응용통계연구 제 **14**권 2호, *2001*년, *pp. 289-304*

2001년 국민건강 · 영양조사 표본설계

류제복[1] 이계오[2] 김영원[3]

요 약

2001년에 실시할「국민건강 · 영양조사」를 위한 새로운 표본설계를 하였다. 본 표본설계에서는 표본의 대표성을 높이기 위해서 기존의 표본가구수는 유지하면서 표본조사구를 증가시키고 새로운 지역 층화변수를 추가로 도입하였다. 또한 추정량의 추정오차 공식을 유도하여 추정의 신뢰성을 측정할 수 있도록 설계하였다.

주요용어: 층화변수, 계통추출, 비례배분, 가중표본합계치.

1. 서론

매 3년 주기로 실시되고 있는「국민건강 · 영양조사」는 국민의 주관적 객관적 건강상태, 건강에 관한 의식 및 행태와 식품섭취현황 등 건강과 관련되는 사항을 종합적이고 다각적으로 파악하고 식품섭취, 영양상태, 주요 질병간의 연관관계를 분석할 자료제공을 주 목적으로 하고 있다. 이 자료는 국민의 질병예방, 영양개선, 건강수준 향상을 위한 보건정책 수립 및 평가에 중요한 기초자료가 된다.

지난 1998년에는 1995년도 인구주택총조사시 사용된 조사구와 1995년 인구주택총조사 이후의 신축 아파트로부터 표본조사구를 추출하여 조사하였다. 그러나 매년 약 30만 가구의 아파트가 신축되고 그간 가구들의 이동이 누적되는 등 현행 모집단이 너무 낙후되었다. 따라서 이들 모집단을 2001년에 실시할「국민건강 · 영양조사」의 모집단으로 사용하기가 적절치 않아 모집단을 잘 대표할 수 있도록 현행 모집단을 새로운 모집단으로 교체하였다.

본 연구에서는 2000년 인구주택총조사에 사용된 조사구를 새로운 조사구모집단으로 하고, 이에 적합한 표본설계를 하였다. 새로운 표본설계에서는 표본의 대표성을 높이기 위해서 기존의 표본가구수는 유지하면서 표본조사구를 증가시키고 새로운 지역 층화변수를 추가로 도입하였으며 추정량의 추정오차 공식을 유도하여 추정의 신뢰성을 측정할 수 있도록 설계하였다.

1) (360-764) 충북 청주시 상당구 내덕동 36번지, 청주대학교 자연과학부 통계학전공, 교수
E-mail: jbryu@chongju.ac.kr
2) (363-849) 충북 청원군 남일면 쌍수리 사서함335-2호, 공군사관학교 전산통계학과, 교수
E-mail: kayolee@afa.ac.kr
3) (140-742) 서울시 용산구 청파동 2가, 숙명여자대학교 통계학과, 교수
E-mail: ywkim@sookmyung.ac.kr

2. 현행 표본설계에 대한 검토

2.1. 현행 표본설계의 개요

각 시 · 도의 섬지역을 제외한 전국을 조사대상지역으로 하였다. 1995년도 인구주택총조사시 사용된 조사구와 1995년 인구주택총조사 이후 1997년 10월말까지의 신축아파트 가구를 조사구모집단으로 하였다. 표본규모는 건강면접조사의 경우는 200개 표본조사구에서 13,523가구를 표본으로 하였고, 보건의식행태조사, 영양조사, 건강검진조사 등 3개 조사는 건강면접조사 표본조사구내의 가구 중에서 약 1/3인 4,828가구를 표본으로 하였다.

표본조사구 배분을 위해서 1995년 인구주택총조사의 조사구모집단을 5개 층(제1층: 6대 시, 제2층: 기타 시의 동, 제3층: 시의 읍 · 면, 제4층: 군의 읍, 제5층: 군의 면)으로, 1995년 인구주택총조사 이후 1997년 10월말까지의 신축아파트를 3개의 층(제6층: 6대 시의 동, 제7층: 기타 시의 동, 제8층: 읍 · 면)으로 층화하여 총 8개의 층에 확률비례배분하였다.

1995년 인구주택총조사의 총 가구수는 12,885,650가구이고 1995년 인구주택총조사 이후 1997년 10월말까지 신축된 아파트가 596,900가구이므로 200개 표본조사구를 가구수에 비례하도록 191개 조사구와 9개 조사구를 각각 할당하였다. 배분의 기준으로 각 조사구의 일반가구수를 10으로 나누어 반올림한 결과를 크기의 측도로 사용하였는데, 200개 표본조사구를 가구수에 비례배분하면 읍 · 면지역에 표본조사구가 너무 적게 배분되기 때문에 읍 · 면 지역인 제3층, 제4층, 제5층과 제8층에 2배의 가중치를 적용하였다. 최초의 표본크기는 1995년 당시의 조사구에 따라 정하였으나, 그후 변동으로 인해 추가, 제외된 가구를 확인하여 조정하였다.(참고; 표 2.1, 보건복지부(1999))

제3층을 제외한 층에서는 각 층 내에서 주택 특성에 따라서 조사구를 분류하고 제3층에서 도시화 정도에 따라 조사구를 정렬하였으며 신축 아파트 가구들의 층에 대해서도 유사한 방법으로 조사구를 정돈하였다. 그리고 나서 크기 측도에 따라 확률비례계통추출법으로 각 층별로 표본조사구를 선정하였다.

표 2.1: 표본조사구 가구수 조정결과의 크기의 측도

자 료	층	모집단 크기의 측도	표본 크기의 측도		표본비율	
			최초	조정	최초	조정
인구주택 총조사	(1) 6대 시의 동	608,395	486	540	1/1,252	1/1,127
	(2) 기타 시의 동	408,580	328	345	1/1,246	1/1,184
	(3) 시의 읍 · 면	112,966	191	195	1/591	1/579
	(4) 군의 읍	61,178	102	101	1/600	1/606
	(5) 군의 면	108,048	176	179	1/614	1/604
신축 아파트	(6) 6대 시의 동	26,135	28	28	1/913	1/933
	(7) 기타시의 동	24,282	21	21	1/1,156	1/1,156
	(8) 읍 · 면	9,271	13	13	1/713	1/713
전 국		1,358,855	1,345	1,422	1/1,010	1/956

1998년 국민영양조사의 조사결과에 의한 추정은 가중치를 이용한 방법을 사용하였다. 이때의 가중치는 표본조사구별 추출률, 조사구별 조사미완률, 조사구내 조사가구수 등을 고려한 가중표본합계방법을 적용하여 평균과 구성비를 추정하였다. 건강면접조사에서는 전체 표본조사구의 조사대상 가구수 기대치(식(2-1)의 우측 분자)를 조사모집단의 조사대상 가구수 추정치(식(2-1)의 우측 분모)로 나누고 총수 추정용 조사구별 승수(M_{hi})를 곱해서 아래와 같은 가중표본합계치 산출용 조사구별 가중치를 사용하였다.(참고: 보건복지부(1999))

$$W_{hi} = M_{hi}\left[\frac{\sum_h \sum_i B_{hi}(A_{hi}/A'_{hi})}{\sum_h \sum_i M_{hi}B_{hi}}\right] \tag{2.1}$$

단, $M_{hi} = \frac{S_h}{n_h S_{hi}} \times \frac{A_{hi}}{A'_{hi}}$

$A_{hi} = h$층의 i번째 조사구내의 가구수

$A'_{hi} = h$층의 i번째 조사구내의 조사완료 가구수

$B_{hi} = h$층의 i번째 조사구내의 표본가구수

$S_h = h$층의 크기의 측도

$S_{hi} = h$층의 i번째 조사구의 크기의 측도

$n_h = h$층의 표본조사구 수

가중표본합계방법에 의한 추정치 계산은 다음과 같다.

$$\hat{Y} = \sum_h \sum_i W_{hi}\hat{Y}_{hi} \tag{2.2}$$

여기서 $\hat{Y}_{hi}$는 h층의 i번째 조사구에서 변수(y)에 대한 총계추정치이다.

2.2. 현행 표본설계 자료의 분석

2001년 표본설계에 필요한 정보를 얻기 위해서 1998년도 건강면접조사의 가구조사표에 포함된 17개 문항 중에서 건강상태측정에 핵심이 되는 8개 문항을 분석하여 도시지역(동부)과 읍 · 면 지역(읍면부)의 특성을 살펴보았다(표 2.2). 여기서 이환여부는 만성이나 급성질환을 앓았는지의 여부를, 제한 정도는 질병, 손상, 장애 등으로 3개월 이상 주요 활동의 제한 정도를 나타낸다.

2001년 표본조사설계에서 적용할 층화방법을 1998년 건강면접조사의 자료분석에 적용하기 위해서 전국을 6대 광역시와 6개 광역지방자치단체(경기도, 강원도, 충청도, 경상도, 전라도와 제주도)로 구분하였으며 12개 그룹 내에서 도시지역과 읍 · 면지역으로 층화해서 전체적으로 20개 층을 구성하였다.

2001년 표본조사설계에서 1차 추출단위(psu: primary sampling unit)는 조사구이므로 분석단위를 조사구로 한 자료분석은 필수적이다. 1998년 조사에 사용한 200개 표본

조사구를 20개 층으로 분류한 후에 각 조사구에서 건강면접조사한 가구별 자료를 조사구 단위로 합산 정리하여 3개 조사문항을 분석하였다. 표 2.3의 두 지역 모평균 차에 대한 대표본 근사 검정통계량값과 p-값에 의하면 조사구당 가구원수와 월평균소득은 도시지역과 읍·면 지역간에 큰 차이가 있음을 알 수 있다.

표 2.2: 문항별 지역 특성

		동부		읍면부	
		평균	표준편차	평균	표준편차
가구원수		3.33	1.56	3.00	1.21
성별	남	1.63	1.06	1.44	0.76
	여	1.69	1.08	1.56	0.77
이환 여부	Y	2.16	1.47	2.21	0.98
	N	1.17	1.40	0.79	0.90
제한 정도	1	0.01	0.14	0.03	0.16
	2	0.03	0.22	0.07	0.23
	3	0.09	0.38	0.20	0.39
	4	3.18	1.62	2.70	1.31
제한 원인	1	0.03	0.20	0.06	0.21
	2	0.00	0.08	0.01	0.09
	3	0.01	0.11	0.02	0.11
	4	0.01	0.11	0.02	0.10
	5	3.28	1.57	2.90	1.23
2주간 외래	Y	1.00	1.15	0.97	0.79
	N	2.32	1.60	2.03	1.18
상용 치료원	1	0.76	1.80	0.62	1.14
	2	0.06	0.57	0.03	0.29
	3	0.02	0.30	0.13	0.52
	4	0.26	1.15	0.18	0.69
	5	2.22	2.18	2.03	1.52
	6	0.00	0.13	0.00	0.07
월평균소득		138.31	118.27	90.53	68.28

표 2.3: 분석결과(조사구)

	동부		읍면부		z	p
	평균	표준편차	평균	표준편차		
가구원수	201.4	31.8	185.1	22.9	4.17	0
이환여부	130.55	26.08	135.94	15.56	1.83	0.067
월평균소득	8367.9	3022.1	5578.7	1772.2	8.22	0

3. 새로운 표본설계

3.1. 모집단 분석

2000년 인구주택총조사에서 사용된 일반조사구수는 24만 6천여 개이며 총 가구수는 1480만 가구이고 13개 광역층 내에서 동부와 읍면부로 나누고 또 다시 아파트와 보통 조사구로 구분하여 정리한 내용이 표 3.1에 있다.

표 3.1: 지역별 조사구 분포

지역	계	합 계		계	동 부		계	읍면부	
		아파트	보통		아파트	보통		아파트	보통
서울	54830	16253	38577	54830	16253	38577	0	0	0
부산	19323	6989	12334	18950	6880	12070	373	109	264
대구	13122	4966	8156	12375	4669	7706	747	297	450
인천	12733	5551	7182	12302	5540	6762	431	11	420
광주	6987	3571	3416	6987	3571	3416	0	0	0
대전	7090	3196	3894	7090	3196	3894	0	0	0
울산	5274	2377	2897	4386	1994	2392	888	383	505
경기	44970	18510	26460	35513	16174	19339	9457	2336	7121
강원	8388	2677	5711	4800	2181	2619	3588	496	3092
충청	17925	5431	12494	7547	3732	3815	10378	1699	8679
전라	21751	6101	15650	10668	5313	5355	11083	788	10295
경상	31119	9398	21721	16058	6833	9225	15061	2565	12496
제주	2585	341	2244	1778	336	1442	807	5	802
전국	246097	85361	160736	193284	76672	116612	52813	8689	44124

3.2. 새로운 표본설계의 특징

과거의 표본설계와 비교해서 새로운 표본설계가 갖는 특징은 다음과 같다.

(1) 기존(1998년조사)에는 1995년 인구주택총조사의 조사구와 1995년 인구주택총조사 이후의 신축아파트를 조사구모집단으로 하였다. 그러나 새로운 표본설계에서는 2000년 인구주택총조사의 조사구에서 10%표본조사구를 제외한 나머지 90%를 조사구모집단으로 하였다. 이는 통계청에서 실시하고 있는 각종 조사(경제활동인구조사, 도시가계조사 등)에 10% 표본조사구를 모집단으로 사용함으로써 표본조사구의 중복 가능성이 있어 조사 수행의 어려움을 감안 한 것이다.

(2) 종전의 면접조사는 200개 표본조사구내에 있는 모든 가구를 대상으로 실시하고 보건의식행태조사, 영양조사, 건강검진조사는 표본조사구내의 가구 중에서 1/3을 표본으

로 추출하여 실시하였다. 그러나 새로운 표본설계에서는 표본조사가구의 크기는 종전과 같은 수준으로 유지하되 표본의 대표성과 정도를 높이기 위해서 면접조사 표본조사구를 600개로 증가시켰고 나머지 3개 조사는 600개 표본조사구 중에서 200개 조사구를 추출하여 사용한다. 표본가구는 4개 조사 모두 표본조사구내 가구 중에서 1/3을 추출하여 실시한다.

(3) 과거의 표본설계에서는 6대 광역시의 동, 기타 시의 동, 시의 읍 · 면, 군의 읍, 군의 면을 기본 층화변수로 사용하였으나 새로운 표본설계에서는 지역 층(7대 광역시, 경기, 강원, 충청, 전라, 경상, 제주)과 행정구역(도시지역, 읍 · 면지역)을 층화변수로 사용하였다. 그리고 표본배분을 위해서 읍 · 면 지역은 크기의 측도(가중치)를 2배로 하여 층별 비례배분하였으나 이는 이론적 배경이 미흡하므로 새로운 표본설계에서는 각 층의 조사구수에 비례하도록 표본을 비례배분하였다.

(4) 표본조사구는 조사구를 주택특성에 따라 단독주택이 많은 조사구, 아파트가 많은 조사구, 연립 및 다세대주택이 많은 조사구, 그리고 기타 조사구 순서로 분류하여 크기의 측도에 따라 확률비례계통추출하였다. 그러나 새로운 표본설계에서는 간단히 아파트지역과 비아파트지역으로 나누고, 지역적인 안배를 고려하여 각 층내에서 행정구역에 따라 조사구를 정렬한 후 표본조사구를 계통추출하였다.

(5) 표본가구는 확률추출에 근거를 두지 않은 관계로 조사원들에 의해 표본가구가 변경될 가능성이 있었다. 그러나 새로운 표본설계에서는 조사구내의 표본을 확률추출에 의해서 선정하고 추출된 가구가 누락되거나 응답거부일 경우를 대비해서 예비표본을 선정하였다. 지난 조사에서는 거처를 단위로, 선정된 조사대상 가구 중 비혈연가구와 조사대상가구의 가구원 중에서 비혈연가구원을 제외하였으나 이번에는 비혈연가구원은 종전과 같이 조사에서 제외하나 비혈연가구는 다른 가구로 교체함으로써 결측자료를 줄이도록 하였다.

(6) 새로운 표본설계에서는 가중치와 이를 이용한 추정량은 물론 추정오차공식도 유도하였다.

3.3. 층화 및 효율성비교

2001년 표본조사설계에서 사용한 조사구는 2000년 11월 1일을 기준으로 실시된 인구주택총조사를 위해 작성된 것으로, 조사구의 특성은 주택형태(아파트, 일반주택)와 가구수만 알 수 있다. 따라서 표본크기 결정이나 신뢰수준의 결정은 다른 차원에서 검토해야 할 것이다. 2001년 표본조사설계에서는 행정구역을 층화 기준으로 하여 광역자치단체별 주요통계 작성을 염두에 두었고 도시지역과 읍 · 면지역을 층화작업시 고려하였다. 전국을 7대 광역시와 6개 광역지방자치단체(경기도, 강원도, 충청도, 경상도, 전라도와 제주도)로 구분하였으며 13개 그룹 내에서 도시지역과 읍 · 면지역으로 층화해서 전체적으로 20개 층으로 구성하였다.

기존 표본설계에 대한 새로운 표본설계의 효율성을 비교하기 위해서 아래의 층화추출의 분산 공식을 사용하였다.

$$\hat{V}(\bar{y}_{st}) = \sum_{h}^{L} W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) s_h^2$$

기존 표본설계에서는 층을 8개로($L = 8$)하였고 새로운 표본설계에서는 층을 20개로 하였다. 분석단위를 조사구로 한 경우, 1998년도 자료를 근거로 한 기존 표본설계와 새로운 표본설계의 분산과 상대효율이 표 3.2에 있다. 분석단위를 가구로 할 때도 결과는 거의 유사하다. 전체적으로 볼 때 새로운 표본설계로 인한 효율성이 기존의 경우보다 60% 정도 증가하였다.

표 3.2: 분석단위가 조사구일 때 상대효율

구분	가구원수	이환여부	월평균소득
기존 표본설계	3.99942	2.545384	33,952.83
신규 표본설계	2.44395	1.583857	22,039.30
상대효율(%)	163.65	160.71	154.06

<참고> 상대효율 $= \frac{\text{기존표본설계분산}}{\text{신규표본설계분산}} \times 100$

3.4. 표본크기의 결정 및 배분

표본크기는 확보된 예산과 행정업무처리능력을 고려하고 추정치의 신뢰수준을 감안하여 건강면접조사에서는 600개 조사구를 표본조사구로 하고 각 조사구에서 20여 가구(조사구내 가구의 1/3)를 계통추출하여 면접조사 한다. 영양조사와 건강검진조사에서는 600개 조사구에서 200개 조사구를 랜덤하게 선정하며 선정된 조사구 내에서 건강면접조사시에 표본가구로 선정된 가구를 대상으로 조사를 실행한다.

600개의 조사구를 배분하는 방법으로 Neyman배분과 식(3-1)를 사용하는 비례배분을 적용할 수 있다.

$$n_h = \frac{n \cdot N_h}{N}, \quad h = 1, 2, \cdots, L \tag{3.1}$$

두 방법을 비교하고 연구진과 자문위원들과의 논의를 통해서 비례배분법을 사용하기로 하였다. 이는 2000년 인구주택총조사에서 사용한 조사구와 1998년 건강면접조사의 표본조사구는 조사변수의 특성에서 차이가 있을 것으로 생각할 수 있을 뿐만 아니라 2001년 건강면접조사에서 각 층의 추정량에 영향을 주는 것은 각 층의 크기(N_h)와 주택 형태가 될 것이므로 이를 고려하여 각 층의 조사구 수에 비례하도록 표본조사구를 배분하는 것이다.

2001년 조사에 사용될 20개 층에 배분하기 위해서 1998년 건강면접조사에서 표본조사구가 하나인 층에 대해서는 3개의 표본조사구를 사전에 할당한 후에 나머지 조사구를 나머지 층에 배분하였다. 경상남도와 울산광역시의 표본배분은 2000년 인구주택총조사의 조사구수에 비례하도록 배분하였다. 앞에서 언급한 바와 같이 1차 추출단위는 조사구이므로 각 층의 조사구 수에 비례하도록 표본조사구를 배분하였다(표 3.3 참조). 그리고 층별 조사구는 2000년 인구주택총조사에 사용된 90%조사구로부터 600개의 표본조사구를 계통추출하였으며 표본조사를 위한 대상가구도 각 조사구로부터 전체 가구의 1/3을 표본가구로 계통추출하였다. 한편 표본가구 중에서 표본으로 사용할 수 없는 가구가 있을 경우에는 교체표본을 사용하는 데 이를 위해서 예비표본가구를 선정하였다.

표 3.3: 조사구수 기준 비례배분

지역	계	합 계		계	동 부		계	읍면부	
		아파트	보통		아파트	보통		아파트	보통
서울	129	36	93	129	36	93	0	0	0
부산	48	18	30	48	18	30	0	0	0
대구	30	12	18	30	12	18	0	0	0
인천	30	15	15	30	15	15	0	0	0
광주	18	9	9	18	9	9	0	0	0
대전	18	9	9	18	9	9	0	0	0
울산	15	6	9	12	6	6	3	0	3
경기	108	45	63	84	39	45	24	6	18
강원	21	6	15	12	6	6	9	0	9
충청	45	15	30	18	9	9	27	6	21
전라	54	18	36	27	15	12	27	3	24
경상	75	24	51	39	18	21	36	6	30
제주	9	0	9	6	0	6	3	0	3
전국	600	213	387	471	192	279	129	21	108

3.5. 추정방법

2001년「국민건강 · 영양조사」의 조사결과에 의한 각종 통계치는 전국 모집단 평균이나 구성비이다. 아울러 필요에 따라 1998년 실시된 동일 조사의 총괄보고서와 단순비교가 가능하도록 본 조사결과를 활용하여 가중표본합계치(weighted sample total)를 산출하는 방법도 함께 제시한다.

추정방법은 본 조사의 표본추출방법인 층화이단집락추출법을 통하여 얻어진 표본조사자료를 이용하여 모집단에 대한 통계치(전국 가구별 또는 개인별 평균 또는 구성비)를 산출하는 방법을 제시하고 있으며, 표준오차 추정방법은 실제 표본설계에서는 최종 추출단계에서 계통추출방법을 사용했지만 이를 단순임의추출한 것으로 간주하고 계산한 것이기

때문에 실제 표준오차를 약간 과대추정(over-estimation)할 수 있다.

건강면접조사의 경우 최종 추출 및 조사단위는 가구이며, 표본추출방법을 기초로 한 추정방법, 표준오차 계산방법, 그리고 실제 자료처리에 있어서 가구별 가중치를 사용한 추정방법을 함께 제시하였다. 이들 중 어떤 추정방법을 사용하더라도 결과적으로 동일한 통계치를 산출하게 된다. 한편 보건의식행태조사, 영양조사, 건강검진조사에 대한 추정문제는 한국조사연구학회(2000)를 참고하기 바란다.

3.5.1. 표본추출방법에 의한 모집단 평균 추정

(1) 총계 추정

본 표본설계에서 적용한 층화이단집락추출법에 따른 식(3-2)로 전국 총계를 추정할 수 있다.

$$\hat{\tau} = \sum_{h=1}^{L} \frac{N_h}{n_h} \sum_{i=1}^{n_h} \frac{M_{hi}}{m_{hi}} \sum_{j=1}^{m_{hi}} y_{hij} = \sum_{h=1}^{L} \frac{N_h}{n_h} \sum_{i=1}^{n_h} M_{hi} \bar{y}_{hi} \tag{3.2}$$

이 공식에서,

N_h : h층의 총 조사구수

n_h : h층의 표본조사구수

M_{hi} : h층 i번째 표본조사구의 총 가구수

m_{hi} : h층 i번째 표본조사구에서 추출된 표본가구수

y_{hij} : h층 i번째 표본조사구 j번째 가구에서 변수(y) 관측값

$\bar{y}_{hi} = \frac{1}{m_{hi}} \sum_{j=1}^{m_{hi}} y_{hij}$: h층 i번째 표본조사구에서 추출된 표본가구들의 관측값 평균

(2) 가구당 평균 추정

전국 가구당 평균은 위에서 구한 $\hat{\tau}$을 이용하여 다음 식으로 산출한다.

$$\hat{\mu} = \hat{\tau}/M \tag{3.3}$$

여기서 M은 전국 총 가구수이고, 이에 대한 정확한 값을 확보하는 데 현실적으로 상당한 어려움이 예상되기 때문에 본 조사결과를 이용하여 전국 총 가구수(M)를 다음과 같이 추정하여 사용한다.

$$\hat{M} = \sum_{h=1}^{L} \frac{N_h}{n_h} \sum_{i=1}^{n_h} M_{hi} \tag{3.4}$$

따라서 전국 가구당 평균은 다음과 같이 산출한다.

$$\hat{\mu}^* = \frac{\hat{\tau}}{\hat{M}} \tag{3.5}$$

(3) 가구당 평균에 대한 표준오차 추정

가구당 평균에 대한 표준오차를 설명해 주는 추정치에 대한 추정분산은 층화이단집락추출법을 사용하였으므로 다음과 같이 된다.

$$\hat{V}(\hat{\mu}) = (\frac{1}{M^2}) \left[\sum_{h=1}^{L} (\frac{N_h - n_h}{N_h})(\frac{N_h^2}{n_h}) s_h^2 + (\frac{N_h}{n_h}) \sum_{i=1}^{n_h} M_{hi}^2 (\frac{M_{hi} - m_{hi}}{M_{hi}})(\frac{s_{hi}^2}{m_{hi}}) \right] \tag{3.6}$$

여기서,

$s_h^2 = (\frac{1}{n_h - 1}) \sum_{i=1}^{n_h} (M_{hi}\bar{y}_{hi} - \bar{M}_h\hat{\mu}_h)^2$,

$s_{hi}^2 = (\frac{1}{m_{hi}-1}) \sum_{j=1}^{m_{hi}} (y_{hij} - \bar{y}_{hi})^2$

이고

$\bar{y}_{hi} = \frac{1}{m_{hi}} \sum_{j=1}^{m_{hi}} y_{hij}$: h층 i번째 표본조사구에서 추출된 표본가구들의 관측값 평균

$\bar{M}_h = \frac{M_h}{N_h}$: h층 조사구들의 평균 가구수

$\hat{\mu}_h = (\frac{1}{M_h})(\frac{N_h}{n_h}) \sum_{i=1}^{n_h} \frac{M_{hi}}{m_{hi}} \sum_{j=1}^{m_{hi}} y_{hij}$.

한편 가구당 평균에 대한 상대표준오차를 설명해 주는 변동계수(coefficient of variation)는 다음과 같이 추정된다.

$$cv(\hat{\mu}) = \frac{\sqrt{\hat{V}(\hat{\mu})}}{\hat{\mu}} \times 100\% \tag{3.7}$$

표준오차 계산에 있어서 전국 총가구수 M과 h층의 총 가구수 M_h가 필요한 데 이에 대한 정확한 값은 구할 수 없으므로 전국 총 가구수 M은 위의 식(3-4)에서 산출된 값을 사용하고, M_h는 다음과 같이 추정하여 사용한다.

$$\hat{M}_h = \frac{N_h}{n_h} \sum_{i=1}^{n_h} M_{hi}$$

이 경우 위 식에서

$$\hat{\mu}_h = \frac{\sum_{i=1}^{n_h} \frac{M_{hi}}{m_{hi}} \sum_{j=1}^{m_{hi}} y_{hij}}{\sum_{i=1}^{n_h} M_{hi}}.$$

즉, 비추정량(ratio estimator)을 사용한다.

3.5.2. 가중평균을 사용한 모집단 평균 추정

모집단 평균을 산출하기 위한 가중치는 다음과 같다.

(1) 가구별 확대승수 산출

각 표본가구가 표본으로 추출된 확률의 역수에 해당하는 총수 추정용 가구별 확대 승수를 산출한다.

$$V_{hi} = \frac{N_h}{n_h}\frac{M_{hi}}{m_{hi}}$$

실제 각 가구별 가중치를 전산파일의 가구별 자료에 적용하는 대신 조사구별로 한번만 가중치를 적용하는 것이 계산상 편리하다고 판단되면 다음의 가중치를 가구별 관측값 대신에 조사구별 변수의 표본평균($\bar{y}_{hi}$)에 적용하여 사용할 수 있다.

$$V^*_{hi} = \frac{N_h}{n_h}M_{hi}$$

(2) 모집단 총 가구수($\hat{M}$)를 반영한 모집단 평균을 산출하기 위해서 가중치를 변환한다.

$$W_{hi} = (\frac{1}{\hat{M}})V_{hi}$$

한편, 조사구별 평균에 확대승수를 적용하는 경우 다음 가중치를 사용한다.

$$W^*_{hi} = (\frac{1}{\hat{M}})V^*_{hi}$$

여기서 전국 총 가구수에 대한 추정치 $\hat{M}$은 식(3-4)에서 구한 것을 사용한다.

(3) 이와 같이 산출된 가중치를 각 변수에 적용하여 합산하는 방법으로 모집단 평균에 대한 추정치를 산출한다.

$$\hat{\mu}^* = \sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} W_{hi}y_{hij} = \sum_{h=1}^{L}\sum_{i=1}^{n_h} W^*_{hi}\bar{y}_{hi}$$

가중평균을 적용해서 산출된 모집단 평균에 대한 추정치는 식(3-5)의 추정치와 동일하다.

3.5.3. 가중평균을 사용한 가중표본합계치 산출

앞에서 언급한 것과 같이 1998년도 총괄보고서(보건복지부(1999))와 단순비교를 위한 가중표본합계치(weighted sample total)를 산출하기 위한 가중치는 다음과 같다.

(1) 조사대상 표본가구수

조사완료(교체 포함)된 표본가구수를 다음과 같이 산출하고, 다음 단계에서 가중표본합계를 조사대상 표본가구 총수와 일치되도록 가중치를 조정하는데 사용한다.

$$m = \sum_{h=1}^{L} \sum_{i=1}^{n_h} m_{hi}$$

(2) 모집단 총 가구수($\hat{M}$)를 실제 조사대상 표본가구수로 변환하기 위해 가구별 확대승수를 가중표본합계 산출용 가중치로 변환

$$W'_{hi} = (\frac{m}{\hat{M}})V_{hi}$$

한편, 조사구별 평균에 확대승수를 적용하는 경우 다음 가중치를 사용한다.

$$W^{*\prime}_{hi} = (\frac{m}{\hat{M}})V^*_{hi}$$

이와 같이 산출된 가중치를 각 변수에 적용하여 합산하는 방법으로 가중표본합계를 산출한다.

$$\hat{Y}_s = \sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} W'_{hi} y_{hij} = \sum_{h=1}^{L} \sum_{i=1}^{n_h} W^{*\prime}_{hi} \bar{y}_{hi}$$

여기서 W'_{hi}는 가구별 가중치로 각 가구별 관측값에 적용한 것이고, $W^{*\prime}_{hi}$는 조사구별 가중치로 각 조사구별 표본평균에 적용한 가중치이다. 따라서 모든 표본 가구에서 $y_{hij} = 1$이면 $\hat{Y}_s = m$, 즉 조사대상 표본가구수가 산출된다. 한편 가중표본합계 $\hat{Y}_s$는 위 표본추출방법에 의한 추정식에서 $m\hat{\mu}^*$에 해당한다.

4. 표본관리 및 대체

새로운 표본설계는 2000년 인구주택총조사에 사용된 조사구를 새로운 조사구모집단으로 하였다. 조사가 1회 조사가 아닌 경우(계절별조사)에는 조사대상이 시간의 경과에 따라 변동하므로 이에 따른 표본관리가 필요하며, 변동된 표본의 특성을 감안한 표본의 대체문제를 고려해야 한다. 또한 조사시에 응답자들로부터 완전한 응답을 얻는다는 것은 현실적으로 불가능하므로 응답자들의 응답거부나 불성실한 응답으로 인한 무응답오차의 발생을 최소화하도록 한다. 특히 본 연구의 조사는 4개의 조사가 함께 이루어지는 복합조사로 조사주관 기관이 다르고 조사 항목이 다양한 관계로 무응답률의 증가가 우려된다. 이를 위해서 조사원의 선발과 훈련, 그리고 조사기관들간의 유기적인 협조체계를 구축하는 것이 중요하다.

4.1. 표본관리

모집단을 구성하고 있는 조사구들은 약 60가구로 구성되어 있고 이들 조사구로부터 표본가구를 추출하여 조사를 하게 된다. 1회 조사가 아닌 경우 모집단은 시간이 경과함에 따라 변동하게 된다. 기존의 가구가 없어지거나 새로운 가구가 생기는 등으로 조사구내 변동이 생길 수 있고, 기존의 아파트나 가구들이 철거되고 새로운 아파트나 주택들이 신축되는 경우에도 조사구의 수정 및 보완은 필요하다. 따라서 모집단의 변동을 수시로 파악하여 이들을 조사에 반영해야 한다.

모집단에 변동이 생기면 이를 즉시 표본에 반영하여야 모집단추정의 정도를 유지해서 연구의 목적을 달성할 수 있다. 모집단의 변동이 크면 모집단에 대한 새로운 정보를 추가하여 모집단을 개편하고 이에 따라 표본설계도 변경하여야 한다. 따라서 전체적인 표본조사구 수와 표본가구의 수도 변하게 된다. 예를 들면 표본조사구내에서 표본가구의 추출률의 역수보다도 많은 가구의 변동이 있으면 변동에 비례해서 표본가구를 증감시킨다.

4.2. 표본대체

이번 조사에 사용될 모집단은 2000년 인구주택총조사에 사용된 조사구이므로 조사구의 변동은 없을 것으로 보인다. 그러나 실제로 조사를 실시하면 표본조사구내 가구들의 변동이 부분적으로 있고, 표본으로 선정된 가구라 할지라도 조사에 불응하거나 불성실한 응답을 하는 등의 문제가 생기게 된다. 즉, 표본조사자료가 결측값이 있는 불완전한 자료가 된다.

통계조사에서는 연구모집단의 특성인 모집단 평균이나 총계에 대한 정확한 추정치를 얻는 것이 주요 관심사인데 조사에서 얻은 자료가 결측값이 있는 불완전한 자료인 경우에는 추정치에 편향이 생기고 추정의 효율이 떨어지게 된다. 따라서 표본결측 자료에 대한 대체문제를 고려하여야 한다.

본 연구에서는 표본조사구내 가구들이 부재이거나, 표본으로 선정된 가구가 조사에 불응 또는 불성실한 응답을 하는 경우에 표본조사가구를 교체하는 방법을 사용한다. 이를 위해서 예비표본조사가구를 설정하였다.

5. 제안

우리 나라 보건복지통계의 중요한 자료인「국민건강 · 영양조사」의 발전과 조사자료의 효율적인 활용을 위해서 다음과 같은 몇 가지 사항에 대한 연구검토를 제안한다.

(1) 기존('98년도)에는 모집단을 1단계 층화 후 계통확률비례추출(표본조사구 추출시)방법을 사용하였다. 그러나 향후에는 모집단을 적절한 층화변수들로 층화하고 층별 또는 보다 작은 지역별 통계를 생산하기 위해서 소지역을 조사구들로 구성된 집락으로 한 다단계 층화집락추출방법을 사용하고 표본조사구와 최종표본가구는 계통추출법을 사용하는 것이 바람직하다.

(2) 표본개편 시기를 인구주택 총조사가 실시되고 조사결과를 충분히 활용할 수 있는 시점(약 1년 6개월 후)으로 변경하는 것이 바람직하다. 특히 「국민건강 · 영양조사」는 일반국민의 건강과 관련된 중요한 전국단위의 조사로 국내에서 유일하다. 이러한 전국단위의 국민건강 관련자료는 정부뿐만 아니고 민간의료기관과 학술단체 그리고 외국기관에서도 절대적으로 요구되는 자료이므로 조사의 주기를 3년에서 매년 실시로 전환해야 보다 정확한 통계자료를 적시에 얻을 수 있다.(물론, 일부 조사의 경우는 분기별 실시가 필요하다(rotation sampling 기법을 이용))

(3) 기존의 「국민건강 · 영양조사」보고서에는 표본설계시 사용된 층화변수에 따른 분석이 미비하고 생산된 통계의 비교분석에 있어서 통계적 유의성 문제가 언급되어 있지 않다(단순히 빈도 분석에 그침). 따라서 지역별, 또는 층별변수를 도입한 상세한 통계적 분석이 요구된다.

(4) 지난 조사의 결과분석에는 가구단위든 가구원단위든 간에 추정치와 추정치에 조사단위수를 곱한 추정치를 제시하고 있다. 그러나 추정치는 모집단에 대한 예측치로서 의미가 있는데 여기서 제시하고 있는 추정치는 단순히 표본조사자료에 대한 예측치를 뜻한다. 따라서 보고서에 제시한 추정치는 의미가 없으며, 자칫하면 모집단의 추정치로 오인할 우려가 있으므로 이러한 추정치의 사용을 제외하는 것이 바람직하다.

(5) 현실적으로 표본크기가 작은 관계로 유의적인 소지역별 통계자료의 생산이 어려우므로 조사자료와 기타 자료를 바탕으로 한 소지역추정기법을 사용한 지역별 통계자료의 생산이 필요하다.

참고문헌

[1] 고려대학교 통계연구소 (1999). *임금구조 기본통계조사 표본개편 연구보고서.*

[2] 류제복 (2000). 무응답 대체방안, *산업경영연구*, Vol. 23, No. 2, 227-244.

[3] 보건복지부 (1999). *98 국민건강 · 영양조사 총괄보고서.*

[4] 서울대학교 통계연구소 (1996). *노동통계 표본설계.*

[5] 서원대학교 통계연구소 (2000). *임업 업종별 경영실태조사를 위한 표본설계.*

[6] 조사통계연구회 (2000). *무응답오차*, 자유아카데미.

[7] 한국조사연구학회 (2000). *2001년도 국민건강 · 영양조사 표본설계 및 표본조사구 추출.*

[8] 한국통계학회 (1999). *'99 소규모사업체 근로실태조사 표본설계.*

[9] Kish, L. (1965). *Survey Sampling*, John Wiley & Sons, Inc.

[10] Kish, L. (1992). Weighting for unequal P_i, *Journal of Official Statistics*, Vol. **8**, No. 2, 183-200.

[11] Lent, J., Miller, S.M., Cantwell, P.J. and Duff, M. (1999). Effects of composite weights on some estimates from the Current Population Survey, *Journal of Official Statistics*, Vol. **15**, No. 3, 431-448.

[12] Lessler, J.T. and Kalsbeek, W.D. (1992). *Nonsampling error in surveys*, John Wiley & Sons, Inc.

[13] Madow, W.G., Nisselson, H., Olkin, I. and Rubin, D.B. (1983). *Incomplete data in sample surveys*, Vol. **1** - Vol. **3**, New York : Academic Press.

[14] Oh, H.L. and Scheuren, F.J. (1983). Weighting adjustment for unit nonresponse, In Madow, W.G., Olkin, I. and Rubin, D.B., eds., *Incomplete data in sample surveys*, Vol. **2**, Academic Press, Inc., 143-184.

[15] U.S. Department of Health and Human Services (1989). *Vital and Health Statistics - Design and Estimation for the National Health Interview Survey, 1985-94*, Series 2, No. **110**.

[16] U.S. Department of Health and Human Services (1999). *Vital and Health Statistics - National Health Interview Survey: Research for the 1995-2004 Redesign*, Series 2, No. 126.

[17] U.S. Department of Health and Human Services (2000). *Vital and Health Statistics - Design and Estimation for the National Health Interview Survey, 1995-2004*, Series 2, No. 130.

[18] Verma, V. (1991). *Sampling Methods*, Training Handbook Statistical Institute for Asia and the Pacific, Tokyo.

[2001년 5월 접수, 2001년 8월 채택]

A Sampling Design for the 2001 National Health · Nutrition Survey

Jea-Bok Ryu[1] Kay-O Lee[2] Young-Won Kim[3]

ABSTRACT

We propose a new sampling design for the 2001 National Health·Nutrition Survey. In this design, an area variable is introduced as a stratification variable. We increase the number of ED(enumeration district) in order to improve the representation of the sample while keeping the number of sample house as before. And also, we design to measure the reliability of the estimator as deriving the formula of the estimation error.

Keywords: Stratification Variable; Systematic Sampling; Proportional Allocation; Weighted Sample Total.

1) Professor, Department of Statistics, Chongju University.
E-mail: jbryu@chongju.ac.kr
2) Professor, Department of Computer Science and Statistics, Korea Air Force Academy.
E-mail: kayolee@afa.ac.kr
3) Professor, Department of Statistics, Sookmyung Women's University.
E-mail: ywkim@sookmyung.ac.kr