이중추출틀 RDD 전화조사를 위한 새로운 가중치 조정 방법

New Weighting Adjustment Methods for Dual-Frame RDD Telephone Surveys

Survey & Health Policy Research Center
Technical Report

April 2018

센터장 김 선 웅

■ 개요

○ 본 연구보고서에서 제시되는 “새로운 가중치 조정 방법”은 2005년부터 2016년까지 동국대학교 서베이앤헬스폴리시리서치센터에서 사용되어온 기존 가중치 조정 방법(부록 참조)을 대체하기 위한 것으로서, 센터 내부적으로는 RDD 표본설계와 전화조사방법론의 개선, 외부적으로 국내 전화번호 체계의 변경 등에 따른 후속조치로서 개발된 것임. 이 새로운 가중치 조정 방법은 CATI(computer-assisted telephone interviewing, 컴퓨터를 이용한 전화조사)로 센터에서 진행한 전국 규모의 “2017 흡연실태조사”의 데이터 분석 시부터 사용됨

○ 새로운 가중치 조정 방법은 이중추출틀, 즉 일반전화(집전화) RDD 표본추출틀과 휴대전화 RDD 표본추출틀로부터 각각 추출된 일반전화 번호 표본과 휴대전화 번호 표본을 가지고 진행되는 CATI 전화조사를 진행하는 과정에서 수집되는 각 표본 전화번호의 다양한 조사상황자료(패러데이터)를 기존 방법에 비해 엄밀하게 이용할 수 있도록 설계되었으며, “응답자 개인별 통계적 가중치(statistical weights)”를 보다 정확하게 산출할 수 있음. 이러한 “응답자 개인별 가중치”를 사용하여 전화조사 데이터를 분석함으로써, 우리나라 전체 성인 인구(모집단)의 특성치(모수: 모비율, 모평균 등)에 대한 추정의 정확성을 극대화할 수 있음

○ 성인을 대상으로 한 전화조사에서 “가중치(weights)”란 ‘응답자 1명이 본인을 포함하여 대표하는 성인 인구수’로서 기본적으로는 표본추출확률의 역수로 나타내며, 표본으로 추출되는 개인이 모두 동일한 확률로 추출되는 경우 모두 동일한 가중치(self-weighting)를 가짐. 하지만 전화조사에서 응답자 개인은 근본적으로 동일한 추출확률로 추출될 수 없음. 예를 들어, 이것은 개인의 전화보유 유형(예: ①일반전화만 보유, ②휴대전화만 보유, ③일반전화와 휴대전화 모두 보유)이 달라 어떤 개인은 일반전화 RDD 표본추출틀에서만 추출될 수 있고, 어떤 개인은 휴대전화 RDD 표본추출틀로부터만 추출될 수 있으며, 또 어떤 개인은 두 표본추출틀 모두에서 추출될 수 있어 추출확률이 서로 다르게 되기 때문임

○ 또한 표본으로 추출된 응답자 중 일부는 비접촉이나 응답거부 등으로 무응답이 발생할 수 있으므로 무응답에 따른 가중치 조정이 반드시 필요함. 따라서 모집단의 중요한 인구학적 특성들을 반영할 수 있도록 사후층화 (post-stratification) 등을 실시하여 가중치 조정을 해야 함

○ “새로운 가중치 조정 방법”은 다음 표1의 ‘5개 단계’의 가중치 조정방법을 적용하는 것이며 이를 통해 응답자별 최종 가중치(final weight)를 산출함

표1. 5단계 가중치 조정

단계1	이중추출틀(Dual Frame) RDD 초기 가중치 및 개인 가중치 산출
단계2	비적격 번호 가중치 조정
단계3	인포먼트 무응답 가중치 조정
단계4	개인 무응답 가중치 조정
단계5	사후층화(지역, 성별, 연령) 가중치 조정

■ 단계별 가중치 조정

○ 각 단계별 가중치 수식에서 사용되는 기호는 다음과 같음

전화 구분: L - 일반전화(landline phones), C - 휴대전화(cell phones)

h : 지역(서울특별시, 6개 광역시, 9개 도 및 세종시의 17개 지역)

i : 각 RDD 표본 전화번호

○ 본 센터에서는 전화조사 진행 시 “일반전화와 휴대전화 구분 없이” 각 표본 전화번호를 같이 사용하는 성인(만 19세 이상)이 ‘2인 이상’인 경우 이들 중 1명을 면접원이 CATI 시스템 상에서 랜덤하게 추출하므로, 이를 단계별 가중치 조정에 적용함. 여기서 “일반전화와 휴대전화 구분이 없이”는 ‘일반전화’ 또는 ‘휴대전화’가 모두 가구용(household phone - 가구원 모두가 사용) 또는 공용(shared phone - 가구원들 중 일부 사람들만 같이 사용) 또는 개인용(individual phone)으로 사용될 수 있음을 전제로 하는 것임

○ 각 단계별 가중치 조정 방법의 구체적인 내용은 다음과 같음

[단계1] 이중추출틀(dual frame) RDD 초기 가중치 및 개인 가중치 산출

이중추출틀, 즉 일반전화 RDD 표본추출틀과 휴대전화 RDD 표본추출틀 각각에서 표본번호들이 동일확률로 추출되지만 일반전화와 휴대전화를 함께 사용하는 사람인 경우 각 표본추출틀에서 동시에 추출될 수 있어 ‘표본추출틀 간의 중복 문제(overlap problem)’가 발생하여 ‘특성치(모수)’ 추정이 복잡해지므로 이런 문제를 해결하기 위해, 다음과 같이 일반전화 RDD 표본과 휴대전화 RDD 표본이 서로 구분되는 가중치들이 사용됨. 단, 이 가중치들을 표현하는 수식에서 아래의 추출확률 기호들을 기본적으로 사용

함

$p_{Li,h}$ = 각 지역(h)별 일반전화 RDD 표본 번호의 추출확률(selection probability)
p_{Ci} = 휴대전화 RDD 표본 번호의 추출확률

▸ 일반전화 RDD 표본

1) 표본 번호가 비적격 번호(업무 전용 번호나 결번 등)이거나 전화를 걸었을 때 전화를 받은 사람(informant)으로부터 아무런 응답을 얻지 못해 조사가 진행되지 않은 경우는 다음 "초기 가충치"를 사용:

$$\text{초기 가중치: } W_{D,h} = \frac{1}{p_{Li,h}}$$

여기서 $p_{Li,h} = \frac{n_{Lh}}{N_{Lh}}$

N_{Lh} = 각 지역별 일반전화 RDD 표본추출틀 크기

n_{Lh} = 각 지역에서 실제 추출된 일반전화 RDD 표본 번호의 전체 개수

2) 표본 번호가 "landline only person(일반전화만 사용하는 사람)"에 해당되는 경우 다음 "개인 가중치"를 사용:

전화를 받은 사람(informant)이 제공한 "표본 번호를 같이 사용하는 만 19세 이상인 개인들의 리스트(목록)"으로부터 랜덤하게 추출된 응답자(1명)가 "일반전화만 사용하는 사람(landline only person)"인 경우

$$\text{개인 가중치: } W_{D,h} = \frac{1}{\sum_{j=1}^{\alpha_{Li}} \frac{p_{Li,h}}{\beta_{Lj}}}$$

여기서 α_{Li} =응답자가 사용하는 일반전화 번호의 개수

β_{Lj} = 응답자가 사용하는 각 일반전화 번호를 함께 사용하는 성인들의 수

3) 표본 번호가 "landline and cell person(일반전화와 휴대전화 둘 다 사용하는 사람)"에 해당 되는 경우 다음 "개인 가중치"를 사용:

전화를 받은 사람(informant)이 제공한 "표본 번호를 같이 사용하는 만 19세 이상인 개인들의 리스트(목록)"으로부터 랜덤하게 추출된 응답자(1명)가 일반전화와 휴대전화 둘 다 사용하는 사람(landline and cell person)인 경우

개인 가중치: $W_{D,h} = \dfrac{1}{\sum_{j=1}^{\alpha_{Li}} \frac{p_{Li,h}}{\beta_{Lj}} + \sum_{j=1}^{\alpha_{Ci}} \frac{p_{Ci}}{\beta_{Cj}} - \sum_{j=1}^{\alpha_{Li}} \frac{p_{Li,h}}{\beta_{Lj}} \sum_{j=1}^{\alpha_{Ci}} \frac{p_{Ci}}{\beta_{Cj}}}$

여기서 "C"가 사용된 기호(수식)은 다음의 "휴대전화 RDD 표본" 참조

▸ 휴대전화 RDD 표본

1) 표본 번호가 비적격 번호(개인용이 아닌 업무전용 또는 결번 등)이거나 전화를 받은 사람(informant)으로부터 아무런 응답을 얻지 못해 조사가 진행되지 않은 경우 다음 "초기 가중치"를 사용

초기 가중치: $W_D = \dfrac{1}{p_{Ci}}$

여기서 $p_{Ci} = \dfrac{n_C}{N_C}$

N_C = 전국 휴대전화 RDD 표본추출틀 크기

n_C = 실제 추출된 휴대전화 RDD 표본 번호의 전체 개수

2) 표본 번호가 "cell only person(휴대전화만 사용하는 사람)"에 해당 되는 경우 다음 "개인 가중치" 사용:

전화를 받은 사람(informant)이 제공한 "표본 번호를 같이 사용하는 만 19세 이상인 개인들의 리스트(목록)"으로부터 랜덤하게 추출된 응답자(1명)가 휴대전화만 사용하는 사람(cell only person)인 경우

개인 가중치: $W_D = \dfrac{1}{\sum_{j=1}^{\alpha_{Ci}} \frac{p_{Ci}}{\beta_{Cj}}}$

여기서 α_{Ci} =응답자가 사용하는 휴대전화 번호의 개수

β_{Cj} = 응답자가 사용하는 각 휴대전화 번호를 함께 사용하는 성인들의 수

3) 표본 번호가 "landline and cell person(휴대전화와 일반전화 둘 다 사용하는 사람"에 해당되는 경우 다음 "개인 가중치" 사용:

전화를 받은 사람(informant)이 제공한 "표본 번호를 같이 사용하는 만 19세 이상인 개인들의 리스트(목록)"으로부터 랜덤하게 추출된 응답자(1명)가 휴대전

화와 일반전화 둘 다 사용하는 사람(landline and cell person)인 경우

$$\text{개인 가중치: } W_D = \frac{1}{\sum_{j=1}^{\alpha_{Li}} \frac{p_{Li,h^*}}{\beta_{Lj}} + \sum_{j=1}^{\alpha_{Ci}} \frac{p_{Ci}}{\beta_{Cj}} - \sum_{j=1}^{\alpha_{Li}} \frac{p_{Li,h^*}}{\beta_{Lj}} \sum_{j=1}^{\alpha_{Ci}} \frac{p_{Ci}}{\beta_{Cj}}}$$

여기서 h^*는 응답자가 보고한 거주 지역

[단계2] 비적격 번호 가중치 조정

표본 번호가 [단계1]에서 '비적격(out of scope)' 번호는 다음과 같이 [단계1]에서의 "초기 가중치"를 재조정함. 여기서 '비적격'은 일반전화번호에서 업무 전용 번호나 결번 등을 의미하며 휴대전화에서는 개인용이 아닌 업무전용 또는 결번 등을 의미함. 수식에서 "unresolved"는 적격 번호인지 확인할 수 없는 경우로, 이런 번호에 대해서는 일반전화 전체 표본번호(휴대전화 전체 표본번호)에서 확인된 번호들 중 적격인 번호의 비율($0 < P_{in-scope} < 1$)을 [단계1]에서의 산출된 가중치에 곱함

▸ 일반전화 RDD 표본

$$W_{D,h}\, A_{1h} = \frac{1}{p_{Li,h}}\, A_{1h}$$

$$\text{여기서 } A_{1h} = \begin{cases} 0 & \text{if } out\ of\ scope \\ P_{in-scope} & \text{if } unresolved \\ 1 & otherwise \end{cases}$$

▸ 휴대전화 RDD 표본

$$W_D\, A_1 = \frac{1}{p_{Ci}} A_1$$

$$\text{여기서 } A_1 = \begin{cases} 0 & \text{if } out\ of\ scope \\ P_{in-scope} & \text{if } unresolved \\ 1 & otherwise \end{cases}$$

[단계3] 인포먼트 무응답 가중치 조정

전화를 받은 사람(informant)으로부터 응답을 얻지 못해 조사가 진행되지 않은 표본 번호들이 발생하므로 [단계1]과 [단계2]에서 얻은 각 표본 번호의 가중치를 이용하여 다음과 같이 가중치를 조정함. 수식에서 “all sampled landline (cell) numbers”는 실제 표본으로 추출된 모든 일반전화(휴대전화) 번호들을 의미하며 이들 각 일반전화(휴대전화) 번호의 가중치는 [단계1]에서의 얻은 “초기 가중치” 또는 “개인 가중치”임. 단, 비적격 번호의 경우 [단계2]에서 얻은 가중치를 사용함. 또한 수식에서 “informant landline (cell) numbers”는 전화를 받은 사람(informant)으로부터 응답을 얻은 번호들을 의미하며, 이들 번호에는 [단계1]에서 “개인 가중치”를 갖는 번호들(조사에 응한 응답자 번호들)도 당연히 포함됨

▸ 일반전화 RDD 표본

$$A_{2h} = \frac{sum\ of\ weights\ for\ all\ sampled\ landline\ numbers}{sum\ of\ weights\ for\ informant\ landline\ numbers}$$

▸ 휴대전화 RDD 표본

$$A_{2} = \frac{sum\ of\ weights\ for\ all\ sampled\ cell\ numbers}{sum\ of\ weights\ for\ informant\ cell\ numbers}$$

[단계4] 개인 무응답 가중치 조정

전화를 받은 사람(informant)으로부터 응답을 얻어 일반전화와 휴대전화 구분 없이 각 RDD 표본전화번호를 같이 사용하는 만 19세 이상인 개인들(eligible persons) 중 1명을 랜덤하게 추출하였으나 이 추출된 사람(응답자)이 응답을 하지 않는 경우들이 발생하므로 다음과 같이 가중치를 조정함. 성별, 연령의 각 범주는 아래 “Categories for a Post-Stratum” 표2 참조.

▸ 일반전화 RDD 표본

$$A_{3h} = \frac{\begin{array}{c}sum\ of\ weights\ for\ all\ selected\ landline\ persons\ in \\ an\ sex\text{-}age\text{-}number\ of\ eligible\ persons\ category\end{array}}{\begin{array}{c}sum\ of\ weights\ for\ landline\ respondents\ in \\ an\ sex\text{-}age\text{-}number\ of\ eligible\ persons\ category\end{array}}$$

▸ 휴대전화 RDD 표본

$$A_3 = \frac{\textit{sum of weights for all selected cell persons in an sex-age-number of eligible persons category}}{\textit{sum of weights for cell respondents in an sex-age-number of eligible persons category}}$$

[단계5] 사후층화(지역, 성별, 연령) 가중치 조정

지역, 성별, 연령대에 따른 모집단 인구 크기 추정치(통계청 인구주택총조사 결과)를 이용하여 가중치를 조정함. 단, 일반전화의 경우 사후층(post-stratum)은 각 지역 내에서의 두 변수(성별과 연령)에 따른 각 범주를 의미하며, 휴대전화의 경우는 3가지 변수(응답자가 보고한 거주 지역, 성별, 연령)에 따른 범주를 사용함 - 지역, 성별, 연령의 각 범주는 아래 표2 “Categories for a Post-Stratum” 참조

▸ 일반전화 RDD 표본

$$A_{4h} = \frac{\textit{population estimate for a post-stratum (sex-age)}}{\textit{sum of weights of landline respondents in a post-stratum (sex-age)}}$$

▸ 휴대전화 RDD 표본

$$A_4 = \frac{\textit{population estimate for a post-stratum (self-report location-sex-age)}}{\textit{sum of weights of cell respondents in a post-stratum (self-report location-sex-age)}}$$

표2. Categories for a Post-Stratum

성별	남
	여
연령	19-29
	30-39
	40-49
	50-59
	60-69
	70세 이상
지역	서울특별시
	부산광역시
	대전광역시
	대구광역시
	광주광역시
	울산광역시
	인천광역시
	경기도
	강원도
	충청남도
	충청북도
	전라남도
	전라북도
	경상남도
	경상북도
	제주도
	세종시

■ 최종 가중치

○ 각 단계별 가중치 조정을 적용함으로써 얻어지는 응답자별 '최종 가중치'는 다음과 같음

▸ 일반전화 RDD 표본

최종 가중치: $W_{final} = W_{D,h} \times A_{2h} \times A_{3h} \times A_{4h}$

▸ 휴대전화 RDD 표본

최종 가중치: $W_{final} = W_D \times A_2 \times A_3 \times A_4$

부록

2017년까지 센터에서 사용된 가중치 조정 방법

○ 2017년까지 본 센터에서는 다음 '6개 단계의 가중치 조정 방법'을 사용해왔음

그림. 6단계 가중치 조정 방법

○ 단계별 가중치 및 가중치 조정 방법은 각 지역별(일반전화에 해당)로 적용되며 구체적인 내용은 다음과 같음

$i = 1$: 일반전화

$i = 2$: 휴대전화

h : 지역

(1) RDD 초기 가중치

▸ Landline

$$W_{초기,h} = \frac{total\ number\ of\ telephone\ numbers\ in\ the\ sample\ frame}{total\ number\ of\ telephone\ numbers\ that\ were\ randomly\ sampled\ from\ sampling\ frame}$$

▸ Cell

$$W_{초기} = \frac{total\ number\ of\ telephone\ numbers\ in\ the\ sample\ frame}{total\ number\ of\ telephone\ numbers\ that\ were\ randomly\ sampled\ from\ sampling\ frame}$$

(2) 비적격 전화번호 가중치 조정

▸ Landline

$$A_{1h} = \begin{cases} 0 & \text{if } out\ of\ scope \\ P_{in-scope} & \text{if } unresolved \\ 1 & otherwise \end{cases}$$

▸ Cell

$$A_1 = \begin{cases} 0 & \text{if } out\ of\ scope \\ P_{in-scope} & \text{if } unresolved \\ 1 & otherwise \end{cases}$$

(3) 무응답 가중치 조정

▸ Landline

$$A_{2h} = \frac{sum\ of\ weights\ for\ all\ sampled\ households}{sum\ of\ weights\ for\ respondent\ households}$$

▸ Cell

$$A_2 = \frac{sum\ of\ weights\ for\ all\ sampled\ households}{sum\ of\ weights\ for\ respondent\ households}$$

(4) 개인 가중치 조정

In case of landline or cell only person,

$$A_3 = \frac{1}{probability\ of\ within\text{-}household\ selection (= \pi_{ik})} ,$$

In case of landline and cell person,

$$A_3 = \frac{1}{\pi_{1k} + \pi_{2k} - \pi_{1k}\pi_{2k}} \quad ,$$

$$\pi_{ik} = \sum_{j=1}^{\alpha_{ik}} \frac{1}{\beta_{ijk}},$$

α_{ik} : Number of phone i's to be reached to the respondent k

β_{ijk} : Number of adults who use jth phone i with respondent k

(5) 개인 무응답 가중치 조정

$$A_4 = \frac{\textit{sum of weights for all sampled selected members}}{\textit{sum of weights for respondent selected members}}$$

(6) 사후층화(연령, 성별, 지역) 가중치 조정

$$A_5 = \frac{\textit{population estimate for a post-stratum}}{\textit{sum of weights of respondent selected members in a post-stratum}}$$

지역, 성별, 연령의 각 범주는 본문의 표2 “Categories for a Post-Stratum”와 동일함

○ 이에 따른 개인별 최종 가중치는 다음과 같음

Landline 최종 가중치: $W_{final} = W_{\in ial,\, h} \times A_{1h} \times A_{2h} \times A_3 \times A_4 \times A_5$

Cell 최종 가중치: $W_{final} = W_{\in ial} \times A_1 \times A_2 \times A_3 \times A_4 \times A_5$