한국의 가구형태에 따른 Kish-격자의 조정

손창균[1], 홍기학[2], 이기성[3]

요 약

하나의 가구에서 대표되는 사람을 뽑는 문제가 조사의 마지막 단계에서 종종 발생한다. 일반적으로 가구내에서 성별과 연령에 따라 최종 조사단위로 선정되는 표본은 대표성에 문제가 있기 때문에 이를 해결하기 위한 방법으로서 Kish-격자를 이용한 방법을 사용하게 된다. 본 논문에서는 한국의 가구형태에 따라 기존의 Kish-격자를 수정하여 대표성 있는 표본을 선택하는 문제를 다루었다.

주요용어 : Kish-격자, 대표성, 추출표, 비선형방정식, 거리함수

1. 서론

일반적으로 가구표집과정의 대부분 1단계에서 1단계 추출단위(psu)를 추출하고, 2단계에서는 1단계에서 추출된 각 단위 내에서 2단계 추출단위(ssu)를 추출하는 2단계로 표본을 추출한다. 모집단으로부터 직접 응답자를 선정하는 방법대신 이와 같이 2단계표집설계를 이용하는 경우는 표본으로 취해진 성인 목록을 직접적으로 이용할 수 없기 때문이다.

가구 내에서 응답자를 선택하는 또 다른 방법으로는 지역표집 방법이 있다. 이 방법은 목표모집단이 시와 같은 지리적 영역에 위치한 경우에 사용된다. 한 시의 모집단을 연구하기 위해 추출틀은 1단계에서는 일정 구역의 목록으로 구성되고, 다음으로 거리목록으로 구성되고, 다시 블록 목록으로, 마지막으로 가구로 구성되게 된다. 그리고 마지막 단계에서는 응답자 표본을 표본가구로부터 얻게 된다. 이때, 문제점은 가구표본을 성인표본으로 변환함으로서 발생되는데, 전화조사에서 종종 발생하게 된다. 최종 표본 가구내에 거주하는 성인표본을 적절히 선택함으로서 표본의 대표성과 확률성을 유지하도록 하는 방법으로 Kish-격자를 적용할 수 있다. 그러나 한국의 가구 표본에 대한 조사의 경우 발표된 Kish-격자는 한국의 가구 특성을 잘 반영하지 못함에 따라 Kish-격자 적용의 본래 의도를 살리지 못하는 문제점이 발생하게 된다.

따라서 본 논문에서는 위와 같은 문제점을 보완한 한국의 가구형태에 적합한 Kish-격자의 조정을 제안하고, 이때 거리함수를 다양하게 적용하여 이들 간의 효율성을 살펴보고자 한다.

2. Kish-격자

Kish-격자는 최종 조사 표본의 선택을 제시하는 것으로서 격자를 이용하여 동일한 확률로 가구내에서 가구원을 선택하도록 한 것이다. 다음의 <표 2.1>은 임의의 가구내에서 조사에 참여하는 사람들을 목록으로 표현한 것이다. 표의 작성과정은 우선 가구의 세대주와의 관계를 첫

1) 협성대학교 교양학부 전임강사, 경기도 화성시 봉담읍 상리 14번지
2) 동신대학교 컴퓨터학과 교수, 전남 나주시 대호동 25번지
3) 우석대학교 전산정보학부 부교수, 전북 완주군 삼례읍 후정리

번째 열에 나타내고, 다음 두열에는 조사자가 필요하다면, 성별과 연령을 기입한다. 그 다음에 각 성인들에게 번호를 할당한다. 첫 번째에는 남성들을 나이가 많은 사람부터 적은 순서대로 번호를 부여하고, 다음으로 여성들에 대해 같은 방법으로 번호를 부여한다. 다음으로 추출표를 이용하여, 조사할 성인의 번호를 나타내준다. 예를 들어 가구에 6명의 성인이 거주하고 있다면, 4번의 성인을 추출하도록 나타낸다.

<표 2.1> 조사에 참여 가능한 성인목록의 예

관계	성별	연령	번호	선택 여부
세대주	남		2	
아내	여	54	4	V
세대주의 부	남		1	
아들	남		3	
딸1	여	21	5	
딸2	여	19	6	

<표 2.2>의 추출표는 다음의 <표 2.3>과 같이 8가지 추출가능한 표들 중 하나이다. 각각의 표에는 추출확률을 나타내고 있으며, 이와 같은 작업을 하는 근본적인 이유는 가구내에서 가구원들을 추출할 확률을 동일하게 하기 위해서이다.

<표 2.2> 8가지의 추출표 중 한 가지(Kish, 1965)

추 출 표-D	
가구에 있는 성인의 수	선택된 성인의 번호
1	1
2	2
3	2
4	3
5	4
6인 이상	4

<표 2.3> 8개의 추출 표의 요약(Kish, 1965)

할당된 표의 비율	표의 구분	가구 내 성인의 수					
		1	2	3	4	5	6인 이상
		추출 번호					
1/6	A	1	1	1	1	1	1
1/12	B1	1	1	1	1	2	2
1/12	B2	1	1	1	2	2	2
1/6	C	1	1	2	2	3	3
1/6	D	1	2	2	3	4	4
1/12	E1	1	2	3	3	3	5
1/12	E2	1	2	3	4	5	5
1/6	F	1	2	3	4	5	6

다음의 <표 2.4>는 가구내의 모든 성인들에 대해 추출확률을 나타낸 것이다. <표 2.4>에서 알 수 있듯이 6명 이상의 성인이 있는 경우는 허용하지 않으며, 1명-4명, 그리고 6명의 성인이 있는 가구의 각 성인별 추출확률은 정확히 동일한 확률을 나타낸 반면, 5명의 성인이 있는 가구의 경우 5번의 성인이 과대 대표(over-representation)하고 있다.

<표 2.4> 선택확률의 요약(Kish, 1965)

성인번호	가구내 있는 성인의 수					
	1	2	3	4	5	6인 이상
1	1	1/2	1/3	1/4	1/6	1/6
2		1/2	1/3	1/4	1/6	1/6
3			1/3	1/4	1/4	1/6
4				1/4	1/6	1/6
5					1/4	1/6
6						1/6
7인 이상						0

이와 같은 절차는 많은 연구자들에 의해 수정되었으며, Kish 자신도 특별한 경우 표들을 수정할 수 있음을 언급하였다. 기입용 설문조사의 경우 조사자는 위에서 제시한 격자를 이용할 수 있으며, 컴퓨터를 이용한 전화조사의 경우 제시된 비율에 따라 컴퓨터를 이용하여 가구를 임의로 할당하면 된다.

3. 대표성

앞에서 언급한 대표성은 표본의 바람직한 특성이다. 이는 어떤 관심특성에 대해 표본과 모집단간의 유사성을 나타내며, 추정의 정당성을 확보하기위해 대표성으로 표본을 평가하는 것이 일반적이다. Kish-격자를 사용하여 얻은 표본에 대해 대표성을 평가할 때, 많은 저서에서 남성을 과소추출(under-sampling)하며, 고령자들의 표본은 과대추출(over-sampling)하는 것으로 나타났다. 이러한 문제점은 실제 인터뷰가 이루어질 때 주로 발생되는데, 남성이 과소추출되는 경우는 이들이 인터뷰 당시 집에 없거나, 조사에 대한 참여도가 소극적이기 때문이다.

Kish-격자에 대한 논문에서는 응답자의 분포를 체크하였고, 위에서 언급한 실질적인 문제들로서 "가구에서 남성을 찾기가 어렵다"는 것으로 남성의 과소 대표성을 설명하였다(1949, Kish).

가구가 등확률로 추출될 때 가구내에서 추출확률이 동일하면 한명의 성인이 추출될 기회는 가구내에 있는 성인들의 수의 역수로 비례한다. 따라서 전체 추출확률은 등확률이 아니다. 만일 추출확률이 가구의 크기에 비례하고, 가구의 크기는 가구원의 인구학적 특성에 의존한다면 추출설계 자체가 대표성 문제의 근원이 된다. 심지어 완전한 확률 가구 표본이고 100% 응답률을 가진다고 해도, 그 표본은 대표성을 갖지 못한다. Kish는 격자를 이용하여 얻은 표본이 중요한 인구학적 특성에 대해 모집단 자료와 거의 일치한다는 사실을 밝혔다. Kish가 1950년대에 미국에서 적용한 격자로부터 추출확률의 분산이 작음을 강조하였고, 이러한 이유는 작은 범위에서 가구의 크기가 높게 집중되었기 때문이다. 그 당시 70%이상이 2인의 성인으로 구성된 가구였으며 다음의 <표 3.1>과 같다.

<표 3.1> 1957년 당시 미국의 가구 형태(Kish, 1965)

가구의 성인 수	1	2	3	4	5	6인 이상
비율(%)	14.6	73.0	9.0	2.8	0.4	0.2

지금까지의 결과로서 대표성은 현재 가구형태의 함수이며, 격자의 우월성은 그것을 언제, 어디에 적용하느냐에 따라 달라진다는 것이다. 이러한 이유로 현재 한국의 가구형태와 Kish가 관찰했던 당시의 가구형태를 비교할 필요가 있다. <표 3.2>에서와 같이 오늘날 한국에서는 24.3%가 1인 성인 가구이며, 2인 성인 가구의 비율은 64.7%, 3인 성인가구는 7.8% 등으로 나타났다.

<표 3.2> 2000년 현재 한국의 가구 형태

성인 수	1	2	3	4	5	6인 이상
비율(%)	24.3	64.7	7.8	2.5	0.6	0.1

격자의 적절성을 검증하기 위해 <표 3.3>와 같이 모집단의 연령과 성별 분포를 구하면 다음과 같다.

<표 3.3> 2000년 현재 한국의 모집단 성별과 연령의 분포(%)

연령	성 별		
	남성	여성	합 계
19-39세	24.13	25.44	49.56
40-59세	17.35	17.24	34.59
60세 이상	6.52	9.32	15.85
합 계	48.00	52.00	100.00

격자의 이용은 이러한 모집단으로 표본의 연령과 성별분포에 관련하여 검증될 수 있으며, 표본의 기대성별과 연령의 비율을 다음과 같이 정의할 수 있다. p_{kl}을 크기 k인 가구에 살고 있는 성인 l의 추출확률이라 하고, $(k=1,2,\cdots,6, l=1,2,\cdots,k)$ 가구는 미리 추출되었다고 가정하자. 가구들이 등확률로 추출됨으로서 크기 k인 가구를 선택할 기회는 이들 가구의 비율과 같게 되며, H_k를 가구의 크기가 k인 가구의 비율이라 하자.

기대성별과 연령그룹의 결합분포를 3×2 행렬로 나타낼 수 있으며, 이 행렬을 a라 하면, $a[11]$은 젊은 남성의 비율, $a[21]$은 중년남성의 비율, $a[32]$는 고령여성의 비율 등을 나타낸다.

가구들의 구성에 관한 정보가 필요한데, 크기가 k인 가구에서 번호가 l인 사람을 선택한 후 그가 남성 또는 여성일 확률, 또는 그가 젊은이거나 중년, 고령일 확률을 알아야 한다. 이를 a_{kl}이라 하고, 3×2행렬이라 하자. $(k=1,2,\cdots,6, l=1,2,\cdots,k)$ 앞에서와 같은 방법으로 $a_{kl}[11]$은 크기가 k인 가구에서 번호가 l인 사람 중에서 젊은 남성의 비율, $a_{kl}[21]$은 크기가 k인 가구에서 번호가 l인 사람 중 에서 중년 남성의 비율 등을 나타낸다.

기대성별과 연령의 결합분포는 다른 모수들의 함수이다. H_k, a, a_{kl}등은 기지의 모수인데, 이 값들은 모집단에 관한 정보로부터 알 수 있다.

$$a[ij]=\sum_k H_k\left(\sum_l p_{kl}a_{kl}[ij]\right),\ i=1,2,3,\ j=1,2 \tag{3.1}$$

식(3.1)에서 기지 모수값들로 대체하여, 기대분포를 얻을 수 있다.

4. Kish-격자의 조정

이 절에서는 한국의 2000년 현재 가구형태에 따라 Kish-격자의 조정에 대해 다루고자 한다. 이때, 의도하는 바는 대표성 또는 적어도 기대 표본이 대표성을 유지하도록 하는 것이며, 이를 위해 격자는 추출표를 변화시킴으로서 조정이 가능하다. 이와 같은 조정과정은 앞에서 언급하였듯이 Kish 자신도 필요한 경우 추출표를 조정할 수 있음을 지적한 바 있다.

추출표의 확률을 조정하기 위해 모든 표본은 고정이며, 다음과 같은 조건을 만족하도록 한다.

① 각 가구는 동일한 추출기회를 갖는다.
② 가구당 단 한명에 대해 면접이 이루진다.
③ 추출표는 가구 구성원의 목록에 기초한다.
④ 가구원에 대한 순서화는 성별과 연령에 따라 이루어진다.
⑤ 조사될 모집단은 앞에서 언급한 모집단이다.
⑥ 12개의 추출표를 사용한다.
⑦ 동일한 규칙을 6명 이상의 가구원을 가진 가구에 대해 적용한다.

이 경우 문제점은 모집단 자료와 가장 근접한 표본을 제공해주는 추출표를 작성하는 것이다. 따라서 가능한 한 <표 3.3>에서 제시한 분포와 가장 가까운 대표성 있는 기대표본을 얻는 것이 주된 목적이라 할 수 있다.

이를 위해 A를 모집단의 성별과 연령의 분포를 나타내는 3×2 행렬이라 하자. 앞에서와 마찬가지로 $A[11]$은 젊은 남성의 비율을 나타낸다. 식(3.1)의 기호를 적용하여 H_k와 a_{ij}는 기지의 모수이며, a는 p_{kl}의 함수로 결정되며, 따라서 A를 재 계산한다. 이를 위해 다음의 식(4.1)을 다음의 조건하에서 풀어야 한다.

$$\sum_{i,j} |a[ij] - A[ij]| = 0 \tag{4.1}$$

조건식(1) $\sum p_{ij} = 1$

조건식(2) $p_{ij} > 0$

조건식(3) $p_{ij} = k_{ij}/12$, 모든 i,j에 대해 k_{ij}는 자연수

조건식은 가구당 최소한 한명이 필요하며, 또한 12개의 추출표가 필요하기 때문에 확률은 1/12로 주어진다. 이와 같은 모형은 부등식과 자연수 조건식을 갖는 비선형 방정식이다. 그러나 이 비선형 방정식의 해는 존재하지 않는다. 이 경우 필요한 표의 수를 제한하거나 표본의 수를 증가시켜 해를 구할 수는 있을지 모르지만, 그에 따르는 비용의 문제와 표본의 대표성의 문제가 발생하게 된다. 이와 같은 문제점을 보완하기 위해 거리함수를 도입하기로 한다. 즉, 모집단의 분포와 가장 근접한 표본을 추출하기 위한 방법으로 거리함수를 최소로 하는 가장 근접한 해를 제시하도록 하는 방법을 적용하자.

우선 다음과 같이 선형거리함수를 고려하면,

$$LSD(a) = \sum_{i,j} \frac{(a[ij] - A[ij])^2}{A[ij]} \tag{4.2}$$

으로서 $LSD(\cdot)$를 최소로 하는 해를 구하면 된다. 이와 같이 고려할 수 있는 거리함수로는 다음과 같은 것들이 있다.

$$RAKE(a) = \sum_{i,j} -a[i,j]\log\left(\frac{a[ij]}{A[ij]}\right) - \sum_{i,j}(a[ij] - A[ij]) \tag{4.3}$$

$$HELL(a) = \sum_{i,j}\left(\sqrt{a[ij]} - \sqrt{A[ij]}\right)^2 \tag{4.4}$$

$$MIEN(a) = \sum_{i,j} -A[ij]\log\left(\frac{a[ij]}{A[ij]}\right) + \sum_{i,j}(a[ij] - A[ij]) \tag{4.5}$$

$$CHI(a) = \sum_{i,j} \frac{(a[ij] - A[ij])^2}{a[ij]} \tag{4.6}$$

5. 결론

Kish-격자를 한국의 가구형태에 맞추어 조정함으로서 가구표집에 대한 최종 조사표본의 대표성 문제를 고려하였다. 현재 한국의 가구형태는 급격한 핵가족화의 진행으로 1인 가구의 비율이 점점 증가하고 있으며, 또한 고령화 사회로 접어들면서 기존의 표본선정 방법의 적용에는 적잖은 문제가 발생한다. 이를 위해 기존의 Kish-격자를 한국의 가구형태에 맞도록 조정하였으며, 2000년 인구주택 총 조사 모집단의 분포를 고려하여 최적의 해를 구하여 보았다.

참고문헌

[1] D. Binson, J.A.Canchola, and J.A.Catania. (2000). Random selection in a telephone survey : A comparison of the Kish, next-birthday, and last-birthday methods, *Journal of Official Statistics*, 16, 53-59.

[2] R.M. Groves, P.P. Biemer, L.E.Lyberg, J.T. Nicholls, and J. Waksberg.(1988). *Telephone Survey Methodology*. John Wiley & Sons, Inc. New York.

[3] J.M..Kenedy.(1993). A comparison of telephone survey respondent selection procedure, http://www.indiana.edu/csr/aapor93.html.

[4] L. Kish. (1949). A procedure for objective respondent selection within the household. *Journal of American Statistical Association*, 380-387.

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

[6] 통계청 DB (2004), http://kosis.nso.go.kr