비확률표본을 위한 통계적 추론

권순필
통계개발원 사무관
pilsogood@korea.kr

오늘날 우리는 방대한 양의 데이터를 활용할 수 있는 시대에 살고 있다. 인터넷, 소셜미디어, 모바일 앱, 각종 센서와 기기가 생성하는 데이터는 엄청난 속도로 축적되고 있으며, 웹조사 같은 조사도구의 발전 덕분에 과거보다 훨씬 더 많은 응답자를 대상으로, 저렴한 비용으로 조사가 가능해졌다. 이를 바탕으로 우리 사회와 세상에 대한 통계적 분석도 한층 더 깊어지고 있다. 하지만 이러한 데이터들은 대부분 확률적으로 수집된 표본이 아니다.

확률표본(probability sample)은 모집단의 모든 조사 단위가 알려진 확률로 표본에 포함될 가능성을 가지는 표본을 의미한다. 이를 통해 모집단의 특성을 정확하게 추론할 수 있기 때문에 공식 통계를 비롯한 다양한 연구 분야에서 오랫동안 널리 사용되어 온 방식이다. 또한 확률표본을 사용하면 비교적 작은 표본만으로도 대규모 모집단에 대한 유효한 통계적 추론이 가능하다.

반면 자발적으로 참여한 사람들, 편의에 의해 선택된 집단, 또는 불규칙적으로 수집된 비확률표본(non-probability sample)은 확률표본보다 대표성이 떨어지기 때문에 이들 표본으로 모집단을 추정하게 되면 불확실성이 커질 수밖에 없다. 그럼에도 불구하고 비확률표본의 활용은 계속해서 중요해지고 있으며, 관련 통계적 추론 방법의 발전도 필수적이다. 이 글에서는 최근 3년간(2022~2024) 통계개발원에서 수행한 비확률표본을 이용한 통계적 추론의 다양한 방법론을 소개하고, 연구의 필요성과 한계 등을 정리한다.

이론적으로 확률표본이란 ① 표본추출틀의 조사 단위가 확률적으로 선정되며, ② 모든 조사 단위의 포함확률이 양수로, ③ 표본의 포함확률이 계산 가능해야 한다. 이를 위해 우리는 모집단을 포괄하는 표본추출틀이 필요하며, 표본설계에 따라 각 조사 단위는 알려진 확률로 선정되어야 한다(그림 1 참고).

확률표본 방식은 변수가 특정 분포를 따른다는 가정을 하지 않으며, 표본 추출 과정에 연구자의 주관이 개입하지 않기 때문에 모집단을 대표할 수 있는 신뢰성 높은 결과를 제공한다. 다만 확률표본은 고비용 구조로 인해 통계청이나 국책연구소 같은 대규모 기관에서 주로 수행된다.

확률표본 조사는 지난 100년간 견고한 이론을 바탕으로 실무적으로 광범위하게 발전해 왔기 때문에 이 시기를 "확률표본조사의 황금기"로 부르기도 한다. 그러나 최근 들어 확률표본의 유지가 어려워지고 있다. 표본추출틀의 포함범위 감소, 무응답 증가, 조사비용의 급증, 그리고 코로나19 팬데믹과 같은 외부 환경 변화가 이러한 어려움을 가중시키고 있다.

[그림 1] 확률표본과 비확률표본

확률표본	비확률표본
표본추출틀의 조사단위는 확률적으로 선정	표본추출틀의 조사단위는 비확률적으로 선정(선택편향)
모든 조사단위의 포함확률이 양수	일부 조사단위의 포함확률이 0 (과소포함)
포함확률 계산 가능	포함확률 계산 불가능 (미지의 추출확률)

2023년 상반기 가계동향조사의 응답률은 55.6%로 2022년보다 약 7%p가 감소하였다. 영국 통계청은 고용통계 응답률 하락으로 인해 세부 통계 발표를 중단하기도 했다. 2014년에 50% 안팎이던 응답률이 2023년 말에 14.6%까지 하락했기 때문이다. 이런 일련의 상황 때문에 확률표본을 기반으로 하는 공식통계의 신뢰성에 대한 우려가 커지고 있다.

확률표본의 위기와 함께 비확률표본에 대한 관심과 활용 가능성도 커지고 있다. 비확률

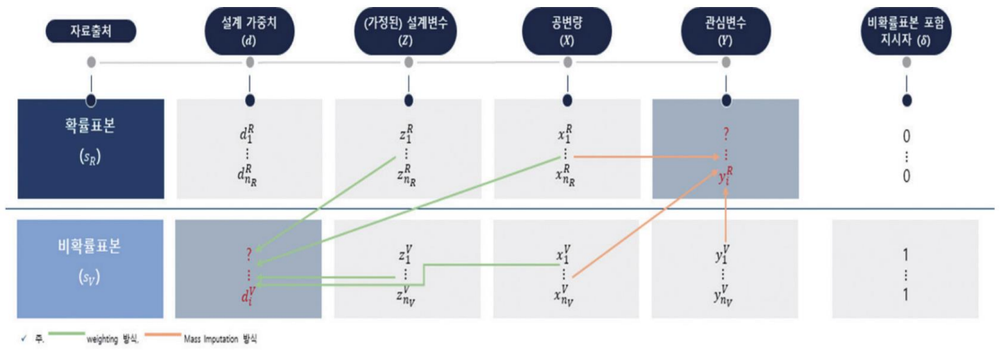
표본은 확률적으로 표본을 추출하지 않은 모든 표본을 말한다. 할당표본, 편의표본, 웹표본 외에도 행정자료, 거래자료, 센서자료, 인터넷자료와 같은 다양한 데이터들이 비확률표본으로 분류될 수 있다.

비확률표본은 확률표본에 비해 저비용으로 대량의 정보를 실시간으로 얻을 수 있는 장점이 있다. 급변하는 사회변화에 대응하기 위해 시의성 있는 통계 수요가 급증하면서, 비확률표본을 활용한 공식통계의 생산과 연구가 더욱 주목받고 있다.

그러나 비확률표본은 모집단에 대한 대표성을 장담할 수 없으며, 잘못된 방법으로 다룰 경우 표본 선택 편향(selection bias) 문제가 발생할 수 있다. 편향이 보정되지 않으면 데이터가 클수록 오히려 잘못된 결론에 이르는 빅데이터의 역설[1]에 빠질 수 있다.

통계개발원은 비확률표본의 대표성과 편향 문제를 해결하기 위해 다각도로 연구를 수행하고 있다. 고품질의 확률표본을 비확률표본과 통합(그림 2 참고)하여 비확률표본의 선택 편향을 보정하고, 미지의 추출 확률을 추정하는 방식[2]이 그 예이다. 이를 위해 확률표본과 비확률표본 사이에 유용한 공변량이 존재하는 것이 전제된다.

[그림 2] 확률표본과 비확률표본을 통합하는 통계적 추론 방안



1) "Big data paradox: the bigger the data, the surer we fool ourselves"(Meng, 2018).

2) 구체적인 방법론에 대해서는 통계개발원의 연구보고서 권순필 등(2023). 「비확률표본을 위한 통계적 추론」, 권순필 등(2024). 「비확률표본을 위한 통계적 추론: 실증연구」, 권순필 등 (2025) 「비확률표본을 위한 통계적 추론: 시계열 안정성 검토(발간예정)」 등을 참고할 수 있다.

〈그림 2〉를 통해서 우리는 확률표본과 비확률표본을 통합하는 것이 마치 확률표본의 결측 문제를 처리하는 것과 유사함을 알 수 있다. 실제로 비확률표본을 활용한 통계적 추론 방법들은 결측을 처리하는 형식으로 제안되어 발전해 왔다. 비확률표본으로 모집단을 추론하기 위한 접근 방식으로는, 첫째로 〈그림 2〉의 초록색 화살표처럼 비확률표본과 확률표본의 보조정보(x, z)를 이용해 비확률표본의 가중치를 추정하는 방법이 있다. 대표적으로는, 추정된 비확률표본의 성향점수(propensity score)의 역수를 가중치로 사용하는 방법(inverse probability weighting; ipw)과 알려진 모집단 혹은 확률표본의 분포에 비확률표본의 분포를 맞추는 보정(calibration; cal) 방법이 있다. 둘째로 주황색 화살표처럼 비확률표본의 관심변수(y)와 보조변수(x)의 관계를 활용해 확률표본의 미관측 관심변수를 통으로 대체하는 방법(Mass Imputation)이 있다. y와 x의 관계는 주로 회귀모형(regression model; reg)을 사용하여 식별한다. 셋째로 성향점수모형과 회귀모형을 결합하는 방법이 있다. 이 방법은 성향점수모형이나 회귀모형의 오식별에 보다 강건(doubly robust; dr)할 수 있다.

실증 검토를 위해서 가계금융복지조사를 이용한 모의실험을 수행하였다. 가계금융복지조사는 전국의 약 2만 가구를 표본으로 하는 대규모 조사로 연속형, 범주형 등 다양한 데이터를 수집한다. 이를 모집단으로 하여 확률표본은 단순임의표본(Simple Random Sample), 비확률표본은 연령이 높을수록 참여확률이 낮아지게 추출하였다. 관심값은 연간가구경상소득의 평균이며, 보조변수는 가구주의 인구특성과 가구원수 등이다. 확률표본과 비확률표본의 크기(400, 800, 1,000) 등을 변화시키는 다양한 시나리오를 가정하고, 시나리오별로 추정값의 편향(%RB), 평균제곱오차(MSE), 추정된 평균의 95% 신뢰구간이 모평균을 얼마나 포함하는지(%CP)를 계산했다. %RB와 MSE는 작을수록, %CP는 95%에 가까울수록 추정량의 성능이 좋다고 말할 수 있다.

〈그림 3〉은 모의실험 결과로, 각각 확률표본 크기가 고정된 상태($n_R = 400, 800, 1{,}000$)일 때 비확률표본의 크기 변화($n_V = 400, 800, 1{,}000$)에 따른 추정 결과를 나타낸다. 예상한 대로 비확률표본을 SRS인 것처럼 다루는 경우(naive)에는 선택편향 문제가 심각하게 발생하고, 편향을 조정한 4개의 평균 추정량(ipw, reg, dr, cal)은 모두 상대편향 및 평균제곱오차가 감소하였다. 그중에는 dr 추정량과 cal 추정량의 편향이 가장 작았다. 그러나 표본의 크기에

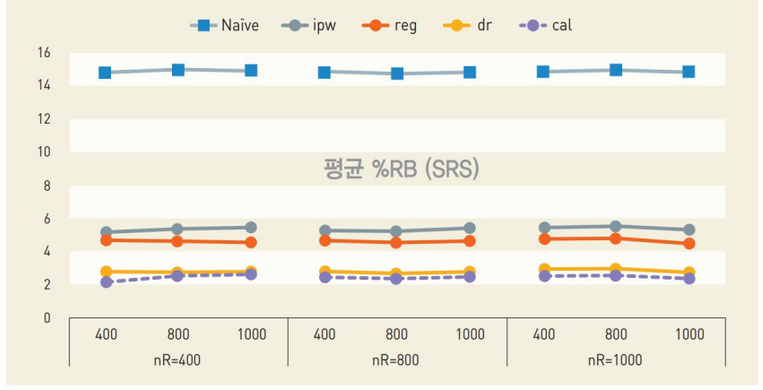
따른 편향 감소 변화는 크지 않았다.

추정량의 95% 신뢰구간의 모평균 포함확률은 80% 전후로 나타났는데, 표본의 크기가 커질수록 포함확률이 크게 감소하는 것을 볼 수 있다. 편향이 남아 있는 상태에서 추정된 신뢰구간은 모수를 포함할 확률이 줄어드는데, 비확률표본의 크기가 증가할 때는 편향보다 분산의 감소 속도가 훨씬 빨라서 이런 경향이 더 크게 나타나기 때문이다. 본 사례에서도 비확률표본 추론 시 비편향을 전제로 한 신뢰구간을 품질지표로 사용한다면 표본의 크기가 클 때 오히려 잘못된 정보를 제공할 수 있다는 빅데이터 역설(bigdata paradox)이 나타난 것을 알 수 있다. 이는 비확률표본의 특성에 맞는 품질지표에 대한 논의가 필요하다는 것을 시사한다.

[그림 3] 비확률표본 평균 추론 모의실험 결과

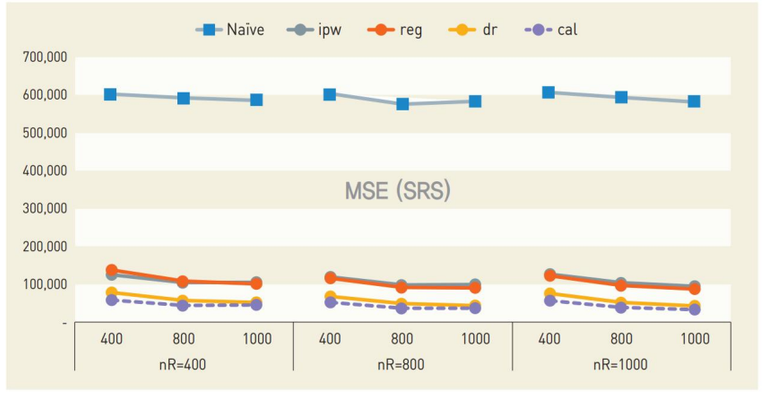
〈(a) 상대편향〉

(단위: %)



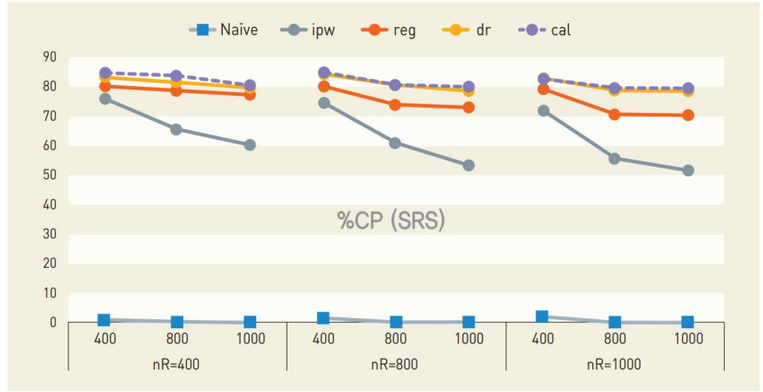
〈(b) 평균제곱오차〉

(단위: 만 원)



〈(c) 95% 신뢰구간의 모평균 포함확률〉

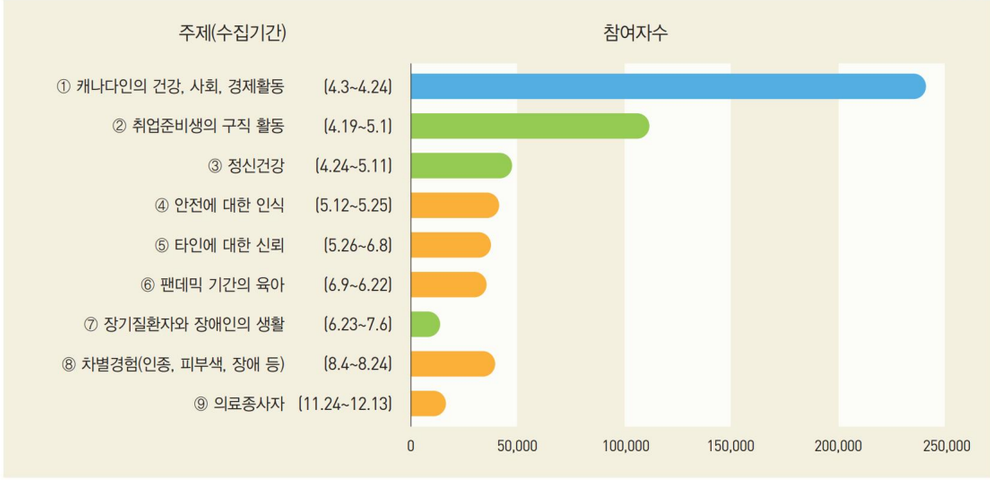
(단위: %)



비확률표본을 이용한 모집단 추론에 대한 연구는 통계 선진국에서도 진행하고 있다. 캐나다 통계청(StatCan)은 다양한 사용자의 통계자료 요청에 신속하게 대응하기 위해 Crowdsourcing을 통해 자원자 표본을 대상으로 웹조사를 수행하였다. [그림 4]와 같은 자료 수집을 통해 코로나 팬데믹 기간에 코로나19가 다양한 캐나다인 그룹의 삶과 복지에 어떤 영향을 미치는지 알 수 있었다.

[그림 4] 캐나다 Crowdsourcing 주제 및 참여자

(단위: 명)



StatCan은 결과 해석과 관련하여 주의사항을 통해, 자원자 표본을 이용한 웹조사의 경우 관심변수는 비율(proportion)로 제한해야 하고, 변이계수(CV), 신뢰구간, 오차한계 등과 같은 데이터 품질지표가 없기 때문에 모집단 대표성을 주장할 수 없다는 한계점을 명시하고 있다.

비확률표본을 이용한 통계적 추론은 빠르게 변화하는 데이터 환경에서 더욱 중요해지고 있다. 확률표본을 유지하기 어려운 상황에서 비확률표본의 유용성은 커지고 있으며, 비확률표본을 적절하게 보정하고 활용할 수 있는 방법론들이 계속해서 발전하고 있다. 100년 전 전수조사에서 확률표본조사로의 전환이 시의성과 비용 효율 요구에서 이루어졌듯, 오늘날에도 외부의 여러 요인 때문에 비확률표본이 부각될 수밖에 없을 것이다.

비 확 률 표 본 을 위 한 통 계 적 추 론