조사연구를 위한 표본크기 결정

박 진 우* · 정 미 숙**

사회조사를 위한 표본설계를 할 때 표본의 크기를 얼마로 할 것인지를 결정하는 문제는 조사연구자에게 고민거리가 된다. 사회조사 중에서 4점 또는 5점 척도로 된 여러 개의 개별 문항들로 구성된 설문지를 사용하는 경우가 많다. 이런 경우 개개의 문항 자체를 직접적으로 하나의 변수로 사용하지 않고 여러 개 문항들을 결합하여 새로운 척도를 만들어 사용하는 것이 일반적이다. 본 연구의 목적은 리커트 척도가 관심변수인 조사연구에서 표본크기를 결정하는 방법을 제공하는 것이다.

리커트 척도를 만들고자 할 때 4점 혹은 5점 척도로 구성된 여러 문항변수들은 일반적으로 서로 양의 상관관계를 가지게 된다. 본 연구에서는 개별 문항변수들은 각각 동일한 분포를 가지며, 각각의 변수들은 서로 동일한 크기의 상관관계를 갖는다는 가정을 한다. 주어진 가정 하에서 새로운 척도의 표본분포를 유도한 후 이를 이용하여 다양한 상황에서의 표본의 크기를 계산한 결과를 표로 제시하게 되는데 표본이론을 잘 모르는 조사연구자들은 이 표를 이용하여 원하는 표본크기를 결정할 수 있을 것이다.

주제어: 리커트 척도, 표본오차, 표본크기, 변동계수

I. 서 론

사회조사를 위해 표본설계를 할 때 표본의 크기를 얼마로 할 것인지를 결정하는 것은 조사연구자들이 만나는 중요한 문제들 중 하나이다. 표본의 모집단에 대한 대표성을 보장하는 동시에 관심모수 추정량의 표본오차가 조사목적에 합당한 수준 이내가 되도록 표본의 크기를 정해주는

*수원대학교 통계정보학과 교수
**수원대학교 통계정보학과 강사

것이 필요하다. Fowler (1993)는 조사연구자가 표본크기를 결정할 때 흔히 범하는 잘못들을 지적하였다.

전통적인 표본조사 관련문헌에서는 표본의 크기를 정하는 방식으로 추정하고자 하는 통계량에 대해 원하는 신뢰도에서의 오차의 한계의 공식을 유도한 후 그것을 연구자가 원하는 목표오차에 일치시켜 표본크기를 계산하는 방법을 소개하고 있다 (박홍래, 2000; 홍두승, 2000; Kish, L., 1965; Cochran, W., 1977). 단순랜덤표집에서 표본평균에 대한 $100(1-\alpha)\%$ 신뢰구간에서의 오차의 한계를 구한 후 이것을 목표오차의 한계와 일치시켜 표본크기를 계산하는 방법이 대표적인 방법이다. 이 방법을 사용하려면 관심변수의 모분산에 대한 사전정보가 필요한데 표본이론에 익숙지 않은 조사연구자들이 이러한 공식을 이용하여 직접적으로 표본크기를 계산하는 일은 쉽지 않다. 이런 까닭에 조사연구자들은 모비율 추정의 문제에서 표본크기를 구하는 공식을 사용하여 표본크기를 정하는 방법을 널리 사용하는데 이 방법은 계산이 간편할 뿐 아니라 개인이 직접 계산하지 않고도 사용할 수 있도록 표의 형태로 제공되고 있기도 하기 때문이다 (Yamane, 1965). 그런데 선거여론조사를 제외한 일반적인 사회조사에서 모비율이 관심모수인 경우는 거의 없고 그보다는 4점 척도 또는 5점 척도를 사용하는 여러 개의 문항으로 설문지를 구성한 후 여러 개 문항들을 결합하여 만든 새로운 척도가 관심모수일 때가 많다. 리커트 척도 (Likert, 1970)는 가장 단순한 형태의 척도인 동시에 일반적으로 가장 널리 쓰이는 척도이기도 하다 (Babbie, 1979). 이런 종류의 사회조사를 위해 표본크기를 결정할 때 모비율 추정 문제에서 표본크기 결정 방식을 이용하는 것은 적절하지 않다.

본 연구의 목적은 리커트 척도가 관심변수인 조사연구에서 표본크기를 결정하는 방법을 제공하는 것이다. 리커트 척도를 만들고자 할 때 4점 혹은 5점 척도로 구성된 여러 문항변수들은 일반적으로 서로 양의 상관관계를 가지게 된다. 본 연구에서는 개별 문항변수들은 각각 동일한 분포를 가지며, 각각의 변수들은 서로 동일한 크기의 상관관계를 갖는다는 가정을 한다. 주어진 가정 하에서 새로운 척도의 표본분포를 유도한 후 이를 이용하여 다양한 상황에서의 표본의 크기를 계산한 결과를 표로 제시하게 되는데 표본이론을 잘 모르는 조사연구자들은 이 표를 이용하여 원하는 표본크기를 결정할 수 있을 것이다.

2절에서는 간단한 통계적 가정을 하고 이 때 리커트 척도의 표본분포를 유도함으로 주어진 목표오차 하에서의 표본크기를 구하는 공식을 나타낸다. 3절에서는 여러 가지 모수값과 표본오차를 가정하는 경우의 표본크기를 계산하여 일목요연하게 표로 제시한다. 여기서 모비율 추정에서의 표본크기를 결정하는 방식은 본 연구에서 제시하는 방식의 한 특별한 경우라는 사실과, 리커트 척도를 사용하는 조사에서 모비율 추정을 위한 표본크기 결정 방식을 사용하게 되면 실제 목표오차에 비해 과다한 표본크기가 결정된다는 사실을 밝힌다. 마지막 절에서는 본 연구 결과를 요약한다.

Ⅱ. 리커트 척도일 때의 표본크기 결정

조사문항들이 각각 "매우 찬성", "약간 찬성", "보통", "약간 불만", "매우 불만"과 같은 4점 또는 5점 척도들로 구성되는 설문조사를 고려하자. 이것을 좀더 일반적으로 표기해보자. 각 문항들은 M개의 응답범주를 가지는데 응답범주별로 각각 0에서 $(M-1)$점의 점수를 부여한다. i번째 응답자가 k개의 조사문항에 대해 응답한 것에 대해 0에서 $(M-1)$점의 점수를 부여한 문항별 응답점수를 $X_{i1}, X_{i2}, \cdots, X_{ij}, \cdots, X_{ik}$라는 변수로 나타내기로 한다. 표본의 크기를 n이라고 하면 i, j는 각각 $(i=1,2,\cdots,n)$, $(j=1,2,\cdots,k)$ 사이의 값을 갖는다. i번째 응답자에 대한 리커트 척도를 Y_i라고 표기하면 이것은 다음의 (1) 식으로 나타낼 수 있다:

$$Y_i = \frac{1}{k}\sum_{j=1}^{k} X_{ij} \ . \tag{1}$$

개별 응답자의 항목별 응답변수 $X_{ij}(i=1,2,\cdots,n; j=1,2,\cdots,k)$들은 각각 동일한 평균 μ와 분산 σ^2을 갖는다고 가정하자. 아울러 모든 응답자의 서로 다른 문항에 대한 응답변수들은 모두 동일한 상관계수인 ρ를 갖는다고 가정하자. 여기서의 가정을 식으로 표현하면 아래의 (2)식과 같이 나타낼 수 있다.

$$\begin{aligned} &X_{ij}\ (i=1,2,\cdots,n; j=1,2,\cdots,k) \sim (\mu, \sigma^2) \\ &Corr(X_{ij}, X_{ij'}) = \rho,\ (j \neq j') \end{aligned} \tag{2}$$

위의 가정이 일반적인 설문조사에서 타당한 가정인지를 살펴보기 위해 아래 수원대학생들의 만족도조사 예를 들어보기로 하자.

<예제> 수원대학교 학생들의 전공만족도 조사

수원대학교 학생들의 학교생활만족도를 측정하기 위해 <표 1>에 나온 바와 같이 다섯 문항을 고려하였다. 각각의 조사변수는 5점 척도로 구성되며 이것들을 합한 리커트 척도를 만들어 학교생활만족도 척도로 사용하려 한다. 표본은 약 10,000명의 전체 재학생 중 500명을 층화추출하여 조사를 실시하였다.

<표 1> 수원대학생들의 전공만족도조사 설문 문항

설 문 문 항	변수
1. 당신은 자신의 전공에 대해 만족합니까?	X_1
2. 당신의 학과에서 제공하는 전공 커리큘럼이 잘 구성되어 있다고 생각합니까?	X_2
3. 당신은 학과 교수님들의 전공강의에 대하여 만족합니까?	X_3
4. 당신은 학과의 학습 분위기에 대해 만족합니까?	X_4

<표 2>는 조사된 데이터를 가지고 개별 문항변수별로 평균, 표준편차 및 변동계수를 계산한 결과이다. 개별 변수들은 0에서 4까지의 값을 가질 수 있는데 0은 '매우 불만', 4는 '매우 만족'을 나타낸다. 조사 데이터를 통해 얻어진 변수별 추정평균은 1.96에서 2.29 사이 임을 알 수 있다. 한편 변수별 표준편차는 0.78에서 0.89 사이의 값으로 계산되었다. 표본설계 단계에서 표본크기를 결정하려 할 때 아주 정교한 값을 구하기보다 보수적인 관점에서 여유 있게 정하는 것이 일반적이므로 이러한 계산 결과는 개별 변수들에 대해 동일한 분포를 가정한 앞의 가정이 무난하게 받아들여질 수 있음을 보여준다.

<표 2> 문항변수 별 기초통계값

변수	X_1	X_2	X_3	X_4
관측수	500	500	500	500
평균	2.29	2.14	1.96	2.05
표준 편차	0.88	0.89	0.78	0.85

다음으로 개별 문항변수 상호간의 상관관계에 대한 가정을 검토하기 위해 변수들 간의 상관계수를 계산한 결과가 <표 3>에 나와 있다. 변수들 간의 상관계수는 0.29에서 0.67 사이의 값을 갖는데 대체로 0.4 내외의 값을 가지는 것으로 보인다. 이를 통해서도 변수들 간의 상관계수가 모두 같다는 가정이 현실적으로 그다지 무리한 것이 아님을 짐작할 수 있다.

<표 3> 변수들간의 상관계수

	X_1	X_2	X_3	X_4
X_1	1	0.67	0.42	0.29
X_2	0.67	1	0.48	0.33
X_3	0.42	0.48	1	0.50
X_4	0.29	0.33	0.5	1

참고로 위 다섯 개 문항을 이용하여 종합만족도 (Y)라는 리커트 척도를 구성한 후 이 척도의 모평균을 추정한 결과 95% 신뢰수준 하에서 오차의 한계는 ±2.75%로 계산되었다. 만일 표본크기 500개를 결정할 때 모비율 추정을 위한 표본크기 결정 방식으로 결정한 것이라면 애초 목표오차의 한계는 ±8.9%로 한 셈인데 실제 달성된 오차의 한계와 매우 차이가 난다. 리커트 척도를 사용하는 조사연구에서 표본크기를 결정할 때 모비율 추정을 위한 표본크기 결정 방식을 사용하면 실제 요구되는 것에 비해 표본크기를 불필요하게 크게 할 우려가 있게 된다.

중심극한정리를 이용하면 리커트 척도 Y_i 의 근사적인 분포는 다음과 같은 형태의 정규분포를 갖는다는 것을 쉽게 보일 수 있다.

$$Y_i = \frac{1}{k} \sum_{j=1}^{k} X_{ij} \sim N(\mu, \sigma*^2)$$

여기서 분산은 $\sigma*^2 = \frac{\sigma^2}{k}(1+(k-1)\rho)$이다.

조사를 통해 구하고자 하는 추정량은 리커트 척도의 표본평균인 $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ 이다. $Y_1, \cdots, Y_n$은 서로 독립이며 동일한 분포를 갖는 확률변수이므로 표본평균은 평균 μ, 분산 $\sigma*^2/n$ 인 정규분포를 따른다는 사실을 알 수 있다.

표본크기를 결정하기 위해서는 먼저 목표 오차의 한계를 정해주어야 하는데 오차의 한계를 B라고 나타내자. 그러면 리커트 척도의 표본평균의 분포를 이용하여 (3)식과 같이 나타낼 수 있다.

$$\begin{aligned} B &= z_{\alpha/2} \cdot \sqrt{V(\overline{Y})} \\ &= z_{\alpha/2} \cdot \sqrt{\frac{\sigma^2}{nk}[1+(k-1)\rho]} \end{aligned} \tag{3}$$

여기서 z는 표준정규분포 아래서 구해지는 신뢰계수를 나타낸다.

위의 식을 n에 대해 풀어냄으로 표본의 크기를 결정할 수 있다. n의 계산공식을 나타내는 (4)식을 보면 표본크기를 결정하기 위해서는 신뢰도 α와 목표 오차의 한계 B 외에 각 문항변수 X_{ij}의 분산인 σ^2과 서로 다른 문항변수들 간의 상관계수인 ρ값이 주어져야 함을 알 수 있다.

$$n = \frac{z^2_{\alpha/w}\sigma^2[1+(k-1)\rho]}{kB^2} \tag{4}$$

일반적으로 표본설계 단계에서 조사연구자가 목표 오차의 한계의 크기를 정하는 것이 쉽지 않다. 그러므로 이런 경우에는 상대 목표오차의 크기를 고려하는 것이 편리하다. 상대 목표오차

란 앞의 목표오차인 B를 추정모수인 μ로 나눈 값을 의미하는데 여기서는 D로 표기하기로 한다. 그러면 (4)식은 다음의 (5)식으로 표현된다.

$$\begin{aligned} D=\frac{B}{\mu} &= \frac{z_{\alpha/2}\cdot\sqrt{V(\overline{Y})}}{\mu} \\ &= z_{\alpha/2}\cdot\sqrt{\frac{\sigma^2}{\mu^2}\cdot\frac{1}{nk}[1+(k-1)\rho]} \\ &= z_{\alpha/2}\cdot\sqrt{C^2\cdot\frac{1}{nk}[1+(k-1)\rho]} \end{aligned} \tag{5}$$

위의 식에서 C는 설문지의 개별 문항변수인 X_{ij}의 변동계수(coefficient of variation)를 나타낸다.

한편 (5)식을 n에 대해 풀어 계산하면 (6)식과 같다.

$$n=\frac{z_{\alpha/w}^2\cdot C^2[1+(k-1)\rho]}{kD^2} \tag{6}$$

Ⅲ. 상황별 표본크기 계산

앞 절에서는 간단한 가정 하에서 리커트 척도를 사용할 경우 목표오차의 크기에 따라 표본크기를 계산하는 공식을 유도하였다. 그런데 (6)식을 사용하여 표본크기를 결정하려면 신뢰도 $1-\alpha$와 목표오차의 크기 D 이외에도 리커트 척도를 구성하기 위해 사용한 변수의 수 k, 모집단의 변동계수 C, 개별 조사변수들 간의 상관계수 ρ 값이 주어져야만 한다.

여기서는 이런 각각의 값들을 여러 가지로 변화시켜 가면서 경우별 표본크기를 계산한 결과를 몇 가지 표로 제시하기로 한다. 실제 일반적인 조사연구자가 표본크기를 결정하기 위해 (6)식을 사용하는 것이 쉽지 않다고 생각하여 다양한 상황에서 계산한 표를 제시하고 이를 이용하여 표본크기 결정할 수 있도록 간단한 지침을 제시한다.

먼저 리커트 척도 구성을 위해 사용된 문항의 수 k는 1에서 10까지를 고려하였다. k가 1이고 응답범주가 2개인 경우라면 모비율 추정의 문제가 된다. 따라서 본 연구에서 제시하는 표본크기 결정 방법은 널리 사용하는 모비율 추정식을 이용한 표본크기 결정 방식을 보다 일반적인 경우로 확장한 것이다. 개별 조사변수들의 변동계수 C는 0.1에서 1.0 사이의 값, 조사변수들 간의 상관계수 ρ는 0.1에서 0.7 사이의 값을 고려하였는데 이러한 범위는 대부분의 현실적인 상황을 충분히 포괄할 수 있다. 한편 목표오차의 한계 $D(=B/\mu)$는 1%, 2%, 3%, 4%, 5%, 10%에

대해 고려하였다.

<표 4>는 변동계수 $C=1.0$이면서 상관계수 ρ는 0.3인 경우 표본크기를 계산한 표이다. 변동계수 $C=1.0$ 인 경우는 모표준편차와 모평균의 크기가 같은 경우로서 $k=1$인 경우는 $p=0.5$인 모비율 추정을 위한 표본크기를 결정하는 문제와 동일하다. 먼저 <표 4>를 관찰해 보면 k값이 커질수록 주어진 목표오차를 만족시키기 위한 표본크기가 줄어드는 것을 알 수 있다. 가령 $k=5$일 때에는 $k=1$인 때에 비해 요구되는 표본의 크기가 반 이하이다. 다시 말하자면 리커트 척도를 구성하기 위해 5개 이상의 조사변수를 사용하는 사회조사에서 표본크기를 정할 때 모비율 추정을 위한 표본크기 결정 공식을 사용하면 실제 필요한 표본크기에 비해 불필요하게 2배 이상의 표본크기를 추출하게 된다는 뜻이다. 그러나 본 연구에서 제시한 방법을 사용할 경우 불필요하게 표본크기를 크게 할 필요가 없으므로 보다 효율적이라고 할 수 있다.

<표 4> $C=1.0$, $\rho=0.3$ 인 경우의 표본크기

D \ k	±1%	±2%	±3%	±4%	±5%	±10%
1	38416	9604	4268	2401	1537	384
2	24970	6243	2774	1561	999	250
3	20489	5122	2277	1281	820	205
4	18248	4562	2028	1140	730	182
5	16903	4226	1878	1056	676	169
6	16007	4002	1779	1000	640	160
7	15366	3842	1707	960	615	154
8	14886	3722	1654	930	595	149
9	14513	3628	1613	907	581	145
10	14214	3553	1579	888	569	142

<표 5>는 변동계수 $C=1.0$이면서 상관계수 ρ는 0.5인 경우 표본크기를 계산한 표로써 앞의 <표 4>와는 상관계수가 다르다. 두 표를 비교해보면 조사변수들끼리의 상관계수가 클수록 요구되는 표본의 크기가 늘어난다는 것을 알 수 있다. 참고로 (6)식을 보면 표본크기는 상관계수의 크기와 양의 직선 관계를 가진다.

<표 5> $C=1.0$, $\rho=0.5$ 인 경우의 표본크기

D \ k	±1%	±2%	±3%	±4%	±5%	±10%
1	38416	9604	4268	2401	1537	384
2	28812	7203	3201	1801	1152	288
3	25611	6403	2846	1601	1024	256
4	24010	6003	2668	1501	960	240
5	23050	5762	2561	1441	922	230
6	22409	5602	2490	1401	896	224
7	21952	5488	2439	1372	878	220
8	21609	5402	2401	1351	864	216
9	21342	5336	2371	1334	854	213
10	21129	5282	2348	1321	845	211

한편 다음의 <표 6>은 변동계수 $C=0.5$이면서 상관계수 ρ는 0.3일 때의 표본크기를 계산한 것이다. (6)식을 보면 표본크기는 변동계수의 제곱과 비례하는 것을 알 수 있는데 이런 사실은 <표 6>을 앞의 <표 5>와 비교해보면 확실히 알 수 있다. 설문조사에서 각각의 문항변수들이 4점이나 5점 척도일 때 일반적으로 중간항목의 비율이 높은 편이므로 이 경우 변동계수는 1.0보다 훨씬 작게 된다. 앞 절에서 소개한 수원대학생들의 만족도조사 예에서는 문항별 변동계수가 대체로 0.3에서 0.5 사이로 나타났는데 이는 일반적인 조사에서도 마찬가지이다.

<표 6> $C=0.5$, $\rho=0.5$ 인 경우의 표본크기

D \ k	±1%	±2%	±3%	±4%	±5%	±10%
1	9604	2401	1068	601	385	97
2	7203	1801	801	451	289	73
3	6403	1601	712	401	257	65
4	6003	1501	667	376	241	61
5	5763	1441	641	361	231	58
6	5603	1401	623	351	225	57
7	5488	1372	610	343	220	55
8	5403	1351	601	338	217	55
9	5336	1334	593	334	214	54
10	5283	1321	587	331	212	53

앞 절에서 소개한 수원대학교 만족도조사의 예에서 살펴본 바와 같이 일반적인 사회조사에서 각 문항변수들의 변동계수 C는 0.3에서 0.5 사이의 값을 갖는다. 한편 하나의 척도 안에 포함되는 개별 문항변수들 간의 상관계수 ρ는 변수에 따라 차이가 나지만 평균적으로 0.2에서 0.5 내외인 것이 보통이다. C나 ρ 값 모두 커질수록 요구되는 표본의 크기가 커지는데 상대적으로 C의 값이 표본크기에 더 큰 영향을 미친다. 일반적으로 조사연구의 표본설계 단계에서 표본크기를 결정할 때 가능한 한 보수적으로 하는 것이 바람직하므로 변동계수 C는 0.5 정도, 상관계수 ρ는 0.5나 0.6 정도를 사용할 것을 제안한다. 그렇게 할 경우 기대오차 수준을 달성하기에 충분하면서도 불필요하게 많은 표본을 추출하지 않게 되어 경제적이고 효율적인 표본설계를 할 수 있으리라 생각한다. 참고로 다음의 <표 7>은 $C=0.5$, $\rho=0.6$인 경우 표본크기를 나타내는 표이다.

<표 7> $C=0.5$, $\rho=0.6$ 인 경우의 표본크기

D \ k	±1%	±2%	±3%	±4%	±5%	±10%
1	9604	2401	1068	601	385	97
2	7684	1921	854	481	308	77
3	7043	1761	783	441	282	71
4	6723	1681	747	421	269	68
5	6531	1633	726	409	262	66
6	6403	1601	712	401	257	65
7	6312	1578	702	395	253	64
8	6243	1561	694	391	250	63
9	6190	1548	688	387	248	62
10	6147	1537	683	385	246	62

Ⅳ. 맺음말

표본크기 결정의 문제는 조사연구자가 표본설계 단계에서 가장 먼저 부딪히게 되는 문제 중의 하나이다. 표본크기는 조사비용과 시간, 추정의 정도(precision) 등과 밀접하게 연관되어 있으므로 조사목적에 합당하도록 결정하는 것이 필요하다. 그러나 표본이론에 능숙하지 않은 일반적인 연구자에게는 결코 쉽지만은 않은 문제이기도 하다.

본 연구에서는 최종적인 연구모수가 리커트 척도의 형태로 표현되는 조사에서 표본크기를

결정하는 방법을 제시하였다. 상식적으로 받아들일 수 있을 만한 가정 아래에서 표본크기 결정 공식을 제시하였는데 이 때 개별 문항들 사이의 상관계수(ρ), 개별 문항들의 변동계수(C) 등의 모수가 표본크기에 영향을 미치는 것을 알 수 있었다. 본 연구에서는 여러 종류의 상관계수 및 변동계수에 따른 표본크기를 계산하여 표로 제시하였다. 아울러 일반적인 연구에서 $C=0.5$, $\rho=0.6$ 정도로 하여 표본크기를 결정할 것을 제안하기도 하였다.

본 연구의 결과 리커트 척도를 사용하는 조사연구임에도 불구하고 모비율 추정 문제의 표본크기 결정 방식을 사용할 경우 실제 목표하는 정도에 비해 불필요하게 많은 표본을 추출하게 된다는 사실을 알 수 있었다. 실제 본 연구에서 제안하는 방식으로 표본크기를 결정할 경우 모비율 추정에서의 표본크기 결정 방식을 사용하는 것에 비해 30% 내지 40% 정도 표본크기를 줄일 수 있게 된다.

참 고 문 헌

박홍래. 2000. 《통계조사론》. 영지문화사.

홍두승. 2000. 《사회조사분석》. 다산출판사.

Fowler, F. J. 1984. Survey Research Methods, Applied Social Reseach Methods Series Vol. 1, Beverly Hills, CA: Sage.

Kish, L. 1965. Survey Sampling. Wiley.

Yamane, T. 1967. Elementary Sampling Theory. Prentice-Hall.

Cochran, W. G. 1977. Sampling Techniques. John Wiley & Sons, New York.

Babbie, E. R. 1989. The Practice of Social Research. Wadworth.

Likert, R. 1970. "A Technique for the Measurement of Attitudes". Chap. 7 in Attitude Measurement.