한국통계학회논문집 제3권 1호
1996년 4월 pp. 101-109

네트웍 샘플링에서 응답오차를 고려한 중복수 추정량

김규성[1], 이기재[2], 박진우[3], 김영원[4]

요 약

네트웍 샘플링은 희귀한 속성을 갖는 모집단에서 유용한 표본조사방법이다. 기존의 중복수 추정량(multiplicity estimator)은 네트웍 샘플링의 특징을 반영하는 추정량으로 응답오차를 고려하지 않은 경우에 이용되었다. 본 논문에서는 응답오차를 고려한 경우에 이용할 수 있는 수정된 중복수 추정량을 제안하였다. 그리고 제안된 추정량의 기대값과 근사기대분산(approximate expected variance)을 유도하였으며, 제안된 추정량이 기존의 모총수 추정량보다 효과적임을 가상모집단을 통하여 보였다.

1. 서 론

희귀속성을 갖는 모집단(rare populations)을 대상으로 하는 표본조사에서 발생하는 문제점은 희귀속성을 포함하는 조사단위가 표본으로 추출되기 어렵다는 점이다. 전체 조사단위중에서 희귀속성을 갖는 조사단위는 일부에 불과하기 때문에 표본조사를 통하여 희귀속성의 총수 혹은 평균을 추정하려면 표본의 크기를 상당히 크게 해야 희귀속성을 갖는 조사단위가 표본에 포함되게 되고, 따라서 원활한 추정이 이루어진다.

예를 들어 전국의 암환자에 관한 표본조사를 고려하자. 표본조사단위로 가구를 이용한다고 했을 때, 암환자가 살고 있는 가구는 모집단 전체 가구에 비하여 소수에 불과할 것이다. 전통적인 표본이론에 근거하여 표본추출을 한 후, 표본가구를 방문하여 암환자의 거주 여부를 조사하면 대부분의 표본조사가구에서 암환자를 발견하지 못할 가능성이 크다. 이는 조사하고자 하는 속성이 희귀하기 때문이며, 이러한 조사를 원활히 수행하기 위해서는 표본의 크기를 상당히 크게 해야 한다.

네트웍 샘플링(Network sampling)은 희귀속성을 갖는 모집단을대상으로 하는 조사에서 이용할 수 있는 유용한 표본조사방법이다. 암환자에 관한 표본조사의 경우 표본조사가구로 부터 그 가구에 암환자가 거주하는지 여부를 조사하고, 더 나아가 그 가구의 구성원으로부터 그들과 연관이 있는 다른 가구에 암환자가 살고 있는지 여부를 추가로 조사하면 표본수 증가의 효과를 거둘 수 있다. 기존의 표본조사에서는 표본으로 추출된 조사단위의 구성원에 대한 정보만 응답을 얻은 반면, 네트웍 샘플링에서는 표본조사단위의 구성원 자신에 대한 정보와 그들이 연관되어 있는 다른 조사단위에 대한 정보까지도 조사를 한다. 따라서 서로 다른 조사단위지만 정보를 공유하는 구성원들의 모임을 네트웍(network)이라 할 때, 실제 관찰값을 얻는 관찰단위는 표본추출단위와 연관된

1) (130-743) 서울특별시 동대문구 전농동 90 서울시립대학교 전산통계학과 전임강사.
2) (110-791) 서울특별시 종로구 동숭동 169번지 한국방송통신대학교 응용통계학과 조교수.
3) (445-743) 경기도 화성군 봉담면 와우리 수원대학교 응용통계학과 조교수.
4) (140-742) 서울특별시 용산구 청파동 2가 숙명여자대학교 통계학과 부교수.

네트웍이 되고, 이같은 네트웍의 속성을 적절히 이용할 경우 보다 효과적인 모수 추정을 기대할 수 있다.

이제까지 대부분의 네트웍 샘플링에 대한 연구는 네트웍내의 구성원간의 정보의 상호 공유를 밑바탕으로 하고 있다. Sirkin(1970, 1972, 1974)의 일련의 연구는 위의 가정을 충실히 지키고 있으며, 조사상의 응답오차는 없다고 가정하고 중복수 추정량에 대한 여러가지 성질을 보여주고 있다. Levy(1977)도 역시 응답오차는 고려하지 않고 네트웍 샘플링에서 층화및 표본배정의 문제를 다루고 있다. 그러나 Kalton과 Anderson(1986)이 지적하였듯이 네트웍내의 정보를 동등하게 공유하기는 어려우며, 네트웍 샘플링의 장점이 표본수 증가의 효과에 있다고 한다면, 반대로 다른 구성원의 정보를 응답하는 과정에서 응답오차(response error)가 추가되면 도리어 응답오차는 커질 수 있는 단점이 있다. Czaja외 2인(1986)은 응답오차를 베르누이 분포를 이용하여 설명하고, 기존의 중복수 추정량을 사용했을 경우에 모수를 과소추정하는 경향이 있음을 보였으며, 기존의 모총수 추정량과 비교하여 응답오차의 추가로 응답편의(response bias)는 증가하지만 표본수 증가의 효과로 평균제곱오차는 감소함을 암환자에 대한 표본조사를 통하여 보여주고 있다.

본 논문에서는 응답오차를 고려한 경우에 이용할 수 있는 수정된 중복수 추정량을 제안한다. 그리고 베르누이 응답모형 아래서 제안된 추정량은 모총수의 불편추정량이 됨을 보이고, 기대분산(expected variance)의 근사값을 유도한다. 또한 몇가지 예제를 통하여 제안된 추정량이 기존의 모총수 추정량보다 효과적일 수 있음을 보인다.

2. 중복수 추정량

모집단 U는 M개의 조사단위로 구성되어 있고, $U = \{ u_i | i=1,\ldots,M \}$, 각 조사단위 u_i는 여러 명의 구성원을 포함하고 있다. 구성원들의 정보공유체인 네트웍은 A_j로 나타내며, A_j내의 구성원들은 서로의 정보를 동등하게 공유함을 가정한다. 네트웍들의 집합을 A로 표시하면 $A = \{A_j | j=1,\ldots,N\}$가 되고, U와 A는 $\bigcup_{i=1}^{M} u_i = \bigcup_{j=1}^{N} A_j$의 관계가 있으며 다음의 네 가지로 분류할 수 있다.

i) 하나의 조사단위가 하나의 네트웍에만 포함되는 경우 :
즉, $M = N$이고, $(u_1,\ldots,u_M) = (A_{i_1},\ldots,A_{i_N})$.

ii) 하나의 네트웍에 여러개의 조사단위가 포함되는 경우 :
즉, $M > N$ 이고, 몇 개의 u_i 로 하나의 A_j 구성.

iii) 하나의 조사단위에 여러개의 네트웍이 포함되는 경우 :
즉, $M < N$ 이고, 몇 개의 A_j로 하나의 u_i 구성.

iv) 나머지 관계들.

위의 네 가지 관계는 형식논리상 존재가능하나, 본 논문에서는 회귀속성을 갖는 모집단을 다루고자 하므로 iii) 과 iv)의 경우는 사실상 별 의미가 없고, i) 과 ii) 의 관계만 의미를 갖는다. 따라서 본 논문에서는 i)과 ii)의 관계만을 고려하기로 한다.

회귀속성을 갖는 구성원을 I_α 로 나타내고, 이들의 집합을 $I = \{I_\alpha | \alpha=1,\ldots,C\}$로 나타내

면, C는 희귀속성을 가진 구성원의 총수를 의미하며, 우리가 추정하고자 하는 모수가 된다. 희귀속성을 갖는 구성원 I_α는 추출단위 u_i에 여러명 포함될 수도 있고, 네트웍 A_j에 여러명 포함될 수도 있다. 그러나 우리가 다루는 모집단이 희귀속성을 갖는 모집단이므로 u_i나 A_j에는 하나의 I_α만이 포함된다고 생각할 수 있다. 이러한 점들을 고려하여 모집단과 네트웍의 관계에 대하여 다음의 두 가지를 가정한다.

가정1) $A_j=\bigcup_{i=1}^{k_j} u_i$, $\sum_{j=1}^{N} k_j=M$ 이며, $A_j\cap A_k=\phi,(j\neq k)$ 이다. (2.1)

가정2) $(I_\alpha\in u_i,\ I_\beta\in u_i)$ 혹은 $(\ I_\alpha\in A_j, I_\beta\in A_j\ )$ 이면 $\alpha=\beta$ 이다. (2.2)

2.1 중복수 추정량

조사단위 u_i 와 네트웍 A_j는 희귀속성을 갖는 구성원 I_α에 의하여 연결될 수 있으며, 다음의 기호를 사용한다.

$$\begin{aligned}
\delta_{ij} &= \begin{cases}1 & \text{만일 } u_i\subset A_j \text{ 이고, } I_\alpha\in A_j\\ 0 & \text{기타,}\end{cases}\\
\nu_{ij} &= \begin{cases}1 & \text{만일 } u_i\subset A_j \text{ 이고, } I_\alpha\in u_i\cap A_j\\ 0 & \text{기타,}\end{cases}\\
\mu_{ij} &= \begin{cases}1 & \text{만일 } u_i\subset A_j \text{ 이고, } I_\alpha\in u_i^c\cap A_j\\ 0 & \text{기타.}\end{cases}
\end{aligned} \tag{2.3}$$

즉, 희귀속성을 지닌 구성원 I_α가 A_j에 속할 때, 표본조사단위에 I_α가 있다고 응답하면 $\nu_{ij}=1$ 이 되고, 표본조사단위에는 I_α가 없으나 네트웍에 연계된 다른 조사단위에 I_α가 있다고 응답하면 $\mu_{ij}=1$이 된다. $\delta_{ij}=\nu_{ij}+\mu_{ij}$이므로 $\delta_{ij}=1$인 경우는 I_α가 u_i내에서 응답되는 경우를 뜻한다.

A_j에 연관된 u_i의 수는 I_α의 중복수(multiplicity)를 의미하며, $s_j=\sum_{i=1}^{M}\delta_{ij}$로 표시한다. u_i에서 얻을 수 있는 I_α들의 가중평균을 $\lambda_i=\sum_{j=1}^{C}\frac{\delta_{ij}}{s_j}$로 나타내기로 한다. 희귀속성을 갖는 구성원의 총수 C의 중복수 추정량은 Sirkin(1970)에 의하여 다음과 같이 제안되었다.

$$\hat{C}=\frac{M}{m}\sum_{i\in s}\lambda_i=\frac{M}{m}\sum_{i\in s}\sum_{j=1}^{C}\frac{\nu_{ij}+\mu_{ij}}{s_j}. \tag{2.4}$$

여기에서 m은 표본수를 의미한다.

2.2 응답 모형

중복수 추정량 $\hat{C}$는 A_j 구성원들 간의 정보의 상호 공유를 전제로 하고 있으며 올바른 응답을 가정하고 있다. 이러한 정보의 상호 공유가 완벽하지 못할 경우, 즉 일부 구성원이 불확실한 정보를 알고 있거나 거짓 응답을 하는 경우, 얻어진 응답을 100% 신뢰하기 어려우므로 이 때에는 응답의 신뢰도에 대한 고려가 필요하다.

희귀속성을 지닌 구성원 I_α가 포함된 네트웍에서 다음과 같은 조건부 베르누이 분포를 가정한다.

가정3) $$a_{ij} \mid (\nu_{ij}=1) \sim B(p_1), \qquad (2.5)$$
$$b_{ij} \mid (\mu_{ij}=1) \sim B(p_2).$$

즉, I_α가 $u_i \cap A_j$ 포함될 때 올바로 응답될 확률을 p_1, I_α가 $u_i^c \cap A_j$ 포함될 때 올바로 응답될 확률을 p_2로 가정한다. 그리고 위의 두가지 응답은 서로 독립임을 가정한다. 그러면 다음과 같은 결과를 얻게 된다.

$$\begin{aligned} E_\xi(a_{ij}) &= p_1\nu_{ij}, \quad Var_\xi(a_{ij}) = p_1(1-p_1)\nu_{ij}. \\ E_\xi(b_{ij}) &= p_2\mu_{ij}, \quad Var_\xi(b_{ij}) = p_2(1-p_2)\mu_{ij}. \\ & Cov_\xi(a_{ij}, b_{kl}) = 0, \ (i \neq k). \end{aligned} \qquad (2.6)$$

단순임의표본에서 응답오차를 고려하면 $a_i \mid (\nu_{ij}=1) \sim B(p_1)$가 되고 단순임의표본 s 에 의한 통상적인 모총수 추정량 $\hat{C}_{srs}$는 다음과 같이 쓸 수 있다.

$$\hat{C}_{srs} = \frac{M}{m}\sum_{i\in s} a_i, \qquad (2.7)$$

$M \approx M-1$과 $\frac{m}{M} \approx 0$을 가정하면, 표본추출방법과 응답모형을 동시에 고려한 기대값(expected value)과 기대평균제곱오차(expected mean squared error)는 다음과 같이 구해질 수 있다.

$$E_\xi E_p(\hat{C}_{srs}) = p_1 C, \qquad (2.8)$$

$$E_\xi E_p(\hat{C}_{srs} - C)^2 = \frac{M}{m} p_1 C\left\{1 - \frac{1}{M}(1 + p_1(C-1))\right\} + C(1-p_1)\{p_1 + C(1-p_1)\}. \qquad (2.9)$$

첨자 p는 표본추출방법을 의미하며, ξ는 응답모형을 의미한다. (2.8)로 부터 응답오차를 고려한 경우 통상적인 모총수 추정량 $\hat{C}_{srs}$은 편의추정량(biased estimator)이 되며, p_1이 작을수록 편의는

커짐을 알 수 있다.

2.3 수정된 중복수 추정량

응답에 오차가 개입되는 경우 (2.4)에서 제시되었던 중복수 추정량을 이용하는 것은 불가능하다. 본 절에서는 오차가 포함된 베르누이 확률변수 a_{ij}, b_{ij}를 기초로 하는 새로운 중복수 추정량을 제안한다. 2.2절의 응답 모형하에서 새로운 중복수 추정량 $\widehat{C_M}$을 다음과 같이 정의한다.

$$\widehat{C_M} = \frac{M}{m}\sum_{i\in s} w_i$$

$$= \frac{M}{m}\sum_{i\in s}\sum_{j=1}^{C}\frac{a_{ij}+b_{ij}}{t_j}\ ,\quad t_j = \sum_{i=1}^{M}(a_{ij}+b_{ij}). \tag{2.10}$$

위에서 제안된 수정된 중복수 추정량 $\widehat{C_M}$은 응답오차를 고려했을 때에도 모총수 C에 대한 불편추정량이 된다는 것과 그것의 기대분산의 근사값을 구한것이 아래의 정리에 나와 있다.

정리 1. 세 가지 가정 (2.1), (2.2), (2.5) 아래에서 비복원 단순임의표본 s를 이용할 때, (2.10)과 같이 정의된 수정된 중복수 추정량 $\widehat{C_M}$의 기대값(expected value)과 근사기대분산(approximate expected variance)은 다음과 같다.

$$E_\xi E_p(\widehat{C_M}) = C, \tag{2.11}$$

$$E_\xi V_p(\widehat{C_M}) \approx \frac{M}{m}\Big\{3p_2\sum_{j=1}^{C}\frac{1}{s_j} + (3p_1-5p_2-8p_1p_2)\sum_{j=1}^{C}\frac{1}{s_j^2}$$

$$+[\ 8p_1p_2-2(p_1-p_2)]\ \sum_{j=1}^{C}\frac{1}{s_j^3} - \frac{C^2}{M}\Big\}. \tag{2.12}$$

증명. 기대값은 $E_\xi E_p(\widehat{C_M}) = E_\xi(C) = C$로 간단히 구할 수 있다. 기대분산은

$$E_\xi V_p(\widehat{C_M}) \ = \ M^2\frac{1}{m}(1-\frac{m}{M})\frac{1}{M-1}\Big\{\sum_{i=1}^{M}E_\xi(w_i^2) - \frac{C^2}{M}\Big\} \tag{2.13}$$

으로 표현할 수 있고, $\sum_{i=1}^{M}E_\xi(w_i^2)$은 아래와 같은 방법에 의해 근사식을 구할 수 있다.

$$\sum_{i=1}^{M}E_\xi(w_i^2) \ = \ \sum_{i=1}^{M}E_\xi\Big\{\sum_{j=1}^{C}\frac{a_{ij}+b_{ij}}{t_j}\Big\}^2$$

$$= \sum_{i=1}^{M}E_\xi\Big\{\sum_{j=1}^{C}\frac{a_{ij}+b_{ij}}{s_j}(1+\Delta_j+\Delta_j^2+\cdots\cdots)\Big\}^2, \qquad \Delta_j = \frac{s_j-t_j}{s_j}$$

$$\approx \sum_{i=1}^{M} E_\xi\Big\{ \sum_{j=1}^{C} \Big(\frac{a_{ij}+b_{ij}}{s_j}\Big)^2 (1+2\Delta_j) + \sum_{k\neq l} \frac{a_{ik}+b_{ik}}{s_k} \frac{a_{il}+b_{il}}{s_l} (1+\Delta_k+\Delta_l) \Big\}. \tag{2.14}$$

모든 (i, j)에 대하여 $\nu_{ij}\,\mu_{ij}=0$ 이므로 $E_M(a_{ij}\ b_{ij})=0$이 되고, $s_j=1+\sum_{i=1}^{M}\mu_{ij}$ 를 이용하면 (2.14)식의 첫번째 항은 다음과 같이 표현된다.

$$\sum_{i=1}^{M} E_\xi\Big\{ \sum_{j=1}^{C} \frac{a_{ij}+2a_{ij}\ b_{ij}+b_{ij}}{s_j^2} \Big\} = \sum_{i=1}^{M}\sum_{j=1}^{C} \frac{p_1\nu_{ij}+p_2\mu_{ij}}{s_j^2}$$

$$= \sum_{j=1}^{C} \frac{1}{s_j^2}(p_1+p_2(s_j-1))$$

$$= p_2 \sum_{j=1}^{C} \frac{1}{s_j} + (p_1-p_2) \sum_{j=1}^{C} \frac{1}{s_j^2}. \tag{2.15}$$

또한, $1+2\Delta_j=3-2\frac{t_j}{s_j}$과 $\sum_{i\in A_j}\sum_{k\in A_j}(\nu_{ki}\ \mu_{ij}+\nu_{ij}\ \mu_{kj})=2(s_j-1)$을 이용하면 다음을 얻을 수 있다.

$$\sum_{i=1}^{M} E_\xi\Big\{ \sum_{j=1}^{C} \Big(\frac{a_{ij}+b_{ij}}{s_j}\Big)^2 \frac{t_j}{s_j} \Big\} = \sum_{i=1}^{M}\sum_{j=1}^{C} E_\xi\Big\{ \frac{(a_{ij}+2a_{ij}b_{ij}+b_{ij}) \sum_{k=1}^{M}(a_{ij}+b_{ij})}{s_j^3} \Big\}$$

$$= \sum_{i=1}^{M}\sum_{j=1}^{C} \frac{1}{s_j^3}\{p_1\nu_{ij}+p_2\mu_{ij}+2p_1p_2 \sum_{k\in A_j}(\nu_{ki}\mu_{ij}+\nu_{ij}\mu_{ki})\}$$

$$= \sum_{j=1}^{C} \frac{1}{s_j^3}\{p_1+p_2(s_j-1)\}+2p_1p_2 \sum_{j=1}^{C} \frac{1}{s_j^3} 2(s_j-1)$$

$$= \sum_{j=1}^{C} \frac{1}{s_j^2}(p_2+4p_1p_2) + \sum_{j=1}^{C} \frac{1}{s_j^3}(p_1-p_2-4p_1p_2). \tag{2.16}$$

$\nu_{ik}\nu_{il}=\nu_{ik}\mu_{il}=\mu_{ik}\nu_{il}=\mu_{ik}\mu_{il}=0$, $(k\neq l)$, 로 부터 (2.14)식의 두 번째 항은 0 이 된다. (2.15)와 (2.16)를 이용하고, $M-1\approx M$, $\frac{m}{M}\approx 0$ 을 가정하면 다음과 같은 근사 기대 분산식을 얻을 수 있다.

$$E_\xi V_p(\widehat{C_M}) \approx \frac{M}{m}\Big\{ \sum_{i=1}^{M} E_\xi(w_i^2) - \frac{C^2}{M} \Big\}$$

$$\approx \frac{M}{m}\Big\{ 3p_2 \sum_{j=1}^{C} \frac{1}{s_j} + (3p_1-5p_2-8p_1p_2) \sum_{j=1}^{C} \frac{1}{s_j^2} + [\ 8p_1p_2-2(p_1-p_2)] \sum_{j=1}^{C} \frac{1}{s_j^3} - \frac{C^2}{M} \Big\}.$$

2.4 효율성 비교

수정한 중복수 추정량 $\hat{C}_M$의 효율을 살펴보기 위하여 다음과 같은 가상적인 모집단을 구성한다. 모집단의 크기는 M=1,000으로 하고, 모집단내 희귀속성의 비율을 각각 1%, 2%, 3% 로 하였다. 이때 중복수 s_j 의 분포는 1~3 의 범위를 고려하여 희귀속성의 비율을 만족하도록 s_j의 분포를 결정하였다. <표1>에 가상적으로 구성한 3개의 모집단 분포가 나타나 있다.

<표1> 효율성 비교를 위한 모집단 구성

모집단	희귀속성의 비율	s_j 의 분포			
		$s_j=1$	$s_j=2$	$s_j=3$	$\bar{s}$
1	0.01	3	5	2	1.90
2	0.02	6	9	5	1.95
3	0.03	11	12	7	1.87

응답의 오차가 존재한다고 가정하여 응답의 정확도를 나타내는 확률인 p_1, p_2의 값은 각각 0.80, 0.85, 0.90, 0.95 인 경우를 고려하였고, 표본의 크기는 $m=50, 100, 150$ 인 경우를 생각하였다.

이러한 각각의 조건에서 기대평균제곱오차의 상대비율 $E_\xi E_p(\hat{C}_M-C)^2/E_\xi E_p(\hat{C}_{srs}-C)^2$ 을 구한 것이 <표2>에 나타나 있다. 여기에서 $\hat{C}_M$의 기대평균제곱값은 (2.12)에 있는 근사값을 이용하였다.

<표2>에서 알 수 있는 바와 같이 희귀속성의 비율이 0.01, 0.02, 0.03인 경우에 제안된 추정량 $\hat{C}_M$ 의 기대평균제곱오차는 $\hat{C}_{srs}$의 그것에 비해 50% 이내로 $\hat{C}_M$이 $\hat{C}_{srs}$보다 효율적임을 알 수 있다. 특히 희귀속성의 비율이 작을수록 $\hat{C}_M$의 효율은 현격히 증가하며, 응답확률 p_2가 클수록 마찬가지로 $\hat{C}_M$의 효율은 증가한다. 응답오차를 고려한 경우에 $\hat{C}_{srs}$는 편의추정량이기 때문에 모수를 과소추정하는 경향이 있고, 응답오차 p_1이 클수록 편의는 줄어들며 $\hat{C}_{srs}$의 표본오차는 감소한다. 또한 $\hat{C}_M$이 불편추정량이기 때문에 표본의 수가 증가할수록 편의추정량인 $\hat{C}_{srs}$보다 기대평균제곱오차의 측면에서 효율이 증가함을 알 수 있다.

<표2> 기대평균제곱오차의 상대 비율 (M=1,000)

회귀속성비율	표본수		응답확률			
			p_2=0.80	p_2=0.85	p_2=0.90	p_2=0.95
0.01	m=50	p_1=0.80	0.19675	0.19253	0.18830	0.18408
		0.85	0.19319	0.18821	0.18324	0.17826
		0.90	0.18887	0.18322	0.17758	0.17193
		0.95	0.18395	0.17771	0.17147	0.16523
	m=100	p_1=0.80	0.18606	0.18206	0.17807	0.17407
		0.85	0.18694	0.18213	0.17731	0.17250
		0.90	0.18587	0.18031	0.17475	0.16919
		0.95	0.18300	0.17679	0.17058	0.16437
	m=150	p_1=0.80	0.17647	0.17268	0.16889	0.16510
		0.85	0.18109	0.17642	0.17176	0.16709
		0.90	0.18297	0.17749	0.17202	0.16655
		0.95	0.18206	0.17588	0.16970	0.16353
0.02	m=50	p_1=0.80	0.42242	0.41501	0.40759	0.40017
		0.85	0.41456	0.40567	0.39677	0.38787
		0.90	0.40511	0.39488	0.38465	0.37443
		0.95	0.39437	0.38297	0.37157	0.36016
	m=100	p_1=0.80	0.39946	0.39245	0.38543	0.37842
		0.85	0.40116	0.39255	0.38394	0.37533
		0.90	0.39867	0.38861	0.37854	0.36848
		0.95	0.39233	0.38098	0.36964	0.35830
	m=150	p_1=0.80	0.37887	0.37221	0.36556	0.35891
		0.85	0.38860	0.38026	0.37192	0.36358
		0.90	0.39244	0.38253	0.37262	0.36272
		0.95	0.39031	0.37902	0.36774	0.35645
0.03	m=50	p_1=0.80	0.45252	0.44606	0.43960	0.43314
		0.85	0.45358	0.44572	0.43785	0.42999
		0.90	0.45098	0.44185	0.43272	0.42359
		0.95	0.44511	0.43486	0.42462	0.41437
	m=100	p_1=0.80	0.41895	0.41297	0.40699	0.40101
		0.85	0.43354	0.42603	0.41851	0.41100
		0.90	0.44137	0.43243	0.42350	0.41456
		0.95	0.44219	0.43201	0.42183	0.41166
	m=150	p_1=0.80	0.39002	0.38445	0.37888	0.37331
		0.85	0.41520	0.40801	0.40081	0.39361
		0.90	0.43215	0.42340	0.41465	0.40591
		0.95	0.43931	0.42920	0.41909	0.40898

3. 결 론

희귀속성의 모총수 C에 대한 추정량으로서 연구된 중복수 추정량은 응답오차를 고려하지 않은 경우에 이용될 수 있는 추정량이다. 응답오차가 없다고 가정하는 경우에 중복수 추정량은 표본수 증가의 효과에 힘입어 표본오차의 감소를 보여준다. 그러나 네트웍내의 정보의 상호 공유가 불완전한 경우 응답오차를 수반하게 되고, 이러한 비표본오차의 증가 가능성은 네트웍 샘플링의 문제점으로 제기되고 있다.

본 논문에서는 응답오차를 고려한 경우에도 사용할 수 있는 수정된 중복수 추정량을 제안하였다. 응답오차에 대한 확률모형으로서 베르누이 분포를 가정하였고, 표본추출방법으로서 비복원 단순임의추출을 고려하였다. 이 때에 제안된 추정량은 불편추정량이 됨을 보였고, 그 추정량에 대한 기대평균제곱의 근사식을 유도하였다.

효율성비교를 위하여 희귀속성의 비율이 1%, 2%, 3%인 희귀모집단을 구성하였고, 각각의 경우에서 제안된 추정량이 기존의 모총수 추정량보다 효율적임을 보였는데, 특히 희귀속성의 비율이 작을수록 효율이 증가함을 살펴볼 수 있었다.

참 고 문 헌

[1] Czaja, R. F. Snowdon, C. B. and Casady, R. J. (1986). Reporting bias and sampling errors in a survey of a rare population using multiplicity counting rules. *Journal of the American Statistical Association,* Vol. 81, 411-419.

[2] Kalton, G. and Anderson, D. W. (1986). Sampling Rare Populations. *Journal of Royal Statistical Society,* Vol. A149, 65-82.

[3] Levy, P. S. (1977). Optimum allocation in stratified random network sampling for estimating the prevalence of attributes in rare populations. *Journal of the American Statistical Association,* Vol. 72, 758-763.

[4] Nathan, G. (1976). An empirical study of response and sampling errors for multiplicity estimators with different counting rules. *Journal of the American Statistical Association,* Vol. 71, 808-815.

[5] Sirken, M. G. (1970). Household surveys with multiplicity. *Journal of the American Statistical Association,* Vol. 63, 257-266.

[6] Sirken, M. G. (1972). Variance components of multiplicity estimators. *Biometrics,* Vol. 28, 869-873.

[7] Sirken, M. G. and Levy, P. S. (1974). Multiplicity estimation of populations based on ratios of random variables. *Journal of the American Statistical Association,* Vol. 69, 68-73.