Journal of Korean
Data & Information Science Society
2003, Vol. 14, No.1 pp. 55~65

The Three-Stage Cluster Unrelated Question Model[1)]

Seung-Chul Ahn[2)] · Gi-Sung Lee[3)]

Abstract

In this study, we systemize the theoretical validity for applying unrelated question model to three-stage cluster sampling method and derive the estimate and it's variance of sensitive parameter. We derive the minimum variance form under the optimal values of the subsample sizes when the cost are fixed. Under the some given precision, we obtain the optimal values of the subsample sizes and derive the minimum cost form by using them.

Keywords : 민감한 정보, 무관질문모형, 3단계 집락추출법

1. 서론

음주운전, 낙태경험, 환각제사용, 동성연애 및 탈세여부 등과 같은 사회적으로나 개인적으로 매우 민감한 문제에 관한 조사에서 기존의 직접질문방식을 그대로 사용할 경우 응답자들이 응답을 회피하거나 거짓으로 응답하는 경향이 있다. 이는 응답자들이 민감한 질문에 응답함으로써 불이익을 받거나 사생활이 보장되지 않는다고 생각하기 때문이다. 민감한 질문에 대한 조사에서 발생하는 응답편향을 줄이기 위하여 1965년 Warner는 응답자들에게 직접적인 응답을 요구하는 것이 아니라 확률장치를 통한 간접적인 응답만을 요구함으로 응답자들의 신분을 보장해 주는 획기적인 확률화응답모형(randomized response model)을 제시하였다. 이 때, Warner는 민감한 질문과, 민감한 질문과 배반되는 질문에 대해 확률장치를 사용하는 관련질문모형(related question model)을 제안하여 민감한 속성에 대한 정보를 얻었다. 그리고, Greenberg et al.(1969)은 민감한 질문과 배반되는 질문 대신에 민감한 질문과 전혀 관계가 없는 질문을 사용하는 무관질문모형(unrelated question model)을 제안하여 그 이론적 체계를 구축하였다. 그 후 미국, 캐나다, 영국, 호주 등 서구 여러 나라와 일본, 인도 등 몇몇 아시아 국가에서도 이 분야에 대한 연구가 활발히 진행되고 있다. 특히, Fox와

1) 이 논문은 우석대학교 교내학술연구비 지원에 의하여 연구됨.
2) 전북 완주군 삼례읍 후정리 490 우석대학교 전산정보학부 부교수
E-mail : scahn@woosuk.ac.kr
3) 전북 완주군 삼례읍 후정리 490 우석대학교 전산정보학부 부교수

Tracy(1986), Chaudhuri와 Mukerjee(1988)는 확률화응답모형들을 정리, 요약하여 체계화하였으며, 류제복, 홍기학과 이기성(1993)들이 확률화응답모형에 관한 책을 출간하여 그 중요성을 강조하였다. 그러나, 이러한 확률화응답모형들은 직접 질문을 사용하는 경우보다 시간·비용과 노력을 더 필요로 하며, 특히 모집단이 큰 경우에 단순임의추출법을 이용하여 응답자들을 추출하여 조사를 하는 데에는 여러 가지 어려움이 따르게 된다. 이러한 문제점을 해결하기 위하여 이기성과 홍기학(1998)은 매우 민감한 조사에서 모집단이 여러 개의 집락으로 구성되어 있을 때, 집락표본추출의 경제성을 유지하면서 효율을 높이는 방법으로 추출된 표본 집락 내에서 모든 조사단위를 조사하는 대신에 각 표본 집락 내에서 다시 조사단위를 추출하는 2단계 집락추출법을 확률화응답모형에 적용하였다. 하지만, 대부분의 실제조사에서는 2단계 이상 3단계 집락추출법을 요구하고 있는 경우가 많이 있다.

이에 본 연구에서는 매우 민감한 조사에서 모집단이 여러 개의 집락으로 구성되어 있을 때, 3단계 집락추출법을 무관질문모형에 적용한 3단계 집락 무관질문모형을 제안하고자 한다. 첫째, 제안한 3단계 집락 무관질문모형에서 민감한 속성에 대한 모수의 추정치와 분산 및 분산추정량을 구하고자 한다. 둘째, 일정한 비용 하에서 분산을 최소로 하는 1차 추출단위와 2차 추출단위 및 3차 추출단위에 대한 최적값을 구하여 최소분산의 형태를 도출하고, 셋째, 일정한 정도 하에서 비용함수를 최소로 하는 1차 추출단위와 2차 추출단위 및 3차 추출단위의 최적값을 구하여 최소비용의 형태를 구해보고자 한다.

2. 3단계 집락 무관질문모형

이 장에서는 3단계 표본추출에 있어서 모집단이 민감한 속성을 가지고 있는 집락으로 구성되어 있을 때, 3단계 집락추출법에 무관질문모형을 적용하여 민감한 속성에 대한 정보를 얻을 수 있을 수 있는 3단계 집락 무관질문모형을 제안하고자 한다.

3단계 표본추출에 있어서 모집단은 N개의 1차 추출단위(psu)가 있고, 각 단위는 M개의 2차 추출단위(ssu)로 구성되고, 각 2차 추출단위는 K개의 3차 추출단위(tsu)로 구성되어 있다고 하자. 이에 대응되는 표본단위의 개수는 각각 n, m, k라 하자.

3단계 집락추출법에 의해 추출된 응답자들은 다음과 같은 두 개의 설문으로 구성되어 있는 무관질문모형의 확률장치에 의해서 선택된 설문에 대해 “예” 또는 “아니오”라고 응답한다.

설문 1 : 당신은 민감한 그룹 A에 속합니까?
설문 2 : 당신은 무관한 그룹 Y에 속합니까?

여기서, 설문 1이 선택될 확률은 p이고, 설문 2가 선택될 확률은 $1-p$이다.

예를 들어, 청소년들의 흡연실태를 조사하기 위하여 다음과 같은 설문으로 구성된 무관질문모형을 생각해 볼 수 있다.

설문 1 : 당신은 흡연을 습관적으로 합니까?
설문 2 : 당신은 야구를 좋아합니까?

위와 같은 예에서 청소년들이 야구를 좋아하는 모비율을 기존의 비슷한 조사나 사전조사에 의해 알고 있는 경우에는 무관한 그룹 Y에 속하는 모비율로 알고 있다고 가정할 수 있으나, 무관한 속성에 관한 정보가 전혀 없을 경우에는 두 개의 표본을 이용하는 이표본 무관질문모형(two-sample unrelated question model)을 고려해 볼 수 있다. 여기서는 무관한 속성의 모비율을 알고 있는 경우에 대하여 다루어 보고자 한다.

따라서, $i(i=1,2,\cdots,N)$번째 1차 추출단위 내 $j(j=1,2,\cdots,M)$번째 2차 추출단위 내에서 $l(l=1,2,\cdots,K)$번째 응답자가 "예"라고 응답할 확률을 구해 보면 다음과 같다.

$$\lambda_{ij}=p\pi_{ij}+(1-p)\pi_{y_{ij}}. \tag{2.1}$$

여기서, π_{ij}는 i번째 1차 추출단위 내 j번째 2차 추출단위 당 민감한 그룹에 속하는 비율이며, $\pi_{y_{ij}}$는 i번째 1차 추출단위 내 j번째 2차 추출단위 당 무관한 그룹 Y에 속하는 모비율이다.

$i(i=1,2,\cdots,n)$번째 1차 추출단위 내 $j(j=1,2,\cdots,m)$번째 2차 추출단위 내 $l(l=1,2,\cdots,k)$번째 추출단위의 관찰치를 z_{ijl}이라 하고, 최종단위인 응답자가 "예"라고 응답하면 $z_{ijl}=1$, "아니오"라고 응답하면 $z_{ijl}=0$이라고 정의하자.
이 때, i번째 1차 추출단위 내 j번째 2차 추출단위 내 l번째 3차 추출단위로 추출된 응답자들 중에서 "예"라고 응답한 사람의 수를 $Z_{ij}=\sum_{l=1}^{k} z_{ijl}$이라 하면, $\hat{\lambda}_{ij}=\frac{Z_{ij}}{k}$ 이 되므로, 식(2.1)로부터 π_{ij}의 추정량 $\hat{\pi}_{ij}$와 그 분산은 다음과 같다.

$$\hat{\pi}_{ij}=\frac{\hat{\lambda}_{ij}-(1-p)\pi_{y_{ij}}}{p}, \tag{2.2}$$

$$\begin{aligned} V(\hat{\pi}_{ij}) &= \frac{\lambda_{ij}(1-\lambda_{ij})}{kp^2} \\ &= \frac{\{p\pi_{ij}+(1-p)\pi_{y_{ij}}\}\{1-p\pi_{ij}-(1-p)\pi_{y_{ij}}\}}{kp^2} \\ &= \frac{\pi_{ij}(1-\pi_{ij})}{k}+\frac{(1-p)\{p\pi_{ij}(1-2\pi_{y_{ij}})-(1-p)\ \pi_{y_{ij}}{}^2+\pi_{y_{ij}}\}}{kp^2}. \end{aligned} \tag{2.3}$$

3단계 집락추출법에 있어서 민감한 그룹에 속하는 3차 추출단위 당 모비율 π는 다음과 같다.

$$\pi = \frac{1}{N}\sum_{i=1}^{N}\pi_i . \tag{2.4}$$

이 때, π_i는 i번째 1차 추출단위 당 민감한 그룹에 속하는 모비율이며, 다음과 같이 표현된다.

$$\pi_i = \frac{1}{M}\sum_{j=1}^{M}\pi_{ij} . \tag{2.5}$$

$i(i=1, 2, \cdots, n)$번째 1차 추출단위 내 $j(j=1, 2, \cdots, m)$번째 2차 추출단위에서 민감한 그룹에 속하는 3차 추출단위 당 모비율 π의 추정량 $\hat{\pi}$는 다음과 같이 정의할 수 있다.

$$\hat{\pi} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\hat{\pi}_{ij} . \tag{2.6}$$

이 때, 추정량 $\hat{\pi}$는

$$\begin{aligned} E(\hat{\pi}) &= E_1E_2E_3\left(\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\hat{\pi}_{ij}\right) \\ &= E_1E_2\left(\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\pi_{ij}\right) \\ &= E_1\left(\frac{1}{n}\sum_{i=1}^{n}\pi_i\right) \\ &= \pi . \end{aligned}$$

이므로, 모비율 π의 비편향추정량이다.

<정리 1> 3단계 추출에 있어서 각 단계마다 단순임의비복원추출을 한다면, 민감한 그룹에 속하는 3차 추출단위 당 모비율 π의 추정량 $\hat{\pi}$의 분산은 다음과 같다.

$$\begin{aligned} V(\hat{\pi}) &= \frac{N-n}{N(N-1)n}\sum_{i=1}^{N}(\pi_i-\pi)^2 + \frac{M-m}{NM(M-1)nm}\sum_{i=1}^{N}\sum_{j=1}^{M}(\pi_{ij}-\pi_i)^2 \\ &\quad + \frac{K-k}{NMKnmk}\sum_{i=1}^{N}\sum_{j=1}^{M}\left[\pi_{ij}(1-\pi_{ij}) + \frac{(1-p)\{p\pi_{ij}(1-2\pi_{y_{ij}}) - (1-p)\ \pi_{y_{ij}}{}^2 + \pi_{y_{ij}}\}}{p^2}\right]. \end{aligned} \tag{2.7}$$

(증명)

$$V(\hat{\pi}) = V_1E_2E_3(\hat{\pi}) + E_1V_2E_3(\hat{\pi}) + E_1E_2V_3(\hat{\pi})$$

에서

$$\begin{aligned} V_1E_2E_3(\hat{\pi}) &= V_1E_2\left(\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\pi_{ij}\right) \\ &= V_1\left(\frac{1}{n}\sum_{i=1}^{n}\pi_i\right) \\ &= \frac{N-n}{N(N-1)n}\sum_{i=1}^{N}(\pi_i-\pi)^2 \end{aligned}$$

이고,

$$\begin{aligned} E_1V_2E_3(\hat{\pi}) &= E_1V_2\left(\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\pi_{ij}\right) \\ &= E_1\left[\frac{M-m}{M(M-1)n^2m}\sum_{i=1}^{n}\sum_{j=1}^{M}(\pi_{ij}-\pi_i)^2\right] \\ &= \frac{M-m}{NM(M-1)nm}\sum_{i=1}^{N}\sum_{j=1}^{M}(\pi_{ij}-\pi_i)^2 \end{aligned}$$

이며,

$$\begin{aligned} &E_1E_2V_3(\hat{\pi}) \\ &= E_1E_2\left[\frac{K-k}{K(nm)^2k}\sum_{i=1}^{n}\sum_{j=1}^{m}\left\{\pi_{ij}(1-\pi_{ij})+\frac{(1-p)\{p\pi_{ij}(1-2\pi_{y_{ij}})-(1-p)\ \pi_{y_{ij}}{}^2+\pi_{y_{ij}}\}}{p^2}\right\}\right] \\ &= E_1\left[\frac{K-k}{MKn^2mk}\sum_{i=1}^{n}\sum_{j=1}^{M}\left\{\pi_{ij}(1-\pi_{ij})+\frac{(1-p)\{p\pi_{ij}(1-2\pi_{y_{ij}})-(1-p)\ \pi_{y_{ij}}{}^2+\pi_{y_{ij}}\}}{p^2}\right\}\right] \\ &= \frac{K-k}{NMKnmk}\sum_{i=1}^{N}\sum_{j=1}^{M}\left[\pi_{ij}(1-\pi_{ij})+\frac{(1-p)\{p\pi_{ij}(1-2\pi_{y_{ij}})-(1-p)\ \pi_{y_{ij}}{}^2+\pi_{y_{ij}}\}}{p^2}\right] \end{aligned}$$

이므로, 추정량 $\hat{\pi}$의 분산은 식(2.7)과 같다.

<정리 2> $V(\hat{\pi})$의 비편향추정량은 다음과 같다.

$$\begin{aligned} \hat{V}(\hat{\pi}) &= \frac{N-n}{Nn(n-1)}\sum_{i=1}^{n}(\hat{\pi}_i-\hat{\pi})^2+\frac{M-m}{NMnm(m-1)}\sum_{i=1}^{n}\sum_{j=1}^{m}(\hat{\pi}_{ij}-\hat{\pi}_i)^2 \\ &\quad+\frac{K-k}{NMKnmk}\sum_{i=1}^{n}\sum_{j=1}^{m}\left[\hat{\pi}_{ij}(1-\hat{\pi}_{ij})+\frac{(1-p)\{p\hat{\pi}_{ij}(1-2\pi_{y_{ij}})-(1-p)\ \pi_{y_{ij}}{}^2+\pi_{y_{ij}}\}}{p^2}\right]. \end{aligned} \tag{2.8}$$

(증명)

우선, $E\left[\frac{1}{n-1}\sum_{i=1}^{n}(\hat{\pi}_i-\hat{\pi})^2\right]$이 다음과 같이 됨을 보이기로 하자.

$$E\left[\frac{1}{n-1}\sum_{i=1}^{n}(\hat{\pi}_i-\hat{\pi})^2\right]=\frac{1}{N-1}\sum_{i=1}^{N}(\pi_i-\pi)^2+\frac{M-m}{NM(M-1)m}\sum_{i=1}^{N}\sum_{j=1}^{M}(\pi_{ij}-\pi_i)^2$$
$$+\frac{K-k}{NMKmk}\sum_{i=1}^{N}\sum_{j=1}^{M}\left[\pi_{ij}(1-\pi_{ij})+\frac{(1-p)\{p\pi_{ij}(1-2\pi_{y_{ij}})-(1-p)\ \pi_{y_{ij}}{}^2+\pi_{y_{ij}}\}}{p^2}\right]. \quad (2.9)$$

$E\left[\frac{1}{n-1}\sum_{i=1}^{n}(\hat{\pi}_i-\hat{\pi})^2\right]$을 증명하기 위하여 변수 $\hat{\pi}_{iK}$를 i번째 1차 추출단위내 m개 2차 추출단위의 민감한 속성에 대한 표본비율이라 하자. 이 때, $\hat{\pi}_{iK}$는 2차 추출단위내의 모든 K개 3차 추출단위를 관찰하게 되므로 $\hat{\pi}_{iK}=\frac{1}{m}\sum_{j=1}^{m}\pi_{ij}$이다. 그리고, $\hat{\pi}_K$는 n개의 $\hat{\pi}_{iK}$의 평균 즉, $\hat{\pi}_K=\frac{1}{n}\sum_{i=1}^{n}\hat{\pi}_{iK}$이다.

2단계 추출로부터 다음 관계가 성립됨을 보일 수 있다.

$$E\left[\frac{1}{n-1}\sum_{i=1}^{n}(\hat{\pi}_{iK}-\hat{\pi}_K)^2\right]=\frac{1}{N-1}\sum_{i=1}^{N}(\pi_i-\pi)^2+\frac{M-m}{NM(M-1)m}\sum_{i=1}^{N}\sum_{j=1}^{M}(\pi_{ij}-\pi_i)^2. \quad (2.10)$$

$\hat{\pi}_i$를 i번째 1차 추출단위의 민감한 속성에 대한 표본비율이라 하면 다음과 같이 쓸 수 있다.

$$(\hat{\pi}_i-\hat{\pi})=(\hat{\pi}_{iK}-\hat{\pi}_K)+[(\hat{\pi}_i-\hat{\pi}_{iK})-(\hat{\pi}-\hat{\pi}_K)]. \quad (2.11)$$

식(2.11)의 오른쪽 둘째항의 제곱합으로부터 기대값을 다음과 같이 구할 수 있다.

$$E\left[\frac{1}{n-1}\sum_{i=1}^{n}\{(\hat{\pi}_i-\hat{\pi}_{iK})-(\hat{\pi}-\hat{\pi}_K)\}^2\right]$$
$$=\frac{n}{n-1}[E(\hat{\pi}_i-\hat{\pi}_{iK})^2-E(\hat{\pi}-\hat{\pi}_K)^2]$$
$$=\frac{n}{n-1}\left[E_1E_2V_3\left(\frac{1}{m}\sum_{j=1}^{m}\hat{\pi}_{ij}\right)-E_1E_2V_3\left(\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\hat{\pi}_{ij}\right)\right]$$

$$= \frac{n}{n-1}\left[\frac{K-k}{NMKmk} \sum_{i=1}^{N}\sum_{j=1}^{M}\left\{ \pi_{ij}(1-\pi_{ij}) + \frac{(1-p)\{p\pi_{ij}(1-2\pi_{y_{ij}}) - (1-p)\ \pi_{y_{ij}}{}^2 + \pi_{y_{ij}}\}}{p^2} \right\} \right.$$
$$\left. - \frac{K-k}{NMKnmk} \sum_{i=1}^{N}\sum_{j=1}^{M}\left\{ \pi_{ij}(1-\pi_{ij}) + \frac{(1-p)\{p\pi_{ij}(1-2\pi_{y_{ij}}) - (1-p)\ \pi_{y_{ij}}{}^2 + \pi_{y_{ij}}\}}{p^2} \right\} \right]$$
$$= \frac{K-k}{NMKmk} \sum_{i=1}^{N}\sum_{j=1}^{M}\left[\pi_{ij}(1-\pi_{ij}) + \frac{(1-p)\{p\pi_{ij}(1-2\pi_{y_{ij}}) - (1-p)\ \pi_{y_{ij}}{}^2 + \pi_{y_{ij}}\}}{p^2} \right]. \tag{2.12}$$

따라서, 식(2.10)과 식(2.12)를 함께 고려하면, $E\left[\frac{1}{n-1} \sum_{i=1}^{n}(\hat{\pi}_i - \hat{\pi})^2 \right]$의 식(2.9)를 얻을 수 있다.

마찬가지 방법으로

$$E\left[\frac{1}{n(m-1)} \sum_{i=1}^{n}\sum_{j=1}^{m}(\hat{\pi}_{ij} - \hat{\pi}_i)^2 \right]$$
$$= \frac{1}{N(M-1)} \sum_{i=1}^{N}\sum_{j=1}^{M}(\pi_{ij} - \pi_i)^2$$
$$+ \frac{K-k}{NMKk} \sum_{i=1}^{N}\sum_{j=1}^{M}\left[\pi_{ij}(1-\pi_{ij}) + \frac{(1-p)\{p\pi_{ij}(1-2\pi_{y_{ij}}) - (1-p)\ \pi_{y_{ij}}{}^2 + \pi_{y_{ij}}\}}{p^2} \right], \tag{2.13}$$

$$E\left[\frac{1}{nm} \sum_{i=1}^{n}\sum_{j=1}^{m}\left\{ \hat{\pi}_{ij}(1- \hat{\pi}_{ij}) + \frac{(1-p)\{p\ \hat{\pi}_{ij}(1-2\pi_{y_{ij}}) - (1-p)\ \pi_{y_{ij}}{}^2 + \pi_{y_{ij}}\}}{p^2} \right\} \right]$$
$$= \frac{1}{NM} \sum_{i=1}^{N}\sum_{j=1}^{M}\left[\pi_{ij}(1-\pi_{ij}) + \frac{(1-p)\{p\pi_{ij}(1-2\pi_{y_{ij}}) - (1-p)\ \pi_{y_{ij}}{}^2 + \pi_{y_{ij}}\}}{p^2} \right] \tag{2.14}$$

를 구할 수 있다. 그러므로, 식(2.8)에서 식(2.9)와 식(2.13) 그리고 식(2.14)를 이용하여 식(2.7)를 얻을 수 있다.

3. 부차표본 m과 k의 최적값

3.1 일정한 비용 하에서의 m과 k의 최적값

이 절에서는 표본의 최적배분을 위해 일정한 비용 하에서 표본의 정도를 최대로 하는 m과 k의 값을 결정해 보고자 한다.

먼저 비용함수를 고려해야 하는데, 3단계 추출의 경우 비용함수는 대개 다음과 같

은 형태를 취한다.

$$C' = C - c_0 = c_1 n + c_2 nm + c_3 nmk . \tag{3.1}$$

여기서, C는 총비용이고, c_0는 고정비용으로 조사행정비, 표본설계비용 등을 포함하며 표본의 크기와는 관계없이 소요되는 비용이다. c_1은 표본 1차 추출단위 당 비용으로 집락 당 소요비용을 의미하며, 각 표본 1차 추출단위에서 2차 추출단위를 추출하기 위한 리스트 작성비와 1차 추출단위의 추출작업 등에 필요한 비용을 포함한다. c_2는 표본 2차 추출단위 당 비용이며 c_3는 표본 3차 추출단위 당 소요비용을 의미하며, 표본 3차 추출단위의 추출 및 확인에 소요되는 비용, 확률장치를 이용한 면접 또는 실측비용, 조사자료의 집계분석비용 등을 포함한다.

분산 식(2.7)을 달리 표현해 보면 다음과 같다.

$$V(\hat{\pi}) = \frac{S_a^2}{n} + \frac{S_b^2}{nm} + \frac{S_c^2}{nmk} - \frac{1}{N(N-1)} \sum_{i=1}^{N} (\pi_i - \pi)^2 . \tag{3.2}$$

여기서,

$$S_a^2 = \frac{1}{N-1} \sum_{i=1}^{N} (\pi_i - \pi)^2 - \frac{1}{NM(M-1)} \sum_{i=1}^{N} \sum_{j=1}^{M} (\pi_{ij} - \pi_i)^2 ,$$

$$S_b^2 = \frac{1}{N(M-1)} \sum_{i=1}^{N} \sum_{j=1}^{M} (\pi_{ij} - \pi_i)^2 - \frac{S_c^2}{K} ,$$

$$S_c^2 = \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} \left[\pi_{ij}(1-\pi_{ij}) + \frac{(1-p)\{p\pi_{ij}(1-2\pi_{y_{ij}}) - (1-p)\ \pi_{y_{ij}}^2 + \pi_{y_{ij}}\}}{p^2} \right]$$

이다.

일정한 비용 하에서 분산을 최소로 하는 n과 m 및 k의 값을 라그랑즈(Lagrange) 승수법을 이용하여 구해보기로 하자. 이 때, 최소화하는 함수 $\varnothing$는 다음과 같이 표현된다.

$$\begin{aligned} \varnothing = & \frac{S_a^2}{n} + \frac{S_b^2}{nm} + \frac{S_c^2}{nmk} - \frac{1}{N(N-1)} \sum_{i=1}^{N} (\pi_i - \pi)^2 \\ & + \lambda(c_1 n + c_2 nm + c_3 nmk - C') . \end{aligned} \tag{3.3}$$

여기서, λ는 라그랑즈 승수이다.

식(3.3)으로부터 m의 최적값 m_{opt}와 k의 최적값 k_{opt}를 구해보면 다음과 같다.

$$m_{opt} = \frac{S_b}{S_a} \sqrt{\frac{c_1}{c_2}} , \tag{3.4}$$

$$k_{opt} = \frac{S_c}{S_b}\sqrt{\frac{c_2}{c_3}} \ . \tag{3.5}$$

또한, 식(3.4)와 식(3.5)의 최적값을 비용함수의 식(3.1)에 대입하여 n에 대해 풀면, n의 최적값 n_{opt}를 다음과 같이 구할 수 있다.

$$n_{opt} = \frac{(C-c_0)\sqrt{S_a^{\,2}/c_1}}{S_a\sqrt{c_1}+S_b\sqrt{c_2}+S_c\sqrt{c_3}} \ . \tag{3.6}$$

식(3.2)에 k와 m 및 n의 최적값인 식(3.5)와 식(3.4) 및 식(3.6)를 대입하면 다음과 같은 일정비용 C에 대한 $\hat{\pi}$의 최소분산 식을 유도할 수 있다.

$$V_{\min}(\hat{\pi}) = \frac{(S_a\sqrt{c_1}+S_b\sqrt{c_2}+S_c\sqrt{c_3})^2}{C-c_0} - \frac{1}{N(N-1)}\sum_{i=1}^{N}(\pi_i-\pi)^2 . \tag{3.7}$$

3.2 일정한 정도 하에서의 m과 k의 최적값

이 절에서는 표본의 최적배분을 위해 정도를 미리 일정한 값 V_0로 고정하였을 때, 비용을 최소로 하는 n과 m 및 k의 값을 결정해 보고자 한다.

이들 값 역시 라그랑즈 승수법을 이용하여 구할 수 있으며, 이 때, 최소화하는 함수 $\varnothing$는 다음과 같이 표현된다.

$$\begin{aligned}\varnothing &= c_0 + c_1 n + c_2 nm + c_3 nmk \\ &\quad + \lambda\left(\frac{S_a^{\,2}}{n} + \frac{S_b^{\,2}}{nm} + \frac{S_c^{\,2}}{nmk} - \frac{1}{N(N-1)}\sum_{i=1}^{N}(\pi_i-\pi)^2 - V_0\right).\end{aligned} \tag{3.8}$$

식(3.8)로부터 m의 최적값 m_{opt}와 k의 최적값 k_{opt}를 구해보면 다음과 같다.

$$m_{opt} = \frac{S_b}{S_a}\sqrt{\frac{c_1}{c_2}} \ , \tag{3.9}$$

$$k_{opt} = \frac{S_c}{S_b}\sqrt{\frac{c_2}{c_3}} \ . \tag{3.10}$$

또한, 식(3.9)와 식(3.10)의 최적값을 분산 식(3.2)의 일정한 값 V_0에 대입하여 n에 대해 풀면, n의 최적값 n_{opt}를 다음과 같이 구할 수 있다.

$$n_{opt} = \frac{S_a\sqrt{c_1} + S_b\sqrt{c_2} + S_c\sqrt{c_3}}{\left(V_0 + \frac{1}{N(N-1)}\sum_{i=1}^{N}(\pi_i - \pi)^2\right)\sqrt{c_1/\ S_a^2}} \ . \tag{3.11}$$

식(3.1)에 k와 m 및 n의 최적값인 식(3.10)과 식(3.9) 및 식(3.11)를 대입하여 정리하면, 분산 V_0를 확실하게 보장하는 최소비용은 다음과 같다.

$$C = c_0 + \frac{(S_a\sqrt{c_1} + S_b\sqrt{c_2} + S_c\sqrt{c_3})^2}{V_0 + \frac{1}{N(N-1)}\sum_{i=1}^{N}(\pi_i - \pi)^2} \ . \tag{3.12}$$

4. 결론

본 연구에서는 모집단이 여러 개의 집락으로 구성되어 있을 때, 모집단으로부터 1차 추출단위를 추출하여 표본으로 선정된 2차 추출단위 내에서 다시 조사단위를 추출하는 3단계 집락추출법에 무관질문모형을 적용하여 민감한 속성에 대한 정보를 얻을 수 있는 3단계 집락 무관질문모형을 제안하였다. 그리고, 일정한 비용 하에서 분산을 최소로 하는 1차 추출단위와 2차 추출단위 및 3차 추출단위에 대한 최적값을 구하여 최소분산의 형태를 도출하였다. 또한, 일정한 정도 하에서 비용함수를 최소로 하는 1차 추출단위와 2차 추출단위 및 3차 추출단위의 최적값을 구하여 최소비용의 형태를 도출하였다.

한편, 이러한 3단계 집락 무관질문모형이 사회학, 경제학, 의학, 경영학 등 여러 분야의 연구조사에서 적절히 사용될 수 있다면 관련분야의 연구발전에 도움을 줄 수 있으리라 생각된다.

참고문헌

1. 류제복, 홍기학, 이기성 (1993). 확률화응답모형, 자유아카데미, 서울.
2. 이기성, 홍기학 (1998). 2단계 집락추출법에 의한 확률화응답모형, 한국통계학회논문집, 제 5권 1호, 99-105.
3. Chaudhuri, A. and Mukerjee, R. (1988). *Randomized Response : Theory and Techniques*, Marcel Dekker, Inc., New York.
4. Fox, J. A. and Tracy, P. E. (1986). *Randomized Response : A Method for Sensitive Survey,* Sage Publications.
5. Greenberg, B. G., Abul-Ela, Abdel-Latif A., Simmons, W. R., and Horvitz, D. G. (1969). The Unrelated Question Randomized Response Model : Theoretical Framework, *Journal of the American Statistical Association*, 64, 520-539.
6. Warner, S. L. (1965). Randomized Response ; A Survey Technique for Eliminating Evasive Answer Bias, *Journal of the American Statistical*

Association, 60, 63-69.

[2002년 12월 접수, 2003년 2월 채택]