한국노동패널조사

표본 추가 연구 용역
최종 보고서

2009. 3. 24.

미국 아이오와 주립대학교 통계학과
김재광

【목 차】

【표 목 차】

1. 서 론

본 연구는 노동연구원에서 실시하는 한국 노동 패널 조사(KLIPS)에 사용될 표본을 추가하는 것을 목적으로 한다. 노동 패널 조사 사업은 1998년부터 도시 지역의 5,000 개의 표본 가구를 선정하여 지금까지 수행되어 왔으며 2008년 11차조사에서는 원가구 5,000 가구 및 분가가구 2,531 가구의 총 7,531 패널 가구 중 소멸가구및 5차년도 이후 무응답을 제외한 나머지 6,027 가구가 패널 표본으로 유지되어 80.03%의 표본 유지율을 기록하고 있다.

그러나, KLIPS 의 초기 표집 과정에 대한 비판과 표본 이탈의 문제는 결국 현 표본의 대표성에 대한 한계를 나타내었고 또한 지역 통계 생산 및 보다 심층적인 분석을 위해서는 현행 표본의 규모가 부족하며 이를 해결하기 위해서는 현재 모집단에 대한 대표성을 보완하도록 표본이 추가되어야 할 것으로 판단 되었다. 또한 지금까지의 KLIPS 의 목표 모집단을 도시 지역 가구로 한정지움으로써 현행 표본이 비도시에서 도시로 이동하는 가구에 대한 대표성을 확보하지 못한다는 단점이 있고 이에 대한 보완책으로써 목표 모집단을 도시 지역 가구가 아닌 전국 가구로 확장하고 그 모집단 정의에 대해 대표성을 가지도록 표본 추가의 필요성이 대두되었다.

본 연구에서는 이러한 표본 추가 작업을 위하여 2005년 인구주택 총조사 자료를 바탕으로 표본 추가 작업을 실시하였고 이에 대한 진행 과정을 기술하고자 한다. 본 연구는 다음과 같은 특징을 가지고 있다.

(1) 본 표본 추가 작업에서는 전국의 가구를 대상으로 모집단을 확장하여 이에 따른 표본 추가를 실시하였다.

(2) 본 표본 추가 작업에서는 시도별 동부 읍면부별 표본 가구수가 2005년 센서스

값과 일치하도록 비례 배정하였다.

(3) 기존의 동부 표본에서 단독주택과 1-2인 가구가 과소 표집되었으므로 이번 표본 추가에서는 동부 자료의 경우 단독 주택과 1-2인 가구가 뽑힐 확률을 더 높게 부여하여 전체적인 표본의 주택 유형별 및 가구원수 분포가 모집단과 비슷해지도록 하였다. 읍면부에서는 기존 표본의 비중이 크지 않으므로 이러한 방법을 사용하지 않고 전통적인 가구수에 비례하는 확률비례 추출법을 사용하였다. 이렇게 해서 조사구가 뽑힌 후에는 각 조사구 별로 일정수의 가구 (5 가구)를 계통 추출법을 사용하여 추출하였다.

2. 표본 추출

1) 조사 모집단 및 조사 대상 가구 설정

본 표본추가 작업에서는 2009년 3월 현재 대한민국 영토 (제주도 포함) 내의 일반가구에 거주하고 있는 가구원을 조사 모집단으로 결정하였다. 이러한 조사 모집단에 대한 추출 프레임으로는 2005년 센서스 조사구 자료를 사용하였다. 따라서 실제로 추출되는 가구는 2005년 센서스 조사구 자료에 주소지가 존재하는 가구가 된다.

이렇게 해서 2005년 센서스 조사구 자료에 주소지가 존재하는 가구는 원가구로써 조사 대상이 되고 이 중 표본으로 뽑힌 가구에서 분가된 가구도 추출할 것인지에 대해서는 추후 논의가 필요할 것으로 보인다.

2) 표본수의 결정 및 표본 배정

먼저 센서스 모집단 자료 및 노동패널 자료에서 동부와 읍면부 비율은 다음과 같다. 노동패널 자료는 2005년 Census 자료와의 비교를 위하여 6,027의 현 패널가구 중 2005년 이후 분가가구를 제외한 5,585가구를 분석의 대상으로 한다.

☐ 2005년 Census 모집단의 동부 읍면부 별 분포

	가구 수	%
동부	12,744,940	80.22
읍면부	3,142,188	19.78
전국	15,887,128	100.00

☐ 노동패널의 동부 읍면부 별 분포

	가구 수	%
동부	4,837	86.61
읍면부	748	13.39
전국	5,585	100.00

따라서 동부에서 900가구, 읍면부에서 600가구의 표본 추가를 통해서 이루어 내고자 하는 목표 표본 수는 다음과 같이 결정하였다.

☐ 표본추가 후 노동패널의 동부 읍변부 별 분포

	가구 수	%
동부	5,737	80.97
읍면부	1,348	19.03
전국	7,085	100.00

이 후 지역별 (동/읍면부, 시도)분포를 고려하여 비례 배정으로 표본을 배정한 결과 동부와 읍면부에서 각각 다음과 같은 가구 표본수가 배정되었다.

< 동부 가구 표본 배정 - 가구 단위 >

광역시/도	census		비례 배정 총 표본 수	현 패널자료		추가 표본 수
	가구 수	%		가구 수	%	
서울특별시	3,309,890	25.97	1,490	1,303	26.94	185
부산광역시	1,162,029	9.12	523	461	9.53	65
대구광역시	767,541	6.02	345	330	6.82	20
대전광역시	478,865	3.76	216	168	3.47	30
인천광역시	796,926	6.25	359	319	6.59	40
광주광역시	460,090	3.61	207	180	3.72	25
울산광역시	285,696	2.24	129	107	2.21	25
경기도	2,757,408	21.64	1,241	1,004	20.76	240
강원도	315,406	2.47	142	120	2.48	25
충청북도	307,284	2.41	138	108	2.23	30
충청남도	236,823	1.86	107	80	1.65	30
전라북도	407,984	3.20	184	161	3.33	25
전라남도	260,479	2.04	117	102	2.11	15
경상북도	462,200	3.63	208	158	3.27	50
경상남도	607,248	4.76	273	234	4.84	40
제주도	129,071	1.01	58	2	0.04	55
합	12,744,940	100.00	5,737	4,837	100.00	900

< 읍면부 가구 표본 배정 - 가구 단위 >

광역시/도	census		비례 배정 총 표본 수	현 패널자료		추가 표본 수
	가구 수	%		가구 수	%	
부산광역시	24,349	0.77%	10	9	1.20	0
대구광역시	47,044	1.50%	20	18	2.41	0
인천광역시	26,097	0.83%	11	7	0.94	5
울산광역시	53,399	1.70%	23	20	2.67	5
경기도	571,769	18.20%	245	194	25.94	50
강원도	205,222	6.53%	88	29	3.88	60
충청북도	197,919	6.30%	85	17	2.27	70
충청남도	423,048	13.46%	182	95	12.70	85
전라북도	211,974	6.75%	91	60	8.02	30
전라남도	405,840	12.92%	174	51	6.82	125
경상북도	476,640	15.17%	204	139	18.58	65
경상남도	448,759	14.28%	193	108	14.44	85
제주도	50,128	1.60%	22	1	0.13	20
총합	3,142,188	100.00%	1,348	748	100.00	600

위의 테이블은 가구 단위의 표본 배정이고 본 연구에서는 센서스 조사구 (ED)를 일차 표본 추출 단위로 하고 가구를 이차 표본 추출 단위로 하는 이단계 집락 추출을 사용하였으므로 실제로 배정된 조사구 단위 표본수는 다음과 같다. 여기서는 각 조사구당 5개의 표본 가구를 추출하는 것을 원칙으로 하였다.

< 최종 배정 표본 조사구수 >

광역시/도	동부	읍면부
서울특별시	37	0
부산광역시	13	0
대구광역시	4	0
대전광역시	6	0
인천광역시	8	1
광주광역시	5	0
울산광역시	5	1
경기도	48	10
강원도	5	12
충청북도	6	14
충청남도	6	17
전라북도	5	6
전라남도	3	25
경상북도	10	13
경상남도	8	17
제주도	11	4
총 조사구수	180	120

3) 표본 추출

표본 추출은 동부와 읍면부에서 다르게 적용하였다. 동부에서는 기존 자료의 가구원수 분포와 주택유형별 분포가 센서스의 분포와 지나치게 차이가 나므로 이를 보정해주는 표본 추출법을 사용하였다. 이를 위하여 조사구별로 1-2인 가구수와 단독주택 가구수의 합을 특성치로 사용하였고 그 특성치 값에 비례하는 PPS 샘플링을 각 층별로 독립적으로 실시하였다. 읍면부에서는 기존 표본의 비중이 크지 않으므로

이러한 방법을 사용하지 않고 전통적인 가구수에 비례하는 확률비례 추출법을 사용하였다. 이렇게 해서 조사구가 뽑힌 후에는 각 조사구 별로 일정수의 최종 표본 가구 (5 가구)를 계통 추출법을 사용하여 추출하였다.

□ 가구 규모별 분포 비교 (동부)

가구원수	2005년 census (%)	노동패널 (%)
1인/2인	42.12	32.3
3인/4인	47.93	56.17
5인이상	9.95	11.53

□ 주택 유형별 분포 비교 (동부)

주택 유형	2005년 census (%)	노동패널 (%)
단독주택	44.46	31.64
아파트	41.73	43.53
기타	13.81	24.83

동부 자료에서는 1-2인 가구와 단독 주택 가구를 많이 추출하기 위하여 다음과 같은 방법을 사용하였다.

step 1 : 센서스 2005년 자료에서 조사구 단위로 지역(시도), 총가구수, 가구규모별 가구수, 주택 유형별 가구수를 얻어낸다.

조사구별 특성치를 다음과 같이 계산한다.

특성치 = ((1-2인 가구수)+단독주택 가구수)

step 2 : 지역 내에서 모집단 조사구를 지역 내에서 특성치에 비례하는 PPS 추출을 실시한다. 즉, 각 지역 내에서 조사구가 뽑힐 확률이 특성치에 비례하도록 뽑는다.

step 3 : 이렇게 해서 얻어진 표본 조사구에서 5 개의 가구를 계통 추출로 뽑아 본표본으로 지정하고 나머지에서 10 개 가구를 예비 표본으로 지정한다.

읍면부 추출 역시 층화 이단계 집락 추출을 사용하였으나 동부에서 사용한 특성치를 사용하지 않고 조사구별 가구수에 비례하는 PPS 추출을 사용하였다. 각 표본 조사구에서 5 개의 가구를 계통 추출로 뽑아 본표본으로 지정하고 나머지에서 10 개 가구를 예비 표본으로 지정한다.

3. 가중치 결정

이렇게 해서 얻어진 표본 가구를 기존 표본에 포함하여 분석하기 위해서는 추가된 표본 가구 및 가구원에 가중치를 결정하고 그 가중치를 기존 분석에 어떻게 사용할 것인가를 결정해야 할 것이다. 여기서는 크게 추가 표본에 대한 가중치 결정과 기존 표본에 통합되어 분석할 때 사용되는 방법론에 대한 내용을 다루고자 한다.

(1) 추가 표본의 가중치 결정

추가 표본의 가중치는 추가 표본의 표본 추출 확률의 역수로 계산된다. 표본 추출이 가구를 최종 추출 단위로 한 층화 이단계 집락 추출을 사용하였으므로 각 가구

에 배정되어지는 기본 가중치(w_{hij})는 원칙적으로 다음과 같이 계산되어질 수 있을 것이다.

$$w_{hij} = C_h \frac{1}{X_{hi}} \frac{M_{hi}}{m_{hi}}$$

(식 1)

여기서,

h: 층(시도)을 나타내는 인덱스

i: 조사구를 나타내는 인덱스

j: 가구를 나타내는 인덱스

X_{hi}: 층 h의 조사구 i 에서의 특성치값

M_{hi} : 층 h의 조사구 i 에서의 (요도 상에서의) 총 가구수

m_{hi} : 층 h의 조사구 i 에서의 표본 가구수 (=5)

이고 C_h 값은 층 내에서의 (분가 가구들을 제외한) 표본 가구들의 가중치 합들이 2005년 센서스에서의 층내 가구수와 같아지도록 결정한다. 여기서 특성치값은 동부의 경우에는 1-2인 가구와 단독 주택 가구수의 합으로 계산되고 읍면부의 가구에는 그냥 총가구수가 된다.

기존 표본에서 가구 가중치가 있는 경우에 비례 상수 C_h 의 결정은 다음의 식을 통해서 구현할 수 있을 것이다.

$$\frac{n_{h0}}{n_h} \times \sum_{(ij) \in S_{h0}} w_{hij} + \frac{n_{h1}}{n_h} \times \sum_{(ij) \in S_{h1}} w_{hij} = M_h$$

(식 2)

여기서 ,

n_{h0} : 층 h 내에서의 원 패널 표본 가구수 (즉, 5585 가구 중 층 h에 속한 가구수)

n_{h1} : 층 h 내에서의 추가 패널 표본 가구 수

$n_h = n_{h0} + n_{h1}$: 층 h 내의 최종 표본 가구 수

S_{h0} : 층 h 내에서의 원 패널 표본가구의 집합 (2005년 이후 분가 가구 제외)

S_{h1} : 층 h 내에서의 추가 패널 표본가구의 집합 (2005년 이후 분가 가구 제외)

M_h : 층 h 내에서의 총 가구 수 (2005년 센서스 기준)

이렇게 해서 추가 표본의 가중치가 결정되면 분가된 가구의 가중치는 원가구 가중치를 그대로 가져다 사용한다. 이렇게 해 줌으로써 2009년 표본 가구의 가중치는 2005년 이후의 분가 상황에 대해 편향 없이 추정이 가능하게 된다.
(식 1)로 표현되는 가중치 값은 이 보고서에는 포함하지 않고 별로의 파일로 제공하게 될 것이다.

(2) 원 자료와 통합된 추정 방법

원 자료와 통합된 추정을 위해서는 원 표본자료의 가중치와 추가 표본 자료의 가중치를 함께 바꾸어주어야 하는데 (식 2)에서 그 힌트를 찾을 수 있다. 즉, 각 층별로 기존 가구 수 (n_{h0})와 추가된 가구 수 (n_{h1})를 계산한 후 각 층별로 기존 표본의 가중치에는 승수 $R_h = n_{h0}/(n_{h0} + n_{h1})$ 를 곱해주고, 추가 가중치에는 (식 1)의 가중치에 승수 $n_{h1}/(n_{h0} + n_{h1})$ 를 곱해주면 된다. 이러한 보정 승수를 사용해 줌으로써 추가 표본이 차지하는 비중이 표본 수에 비례하도록 하였고 이는 평균적으로 원표본의 가중치와 추가 표본의 가중치가 비슷해지도록 하는 효과를 가져온다. 이러한 승수(R_h)에 대한 테이블은 다음과 같다.

< 층별 가중치 승수 R_h 분포 표>

광역시/도	동부	읍면부
서울특별시	0.8757	
부산광역시	0.8764	1.0000
대구광역시	0.9429	1.0000
대전광역시	0.8485	
인천광역시	0.8886	0.5833
광주광역시	0.8780	
울산광역시	0.8106	0.8000
경기도	0.8071	0.7951
강원도	0.8276	0.3258
충청북도	0.7826	0.1954
충청남도	0.7273	0.5278
전라북도	0.8656	0.6667
전라남도	0.8718	0.2898
경상북도	0.7596	0.6814
경상남도	0.8540	0.5596
제주도	0.0351	0.0476

4. 기타 사항 - 표본 대체

이렇게 해서 얻어진 표본 가구는 실사를 통해서 조사 가구로 포함된다. 실사를 한 결과 조사 가구로 포함할 수 없는 경우 (예: 주소지가 없어짐, 기존 노동패널 조사 가구) 에는 동일 조사구 내에서의 예비 표본에서 얻어진 가구를 사용한다. 또한 각 표본 조사구 내에서는 최소 표본 가구수인 5 가구 이상을 유지해야 한다. 만약 조사구내 가구수가 4이하가 되면 (식 1)의 가구 가중치가 지나치게 커질 위험이 있기에 바람직하지 않다.

조사구가 없어져서 조사 가구를 얻어낼 수 없는 경우에는 해당 층에서 조사구를 새로 추출하거나 아니면 해당 층의 다른 표본 조사구에서 추가로 표본 가구를 추출한다. 전자의 경우에는 조사구 요도를 추가로 구입하여야 하는 절차상의 번거로움이 있으므로 후자의 경우인 해당 층의 다른 표본 조사구에서 추가로 표본 가구를 추출한다. 이 경우에도 표본 조사구는 PPS 추출을 따른 것이므로 조사구의 대표성은 훼손되지 않는다. (즉, 5개의 조사구를 PPS 로 뽑은 후에 임의로 하나를 제거해도 4개의 조사구의 대표성은 훼손되지 않는다.) 그 후 나머지 조사구에서 표본 가구를 추가로 랜덤하게 추출한다. 이 경우 추가로 표본이 추출되는 조사구는 조사구의 크기가 큰 것 위주로 추출하는 것이 가중치 작업상 바람직할 것으로 판단되며 통합된 추정을 위한 승수 R_h 값도 그에 맞게 바뀌어야 할 것이다.

5. 결 론

노동 패널 조사 표본의 전국 대표성을 높이기 위해 표본 추가를 실시하였다. 표본 추가에 사용된 샘플링 프레임은 2005년 센서스 자료였으며 과거 패널 자료를 분석한 결과 동부의 경우 1-2인 가구와 단독 주택 가구가 과소 표집된 것을 발견하고 이를 보정하기 위해 특성치를 사용한 PPS 추출을 사용하였다. 또한 시도별로 모집단 가구 수에 비례하도록 표본 배정을 실시하였다.

가중치는 표본 추출 확률을 반영하는 기본 가중치 작업을 실시하였다. 이를 위하여 먼저 추가 표본에 한하여 추출 확률을 반영하는 가중치 작업을 실시한 후에 추가 표본이 차지하는 비중이 표본 수에 비례하도록 보정하였다. 이러한 가중치 작업은 추출 확률만을 반영한 것이고 인구의 성별 연령별 분포를 반영하지는 않았다. 이러한 추가적인 가중치 보정은 calibration 이라는 방법을 통해서 가능한데 여기서는 본 용역의 범위를 벗어나므로 이를 고려하지 않았다.