

가중치 산정

가. 가중치 산정 개요

통계조사에서 가중치 부여는 표본추출에 따른 추출률의 차이와 응답률 및 모집단에 대한 정보 등을 이용하여 모집단의 구조와 표본 구조를 맞추으로써 추정의 정확도를 높이는 것을 목적으로 한다. 추정단계에서 가중치를 이용하면 모집단에 대한 특성치인 모수에 대한 비편향 추정량(unbiased estimator)을 얻을 수 있다. 만약 통계 분석 과정에서 가중치를 무시하고 분석한 추정치는 심각한 편향(bias)이 발생할 수 있다. 표본의 크기가 큰 대규모 조사에서 문제가 되는 것은 추정량의 편향이기 때문에 추정과정에서 반드시 가중치를 이용해야 한다.

일반적으로 표본조사의 가중치는 ㉠ 설계가중치 산정, ㉡ 무응답에 대한 조정¹⁾, ㉢ 모집단 정보를 이용한 조정 등의 세 가지 과정을 통해서 산정된다. 가중치 조정단계에서 이용하는 모집단 정보는 2023년 가구 및 인구 추계값이다.

표본설계에서는 각 시도 구분과 도 지역 내의 동 및 읍·면 구분을 이용한 세부 층화를 통해서 전체 층을 구성하였다. 표본추출은 층화 2단계 확률비례계통추출법을 적용하였는데, 1차 추출단위는 조사구이고, 2차 추출단위는 가구 및 만 15세 이상 가구원이다. 각 층에서 배정된 표본 조사구 수만큼을 조사구 내의 가구 수에 비례하는 확률비례계통추출법에 따라 추출하였다. 이 조사의 표본가구로 선정된 경우에 적격 가구원이 2명 이상인 경우에는 최근 생일법에 의해서 한 명을 조사하도록 하였다.

본 연구에서 설계가중치는 표본추출 과정에서 층화2단계 확률비례계통추출법을 적용함에 따라 나타나는 조사구 내의 가구별 추출확률의 차이를 반영할 수 있도록 산정한다.

가중치 작성 과정에서 사용될 기호들을 정리하면 다음과 같다.

- L : 층의 수
- N_h : 층 h 의 모집단 조사구 수
- n_h : 층 h 의 표본 조사구 수
- S_{hi} : 층 h 의 i 번째 조사구에 대한 크기의 측도(해당 조사구의 총 가구 수)
- $S_h = \sum_{i=1}^{N_h} S_{hi}$: 층 h 에서 크기의 측도에 대한 총합

1) 무응답에 대한 조정은 대체표본을 이용하기 때문에, 실제 무응답 조정이 발생하지 않을 수도 있음

- M_{hi} : 층 h 의 i 번째 조사구 내 가구 수(조사완료+조사미완+조사미착수)
- m_{hi} : 층 h 의 i 번째 조사구 내 조사접촉 가구 수(응답+거절)
- r_{hi} : 층 h 의 i 번째 표본조사구 내 조사완료 가구 수(응답)

나. 조사구 추출에 대한 가중치 작성

국민여행조사의 조사는 월별로 진행되기 때문에 표본도 월별로 추출하고, 추정도 월별로 진행하고 있다. 단, 공표기준이 잠정치는 분기, 확정치는 1년 단위로 하는데, 이때 분기와 연 단위로 별도의 가중값을 산출하지 않고, 월별로 추정된 값을 결합하여 사용한다. 따라서 여기서 제시하는 가중값 산출방안은 월별로 산출하는 것을 기준으로 제시한 값이다.

(1) 설계가중치

이 조사의 설계가중치는 각 표본조사구에 대한 표본추출률의 역수와 표본조사구에서 가구조사 착수율의 역수를 곱하여 다음과 같이 산출한다.

$$\text{설계가중치} = \frac{S_h}{n_h S_{hi}} \times \frac{M_{hi}}{m_{hi}}$$

원칙적으로 각 표본 조사구에서는 10가구씩을 표본으로 조사하였기 때문에 $m_{hi} = 10$ 이다. $S_{hi} \approx M_{hi}$ 를 가정할 수 있는 경우(표본추출률 상의 조사구 내 가구 수와 실제 가구 수에 차이가 작은 경우) 각 층에서 설계가중치는 해당 층 내에서 일정한 값이 되어 설계가중치 = $\frac{S_h}{n_h \times 10}$ 으로 표현할 수 있다. 이때 각 지역 내 층에서

표본 가구들은 모두 동일한 설계가중치를 갖게 된다.

실제 조사과정에서는 표본추출률의 조사구 내 가구 수와 실제 가구 수에 차이가 있어 각 층 내에서 설계가중치는 동일하지는 않지만 비슷한 값을 갖게 된다.

(2) 무응답 조정

해당 지역 내 세부 층에서 조사 가구에 대한 설계가중치는 원칙적으로 같다. 본 조사에서 무응답 조정은 표본 조사구 단위로 진행되었다. 무응답 조정계수는 다음 식에 따라 구한다.

$$\text{무응답 조정계수} = \frac{m_{hi}}{r_{hi}}$$

(3) 표본 가구 내 추출률 반영

이 조사의 표본가구로 선정된 경우에 적격 가구원이 2명 이상인 경우에는 ‘최근 생일자법’에 의거하여 조사시점과 가장 근접하게 생일인 사람을 최종 적격자로 선정하였다. 따라서 표본 가구 내의 조사 적격자가 몇 명인가에 따라 추출률에 차이가 발생한다. 표본 가구의 적격자 추출률은 다음과 같다.

$$\text{가구 내 추출률} = \frac{1}{\text{표본 가구의 만 15세 이상 전체 가구원 수}}$$

앞서 구한 가구 내 추출률의 역수를 설계가중치에 곱하여 가구 내 추출률 차이를 가중치 작성 과정에 반영한다.

(4) 「인구총조사」 결과를 이용한 조정

모집단 정보를 이용한 가중치 조정은 모집단에 대한 정보를 이용하여 모집단의 구조와 표본 구조를 유사하게 맞추므로써 추정의 정확도를 높이는 것을 목적으로 한다. 가중치 작성 단계에서 사용한 모집단에 대한 정보는 2023년 인구추계결과이다. 본 연구에서는 인구총조사 결과에 대해서 레이킹 비 접근법(Raking Ratio Method)을 적용하여 가중치를 조정하였다. 본 연구에서 사용한 모집단에 대한 정보는 2023년 기준의 인구추계 결과 중 시·도별(17)×성별(2) 구분과 지역(동부, 읍·면부)×연령대(7)×성별(2) 구분에 대한 만 15세 이상 인구 현황이다. 즉, 모집단 정보를 이용한 조정 단계에 사용되는 정보는, 2023년 인구추계결과의 시·도별(17)×성별(2)×연령대(7) 구분에 대한 만 15세 이상 인구 현황이다. 다만, 연령대를 세분할 것인가는 최종 조사데이터의 성 및 연령별 분포를 살펴서 최종 결정된다. 최종 가중치는 다음 식에 따라 산정되었다.

$$\text{최종 가중치} = \text{설계가중치} \times \text{무응답 조정계수} \times (1/\text{가구내추출률}) \times \text{조정계수}$$

(5) 극단 가중치 조정

일반적으로 가중치의 과도한 변동은 추정량의 분산을 크게 만들어 추정의 정확도를 떨어뜨릴 수 있다. 몇 개의 과도하게 큰 가중치를 갖는 조사값이 잘 설계되고 수행된 통계조사에서 얻어지는 조사의 정확성을 낮추게 하는 경우도 발생할 수 있다. 실제 가중치 작성 과정에서는 지나치게 큰 가중치를 표본설계가 끝난 후 사후적으로 제한하거나 조정하는 방법이 사용되고 있다. 대표적인 방법으로는 다음 두 가지로 구분할 수 있다.

첫째, 가중치 작성 과정에서 무응답 조정, 모집단 정보를 이용한 조정 등과 같은 가중치 작성 단계별로 지나치게 큰 조정값이 나오는 것을 막기 위하여 지나치게 큰

가중치의 경우에는 절단(trimming)하거나 제한을 두는 방안이다.

둘째, 가중치 작성의 전체 과정을 마친 후에 과도하게 큰 가중치를 찾아서 절단(trimming)하고, 절단으로 인하여 부족해진 가중치를 보충하는 과정을 거치는 방안이다.

「2023 국민여행조사」의 특이 가중치 조정은 두 번째 방안을 사용되었는데, 전체 응답자의 가중치 분포를 고려하여 상·하위 5%씩을 특이 가중치로 간주하여 조정하였다.