

의 11.8%보다 적게 나타난 것 외에는 실험전집과 센서스자료의 가구크기 분포가 대체로 비슷하게 나타났다. 본 연구의 실험전집에서 성인 1인가구 비율이 낮게 나타난 것은 실험전집이 학생들을 대상으로 해서 추출되었기 때문에 미혼 내지 독신자가 제외되었기 때문인 것으로 보인다.

자료분석에서는 이상과 같은 1,345명의 연령별 자료를 실험전집으로 하고, 여기에서 일정한 크기의 표본을 추출할 때, 그 연령 추정치의 평균오차 자승화가 표집방법에 따라 어떻게 달라지는가를 살펴보았다. 여기에서 연령별 비율 추정을 통해 표집방법을 비교한 것은 이에 대한 결과를 토대로해서 다른 사회과학적 변인들에 대한 영향도 어느 정도 알 수 있을 것으로 판단되었기 때문이다.

(2) 추정상의 정확성 비교

먼저 전집 추정치의 정확성(accuracy)을 비교해 보았다. 추정치의 정확성은 반복적으로 같은 크기의 표본을 추출하여 전집치를 추정하는 경우 그 추정치들이 얼마나 전집치에 근접해 있는가를 나타내 주는 것으로서 정확성이 높다는 것은 곧 표본의 통계치들이 전집치에 근접할 가능성이 그만큼 높다는 것을 의미한다. 이러한 전집 추정치의 정확성은 불편추정치(unbiased estimator)일 경우에는 표준오차의 크기로서 그리고 편파 추정치(biased estimator)일 경우에는 평균오차자승화(MSE)로서 알 수 있는데 각 표집방법에 의한 표준오차 및 평균오차자승화는 다음과 같다.

a. CNU방법에 의한 표준오차 및 평균오차자승화

가. 단순평균치일 경우[5]

5) $MSE(\bar{x}) = V(\bar{x}) + \mathrm{Bias}^2(\bar{x})$
$= E(\bar{x} - \bar{X}_e)^2 + (\bar{X}_e - \bar{X})^2$
$\bar{x} - \bar{X}_e = \left(\bar{x} - \frac{1}{m}\sum \bar{X}_i\right) + \left(\frac{1}{m}\sum \bar{X}_i - \bar{X}_e\right)$
$= A + B$

$$MSE(\bar{x})=\frac{1}{mM}\sum^{M}\left(\frac{N_i-1}{N_i}S_i^2\right)+\frac{1}{mM}\sum^{M}(\bar{\bar{X}}_i-\bar{\bar{X}}_e)^2+(\bar{\bar{X}}_e-\bar{\bar{X}})^2$$

여기서 m 표본의 가구 수

M 전집의 가구 수

N_i 가구 i의 구성원 수

$\bar{\bar{X}}_e$ 단순 평균치의 기대값

$\bar{\bar{X}}$ 실제 전집 평균치

$\bar{\bar{X}}_i$ 가구 i에서의 평균치

$$S_i=\frac{\sum^{Ni}(X_{ij}-\bar{\bar{X}}_i)^2}{N_i-1}$$

$$E(\bar{x}-\bar{\bar{X}}_e)^2=E(A^2)+E(B^2)+E(2AB)$$

$$2AB=2\left(\frac{1}{m}\sum^{m}\bar{x}_i-\frac{1}{m}\sum^{m}\bar{\bar{X}}_i\right)\left(\frac{1}{m}\sum^{m}\bar{\bar{X}}_i-\bar{\bar{X}}_e\right)$$

$$=\frac{2}{m}\Sigma(\bar{x}_i-\bar{\bar{X}}_i)\left(\frac{1}{m}\sum^{m}\bar{\bar{X}}_i-\bar{\bar{X}}_e\right)$$

$$E_i[E_j(2AB)]=E_i\left[E_j\ 2\frac{1}{m}\sum^{m}(\bar{x}_i-\bar{\bar{X}}_i)\left(\frac{1}{m}\Sigma\bar{\bar{X}}_i-\bar{\bar{X}}_e\right)\right]$$

$$E_i(2AB)=0$$

$$A^2=\left(\bar{x}-\frac{1}{m}\sum^{m}\bar{\bar{X}}_i\right)^2$$

$$=\left(\frac{1}{m}\sum^{m}\bar{x}_i-\frac{1}{m}\sum^{m}\bar{\bar{X}}_i\right)^2$$

$$=\frac{1}{m^2}[\Sigma(\bar{x}_i-\bar{\bar{X}}_i)^2+2\sum_{i}^{m}\sum_{i\neq i'}^{m}(\bar{x}_i-\bar{\bar{X}}_i)(\bar{x}_{i'}-\bar{\bar{X}}_i)]$$

$$E_i(E_jA^2)=E_i\left[E_j\frac{1}{m^2}\Sigma(\bar{x}_i-\bar{\bar{X}}_i)^2+E_j\cdot 2\Sigma\Sigma(\bar{x}_i-\bar{\bar{X}}_i)(\bar{x}_{i'}-\bar{\bar{X}}_i)\right]$$

$$E_i\left[E_j\frac{1}{m^2}\sum^{m}(\bar{x}_i-\bar{\bar{X}}_i)^2\right]$$

$$=E_i\left[\frac{1}{m^2}\Sigma\frac{N_i-1}{N_i}\cdot\ S_i^2\ \right]$$

$$=\frac{1}{m^2}\cdot m\sum^{M}\frac{N_i-1}{N_i}\ \ S_i^2\cdot\frac{1}{M}$$

$$=\frac{1}{mM}\Sigma\left(\frac{N_i-1}{N_i}\cdot\ S_i^2\ \right)$$

$$B^2=\left(\frac{1}{m}\sum^{m}\bar{\bar{X}}_i-\bar{\bar{X}}_e\right)^2$$

$$=\frac{1}{m^2}[\sum^{m}(\bar{\bar{X}}_i-\bar{\bar{X}}_e)^2+2\sum_{i}\sum_{i\neq i'}(\bar{\bar{X}}_i-\bar{\bar{X}})(\bar{\bar{X}}_{i'}-\bar{\bar{X}})]$$

$$E(B^2)=\frac{1}{m^2}E_i[\sum^{m}(\bar{\bar{X}}_i-\bar{\bar{X}}_e)^2]$$

$$=\frac{1}{Mm}\sum^{M}(\bar{\bar{X}}_i-\bar{\bar{X}}_e)^2$$

$$\therefore\ V^2(\bar{x})=\frac{1}{mM}\sum^{M}\left(\frac{N_i-1}{N_i}S_i^2\right)+\frac{1}{mM}\sum^{M}(\bar{\bar{X}}_i-\bar{\bar{X}}_e)^2$$

$$\therefore\ MSE(\bar{x})=V(\bar{x})+(\bar{\bar{X}}_e-\bar{\bar{X}})^2$$

나. 가중 평균치일 경우[6]

$$V(\bar{x})=\frac{1}{N^2}\left(\frac{M^2}{m}\ \frac{M-m}{M}\ \frac{\sum^{M}N_i^2(\bar{\bar{X}}_i-\bar{\bar{X}})^2}{M-1}+\frac{M}{m}\sum^{M}N_i(N_i-1)S_i^2\right)$$

b. 비례확률 표집방법에 의할 경우[7]

$$V(\bar{x})=\frac{1}{mN}\sum^{M}N_i(\bar{\bar{X}}_i-\bar{\bar{X}})^2+\frac{1}{mN}\sum^{M}(N_i-1)S_i^2$$

위 식에서 볼 수 있듯이 CNU 방식에 의해 표본을 추출한 다음 단순 평균치로서 전집치를 추정하는 경우의 평균오차자승화는 추정치의 변량에 편파(bias)를 합한 것이 된다. 그런데 이 편파는 앞서 살펴보았듯이 표본의 크기와는 무관하고 다만 가구크기 N_i와 대상 변인의 가구내 평균치 $\bar{\bar{X}}_i$간의 상관도에 의해 결정된다. 이러한 편파를 제외한 추정치의 변량만을 보면 단순평균치가 가중 평균치의 경우보다 적게 된다. 그것은 가중 평균치의 경우 추정치의 변량은 위 식에서 보듯이 $N_i^2(\bar{\bar{X}}_i-\bar{\bar{X}})^2$에 의해 결정되고 있지만 단순평균치의 변량은 $(\bar{\bar{X}}_i-\bar{\bar{X}})^2$에 의해 결정되고 있기 때문이다. 비례확률표본에 의한 추정치의 경우 역시 같은 이유에서 가중평균치의 표준오차보다 적다. 그리고 비례확률 표본의 경우에는 그 추정치가 불편 추정치이기 때문에 CNU방법에 의한 단순평균치의 표준오차보다 항상 적게 된다. 따라서 비례확률 표본의 정확도가 가장 높고 그 다음은 편파의 크기가 어떻게 되느냐에 따라 가중 평균치 또는 단순평균치가 된다. 그러나 가중 평균치의 표준오차는 표본의 크기가 증가할 수록 감소하게 되지만 단순 평균치로 추정할 경우의 편파는 표본 크기와 무관하다. 따라서 표본의 크기가 클 경우에는 가중 평균치를 이용한 추정치가 단순평균치 보다 정확성이 높게 될 것이다.

그러면 본 연구의 실험전집에서 n=50의 표본을 추출한 다음 각 연령층의 비율을 추정할 경우의 표준오차는 구체적으로 어떻게 되는가? 다음 〈표 6〉에서 보듯이 비례확률표집 방

〈표 6〉 표집 방법의 정확도 비교

연령	CNU 방법		비례확률 표집방법
	단순평균치 $MSE(\bar{x})$	가중평균치 $V(\bar{x})$	$V(\bar{x})$
20대	7.94%	6.80	5.70
30대	9.44	6.08	6.24
40대	8.02	6.90	6.69
50대	5.19	5.61	4.78
60세 이상	3.40	4.10	3.40

6) William G. Cochran, *Sampling Techniques*(New York: John Wiley & Sons, 1977), p.292~322 참조.

7) Taro Yamane, *Elementary Sampling Theory*(New York: New York University Press), pp. 259~260.

법에 의한 정확도가 가장 높게 나타나고 있고 그 다음이 CNU방법에 의한 가중 평균치, CNU방법에 의한 단순 평균치의 순으로 나타났다. 즉 연령의 경우 가구크기와 가구내 성원들의 연령간 공변관계가 높았고, 따라서 비교적 그 추정치에 편파가 많았기 때문에 가중평균치의 정확도가 단순평균치의 정확도보다 높게 나타났다.

(3) 실용성의 비교

그러면 표집에 소요되는 시간 등 실용성의 측면에서는 어떠한가? CNU방법과 비례확률 표집방법은 일단 응답자가 추출된 다음의 과정은 동일하다. 다만 CNU방법에서는 추출된 전화번호 모두에서 응답자를 추출하지만 비례확률 표집방법에서는 그 중 일부에서만 응답자를 추출하게 된다. 즉 전화를 걸고도 표본을 추출하지 않는 경우가 있기 때문에 같은 크기의 표본을 추출하기 위해 소요되는 전화번호의 수 즉 전화통화회수는 비례확률 표집방법이 CNU방법에 비해 더 많게 된다. 실제 한국의 가구 구성분포에서 이와 같이 비례확률 표집방법을 이용했을 경우, 추출된 전화번호의 68.5%에서만 실제 응답자를 추출하게 되며, 따라서 소요되는 전화번호의 수는 실제 표본 크기의 1.46배에 달할 것으로 보인다. 즉 비례확률 표집 방법을 이용할 경우 정확도는 높아지는 대신에 실제 소요 전화번호는 그만큼 증대된다는 것이다.

〈표 7〉 비례확률 표집방법에서의 유효 전화번호 비율

가 구 크 기	구 성 비	표 본 포 함 비
0명	1.0%	0.0%
1	11.8	3.0
2	54.0	27.0
3	18.8	14.1
4명 이상	14.4	14.4
계	100.0%	68.5%

그렇다면 같은 정도의 정확도를 얻는데는 어느 표집 방법이 우월한가? 이를 알기 위해서는 각 표집방법별로 해당 변인의 표준오차를 알아야 하기 때문에 정확하게 추정할 수는

〈표 8〉 CNU방법(가중 평균치)과 비례확률 표집 방법의 상대적 효율성 비교

연 령	CNU방법 : 가중평균치 $V(\bar{x})$	비례확률 표집방법 $V(\bar{x})$	표 준 오 차 비	표본크기의 비
20대	6.80	5.70	1.19	1.21
30대	6.08	6.24	.97	0.94
40대	6.90	6.69	1.03	1.06
50대	5.61	4.78	1.17	1.21
60세 이상	4.10	3.40	1.20	1.44

없지만 본 연구에서의 실험전집에서 연령을 추정하는 경우를 보면 비례확률 표집방법이 우월한 것으로 나타났다. 즉 다음 〈표 8〉에서 보듯이 CNU방법을 이용하여 가중치를 부여하는 방법이 비례확률 추출방법을 이용하는 경우보다 그 표본이 적게는 0.94배에서 많게는 1.44배 정도 더 클 때 동일한 수준의 정확도를 얻을 수 있는 것으로 나타났다. 그런데, 비례확률 표집에서 실제 표본의 크기보다 더 필요한 전화번호에서는 그 가구의 성인 수와 성인 남자 수만을 확인하면 되므로 실제 소요되는 시간은 면접에 소요되는 시간에 비해 미미하다고 볼 수 있다. 따라서 46%의 전화번호를 더 추출한 다음 이들의 가족 구조만을 확인하는 것이 실제 표본의 크기를 20%정도 늘리는 것보다는 훨씬 유리하다고 볼 수 있다. 이렇게 본다면 비례확률 표집 방법이 가중치를 부여하는 방법보다도 우월하다고 볼 수 있다. 물론 가구크기와 조사대상 변인간의 상관도가 전혀 없는 경우에는 비례확률 표집 방법보다는 CNU방법이 유리하게 된다. 왜냐하면 이 경우 가중치를 사용할 필요가 없이 단순평균치로서 전집치를 추정하면 되는데, 단순 평균치의 표준오차는 비례확률 표집방법과 대체로 비슷하기 때문이다.

4. 결 론

이상에서 보았듯이 한국의 경우 가구크기와 연령이 밀접한 관계에 있기 때문에 가구크기에 따라 구성원의 선정확률에 차이가 있는 CNU방법을 사용하게 될 경우, 편파된 추정치를 얻게 될 가능성이 높다. 이러한 추정상의 편파를 제거하는 방안으로서 CNU방법을 사용하되, 다만 전집치 추정에서 가중치를 부여하는 통계적 방법과 가구 선정확률을 달리하여 동일확률을 부여토록 하는 비례확률 표집방법을 비교해 본 결과, 비례확률 표집방법이 그 정확성 및 실용성에서 CNU방법보다 우월한 것으로 나타났다. 따라서 가구크기와 공변관계에 있을 것으로 보이는 변인 즉 최소한 연령과의 상관도가 높은 변인을 추정하고자 할 경우에는 CNU방법보다는 본 연구에서 제시한 비례확률 표집 방법을 이용하는 것이 보다 적절할 것으로 판단된다.

그러나 본 연구에서는 이러한 가구크기와 관련된 변인에 어떤 것들이 있는지에 대해서는 명확히 밝혀보지 못했다. 그리고 실험전집을 이용했기 때문에 그 실용성을 정확하게 측정해 볼 수 없었다. 따라서 이러한 문제점을 보다 명확하게 규명하는 연구가 필요할 것으로 보인다. 즉 가구크기와 어떤 변인들이 어느 정도의 공변관계에 있는지 그리고 있다면 그로 인한 편파의 정도와 그것을 극복하는 표집방법들의 상대적 효율성은 어떠한지를 분명하게 밝힐 필요가 있다고 하겠다. 왜냐하면 가구크기와 상관도가 낮을 경우에는 소요 시간면에서 CNU방법에서의 단순평균치로 추정하는 것이 비례확률 표집방법보다 경제적이기 때문이다.

〈부록〉 비례확률 표집방법에서의 선정표

(#1)

		성인수			
		1명	2명	3명	4명 이상
남자수	0명	여 자	나이 많은 여자	제일 나이 많은 여자	제일 나이 많은 여자
	1명	(조사안함)	(조사안함)	남 자	남 자
	2명		나이 많은 남자	나이 많은 남자	나이 많은 남자
	3명			제일 나이 많은 남자	제일 나이 많은 남자
	4명 이상				제일 나이 많은 남자

(#2)

		성인수			
		1명	2명	3명	4명 이상
남자수	0명	(조사안함)	(조사안함)	둘째로 나이 많은 여자	둘째로 나이 많은 여자
	1명	남 자	남 자	(조사안함)	제일 나이 많은 여자
	2명		(조사안함)	나이 적은 남자	나이 적은 남자
	3명			둘째로 나이 많은 남자	둘째로 나이 많은 남자
	4명 이상				둘째로 나이 많은 남자

(#3)

		성인수			
		1명	2명	3명	4명 이상
남자수	0명	(조사안함)	나이 적은 여자	제일 나이 적은 여자	세째로 나이 많은 여자
	1명	(조사안함)	(조사안함)	나이 많은 여자	둘째로 나이 많은 여자
	2명		나이 적은 남자	(조사안함)	제일 나이 많은 여자
	3명			제일 나이 적은 남자	제일 나이 적은 남자
	4명 이상				세째로 나이 많은 남자

(#4)

		성인수			
		1명	2명	3명	4명 이상
남자수	0명	(조사안함)	(조사안함)	(조사안함)	제일 나이 적은 여자
	1명	(조사안함)	여 자	나이 적은 여자	제일 나이 적은 여자
	2명		(조사안함)	여 자	제일 나이 적은 여자
	3명			(조사안함)	제일 나이 적은 여자
	4명 이상				제일 나이 적은 남자