Journal of the Korean Data & Information Science Society
2009, **20**(6), 1155–1167

한국데이터정보과학회지

경북인의 생활과 의식조사 표본설계[†]

김달호[1] · 조길호[2] · 황진섭[3] · 정경하[4]

[123]경북대학교 통계학과 · [4]경북도청 기획조정실

접수 2009년 10월 5일, 수정 2009년 11월 2일, 게재확정 2009년 11월 7일

요 약

2007년 경북인의 생활과 의식조사를 위한 표본설계를 연구하였다. 기존 조사에 대한 분석을 바탕으로 새로운 표본설계를 위한 여러 가지 사항을 검토하였다. 최근 시행된 2005년 인구주택총조사의 10% 표본조사자료를 조사모집단으로 사용하였고, 2006년 조사결과를 바탕으로 3가지 주요 항목 (경제활동상태, 연간소득수준, 주택소유)을 이용하여 표본조사구수에 대한 추정의 정도를 제시하고, 여러 가지 층별 표본 배분을 검토한 후 비례배분을 사용하여 층별로 표본을 배분하고 적절한 표본의 크기를 결정하였다. 새로운 표본설계에서는 가중치를 계산하였고 이를 이용한 추정량과 추정오차 공식을 유도하여 기존의 단순집계를 벗어나 시군별 그리고 특성별 추정과 추정의 정도에 대한 평가를 가능하게 하였다.

주요용어: 계통추출, 네이만배분, 복합표본조사, 비례배분, 집락, 층화.

1. 서 론

경상북도 도민의 사회경제적 지위와 복지수준을 측정하여 각종 도청 시책개발에 활용하고자 1997년부터 경북인의 생활과 의식조사가 실시되었다. 이 조사는 가족, 소득과 소비, 주거와 교통, 정보와 통신, 노동, 환경, 복지, 정부와 사회참여, 교육, 보건, 문화와 여가, 안전 등 12가지 부문에 대해 1년에 4개 부문씩 순환하면서 매년 조사되어 도민의 생활실태와 의식수준을 파악하여 각종 주민복지나 지역개발 정책의 방향설정과 계획수립을 위한 기초 자료로 사용된다.

2006년까지 지난 5년간 실시된 조사에서는 2000년 인구주택총조사의 10% 표본조사자료를 조사모집단으로 하였다. 그러나 기존의 모집단은 낙후되고 모집단의 변동이 생겨서 이들 모집단을 2007년에 실시할 경북인의 생활과 의식조사의 조사모집단으로 사용하기가 적절치 않기에 최근 2005년에 실시된 인구주택총조사의 10% 표본조사자료를 조사모집단으로 교체하고자 한다.

새로운 표본설계에서 표본추출방법과 표본규모를 결정하고 지역별, 특성별 표본을 배분하기 위해서 새로운 모집단의 특성을 분석하고, 표본조사구와 표본조사구로부터 표본 가구를 추출하는 방법을 제시하고자 한다. 그리고 지금까지의 조사결과 분석은 단순히 표본자료의 집계에만 그쳐 필요한 시군별 그리고 특성별 추정치와 추정치의 정도에 대한 평가가 불가능 하였다. 따라서 본 연구에서는 시군별 그리고 특성별 추정이 가능하도록 추정량과 추정량의 분산 식을 유도하고자 한다.

[†] 이 논문은 2007년도 경상북도 연구용역에 의하여 수행되었음.
[1] 교신저자: (702-701) 대구광역시 북구 산격동 1370, 경북대학교 통계학과, 교수.
E-mail: dalkim@knu.ac.kr
[2] (702-701) 대구광역시 북구 산격동 1370, 경북대학교 통계학과, 교수.
[3] (702-701) 대구광역시 북구 산격동 1370, 경북대학교 통계학과, 박사수료.
[4] (450-701) 대구광역시 북구 연암로 60, 경북도청 기획조정실, 정책기획관.

2. 현행 조사에 대한 분석

경북인의 생활과 의식조사는 1997년부터 시작되었으며 2002년 조사부터 기존 읍면동 통계담당자를 동원하던 조사방식에서 탈피하여 보다 체계적인 표본설계를 통해 조사구 규모를 618개에서 417개로 대폭 줄이고 조사원에 의한 전체 조사를 수행하는 등 실질적 통계조사가 수행되도록 크게 개선되었다. 2001년∼2003년 조사의 표본설계에서는 2000년 인구주택총조사의 전체조사구를 추출틀로 사용하였으며, 시군별로 층화하여 표본조사구를 1차계통추출한 후 추출된 조사구내에서 다시 표본가구를 2차계통추출 하였다.

그러나 2002년 태풍 매미로 인해 김천, 상주 등의 지역에서 기존 표본조사구를 유지할 수 없는 상황이 발생하여 기존 표본의 전면적인 개편 필요성이 대두되었다. 또한 2001년 이후 경북인의 생활과 의식조사는 거의 동일한 조사구와 동일한 가구를 대상으로 실시되었으므로 응답자의 응답부담을 줄이는 차원에서 표본개편이 필요하게 되었다.

2004년 개편된 표본설계는 기존의 2000년 인구주택총조사의 전체 일반가구 대신 2000년 인구주택총조사의 10% 표본자료를 사용하였다. 따라서 기존 일반가구에서 얻을 수 없는 주택특성, 산업별 종사자수 등의 조사구별 특성자료를 이용할 수 있게 됨에 따라 조사구별 모집단의 특성을 잘 반영할 수 있게 되었다.

2004년의 조사모집단은 2000년 인구주택총조사의 10% 표본조사구 중에서 섬지역, 특수시설 조사구 등을 제외한 보통조사구에 거주하는 가구와 가구원을 조사모집단으로 하였으며, 표본규모는 주어진 예산을 고려하여 결정하여, 도전체 417개 표본조사구에서 8,340가구를 조사대상 가구로 선정하였다.

표본조사구 추출을 위한 추출명부는 1차적으로 읍면지역과 동지역으로 나누어 시지역은 20개, 군지역은 13개로 구성된 총 33개의 층으로 작성하였고, 조사구의 정렬은 1차는 주택특성에 따라, 2차는 농림어업 종사율, 3차는 동지역은 광공업종사율, 읍면지역은 서비스업 종사율에 따라 정리하였다.

표본조사구 추출은 33개 층별로 정리한 표본추출명부에서 각 지역별로 설정된 표본규모수 만큼 크기에 비례하는 확률계통추출방법에 따라 표본조사구를 추출하였으며, 이런 추출방법은 2원층화비례계통추출법으로 표본조사구를 추출한 것으로 간주 할 수 있다. 표본가구는 표본으로 선정된 표본조사구내에서 20가구씩 계통적으로 추출하였으며 시군별 표본배분 현황은 표 2.1과 같다.

층화2단집락추출을 통하여 얻은 표본이 과연 조사모집단과 어느 정도 유사한지를 파악하기 위해서 1차 추출단위 (PSU: Primary Sampling Unit)인 표본조사구의 특성을 살펴보기 위해 2004년 표본설계에서 표본추출틀과 표본조사구간의 주택특성, 산업별 종사율에 대한 비교를 정리하면 표 2.2와 같다.

표 2.2에 주어진 바와 같이 표본추출틀 조사구와 표본조사구간의 주택특성과 산업별 종사율을 비교해 보면 거의 비슷함을 알 수 있다. 그러나 표본추출틀은 2000년 인구주택총조사의 10%표본자료인 반면 표본조사구의 사용은 2004년도이므로 주택특성은 포항, 구미, 경산지역의 경우 아파트의 비중이 상대적으로 더 높아졌다고 보아야 한다.

경북인의 생활과 의식조사의 표본설계 연구에 필요한 정보를 얻기 위해서 2006년도 조사에서 주요 항목을 선정하여 이들을 분석하였다. 실무적 협의를 통해 결정된 주요 항목은 경제활동상태, 연간가구소득, 주택소유 등이며, 이 중 경제활동상태는 분석단위가 가구원이고 연간 가구소득과 주택소유는 분석단위가 가구이다. 지역별 주요 항목에 대한 범주별 빈도수와 구성비율을 살펴보니, 시군별 경제활동상태에서 '취업' 범주의 구성비율은 대체로 50 80%정도에 머물고 있으며 그 중 군위군과 울릉군이 78%로 가장 높게 나타났다. 시군별 연간소득소준은 대체로 시지역이 군지역보다 높게 나타났으나 공단이나 산업단지를 가지지 못한 농업기반 시지역들은 군지역과 별로 차이가 나지 않음을 알 수 있었다. 주택소유에 대해서는 대체로 군지역이 '자가'범주의 구성비율이 높고, 몇몇 시지역과 칠곡군에서는 '전세'범주의 구성비율도 상당하였다.

표 2.1 2004년 표본설계에서 시군별 표본추출 현황

시군	표본조사구수	표본가구수	조사대상인원
전체	417	8,340	17,466
포항시	33	660	1,491
경주시	27	540	1,212
김천시	21	420	935
안동시	24	480	1,033
구미시	27	540	1,129
영주시	21	420	880
영천시	21	420	904
상주시	21	420	834
문경시	21	420	847
경산시	24	480	1,056
군위군	12	240	464
의성군	15	300	605
청송군	12	240	492
영양군	12	240	489
영덕군	15	300	596
청도군	15	300	614
고령군	12	240	412
성주군	15	300	625
칠곡군	18	360	790
예천군	15	300	572
봉화군	12	240	491
울진군	15	300	611
울릉군	9	180	384

표 2.3은 분류변수와 층화변수의 선정을 위해 지역 × 주요항목, 성별 × 주요항목, 주거형태 × 주요항목 등의 분할표를 작성하고 카이제곱 통계량을 계산하여 통계적 유의성검정을 실시한 결과로써 지역, 성별, 주거형태의 3개의 주요항목에 대해 모두 매우 유의한 차이를 보이고 있다.

3. 새로운 표본설계

3.1. 새로운 표본설계의 특징

2005년 인구주택총조사의 10% 표본조사구를 1차 추출단위 (PSU)로 사용하므로 표본조사구의 크기는 60가구를 기준으로 종전과 같은 수준으로 유지하며, 조사결과의 시계열을 유지하기 위해서 표본조사구 수를 종전과 비슷하게 425개로 유지하였다.

조사에서 표본가구는 종전과 같이 표본 조사구내 가구 중에서 평균 1/3가구에 해당하는 20가구를 추출하는 방법을 그대로 유지하였으며, 2005년 인구주택총조사 이후의 변화된 모집단의 특성은 사후추정방법을 사용하여 조정해 준다. 종전과 같이 경북 전체와 시군별, 그리고 주요 특성별 통계생산에 중점을 두었다.

층별 표본조사구 배분은 비례배분과 네이만 최적배경을 검토한다. 표본조사구 추출시 조사구의 특성을 충분히 반영하기 위해 모집단조사구를 주택특성과 산업특성 그리고 행정구역에 따라 정렬한 후 각 지역별로 설정된 표본규모수 크기에 비례하는 확률비례계통추출법을 사용하였으며, 추출된 조사구목록을 작성하고 조사구 요도를 확보하며, 표본조사구내의 가구목록을 작성해서 조사를 수행하기 전에 조사구와 가구의 변동을 반영하면서, 추가로 예비표본조사구를 추출한다.

표 2.2 2004년 표본설계에서 시군별 표본추출틀과 표본조사구의 특성 비교
(표본추출틀비율/표본비율)

구분	주택특성			산업별 특성		
	단독	아파트	연립	농림어업	광공업	서비스업
포항시	39.7/40.7	44.1/49.4	6.8/4.5	5.1/5.5	9.1/9.0	85.8/85.4
경주시	60.3/59.3	29.9/35.9	1.0/0.0	10.7/8.8	8.1/7.9	81.1/83.3
김천시	63.8/67.0	27.8/25.1	2.0/4.4	22.3/23.2	7.0/7.2	70.7/69.7
안동시	64.0/66.0	26.7/25.0	3.1/1.3	18.3/18.5	2.0/1.9	79.7/79.6
구미시	35.6/31.0	52.2/55.4	3.4/2.8	4.2/2.6	18.6/18.9	77.2/78.5
영주시	63.4/67.5	22.2/19.1	6.1/5.1	16.4/17.0	3.4/4.0	80.2/78.5
영천시	65.7/62.7	25.5/27.1	0.0/0.0	22.6/22.1	8.5/8.3	68.9/69.6
상주시	80.5/77.0	13.6/14.6	1.8/4.7	34.4/31.1	2.0/2.1	63.1/66.8
문경시	74.5/71.3	15.9/21.0	1.9/1.8	19.8/20.5	3.2/3.4	77.0/76.1
경산시	47.7/45.1	46.2/44.9	1.4/3.1	7.7/7.2	9.6/9.6	82.7/83.2
군위군	86.4/90.1	0.0/0.0	2.0/1.4	43.3/43.3	3.6/3.2	53.1/53.6
의성군	88.4/91.9	0.0/0.0	3.2/1.3	44.0/44.8	2.7/2.2	52.8/53.0
청송군	74.8/74.5	7.0/7.5	1.6/3.5	35.1/34.7	1.4/2.0	63.5/63.2
영양군	86.8/86.6	1.6/3.3	2.5/4.2	39.0/42.4	1.8/1.7	59.2/55.8
영덕군	82.7/84.2	2.1/4.2	3.5/4.9	21.3/23.2	4.3/2.8	74.5/74.0
청도군	84.0/81.0	7.7/6.0	1.0/3.0	36.6/35.8	3.5/3.7	59.9/60.4
고령군	83.1/72.4	12.9/16.9	0.5/1.4	30.9/32.0	9.0/7.9	60.1/60.0
성주군	88.7/90.2	0.0/0.0	1.6/1.4	42.3/39.6	3.8/5.0	53.9/55.4
칠곡군	47.0/41.7	43.7/50.1	0.3/0.0	9.0/9.4	19.0/19.3	72.0/71.3
예천군	85.2/85.2	5.4/4.2	1.0/2.2	41.7/43.6	1.7/1.2	56.6/55.3
봉화군	82.0/75.8	6.1/9.0	0.0/0.0	39.7/37.4	3.7/1.8	56.6/60.7
울진군	73.7/75.6	13.3/13.6	3.1/3.1	19.9/18.1	4.1/4.0	76.0/77.9
울릉군	71.8/73.4	0.0/0.0	4.4/0.4	13.8/12.8	5.1/4.8	81.1/82.4

표 2.3 분류/층화변수×주요항목별 유의성 검정
(카이제곱값/유의확률)

분류/층화변수	주요항목		
	연간가구소득	주택소유	경제활동
읍면동	288.57(<0.0001)	413.69(<0.0001)	611.41(<0.0001)
성별	1,866.61(<0.0001)	80.98(<0.0001)	2,013.13(<0.0001)
주택형태	1,193.52(<0.0001)	877.82(<0.0001)	306.57(<0.0001)

과거에는 추정의 문제를 다루지 않았지만, 새로운 표본설계에서는 가중치 계산과 이를 이용한 추정량은 물론 추정오차 공식도 유도한다.

3.2. 모집단 분석

새로운 표본설계에서 사용할 모집단조사구는 2005년도에 실시한 인구주택총조사의 10% 표본조사구에서 통계청에서 실시하고 있는 각종 조사 (경제활동인구조사, 사회통계 등)에 사용되고 있는 조사구를 제외한 2,439개로 구성되었다. 이는 통계청에서 사용하고 있는 표본조사구가 중복될 경우 조사 수행의 어려움을 감안한 것이다.

아파트와 일반 조사구를 제외한 조사구에서의 조사가 현실적으로 어려우므로 본 표본설계에서는 1,765개의 아파트조사구와 일반조사구만을 모집단조사구로 사용한다. 표 3.1에 의하면 아파트조사구와 일반조사구가 전체의 72.4%를 차지하고 있으나, 가구수와 인구수는 각각 전체의 99.3%와 78.0%를 차지하고 있다.

표 3.1 조사구 종류별 분포

10% 표본	합 계	아파트	일 반	기숙시설	특수사회시설	기 타
조사구수(%)	2,439(100.0)	447(18.3)	1,318(54.1)	506(20.7)	156(6.4)	12(0.5)
가구수(%)	103,120(100.0)	26,824(26.0)	75,563(73.3)	526(0.5)	195(0.2)	12(0.0)
인구수(%)	329,256(100.0)	81,958(24.9)	174,913(53.1)	59,042(17.9)	13,230(4.0)	113(0.1)

표 3.2 조사모집단의 거처별 가구수 현황

조사모집단	합 계	아파트	단독주택	연립/다세대주택	기 타
가구수(%)	102,387(100.0)	27,224(26.6)	67,624(66.0)	4,416(4.3)	3,123(3.1)

표 3.2의 조사모집단의 거처종류별 가구수 현황을 살펴보면 단독가구가 전체의 66.0%인 67,624가구이고, 아파트가 전체의 26.6%인 27,224가구를 차지하고 있어 단독과 아파트가 전체가구의 92.6%가 된다.

3.3. 층화

새로운 표본설계에서 사용할 모집단조사구는 2005년 인구주택총조사의 10% 표본조사 자료이다. 본 조사를 통해 경상북도 전체와 지역별, 그리고 특성별 통계를 얻고자 한다. 따라서 적절한 층화변수를 설정하여 추정치의 생성은 물론 표본크기를 정하고 표본배분에 반영해야 한다.

2006년도의 조사자료를 바탕으로 한 카이제곱 분석에 의하면 3개의 주요항목과의 독립성 검정에 대해 유의성이 있는 변수로는 성별, 지역 (동부와 읍·면부), 그리고 주거형태가 있다. 이들을 층화변수로 사용하기 위해서는 모집단조사구의 정보가 이들을 어느 정도 반영할 수 있는지를 파악해야 한다. 또한 여러 변수를 사용하여 층화를 하게 되면 층의 수가 많게 되어 표본수가 작을 경우에는 상당수의 층에 표본이 배분되지 않는 문제가 발생한다.

2007년 표본설계에서는 조사항목에 대한 시군별 지역통계 작성을 염두해 두어 행정구역을 1차적인 층화 기준으로 하고자 한다. 이런 관점에서 울릉군과 같은 작은 지역의 표본조사구 배분에 있어서 최소한의 표본 조사구수를 확보할 수 있도록 지역별로 표본조사구수를 배분하여야 한다. 전체적으로 경상북도를 10개 시와 13개 군으로 구분하여 23개 지역 층을 구성한 후, 다시 10개 시지역층을 동지역과 읍·면부지역으로 나누어 결과적으로 전체 조사구를 33개 층으로 구분하고자 한다.

주거형태 변수는 조사구 추출시 모집단 특성을 반영하기 위한 분류변수로 사용하고, 성별 변수는 조사 후 사후층화방법을 적용해서 추정치의 보정에 반영하는 것이 바람직하다.

3.4. 표본크기 결정 및 배분

표본규모와 표본조사구수의 결정은 표본의 대표성과 추정의 정도에 영향을 주게 된다. 새로운 표본설계에서 표본크기는 확보된 예산과 조사업무 관리 및 조사일정을 고려하는 동시에 추정값의 신뢰수준을 감안하여 결정한다.

2006년도 조사결과에서 3개 주요항목 중 주택소유 변수는 CV가 상당히 크나 경제활동상태 변수는 CV가 아주 작으며 연간가구소득 변수는 중간 정도의 CV를 가지므로 안정적이다. 통상적으로 CV가 큰 것을 사용할 경우 표본규모가 너무 크게 되어 주어진 조사환경 (비용과 시간 및 관리 등)에 적용할 수가 없게 된다. 따라서 표본 크기 결정에 CV면에서 다소 차이가 나는 변수인 연간가구소득과 경제활동상태 변수를 사용하여 그 결과를 살펴보고자 한다.표본규모를 결정하기 위해서 비례배분과 네이만배분

을 고려하여 표본크기를 계산하였다.

$$\text{비례배분} : n = \frac{N \sum N_h P_h Q_h}{N^2 V' + \sum N_h P_h Q_h} \tag{3.1}$$

$$\text{네이만배분} : n = \frac{(\sum N_h \sqrt{P_h Q_h})^2}{N^2 V' + \sum N_h P_h Q_h} \tag{3.2}$$

여기서 V'는 허용오차로 $V' = (C' \sum W_h P_h)^2$이고 C'는 경북전체의 통계를 생산하는데 사용할 변동계수로 일정한 값을 갖으며, P_h 대신에 2006년도 자료를 이용한 추정치를 사용한다. 그리고 N과 N_h는 조사모집단 자료를 사용한다.

표 3.3 연간가구소득에 대한 표본규모 (가구수)

C.V.(%)	비례배분						
	500만원미만	5백~1천만원미만	1천~2천만원미만	2천~3천만원미만	3천~4천만원미만	4천~5천만원미만	5천만원이상
1	29,069	35,219	29,419	30,440	38,721	57,636	56,741
2	**9,233**	**11,866**	**9,375**	**9,794**	**13,513**	24,937	24,275
3	**4,320**	**5,637**	**4,390**	**4,597**	**6,481**	**12,818**	**12,425**
4	2,476	3,249	2,517	2,638	3,749	**7,628**	**7,381**
5	1,598	2,103	1,625	1,704	2,432	5,016	4,850
6	1,115	1,470	1,134	1,189	1,701	3,536	3,417
7	822	1,084	836	876	1,255	2,622	2,533
8	630	832	641	672	964	2,020	1,951
9	499	659	507	532	763	1,602	1,548
10	404	534	411	431	619	1,302	1,257
11	334	442	340	357	512	1,078	1,041
12	281	371	286	300	431	908	876
13	240	317	244	256	367	774	748
14	207	273	210	221	317	668	645
15	180	238	183	192	276	583	563

C.V.(%)	네이만배분						
	500만원미만	5백~1천만원미만	1천~2천만원미만	2천~3천만원미만	3천~4천만원미만	4천~5천만원미만	5천만원이상
1	28,103	34,627	29,305	30,034	37,755	54,680	51,547
2	**8,926**	**11,667**	**9,339**	**9,663**	**13,176**	23,658	22,053
3	**4,177**	**5,542**	**4,373**	**4,536**	**6,319**	**12,160**	**11,288**
4	2,394	3,194	2,507	2,603	3,656	**7,236**	**6,706**
5	1,545	2,068	1,619	1,681	2,371	4,759	4,406
6	1,078	1,445	1,130	1,173	1,659	3,355	3,105
7	794	1,066	832	865	1,224	2,488	2,301
8	609	818	638	663	940	1,916	1,772
9	482	647	505	525	744	1,520	1,406
10	391	525	410	426	603	1,235	1,142
11	323	434	339	352	499	1,023	946
12	272	365	285	296	420	861	796
13	232	311	243	252	358	735	679
14	200	269	209	218	309	634	586
15	174	234	182	190	269	553	511

표 3.3은 CV의 크기에 따라 연간가구소득의 범주별로 비례배분과 네이만배분을 사용한 경우의 가구표본의 크기이며, 표 3.5는 경제활동상태의 범주별로 비례배분과 네이만배분을 사용한 경우의 가구원의 크기이다.

표 3.4 연간가구소득에 대한 예상 CV(%)

	500만원미만	5백∼1천만원미만	1천∼2천만원미만	2천∼3천만원미만	3천∼4천만원미만	4천∼5천만원미만	5천만원이상
예상 CV	2.09	2.41	2.11	2.16	2.59	3.77	3.71
예상 CV	2.05	2.38	2.11	2.15	2.56	3.66	3.52

표 3.5 경제활동상태에 대한 표본규모 (인구수)

C.V.(%)	비례배분					
	취업	구직	가사	통학	연로/장기질병	기타
1	6,128	148,785	**56,980**	81,813	68,169	85,447
2	1,560	65,766	**17,088**	**26,872**	**21,277**	**28,463**
3	696	**34,076**	7,886	**12,680**	**9,913**	**13,480**
4	392	**20,349**	4,496	7,290	5,672	7,761
5	251	13,406	2,896	4,714	3,659	5,021
6	174	9,460	2,018	3,292	2,552	3,508
7	128	7,019	1,486	2,427	1,880	2,587
8	98	5,409	1,139	1,862	1,442	1,985
9	77	4,292	901	1,474	1,141	1,571
10	63	3,488	730	1,195	925	1,274
11	52	2,889	604	988	765	1,054
12	44	2,432	507	831	643	886
13	37	2,075	433	708	548	755
14	32	1,791	373	611	473	652
15	28	1,562	325	532	412	568

C.V.(%)	네이만배분					
	취업	구직	가사	통학	연로/장기질병	기타
1	6,119	133,920	**55,426**	78,331	**65,923**	84,762
2	1,558	59,195	**16,622**	**25,729**	**20,576**	**28,235**
3	695	**30,672**	7,671	**12,141**	9,586	**13,372**
4	391	**18,316**	4,374	6,980	5,485	7,698
5	251	12,066	2,817	4,513	3,538	4,981
6	174	8,515	1,963	3,152	2,468	3,480
7	128	6,318	1,445	2,324	1,818	2,566
8	98	4,868	1,108	1,783	1,394	1,969
9	77	3,864	876	1,411	1,103	1,558
10	63	3,139	710	1,144	894	1,264
11	52	2,601	587	946	739	1,045
12	44	2,189	494	796	622	879
13	37	1,868	421	678	530	749
14	32	1,612	363	585	457	646
15	28	1,406	316	510	398	563

표 3.6 경제활동상태에 대한 예상 CV (%)

	취업	구직	가사	통학	연로/장기질병	기타
예상 CV	0.52	3.91	1.78	2.28	2.00	2.35
예상 CV	0.52	3.69	1.75	2.22	1.97	2.34

조사모집단 분석에 의하면, 조사구당 평균가구수는 58이고 인구수는 2.5명이었으며, 조사여건을 고려해서 경상북도와 협의를 거쳐 적절한 표본가구 규모를 8,500가구로 결정하였다. 따라서 425개의 조사구를 표본조사구로 하고, 각 조사구에서 20가구 (평균적으로 조사구내 가구 중 1/3에 해당하는 가구

수)를 표본가구로 결정하였다. 이에 따라 표본의 크기는 최종 8,500가구 (가구원 21,250명)로 정하였다. 표 3.4는 비례배분과 네이만배분을 사용한 경우 표본가구 크기가 8,500일 때의 각변수에 대한 예상 CV이며, 표 3.6은 비례배분과 네이만배분을 사용한 경우 표본가구원 크기가 21,250일 때 경제활동상태에 대한 예상 CV이다.

한편, 425개의 표본조사구를 배분하는 방법으로 비례배분과 네이만배분을 고려하였다. 배분공식은 일반적으로 다음과 같은 공식을 사용한다. 만약 주요항목이 비율 변수이면 $S_h = \sqrt{P_h Q_h}$로 주어진다.

$$\text{비례배분} : n_h = n \cdot \frac{N_h}{N} \tag{3.3}$$

$$\text{네이만배분} : n_h = n \cdot \frac{N_h S_h}{\sum_h N_h S_h} \tag{3.4}$$

여기서 N_h는 h층의 모집단조사구 수이고 S_h는 전년도 조사의 주요 항목별로 계산한 표준편차이고, n_h는 배분된 층별 조사구를 나타낸다.

비례배분은 층별 조사구수를 기준으로 배분하므로 포항, 구미, 경산 등과 같은 조사구가 많은 곳에 표본조사구가 많이 배분되는 반면, 울릉군 등 작은 농어촌 군지역은 표본 조사구가 너무 적게 배분되어 추정값의 신뢰성에 문제가 발생할 소지가 있다. 따라서 이를 방지하기 위해 최저조사구수를 확보하여 배분하는 것을 고려해야 한다.

한편, 3개의 주요 조사항목에 대해 계산한 네이만 배분의 경우는 주요 3개 조사항목에 대한 네이만 배분의 절충형 (평균)을 적용하여 표본조사구를 구해본 결과 비례배분보다 더 심하게 포항, 구미, 경산 등 조사구수가 많은 시지역에 표본조사구가 많이 배분되는 것으로 나타났다. 이러한 문제점을 개선하게 위해 크기의 제곱근 혹은 세제곱근에 비례해서 배분하는 방법을 고려하였다.

추정의 정도 (신뢰도) 측면에서 필요한 최저조사구수를 확보하는 관점에서 보면 층별 조사구수의 세제곱근에 비례하게 표본조사구를 배분하는 방법이 가장 바람직한 것으로 사료된다. 그리고 세제곱근에 비례하게 표본조사구를 배분한 결과는 이전에 이미 시행했던 2004년도 표본조사구의 배분과 유사하므로 조사결과의 시계열적 안정성 측면에서 바람직하다.

층별 조사구수의 세제곱근에 비례하게 표본조사구를 배분하는 방법을 선택하여 최종적으로 층별, 동부와 읍·면부에 배분된 표본조사구는 표 3.7에 주어져 있다.

3.5. 표본 추출

23개 층 (시와 군)으로부터 표본조사구를 추출할 때 층내를 동부와 읍·면부로 나누어 계통추출법을 적용한다. 계통추출을 위해서 조사구를 주택특성과 농림어업 종사율, 광어업 종사율, 서비스업 종사율, 행정구역번호와 조사구번호에 따라 분류하여 정렬한다. 33개 층별로 분류지표에 따라 정리한 표본추출명부에서 각 지역별로 설정된 표본규모수 크기에 비례하는 확률계통추출방법에 따라 표본조사구를 추출한다. 이러한 추출방법은 표본조사구가 모집단조사구의 특성을 가장 닮을 뿐만 아니라 행정구역별로 고르게 추출되어서 표본의 대표성이 커진다.

표본조사를 위한 조사대상 가구는 표본조사구로부터 추출한다. 조사모집단 자료에 의하면 조사구는 평균 58가구로 구성되어 있다. 이번 조사에서는 표본조사구에 있는 가구들로부터 20가구를 표본가구로 추출하므로, 이를 위해 표본조사구에 대한 가구목록을 사용하는데, 통계청에서 제공한 가구목록을 사용하기 전에 조사팀은 표본조사구를 사전에 방문해서 가구목록과 실제 거주하고 있는 가구들을 확인하고, 목록에 누락되었거나 추가되는 가구를 파악해서 가구목록을 새로이 작성하여 실제 조사에 사용한다. 표본가구추출은 425개의 표본조사구 각각에서 계통추출에 의해 표본조사 가구를 추출하며, 표본가구 중에

표 3.7 동부와 읍·면부의 표본조사구

	합계	동부	읍·면부
경북도	425	138	287
포항시	35	19	16
경주시	29	15	14
김천시	26	12	14
안동시	27	14	13
구미시	31	19	12
영주시	24	12	12
영천시	24	11	13
상주시	25	11	14
문경시	21	10	11
경산시	29	15	14
군위군	11	0	11
의성군	15	0	15
청송군	11	0	11
영양군	10	0	10
영덕군	12	0	12
청도군	12	0	12
고령군	11	0	11
성주군	12	0	12
칠곡군	14	0	14
예천군	13	0	13
봉화군	12	0	12
울진군	13	0	13
울릉군	8	0	8

서 표본으로 사용할 수 없는 가구가 발생하는 경우에는 교체표본을 사용하는데 이를 위해서 예비표본가구를 선정한다.

4. 추정

본 조사의 표본추출방법인 층화이단집락추출법을 통하여 얻어진 표본조사자료를 이용하여 모집단에 대한 추정치 (시군별, 특성별)을 산출하는 방법을 제시하고자 한다. 일반적으로 본 표본설계와 같은 복합표본조사 (complex sample survey)의 경우 최종추출단위 (가구 또는 가구원)에 대해 적절한 가중치를 산출하여 이를 적용한 추정방법을 사용한다.

본 설계에서 사용된 조사모집단은 2005년 인구주택 총조사의 10% 표본자료를 사용한 관계로 시군별 (층별) 추정에 있어서는 가중치를 고려해야 한다. 가중치는 가구수 기준과 인구수 기준의 두 경우를 사용할 수 있다.

실제 조사를 마치고 얻은 자료들에는 무응답, 부재, 응답 거절 등의 이유로 결측치가 생긴다. 또한 조사 결과의 특성치별 (성별, 도시·농촌별 등)비율이 모집단의 비율을 따르지 않는 경우 가중치를 조정하여 추정에 반영한다.

4.1. 시군별 추정

본 연구에서 제시된 평균 추정식 및 표본오차와 관련된 모든 계산은 SAS에서 제공하는 PROC SURVEYMENAS 모듈을 통해 간편하게 수행할 수 있다. 추정식을 이용하기 위해서는 각 관측값에 대한 적절한 가중값 산출이 필요하며, 가중값은 조사 (추출)단위가 가구인 경우와 조사 (추출)단위가 가구원인

경우를 구분할 필요가 있다. 조사단위가 가구인 경우 시군별 평균추정량은 다음 식과 같다.

$$\bar{y}_h = \frac{\sum_{i=1}^{n_h}\sum_{j=1}^{m} w_{hij} y_{hij}}{\sum_{i=1}^{n_h}\sum_{j=1}^{m} w_{hij}} \tag{4.1}$$

여기서,

y_{hij} : h 층 i 조사구 j 번째 표본 가구에 대한 관측값
w_{hij} : h 층 i 조사구 j 번째 표본 가구 (가구원)에 대한 가중값

$$w_{hij} = W_h \cdot \frac{N_h}{n_h} \cdot \frac{M_{hi}}{m_{hi}} \tag{4.2}$$

N_h : h 층의 전체 조사구 수
n_h : h 층의 표본 조사구 수
M_{hi} : h 층 i 번째 조사구의 전체 가구 수
m_{hi} : h 층 i 번째 조사구의 표본 가구 수 ($m_{hi} = m = 20$)
W_h : h 층에 대한 가중값 (모집단 확대 인자)

이다. 참고로 W_h는 가구기준 가중값과 인구기준 가중값을 생각할 수 있는데, 최종단위가 가구이면 가구기준 가중값을 사용하고 최종단위가 가구원이면 인구기준 가중값을 사용한다.

여기서 h층의 i번째 표본조사구의 j번째 대상이 어떤 특성을 가지면, $y_{hij} = 1$, 그 외에는 0이라 두면, 모비율의 추정치로도 사용할 수 있다. 경북인의 생활과 의식조사에서 대부분의 조사항목이 범주형으로 주어져 있으며 이 경우 각 범주별 구성비율의 추정이 관심사이다. 만약 범주형 변수 C의 범주가 q개 있으면 이를 $c_1, c_2, \cdots, c_q$로 표시하면, 이 경우 관측값 y_{hij}는 k번째 범주에 해당하는 값으로 나타내면 다음과 같이 주어진다.

$$y_{hij}^{(k)} = I_{C=c_k}(h,i,j) = \begin{cases} 1, & \text{if } C_{hij} = c_k \\ 0, & \text{otherwise} \end{cases} \tag{4.3}$$

구성비의 추정치를 계산하고자 한다면 y_{hij} 대신 $y_{hij}^{(k)}$를 사용하면 된다.

시군별 표본평균에 대한 분산추정량은 다음과 같이 주어진다.

$$\widehat{V}(\bar{y}_h) = \frac{\dfrac{n_h}{n_h - 1}(1 - f_h)\sum_{i=1}^{n_h}\left[W_{hi}(\bar{y}_{hi} - \bar{y}) - \dfrac{1}{n_h}\sum_{s=1}^{n_h} W_{hs}(\bar{y}_{hs} - \bar{y})\right]^2}{(\sum_{i=1}^{n_h} W_{hi})^2} \tag{4.4}$$

여기서, $W_{hi} = \sum_{j=1}^{m_{hi}} w_{hij}$, $\bar{y}_{hi} = (\sum_{j=1}^{m_{hi}} w_{hij} y_{hij})/(\sum_{j=1}^{m_{hi}} w_{hij})$이다.

평균추정량에 대한 표준오차 (SE)와 상대표준오차를 나타내는 변동계수 (CV)는 각각 다음과 같이 계산된다.

$$\widehat{SE}(\bar{y}_h) = \sqrt{\widehat{V}(\bar{y}_h)}, \quad \widehat{CV}(\bar{y}_h) = \frac{\sqrt{\widehat{V}(\bar{y}_h)}}{\bar{y}} \times 100(\%) \tag{4.5}$$

4.2. 경북전체 추정

경북도 전체평균 $\bar{y}$에 대한 추정치와 분산은 다음과 같이 계산된다.

$$\bar{y} = \frac{\sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{m} w_{hij}y_{hij}}{\sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{m} w_{hij}} \tag{4.6}$$

$$\widehat{V}(\bar{y}) = \frac{\sum_{h=1}^{L}\frac{n_h}{n_h-1}(1-f_h)\sum_{i=1}^{n_h}\left[W_{hi}(\bar{y}_{hi}-\bar{y}) - \frac{1}{n_h}\sum_{s=1}^{n_h} W_{hs}(\bar{y}_{hs}-\bar{y})\right]^2}{(\sum_{h=1}^{L}\sum_{i=1}^{n_h} W_{hi})^2} \tag{4.7}$$

한편, 가구원대상 조사변수에 대한 추정은 다음과 같다. 개인 가구구성원에 대해 조사된 각종 변수에 대한 평균추정을 위한 가중값은 다음과 같이 산정된다. 먼저 결측값이 전혀 없다고 가정하는 경우 표본가구내의 조사대상자를 모두 표본으로 추출하기 때문에 동일 가구내의 모든 개인별 관측값 y_{hijk}에 대한 가중값으로는 앞에서 사용한 w_{hij}를 반복 적용하면 된다.

$$w_{hijk} = W_h \cdot \frac{N_h}{n_h} \cdot \frac{M_{hi}}{m_{hi}} \tag{4.8}$$

여기서 W_h는 최종단위가 가구원이므로 인구기준 가중값을 사용한다.

이 경우 앞에서 주어진 시군구별 표본가중평균은 다음과 같이 표현된다.

$$\bar{y}_h = \frac{\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}}\sum_{k}^{K_{hij}} w_{hijk}y_{hijk}}{\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}}\sum_{k}^{K_{hij}} w_{hijk}} \tag{4.9}$$

표본분산 공식은 앞에서 y_{hij} 대신 $y^*_{hij} = \sum_{k}^{K_{hij}} y_{hijk}$를 사용하여 구하면 된다. 경북 전체 추정식도 마찬가지로 적용하면 된다.

5. 결론

2007년도 경북인의 생활과 의식조사를 위한 새로운 표본 설계는 2000년도의 낙후되고 변동된 조사모집단을 2005년에 실시된 인구주택총조사의 10% 표본조사자료를 조사모집단으로 교체하였다. 층별 조사구수의 배분을 위해 비례배분과 네이만배분을 비교하여 적절한 세제곱근 비례배분을 고려하였으며, 과거의 조사결과에 대한 단순한 집계를 벗어나 새로운 표본설계로 인해 추정이 가능하였고 추정의 정확성을 측정하기위해 추정량의 분산 식을 유도하였다.

조사 단위들은 시간이 경과함에 따라 변동하므로 이에 따른 관리가 필요한다. 변경된 조사구와 가구들을 파악해서 이들을 조사에 반영해 주어야 한다. 또한 표본을 사용함으로써 발생하는 표본오차와 무응답, 응답거부, 측정과정 등에서 발생하는 비표본오차들을 살펴서 오차의 발생을 최소화 한다. 특히 조사환경의 악화로 무응답률 증가, 인위적인 표본교체 등에 따른 다양한 형태의 비표본오차가 나타날 가능성이 매우 높으므로, 이런 비표본오차를 줄이기 위해서는 조사원의 선발과 교육이 매우 중요하다고 판단된다.

그리고 경상북도의 각종 주민복지와 지역개발정책의 방향설정과 계획수립을 위한 기초자료가 되는 경북인의 생활과 의식조사의 발전을 위해 관심사인 연동표본과 다목적 표본에 대한 검토를 제안한다.

참고문헌

김영원, 류제복, 박진우, 홍기학 (1998). <표본조사의 이해와 활용>, 자유아카데미, 서울.
한국보건사회연구원 (2003). <2004년도 국민 영양조사 및 표본설계 연구용역 최종보고서>, 서울.
한국조사연구학회 (2004). <제7차 전국 장내 기생충감염 실태조사 표본설계 연구용역 최종보고서>, 서울.
Eubank, R. L. and Spiegelman, C. H. (1990). Testing the goodness of fit of a linear model via nonparametric regression techniques. *Journal of the American Statistical Association*, **85**, 387-392.
Cochran, W. G. (1977). *Sampling techniques*, Wiley, New York.
Kish, L. (1965). *Survey sampling*, Wiley, New York.
Lessler, J. T. and Kalsbeek, W. D. (1992). *Nonsampling errors in surveys*, Wiley, New York.
Lohr, S. (1999). *Sampling: Design and analysis*, Duxbury Press, Belmont, California.
SAS (1999). *SAS/STAT User's guide*, Version 8, SAS Institute Inc., Cary, North Carolina.

Journal of the Korean Data & Information Science Society
2009, **20**(6), 1155–1167

한국데이터정보과학회지

A sample design for life and attitude survey of Gyeongbuk people[†]

Dal Ho Kim[1] · Kil Ho Cho[2] · Jin Seub Hwang[3] · Kyung Ha Jung[4]

[123]Department of Statistics, Kyungpook National University
[4]Department of Planning and Coordinating, Gyeongsangbuk-do

Received 5 October 2009, revised 2 November 2009, accepted 7 November 2009

Abstract

We made a new sample design for life and consciousness survey of Kyungpook people in 2007. We used the 10% sample survey data of 2005 population and housing census as a survey population. After stratification, we allocate proportionally samples within strata after examining various characteristics in previous survey, which includes economic activity state, an income level per year, and housing possession. And we calculated weight in a new sample design and derived estimators and a formula of standard error using the weights.

Keywords: Cluster, complex sample survey, Neyman allocation, proportional allocation, stratification, systematic sampling.

[†] This paper was supported by the research project fund by Gyeongsangbuk-do, 2007.
[1] Corresponding author: Professor, Department of Statistics, Kyungpook National University, Daegu 702-701, Korea. E-mail: dalkim@knu.ac.kr
[2] Professor, Department of Statistics, Kyungpook National University, Daegu 702-701, Korea.
[3] Lecturer, Department of Statistics, Kyungpook National University, Daegu 702-701, Korea.
[4] Policy Planning Officer, Department of Planning and Coordinating, Gyeongsangbuk-do 450-701, Korea.