조사연구(2017. 2) 18권 1호: 1−29
DOI http://dx.doi.org/10.20997/SR.18.1.1

연구논문

표본조사에서 비확률표집 방법론 고찰*

A Study of Non-probability Sampling Methodology in Sample Surveys

김규성a)
Kyu-Seong Kim

본 연구에서는 표본조사에서의 비확률표집 방법론을 고찰하였다. 포함범위 감소, 무응답률 증가, 그리고 조사비용의 증가 등으로 인하여 표본조사의 주류이론이었던 확률표집이론이 심각한 도전에 직면한 상태이고, 현실적으로는 웹조사와 같은 비확률표본에 의한 표본조사가 점점 늘고 있는 상황에서 비확률표본을 이용한 표본조사가 수용 가능한 확률표본 기반의 표본조사의 대안이 될 수 있는지에 대한 관심이 커지고 있다. 비확률표본의 약점은 연구자의 선택편향을 제어하지 못한다는 점과 연계된 확률분포가 뚜렷하지 않아 통계적 추론이 어렵다는 점이다. 선택편향과 관련하여 무시가능한 표집설계를 사용하거나 관측연구나 임상실험에서와 같이 표본대응 방법을 이용하면 선택편향을 줄일 수 있음을 살펴보았고, 통계적 추론 방법으로는 균형표집 방법, 모형기반추론 방법, 그리고 의사설계기반 방법을 고찰하였다.

주제어: 무시가능한 표본설계, 선택편향, 의사설계기반 방법, 표본대응

Non-probability sampling methodology in sample surveys is investigated. Concerns about whether a sample survey with non-probability sample might be an acceptable alternative to the survey

* 이 논문은 2015년도 서울시립대학교 교내학술연구비에 의하여 연구되었음.

a) 서울시립대학교 통계학과 교수 김규성.
E-mail: kskim@uos.ac.kr.

with probability sample have been increasing in the situations that the probability sampling theory is faced with a great challenge due to decreasing coverage rate and increasing non-response rate coupled with rising costs and practically the number of sample surveys using non-probability samples like web survey is growing.

The weakness of surveys with non-probability samples is the lack of control of selection bias as well as the difficulty of statistical inference based on the non-probability sample. With regard to selection bias, it is examined that if an ignorable sampling design or the method of sample matching is used, then the selection bias can be ignored or may be reduced. In addition, the balanced sampling method, the model-based method and the pseudo-design based method are studied for the inferential method based on non-probability samples.

Key words: ignorable sampling design, pseudo-design based method, sample matching, selection bias

Ⅰ. 서론

확률표집이론은 20세기 중반 이후 표본조사에서 표본추출 및 모수추정에 관한 주류 방법론으로 자리를 잡아왔다(Baker et al. 2013: 90). 특히 국가기관이나 국책연구소 등에서 실시되는 대규모 표본조사에서는 확률표집이론을 바탕으로 조사를 수행하는 경우가 많았는데, 그 이유는 확률표집이론에 의하여 표본을 선정하면 연구자의 주관적 성향이 표본선정 과정에 개입되지 않고 객관적인 통계를 생산할 수 있다는 점이 국가기관이나 국책연구소에게는 매우 매력적이었기 때문이다. 확률표집은 잘 정비된 표본추출틀과 정교한 표집설계, 표집설계에 의한 표본추출, 그리고 완전한 응답을 전제로 한다. 그런데 만일 완비된 표본추출틀

이 준비되지 못하거나 완전한 응답을 받지 못하면 확률표집이론의 바탕이 되는 확률분포에 왜곡이 생기게 되고, 이 때문에 표본의 대표성이 훼손되고 통계의 객관성이 약화되는 결과가 초래된다. 확률표집이론의 최대 강점인 표본의 대표성과 통계 결과의 객관성에 치명적인 문제가 제기될 수 있는 것이다.

그런데 최근 표본조사의 주류 이론인 확률표집이론이 중대한 도전을 맞이하는 상황이 형성되고 있다. 표본추출틀의 포함범위 감소와 무응답의 증가가 점점 더 많이 발생하고 있는 것이다. 표본추출틀의 포함률과 응답자의 응답률을 높이기 위해서는 조사비용을 더 늘리고 조사기간을 더 길게 잡아야 한다. 통상적으로 표본조사를 통하여 얻으려고 하는 정보의 양이 많아질수록 조사항목 수는 증가하고 세분화 되며 조사내용의 난도는 더 높아지게 되므로, 이를 구현하기 위한 조사비용은 더 증가하게 된다. 표본조사를 통하여 획득하고자 하는 정보의 양이 증가할수록 조사비용은 더 많이 필요해지는 것이다. 이러한 상황에서 표본추출틀을 보완하고 응답률을 높이기 위하여 조사비용을 추가로 더 부담해야 한다면 이는 곧바로 확률표집이론의 채택을 머뭇거리게 하는 현실적인 요인이 될 수도 있다.

확률표집이론에 근거한 표본조사뿐만 아니라 비확률표본을 이용하는 표본조사도 꾸준히 실시되어 왔다. 예를 들면 웹 조사가 그렇다. 웹 조사는 비록 비확률표본을 이용하는 조사이지만 데이터 수집 속도가 빠르고 조사비용이 저렴하며 경우에 따라서는 잠재 응답자에 대한 접근이 용이하기 때문에 현실적으로 매우 선호되는 조사 중의 하나이다. 그런데 이러한 장점과는 달리 웹 조사는 과소 포함범위 문제, 자기선정(self-selection) 문제, 그리고 미지의 추출확률 문제 등을 내포하고 있어(Bethleham 2009: 24), 웹 조사에 기초한 통계적 추론 방법론은 아직까지도 논쟁적이다. 확률표집이론을 선호하는 연구자들에게는 “만일 연구목적이 모집단 수치를 정확히 추정하는 것이라면 연구자들은 비확률 온라인 패널은 반드시 피해야 하는 것(Baker et al. 2010: 714)”이라거나, 혹은 “확률표집 없는 통계적 추론은 불가능한 것이거나 혹은 비확률표집 방법은 통계적 추론에 적절하지 않은 것(Baker et al. 2013: 92)”이라는 견해가 지배적이다.

비확률표본을 이용하는 표본조사는 경우에 따라서는 완비된 표본추출틀을 요구하지도 않을 뿐만 아니라 표본선정이 간단하고 빠르며 상대적으로 조사비용도

저렴하여(Statistics Canada 2003: 89) 확률표집이론에 근거하면 매우 어려울 수 있는 조사도 상대적으로 수월하게 실시할 수 있는 장점이 있다. 비확률표본에 의한 표본조사가 매우 유혹적인 이유이다. 이러한 장점과는 달리 앞서의 웹 조사와 마찬가지로 비확률표본을 이용한 조사의 약점은 주로 통계적 추론에서 나타난다. 확률표집이론의 관점에서 보면 비확률표본에는 표본의 대표성을 부여하기도 어려울 뿐만 아니라 추출확률분포가 없으므로 이에 근거한 추정값이나 표준오차 값 등도 계산할 수가 없다. 따라서 비확률표본을 조사해서는 통계적 추론을 제대로 할 수가 없는 것이다.

표본조사가 아닌 다른 분야, 예를 들면 사례조절(case-control) 연구, 임상실험(clinical trial), 관측연구(observational study) 등에서 사용하는 표본도 확률표집이론의 기준에 의하면 비확률표본인 경우가 대부분이다. 그 이유는 조사나 실험의 편의성 혹은 불가피성으로 인하여 비확률표본을 사용할 수밖에 없는 상황이 많기 때문이다. 이러한 분야에서는 미지의 모집단에 가상의 데이터 분포를 가정하고 만일 데이터 획득 과정에서 실험자 혹은 조사자의 선택편향이 발생하지 않았다면 획득된 데이터는 타당한 데이터로 간주하고 가정된 분포에 기초하여 통계적 추론을 하는 방법론을 채택한다. 그런데 이러한 방법론이 표본조사에서 주류 방법론으로 채택되지 못한 이유는 모집단에 부여한 분포가정이 현실적이지 않은 경우가 많기 때문이다. 현실적인 유혹에도 불구하고 표본조사에서는 비확률표본을 이용하는 조사를 권장할 준비가 아직 안 되어 있는 것이다.

그러나 비확률표본조사의 사례가 많아지면 비확률표본에 근거한 추론방법을 개발할 필요성이 점점 더 커질 것이다. 전통적으로 표본조사 분야에서는 이론개발이 현실의 수요를 선도하기보다는 뒤쫓아 가는 경우가 많았다. 대표적으로 20세기 초에 조사방법이 전체를 다 조사하는 완전계수(complete enumeration)에서 부분만 조사하는 표본조사로 이행된 것은 표본조사가 완전계수보다 개념적으로 우월하거나 이론이 더 완벽해서가 아니라 폭증하는 정보 수요에 시간과 비용이 많이 드는 완전계수보다는 미흡하지만 신속한 결과를 제공하는 표본조사가 더 잘 부응했기 때문이다. 그리고 표본조사이론은 시간이 지나면서 서서히 체계를 갖추어 갔다. 비확률표본에 의한 표본조사도 이와 유사한 과정을 거칠 것으로 보인다. 비확률표본에 의한 표본조사의 수요가 늘어나면서 이에 부응하

는 이론 개발이 뒤따를 것으로 예상되는 것이다. 이러한 예상이 희망적인 것은 표본조사에서 확률표집이론은 도그마가 아니라 전략이라는 점이다(Kish 1965: 29). 즉, 표본조사에서 확률표집이론은 절대적인 원리라기보다는 객관적인 통계결과를 생산하기 위한 훌륭한 전략의 하나인 것이다. 따라서 전략으로서의 확률표집의 원리를 충분히 이해하면 비확률표본을 이용한 방법론도 모색해 볼 수 있을 것이다.

본 논문은 다음과 같이 구성되어 있다. 다음 장에서는 확률표집이론의 바탕이 되는 확률과 확률화의 개념을 살펴볼 것이다. 제3장에서는 비확률표본이 역사적으로 활용된 사례를 부분적으로나마 고찰함으로써 비확률표본에 관한 논점들을 검토해 볼 것이다. 제4장에서는 비확률표본의 주요 방법론을 고찰하고 마지막으로 제5장에서는 요약과 더불어 향후 연구방향을 제시할 것이다.

Ⅱ. 확률화와 비확률표본

표본조사에서 비확률표본 문제를 다룰 때 확률과 확률화(randomization)[1]의 개념을 구분하여 이해할 필요가 있다. 먼저 통계적 실험에서 확률화는 실험단위를 무작위로 배치하는 과정을 뜻하고 표본조사에서 확률화는 조사단위를 선정할 때 조사단위를 확률적으로 선정하는 과정을 뜻한다. 과학의 한 분야로서 통계학이 과학에 제공한 매우 중요한 기여 중의 하나로 확률화를 꼽는 견해(Smith

1) ‘random’의 사전적 의미는 ‘어떤 일이 일어날지 미리 생각하거나 결정하지 않은 상태에서 일이 이루어지거나 선택되는(Oxford Advanced Learner’s Dictionary) 것’이므로 단어 그대로 번역하면 ‘randomization’은 ‘무작위(無作爲)’ 즉 ‘통계의 표본 추출에서, 일어날 수 있는 모든 일이 동등한 확률로 발생하게 함(국립국어원 표준국어대사전)’에 가깝다. 따라서 무한모집단을 대상으로 하는 통계적 실험에서는 ‘randomization’는 ‘무작위’로 번역되는 것이 자연스럽다. 그러나 표본조사에서는 사전에 계획된 표본설계에 의하여 매우 고의적이고 체계적으로 표본선정이 이루어지고 표본추출확률도 동등하지 않으므로 ‘randomization’은 ‘무작위’보다는 ‘확률화’에 더 가깝다. 이 논문에서는 ‘randomization’을 ‘확률화’로 표현하되 확률의 개념이 내포되지 않은 상황에서는 ‘무작위’로 표현하기로 한다.

1983: 394)가 있을 정도로 확률화는 통계학을 특징짓는 매우 중요한 개념 중의 하나이다. 그 이유는 통계적 실험에서 확률화는 연구자의 개인적 선택을 제거함으로써 연구자의 주관적인 선택편향의 가능성을 제거하기 때문이다. 결과적으로 통계 결과의 객관성은 확률화에 의하여 보장받게 된다. 확률화의 또 다른 기여는 확률화로 인하여 생성된 확률분포가 통계적 추론의 근거를 제공한다는 점이다 (Smith 1983: 394; Cox 2006: 192). 이러한 확률분포에 근거한 통계적 추론을 확률화 기반(randomization-based) 추론 혹은 설계기반(design-based) 추론이라고 부른다. 엄밀한 의미에서 확률화에 의해 생성된 확률분포는 표본단위 선정을 위한 확률분포이므로 조사 대상의 불확실성을 표현하는 일반적인 확률분포와는 그 의미가 다르다. 이러한 이유 때문에 확률화에 의해 생성된 확률분포는 비록 확률적으로는 타당할지 모르나 이러한 확률분포를 기초로 하는 추론은 타당하지 않다는 주장도 있다(Royall 1983: 794).

이제 표본조사에서 사용되는 확률의 개념을 살펴보자. 표본조사에서 발생할 수 있는 확률의 원천은 다음의 세 가지를 포함한다(Cassel et al. 1977: 1). 첫째, 조사단위를 선정하는 과정에서 표본선정확률이 발생될 수 있다. 이 확률은 앞에서 언급했던 확률화에 의하여 생성되는 확률과 동일한 확률이다. 둘째, 선정된 조사단위에서 관측 혹은 응답을 받는 과정에서 응답확률이나 측정확률이 발생된다. 셋째, 주어진 조사단위에서 데이터 생성 과정에 관한 확률이 있다. 응답값 혹은 관측값에 대한 데이터 생성확률이 발생된다.

포괄적으로 보면 표본조사에서는 위의 세 가지 확률이 모두 발생할 가능성이 있다. 그러나 표본조사를 통하여 추론하고자 하는 모수가 분석적(analytic) 모수가 아니라 기술적(descriptive) 모수인 경우에 측정오차를 최소화하고 조사단위가 갖는 값을 상수로 가정하면 측정확률과 데이터 생성확률은 발생하지 않는 것으로 간주할 수 있다. 따라서 이러한 가정이 성립되면 표본조사에서 다루어야 할 주요한 확률로는 표본추출확률만 남는다. 따라서 확률표집이론에서 별다른 언급이 없으면 확률은 표본추출확률을 뜻하게 되는 것이다. 그리고 확률표집은 확률화에 의하여 생성된 확률분포를 이용하여 모집단으로부터 표본을 선정하는 것을 말한다. 좀 더 기술적으로 말하면 확률표집은 다음의 두 가지 기준을 만족해야 한다. 첫째, 조사단위는 확률적으로 선정되어야 한다. 둘째, 표본추출틀의 모

든 조사단위는 양수의 포함확률을 가져야 하고, 또한 이러한 확률은 계산이 가능해야 한다(Statistics Canada 2003: 91; Sandal et al. 1992: 32). 비확률표집은 확률표집이 아닌 표집으로 정의될 수 있다. 이러한 정의에 의하면 만일 어떤 표본이 확률적으로 선정되지 않거나, 혹은 확률적으로 선정되더라도 조사단위의 포함확률을 계산할 수 없으면 그 표본은 비확률표본으로 간주된다. 또한 표본추출틀의 일부 단위가 표본에 포함될 수 없는 상황에서 추출된 표본은 비확률표본으로 간주된다. 간단하게 비확률표집을 "모집단으로부터 조사 단위를 주관적으로 선정하는 표본추출방법" (Statistics Canada 2003: 87)으로 정의하기도 하는데, 여기에서 주관적이라는 말의 의미는 위에서 설명한 바와 같은 내용을 포함한다.

비확률표집의 대표적인 예인 할당표집(quota sampling)에서는 할당 영역별로 사전에 정해진 수만큼만 표집을 하면 될 뿐 각 영역에서 할당된 수를 확률표집할 필요는 없다. 따라서 각 영역에서 뽑히는 표본은 비확률표본도 가능하다. 판단표집(judgement sampling)에서는 연구자의 주관적인 판단에 의하여 표본을 선정하므로 표본추출확률은 정의하기 어렵거나 계산이 불가능하다. 혹은 주관성에 의하여 표본추출틀의 일부 단위는 표본에 포함될 기회를 부여받지 못할 수도 있다. 자원자 표집(volunteer sampling)은 자원자의 주관적인 판단에 의하여 스스로를 표집하는 것이므로 포함확률이 1이거나 혹은 0이다. 이 경우 표집이 확률적이지도 않고 자원자가 거부하면 표본에 포함되지 않으므로 자원자 표본은 비확률표본이 된다.

Ⅲ. 비확률표본 활용의 역사적 사례

지금은 확률표본이 널리 받아들여지는 시대이기 때문에 표본이라고 하면 확률표본을 떠올리지만 표집이론이 발달해 온 과정을 돌이켜 보면 확률표본보다는 도리어 비확률표본이 더 먼저 사용되었음을 알 수 있다. Graunt(1662), Laplace(1814), Kiaer(1897) 등이 표본조사를 통하여 모집단 특성치를 추정한 사례가 문헌에 기록되어 있는데 이들이 사용한 표본을 지금의 관점에서 보면 모두 비확률표본이다. Graunt & Laplace의 예는 현대적 의미에서의 표본조사 이론체계가

성립되기 전의 일이므로 두 예에서 사용된 표본을 비확률표본이라고 구분하는 것은 적절하지 않다고 볼 수도 있겠으나 굳이 구분하자면 비확률표본이다. 현대적 의미의 표본조사를 시작한 주역으로 Kiaer를 언급하는 문헌이 많은데, Kiaer가 사용한 표본도 굳이 구분하자면 비확률표본이다. 표본조사에서 확률표집 개념이 공식적으로 언급된 해는 Lucien March가 논문을 발표한 1903년이라고 한다(Kruskal & Mosteller 1980: 177). 그리고 Rao(2005)는 'Hubback(1927)이 생산량 조사에서 확률표집의 필요성을 인식하였다'고 쓰고 있으며 잘 알려진 대로 표본조사에서 확률표집이 이론적 체계를 갖추기 시작한 것은 Neyman(1934)에 이르러서부터이다. 그 후에 확률표집이 주류가 되었고 별다른 언급이 없으면 표본은 확률표본을 의미하는 것으로 인식하였다. 그러나 표집이론 발달 초기에는 확률표본과 비확률표본이 병행하여 사용되었고, 한동안 확률표본이 대세를 이루고 있다가 최근에 이르러 비확률표본의 활용이 점점 더 늘어나게 된 것이다.

이번 절에서는 초기의 잘 알려진 비확률표본의 활용 사례를 간단하게 살펴보고, 왜 그 당시에 비확률표본이 받아들여지지 않았는지 고찰한다. 비확률표본이 받아들여지지 않은 이유를 이해하면 이를 통하여 최근에 당면하고 있는 비확률표본의 활용방안을 모색해 볼 수 있기 때문이다.

1. Graunt의 런던인구 추정

영국 상인이었던 John Graunt(1662)는 그의 사망자통계표에 관한 소책자에서 런던시민의 수를 추정하는 방법을 설명하고 있다. 당시 런던 시민의 수에 대해서는 경험에 의한 나름의 수치들이 언급되고 있었는데 Graunt는 이 수치에 의심을 품고 나름의 방법으로 런던 시민의 수를 추정하려고 시도하였다.

당시에는 교회에 출생, 사망 기록이 비교적 잘 정리되어 있었기 때문에 Graunt는 교회의 교구(parish)를 표본 단위로 하여 교구 내에 거주민 수와 사망자 수가 잘 정리된 교구의 거주민 수와 사망자 수를 조사하였다. 그의 조사에 의하면 한해 기준으로 평균적으로 11가족에서 3명이 사망한 것으로 나타났다. 그리고 한해 런던의 총 사망자 수는 약 13,000명인 것으로 조사되었으므로 이를 비례하여 계산하면 런던의 총 가족의 수는 약 $13{,}000 \times 11/3 = 48{,}000$이 된다. 그리고 평균적

인 가족 규모를 부부 2명, 자녀 3명, 하인이나 동숙자 3명 등 8명으로 간주하여 Graunt는 런던의 인구를 약 48,000 × 8 = 384,000명으로 추산하였다.

지금의 관점에서 보면 Graunt는 런던 인구의 추정에 표본조사 기법을 활용한 것이다. 표집단위는 교회의 교구였는데 몇 교구를 어떻게 선정하였는지에 대한 구체적인 언급은 없다. 단지 출생자 수와 거주민 수의 기록을 잘 보관하고 있는 교구를 표본으로 하였을 것이라고 짐작된다(Bethlehem 2009: 6). 응답 단위는 가족이었고 11가구에서 평균 3명이 사망하였으므로 표본가족의 평균 사망자 수는 $\bar{x}_s$ = 3/11, 그리고 런던의 사망자 총수는 약 t_x=13,000명이라고 하였으므로, 런던 시민의 총수(t_y)는 비추정법을 이용하여 추정하면

$$\hat{t}_y = \frac{\bar{y}_s}{\bar{x}_s} \times t_x = \frac{\bar{y}_s}{3/11} \times 13{,}000 = 48{,}000 \times \bar{y}_s$$

으로 계산된다. 여기에 당시 사회의 통념적인 가족규모, 즉 부부 2명, 자녀 3명, 하인 3명의 8명, $\bar{y}_s$ = 8,을 대입하면 Graunt가 계산한 수치인 384,000명을 얻는다.

런던 인구 추정치 384,000명의 신뢰도에 대한 언급은 없다. Graunt가 사용한 표본은 비확률표본이고 가족 규모 8명은 통념에 근거한 주관적인 수치이므로 이를 근거로 Graunt의 추정량 $\hat{t}_y$의 편향성이나 신뢰도를 측정하기는 어렵다. 그러나 Graunt는 다음과 같은 두 가지 새로운 개념을 발명하여 표집이론에 제공하는 기여를 하였다. 첫째, Graunt는 인구나 사회지표가 시간적으로나 공간적으로 안정성이 유지된다는 사실을 관찰하고 런던 인구 계산에 적용하였다. Graunt는 거주민 수와 사망자 수의 비율이 안정적으로 유지되고 있다는 사실을 간파하고 이러한 사실을 비추정량에 활용하였다. 이러한 안정성 가정은 지금도 매우 중요한데, 만일 이러한 안정성 가정이 없다면 데이터에 기초한 통계적 추론은 그 정당성을 부여받지 못할 수도 있다. 둘째, Graunt는 위의 안정성 가정을 전제로 구체적인 평균 지표를 제시하였는데, 그가 제시한 지표로는 가족당 사망자 수는 3/11명이라는 것과 평균 가족 규모는 8명이라는 것이다(Bethlehem 2009: 6). 지금은 통계지표가 범람하는 시대이기 때문에 통계지표를 당연한 것처럼 받

아들이지만 당시로서는 매우 획기적인 제시였다.

2. Laplace의 프랑스 인구 추정

표본조사를 통하여 통계 수치가 제공된 또 다른 사례는 Graunt 이후 100여년 후에 나타났다. 프랑스의 저명한 수학자인 Laplace가 제공한 프랑스 인구 추정이 그 사례이다. Laplace는 프랑스 인구를 파악하기 위해 Graunt가 사용한 방법과 유사한 방법을 시도하였다. 먼저 프랑스 전역에서 30개 지역을 선정하고, 선정된 지역에서 다시 지역 공동체(commune)를 선정하였다. 그 과정에서 두 가지 선정 기준이 고려되었는데 첫째, 모든 종류의 서로 다른 기후가 고려되었고, 둘째, 정확한 정보를 제공하는 지방정부가 고려되었다. Laplace는 기후대가 다르면 거주민의 밀도가 달라질 수 있다고 본 것 같고, 통계의 정확도를 위해서는 인구 관련 정보를 정확하게 제공받아야 한다고 판단한 것으로 보인다. 이렇게 해서 조사한 결과 1802년 9월 23일을 기준으로 표본 지역 공동체에 거주하는 주민의 수는 2,037,615명이었다. 또한 1800년, 1801년, 1802년 동안 출생한 신생아 수는 남자 110,312명, 여자 105,287명으로 총 215,599명이었다. 연 평균 215,599/3≈71,866.3명이 출생한 셈이다. 따라서 신생아 1명당 거주민 수는 2,037,615/71,866.3≈28.352명으로 계산된다. 따라서 만일 프랑스에서 한해 1,000,000명이 출생하고 Laplace의 신생아 1명당 거주민의 수 28.352명을 적용하면 거주민 총수는 28,352,000명으로 추정된다(Laplace 1812, 영역본 1951: 67).

현대 표집이론의 관점으로 보면 Laplace는 표집방법으로 2단계 집락표집 방법을 사용한 셈이다. 1단계에서는 기후대를 고려하여 지역을 선정하였고, 2단계에서는 선정된 각 지역에서 자료 제공 협조도가 높은 지역 공동체를 선정하였다. 그리고 추정 방법으로는 비추정법을 사용하였는데, 인구 추정에 대한 보조변수로 Graunt가 사망자 수를 이용한데 반하여 Laplace는 출생자 수를 이용한 점이 다르다. 보조변수 x_{tk}를 t연도의 k지역 공동체의 출생자 수라고 하고 y_{tk}를 t연도의 k지역 공동체의 거주민 수라고 하면 Laplace가 제시한 프랑스 인구 추정치는 다음과 같이 계산된다.

$$\hat{t}_y = \frac{\sum_{k \in s} y_{1802,k}}{(\sum_{k \in s} x_{1800,k} + \sum_{k \in s} x_{1801,k} + \sum_{k \in s} x_{1802,k})/3} \times t_x$$
$$= \frac{2{,}037{,}615}{215{,}599/3} \times t_x = 28.352 \times t_x$$

여기에서 s는 지역 공동체 표본을 뜻한다. Laplace는 구체적으로 프랑스 출생아 총수인 t_x의 값은 제시하지 않았고, 단지 t_x = 1,000,000이라면 $\hat{t}_y$ = 28,352,000이 된다는 예를 보였다.

표집과정의 1단계에서 각 기후대를 고려하여 지역을 선정한 방법은 표본의 유의 선정(purposive selection)에 해당되고, 2단계에서 지방 정부의 협조도가 높은 지역 공동체를 선정한 것은 판단표본에 해당한다. 둘 다 비확률표집 방법이다. 또한 Laplace는 3개년의 출생아 수만을 인구 계산에 반영하였는데, 이러한 가정은 닫힌 모집단(closed population)을 가정하는 경우로 프랑스 경계 내에서 사망과 지역 간 인구의 이동이 조사 기간 동안 없다는 것을 가정하는 것이다 (Write & Farmer 2000: 4). 또한 Laplace는 인구 관련 지표가 안정성이 있다고 가정하였다. Graunt와의 차이는 보조변수로 Graunt가 사망자 수를 사용한 데 반하여 Laplace는 3개년 출생자 수를 사용한 점이다. 결과적으로 수치만 비교하면 Graunt는 사망자 1명당 거주민 수를 8×11/3 = 29.3명으로 산정하였고 Laplace는 출생자 1명당 거주민 수를 28.3명으로 계산하였다. 놀랍게도 불과 1명 정도만 차이가 나는 유사한 수치가 제시된 셈이다.

Graunt에 비하여 Laplace가 더 진보한 점은 추정치의 신뢰도에 대한 언급을 하고 있는 점이다. Laplace(1812, 영역본 1951: 67)는 “그렇다면 결정된 추정치가 주어진 (오차)한계를 넘어서 모집단 참값으로부터 벗어날 확률은 얼마인가?”하는 물음을 던지고 있다. 이 물음에 대한 답이 1812년 소책자에는 구체적으로 제시되어 있지 않는데, Bethlehem(2009: 7)에 의하면 Laplace는 중심극한정리를 이용하여 그의 추정량이 정규분포를 따름을 증명하였고 이를 통하여 오차한계를 계산하였다고 한다. 그러나 그가 사용한 지역 공동체 표본은 중심극한정리에서 가정하는 확률표본이 아니라 기후와 지방정부의 협조를 고려한 비확률표본이기

때문에 적절한 적용이었는지에 관한 의구심은 남는다.

Laplace의 지방 공동체 표본은 비확률표본이므로 확률분포를 가지고 있지 않다. 따라서 Laplace의 비추정량에 대한 표준오차와 같은 신뢰도 지표를 직접 계산하는 것은 불가능하고 대신 다른 적절한 확률분포를 이용하여 신뢰도 지표를 계산하여야 한다. Cochran(1978)은 Laplace의 데이터를 이용하여 신뢰도를 구하는 방법을 설명하였다.

먼저 표준적인 2단계 표집이론을 이용하는 방법을 살펴보자. Laplace의 표본은 2단계 표본이므로 각 단계에서 비록 비확률적인 방법으로 표본을 뽑았지만 확률표본으로 간주할 수 있으면 표준적인 2단계 표집공식을 이용할 수 있을 것이다. 프랑스 전역에서 지역 표본을 기후대를 고려하여 뽑았다고 했는데, 표본이 프랑스 전국에 걸쳐 산재하여 뽑혔다면 확률표본으로 간주할 수 있을 것이고, 각 지역에서 협조도가 높은 지방 공동체를 표본으로 선정했다면 지역 공동체의 분포를 확률분포로 간주할 수 있을 것이다. 그렇다면 근사적으로나마 표준적인 2단표집공식을 활용하여 표준오차 등을 계산하면 된다(Cochran 1978: 6).

다른 방법으로 프랑스 거주민 수에 대한 출생자 수의 비율 p를 구하는 문제를 이항 추정의 문제로 바꾸어 볼 수 있다. 프랑스 인구 대비 표본 수는 상대적으로 매우 작으므로 프랑스를 무한모집단으로 간주하고 지방 공동체 표본은 여기에서 뽑힌 확률표본으로 간주하자. 무한모집단은 전년도에 출생한 사람과 나머지 사람들로 구성되어 있다. 뽑힌 거주민 수를 y라 하고 이 중에 전년도 출생자 수를 x라 하면 출생자 수 비율 p는 $\hat{p}=x/y$로 추정할 수 있다. 그리고 t_x를 전년도 출생자 총수라고 하면, 프랑스 인구 총수는 $\hat{t}_y = t_x y/x = t_x/\hat{p}$로 추정된다. 이제 테일러 근사를 사용하면

$$\hat{t}_y = \frac{t_x}{\hat{p}} \approx t_x\left(\frac{1}{p} - \frac{1}{p^2}(\hat{p}-p) + \frac{1}{p^3}(\hat{p}-p)^2\right)$$

가 되고, $\hat{t}_y$의 근사분산은 $V(\hat{t}_y) \approx t_x^2(1-p)/yp^3$이 되며, p를 $\hat{p}=x/y$로 추정하여 대입하면 다음과 같은 $\hat{t}_y$의 분산추정량을 얻는다(Cochran 1978: 7).

$$\hat{V}(\hat{t}_y) = t_x^2 \frac{y(y-x)}{x^3}$$

이 추정량에 Laplace의 수치인 y = 2,037,615와 x = 71,866을 대입하면

$$\sqrt{\hat{V}\left(\frac{1}{\hat{p}}\right)} = \sqrt{\frac{y(y-x)}{x^3}} = 3.285$$

을 얻는다. 또한 프랑스 추정 인구의 95% 신뢰구간은

$$28.352 \times t_x \pm 1.96 \times 3.285 \times t_x$$

임을 알 수 있다.

3. Kiaer의 대표성 있는 표본

Graunt나 Laplace의 방법은 확률 개념을 포함하지 않았고 후속 연구로 이어지지도 않았다. 다만, 표본만으로도 모집단 총계를 추정할 수 있다는 사례를 제시한 것에 역사적인 의의가 있다. 실제로 19세기 후반까지 정부기관이나 민간기관에서 일반적으로 채택한 전국조사 방법은 완전계수 방법이었다(Bellhouse 1988: 2). 일부 표본조사가 있기는 했으나 매우 예외적인 상황이었고 조사지역 전체를 세어 보는 완전계수 방법이 대부분이었으며 Quetelete 같은 저명한 학자들도 대부분 완전계수 방법을 지지하였다(Stigler 1986: 164).

이러한 분위기에서 완전계수에 이의를 제기하고 나선 대표적인 사람이 노르웨이 통계청의 Anders Kiaer였는데, 그는 1895년까지 노르웨이에서 15년 이상 수많은 표본조사를 성공적으로 수행한 경험이 있었다. 이러한 경험을 바탕으로 그는 만일 표본이 모집단의 축소판(miniature)으로서 모집단 특성을 그대로 반영할 수 있도록 선정된다면 완전계수가 아닌 표본조사만으로도 유용한 정보를 제공할 수 있다고 주장하였다(Kiaer 1897: 51). 그리고 표본이 모집단의 축소판이 되도록 선정하는 절차를 대표성 있는 방법(representative method)이라고 명명

하였다. 그러나 불행하게도 그의 아이디어는 혁신적이었지만 완전계수가 널리 통용되던 당시 분위기를 뚫고 가기에는 너무 주관적이었고, 게다가 확률적으로 정교하지 못했다(Brewer 2013: 250).

Kiaer의 제안에 대한 초기 반응은 대부분 부정적이었는데, 그 이유는 부분조사(partial investigation)는 완전계수를 결코 대신하지 못할 것이라는 믿음에 근거를 두고 있었다(Bellhouse 1988: 3). 당시 사회연구를 하는 학파에서는 사회발전 과정이 너무 복잡하고 다양하기 때문에 Kiaer의 부분조사 방법으로는 완전계수를 대체하지 못한다고 주장하였다. 총조사 학파는 포함범위에 대한 우려를 지적하였고, 사례연구를 하는 모노그래프 학파는 가족과 같은 그룹의 행동을 연구하기 위해서는 가족 전체를 추적조사 해야 하기 때문에 일부조사만으로는 곤란하다는 우려를 나타내었다(Smith 1997: 29). 그럼에도 불구하고 Kiaer는 그가 수행한 표본조사에서 사용한 방법에 대한 논문을 학회 등을 통하여 집요하게 계속적으로 발표하였다(Brewer 2013: 250).

개념적으로 Kiaer의 대표성 있는 표본은 모집단의 근사적인 축소판을 의미한다. 그러나 이러한 표본을 어떻게 선정하는지에 대해서 Kiaer는 구체적으로 기술하지 않았다. 대표성 있는 표본선정 방법에 대한 Kruskal & Mosteller(1980: 175)의 해석은 다음과 같다. 첫째, Kiaer는 사회, 경제 조사에서는 지역, 마을, 도시의 구역, 도로, 그리고 이어서 가구를 계통적으로 선정하는 것을 생각하였다. 둘째, 선정 과정의 모든 수준에서 충분한 표본 크기를 확보할 것을 주장하였다. 셋째, 다양한 기준으로, 특히 지리적으로 표본은 산재할 것을 강조하였다. 이를 현대 표집이론의 용어로 바꾸면. Kiaer의 표본은 다단계 집락화를 한 후 계통선정한 표본이다. 그리고 표본 크기는 공표단위 기준으로 하위 수준의 통계 생산에 충분한 크기이다. Kiaer는 표본선정에서 확률표집과 유사한 언급을 하기는 하였으나(Kiaer 1897: 39) 확률표집 이론으로 발전시키지는 못했다.

4. Bowley의 확률표집과 유의선정

1895년 Kiaer의 방법이 발표된 이후 대표성 있는 표본을 선정하는 방법에 대한 논쟁이 시작되었다. 확률화 방법은 프랑스 통계학자인 Lucien March에 의해서

1903년에 처음 제안되기는 하였으나, 그는 확률표집 옹호자는 아니었다 (Bellhouse 1988: 3). 표본조사에 필요한 표본선정 방법을 실증적으로 연구한 사람은 영국의 경제학자인 A.L. Bowley였다. 그는 확률화 표집방법을 공개적으로 도입하고 단순확률표집 후에 중심극한정리를 적용하는 실증연구를 하였다(Bowley 1906). 또한 영국의 도시 레딩에서 수행한 가난 관련 표본조사에서는 계통표집 방법을 적용하였다(Bowley 1913). Bowley는 1926년에 확률표집과 유의선정에 관한 이론적인 모노그래프를 발표하였는데, Bowley의 확률선정 방법에서는 각 조사단위의 포함확률은 정확하게 같도록 하고 신뢰도 유지를 위하여 추정치에 유의한 편차가 나지 않도록 표본 수를 충분히 크게 하도록 하고 있다. 또한 유의선정에서는 집락표본을 선정하였는데, 표본선정 기준은 공변량의 표본특성과 모집단 특성의 유사성이었다. 그는 표본 특성치와 모집단 특성치를 비교함으로써 대표성 있는 표본인지 여부를 판단하였다. 그밖에 모노그래프는 비례배정을 한 층화표집, 조사변수와 공변량의 상관관계를 이용한 추정 방법 등을 소개하였다. Bowley의 모노그래프를 계기로 Kiaer의 대표성 있는 표본을 선정하는 방법으로 확률표집과 유의선정이 동시에 받아들여졌고 표준적인 표집방법으로 채택되었다(Bellhouse 1988: 7).

1895년에 발표된 Kiaer의 대표성 있는 표본조사 방법이 30년이 지나 Bowley가 논문을 발표한 1926년 이후에야 널리 받아들여진 이유는 현실적인 측면과 이론적인 측면으로 나누어 설명할 수 있다. 먼저 현실적인 측면으로는 표본조사 수요가 폭발적으로 증가한 것을 지적할 수 있다. 제1차 세계대전 이후에 각국 정부는 많은 정보를 필요로 하였고 또한 정부통계에 대한 수요도 폭발적으로 증가하였다. 이러한 수요는 시간과 비용이 많이 드는 완전계수에서 보다 더 신속하게 수행되는 표본조사로 조사의 패턴이 바뀌는 흐름을 견인하였다. 이론적인 측면으로는, Kiaer는 대표성 있는 표집 방법에 대한 명시적인 설명을 하지 않았고 이론적 추론틀을 제공하지 않아 표집으로 인한 불확실성 측도를 계산할 수 없었던 것에 비하여, Bowley는 확률표집과 유의선정 방법을 명시적으로 제안하면서 확률표집에서는 중심극한정리를 이용하고 유의선정에서는 공분산 관계를 이용하는 등 나름의 이론적인 추론틀을 제공하였기 때문이다. 1926년 이후 확률표집 방법과 유의선정 방법은 각각의 이론적 근거를 바탕으로 선택적으로 사용

되었고 1934년까지 심각한 비평에 직면하지는 않았다.

5. 균형표본과 Neyman의 확률표집

유의표본선정에 필요한 공변량으로 인구, 사회, 경제, 지리적 변수들이 이용되었고 이들 변수를 기준으로 모집단과 동일한 평균을 갖는 균형표본(balanced sample)을 선정하는 것이 유의표본선정의 주요 방법이었다. 균형표본 방법은 기준 변수의 평균만 모집단과 비교하면 되므로 유의표본선정방법 중 실행이 수월한 방법이다. 그러나 이러한 균형표본 방법이 늘 성공적인 것은 아니었다(Bellhouse 1988: 7). 예를 들어 Gini & Galvani는 1921년 이탈리아 총조사에서 214개 행정구역 중 29개 행정구역을 7개 공변량을 기준으로 하여 근사적으로 평균이 비슷하도록 선정하였다. 그런데 결과적으로 표본선정에 포함되지 않는 대부분의 다른 변수에서 균형표본의 평균과 모평균이 서로 유사하지 않음이 발견되었다(Langel & Tille 2011: 51). 이 사례는 유의선정된 균형표본이 대표성 있는 표본이 되지 못할 수 있음을 보여주고 있다. 즉, 균형표본 방법은 기준 변수에 대해서는 표본평균과 모평균이 일치하지만, 기준 변수 이외의 다른 변수에 대해서는 표본평균이 모평균과 상당히 다를 수도 있기 때문에 표본이 모집단의 축소판이어야 한다는 표본의 대표성 개념을 만족시키지 못하는 결과가 되는 것이다.

Bowley가 제안한 유의선정 이론은 분명하게 기술되어 있지 않아 해독이 쉽지 않지만(Neyman 1934: 573), Kruskal & Mosteller(1980)의 설명에 의하면 Bowley는 조사변수와 조절변수의 상관관계를 이용하여 유의선정 이론을 정밀하게 전개하려고 시도하였다고 한다. 그러나 불행하게도 그는 유의표본선정방법을 분명하게 기술하지는 못하였다고 한다. Neyman(1934)은 조사변수와 조절변수 사이에 선형관계가 성립할 때 유의선정은 잘 이루어진다고 설명하였다. 그리고 Gini & Galvani의 예를 균형표본에서 조절변수의 표본평균이 모집단 평균과 비록 같아지더라도 조사변수와 조절변수의 선형성은 만족되지 않는 반례로 제시하였다. 즉, 조사변수와 조절변수 사이에 선형성 가정이 만족되지 않을 때에는 유의표본 활용이 적절하지 않으므로 유의표본의 활용 범위는 매우 제한적이라는 설명이다.

확률표집을 사용하면 선택편향 문제는 발생하지 않고 조사변수와 조절변수 간의 선형성 가정 문제도 발생하지 않는다. 게다가 Neyman(1934)은 층화확률표집에서 모평균의 최량선형추정량을 찾아냈는데, 그 추정량은 각 층에 표본 수를 최적배정한 후 층별 표본평균들을 가중평균하여 구한 것이다. 이 과정에서 최량선형추정량은 효율성의 개념을, 최적배정은 불균등 표본추출확률의 개념을 확률표집이론에 도입한 결과가 되었다. 또한 확률화에 의한 표본추출분포로부터 추정량의 신뢰구간을 계산하는 원리가 제시됨으로써 확률화에 의하여 표본추출은 물론 통계적 추론까지 동시에 하는 이론체계가 정립되었다. 1930년대 후반에 Neyman의 아이디어는 널리 받아들여졌고, 조사를 실제로 수행하는 국가통계기관에서 채택함으로써 확률표집이론은 표본조사의 주류 이론으로 자리를 잡게 되었다.

6. 균형표집

확률표집이론이 정립되어 주류 이론으로 자리를 잡아감에 따라 균형표본 방법에 대한 비판은 보다 구체화 되었다. 첫째, Gini & Galvani의 예에서 보듯이 균형표본은 비록 조절변수에 대해서는 선택편향을 제거할 수 있다고 하더라도 표본선정에 관여되지 않는 다른 변수에 대해서는 선택편향의 문제에서 자유롭지 못한 한계가 있다는 점과, 둘째, 균형표본에 대한 확률구조가 분명하지 않아 추론의 근거가 희박하다는 점에서 비판이 제기되었다. 따라서 이에 대한 대응으로 균형표본을 확률적으로 뽑는 방법과 균형표본에 확률구조를 부여하는 방법에 대한 연구가 이어졌다.

균형표본을 확률적으로 선정하는 방법으로 기각표집(rejective sampling) 방법과 균형표집(balanced sampling) 방법이 제안되었다. 기각표집 방법은 Yates (1946) 등이 제안한 방법으로, 확률표본을 뽑는 작업을 반복적으로 시행하여 표본평균과 모평균을 비교한 후 두 평균의 차이가 가장 작은 표본을 균형표본으로 사용하는 방법이다. 이 방법은 많은 표본추출 시행을 필요로 하는 단점이 있으나, 이렇게 표본을 뽑으면 균형표본선정에 관여하지 않는 변수들에 대해서는 확률표집을 한 결과가 되어 선택편향이 제거되는 장점이 있다.

균형표집법은 Deville & Tille(1998, 2004)가 제안한 방법으로서 균형표본을 뽑는 방법을 체계적으로 일반화한 것이다. 종래의 개별 균형표본선정 방식은 먼저 표본을 뽑아서 평균을 비교해 본 후 표본평균이 모평균과 동일하면 균형표본으로 채택하고 그렇지 않으면 선정을 다시 하거나 혹은 처음부터 평균이 일치하는 조사단위를 찾는 것이다. 이를 체계화한 균형표집법에서는 먼저 조절변수를 기준으로 표본평균과 모평균을 같게 놓은 균형조건을 만든 후 균형조건을 만족하는 모든 표본으로 이루어진 균형표집설계(balanced sampling design)를 만든다. 그리고 이 균형표집설계를 이용하여 표본을 뽑는다. Deville & Tille는 균형표집으로 표본을 뽑은 방법으로 큐브(cube) 방법을 제안하였다. 균형표집설계에서는 모든 표본이 균형조건을 만족하므로 균형표집설계에 의해 뽑힌 표본은 모두 균형표본이다. 게다가 균형표집법에서는 균형표집설계가 있으므로 자연스럽게 이를 이용하여 통계적 추론을 하면 된다.

기각표집법과 균형표집법은 다소의 차이는 있으나 균형표본을 확률적으로 뽑음으로써 연구자의 선택편향을 배제하고 표본추출확률로 통계적 추론의 근거를 제공한다는 공통점이 있다. 그리고 균형표집법은 비록 균형표본에서 시작하였지만 확률표집의 성격을 도입함으로써 유의선정과 확률표집의 특징을 동시에 갖는 표집법이다.

Ⅳ. 비확률표집 방법론 고찰

앞 장에서는 비확률표본의 활용 사례를 시간 순으로 간단하게 살펴보면서 비확률표본에 제기된 비판의 내용을 알아보았다. 요약하면, 첫째는 비확률표본은 비확률적으로 선정되기 때문에 연구자의 선택편향으로부터 자유롭지 않다는 점이고, 둘째는 비확률표본이 확률분포와 연계되어 있지 않으므로 통계적 추론이 어렵다는 점이다. 따라서 비확률표집 방법론을 정립하려면 이 두 가지 질문에 대한 답을 해야 한다.

먼저, 선택편향을 살펴보자. 표본조사에서 사전적으로 선택편향을 제어하면서 비확률표본을 선정하는 방법이 알려진 바는 거의 없다. 대신, 비확률표본의

선택편향을 인지한다면 선택편향에 대한 대응은 두 가지로 나누어 생각해 볼 수 있다. 첫째, 표본추출 메카니즘이 추론에 영향을 미치지 않는 경우 비확률표본은 선택편향 문제를 유발하지 않는다(Ⅳ. 1). 둘째, 비확률표본을 사후적으로 분류한 후 분류된 영역별로 추론을 함으로써 선택편향을 줄이는 방법을 모색할 수 있다(Ⅳ. 2).

비확률표본에 확률분포를 연계하는 방법으로 세 가지 방법을 고찰한다. 먼저 Ⅲ. 6에서 소개한 균형표집 방법은 비확률표본에 확률구조를 부여하는 전형적인 해법이다. 조사변수에 확률분포를 가정하고 문제를 풀어가는 모형기반 방법도 잘 알려진 비확률표본 활용법이다. 또한 다른 방법으로는 비확률표본의 조사 단위에 가상의 추출확률을 부여하는 의사설계기반 방법이 있다(Ⅳ. 3).

1. 비확률표본 타당성의 조건

표본조사에서 조사자의 선택편향을 제거해야 하는 본질적인 이유는 선택편향에 의하여 표본의 특성과 비표본의 특성이 달라질 수 있기 때문이다. 표본조사에서는 표본의 특성이 비표본(non-sample)의 특성을 대신해야 하므로 두 집단의 특성이 다르면 표본에 근거한 모집단 추론은 그 타당성을 잃게 된다. 반대로 선택편향이 발생해도 추론에 문제가 발생하지 않는 경우는 조사자의 표본추출이 추론에 영향을 주지 않는 때이다. 추론에서 표집설계의 무시 가능성(ignorability)과 그것의 조건 등에 관해서는 폭넓게 논의된 연구가 있다(예를 들면, Little 1982; Rubin 1976; Smith 1983; Sugden & Smith 1984; Pfeffermann 1993 등).

$Y=(Y_1,\ldots,Y_N)'$을 조사변수 벡터라고 하고, $Y_s=\{Y_k,\ k\in s\}$를 응답된 조사변수, $Y_{\tilde{s}}=\{Y_k,\ k\notin s\}$를 응답되지 않는 조사변수라고 하고, $Z=(Z_1,\ldots,Z_N)$를 설계변수, $I=(I_1,\ldots,I_N)$를 표본 여부를 표현하는 표본 지시변수라고 하자. 통계적 추론의 근거는 응답된 조사변수 Y_s, 설계변수 Z, 그리고 표본 지시변수 I의 함수 $f(Y_s,Z,I)$이다. 표본 지시변수 I에 의하여 표본 s와 여집합인 비표본 $\tilde{s}$의 정보가 추론에 반영된다. 만일 확률화를 하지 않고 비확률표본을 사용하게 되면 표본 지시함수 I의 정보를 알 수 없으므로 추론은 조사변수 Y_s와 설계변수 Z의 함수 $f(Y_s,Z)$ 에 의해서만 이루어진다. 따라서 만일 $f(Y_s,Z,I)$ 에 근거한

추론과 $f(Y_s, Z)$에 근거한 추론이 동일한 결론에 도달한다면, 즉,

$$f(Y_s, Z, I) = f(Y_s, Z)$$

표본 지시함수 I는 무시가능하다. 즉, 표본추출 과정을 추론에서 무시할 수 있다(Pfeffermann 1993).

예를 들어 자원자 표본으로 구성된 웹 조사에서 자원자의 특성과 표본에 들어오지 않은 사람들의 특성이 유사하다면 비록 비확률표본이지만 이에 근거한 추론의 오류는 발생하지 않는다. 비슷한 예로 쇼핑몰 앞 조사에서 만일 쇼핑객이 완전히 무작위로 움직인다면 편의표본에 의한 추론의 오류는 발생하지 않는다. 물론 현실적으로는 표본의 특성과 비표본의 특성을 비교할 수가 없으므로 이론적으로 두 특성이 같은 경우가 언제인가를 따져서 추론을 진행해야 하는 어려움이 있기는 하다.

2. 표본대응

관측연구나 임상실험에서는 확률화에 의한 표본선정이 어려우므로 본질적으로 비확률표본을 이용한 통계적 실험을 하게 되는데 이로 인하여 늘 선택편향의 문제가 제기된다(Rosenbaum 2010: 153 등). 관측연구나 임상실험을 통하여 비교연구(comparative study)를 할 때 선택편향의 문제를 해결하기 위하여 주로 사용하는 방법 중의 하나는 표본대응(sample matching)을 하여 두 그룹의 특성을 비교하는 것이다. 즉, 공변량으로 조절 그룹을 형성한 후 조절 그룹을 기준으로 처리군의 표본과 조절군의 표본을 나누어 대응시키고 그 차이를 비교하는 것이다. 이때 공변량은 조사변수와 연관성이 있는 변수가 선정되어야 한다. 표본대응 방법을 사용하면 비록 처리군과 조절군의 표본 특성이 다소 다르더라도 표본대응을 하여 비교함으로써 선택편향으로 인한 추정의 편향을 줄일 수 있을 것으로 기대된다.

표본조사에서 비확률표본의 선택편향을 줄이는 방법으로 표본대응 방법을 적용하여 사용할 수 있다(Baker et al. 2013: 94). 목표모집단과 비확률표본 사이의 차이를 비교하기 위하여 목표모집단을 처리 그룹으로 간주하고 비확률표본을

대조 그룹으로 간주하여, 공변량을 기준으로 조절 그룹을 나눈 후 각 그룹에서 목표모집단 특성치와 비확률표본 특성치를 대응시키는 것이다. 각 조절 그룹에서 비확률표본을 모집단에 대응시키는 방법으로는 조사 전에 대응시키는 방법이 있고, 조사 후에 대응시키는 방법이 있다. 전자의 경우 대표적인 방법이 할당표집 방법이고 후자의 대표적인 방법이 사후층화 방법이다. 할당표집에서는 인구 통계적 조절변수를 사전에 선정한 후 각 조절 그룹의 모집단 특성치를 계산하고 여기에 대응하는 표본 수를 미리 할당한다. 그리고 각 조절 그룹에 할당된 표본 수 만큼 표본을 조사하는 것이다. 물론 이때 각 조절 그룹에서의 표집은 비확률표집을 포함하므로 할당표집은 비확률표집이다. 사후층화에서는 조사된 표본을 조절변수를 기준으로 사후적으로 조절 그룹으로 분류하고, 각 사후 층에서 표본 특성치와 모집단 특성치의 차이를 비교한다. 조절변수를 기준으로 비확률표본을 사후 층으로 분류한 후 모집단 특성치와의 차이를 비교하면 조절변수로 인하여 선택편향의 크기가 다소 줄어들 것으로 기대된다.

이론적으로 볼 때, 만일 표본 대응에 사용된 조절변수가 조사변수와 매우 밀접하게 연관이 있는 변수여서 각 조절 그룹에서 표본의 분포와 목표모집단의 분포가 유사하게 되면 선택편향은 줄어든다(Rosenbaum & Rubin 1983). 결국 표본대응을 비확률표본에 적용했을 때 선택편향 제거의 성공 여부는 공변량 조절변수가 조사변수와 얼마만큼 연관성이 있는지에 달려 있다. 만일 조사변수와 연관성이 없는 공변량이 조절변수로 사용되면 도리어 선택편향은 증가할 수도 있다(Kish 1987).

표본대응 방법을 표본조사의 비확률표본에 적용하면 비록 이론적으로는 선택편향을 줄일 가능성은 있다고 하더라도, 이 방법은 주로 인과분석 분야에서 개발되고 발전된 방법이므로 표본조사에 적용할 때는 다소의 차이가 있을 수 있음을 인식할 필요가 있다. 먼저, 인과분석에서는 추론의 목적이 분명하고 조사변수나 처리효과와 관련이 있는 대응변수의 식별에 표본대응의 초점이 맞춰진다. 따라서 조사변수 혹은 처리효과에 연관성이 높은 대응변수가 선택될 가능성이 높다. 반면, 표본조사에서는 관심 있는 조사변수의 수가 많고, 각 조사변수와 연관성이 높은 대응변수는 서로 다를 수 있으므로 단일 조사변수인 경우보다는 더 많은 대응변수가 필요할 수도 있다. 이러한 차이는 곧바로 표본조사에서 대응변수의

선택을 어렵게 만든다. 두 번째 차이는 추론의 대상 범위에서 나타난다. 비교연구나 임상실험에서는 비교결과의 내적인 타당성을 높이기 위하여 비교 대상을 매우 구체적인 소그룹으로 축소하려는 경향이 있다. 소그룹으로 비교 대상을 축소하는 것은 처리 그룹과 대조 그룹의 분포를 유사하게 하는 데 도움을 주어 선택편향을 줄이게 된다. 반면, 표본조사에서는 추정의 대상이 모집단 전체이다. 추론의 대상 범위를 모집단 전체로 확대하면 확대된 모집단은 이질적인 소그룹을 다수 포함하게 되어 처리 그룹(모집단)과 대조 그룹(비확률표본)의 분포의 유사성은 줄어들게 된다. 소그룹에서 유효했던 선택편향 상쇄효과가 모집단 전체에서는 제거될 수 있다. 세 번째는 표본조사에서는 주로 총계나 평균과 같은 점추정치가 중요한 반면, 비교연구에서는 처리군에서 계산되는 점추정치보다는 처리군이 대조군보다 효과적인가 하는 비교에 더 비중을 둔다(Baker et al. 2013: 40).

관측연구 분야에서는 표본대응에 관한 연구가 많이 이루어진 반면 표본조사 분야에서는 비록 몇몇 사례가 보고되고 있기는 하지만(예를 들면 Rivers 2007 등) 아직은 체계적으로 표준화 되어 있지 않다. 이는 확률표집이론이 하나의 이론체계를 구축하고 있는 모습과 대비된다.

3. 비확률표본을 이용한 통계적 추정

통계적 추정은 본질적으로 표본이 생성된 확률분포에 기초하므로 만일 어떤 표본이 그 표본을 생성한 확률분포를 알지 못하면 원천적으로 통계적 추정은 불가능하다. 비확률표본을 이용하여 모수 추정을 할 때 발생하는 문제의 원인도 본질적으로는 비확률표본을 생성한 확률분포를 모른다는 데 있다. 따라서 비확률표본을 이용한 추정 문제를 푸는 접근방법은 비확률표본과 연관된 확률분포를 만들어 내거나 아니면 근사적으로라도 찾아내는 것이다.

비확률표본을 반복적으로 생성할 수 있다고 가정해 보자. 그러면 반복추출에 의한 확률분포가 만들어지므로 이 경우는 확률표집에서와 마찬가지로 반복추출 확률분포에 의하여 통계적 추정이 가능해진다. 대표적인 사례가 균형표집이다. 균형조건, 예를 들면 $\sum_{k\in s} x_{kj}/n = \sum_{k=1}^{N} x_{kj}/N,\ j=1,\ldots,p$을 만족하는 표본이 균형표본이고 균형표본으로 이루어진 설계가 균형표집설계이므로, 균형표집설계

에 의하여 뽑힌 표본은 자연스럽게 균형조건을 만족하는 균형표본이다. 즉, 하나의 균형표본은 비확률표본이지만 균형표집에서는 각 균형표본을 확률적으로 추출할 수 있으므로 이 확률을 근거로 통계적 추정이 가능하다(Tille 2011).

비확률표본에 확률분포를 부여하는 다른 방법은 조사변수에 통계적 모형을 가정하는 방법이다. 조사변수 y_k와 설명변수 $(x_1,...,x_p)$ 사이의 관계를 설명하는 통계적 모형이 다음과 같다고 하다.

$$y_k = \beta_0 + \beta_1 x_{k1} + ... + \beta_p x_{kp} + \epsilon_k,\ \epsilon_k \sim (0,\sigma^2),\ k \in s$$

그러면 회귀분석 이론에 의하여 회귀계수 추정량 $\hat{\beta}_0,...,\hat{\beta}_p$을 얻을 수 있고, 모총계 t_y는 다음과 같이 추정할 수 있다.

$$\hat{t}_y = \sum_{k \in s} y_k + \sum_{k \notin s} \hat{y}_k = \sum_{k \in s} y_k + \sum_{k \notin s} \left(\hat{\beta}_0 + \hat{\beta}_1 x_{k1} + ... + \hat{\beta}_p x_{kp}\right)$$

또한 모총계 추정량의 통계적 성질은 회귀모형의 성질로부터 유도할 수 있다(Royall 1970; Valliant et al. 2000 등). 이러한 추정방법을 모형기반 추정법이라고 한다.

이처럼 균형표집법과 모형기반 추정법은 온전한 확률분포를 가지고 통계적 추정을 할 수 있는 장점이 있다. 반면 균형표집법에서는 큐브 방법을 이용하여 방법론 전개가 가능하나 절차가 간단하지 않는 현실적인 문제가 있고, 이미 많이 지적된 대로 모형기반 추정법에서는 가정된 모형이 조사데이터에 잘 부합한다는 가정이 성립할 때만 의미가 있는 단점이 있다.

현실적으로 비확률표본에 폭넓게 적용할 수 있는 방법으로 의사설계기반(pseudo design-based) 방법이 있다. 확률표집에서는 조사단위의 포함확률을 계산할 수 있지만 비확률표본에서는 포함확률이 정의되지 않거나 혹은 계산이 되지 않으므로 비확률표본의 조사단위에 대한 포함확률을 계산할 수가 없다. 따라서 비확률표본에 속하는 단위의 포함확률을 추정한 후, 추정된 포함확률을 사용하여 추론하는 것이 의사설계기반 방법이다. 이 방법은 비록 비확률표본이지만 표본의 각 조사단위가 표본에 뽑히지 않는 조사단위를 대표한다는 가정을 염두에 두고 있

다.

포함확률을 추정하는 방법은 응용사례에 따라 다를 수 있으나, 가장 기본적인 방법은 응답을 몇 개의 범주로 분류한 후 각 범주에서 응답 수와 모집단 수를 비교하여 추출률을 계산하는 방법이다. 범주를 사후층으로 간주하면 사후 층별로 표본추출률을 계산하는 것이 된다(Bakers et al. 2013: 67). 또 다른 예를 들면, 어떤 조사에서 확률표본과 비확률표본이 활용가능한 상황이고 유용한 공변량 정보가 있다고 가정하면 공변량 정보와 확률표본의 포함확률을 이용하여 비확률표본의 포함확률을 계산할 수 있다(Elliott 2009). 만일 I_k^*를 단위 k의 비확률표본 지시자, I_k를 확률표본 지시자로서 포함되면 1의 값, 포함되지 않으면 0의 값을 갖는다고 하자. 그리고 W는 공변량이라고 하면 다음이 성립한다.

$$P(I_k^*=1|W) \propto P(I_k=1|W)\frac{\hat{\phi}(I_k^*=1|W)}{1-\hat{\phi}(I_k^*=1|W)}$$

여기에서 $\hat{\phi}(I_k^*=1|W)$ 는 단위 k가 비확률표본에 포함될 확률을 공변량 W가 주어졌을 때 로지스틱 회귀분석을 통하여 추정한 값이다. 이 공식에 의하면 비확률표본에 속한 단위 k의 포함확률은 확률표본에 포함될 확률에 비례하며, 그 비는 로지스틱 회귀분석 등을 통하여 추정됨을 알 수 있다. 의사설계기반 방법으로 구한 포함확률은 기본기중치를 계산하는 데 사용된다.

Ⅴ. 결론

이 연구에서는 표본조사에서 비확률표집 방법론을 고찰하였다. 포함범위의 감소, 무응답률의 증가, 조사비용의 증가 등으로 표본조사의 주류이론이었던 확률표집이론이 심각한 도전에 직면한 상태이고 현실적으로는 웹 조사와 같은 비확률표본에 의한 표본조사의 수가 점점 늘고 있는 상황에서, 비확률표본을 이용한 표본조사가 수용 가능한 대안이 될 수 있는지 여부를 살펴보았다. 비확률표본의 약점은 연구자의 선택편향을 제어하지 못한다는 점과, 비확률표본과 연계된

확률분포가 뚜렷하지 않아 통계적 추론이 어렵다는 점이다. 이러한 약점에 대하여 무시가능한 표집설계의 경우 선택편향이 발생하지 않으며, 관측연구나 임상실험에서와 같이 표본대응 방법을 이용하면 선택편향을 줄일 수 있음을 고찰하였다. 또한 균형표집 방법과 모형기반추론 방법, 그리고 의사설계기반 방법을 비확률표본을 이용한 통계적 추론 방법으로 사용될 수 있음을 살펴보았다.

20세기 초에 조사 양식(pattern)이 완전계수에서 표본조사로 전이된 것은 이론의 획기적인 개발보다는 완전계수로는 감당하기 어려운 조사 수요의 폭증 때문이었다. 21세기에 진입한 현재 확률표본에 기반한 표본조사의 양도 증가하겠지만 비확률표본을 이용하는 표본조사도 점점 더 증가할 것으로 예상된다. 서론에서 언급하였듯이 표본추출틀의 과소포함, 무응답률 증가, 증가하는 조사비용 등의 복합적인 원인 때문이다. 확률표집이론과는 달리 모든 비확률표집을 아우르는 단일한 추론틀이 아직까지는 정립되어 있지 않기 때문에 비확률표본조사는 아직까지는 논쟁적이다. 그럼에도 불구하고 만일 21세기에 표본조사의 양식이 확률표본을 이용하는 조사에서 비확률표본을 이용하는 조사로 전이된다면 그것은 한 세기 전과 유사하게 비확률표본을 이용하는 조사의 폭증하는 수요 때문일 가능성이 매우 높다. 그리고 이에 부응하여 비확률표집과 관련된 이론은 더 개발되고 보완될 것이다.

현재에도 비확률표본을 선정하는 과정을 투명하게 하고, 표본 대응과 모형설정 등 추론에 활용된 메타 정보를 상세하게 기록, 관리하여 비확률표본을 이용한 표본조사의 선택편향 가능성과 추론의 오류 가능성을 줄이는 노력(Baker et al. 2013)은 필요하다. 그러나 향후 비확률표본을 이용한 방법론이 조사연구 분야에서 폭넓게 수용되려면 앞서 논의한 바와 같이 비확률표집의 최대 약점인 선택편향의 문제를 해결하고 조사 이용자가 수용할 수 있는 내적 일관성(coherent)이 있는 단일한 추론틀이 마련되어야 할 것이다. 그리고 이러한 추론틀은 비확률표본조사 결과의 품질을 평가할 수 있는 측도를 포함하여야 할 것이다

이와 더불어 최근에는 비확률표본을 이용하는 조사만이 아니라 확률표본과 비확률표본을 동시에 이용하는 조사, 조사데이터와 행정데이터의 결합, 혹은 조사데이터와 빅데이터의 결합 등의 수요가 늘어날 전망이다. 사회에서 필요로 하는 정보의 수요는 폭발적으로 증가하는데 가용 재원은 제한되어 있으므로 이와

같은 다양한 형태의 정보수집 방법들이 모색되고 시도되는 것이다. 활발한 연구가 기대된다.

참고문헌

Baker, R., S.J. Blumber, J.M. Brick, M.P. Couper, M. Courtright, J.M. Dennis, D. Dillman, M.R. Frankel, P. Garland, R.M. Grove, C. Kennedy, J. Krosnick, P.J. Lavrakas, S.Lee, M. Link, K. Rao, R.K. Thomas, and D. Zahs. 2010. "AAPOR Report on Online Panels." *Public Opinion Quarterly* 74: 711-781.

Baker, R., J.M. Brick, N.A. Bates, M. Battaglia, M.P. Couper, J.A. Dever, K.J. Gile, and R. Tourangeau. 2013. "Summary Report of the AAPOR Task Force on Non-probability Sampling." *Journal of Survey Statistics and Methodology* 1: 90-143.

Bellhouse, D.R. 1988. "A Brief History of Random Sampling Methods." In *Handbook of Statistics* Vol 6, Sampling. Elsevier.

Bethlehem, J. 2009. "The Rise of Survey Sampling." *Statistics Netherlands Discussion Paper* #09015.

Bowley, A.L. 1906. "Address to the Economic Section of the British Association." *Journal of the Royal Statistical Society* 540-558.

Bowley, A.L. 1913. "Working-class Households in Reading." *Journal of the Royal Statistical Society* 672-701.

Brewer, K. 2013. "Three Controversies in the History of Survey Sampling." *Survey Method* 39: 249-262.

Cassel, C.M., C.E. Sarndal, and J.H. Wretman. 1977. *Foundation of Inference in Survey Sampling*. John Wiley and Sons.

Cochran, W.G. 1978. "Laplace's Ratio Estimator." In *Contributions to Survey Sampling and Applied Statistics*: 3-10.

Cox, D.R. 2006. *Principles of Statistical Inference*. Cambridge.

Deville, J.C. and Y. Tille. 1998. "Unequal Probability Sampling without Replacement through a Splitting Method." *Biometrika* 85: 89-101.

Deville, J.C. and Y. Tille. 2004. "Efficient Balanced Sampling: The Cube Method." *Biometrika* 91: 893-912.

Elliott, M.R. 2010. "Combining Data from Probability and Non-probability Samples Using Pseudo-weights." *Survey practice* 2:1-7.

Graunt, J. 1662. *Natural and Political Observations Mentioned in a Following Index, and Made upon the Bills of Mortality*, London.

Kiaer, A.N. 1897. *The Representative Method of Statistical Surveys*. Kristinia, Oslo.

Kish, L. 1965. *Survey Sampling*. John Wiley and Sons.

Kish, L. 1987. *Statistical Design for Research*. Wiley.

Kruskal, W. and Mosteller, F. 1980. "Representative Sampling, IV: The History of the Concept in Statistics, 1895-1939." *International Statistical Review* 48:169-196.

Langel, M. and Y. Tille, 2011. "Corrado Gini, a Pioneer in Balanced Sampling and Inequality Theory." *Metron LXIX*: 45-65.

Laplace, M.D. 1814(English translation 1951). *A Philosophical Essay on Probabilities*. Dover.

Littel, R.J.A. 1982. "Models for Nonresponse in Sample Surveys." *Journal of the American Statistical Association* 77: 237-250.

Neyman, J. 1934. "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection." *Journal of Royal Statistical Society*, *Series A* 64: 558-625.

Pfeffermann, D. 1993. "The Role of Sampling Weights When Modeling Survey Data." *International Statistical Review* 61: 317-337.

Rao, J.N.K. 2005. "Interplay between Sample Survey Theory and Practice: An Appraisal." *Survey Methodology* 31: 117-138.

Rivers, D. 2007. “Sampling for Web Surveys.” Joint Statistical Meeting, Salt Lake City, Utah.

Rosenbaum, P.R. 2010. *Design of Observational Studies*. Springer.

Rosenbaum, P.R. and Rubin, D.B. 1983. “The Centural Role of the Propensity Score in Observational Studies for Casual Effects.” *Biometrika* 70: 41-55.

Royall, R.M. 1970. “On Finite Population Sampling Theory under Certain Linear Regression Models.” *Biometrika* 57: 377-387.

Royall, R.M. 1983. “Comment on an Evaluation of Model-dependent and Probability-sampling Inferences in Sample Surveys.” *Journal of the American Statistical Association* 78: 794-796.

Rubin, D.B. 1976. “Inference and Missing Data.” *Biometrika* 63: 581-592.

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

Smith, T.M.F. 1983. “On the Validity of Inferences from Non-random Samples.” *Journal of the Royal Statistical Society Series A* 146: 394-403.

Smith, T.M.F. 1997. Social Surveys and Social Science. *The Canadian Journal of Statistics* 25:23-38.

Statistics Canada. 2003. *Survey Methods and Practices*. Catalogue no. 12-587-X.

Stigler, S.M. 1986. *The History of Statistics*. Belknap Harvard.

Sugden, R.A. and T.M.F. Smith. 1984. “Ignorable and Informative Designs in Survey Sampling.” *Biometrika* 71: 495-506.

Tille, Y. 2011. “Ten Years of Balanced Sampling with the Cube Method: An Appraisal.” *Survey Methodology* 37: 25-226.

Valliant, R., A.H. Dorfman, and R.M. Royall. 2010. *Finite Population Sampling and Inference: A Prediction Approach*. Wiley.

Wright, T. and J. Farmer. 2000. “A Bibliography of Selected Statistical Methods and Development Related to Census 2000.” *Bureau of the Census Statistical Research Report Series,* No. RR2000/02.

Yates, F. 1946. “A Review of Recent Statistical Developments in Sampling and Sampling Surveys.” *Journal of the Royal Statistical Society*, *Series A* 109: 12-43.

<접수 2016/11/24, 수정 2016/11/28, 게재확정 2016/12/16>