응용통계연구 제 8권 1호
1995년 3월 pp. 27-38.

층화이중추출법에 의한 양적속성의 무관질문모형

이 기 성[1], 홍 기 학[2]

요 약

본 논문에서는 사회적으로나 개인적으로 매우 민감한 조사에서 모집단이 양적속성을 갖는 여러 개의 층으로 구성되어 있을 때, 층의 크기를 모르는 경우 층화표본을 위하여 이중추출법을 이용하는 층화이중추출법에 의한 양적속성의 무관질문모형을 제안하였다. 그리고, 층화이중추출에 있어서 각 층의 표본배분에 관해 비례배분, 최적배분으로 나누어 각 층의 크기를 알고 있는 경우에 층화추출법에 의한 양적속성의 무관질문모형과 그 효율성을 비교하였다.

1. 서 론

사회 여러 분야의 조사에서 응답자들은 개인의 사생활과 깊은 관련이 있는 민감한 질문을 받았을 경우 자신의 신분이나 비밀을 보호하기 위해 응답을 회피하거나 고의적인 거짓 응답을 하게 된다. 이로 인해 발생하는 비표본오차(nonsampling error)를 줄이기 위해 Warner(1965)는 확률장치를 통한 간접응답으로 응답자의 신분이나 비밀을 노출시키지 않고서 민감한 질문에 대해 정보를 이끌어 낼 수 있는 확률화응답모형(randomized response model ; RRM)을 처음으로 제시하였다.

그 후 Greenberg(1971) 등은 무관질문모형(unrelated question model)을 제안하여 양적속성(quantitative attribute)을 갖는 경우로 확장하였으며, 이러한 양적속성의 확률화응답모형은 그 이후 Eriksson(1973), Poole(1974), 그리고 Pollock와 Bek(1976) 등 많은 학자들에 의해 연구, 발전되어 왔다. 특히, 이기성 외 2인(1993)은 조사하고자 하는 모집단이 양적인 속성을 갖는 여러 개의 층으로 구성되어 각 층의 크기를 알고 있는 경우에, Greenberg 등이 제안한 무관질문모형에 층화추출법을 적용하여 각 층의 평균에 대한 추정뿐만 아니라 모집단 전체 모평균에 대한 추정을 할 수 있는 층화추출법에 의한 양적속성의 무관질문모형을 제안하였다.

본 논문에서는 층화추출에 있어서 각 층의 크기를 모르는 경우에, 층화표본을 위하여 이중추출법(double sampling)을 이용하는 층화이중추출법에 의한 양적속성의 무관질문모형을 제안하여 각 층의 크기를 알고 있는 경우에 층화추출법에 의한 양적속성의 무관질문모형과 그 효율성을 비교해 보고자 한다.

2. 층화이중추출법에 의한 양적속성의 무관질문모형

모집단이 양적인 민감한 속성을 갖는 여러 개의 층으로 구성되어 있는 경우 각 층의 모평균

1) (565-800) 전북 완주군 삼례읍 후정리 480, 전주우석대학교 계산통계학과.
2) (520-714) 전남 나주시 대호동 252, 동신대학교 전산통계학과.

과 전체 모평균에 대한 추정을 필요로 할 때 확률화응답모형에 층화임의추출법을 적용할 수 있으며, 이러한 층화임의추출에 있어서 각 층의 크기를 모르는 경우에 층화표본을 위하여 층화이중추출법을 이용할 수 있다. 이 때, 1단계로 응답자에게 민감한 속성과 무관한 질문을 하고 그 결과에 의하여 층을 나눈 다음, 2단계로 각 층에 확률화응답질문을 하는 층화이중추출법을 이용할 수 있다. 따라서, 본 장에서는 각 층의 크기를 모르는 경우 Greenberg 등의 양적속성을 갖는 무관질문모형에 층화이중추출법을 적용하여 비용에 따른 최소분산과 표본배분을 다루어 보고자 한다.

상호 배반인 L개의 층으로 구성되어 있는 크기 N인 모집단에서 1단계 표본으로 크기 n'인 표본을 단순임의복원추출하여 직접질문을 한다.

예를 들어 응답자들을 태어난 월 별로 층화하고자 할 경우 다음 설문을 직접 질문한다.

설 문 : "나의 생일은 i ($i=1,2,\cdots,12(=L)$)월이다."

1단계 표본을 i개의 그룹(층)으로 분류하고 그룹 i에 속하는 사람의 수를 n_i' 라 하자. 그러면 W_i와 w_i는 다음과 같다.

$$W_i = \frac{N_i}{N} \quad (i=1,2,\cdots,L) \qquad : \text{그룹 } i\text{에 속하는 모집단 비율.}$$

$$w_i = \frac{n_i'}{n'} \quad (i=1,2,\cdots,L) \qquad : \text{그룹 } i\text{에 속하는 표본 비율.}$$

여기서, w_i는 W_i의 불편추정량이다.

2단계에서는 그룹 i에 속하는 사람의 수 n_i' 로부터 2단계 표본 n_i ($n=\sum_{i=1}^{L} n_i$)를 단순임의복원추출하여 민감한 질문에 대해 응답하도록 한다. 이 때, 2단계의 각 층 i ($i=1,2,\cdots,L$)에서 민감한 질문에 대해 민감한 변수 X가 연속인 밀도함수 $g_i(\cdot)$를 갖는다고 가정하고, Y를 밀도함수 $h_i(\cdot)$를 갖는 무관속성의 변수라고 하자. 여기서 각 층의 민감한 변수와 무관속성 변수의 확률함수의 형태는 모두 연속(또는 이산)이고 어느 경우이든 기대값은 반드시 존재한다. 단, 위와 같은 조건하에서 $g_i(\cdot)$와 $h_i(\cdot)$는 층마다 같을 수도 있고 다를 수도 있다. 그리고, 전체 X의 모평균 $\mu_{X(s)}$를 추정하는데 있어서, i층에서의 무관속성의 변수 Y의 모평균 μ_{Yi}를 알고 있다고 가정하자.

각 층에서 응답자들은 민감한 변수 X가 선택될 확률이 p이고 무관속성의 변수 Y가 선택될 확률이 $q=1-p$인 확률장치를 통해 선택된 변수에 대해 응답하게 된다.

이 때, i번째 층에서 j번째 응답자가 Z_{ij}라고 응답하면, Z_{ij}는 다음과 같은 확률밀도함수를 갖는다.

$$f_i(Z_{ij}) = pg_i(Z_{ij}) + qh_i(Z_{ij}) \qquad i=1,2,\cdots,L$$
$$j=1,2,\cdots,N_i .$$

그러므로, i번째 층에서의 응답의 평균 및 분산을 구해 보면 다음과 같다.

$$\begin{aligned}\mu_{Zi} &= E_i(Z_{ij}) \\ &= pE_i(X) + qE_i(Y) \\ &= p\mu_i + q\mu_{Yi} \ ,\end{aligned}$$

$$\begin{aligned}\sigma_{Zi}^2 &= E_i(Z_{ij}^2) - [E_i(Z_{ij})]^2 \\ &= [pE_i(X^2) + qE_i(Y^2)] - [pE_i(X) + qE_i(Y)]^2 \\ &= p(\sigma_i^2 + \mu_i^2) + q(\sigma_{Yi}^2 + \mu_{Yi}^2) - (p\mu_i + q\mu_{Yi})^2 \\ &= p\sigma_i^2 + q\sigma_{Yi}^2 + pq(\mu_i - \mu_{Yi})^2 \ .\end{aligned}$$

여기서, μ_i는 i번째 층에서 민감한 그룹에 속하는 모평균이며, σ_i^2은 i번째 층에서의 X의 모분산이고 σ_{Yi}^2은 i번째 층에서의 Y의 모분산이다.

또한, 층화이중추출법에 있어서 민감한 그룹에 속하는 모평균 $\mu_{X(s)}$는 다음과 같다.

$$\mu_{X(s)} = \frac{1}{N}\sum_{i=1}^{L} N_i\mu_i = \sum_{i=1}^{L} W_i\mu_i \ . \tag{2.1}$$

i번째 층에서 단순임의복원추출된 $n_i(n = \sum_{i=1}^{L} n_i)$명의 응답자들이 z_{ij} $(i = 1,2,\cdots,L;$ $j = 1,2,\cdots,n_i)$라고 응답했다고 하자. 그러면, i번째 층에서 민감한 그룹에 속하는 모평균 μ_i의 추정량 $\hat{\mu}_i$는 다음과 같다.

$$\hat{\mu}_i = \frac{\bar{z}_i - q\mu_{Yi}}{p} \ . \tag{2.2}$$

여기서, $\bar{z}_i = \hat{\mu}_{Zi} = \sum_{j=1}^{n_i} z_{ij}/n_i$ 이다.

따라서, 층화이중추출에 있어서 민감한 그룹에 속하는 모평균 $\mu_{X(s)}$의 추정량 $\hat{\mu}_{X(s)}$는 다음과 같다.

$$\begin{aligned}\hat{\mu}_{X(s)} &= \sum_{i=1}^{L} w_i\hat{\mu}_i \\ &= \sum_{i=1}^{L} w_i\left[\frac{\bar{z}_i - q\mu_{Yi}}{p}\right] \\ &= \sum_{i=1}^{L} \frac{n_i'}{n'}\left[\frac{\bar{z}_i - q\mu_{Yi}}{p}\right] \ .\end{aligned} \tag{2.3}$$

정리 1 각 층에 있어서 $\hat{\mu}_i$ 가 μ_i 의 불편추정량이면 $\hat{\mu}_{X(s)}$ 는 모평균 $\mu_{X(s)}$ 의 불편추정량이다.

증명

$$E(\hat{\mu}_{X(s)}) = E_1\left[E_2(\sum_{i=1}^{L} w_i \hat{\mu}_i \mid w_i)\right] = E_1\left[\sum_{i=1}^{L} w_i \mu_i\right] = \sum_{i=1}^{L} W_i \mu_i = \mu_{X(s)} \quad \blacksquare$$

정리 2 서로 다른 층에서 표본을 독립적으로 단순임의복원추출한다면 민감한 그룹에 속하는 모평균 $\mu_{X(s)}$ 의 추정량 $\hat{\mu}_{X(s)}$ 의 분산은 다음과 같다.

$$Var(\hat{\mu}_{X(s)}) = \frac{1}{n'}\left[\sum_{i=1}^{L} W_i \sigma_{Zi}^2 + \sum_{i=1}^{L} W_i(\mu_{Zi} - \mu_{X(s)})^2\right] + \sum_{i=1}^{L} \frac{W_i}{n' v_i} \frac{\sigma_{Zi}^2}{p^2} . \quad (2.4)$$

여기서, $\sigma_{Zi}^2 = p\sigma_i^2 + q\sigma_{Y_i}^2 + pq(\mu_i - \mu_{Y_i})^2$이고, $0 \le v_i = \frac{n_i}{n_i'} \le 1$ 인 고정값이다.

증명

$$Var(\hat{\mu}_{X(s)}) = Var_1[E_2(\hat{\mu}_{X(s)})] + E_1[Var_2(\hat{\mu}_{X(s)})]$$

여기서, $$Var_1[E_2(\hat{\mu}_{X(s)})] = Var_1\left[E_2(\sum_{i=1}^{L} w_i \hat{\mu}_i)\right], \ (E_2(\hat{\mu}_i) = \mu_i')$$

$$= Var_1\left[\sum_{i=1}^{L} w_i \mu_i'\right]$$

$$= Var_1(\mu')$$

$$= \frac{\sigma^2}{n'}$$

이고, σ^2은 다음과 같이 구해진다.

$$N\sigma^2 = \sum_{i=1}^{L} N_i \sigma_{Zi}^2 + \sum_{i=1}^{L} N_i(\mu_{Zi} - \mu_{X(s)})^2 .$$

$$\sigma^2 = \sum_{i=1}^{L} \frac{N_i}{N} \sigma_{Zi}^2 + \sum_{i=1}^{L} \frac{N_i}{N}(\mu_{Zi} - \mu_{X(s)})^2$$

$$= \sum_{i=1}^{L} W_i \sigma_{Zi}^2 + \sum_{i=1}^{L} W_i(\mu_{Zi} - \mu_{X(s)})^2 .$$

따라서,

$$Var_1[E_2(\hat{\mu}_{X(s)})] = \frac{1}{n'}\left[\sum_{i=1}^{L} W_i \sigma_{Zi}^2 + \sum_{i=1}^{L} W_i(\mu_{Zi} - \mu_{X(s)})^2\right] .$$

또한,

$$E_1[Var_2(\hat{\mu}_{X(s)})] = E_1\left[Var_2\left(\sum_{i=1}^{L} w_i \hat{\mu}_i\right)\right]$$

$$= E_1\left[\sum_{i=1}^{L} w_i^2 \, Var_2\left(\frac{\bar{z}_i - q\mu_{Yi}}{p}\right)\right]$$

$$= E_1\left[\sum_{i=1}^{L} \frac{w_i^2 \sigma_{Zi}^2}{n_i p^2}\right]$$

이 되고, 여기서 $n_i = v_i n_i' = v_i w_i n'$ 이므로

$$E_1[Var_2(\hat{\mu}_{X(s)})] = E_1\left[\sum_{i=1}^{L} \frac{w_i^2}{v_i w_i n'} \frac{\sigma_{Zi}^2}{p^2}\right]$$

$$= E_1\left[\sum_{i=1}^{L} \frac{w_i}{n' v_i} \frac{\sigma_{Zi}^2}{p^2}\right]$$

$$= \sum_{i=1}^{L} \frac{W_i}{n' v_i} \frac{\sigma_{Zi}^2}{p^2} \quad .$$

결과적으로, $Var(\hat{\mu}_{X(s)})$ 은 다음과 같다.

$$Var(\hat{\mu}_{X(s)}) = \frac{1}{n'}\left[\sum_{i=1}^{L} W_i \sigma_{Zi}^2 + \sum_{i=1}^{L} W_i (\mu_{Zi} - \mu_{X(s)})^2\right] + \sum_{i=1}^{L} \frac{W_i}{n' v_i} \frac{\sigma_{Zi}^2}{p^2} \quad \blacksquare$$

$\hat{\mu}_{X(s)}$ 의 분산의 불편추정량은 다음과 같다.

$$\widehat{Var}(\hat{\mu}_{X(s)}) = \frac{1}{n'}\left[\sum_{i=1}^{L} w_i s_{Zi}^2 + \sum_{i=1}^{L} w_i (\hat{\mu}_{Zi} - \hat{\mu}_{X(s)})^2\right] + \sum_{i=1}^{L} \frac{w_i}{n' v_i} \frac{s_{Zi}^2}{p^2}$$

여기서, $s_{Zi}^2 = \sum_{j=1}^{n_i} (z_{ij} - \bar{z}_i)^2/(n_i - 1)$ 이다.

다음은 층화추출에 있어서 표본의 크기 n을 각 층에 배분하는 표본배분에 대해 살펴보자. 비례배분은 $N_i/N = n_i/n$을 만족하도록 각 층에 표본 n_i를 배분하는 것을 의미한다. 이 때, 모집단의 크기 N과 N_i 대신에 1단계 표본 n' 와 n_i' 를 이용하면, 비례배분에 있어서 i층의 표본크기는 $n_i = n \cdot n_i'/n'$ 이다. 따라서, 추정량 $\hat{\mu}_{X(s)}$ 의 분산은 다음과 같다.

$$Var_{(P)}(\hat{\mu}_{X(s)}) = \frac{1}{n'}\left[\sum_{i=1}^{L} W_i \sigma_{Zi}^2 + \sum_{i=1}^{L} W_i (\mu_{Zi} - \mu_{X(s)})^2\right] + \sum_{i=1}^{L} \frac{W_i \sigma_{Zi}^2}{np^2} \quad . \qquad (2.5)$$

다음으로 최적배분은 표본의 크기 n을 각 층에 배분하는데 있어서 일정한 비용 하에서 $Var(\hat{\mu}_{X(s)})$ 을 최소로 하거나, 일정한 분산 $Var(\hat{\mu}_{X(s)})$ 하에서 비용을 최소로 하는 각 층의 표본크기를 결정하는 방법이다. 일정한 비용 하에서 $Var(\hat{\mu}_{X(s)})$ 을 최소화하기 위해 n'

와 v_i의 최적값을 구한다. 이 때, 다음과 같은 비용함수를 사용한다.

$$C = c' n' + \sum_{i=1}^{L} c_i n_i \quad . \tag{2.6}$$

여기서, c' 는 단위당 층 분류비용이고, c_i는 i층내의 표본조사 단위의 조사비용이다.

n_i는 확률변수이므로 n' 와 v_i의 최적값을 구하기 위하여 식(2.6)의 기대값을 최소화하여야 한다. C의 기대값은 다음과 같이 구할 수 있다.

$$\begin{aligned} E(C) = C^* &= c' n' + \sum_{i=1}^{L} c_i E(n_i) \\ &= c' n' + \sum_{i=1}^{L} c_i E[E(n_i | n_i')] \\ &= c' n' + \sum_{i=1}^{L} c_i E(n_i' v_i) \\ &= c' n' + \sum_{i=1}^{L} c_i E(n' v_i w_i) \\ &= c' n' + n' \sum_{i=1}^{L} c_i v_i W_i \, . \end{aligned} \tag{2.7}$$

그리고, 식(2.4)에서

$$\sigma^2 = \sum_{i=1}^{L} W_i \sigma_{Zi}^2 + \sum_{i=1}^{L} W_i (\mu_{Zi} - \mu_{X(s)})^2$$

로 두면, $\hat{\mu}_{X(s)}$ 의 분산은 다음과 같이 된다.

$$Var(\hat{\mu}_{X(s)}) = \frac{\sigma^2}{n'} + \sum_{i=1}^{L} \frac{W_i \sigma_{Zi}^2}{n' v_i p^2} \, . \tag{2.8}$$

따라서, 기대비용과 분산의 곱은 다음과 같다.

$$\left[c' n' + n' \sum_{i=1}^{L} c_i v_i W_i \right] \left[\frac{\sigma^2}{n'} + \sum_{i=1}^{L} \frac{W_i \sigma_{Zi}^2}{n' v_i p^2} \right] . \tag{2.9}$$

식(2.9)를 최소로 하는 최적값 v_i를 구하기 위해

$$a_1 = (c' n')^{1/2} \quad , \quad a_2 = \left[n' \sum_{i=1}^{L} c_i v_i W_i \right]^{1/2} \quad ,$$
$$b_1 = \left[\frac{\sigma^2}{n'} \right]^{1/2} \quad , \quad b_2 = \left[\sum_{i=1}^{L} \frac{W_i \sigma_{Zi}^2}{n' v_i p^2} \right]^{1/2}$$

라 두고, 다음과 같은 Cauchy-Schwarz의 부등식을 적용한다.

$$\left(\sum_{i=1}^{2} a_i^2 \right) \left(\sum_{i=1}^{2} b_i^2 \right) \geq \left(\sum_{i=1}^{2} a_i b_i \right)^2 \quad .$$

이 때, 앞의 등식이 성립하기 위한 필요충분조건인 $a_1/b_1 = a_2/b_2 =$ 상수를 모든 층 i에 대

하여 적용하면 다음과 같은 결과를 얻는다.

$$\frac{c' n'^2}{\sigma^2} = \frac{c_i v_i^2 p^2 n'^2}{\sigma_{Zi}^2} \quad .$$

따라서, 윗 식을 정리하면 최적값 v_i를 다음과 같이 얻을 수 있다.

$$v_i = \left[\frac{c'}{c_i} \frac{\sigma_{Zi}^2}{p^2 \sigma^2} \right]^{1/2} \quad . \tag{2.10}$$

그리고, v_i를 식(2.7)에 대입하여 n' 의 최적값을 구한다.

$$n' = \frac{C^*}{c' + \sum_{i=1}^{L} c_i W_i \left[\frac{c'}{c_i} \frac{\sigma_{Zi}^2}{p^2 \sigma^2} \right]^{1/2}} \quad . \tag{2.11}$$

따라서, 식(2.10)과 식(2.11)에서 구해진 v_i와 n' 의 최적값을 식(2.8)에 대입하면 다음과 같은 $\hat{\mu}_{X(s)}$ 의 최소분산을 얻을 수 있다.

$$Var_{(O)}(\hat{\mu}_{X(s)}) = \frac{1}{C^*} \left[\sqrt{c'}\sigma + \frac{1}{p} \sum_{i=1}^{L} W_i \sigma_{Zi} \sqrt{c_i} \right]^2 \quad . \tag{2.12}$$

3. 효율성 비교

이 장에서는 층화추출법을 양적인 무관질문모형에 적용하는데 있어서, 층의 크기를 알고 있는 경우와 모르고 있는 경우의 추정량의 분산을 이용하여 효율성을 비교하고자 한다.

3.1 층화임의추출법의 비교

층의 크기를 알고 있는 경우 층화임의추출법에 의한 추정량 $\hat{\mu}_X$ 의 분산은

$$Var(\hat{\mu}_X) = \sum_{i=1}^{L} \frac{W_i^2 \sigma_{Zi}^2}{n_i p^2}$$

이고, 층의 크기를 모르고 있는 경우 층화이중임의추출법을 이용하여 얻은 추정량 $\hat{\mu}_{X(s)}$ 의 분산은 다음과 같다.

$$Var(\hat{\mu}_{X(s)}) = \frac{1}{n'} \left[\sum_{i=1}^{L} W_i \sigma_{Zi}^2 + \sum_{i=1}^{L} W_i (\mu_{Zi} - \mu_{X(s)})^2 \right] + \sum_{i=1}^{L} \frac{W_i}{n' v_i} \frac{\sigma_{Zi}^2}{p^2} \quad .$$

따라서, 층화임의추출을 하는데 있어서 층의 크기를 모르는 경우 층화표본을 얻기 위해 생기는 분산의 증가량이 다음과 같음을 알 수 있다.

$$\frac{1}{n'}\left[\sum_{i=1}^{L} W_i\,\sigma_{Zi}^2 + \sum_{i=1}^{L} W_i(\mu_{Zi} - \mu_{X(s)})^2\right] + \frac{1}{p^2}\sum_{i=1}^{L} W_i\,\sigma_{Zi}^2\left[\frac{1}{n' v_i} - \frac{W_i}{n_i}\right] .$$

층이 2인 경우에, 효율성을 비교하기 위하여 $n = 100$, $n_1 = 60$, $n_2 = 40$, $W_1 = 0.7$, $W_2 = 0.3$, $\mu_1 = 5$, $\mu_2 = 10$, $\mu_{Y1} = \mu_{Y2} = 10$, $\sigma_1^2 = 10$, $\sigma_2^2 = 15$, $\sigma_{Y1}^2 = \sigma_{Y2}^2 = 10$ 이라고 가정하고, p와 n' 를 변화시키면서 분산비 $Var(\hat{\mu}_{X(s)})/Var(\hat{\mu}_X)$ 를 계산하여 <표 1>을 작성하였다.

<표 1> 층의 크기를 알고 있는 경우와 층의 크기를 모르고 있는 경우의 효율성 비교

p	n' = 500	n' = 1000	n' = 5000
0.1	1.00	1.00	1.00
0.2	1.01	1.00	1.00
0.3	1.02	1.01	1.00
0.4	1.03	1.01	1.00
0.5	1.05	1.02	1.00
0.6	1.08	1.04	1.00
0.7	1.11	1.05	1.01
0.8	1.15	1.07	1.01
0.9	1.20	1.10	1.02

<표 1>에서 1보다 큰 값은 층의 크기를 알고 있는 경우가 층의 크기를 모르고 있는 경우보다 더 효율적임을 나타낸다. 즉, 층의 크기를 알고 있는 경우가 층의 크기를 모르고 있는 경우보다 효율적임을 알 수 있다. 그러나, 1단계 표본 n' 가 대표본으로 추출되고, p값이 작을수록 두 경우의 효율성이 비슷하게 됨을 알 수 있다.

3.2 층화비례추출법의 비교

먼저, 층의 크기를 알고 있는 경우 층화비례추출법에 의한 추정량 $\hat{\mu}_X$ 의 분산은

$$Var_{(P)}(\hat{\mu}_X) = \sum_{i=1}^{L} \frac{W_i\,\sigma_{Zi}^2}{np^2}$$

이고, 층의 크기를 모르고 있는 경우 층화이중 비례추출법을 이용하여 얻은 추정량 $\hat{\mu}_{X(s)}$ 의 분산은 다음과 같다.

$$Var_{(P)}(\hat{\mu}_{X(s)}) = \frac{1}{n'}\left[\sum_{i=1}^{L} W_i\,\sigma_{Zi}^2 + \sum_{i=1}^{L} W_i(\mu_{Zi} - \mu_{X(s)})^2\right] + \sum_{i=1}^{L} \frac{W_i\,\sigma_{Zi}^2}{np^2} .$$

따라서, 층화비례추출을 하는데 있어서 층의 크기를 모르는 경우 층화표본을 얻기 위해 생기

는 분산의 증가량이 다음과 같음을 알 수 있다.

$$\frac{1}{n'}\left[\sum_{i=1}^{L} W_i \sigma_{Zi}^2 + \sum_{i=1}^{L} W_i(\mu_{Zi} - \mu_{X(s)})^2\right] .$$

3.3 층화최적추출법의 비교

층의 크기를 알고 있는 경우 층화최적추출법에 의한 추정량 $\hat{\mu}_X$ 의 분산은

$$Var_{(O)}(\hat{\mu}_X) = \sum_{i=1}^{L} \frac{W_i \sigma_{Zi} \sqrt{c_i}}{np^2} \cdot \sum_{i=1}^{L} \frac{W_i \sigma_{Zi}}{\sqrt{c_i}}$$

이고, 층의 크기를 모르고 있는 경우 층화이중 최적추출법을 이용하여 얻은 추정량 $\hat{\mu}_{X(s)}$ 의 분산은 다음과 같다.

$$Var_{(O)}(\hat{\mu}_{X(s)}) = \frac{1}{C^*}\left[\sqrt{c'}\sigma + \frac{1}{p}\sum_{i=1}^{L} W_i \sigma_{Zi} \sqrt{c_i}\right]^2 .$$

따라서, 층화최적추출을 하는데 있어서 층의 크기를 모르는 경우 층화표본을 얻기 위해 생기는 분산의 증가량이 다음과 같음을 알 수 있다.

$$\frac{1}{C^*}\left[\sqrt{c'}\sigma + \frac{1}{p}\sum_{i=1}^{L} W_i \sigma_{Zi} \sqrt{c_i}\right]^2 - \sum_{i=1}^{L} \frac{W_i \sigma_{Zi} \sqrt{c_i}}{np^2} \cdot \sum_{i=1}^{L} \frac{W_i \sigma_{Zi}}{\sqrt{c_i}} .$$

층이 2인 경우에, 효율성을 비교하기 위하여 $n = 100$, $n_1 = 60$, $n_2 = 40$, $W_1 = 0.7$, $W_2 = 0.3$, $\mu_1 = 5$, $\mu_2 = 10$, $\mu_{Y1} = \mu_{Y2} = 10$, $\sigma_1^2 = 10$, $\sigma_2^2 = 15$, $\sigma_{Y1}^2 = \sigma_{Y2}^2 = 10$, $c_1 = c_2 = 100$, $c' = 0.01$ 이라고 가정하고, p와 n' 를 변화시키면서 분산비 $Var_{(O)}(\hat{\mu}_{X(s)})/Var_{(O)}(\hat{\mu}_X)$를 계산하여 <표 2>를 작성하였다.

<표 2> 층의 크기를 알고 있는 경우와 층의 크기를 모르고 있는 경우의 효율성 비교

p	$n' = 500$	$n' = 1000$	$n' = 5000$
0.1	1.00	1.00	1.00
0.2	1.00	1.00	1.00
0.3	1.01	1.01	1.00
0.4	1.01	1.01	1.00
0.5	1.01	1.01	1.01
0.6	1.01	1.01	1.01
0.7	1.01	1.01	1.01
0.8	1.02	1.02	1.01
0.9	1.02	1.02	1.01

<표 2>에서 1보다 큰 값은 층의 크기를 알고 있는 경우가 층의 크기를 모르고 있는 경우보다 더 효율적임을 나타낸다. 여기에서도, 1단계 표본 n' 가 대표본으로 추출되고, p값이 작을수록 층의 크기를 알고 있는 경우와 층의 크기를 모르고 있는 경우의 효율성이 비슷해진다는 사실은 층화임의추출법의 경우와 같다. 그리고, i층내의 표본조사 단위의 조사비용인 $c_i(i=1,2)$와 단위당 층 분류비용인 c' 에 따라 효율성의 변화가 나타남을 알 수 있었다.

4. 결 론

본 논문에서는 사회적으로나 개인적으로 매우 민감한 조사에서 모집단이 양적속성을 갖는 여러 개의 층으로 구성되어 있을 때 층의 크기를 모르는 경우 층화표본을 위하여 이중추출법을 이용하는 층화이중추출법에 의한 양적속성의 무관질문모형을 제안하였다. 그리고, 층화이중추출에 있어서 각 층의 표본배분에 관해 비례배분, 최적배분으로 나누어 살펴보았다. 또한, 층의 크기를 알고 있는 경우와 층의 크기를 모르고 있는 경우의 효율성을 비교해 본 결과 층의 크기를 알고 있는 경우가 더 효율적임을 알 수 있으나, 1단계 표본 n' 가 대표본으로 추출되고, p값이 작을수록 효율성이 서로 비슷하게 됨을 알 수 있었다. 이는 비록 각 층의 크기를 모른다 하더라도, 조사에서 응답자 개인이 민감한 질문에 노출될 확률을 작게 해 주는 대신 1단계 표본을 크게 해줌으로써 각 층의 크기를 알고 있는 경우와 비슷한 효율을 이룰 수 있음을 나타낸다고 볼 수 있다. 또한, 층화최적추출법에서는 비용에 따라 효율성의 차이가 나타남을 알 수 있었다.

참고문헌

[1] 류제복, 홍기학, 이기성 (1993). 「확률화응답모형」, 자유아카데미, 서울.
[2] 박홍래 (1989). 「통계조사론」, 영지문화사, 서울.
[3] 김종호, 류제복, 이기성 (1993). 층화이중추출법을 이용한 확률화응답모형, 「춘계학술논문발표」, 한국통계학회.
[4] 김종호, 홍기학, 이기성 (1993). 층화추출법에 의한 양적속성의 확률화응답모형, 「동국 논총」, 제 32집 자연과학편, 87-99.
[5] 홍기학, 염준근, 이화영 (1994). 층화 확률화응답기법, 「응용통계연구」, 제 7권 1호, 141-147.
[6] 이기성 (1992). 2단계 확률화응답모형에 관한 연구, 「박사학위논문」, 동국대학교.
[7] 홍기학 (1991). 확률화응답모형에 관한 연구, 「박사학위논문」, 동국대학교.
[8] Chaudhuri, A. and Mukerjee, R. (1988). Randomized Response : *Theory and Techniques,* Marcel Dekker, Inc., New York.
[9] Cochran, W. G. (1977). *Sampling Techniques,* 3rd ed. John Wiley and Sons, New York.
[10] Greenberg, B. G., Kubler, R. R., Abernathy, J. R., and Horvitz, D. G. (1971).

Applications of the RR Technique in Obtaining Quantitative Data, *Journal of the American Statistical Association,* Vol. 66, 243-250.

[11] Warner, S. L. (1965). Randomized Response ; A Survey Technique for Eliminating Evasive Answer Bias, *Journal of the American Statistical Association,* Vol. 60, 63-69.

Unrelated Question Model with Quantitative Attribute by Stratified Double Sampling

Gi-Sung Lee[3], Ki-Hak Hong[4]

Abstract

In the surveys of sensitive issues of the population that is composed of several unknown-size stratum, we propose the unrelated question model with quantitative attribute by using stratified double sampling.

And, we consider two types of sample allocations under the fixed cost, which are the proportional allocation, the optimum allocation.

In efficiency, the proposed model is inferior to the unrelated question model with quantitative attribute by stratified sampling in case of the size of each stratum is known. But we find that efficiency of the proposed model is increased, when the selecting probability of sensitive question p is small and first stage sample size n' is large.

3) Department of Computer Science & Statistics, Cheonju Woosuk University, Wanju-gun, Chonbuk, 565-800, KOREA.

4) Department of Computer Science & Statistics, Dongshin University, 252, Daeho-dong, Naju, Chonnam, 520-714, KOREA.