표본의 대표성, 비편향성 그리고 효율성

김규성[1)]

요 약

이 논문에서는 표본조사에서 자주 사용되는 표본의 대표성, 비편향성, 그리고 효율성에 개념에 대하여 고찰하였다. 표본의 대표성은 조사단위의 포함확률로 표현되며 조사모집단의 포함범위와 연관이 있는 반면, 비편향성과 효율성은 표집설계와 추정량에 관련된 개념이다. 비편향성과 효율성은 표본의 대표성을 전제로 하며 가중치 부여로 나타난다.

주요용어 : 대표성, 비편향성, 선형성, 효율성.

1. 서론

유한모집단을 조사대상으로 하는 표본조사에서는 모집단의 일부인 표본만 조사하여 모집단 전체에 관한 사항을 추론해야 하므로, 표본은 표본 자신뿐 아니라 표본에 포함되지 않은 단위까지 대표해야 한다. 이러한 맥락에서 '표본의 대표성'이란 용어는 모집단 전체를 잘 대표하는 표본이라는 의미를 함축하고 있는데, 엄밀하게 보면 유한모집단은 식별되는 조사단위의 모임이므로 한 조사단위가 다른 조사단위를 대신하기 위해서는 조사단위간의 연관성이 있어야 하고, 이를 분명히 해야 한다. 조사단위간에 연관성을 부여하는 방법은 확률화의 관점과 모형화의 관점으로 나누어 볼 수 있다.

1.1 확률화 관점의 대표성

표본을 확률추출하면 한 조사단위가 표본에 포함되어 모집단을 대표할 기회는 확률적으로 주어진다. 마찬가지로 다른 조사단위도 표본에 확률적으로 포함되기 때문에 서로가 서로를 대신할 기회는 확률적이다. 이처럼 확률표집(random sampling)과정에서 조사단위간의 연관성이 발생한다. 대표성의 원리(representative principle)는 만일 표본이 확률추출되면 표본에 포함된 조사단위는 자기 자신뿐 아니라 표본에 선정되지 않는 조사단위까지 대표할 수 있고, 표본에 선정되지 않은 단위의 수는 표본에 선정된 단위들의 포함확률로 추정할 수 있다고 하는 것이다(Brewer, 1999, p.36). 이 원리는 전통적인 표집이론이 기초로 하는 원리이기도 하다. 대표성의 원리가 구현되려면 모집단의 모든 조사단위는 양수의 포함확률을 가져야 한다. 만일 일부 단위가 표본에 포함될 확률이 0이면 이 단위들은 표본에 포함될 수 없기 때문에 다른 단위를 대표할 수 없다. 모든 조사단위가 양수의 포함확률을 갖는 표집설계를 확률표집설계(probability sampling design, Sarndal, et al., 1992, p.32)라고 하는데, 이를 이용하면 '대표성 있는 표본'이란 확률표집설계에 의해서 선정된 표본을 의미하게 된다. 표집이론 책에 나오는 대부분의 표집법은 확률표집법의 일종이다(e.g., Cochran, 1977; 박홍래, 1999). 확률표집법은 정의에서부터 이미 표본의 대표성을 확보하고 있기 때문에 확률표집법에서 표본의 대표성을 따지는 것은 무의미하다. 예를 들어 층화표집에서 비례배정과 최적배정을 고려해보자. 이 경우 '비례배정 표본과

1) 서울시립대학교 통계학과, 교수. 서울시 동대문구 전농동 90

최적배정 표본 중 어느 표본이 더 대표성이 있는가' 하는 질문은 의미가 없다. 왜냐 하면 두 표본은 모두 대표성이 있으며, 차이는 추정의 효율성에 있기 때문이다.

확률화 관점에서 표본의 대표성을 저해하는 요인은 확률표집방법에 있는 것이 아니라 포함확률이 0이 되도록 하는 요인에 있다. 예를 들어 전화조사에서 전화번호부가 모집단 전체를 포함하지 않으면 전화번호부에 등재되지 않은 가구는 표본에 포함될 확률이 0이다. 우리나라의 경우 2003년 현재 전화번호부의 포함률이 대략 70% 정도라는 보고가 있다(허명회외 2인, 2004, p.5). 이 경우 30%의 가구는 표본에 포함될 수 없기 때문에 표본의 대표성은 저하될 수밖에 없다. RDD(random digit dialing)는 전화번호부의 약점에 착안하여 표본의 대표성을 개선하기 위한 방법으로 볼 수 있다.

1.2 모형화 관점의 대표성

관심변수에 확률모형을 가정하면 조사단위간의 연관성은 모형을 통하여 발생한다. 예를 들어 다음과 같은 비모형을 고려하자.

$$Y_i = \beta x_i + v(x_i)\varepsilon_i, \quad i=1,\ldots,N \tag{1}$$

여기서 $\varepsilon_i \sim (0, \sigma^2)$이다. 그러면 모든 조사단위에서 관심변수 Y_i와 보조변수 x_i는 선형성을 갖는다는 공통점이 생긴다. 이러한 연관성이 표본과 비표본을 연결해주는 매개체가 된다. 확률화 관점과는 달리 모형화 관점에서는 모든 단위가 양수의 포함확률을 가질 필요는 없다. 왜냐하면 어느 조사단위에서든 관심변수와 보조변수의 연관성은 있고, 표본은 모형을 통하여 비표본을 대신하기 때문에 추론의 효율을 높이는 표본이면 충분하기 때문이다. 예를 들어 비모형의 경우 모평균추정에 적합한 최적의 표본은 보조변수값이 큰 조사단위들로 구성된 표본이다(Royall, 1970). 이러한 표본은 확률추출된 표본이 아니라 유의선정된 표본이다. 유의선정에서 표본에 포함되지 않는 조사단위들은 포함확률이 0이기 때문에 모집단을 대표할 기회가 없지만 유의 선정된 단위들이 모형을 통하여 모형을 대표하므로 대표성의 문제는 발생하지 않는다.

모형을 이용한 추론에서는 대표성의 문제는 발생하지 않으나 가정된 모형이 적합되지 않을 때에는 추론에 오류가 발생할 가능성이 많다. 따라서 모형가정이 틀렸을 때 오류를 적게 범하는 표본이 선호되는데, 그러한 표본이 모형에 강건(robust)한 표본이다. 균형표본(balanced sample, Royall & Herson, 1973, pp.884)이 그 중의 하나이며, 할당 표본(Quota sample)도 강건성을 염두에 둔 표본이라고 할 수 있다. 할당 표본은 미리 정한 할당표에 의하여 표본수를 고정하고 조건에 맞는 단위를 정해진 수만큼 조사하는 것이므로 확률표본은 아니다. 그러나 주요 보조변수를 이용하여 만든 할당표에 의하여 할당표의 셀별로 표본수를 확보하였기 때문에 모형의 오류에 강건하게 반응하는 표본이라고 할 수 있다.

2. 비편향성

표본의 대표성이 표집과정에서 조사단위의 포함여부에 관련된 기준이라면 비편향성과 효율성은 추정과정에서 표집설계와 추정량의 성질에 관련한 기준이다. 일반통계학에서 비편향성은 비중 있게 다루어지는 개념이긴 하지만 절대적인 개념은 아니다. 편향추정량이라도 평균제곱오차가 작으면 추정량으로 선택될 수 있기 때문이다. 그러나 설계기반관점에서 설계비편향성은 거의 절대적인 기준이다. 적어도 근사 비편향성이나 일치성은 가져야 한다. 표집이론에 등장하는 비추정량이나 회귀추정량은 편향추정량이긴 하지만 근사 비편향추정량이기 때문에 표본의 수가 크면 비편향성을 만족시킨다고 할 수 있다. 보조변수를 이용한 일반화회귀추정량도 마찬가지다.

확률표집설계에 의하여 표본을 선정하면 모든 조사단위는 양수의 포함확률을 갖는다. 따라서

Horvitz-Thompson 추정량을 이용하면 비편향 추정량을 만들 수 있다. 즉, 대표성을 확보한 표본으로는 비편향추정량을 만들 수 있다는 뜻이다. 반대로 비편향 추정량은 반대로 비편향 추정량을 만들 수 있으면 대응하는 표집설계는 대표성을 갖는다. 따라서 설계비편향 추정량을 구할 수 있으면 표본의 대표성은 자연스럽게 만족시킨다고 할 수 있다. 설계비편향성이 강조되는 하나의 이유이다.

3. 표본의 효율성

3.1 비례배정과 최적배정

통상적으로 표본의 효율성은 확률표집 방법과 관련이 있다. 예를 들어 층화임의추출에서 비례배정과 최적배정을 고려해보자. 두 표집법 모두 각 층에서 단순임의추출을 하므로 표본의 대표성 측면에서는 두 표집법 모두 만족한다. 차이는 효율성에 있다. 어느 표집법이 더 효율적인가? 일변수인 경우는 최적배정에 의하여 선정된 표본이 더 선호된다. 그러나 이때 하나의 조건이 필요하다. 즉, 층내 분산을 근사적으로 알고 있어야 한다. 왜냐하면 최적배정은 층내 표준편차와 층내 모집단 수의 곱에 비례하는 배정이기 때문이다. 만일 층내 분산을 알지 못하면 최적배정은 활용할 수 없는 이론적인 배정방법이 된다. 다변수인 경우는 비례배정이 더 선호된다. 왜냐하면 각 변수에 대한 층내 근사 분산을 구하기 어려울 수 있고, 설령 층내 분산을 안다고 하더라도 여러 변수를 동시에 만족시키는 최적배정은 복잡한 계산 절차를 필요로 하기 때문이다(e.g., Bethel, 1989). 또한 비례배정은 최적배정을 할 때 범할 수 있는 오류, 즉 잘못 추정된 층내 분산을 사용할 때 발생하는 오류,를 범하지 않는 장점과 계산이 간단한 장점이 있어서 보통의 사회·경제조사에서 선호되는 방법이다. 두 표본을 비교하면, 표본의 대표성 측면에서는 두 표본이 공통적으로 대표성을 가지고 있으며, 차이점은 두 표집설계의 추정의 효율이 서로 다르다는 것이다.

3.2 예제

비례배정과 최적배정의 차이를 비교해 보기 위하여 간단한 예제를 들어본다. 모집단은 Sarndal, et al.(1992)의 부록에 있는 MU284 데이터를 사용하고, 그 중 4개의 변수를 관심변수로 하였다.

P85 : 1985년도 인구수 (단위 1,000)
RMT85 : 1985년도 세금
CS82 : 보수당 의석수
REV84 : 1984년 부동산 가격

<표 1> 층별 표본배정 방법

층번호	층크기	층별 표준편차				표본배정				
		P85	RMT85	CS82	REV84	비례	최적1	최적2	최적3	최적4
1	73	79.25	751.85	6.43	7256	8	13	10	11	13
2	70	32.18	428.53	4.05	2655	7	5	5	7	4
3	141	39.17	575.08	3.70	3763	15	12	15	12	13

지역변수 Reg를 이용하여 3개의 층으로 층화를 하였고(층1:Reg=1,2/층2:Reg=3,4/층3:Reg=5,6,7), 표본의 수는 30으로 하여 층 크기에 비례한 비례배정과 각 변수에 최적배정을 하였다(예 : 최적배정1은 P85에 최적배정). <표 1>은 모집단 상황과 표본 배정에 때른 층별 표본수이다.

아래의 <표 2>은 5가지 표본배정에 대한 층화평균의 변동계수이다.

<표 2> 표본 배정 방법에 따른 변동계수

변수명	변동계수				
	비례배정	최적배정1	최적배정2	최적배정3	최적배정4
P85	0.29854	**0.27831**	0.28507	0.28288	0.27981
RMT85	0.41722	0.42121	**0.41126**	0.41892	0.42190
CS82	0.08732	0.08662	0.08662	**0.08521**	0.08801
REV84	0.26202	0.24438	0.24911	0.24955	**0.24417**

4가지 최적배정법은 해당 변수에서는 가장 낮은 변동계수를 보이며, 비례배정은 P85와 REV84에서 가장 큰 변동계수를 보인다. 위의 경우에 어느 배정법이 가장 효율적인가?

3.3 사후층화

사후층화는 표집 후 조사된 관심변수를 층화하는 방법으로, 표집 전에는 층화변수 값을 알지 못할 때 쓰는 방법이다. 많은 경우 사회조사에서 인구변인에 대한 고려를 표집 전에는 하기 어렵기 때문에 표집 후에 조사된 데이터를 보고 이에 대한 고려를 하게 된다. 통상적으로 사후층화 추정량은 조건부 비편향추정량이며 보조변수의 모집단 값을 추정량에 반영하기 때문에 추정의 효율을 높이는 것으로 알려져 있다.

사후층화 추정량은 일종의 가중치 보정추정량이다. 보조변수를 이용하여 가중치 보정을 하면 가중치 보정을 하기 전보다 효율이 증가하는가? 사후층화 추정량에 대한 효율성은 표집 후 조건부 분포를 기준으로 할 것인지 혹은 표집전 무조건부(unconditional) 분포를 기준으로 할 것인지에 따라 달라진다.

무조건부 분포에 기준한 추론

아래의 <표 3>은 3.2절의 위의 예제에서 층을 사후층으로 간주하여 사후층화추정량의 변동계수를 구한 결과이다. 그리고 분산은 무조건부 분산이다. 4개의 변수 중 3개의 변수가 사후층화를 하지 않는 표본평균의 변동계수보다 더 크게 나타난다. 왜 이런 현상이 생기는가? 사후층화가 효율적이려면 다음의 조건이 만족하여야 함이 알려져 있다(e.g. Cochran, 1977) :

$$\left(\frac{1}{n}-\frac{1}{N}\right)\sum_h(\bar{Y}_h-\bar{Y})^2 > \frac{1}{n^2}\sum_h(1-W_h)S_h^2 \qquad (2)$$

즉, 사후층화가 효과가 있으려면 층별 평균이 상당한 차이가 있어야 함을 의미한다. 극단적으로 사후층 평균이 차이가 없으면 사후층화는 하지 않는 것만 못하다. P85, RMT85, REV84의 경우 층별 평균의 차이가 크지 않기 때문에 사후층화 추정량의 변동계수가 단순임의추출의 경우보다 크게 나타나는 것으로 볼 수 있다.

<표 3> 사후층화추정량의 변동계수

변수명	단순임의추출	사후층화
P85	0.30317	0.31472
RMT85	0.42026	0.43547
CS82	0.09371	0.09204
REV84	0.26624	0.27614

조건부 분포에 기준한 추론

사후층화에서는 표본수가 미리 정해지는 것이 아니고 표집 후 사후적으로 정해진다. 조건부 분포에 기준한 추론은 사후 층화 후 표본수를 고정으로 간주하고 추론을 하는 것이다. 따라서 층별 표본수가 얼마가 될지 모르기 때문에 표집전에 변수별 변동계수를 구해볼 수는 없다. 따라서 표집 후 표본수를 보고 조건부 변동계수를 구하게 된다. 아래의 <표 4>는 층별 표분수에 대한 가능한 여러 경우의 수 중에서 예제로 4가지 경우를 생각한 것이다. 그리고 <표 5>는 4가지 사후 표본수에 대한 조건부 변동계수를 구한 것이다.

<표 4> 층별 표본배정 방법

층번호	층크기	사후 층별 표본수				
		비례	배정1	배정2	배정3	배정4
1	73	8	10	15	7	7
2	70	7	10	7	15	8
3	141	15	10	8	8	15

<표 5> 표본 배정 방법에 따른 변동계수

변수명	변동계수				
	비례배정	배정1	배정2	배정3	배정4
P85	0.29854	0.29763	0.29434	0.34316	0.31051
RMT85	0.41722	0.44252	0.46584	0.50025	0.42646
CS82	0.08731	0.08800	0.09004	0.09839	0.09004
REV84	0.26202	0.26423	0.26215	0.30539	0.27277

조건부 변동계수(<표 5>)를 무조건부 변동계수와 비교해 보면, 배정1, 2와 배정4는 무조건부 변동계수보다 작으나 배정3의 경우는 도리어 무조건부 변동계수보다 크다. 즉, 경우에 따라서 조건부 변동계수의 크기와 무조건부 변동계수의 크기보다 크거나 작다.

4. 결론

본 논문에서는 표본조사에서 흔히 말하여지는 '표본의 대표성'에 관하여 알아보았다. 표본의 대표성은 표집에 관련된 개념이며 조사모집단의 포함률(coverage rate)과 관련이 있다. 그리고

비편향성은 표집설계와 추정량에 관련한 개념으로 표본의 대표성과는 밀접한 연관이 있다. 반면 효율성은 확률표집을 전제로 추정량의 정도와 관련이 있다. 예로써 층화시 표본배정이나 사후층화는 표본의 대표성 보다는 추정의 효율성에 관련한 개념임을 설명하였다.

참고문헌

[1] 박홍래 (1999). 통계조사론. 영지문화사.

[2] 허명회, 강용수, 손은진(2004). 사회조사에서 조사방법에 따른 가중 칸 설정에 관한 연구. 조사연구, 5권 1호, 1-26.

[3] Bethel, J. (1989). Sample allocation in multivariate surveys. Survey Methodology. 15, 47-57.

[4] Brewer, K.R.W. (1999). Design-based or prediction-based inference? stratified random vs stratified balanced sampling. International Statistical Review, 67, 35-47.

[5] Cochran, W.G. (1977). Sampling Techniques. 3rd edition, Wiley.

[6] Royall, R.M. and Herson, J. (1973). Robust estimation in finite populations I. Journal of the American Statistical Association. 68, 880-889.

[7] Sarndal, C.E., Swensson, B. and Wretman, J. (1992). Model assisted survey sampling. Springer-Verlag.