The Korean Communications
in Statistics Vol. 9, No. 2, 2002
pp. 565-576

A Sampling Design for Health Index Survey

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

Abstract

We propose a new sampling design for the 2001 Health Index Survey at Seoul. In this stratified two-stage sampling design, the ED(enumeration district) of 2000 Population and Housing Census is used as primary sampling unit and the Gu is used as stratification variable in order to obtain the sub-domain estimate for 25 Gu's as well as population estimate for Seoul. The sample ED's are systematically selected after the Ed's are ordered by location and property to obtain a representative sample. And also, the imputation methods for item nonresponses are suggested.

Keywords : Classification and regression tree, Optimal allocation, Variance estimation, Imputation

1. 서론

국민의 건강문제를 적극적으로 대처하기 위해서 1995년 1월에 국민건강증진법이 제정·공포되었다. 이 법에는 국가 및 지방자치단체가 건강에 관한 국민의 관심을 높이고 국민건강을 증진할 책임을 지며, 지방자치단체의 장은 국민건강의 증진에 관한 기본시책에 따라 지방자치단체의 실정을 감안하여 주민건강의 증진에 관한 세부계획을 수립·시행할 것을 규정하고 있다. 또한 지역보건법에서도 지방자치단체장이 지역보건의료계획을 수립하고 보건복지부령이 정하는 바에 의하여 지역보건의료계획을 시행하도록 규정하고 있다.

서울시는 광역지방자치단체로서는 처음으로 1997년 9월1일~11월 30일(3개월간)까지 25개 구의 약 20,000가구(6만 7천여 가구원)를 대상으로 서울시민의 건강수준과 의료이용실태에 관한 조사를 실시한 바 있다. 지난 조사의 모집단은 1995년도 인구주택총조사의 보통조사구 자료와 1995년도 인구주택총조사 이후 1997년 4월 30일까지의 신축아파트를 새로운 조사구로 구성한 후 표본조사구를 추출하여 조사하였다. 그러나 기존의 모집단은 그간 변동이 심하고 낙후되어 2001년에 실시한 보건지표조사의 모집단으로 사용하기가 적절치 않다. 따라서 2001년에 실시된 보건지표조사를 위한 새로운 표본설계가 필요했다. 본 논문에서는 현행 모집단을 잘 대표하고 추정량의 정확성을 높이기 위해서 2000년 인

1) Professor, Statistics, Division of Natural Science, Chongju University, Chongju, 360-764, Korea.
E-mail : jbryu@chongju.ac.kr
2) Professor, Department of Computer Science and Statistics, Korea Air Force Academy, Chongju, 363-849, Korea.
E-mail : kayolee@afa.ac.kr
3) Professor, Department of Statistics, Sookmyung Women's University, Seoul, 140-742, Korea.
E-mail: ywkim@sookmyung.ac.kr

구주택총조사에 사용된 조사구를 모집단조사구로 하였다.

새로운 표본설계에서는 서울시 전체는 물론 25개 구별 단위의 통계를 얻기 위해 층화변수로 구를 사용하였다. 구 단위 통계의 정도를 측정해서 요구되는 허용오차수준에 따른 표본크기를 구하고 층에 따라 표본을 배정하였다. 그리고 모집단의 특성을 추정하는 데 사용되는 추정식을 구하고 추정오차를 계산하기 위한 추정량의 분산공식을 유도하였다. 또한 조사결과 무응답이 발생한 문항들에 대한 무응답 대체방법을 다루었다.

2. 현행표본설계에 대한 검토

2.1. 현행 표본설계의 개요

1997년도 시민지표조사를 위한 표본설계에서는 1995년도 인구주택총조사시 사용된 52,047개의 조사구(2,965,794가구)에 1995년 인구주택총조사 이후 1997년 4월 30일까지 준공검사를 받은 신축 아파트 476개 조사구(신축 아파트의 경우 1개 동을 1개의 조사구로 취급 : 49,764가구)를 합한 총 52,523조사구를 모집단조사구로 하였다.

표본규모는 주어진 예산을 고려하여 1,500개 표본조사구에서 확인된 23,511가구 중 22,471 대상가구에 대해서 총 19,787가구를 조사완료하였다(참고: <표 2-1>). 건강수준과 의료이용실태조사는 표본가구의 전체 가구원 67,099명을 대상으로 조사하였으며 보건의식행태조사는 표본가구의 가구원 중 생일이 가장 빠른 15~69세 가구원 1명을 선정하여 조사하였는데 19,787 가구중에서 17,256 가구를 조사완료하였다.

<표 2-1> 구별 조사 가구 수

구	확인 가구수	조사 가구수	구	확인 가구수	조사 가구수	구	확인 가구수	조사 가구수
종 로 구	921	790	도 봉 구	934	796	영등포구	985	790
중 구	889	787	노 원 구	928	804	동 작 구	972	791
용 산 구	950	777	은 평 구	922	783	관 악 구	945	795
성 동 구	913	803	서대문구	930	793	서 초 구	932	793
광 진 구	961	797	마 포 구	954	781	강 남 구	918	783
동대문구	900	808	양 천 구	920	783	송 파 구	933	799
중 랑 구	998	791	강 서 구	919	781	강 동 구	951	803
성 북 구	943	806	구 로 구	971	784	합 계	23,511	19,787
강 북 구	927	792	금 천 구	995	777			

구별 조사구 규모를 60개 조사구의 800가구로 하고, 조사가 완료된 후에 25개 구를 4~5개의 생활권으로 통합하여 추정치를 산출하고, 생활권별로 조사결과를 분석하였다. 표본조사구는 각 구별로 주택특성에 따라 ① 단독주택이 많은 조사구, ② 아파트가 많은 조사구, ③ 연립 및 다세대주택이 많은 조사구, ④ 기타 조사구의 순서로 1차 분류한 후 60개의 표본조사구를 가구 수에 비례

하는 확률로 계통추출하였다. 1차 분류내에는 행정구역 및 조사구 번호 순서로 나열하였다.

한편 1,500개 표본조사구의 가구명부 작성을 위한 업무량이 방대하여 많은 인력과 시간이 소요되므로 주어진 예산과 기간을 고려하면 조사구별 가구파악은 현실적으로 불가능하였다. 따라서 인구주택총조사 당시의 조사구요도에 15~19가구가 포함되는 구역을 설정하고 이 구역에서 13~14가구를 조사하였다. 비혈연가구와 외국인가구로 판명되면 부적격가구로 처리하고 적격가구에 대해서는 개인별 조사를 실시하였다. 보건의식행태조사에서는 비혈연가구원을 제외한 가구원으로 15~69세의 사람 중에서 출생년월에 관계없이 생일이 가장 빠른 사람에 대하여만 조사하였다. 적격가구에 대하여는 재방문을 해서 각 조사구별로 최소한 13가구의 조사표를 작성하도록 하였다.

2.2. 현행 표본설계 자료의 분석

건강수준 및 의료이용실태조사에서 조사된 항목 중에서 관심의 대상이 될 수 있는 6개 항목에 대한 서울시 전체의 분석 결과와 서울시 전체 자료를 조사구단위로 분석한 내용이 <표 2-2>에 있다. 조사된 자료를 분석한 항목 중에서 성별 구성비는 각 구별로 대동소이하였으나 이환여부와 2주간 활동제한 여부는 구별로 유의할 만한 차이를 보이고 있음을 <표 2-3>에서 알 수 있다.

<표 2-2> 건강수준 및 의료이용실태조사 분석

항목		가구수 단위		조사구 단위	
		평균	표준편차	평균(비율)	표준오차
가구원수9		3.40	1.68	44.73(100.0)	0.38
성별	남	1.65	1.04	21.8(48.7)	0.20
	여	1.74	1.11	22.93(51.3)	0.20
이환여부	Y	1.47	1.25	19.32(43.2)	0.21
	N	1.93	1.55	25.42(56.8)	0.28
2주간 활동제한여부	Y	0.15	0.43	1.98(4.4)	0.06
	N	3.24	1.68	42.75(95.6)	0.37
2주간 외래여부	Y	0.84	0.96	11.05(24.7)	0.13
	N	2.56	1.63	33.68(75.3)	0.32
상용 치료원	없음	2.29	2.11	30.12(67.3)	0.38
	병의원	0.83	1.66	10.87(24.3)	0.24
	한방병의원	0.05	0.50	0.72(1.6)	0.05
	보건소	0.01	0.21	0.16(0.4)	0.02
	약국	0.21	0.92	2.71(6.1)	0.11
	기타	0.01	0.15	0.07(0.1)	0.02
	모름	0.01	0.17	0.08(0.2)	0.02

<표 2-3> 이환과 2주간 활동제한 여부에 대한 분석

구	조사구 수	이환 여부		2주간 활동제한여부	
		평균(%)	CV(%)	평균(%)	CV(%)
종로구	60	9.7(47.6)	13.54	1.2(5.8)	51.94
중 구	60	7.5(52.2)	12.35	1.1(7.9)	44.21
용산구	60	11.1(43.7)	14.65	1.2(4.8)	57.57
성동구	60	14.4(38.7)	16.25	1.4(3.7)	66.12
광진구	60	19.7(44.2)	14.50	2.6(5.8)	52.11
동대문구	60	16.4(38.6)	16.28	4.3(10.0)	38.71
중랑구	60	20.1(39.6)	15.94	1.1(2.1)	88.35
성북구	60	22.2(41.9)	15.20	2.1(3.9)	63.73
강북구	60	19.7(48.2)	13.37	1.0(2.5)	80.17
도봉구	60	17.9(46.1)	13.95	1.5(3.9)	64.11
노원구	60	24.2(38.7)	16.25	0.5(0.8)	142.29
은평구	60	30.2(56.0)	11.43	3.9(7.1)	46.55
서대문구	60	16.6(43.3)	14.79	2.4(6.3)	49.62
마포구	60	17.4(41.0)	15.47	1.4(3.3)	70.04
양천구	60	19.1(37.0)	16.84	0.7(1.4)	106.89
강서구	60	25.6(46.9)	13.73	2.2(3.9)	63.69
구로구	60	17.0(40.8)	15.54	1.2(3.0)	74.04
금천구	60	11.2(36.6)	17.00	0.8(2.6)	79.24
영등포구	60	22.6(47.7)	13.52	1.2(2.6)	79.01
동작구	60	19.4(41.9)	15.20	2.9(6.3)	49.95
관악구	60	23.2(39.7)	15.90	3.0(5.1)	55.81
서초구	60	17.2(43.2)	14.81	2.2(5.6)	52.93
강남구	60	25.2(42.0)	15.19	2.6(4.4)	60.18
송파구	60	27.9(38.7)	16.23	3.2(4.5)	95.65
강동구	60	27.8(53.6)	12.01	3.9(7.5)	45.26

3. 새로운 표본설계

3.1. 모집단 분석

새로운 표본설계에서는 2000년 인구주택총조사 후 정리된 인구주택조사구를 모집단으로 사용하였다. 조사구는 아파트, 보통, 특수시설, 기숙사와 외국인 거주 조사구로 구분된다. 그러나 서울시의 보건지표조사에서 실제로 조사 가능한 조사구는 아파트조사구와 보통조사구이므로 조사대상을 이들로 제한한다면 전체 조사구는 54,830개이고 이 중에서 아파트조사구는 16,253개, 보통조사구는 38,577개이다. 10% 표본조사구를 제외하고 실제 조사대상이 되는 구별 조사구수, 가구수, 그리고 인구수에 대한 현황이 <표 3-1>에 있다.

3.2. 새로운 표본설계의 특징

과거의 표본설계와 비교해서 새로운 표본설계가 갖는 특징은 다음과 같다.

(1) 새로운 표본설계에서는 2000년 인구주택총조사 후 정리된 조사구를 모집단으로 하되 10% 표본조사구를 제외한 나머지 90%를 실제 추출대상 모집단조사구로 하였다.

<표 3-1> 각 구별 조사구 수 현황

구	조사구수(비율)	가구수	조사구 당 가구수	인구수	조사구 당 인구수
종로구	942(1.89)	51,322	54.5	148982	158.2
중 구	747(1.50)	41,109	55.0	118341	158.4
용산구	1,284(2.58)	69,754	54.3	204365	159.2
성동구	1,628(3.27)	93,207	57.3	287474	176.6
광진구	2,007(4.03)	110,861	55.2	336457	167.6
동대문구	1,982(3.98)	111,096	56.1	327054	165.0
중랑구	2,171(4.36)	122,148	56.3	390064	179.7
성북구	2,285(4.59)	130,219	57.0	406406	177.9
강북구	1,669(3.35)	93,920	56.3	302195	181.1
도봉구	1,658(3.33)	93,767	56.6	311917	188.1
노원구	2,835(5.70)	162,418	57.3	535972	189.1
은평구	2,276(4.57)	121,169	53.2	389806	171.3
서대문구	1,792(3.60)	100,708	56.2	310564	173.3
마포구	1,979(3.98)	109,374	55.3	324125	163.8
양천구	2,108(4.24)	123,041	58.4	405844	192.5
강서구	2,416(4.86)	139,483	57.7	445678	184.5
구로구	1,965(3.59)	111,922	57.0	346855	176.5
금천구	1,351(2.72)	79,306	58.7	238904	176.8
영등포구	2,005(4.03)	113,723	56.7	344350	171.8
동작구	2,055(4.13)	114,795	55.9	354232	172.4
관악구	2,623(5.27)	144,147	55.0	443011	168.9
서초구	1,884(3.79)	105,349	55.9	323431	171.7
강남구	2,809(5.65)	155,188	55.3	461511	164.3
송파구	3,033(6.10)	174,432	57.5	559554	184.5
강동구	2,255(4.53)	129,858	57.6	418721	185.7
합계	49,759(100.0)	2,802,316	56.3	8,735,813	175.6

(2) 구별 통계작성은 기존의 표본 가구수로는 불충분하여 제한된 경비와 조사인력 하에서 최대 25,000가구를 표본가구로 할 것을 서울시와 보사연이 결정하였다. 구별 통계의 신뢰성을 높이기 위해서 총 2,500개의 표본조사구와 조사구 당 10가구를 표본가구로 추출하였다.

(3) '97년도 조사에서는 구당 표본조사구를 60개씩 균등 배분하였으나 이번에는 1,500개의 조사

구를 각 구별로 60개씩 균등하게 배분하고 나머지 1,000개의 조사구는 최적배분하였다. 이로써 지난번보다 각 구별로 정확도가 높은 신뢰할 수 있는 조사결과를 얻을 수 있을 것으로 기대된다.

(4) '97년도 조사에서는 1개 조사구 요도에서 15~19가구가 포함되는 구역을 설정하고 여기서 13~14가구를 조사하였으나 이번에는 2000년 인구주택총조사의 조사구와 조사구 요도를 변형시키지 않고 그대로 사용하여 조사구 당 10가구를 표본가구로 선정하였다. 이는 설계당시 조사구 당 가구수가 큰 차이가 나지 않고, 지난번과 같이 15~19가구가 포함되는 구역을 설정하는 것에 대한 기준이 모호하기 때문이다.

(5) 종전에는 각 구별로 조사구를 주택특성에 따라 1차 분류한 후 60개의 표본조사구를 가구 수에 비례하는 확률계통추출하였다. 그러나 이번에는 각 구의 조사구를 동(洞)별로 정렬하고 동(洞)내에서는 아파트조사구와 보통조사구 순으로 정돈한 후에 계통추출하여 행정구역 및 조사구 형태를 반영한 대표성있는 표본을 추출하였다.

(6) 조사구내의 표본은 계통추출에 의해서 선정하고 추출된 가구가 누락되거나 조사가 불가능한 경우를 대비해서 예비표본을 선정하였다.

(7) 새로운 표본설계에서는 가중치계산과 이를 이용한 추정량은 물론 추정오차공식을 유도하고, 무응답에 대한 대체방법을 다루었다.

3.3. 표본크기의 결정 및 표본추출

조사비용과 조사원의 동원 능력을 고려하여 조사 가구수를 25,000가구로 사전협의를 통해 정하였다. 한 조사구내의 가구수가 60가구 정도로 대동소이하므로 조사여건과 효율성을 감안하여 한 조사구 당 표본가구수는 10가구로 하였다.

3.3.1. 표본배분

각 구별로 최소 표본조사구를 확보하기 위하여 먼저 1,500개의 조사구를 구별로 60개씩 균등하게 배분한 후 나머지 1,000개의 표본조사구는 이환여부, 음주여부와 흡연여부의 3가지 관심변수 각각에 대해서 최적배분법을 적용하여 이들의 평균치를 최적배분 값으로 사용하였다. 각 구별로 최종 배분된 표본조사구 현황과 3개 변수에 대한 배분결과의 변동계수가 <표 3-2>에 있다. 주어진 배분결과에서 얻어진 3개 변수에 대한 변동계수는 규모가 큰 구와 작은 구간의 차이가 크지 않다는 것을 확인할 수 있다.

3.3.2. 표본추출

구별로 표본조사구는 구의 조사구를 동(洞)별로 정렬하고 동(洞)내에서는 아파트조사구와 보통조사구 순으로 정돈한 후에 추출 간격을 정하여 계통추출하였다. 이러한 방법은 표본조사구를 지역별로 고르게 뽑을 수 있고 또한 동(洞)내에서 아파트조사구와 보통조사구가 균형있게 추출되어서 표본의 대표성이 커진다.

표본조사를 위한 대상가구는 약 60호 정도의 가구로 구성된 표본조사구로부터 10가구씩을 표본가구로 추출하였다. 이를 위해서 2,500개의 표본조사구 각각에 계통추출을 사용하였다. 한편 표본가구 중에서 표본으로 사용할 수 없는 가구가 있을 경우에는 교체표본을 사용하는 데 이를 위해

서 예비표본가구를 선정하였다. 예비표본가구는 주어진 표본조사구에서 표본가구로 선정된 가구를 제외하고 나머지 가구들을 대상으로 필요한 수의 예비표본을 표본가구 선정과 동일한 계통추출방법을 사용하여 얻는다.

<표 3-2> 표본조사구 배분과 변동계수

구	모집단크기	표본배분			변동계수		
		균일배분	최적배분	최종배분	이환여부	흡연율	음주율
종로구	942	60	20	80	11.27	15.44	8.26
중 구	747	60	16	76	10.48	14.85	8.00
용산구	1,284	60	26	86	11.85	14.79	7.78
성동구	1,628	60	33	93	12.67	14.36	7.21
광진구	2,007	60	40	100	10.92	16.07	7.30
동대문구	1,982	60	40	100	12.27	12.83	7.08
중랑구	2,171	60	44	104	11.85	13.55	6.93
성북구	2,285	60	45	105	11.20	14.11	5.92
강북구	1,669	60	34	94	10.38	13.97	8.26
도봉구	1,658	60	33	93	10.86	15.80	7.47
노원구	2,835	60	57	117	11.41	14.13	7.23
은평구	2,276	60	46	106	8.42	14.23	6.68
서대문구	1,792	60	36	96	11.41	13.83	6.76
마포구	1,979	60	40	100	11.71	13.31	6.56
양천구	2,108	60	42	102	12.61	15.15	7.19
강서구	2,416	60	48	108	10.01	15.04	6.86
구로구	1,965	60	39	99	11.81	15.41	6.87
금천구	1,351	60	27	87	13.70	14.35	7.46
영등포구	2,005	60	41	101	10.16	13.33	6.73
동작구	2,055	60	41	101	11.39	17.10	7.48
관악구	2,623	60	53	113	11.35	13.34	7.47
서초구	1,884	60	38	98	11.28	16.33	7.65
강남구	2,809	60	56	116	10.69	14.71	6.96
송파구	3,033	60	60	120	11.26	15.08	7.21
강동구	2,255	60	45	105	8.87	14.05	6.76
합계	49,759	1,500	1,000	2,500			

3.4. 추정치 및 추정오차 계산

2001년 보건지표조사 결과에 의해 작성되는 각종 통계치는 서울시 전체 모집단 또는 각 구별 모집단 평균(비율)이다.

보건지표조사에 사용된 층화이단집락추출법을 통하여 얻어진 표본조사 자료를 이용하여 모집단에 대한 특성치를 추정하고 추정오차를 구한다. 실제 표본설계에서는 최종 추출단계에서 조사구를

행정구역 및 조사구 특성에 따라 배열하고 계통추출법을 사용했지만 이를 단순임의추출한 것으로 간주하고 표본오차를 계산하기 때문에 제시된 분산 추정방법은 실제 분산을 과대추정 할 수 있다.

3.4.1. 구별 평균 추정

본 표본설계에서 각 구별로 적용한 이단집락추출법에 따른 다음의 추정식으로 h구의 평균을 추정할 수 있다.

$$\hat{\mu}_h = \frac{1}{M_h}\frac{N_h}{n_h}\sum_{i=1}^{n_h}\frac{M_{hi}}{m_{hi}}\sum_{j=1}^{m_{hi}} y_{hij} = \frac{1}{M_h}\frac{N_h}{n_h}\sum_{i=1}^{n_h} M_{hi}\,\bar{y}_{hi} \tag{1}$$

이 공식에서

M_h : h구의 총 가구수

N_h : h구의 총 조사구수

n_h : h구의 표본조사구수

M_{hi} : h구 i번째 표본조사구의 총 가구수

m_{hi} : h구 i번째 표본조사구에서 추출된 표본가구수

y_{hij} : h구 i번째 표본조사구의 j번째 가구에서 관심특성을 갖는 관측값(사람수)

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

한편 식(1)에서 M_h는 h구의 총 가구수이고, 조사시점에서 이에 대한 정확한 값을 확보하는 데 현실적으로 상당한 어려움이 예상되기 때문에 본 조사결과를 이용하여 구별 총 가구수(M_h)를 다음과 같이 추정하여 사용한다.

$$\hat{M}_h = \frac{N_h}{n_h}\sum_{i=1}^{n_h} M_{hi} \tag{2}$$

결과적으로 각 구별 평균은 다음 비(ratio)추정법으로 산출된다.

$$\hat{\mu}_h = \frac{\sum_{i=1}^{n_h}\frac{M_{hi}}{m_{hi}}\sum_{j=1}^{m_{hi}} y_{hij}}{\sum_{i=1}^{n_h} M_{hi}} = \frac{\sum_{i=1}^{n_h} M_{hi}\,\bar{y}_{hi}}{\sum_{i=1}^{n_h} M_{hi}} \tag{3}$$

3.4.2. 구별 평균 추정치에 대한 분산 추정

구별 평균 추정치에 대한 근사적인 분산 추정치는 다음과 같다(박홍래(2000), 356쪽).

$$var(\hat{\mu}_h) = \frac{(1-f_h)}{n_h}\left(\frac{1}{\bar{M}_h^2}\right)s_h^2 + \frac{f_h}{n_h^2}\sum_{i=1}^{n_h}\left(\frac{M_{hi}}{\bar{M}_h}\right)^2\frac{(1-f_{hi})}{m_{hi}}s_{hi}^2 \tag{4}$$

이 공식에서,

$$f_h = \frac{n_h}{N_h},\ f_{hi} = \frac{m_{hi}}{M_{hi}}$$

$\overline{M}_h = \frac{M_h}{N_h}$: h구에 속한 조사구들의 평균 가구수

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

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

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

$$cv(\hat{\mu}_h) = \frac{\sqrt{var(\hat{\mu}_h)}}{\hat{\mu}_h} \times 100\% \tag{5}$$

식(4)에 제시된 분산 추정치 계산에 있어서 h구의 조사구당 평균 가구수 $\overline{M}_h$가 필요한 데 이에 대한 정확한 값은 구할 수 없으므로 다음과 같이 추정하여 사용한다.

$$\hat{\overline{M}}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} M_{hi} \tag{6}$$

3.4.3. 서울시 전체 평균 추정

본 표본설계는 서울시의 각 구를 층으로 반영한 층화이단집락추출법에 해당된다. 따라서 서울시 전체 모집단에 대한 평균 추정식은 다음과 같다.

$$\begin{aligned} \hat{\mu} &= \sum_{h=1}^{L} \frac{M_h}{M} \hat{\mu}_h \\ &= \frac{1}{M} \sum_{h=1}^{L} \frac{N_h}{n_h} \sum_{i=1}^{n_h} \frac{M_{hi}}{m_{hi}} \sum_{j=1}^{m_{hi}} y_{hij} \\ &= \frac{1}{M} \sum_{h=1}^{L} \frac{N_h}{n_h} \sum_{i=1}^{n_h} M_{hi} \bar{y}_{hi} \end{aligned} \tag{7}$$

여기서 M은 서울시 총 가구수이고, 조사시점에서 이에 대한 정확한 값을 확보하는 데 현실적으로 어려움이 예상되기 때문에 본 조사결과를 이용하여 서울시 총 가구수 M을 다음과 같이 추정하여 사용한다.

$$\hat{M} = \sum_{h=1}^{L} \frac{N_h}{n_h} \sum_{i=1}^{n_h} M_{hi} \tag{8}$$

3.4.4. 서울시 전체 평균 추정치에 대한 분산 추정

서울시 전체 평균 추정치에 대한 근사적인 분산 추정치는 다음과 같다.

$$var(\hat{\mu}) = \sum_{h=1}^{L} \left(\frac{M_h}{M}\right)^2 var(\hat{\mu}_h) \\ = \frac{1}{M^2} \sum_{h=1}^{L} [N_h^2 \frac{(1-f_h)}{n_h} s_h^2 + \frac{1}{f_h} \sum_{i=1}^{n_h} M_{hi}^2 \frac{(1-f_{hi})}{m_{hi}} s_{hi}^2] \quad (9)$$

여기서 서울시의 총 가구수 M은 식(8)과 같이 추정하여 사용한다.

3.4.5. 보건의식 행태조사

보건의식행태조사의 경우 건강수준 및 의료이용실태조사를 위해 추출된 조사대상 표본가구에서 15~69세의 가족원 중 생일이 가장 빠른 사람을 조사대상으로 선정하여 조사한다. 따라서 이 경우 추정방법은 가구를 조사대상으로 하는 건강수준 및 의료이용실태 조사를 위한 추정식에서 관측결과를 나타내는 y_{hij} 대신에 $K_{hij} y_{hij}$ 를 사용한다. 여기서 K_{hij}는 조사완료된 h구 i조사구 j가구의 보건의식행태조사 대상 적격자수를 나타내며 이는 조사대상 가구에서 한 명의 보건의식행태조사 대상자를 추출하는 단계를 추정식에 반영하기 위한 확대승수에 해당된다.

4. 표본관리 및 무응답 대체

모집단으로부터 추출된 표본들 중에는 시간이 경과함에 따라 변동되는 것이 있을 수 있으므로 이에 따른 표본관리가 필요하며, 변동된 표본의 특성을 감안한 표본의 대체문제도 함께 고려해야 한다. 이번과 같이 조사 항목이 많은 경우에는 조사시에 응답자들로부터 완전한 응답을 얻는다는 것이 현실적으로 불가능하다. 무응답이 발생한 경우 무응답을 효율적으로 대체하기 위해서는 동질적인 단위들로 구성된 적절한 대체층을 구성하는 작업이 매우 중요하다. 한편 무응답 대체를 함에 있어서 많은 조사의 경우 변수들간에 논리적인 관계가 존재할 수 있기 때문에 이런 논리적인 연관성에 문제가 생기지 않는 방안을 강구하는 것이 필요하다.

4.1. 표본관리

서울시에 대한 이번의 조사는 서울시 전체의 통계는 물론 구별 통계를 산출해야 하므로 모집단과 표본을 동시에 관리할 필요가 있다. 본 조사의 경우 1차 추출단위가 조사구이고 2차 추출단위가 가구인 관계로 조사구와 가구의 변동에 세심한 주의를 기울여야 한다. 정기적으로 실시되는 보건지표조사의 경우 모집단은 시간이 경과함에 따라 변동하게 된다. 기존의 가구가 없어지거나 새로운 가구가 생기는 등의 조사구내의 변동이 생길 수 있고, 기존의 아파트나 가구들이 철거되고 새로운 아파트나 주택들이 신축되는 경우에도 조사구의 수정 및 보완은 필수적이다. 따라서 모집단의 변동을 수시로 파악하여 이들을 조사에 반영해야 한다. 모집단의 변동이 크면 모집단에 대한 새로운 정보를 추가하여 모집단을 개편하고 이에 따라 표본설계도 변경하여야 한다. 표본조사구내 가구들이 부재이거나, 표본으로 선정된 가구에 대한 조사가 불가능하여 조사단위의 결측이 생기면 1차적으로 표본조사가구를 예비표본조사가구로 교체한다. 예비표본조사가구는 초기 표본추출방법과 동일한 방법에 의해서 얻는다. 한편 표본조사구내에 새로운 가구가 생겼거나 교체된 표본으로

부터 자료를 얻지 못하는 경우에는 가중치를 조정하여 추정량과 추정오차를 구한다.

4.2. 무응답 대체

보건지표조사는 많은 변수를 포함하고 있기 때문에 조사대상이 되는 모든 변수에 대해 심층적인 분석을 통해 적절한 무응답 대체 방안을 구현하는 것은 현실적으로 불가능하다. 따라서 보사연의 보건통계 전문가와의 협의를 통해 가구조사 및 보건의식행태조사 변수 중 무응답 처리가 필요한 주요변수를 선정하였다.

보건지표조사 자료를 분석해 본 결과 가구조사 및 보건의식행태조사의 문항에 응답한 총 19,360건 중 선정된 주요변수에 대한 무응답 건수는 <표 4-1>과 같다.

<표 4-1>에 의하면 교육수준, 소득, 키, 몸무게에 상당수의 무응답이 발생하고 있다. 이러한 변수들에 대해 대규모 자료에서 변수들 간의 연관성을 규명하기 위해 사용되고 있는 대표적인 데이터 마이닝 기법인 CART(Classification and Regression Tree)기법을 활용하여 대체층을 구성하고 각 대체층에 Hot-deck 방법을 적용하여 무응답을 대체하였다.

<표 4-1> 무응답 발생 건수

	무응답건수	대체방법
교육수준	147명	hot-deck
소득	506가구	hot-deck
키	350명	hot-deck
몸무게	303명	hot-deck
음주여부	2명	최빈도대체
건강상태	2명	최빈도대체
흡연량	12명	최빈도대체
운동빈도	13명	최빈도대체
운동시간	34명	최빈도대체
건강검진	9명	최빈도대체
음주빈도	7명	최빈도대체
음주량	7명	최빈도대체
안전벨트	15명	최빈도대체

Hot-deck 방법을 적용함에 있어서 어떻게 대체층을 구성하는가 하는 것이 성공적인 대체여부를 결정하는 중요한 요소이다. Ryu et. al(2001)의 2000년 인구주택총조사에 대한 연구결과에서 밝혀진 바와 같이, 본 과제에서 적용한 CART를 활용한 대체층 구성방법은 매우 효율적임을 알 수 있다. 아울러 소득, 키, 몸무게 등의 변수를 제외한 음주여부, 건강상태 등 무응답 건수가 적은 변수들에 대해서는 CART를 활용하여 각 변수와 사회경제적 변수의 관련성을 반영한 대체층을 구성하고 각 층에서 최빈값으로 무응답을 대체하는 방법을 적용하였다. 이는 실제 평균대체법에 해당하는 것으로 여기서 다루는 변수가 범주형이기 때문에 평균 대신에 최빈값을 적용한 것이다. 최빈값 대체가 적용된 항목의 무응답률은 0.2%이하로 매우 낮아 무응답으로 인한 추정의 편향은 무시될 수 있다. 하지만 본 연구에서 무응답률이 매우 낮은 항목에 대해서도 최빈값 대체를 적용한 이유는 분석과정을 단순화하기 위해서이다. 실제로 항목별로 무응답이 모두 다르게 발생하기 때문에 무응답 보정을 위한 가중값 조정을 항목별로 하게 되면 분석과정이 매우 복잡해진다(항목별로 별도의 가중값을 적용해야 함). 따라서 각 항목별로 추출확

률에 따른 동일한 가중값을 적용하여 분석과정을 단순화하기 위해 최빈값 대체를 적용하였다.

5. 결론 및 제안

본 설계에서는 서울시 전체는 물론 25개 구별 단위의 통계를 얻기 위해서 요구되는 허용오차수준에 따른 표본크기를 구하고 층(구)에 따라 표본을 배정하였다. 그리고 모집단의 특성을 추정하는 데 사용되는 추정식을 구하고 추정오차를 계산하기 위한 추정량의 분산공식을 유도하였다. 또한 조사결과 무응답이 발생한 문항들에 대한 무응답 대체방법을 다루었다.

서울 시민의 건강상태를 파악하고 이를 근거로 적절한 건강복지정책을 마련하기 위해서 실시하는 보건지표조사의 정확성을 높이고 생산된 통계자료의 효율적인 활용을 위해서 다음 몇 가지 사항을 제안한다.

(1) 시민의 건강에 영향을 미치는 요인은 영양, 환경, 운동 등 매우 다양하다. 따라서 건강변화를 주기적(연도별이나 계절별)으로 파악하기 위해서 계절별 조사를 실시하는 것이 바람직하다.

(2) 보건지표조사를 타 지역의 유사한 조사나 국민건강·영양조사 등과 연계 분석하여 광역자치단체간 또는 전체와의 비교분석을 하고 이를 바탕으로 서울시 특성에 맞는 정책을 세우도록 한다.

(3) 많은 비용과 노력을 기울인 조사 결과에 대한 활용가치를 높이기 위해서 보다 심층적인 통계분석이 요구된다. 요인들간의 관련성이나 영향력 등을 측정해서 서울 시민의 건강증진 프로그램 개발에 적용하고, 이 결과들을 타 분야에서 활용할 수 있도록 한다.

(4) 이번의 조사는 서울시 전체는 물론 구별 통계를 작성할 수 있도록 설계하였다. 하지만 보다 정확한 구별 통계(소지역 통계)를 얻기 위해서는 직접 조사를 통한 자료뿐 아니라 기존의 각종 통계자료를 활용하는 소지역 추정기법의 적용이 필요하다.

참고문헌

[1] 김영원, 류제복, 박진우, 홍기학 공역(1998), 「표본조사의 이해와 활용」, 자유아카데미.

[2] 박홍래(2000),「통계조사론」, 영지문화사

[3] 조사통계연구회(2000),「무응답오차」, 자유아카데미.

[4] 한국통계학회(2001), 시민보건지표조사 및 건강증진 프로그램개발 표본설계 및 표본조사구 추출, 연구용역 최종보고서.

[5] 한국통계학회(2002), 시민보건지표조사 결과분석(승수산정, 무응답처리), 연구용역 최종보고서.

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

[7] 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.

[8] Ryu, Jea-Bok, Kim, Young-Won, Park, Jin-Woo, and Lee, Jae-Won.(2001), Imputation methods for the Population and Housing Census 2000 in Korea, *Bulletin of the International Statistical Institute*, 53rd Session of ISI, August 22-29, 2001, Seoul Korea, 421-422.

[2002년 7월 접수, 2002년 8월 채택]