연구노트

제4기 국민건강 · 영양조사를 위한 순환표본 설계연구

A Rolling Sampling Design for the Korea National Health and Nutrition Examination Survey

이계오* · 박진우**
Lee, Kay−O · Park, Jin−Woo

『국민건강 · 영양조사』는 건강면접 및 보건의식 행태조사, 영양조사, 검진조사로 구성되며 국민의 주관적·객관적 건강상태, 건강에 관한 의식 및 행태와 식품섭취현황 등 건강과 관련되는 제반 사항을 종합적이고 다각적으로 파악하는 것을 목적으로 실시된다. 이 조사는 1998년부터 3년 주기로 시행되고 있는데 현재 제3기 조사 통계가 공표되었으며, 2007년부터 2009년에 걸쳐 질병관리본부 주관으로 제4기 조사가 진행될 예정이다. 본 연구는 질병관리본부에서 결정한 대로 연중에 상시적으로 조사를 수행하고, 매년 국가단위 통계를 생산할 수 있도록 제4기 국민건강 · 영양조사를 위한 새로운 표본을 설계하는 것을 목적으로 한다. 본 연구의 가장 큰 특징은 국내 최초로 순환표본조사(rolling sampling survey)의 개념을 도입한 것이다. 조사여건을 고려할 때 일시에 훈련된 조사요원을 확보하는 것이 어려우므로 수준 높은 조사요원들이 상시적으로 조사를 수행할 수 있게 하도록 함으로써 조사의 품질을 담보할 수 있게 하는 순환조사 방식을 도입하는 것이 바람직하다. 이렇게 하면 측정자 간의 변동을 최소화시킬 수 있으며, 우리나라의 조사 여건을 충분히 감안한 효과적인 조사가 가능할 것으로 기대된다.

주제어: 순환센서스, 순환표본조사, 군집분석, 설계효과, 비례배분

The Korea National Health and Nutrition Examination Survey(KNHANES) consists of Health Interview Survey, Health Behaviour Survey, Nutrition Survey, and Health Examination, and is designed to produce a broad range of descriptive health and nutritional statistics for sex and age subdomains of the population. These data can be used to measure and monitor the health and nutritional status of the population of Korea. The survey has been conducted three times from 1998. The Korea Centers for Disease Control and

* 한국갤럽 자문교수.
** 교신저자(corresponding author): 수원대학교 통계정보학과 교수 박진우.
E−mail: jwpark@suwon.ac.kr

Prevention(KCDC) is preparing for the 4th survey which is to be conducted from 2007 through 2009. This study is to design a sample for the 4th survey. The main new feature of the sampling design is using a rolling sampling design method. Since KCDC has imposed some operational requirements, e.g., the needs of producing the annual national statistics and of year-round data collection by some regular staffs, a rolling sampling design method is introduced. This is the first time in history of applying a rolling sampling design for a national-wide large scale survey in Korea. Bringing in the rolling sampling, measurement variation due to different data collectors may be minimized.

Key words: rolling census, rolling sampling survey, cluster analysis, design effect, proportional allocation

Ⅰ. 서 론

『국민건강·영양조사』는 건강면접 및 보건의식 행태조사, 영양조사, 검진조사로 구성되며 국민의 주관적·객관적 건강상태, 건강에 관한 의식 및 행태와 식품섭취 현황 등 건강과 관련되는 제반 사항을 종합적이고 다각적으로 파악하는 것을 목적으로 실시되는 대규모 통계조사이다. 이 조사는 1995년에 공표된 「국민건강증진법」 제16조에 의거하여 1998년부터 3년 주기로 시행되어 오고 있는데 현재 제3기 조사 통계가 공표되었으며, 2007년부터 2009년에 걸쳐 질병관리본부 주관으로 제4기 조사가 진행될 예정이다. 본 조사를 통해 얻어진 자료는 식품섭취, 영양상태, 주요 질병 간의 연계분석 등 국민 건강 및 영양에 관련된 중요 정보를 도출하기 위한 다양한 심층 분석의 기초 자료로 활용되고 있으며, 결과적으로 국민의 질병예방, 영양개선, 건강수준 향상을 위한 보건정책 수립 및 평가에 필요한 중요한 정보를 제공한다.

기존의 제3기까지의 조사는 주기가 3년이었으므로 통계도 3년에 한 번씩 작성되어 공표되었다. 그런데 3년 간격의 통계 작성주기가 너무 긴 관계로 급변하는 국민들의 건강과 영양 관련 상황을 시의성 있게 파악하는 면에서는 일정한 한계를 지니고 있었다. 한편 검진조사에서는 신체계측, 혈압 및 맥박측정, 혈액 및 소변검사 등을 실시하게 되는데 제3기의 경우 전국에서 4명 단위

의 조사인력(검진조사팀장, 검진조사원, 보건소 간호사, 보건소 임상병리사)으로 구성된 20개 팀에 의해 약 10주 동안 조사가 수행되었다. 전국에서 20개 팀이나 되는 많은 조사팀에 의해 조사가 수행되므로 서로 다른 조사팀에 의해 생길 수 있는 측정에서의 편향 문제가 제기될 여지가 있다. 따라서 질병관리본부에서는 생활환경과 사회의식의 급격한 변화를 시의적절하게 반영하는 국가단위 통계를 매년 생산하기 위해 제4기 조사부터는 조사방법을 매년 조사로 변경하기로 결정하였다. 뿐만 아니라 검진조사와 같이 전문적인 검진조사팀을 운영해야 하는 경우 단기간에 여러 검진조사팀을 구성함으로써 발생할 수 있는 편향을 최소화하기 위해 연중 상시적으로 검진조사를 전문적으로 수행할 4개의 팀을 운영하여 조사를 수행하기로 하였다.

본 연구의 목적은 질병관리본부에서 결정한 대로 연중에 상시적으로 조사를 수행하고, 매년 국가단위 통계를 생산할 수 있도록 제4기 국민건강·영양조사를 위한 새로운 표본을 설계하는 것이다. 본 연구의 가장 큰 특징으로는 국내 최초로 순환표본조사(rolling sampling survey)의 개념을 도입한 것을 들 수 있다. 조사 여건을 고려할 때 일시에 훈련된 조사요원을 확보하는 것이 어려우므로 질적인 조사요원들이 상시적으로 조사를 수행할 수 있게 함으로써 조사의 품질을 담보할 수 있게 하는 순환조사 방식을 도입하는 것이 바람직한 것으로 생각되었다. 이렇게 하면 측정자 간의 변동을 최소화시킬 수 있으며, 우리나라의 조사 여건을 충분히 감안한 효과적인 조사가 가능할 것으로 기대된다. 기본적으로는 기존에 실시하였던 1, 2, 3기 조사 통계들과의 시계열적 연관성을 고려하면서도 대표성 있고 효율적인 표본설계를 하는 것이 목적이다. 2장에서는 순환조사의 기본개념과 배경 등을 소개하며, 3장에서는 제4기 국민건강·영양조사 표본설계의 핵심적인 내용들을 소개한다. 마지막으로 결론을 내린다.

II. 순환조사

1. 개념

인구주택총조사나 국민건강·영양조사와 같은 대규모 전국조사를 일시에

수행하기 위해서는 엄청난 자원과 노력이 필요하다. 비용도 비용이려니와 한꺼번에 수많은 조사를 감당할 수 있는 조사인력이 절대적으로 필요하다. 특히 검진조사와 같이 조사원이 어느 정도의 전문성을 지녀야 하는 경우라면 훈련된 조사원이 부족하여 일거에 전국적인 조사를 하는 것이 불가능할 수도 있다. 이런 경우를 위해 순환조사(rolling survey) 방법이 고안되었는데 순환조사의 개념은 Kish(1990)에 의해 이론적으로 체계화되었다. Kish는 순환 센서스와 순환표본조사의 정의를 다음과 같이 내린 바 있다.

정의 1: 순환 센서스(rolling census)

모집단에 대한 서로 겹치지 않는 F개의 독립적인 순환표본(rolling samples)들을 결합하는 조사를 순환 센서스라고 한다. 이때 각 순환표본은 각각 전체 모집단에 대해 추출률 $f=1/F$인 확률표본이 되도록 추출해야 한다. 따라서 F번째 주기가 지날 경우 그 동안 조사된 표본들을 모두 합하면 전체 모집단에 대한 센서스가 되는 셈이다.

정의 2: 순환표본조사(rolling sampling survey)

전국 단위의 대규모 표본조사에서 전체 표본을 서로 겹치지 않는 F개의 독립적인 순환표본으로 구성한 후 결합하는 조사를 순환 표본조사라고 한다. 이때 각 순환표본은 각각 전체 표본에 대해 추출률 $f=1/F$인 확률표본이 되도록 추출해야 한다. 따라서 F번째 주기가 지날 경우 그 동안 조사된 표본들을 모두 합하면 전체 모집단에 대한 표본조사가 되는 셈이다.

2. 배경 및 특징

대규모 전국조사를 한꺼번에 수행할 수 있는 충분한 예산을 확보하기 어려울 때가 있으며, 경우에 따라서는 예산이 확보되었다고 해도 조사를 위해 훈련된 조사원의 수가 부족하여 조사가 어려운 경우도 있다. 순환표본조사는 조사비용이나 조사인력 등의 한계로 인해 한꺼번에 전국의 표본을 일제히 조사하기 어려운 경우 유용한 조사방법이다.

총조사나 대규모 표본조사의 경우, 예산과 인력의 제약으로 인해 3년이나

5년, 또는 10년 주기로 한 번씩 이루어지는 것이 일반적이다. 이때 조사주기 사이에 끼인 연도의 정보를 얻을 수 없다는 점이 중요한 단점으로 지적된다. 순환표본조사는 전체 표본조사를 마치지 않고 부분적인 순환표본의 조사만 이루어져도 전국의 추정값을 구할 수 있다는 장점이 있으므로, 전체 조사가 끝나야만 일제히 추정하는 경우에 비해 기동성 있게 통계를 생산할 수 있게 한다. 한편, 전체 순환표본들의 조사가 완료되면 보다 자세한 영역(domain)별 통계를 생산할 수 있으므로 센서스나 대규모 일회조사의 효과도 얻을 수 있다. 뿐만 아니라 훈련된 조사인력을 지속적으로 활용할 수 있으므로 조사의 질을 담보할 수 있다는 장점도 지닌다.

미국의 National Centers for Health Statistics(NCHS)에서 실시하는 National Health and Nutrition Examination Survey(NHANES)와 National Health Interview Survey(NHIS), 그리고 미국의 통계청에서 실시하는 The American Community Survey(ACS) 등이 순환표본조사를 적용하고 있는 대표적인 사례들이다.

순환조사를 적용하고자 할 때 각별하게 주의를 기울여야 할 사항들이 있다. 매년 서로 독립적인 다른 표본에 의해 전국 추정을 하게 되는데 이때 각각의 순환표본의 고유한 특성 때문에 추정값이 서로 달라질 수가 있다. 따라서 가능한 한 각 순환표본들 간에 동질적이 될 수 있도록 하기 위해 표본설계 단계에서 세심하게 주의를 기울여야 한다. 각 조사 때마다 가능한 한 조사의 내용이 비슷하고, 표준화되도록 하는 것이 필요하다. 뿐만 아니라 조사내용에 따라 계절성 등이 반영될 수 있는데 이로 인해 교락(confounding)이 생기지 않도록 표본을 설계하고 조사일정을 수립하는 것이 필요하다.

III. 제4기 순환표본 설계

본 설계에 요구된 기본적인 사항을 요약하면 ① 전체 조사가 3년 간 균등하게 배분되어 순환표본조사 방식으로 수행되도록 설계할 것, ② 매년 전국단위의 통계를 생산할 수 있게 할 것, ③ 전국의 성별, 연령별 통계가 일정 수준

이상 정밀하게 생산되도록 할 것, ④ 3년 간의 조사결과를 종합했을 때 시·도별 추정이 가능하게 할 것, ⑤ 전국을 대상으로 4개의 조사팀에 의해 연중 조사가 이루어질 수 있도록 할 것, ⑥ 제3기 조사결과와의 시계열을 유지할 수 있도록 할 것 등이다. 이 장에서는 이러한 요구들을 반영한 표본설계 방안을 단계별로 소개한다.

1. 조사모집단과 추출단위

국민건강·영양조사의 목표모집단(target population)은 대한민국에 거주하는 모든 가구 내의 국민들이다. 그런데 현실적으로 시간이 지남에 따라 국민의 상황은 늘 변하고 있다. 따라서 본 연구에서는, 2005년 11월 1일 현재를 기준으로 한 2005 인구주택총조사 결과에 나타난 모든 가구와 국민들을 조사모집단(survey population)으로 정의한다.

〈표 1〉에는 2000년과 2005년의 성별, 연령대별 모집단 인구수 현황이 나와 있다. 2000년의 우리나라 인구수는 약 4,470만 명 가량이었는데 2005년은 약 4,700만 명으로 5년 동안 약 230만 명이 증가한 것을 알 수 있다. 그런데 연령대별 인구를 살펴보면 모든 연령대에 걸쳐 인구가 골고루 늘어난 것이 아니라 19세 이하의 인구는 줄어드는 반면 장·노년층으로 갈수록 늘어나고 있다는 사실을 발견할 수 있다. 다시 말해, 지난 5년 간 우리나라 국민들의 연령대별 인구분포에도 많은 변화가 생겼다.

본 조사를 통해 얻어지는 건강 및 영양 관련 변수들은 대부분 성별이나 연령과 연관성이 높은 변수들이므로 가급적이면 모집단의 성별, 연령별 구성을 잘 반영하는 표본을 추출하는 것이 바람직하다. 또한 지역별 통계를 생산하기 위해 지역도 함께 고려해야 한다. 본 연구가 진행될 당시 2005년 인구주택총조사의 조사구별 데이터가 공적으로 공표되지 않았기 때문에 1차 추출단위(Primary Sampling Unit: PSU)로는 동·읍·면으로 결정했으며, 2차 추출단위(Secondary Sampling Unit: SSU)는 인구주택총조사의 조사구, 3차 추출단위(Ultimate Sampling Unit: USU)는 가구로 정하였다.

〈표 1〉 성별 · 연령대별 모집단 인구수

구분	2000년			2005년		
	남자	여자	합계	남자	여자	합계
전국	22,021,287	22,689,334	44,710,621	23,465,650	23,575,784	47,041,434
연령별						
15세미만	5,073,507	4,540,850	9,614,357	4,707,847	4,278,281	8,986,128
15-19세	1,768,248	1,696,603	3,464,851	1,626,378	1,474,145	3,100,523
20-24세	1,360,955	1,752,158	3,113,113	1,915,902	1,746,221	3,662,123
25-29세	1,965,205	2,023,874	3,989,079	1,858,332	1,813,515	3,671,847
30-34세	2,034,057	2,017,078	4,051,135	2,059,913	2,036,369	4,096,282
35-39세	2,093,090	2,062,739	4,155,829	2,065,668	2,047,117	4,112,785
40-44세	2,007,571	1,960,258	3,967,829	2,082,427	2,040,614	4,123,041
45-49세	1,479,639	1,451,260	2,930,899	1,961,859	1,939,040	3,900,899
50-54세	1,174,818	1,161,674	2,336,492	1,426,597	1,428,700	2,855,297
55-59세	952,988	1,005,730	1,958,718	1,126,997	1,151,441	2,278,438
60-64세	831,814	949,684	1,781,498	897,384	991,469	1,888,853
65-69세	590,947	779,443	1,370,390	755,949	924,118	1,680,067
70-74세	346,164	566,282	912,446	514,241	738,493	1,252,734
75-79세	209,902	385,262	595,164	270,632	496,238	766,870
80-84세	93,261	206,212	299,473	136,705	295,554	432,259
85세 이상	39,121	130,227	169,348	58,819	174,469	233,288

2005년 인구주택총조사 중 인구수와 가구수 정보를 동 · 읍 · 면별로 정리한 자료를 본 표본설계를 위한 추출틀(sampling frame)로 사용하였다. 한편, 시 · 도별 1차 추출단위의 분포를 나타낸 것이 〈표 2〉에 나와 있는데 2005년 현재 우리나라의 동 · 읍 · 면의 수는 총 3,573개(동: 2,163개; 읍: 210개; 면: 1,200개)인 것으로 나타났다.

〈표 2〉 시·도별 동·읍·면 수

구분	동	읍	면	소 계
전 국	2,163	210	1,200	3,573
서 울	522	0	0	522
부 산	222	2	3	227
대 구	134	3	6	143
인 천	120	1	19	140
광 주	91	0	0	91
대 전	80	0	0	80
울 산	46	4	8	58
경 기	384	30	110	524
강 원	74	24	90	188
충 남	39	24	146	209
충 북	50	13	90	153
전 남	70	31	198	299
전 북	85	14	145	244
경 남	115	22	177	314
경 북	100	35	203	338
제 주	31	7	5	43

2. 층화

1) 층화변수

표본설계를 위한 층화변수를 선정할 때 가장 중요하게 고려해야 할 요소로는 통계의 최소 생산단위, 과거 조사와의 시계열 상의 연계성, 추정의 효율 등을 들 수 있다. 제3기 표본설계에서는 지역(대도시, 중소도시, 읍·면 지역), 거주형태(아파트, 일반주택[1]), 시·도(서울, 부산, 대구, 인천, 광주, 대전, 울산, 경기, 강원, 충청, 경상, 전라/제주)를 층화변수로 사용한 바 있다. 이러한 층화변수들은 모두 추정을 위한 영역(domain) 구분에도 그대로 사용되었다. 한편 층화변수로 고려되지는 않았지만 별도의 추정을 하는 영역변수로는

1) 여기서 일반주택이란 아파트가 아닌 단독, 연립, 다세대 등의 주택유형을 포괄하는 의미로 사용하고 있음.

성, 연령층(1-2세, 3-6세, 7-12세, 13-19세, 20-29세, 30-49세, 50-64세, 65세 이상), 교육수준, 직업, 월 평균 가구소득 등이 있다.

참고로 2005년 국민건강 · 영양조사 중 영양조사 결과의 일부를 다음의 〈표 3〉과 〈표 4〉에 수록해 놓았다(보건사회연구원 2005). 〈표 3〉은 시 · 도별 통계를 구한 예이며, 〈표 4〉는 지역, 성, 연령별 통계를 구한 예이다. 이 두 가지 유형의 통계들은 전체 국민건강 · 영양조사를 통해 작성되는 대부분의 통계 유형들을 대체로 포괄한다. 여기서 섭취비율이란 실제 추정된 섭취량을 기준량으로 나눈 값에 100을 곱한 값이다.

〈표 3〉 영양소별 영양 섭취기준에 대한 평균 섭취비율 (단위: %)

지역 / 영양소	전 국		서 울		부 산		대 구	
	평균	표준오차	평균	표준오차	평균	표준오차	평균	표준오차
에너지	98.4	0.7	97.0	0.7	101.8	1.1	102.0	1.2
단백질	196.0	1.6	168.6	1.5	173.7	2.5	168.1	1.7
칼슘	76.3	1.0	77.0	0.9	83.0	1.4	75.5	1.0
인	174.0	1.5	173.4	1.3	185.7	2.9	171.5	1.4
철	126.3	1.6	121.8	1.4	141.7	2.5	121.5	1.5
칼륨	61.1	0.6	60.2	0.5	66.2	1.0	59.4	0.5
나트륨	376.0	4.0	372.7	3.8	378.0	4.4	368.9	5.4
비타민 A	121.9	2.0	117.7	2.0	135.2	3.8	118.5	2.3
티아민	122.3	1.3	121.1	1.3	125.7	1.9	120.9	1.5
리보플라빈	95.8	1.0	96.2	0.9	101.3	1.4	95.9	1.0
나이아신	121.5	1.4	119.7	1.1	127.2	2.6	118.6	1.3
비타민 C	106.6	1.8	112.6	1.6	112.1	1.8	100.8	2.2

지역 / 영양소	인 천		광 주		대 전		울 산	
	평균	표준오차	평균	표준오차	평균	표준오차	평균	표준오차
에너지	96.1	0.8	103.1	1.5	96.8	0.9	104.0	1.3
단백질	167.6	1.7	179.2	3.6	175.7	3.2	180.1	3.2
칼슘	75.3	1.0	82.9	1.5	82.5	1.5	83.6	1.4
인	171.3	1.5	187.8	4.1	171.9	1.5	188.3	3.5
철	119.6	1.5	139.1	2.8	138.4	2.8	139.4	2.8
칼륨	59.5	0.5	66.0	1.2	65.5	1.1	66.6	1.1
나트륨	372.1	5.2	370.2	5.8	371.7	6.0	372.8	4.7
비타민 A	116.7	2.4	136.7	4.0	136.7	4.2	135.7	4.1
티아민	122.7	1.6	130.5	2.4	129.3	2.3	132.8	2.4
리보플라빈	96.2	1.0	105.0	1.6	103.4	1.7	105.8	1.6
나이아신	118.8	1.3	133.9	5.4	128.1	3.8	136.0	3.7
비타민 C	100.4	2.2	113.9	2.0	101.8	2.4	114.2	2.0

〈표 4〉 영양소 섭취부족과 에너지/지방 과잉섭취 대상자의 비율 (단위 : %)

구분		영양소 섭취부족		에너지/지방 과잉 섭취	
		비율	표준오차	비율	표준오차
전국		11.4	0.5	7.2	0.4
대도시		11.4	0.7	8.3	0.6
중소도시		11.3	0.8	7.0	0.6
읍·면지역		11.8	1.1	5.0	0.8
남자		9.4	0.6	8.2	0.5
여자		13.5	0.7	6.2	0.4
1~2세		9.0	2.7	6.1	1.6
3~6세		7.5	1.3	7.6	1.3
7~12세		6.1	1.0	8.5	1.0
13~19세		17.8	1.4	6.7	0.9
20~29세		15.0	1.4	9.7	1.1
30~49세		9.8	0.7	8.5	0.6
50~64세		10.1	1.0	4.4	0.6
65세 이상		14.7	1.5	2.3	0.7
남자	1~2세	6.1	4.3	4.0	1.5
	3~6세	7.7	1.9	9.0	2.0
	7~12세	5.2	1.0	7.1	1.3
	13~19세	19.3	2.1	6.9	1.3
	20~29세	11.3	1.8	10.5	1.6
	30~49세	6.8	0.9	10.2	1.0
	50~64세	8.6	1.2	5.4	1.0
	65세 이상	13.7	2.3	3.1	1.0
여자	1~2세	12.4	3.2	8.5	1.8
	3~6세	7.3	1.7	6.2	1.6
	7~12세	7.3	1.7	10.1	1.5
	13~19세	16.0	2.1	6.4	1.3
	20~29세	18.9	2.0	8.8	1.6
	30~49세	12.9	1.0	6.7	0.7
	50~64세	11.5	1.5	3.4	0.6
	65세 이상	15.4	1.8	1.7	0.7

위의 표들을 통해, 층화에 사용되지 않은 영역변수별 통계는 전국 단위의 통계만을 생산하므로 각각의 변수를 설계에 별도로 고려하지 않아도 최종 가구 추출단계에서 확률추출의 원리를 적용할 경우 모집단을 잘 대표하는 것으로 나타났고, 추정값의 표준오차도 만족스러운 수준을 유지하고 있음을 확인

할 수 있었다. 과거 조사와의 시계열 상의 연계성을 고려할 때 본 연구에서도 기본적으로는 과거의 층화변수들을 사용하는 것이 바람직하다고 판단된다. 따라서 제4기 표본설계에서도 제3기와 마찬가지로 기본적으로는 시·도, 지역, 거주 등의 변수를 층화변수로 결정한다.

제4기 표본설계를 위해서는 기존의 층화변수 외에 순환표본 설계방식 도입으로 인한 변화, 1차 추출단위(PSU)가 동·읍·면으로 변경된 것 등을 추가적으로 고려해야 할 필요가 생겼다. 국민건강·영양조사의 조사변수들은 대부분 성별, 연령별 특성과 밀접하게 연관을 갖는 변수라고 할 수 있으므로 지역별 인구특성까지 제대로 반영하는 것이 필요하다. 또한 본 연구에서는 추가적으로 순환조사의 특성을 반영해야 하므로 각 순환표본이 가능한 한 동질적일 수 있는 방안을 마련해야 한다. 이런 점들을 감안하여 일차추출을 위한 층화변수로는 권역과 연령대별 인구수를 사용한 후, 2차 추출단위인 조사구를 추출하는 과정에서 시·도, 지역, 거주의 층화변수를 사용하기로 한다.

2) 1차 추출단위의 층화

추출틀에는 1차 추출단위인 각 동·읍·면별, 연령대별 인구수 정보가 포함되어 있다. 따라서 1차 추출단위별 연령대별 인구비율을 변수로 활용하여 군집분석(cluster analysis)으로 지역 내에서 층화를 하기로 했다. 다변량 층화변수가 주어지는 경우 군집분석으로 층을 구분하는 방법은 널리 사용되는 층화방법이다(Golder & Yeomans 1973). 지역은 인구수를 감안하여 서울과 6대 광역시, 경기, 경상/강원, 충청, 전라/제주의 11개 지역으로 구분하였다. 이렇게 할 경우 궁극적으로 4개 권역(수도권, 경상/강원권, 충청권, 전라/제주권)으로 나누어 조사를 관리하는 데 무리가 없을 것으로 판단하였기 때문이다.

군집분석 중 Ward(1963)법은 층별 평균제곱오차(Mean Square Error: MSE)가 최소화되도록 군집들을 나누는 방법이므로 분산을 최소화하는 층화의 개념과 일치한다. 따라서 Ward 법을 사용하여 층을 나누었다. 군집분석법을 사용하여 각 지역별 층을 나눈 결과를 요약하여 정리한 것이 <표 5>에 나와 있다.

서울을 예로 들어 살펴보자. 층 1은 평균에 가까운 인구 구성을 보이는 동

〈표 5〉 지역별 층별 연령대별 평균 인구수

지 역	층	9세 이하	10~19	20~29	30~39	40~49	50~69	60세 이상	전 체	동·읍·면 수
서 울	1	1919	2219	3421	3483	2993	2219	2184	18438	436
	2	2181	3979	3358	3204	4780	2688	2023	22212	59
	3	1050	1639	5816	3022	1859	1486	1650	16521	25
	평균	1907	2391	3529	3429	3141	2237	2140	18774	
부 산	1	674	970	1529	1269	1409	1358	1607	8815	81
	2	1337	1992	2657	2230	2600	2148	2084	15048	96
	3	3162	4232	3956	4577	5313	3141	2696	27078	50
	평균	1502	2121	2541	2404	2773	2085	2049	15474	
대 구	1	992	1372	2116	1745	1772	1554	1875	11424	71
	2	2905	3666	3348	4084	4287	2348	2208	22846	72
	평균	1955	2527	2736	2923	3038	1954	2042	17175	
인 천	1	2458	3023	3119	3679	3790	1961	1987	20017	124
	2	140	174	148	189	332	376	866	2224	16
	평균	2193	2697	2779	3280	3394	1780	1859	17983	
대 전	1	1051	1403	2131	1693	1712	1404	1683	11076	31
	2	2971	3277	3850	4118	3599	2038	1994	21847	40
	3	3127	4901	3028	4003	5545	2248	1738	24590	9
	평균	2245	2733	3091	3166	3087	1816	1845	17982	
광 주	1	758	1142	1882	1166	1237	1149	1466	8801	50
	2	3637	3995	3741	4556	4034	1957	1827	23747	41
	평균	2055	2427	2720	2694	2497	1513	1629	15535	
울 산	1	3819	3849	2955	5185	4487	1948	1636	23879	22
	2	1548	2195	2378	2326	2847	1785	1354	14433	36
	평균	2410	2822	2597	3411	3469	1847	1461	18016	
경 기	1	2182	2468	3369	3481	3164	1982	2136	18783	180
	2	4264	4319	3761	5932	5221	2220	2441	28158	223
	3	500	652	740	734	933	761	1308	5626	121
	평균	2680	2836	2929	3890	3524	1801	2075	19735	
경상 /강원	1	124	171	157	167	295	384	1119	2418	277
	2	1422	1626	1633	1957	1946	1233	1617	11435	562
	평균	994	1146	1146	1366	1401	953	1453	8458	
충청	1	261	369	351	343	536	564	1326	3750	215
	2	2455	2627	2615	3203	2973	1665	2150	17689	129
	3	1147	2318	3881	1696	1569	1175	1829	13615	18
	평균	1087	1271	1333	1429	1456	987	1645	9208	
전라 /제주	1	124	165	132	137	249	341	1109	2257	129
	2	247	318	296	297	447	506	1296	3407	220
	3	1005	1300	1395	1274	1377	1161	1689	9202	168
	4	3692	3643	2911	4463	3845	1848	1887	22290	69
	평균	843	957	883	1032	1070	816	1437	7039	

들로 대부분의 동들이 여기에 포함된다. 층 2는 주로 인구가 많은 동들로 구성되는데, 10대와 40대의 인구가 유달리 많은 동들임을 알 수 있다. 반면, 층 3은 전체적인 인구수는 평균보다 적은 데 반해 20대의 인구가 두드러지게 많은 동들임을 알 수 있다. 이러한 결과를 볼 때 군집분석법을 통한 층화는 동·읍·면별 인구 특성을 적절하게 반영하는 층화라는 사실을 알 수 있다. 서울 이외의 다른 지역들에서도 층별 특성은 서울과 비슷한 양상을 보인다.

참고로 1차 추출단위인 동·읍·면의 연령대별 인구 구성비를 기초로 한 위의 층화 작업은 제3기 표본설계에서는 고려하지 않은 사항이다. 제4기 표본설계에서 이와 같은 층화를 고려하게 된 이유는 가급적이면 서로 다른 순환표본들 간의 변동을 최소화하기 위해서이다. 서로 다른 순환표본을 가지고 매년 전국적인 통계를 독립적으로 작성할 때 서로 다른 표본을 사용함으로 인해 생길 수 있는 변동을 최소화할 수 있는 장치가 필요하기 때문이다. 그런데 군집분석을 이용한 이와 같은 층화는 표본의 대표성을 높이기 위한 장치로는 사용되지만 추정 과정에서는 별도로 고려되지 않는 내재적 층화(implicit stratification) 작업이라고 할 수 있다.

3. 표본크기

1) 표본크기의 결정

표본크기는 확보 가능한 인력 및 예산 등 제반 조사여건들과 작성하고자 하는 영역(domain) 통계들의 표본오차 수준을 종합적으로 감안하여 결정되어야 한다. 제3기 표본의 경우 600개 표본조사구에 대해 조사구 당 20가구씩 약 12,000 가구의 표본크기를 가지고 있었다. 제3기 조사 데이터분석에 따르면, 기존의 표본규모는 각 변수별 전국의 성별, 연령대별 통계와 시·도별 통계를 생산하기 위해 적정한 수준으로 평가되었다. 또한 질병관리본부에서도 기존 수준의 조사를 예상하여 인력, 예산, 조사시스템을 계획하고 있기 때문에 제4기 조사의 표본크기도 기존의 표본크기 수준을 유지하는 것을 기본 방향으로 정하였다.

제4기 표본은 전체 600개의 조사구로 구성되는데, 전체 표본은 각각 200개

조사구로 이루어지는 독립적인 세 개의 순환표본으로 나누어진다. 각각의 순환표본은 독립적인 하나의 전국 표본으로 매년 전국 통계를 작성하여 공표하는 것이 가능하도록 한다. 제3기 조사 데이터를 이용하여 계산한 예상 목표오차는 상대표준오차 기준으로 대부분 5% 이내일 것으로 기대된다.

2) 조사구 내 최적 표본크기 결정

본 표본설계는 층화 다단 집락추출 방법을 사용하게 되는데, 1차 추출로 각 층에서 일정 수의 동·읍·면을 표본으로 추출하고, 2차 추출로 다시 각 동·읍·면에서 조사구를 추출하며, 3차 추출에서는 조사구 내 가구를 표본으로 추출하게 된다. 인구주택조사구는 지리적 위치를 고려하여 60가구 내외의 가구를 묶은 일종의 집락이다. 따라서 각 조사구에서 몇 가구를 표본으로 추출하는 것이 가장 바람직한지를 결정하는 것이 필요하다.

조사구 내 최적 표본크기를 정하는 데 영향을 미치는 요인으로는 조사구 내의 가구들의 속성과 조사비용을 들 수 있다. 만일 조사구 내 가구들의 속성이 동질적이라면 동일한 조사구 내에서 여러 가구를 표본으로 추출하는 것은 효율적이지 못한 반면 조사구 내 가구들의 속성이 서로 이질적이라면 여러 가구를 추출하는 것이 효과적이다. 한편, 조사구를 옮겨 다니며 조사하는 것이 동일 조사구 내 가구들을 조사하는 것에 비해 비용이 많이 든다면 한 조사구 내에서 가급적 여러 가구를 조사하는 것이 비용을 절감하는 방법이 될 것이다. 조사구 내 가구들의 속성을 평가하기 위해 표본조사 이론에서 사용하는 개념으로 설계효과(Design Effect: Deff)가 있다. 설계효과란 집락추출의 분산과 동일한 표본크기 하에서의 단순확률추출(simple random sampling)의 분산을 상대적인 비로 정의하며 설계효과가 클수록 집락추출의 효율은 떨어지는 것을 의미한다.

제3기 국민건강·영양조사에서는 조사구 당 표본크기를 20가구로 정하였다. 주로 조사여건, 조사비용 등을 고려하여 결정한 것으로 보여지는데 과연 이것이 타당한 것인지 실증적으로 검토할 필요가 있다. 각 층 내 최적의 집락크기를 결정하기 위해서는 기존의 층 내 1차 추출단위 개수와 기존의 집락 내 2차 추출단위 개수를 알아야 하는데, 제3기 조사 데이터를 이용하여 분석할 수

있다. 기존의 1차 추출단위 개수, 즉 조사구수(n)와 기존의 조사구 내 2차 추출단위 개수, 즉 가구수(m)를 알고 있으므로 가구단위 특정 조사변수 값을 이용하여 (1)식과 같이 최적 표본크기를 계산하면 된다.

$$m_{opt}=\sqrt{\frac{s_w^2}{s_b^2}\times\frac{c_1}{c_2}} \qquad (1)$$

여기서, $s_b^2=\frac{MS(b)-MS(w)}{m}$, $s_w^2=MS(w)$, $MS(b)=$ 분산분석에서 집락 간 평균제곱, $MS(w)=$분산분석에서 집락 내 평균제곱, $c_1=$ 1차 추출단위 당 조사비용, $c_2=$ 2차 추출단위 당 조사비용을 나타낸다.

본 연구에서는 국민건강 · 영양조사 건강면접조사의 가구조사 변수 중 민간의료보험 가입여부와 월 평균 가구소득 두 변수를 고려하였다. 국민건강 · 영양조사의 관심 통계 중 많은 부분이 비율 추정이므로 의료보험 가입여부라는 변수를 택했고, 관심 통계 중 가장 변동의 정도가 큰 변수로 월 평균 가구소득을 택하였다. 최적 표본크기는 각각 아파트 층과 비아파트 층에 대해 계산하였다. 일반적으로 c_1은 c_2보다 높으므로 여기서는 ($c_1=5$, $c_2=1$), ($c_1=10$, $c_2=1$), ($c_1=15$, $c_2=1$), ($c_1=20$, $c_2=1$)의 네 가지 경우를 고려해 보았다. 또한 각 지역별 설계효과(Deff)도 아울러 계산하였다.

〈표 6〉은 의료보험 가입여부 변수에 대한 계산 결과를 보여준다. 아파트나 비아파트 조사구 모두 지역별로 큰 차이를 보이고 있음을 알 수 있다. 아파트의 경우, 강원이나 인천은 Deff가 1보다 작아 동일한 조사구라고 해도 매우 이질적인 데 반해, 경기의 경우 Deff가 14를 넘어 조사구 내 가구들의 특성이 매우 동질적인 것으로 나타났다. 한편, 비아파트의 경우는 같은 지역이라고 해도 아파트 조사구와는 전혀 다른 양상을 나타낸다. 경기지역은 아파트 조사구의 Deff는 14가 넘었지만 비아파트 조사구의 Deff는 0.88로 나타난 것이 대표적인 경우이다. 이러한 결과를 볼 때 조사구 내 최적 표본크기를 결정하려면 지역별, 조사구 특성별로 모두 다르게 결정하는 것이 필요함을 알 수 있으나 실제 적용하기에는 부적절하다고 할 수 있다. 모든 경우를 포괄할 수 있는 최

〈표 6〉 의료보험 가입여부를 사용하여 계산한 최적표본크기(아파트 조사구)

지역	M_{opt}				Deff
	$C_1=5$	$C_1=10$	$C_1=15$	$C_1=20$	
서울	6.77	9.58	11.73	13.54	6.54
부산	7.17	10.14	12.42	14.35	5.83
대구	7.31	10.33	12.65	14.61	5.62
인천	23.18	32.78	40.15	46.36	0.56
광주	10.10	14.28	17.49	20.20	2.94
대전	5.45	7.71	9.44	10.90	10.10
울산	6.86	9.70	11.88	13.72	6.37
경기	4.52	6.40	7.84	9.05	14.65
강원	19.22	27.18	33.28	38.43	0.81
충청	10.99	15.54	19.03	21.98	2.48
전라	10.83	15.32	18.76	21.67	2.56
경상	8.48	11.99	14.68	16.95	4.17
제주	-	-	-	-	-

적 표본크기를 일률적으로 정하는 것은 어렵지만, 기존의 조사에서와 같이 조사구 당 20가구를 표본으로 추출하는 결정이 그다지 무리한 일은 아니라는 점은 확인할 수 있었다. 따라서 본 표본설계에서도 조사구 당 표본크기를 과거 조사와 마찬가지로 20가구로 결정하기로 한다.

3) 표본배분

앞에서도 설명한 바와 같이 표본크기는 확보 가능한 인력 및 예산 등 제반 조사여건들과 작성하고자 하는 영역통계들의 표본오차 수준을 종합적으로 감안하여 결정되어야 한다. 바로 앞부분에서 조사구 당 20가구씩을 조사하는 것이 나름의 타당성을 가지는 것으로 판단되었으므로 제4기 표본의 크기는 600개 표본 조사구에 대해 조사구 당 20가구씩 약 12,000 가구로 하여 제3기와 같게 결정하기로 한다.

표본크기가 결정되고 나면 다음으로는 표본을 층별로 배분하는 것이 필요하다. 제3기 표본설계 당시 네이먼(Neyman) 배분법의 적용 가능성을 검토한

결과 분석변수에 따라 서로 상충되는 결과가 나타났기 때문에 다양한 관심변수를 고려할 때 비례배분법을 적용하는 것이 합리적이라는 결론을 내렸고 실제 비례배분법을 기초로 표본을 배분한 바 있다(한국조사연구학회 2003). 따라서 제4기 표본설계에서는 비례배분법으로 층별 표본을 배분하기로 한다.

4. 표본추출

〈표 7〉은 각 층별 모집단 수 및 (2)식에 의해 계산된 표본 조사구 수를 나타냈다. 전국의 모집단 조사구 수는 25,393개인데, 아파트 조사구가 11,630개, 일반조사구가 13,763개이다. 반면에 표본 조사구 수는 2007, 2008, 2009년 각각 200개로 총 600개이다. 각 지역별 표본 아파트 조사구와 일반 조사구의 구성비는 모집단의 구성비를 따른다.

$$n_h = n \times \frac{N_h}{\sum_i N_i} \qquad (2)$$

본 표본설계에서 1차 추출단위는 동 · 읍 · 면인데 앞에서 작성된 표본할당 결과를 기준으로 200개 표본을 배분하고 확률비례추출법으로 동 · 읍 · 면을 추출한다.

기존의 3기 조사에서는 추출된 표본 전체에 대해 20여개의 조사팀을 투입하여 약 10주 간에 걸쳐 대대적으로 조사가 이루어졌다. 그러나 새로운 4기 조사에서는 4개의 상시적인 조사팀을 구성하여 3년 내내 조사를 진행할 계획이다. 따라서 매년 전체 표본 중 1/3에 해당하는 순환표본의 조사를 순차적으로 진행할 예정이므로 3개 연도별 표본을 별도로 구분하여 제시한 것이다.

본 연구는 표본설계 전반을 소개하는 논문이므로 연도별 표본을 제시하는데 그치고 있지만 실제 조사 수행을 위해서는 여기에 제시된 표본들을 지역별, 월별로 어떻게 배분할 것인지에 대한 치밀한 전략을 수립하는 것이 반드시 필요하다고 할 수 있다.

〈표 7〉 모집단 및 표본 조사구 수

지역	모집단			2007년 표본		
	아파트 조사구	일반 조사구	소계	아파트 조사구	일반 조사구	소계
전국	11,630	13,763	25,393	79	121	200
서울	1,359	2,830	4,189	11	23	34
부산	984	1,069	2,053	7	8	15
대구	478	575	1,053	4	5	9
인천	683	709	1,392	5	7	12
광주	689	401	1,090	4	3	7
대전	511	368	879	4	3	7
울산	356	290	646	3	2	5
경기	2,710	3,185	5,895	18	21	39
강원	349	328	677	2	5	7
충남	650	740	1,390	3	7	10
충북	561	501	1,062	2	5	7
전남	192	469	661	3	7	10
전북	441	355	796	3	5	8
경남	1,058	884	1,942	5	8	13
경북	517	790	1,307	4	9	13
제주	92	269	361	1	3	4
지역	2008년 표본			2009년 표본		
	아파트 조사구	일반 조사구	소계	아파트 조사구	일반 조사구	소계
전국	79	121	200	79	121	200
서울	11	23	34	11	23	34
부산	7	8	15	7	8	15
대구	4	5	9	4	5	9
인천	5	7	12	5	7	12
광주	4	3	7	4	3	7
대전	4	3	7	4	3	7
울산	3	2	5	3	2	5
경기	18	21	39	18	21	39
강원	2	5	7	2	5	7
충남	3	7	10	3	7	10
충북	2	5	7	2	5	7
전남	3	7	10	3	7	10
전북	3	5	8	3	5	8
경남	5	8	13	5	8	13
경북	4	9	13	4	9	13
제주	1	3	4	1	3	4

5. 추정

이 절에서는 먼저 추정에 사용되는 기호들을 모아 설명한 후, 모집단 총합 및 모평균 추정을 위한 추정량과 분산추정량을 제시한다.

1) 기호 설명

추정식을 제시하기 전에 먼저 필요한 다음의 첨자와 기호들을 설명하기로 한다.

Y : 모집단 총합

$\overline{Y}$: 모평균

h : 시 · 도, 지역별, 거주별 h번째 층($h=1,2,\cdots,H$)

i : i번째 표본 조사구

j : j번째 표본 가구

k : j번째 표본 가구 내 k번째 개인

y_{hij} : h층에 속한 i번째 표본 조사구, j번째 가구의 응답값

$y_{hi\cdot}$: h층에 속한 i번째 조사구의 총합 추정값

w_{hij} : h층에 속한 i번째 조사구 내 j번째 가구의 설계가중값

$w_{hi\cdot}$: h층에 속한 i번째 조사구의 설계가중값

N_h : h층에 속한 모집단 조사구들의 수

n_h : h층에 속한 표본 조사구들의 수

$f_h=\frac{n_h}{N_h}$: h층 표본 조사구의 추출확률

m : 표본 조사구 내 표본 가구수

2) 추정식

본 연구를 통해 얻고자 하는 대부분의 통계들은 가구(또는 개인)별 데이터에 기초하여 계산된다. 먼저, 전국과 층별 총합 추정량과 그 분산추정량, 그리고 상대표준오차의 추정량들은 다음의 식들과 같다.

$$\hat{Y} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m} w_{hij} \cdot y_{hij} = \sum_{i=1}^{H} \hat{Y}_h$$

$$\hat{V}(\hat{Y}) = \sum_{h=1}^{H} \hat{V}(\hat{Y}_h)$$

위 식에서 $\hat{Y}_h$는 h번째 층의 총합을 나타내는데 그것의 분산추정량은 다음과 같다.

$$\hat{V}(\hat{Y}_h) = \frac{n_h(1-f_h)}{n_h-1} \sum_{i=1}^{n_h} (y_{hi\cdot} - \overline{y_h})^2$$

여기서, $y_{hi\cdot} = \sum_{j=1}^{m} w_{hij} \cdot y_{hij}$ 이며, $\overline{y_h} = \frac{\sum_{i=1}^{n_h} y_{hi\cdot}}{n_h}$ 이다.

일반적으로 표본오차를 분산이나 표준오차로 표현할 때 일반인이 오차의 크기를 가늠하기 어렵기 때문에 흔히 상대표준오차(relative standard error)의 형태로 나타내기도 한다. 위 추정량의 상대표준오차의 추정식은 다음과 같다.

$$\hat{CV}(\hat{Y}) = \frac{\sqrt{\hat{V}(\hat{Y})}}{\hat{Y}} \times 100 \quad (\%)$$

참고로, 앞에서 소개한 추정량은 '07, '08, '09년 각각의 순환표본을 이용하여 전국 및 층별 추정을 하기 위한 추정식이다. 만일 3년 동안의 데이터를 모두 합쳐서 추정을 하려면 각 년도별 추정량을 다음과 같이 단순 평균하여 구하면 된다.

$$\hat{Y}^* = \frac{1}{3}(\hat{Y}_{07} + \hat{Y}_{08} + \hat{Y}_{09})$$

$$\hat{V}(\hat{Y}^*) = \frac{1}{9}[\hat{V}(\hat{Y}_{07}) + \hat{V}(\hat{Y}_{08}) + \hat{V}(\hat{Y}_{09})].$$

실제 조사 수행 후의 분석 단계에서는 여기서 제시하는 단순한 추정 외의 다양하고 복잡한 여러 문제들이 생길 것으로 예상된다. 보다 고급 통계기법을 사용하는 다양한 추정 문제, 순환조사를 사용함으로 인해 초래될 조사 시점의 불일치를 어떻게 해결할 것인가의 문제, 모집단의 변동 및 무응답을 반영하기

위한 가중값 조정 문제 등을 예로 들 수 있다. 이러한 부분들은 본 연구의 주제인 표본설계 문제와는 별도로 독립적인 연구주제로 다뤄져야 할 것이다.

IV. 맺음말

국민의 건강상태, 주요 질환의 유병률, 식품섭취와 영양상태 등 국민의 건강과 영양에 관한 종합적인 정보를 제공하는 국민건강·영양조사는 3년 주기로 조사하여 통계를 생산하였다. 하지만 생활여건과 사회의식의 변화가 급격하게 이루어지므로 시의적절하고 효과적인 국민건강 증진정책을 수립하여 시행하기 위해서 제4기 국민건강·영양조사부터 조사주기를 매년으로 조정하지만 조사인력과 소요예산을 고려하여 조사규모를 획기적으로 확대하지 않으면서 국가단위 통계를 매년 생산할 수 있는 표본설계를 연구하였다.

제3기 국민건강·영양조사와 제4기의 조사 간의 시계열 유지관리를 위해서 기존의 표본설계와 조사된 자료를 분석하였는데, 제3기 국민건강·영양조사 표본설계에서 사용한 층화변수가 적절한 것이었음을 확인할 수 있었다. 아울러 매년조사를 위한 순환표본을 마련하기 위해 각각의 순환표본들이 서로 유사하면서도 모집단에 대한 대표성을 지니도록 하기 위해 성과 인구수 변수를 활용한 내재적인 층화작업을 실시하였다. 전체 표본크기는 600조사구, 12,000가구인데, 세 개의 순환표본에 각각 200조사구, 4,000가구가 배당된다. 층별로 비례배분법을 적용하여 표본을 할당했으며, 각 층 내에서는 가구 수 크기 확률비례추출법으로 표본 동·읍·면이 추출되도록 하였다.

본 연구는 국내에서는 드물게 순환표본의 개념을 실제로 적용한 표본설계 연구로서 나름의 의의를 지닌다. 향후 본 연구를 통해 얻어진 표본조사 자료가 축적되고 나면, 한편으로는 순환표본 설계이론의 측면에서 활발한 후속 연구가 이루어질 것을 예상하며, 다른 한편으로는 조사방법의 측면에서 다양한 논의가 이루어질 수 있을 것으로 기대한다.

마지막으로 본 연구는 조사의 시작 단계인 표본설계에 국한되는 연구이다. 향후 본 연구에서 얻어진 표본을 토대로 조사 시스템을 구축하는 문제, 서로

다른 조사시점의 데이터를 결합하는 문제, 연도별 추정량의 불일치 문제, 모집단 변동이나 무응답 등으로 인해 생기는 대체 또는 가중값 조정 문제 등에 대해 별도의 추가적인 연구가 절실히 필요하다.

참고문헌

보건복지부 질병관리본부. 2006. ≪국민건강영양조사 제3기 검진조사≫.

보건사회연구원. 2005. ≪국민건강영양조사 제3기 – 성인보건의식행태–≫.

통계청. 2005. ≪인구주택총조사≫.

한국조사연구학회. 2003. ≪2004년도 국민건강영양조사 표본추출연구 결과보고서≫.

Alexander, C. H. 2002. "Still Rolling: Leslie Kish's "Rolling Samples" and the American Community Survey." *Survey Methodology* 28: 35–41.

Ezzati, T. M., Massey, J. T. , Waksberg, J. , Chu, A , and Maurer, K. R. 1992. *Sample Design: Third National Health and Nutrition Examination Survey*. Vital and Health Statistics 113.

Golder, P. A., and Yeomans, K. A. 1973. "The use of cluster analysis for stratification." *Applied Statistics* 22: 213–219.

Gunning, P., and Horgan, J. M. 2004. "A new algorithm for the construction of stratum boundaries in skewed populations." *Survey Methodology* 30: 159–166.

Jarque, C. M. 1981. "A solution to the problem of optimum stratification in multivariate sampling." *Applied Statistics* 30: 163–169.

Kish, L. 1965. *Survey Sampling*. New York: Wiley.

Kish, L. 1990. "Rolling Samples and Censuses." *Survey Methodology* 16: 63–79.

Kish, L. 2002. "New Paradigms (Models) for Probability Sampling." *Survey Methodology* 28: 31–34.

National Center for Health Statistics. *Plan and Operation of the Health and Nutrition Examination Survey*. 1988–1994. Vital and Health Statistics.

Ward, J. H. (1963). "Hierachical grouping to optimize an objective function." *Journal of the American Statistical Association* 58: 236–244.