Journal of the Korean
Data & Information Science Society
1999, Vol. 10, No. 1, pp. 79 ~ 89

중복수가 있는 다변량 층화임의추출에 관한 연구[1] (층별로 독립인 경우의 배분문제)

김 호 일[2]

요약

중복수가 있는 조사는 추출단위(병원,가구)가 단순임의추출 또는 층화임의추출을 통해 추출되고 추출단위들이 여러 조사단위(환자,사람)들과 서로 연결되어 있는 경우를 말한다. 연결형태에 따른 조사단위의 집합을 network라 정의하면 network는 하나 이상의 추출단위와 연결될 것이고 하나의 추출단위는 하나 이상의 network와 연결이 될 것이다. 본 논문에서는 두 개 이상의 변수가 연결되는 중복수가 있는 다변량 층화임의추출의 경우에 배분문제를 연구하였다.

주제어: 중복수 추출, 다변량 층화추출법, Neyman 최적배분

1. 서 론

흔치않은 특수한 질병을 지닌 환자의 수를 추정할 목적으로 조사를 행하는 경우에는 표본의 추출틀로서 몇몇의 병원을 선택하게 된다. 이때 선택된 병원의 환자의 진료기록부를 통해 특수한 질병을 지닌 환자의 수를 추정할 수 있다. 그러나 특수한 질병을 지닌 환자들은 한 곳의 병원뿐 아니라 여러 곳의 병원으로 옮겨 다니면서 진료를 받을 수도 있다. 예를 들어 암 환자수를 추정하고자 할 경우 한 암환자가 가까운 개인 병원에서 진료를 받다가 큰 병원으로 가게 되고, 또 그것도 안될 경우에는 전문적인 암치료 센터에서 진료를 받을 수 있다. 따라서 한 암환자가 진료받은 병원이 많으면 많을수록 그 암환자가 표본에서 추출될 확률이 높게 된다. 위와 같은 형태의 조사를 중복수 추출(multiplicity sampling) 또는 네트워크 추출(network sampling)이라 한다. 중복수 추출의 첫 연구는 미국 동부의 어느 질병의 확산정도를 추정하기 위한 조사로 1963년에 처음으로 행해졌다. 그 이후 중복수 추출의 추정량에 대한 불편추정량은 Birnbaum과 Sirken(1965)에 의해 연구되었다. 중복수 추출에 관한 논문은 주로 Sirken(1970, 1972)에 의해 주도되었으며 그 이후에도 계속 연구가 되었

[1]본 논문은 1997년도 안양대학교 학술연구비의 지원을 받아 연구되었음
[2](430-714) 경기도 안양시 만안구 안양5동 708-113 안양대학교 정보통계학과 조교수

다(Sirken과 Levy (1974), Nathan(1976), Levy(1977), Czaja, Snowdon과 Casady(1986), Sudman, Sirken과 Cowan(1988)과 Kalton과 Anderson (1986), Faulkenberry과 Garoui(1991)). 그중에서도 Sirken(1972)은 중복성이 있는 층화임의추출을 연구하였고 Levy(1977)은 중복성이 있는 층화임의추출에서 배분문제를 언급하였다.

그러나 이런 중복수 추출의 경우에 한 모집단에서 변수가 둘 이상 조사해야 할 경우가 생길 수 있다. 예를 들면 어느 지역의 당뇨병 환자의 수와 고혈압 환자의 수를 추정하고자 하는 경우가 이 경우에 해당된다. 당뇨병의 합병증으로 고혈압이 생겨 날수 있으므로 어느 정도 상관이 존재한다고 볼 수 있다. 이 경우에 조사변수의 개수는 두 개가 되어 이변량 중복수 추출이 되며 이런 변수가 3개 이상일 때 다변량 중복수 추출이 된다. 이런 중복수가 있는 다변량 층화임의추출에 대해 고려해보고 그것에 따른 배분법을 알아보고자 한다.

2장에서는 일변량의 경우 중복수 추출의 정의와 배분문제를 다루었으며, 3장에서는 중복수가 있는 다변량 층화임의추출의 경우에 기존 각 변수별 Neyman 배분의 평균배분이 적절하다는 것을 근거로 각 변수별 최적배분에 가중치를 부여하는 문제를 다루었다.

2. 일변량 중복수 추출의 배분 문제

2.1 정의

먼저 조사단위란 조사대상이 되는 병원에서의 환자나 가구에서의 조사대상이 되는 사람을 말하고 추출단위란 병원이나 가구같은 조사단위의 집합을 말한다. 따라서 모집단 π에 추출단위 M개가 속해 있다 하고 모집단은 L개의 층으로 나누어져 있으며 h층의 π_h 모집단에는 M_h개의 추출단위가 있다 하자. 또 $Q_{h,T,k}$는 $\pi_h(h=1,\cdots,L)$에서의 $T(T=\tau,\tilde{t},\hat{t})$방법에 따른 $k(k=1,...,M_h)$번째 추출단위라 정의하자. 또 모집단에는 N개의 조사단위가 있다하고 $I_{h,T,\alpha}(\alpha=1,\ldots,N)$는 T의 조사방법에 따른 h 번째 층에서 α번째 조사단위라 정의하자. 여기서 추출방법에는 기존 추출방법과 두 종류의 중복수 추출방법으로 총 세 종류의 추출방법이 가능하다. 이를 각각 $T=\tau,\tilde{t},\hat{t}$로 표기하자. 다음은 세가지의 추출방법에 따른 설명이다.

1) $T=\tau$은 기존조사의 방법으로 조사단위가 하나의 추출단위에서만 보고 되는 경우를 가르킨다. π_h층에서는 M_h개의 추출단위가 어떤 특징을 지닌 조사단위 N_h를 보고하고, 조사단위 $I_{h,\tau,\alpha}$는 각각 하나의 추출단위에서만 보고되고 다른 추출단위에서는 보고가 되지 않는 경우를 말한다.(기존조사)

2) $T=\tilde{t}$는 각 조사단위 $I_{h,\tilde{t},\alpha}$가 그 추출단위뿐 아니라 그 층에 있는 다른 추출단위에서도 보고되는 경우를 말한다.(중복조사1)

3) $T=\hat{t}$는 각 조사단위 $I_{h,\hat{t},\alpha}$가 그 추출단위뿐 아니라 그 층에 있는 추출단위는 물론 다

른 층의 추출단위에서도 보고되는 경우로 모든 보고 가능한 경우를 말한다. (중복조사2) 이 경우 본 논문에서는 다변량의 경우에는 고려하지 않았다.

또 다음과 같은 기호를 정의하자.

$$\delta_{h,T,k,\alpha} = \begin{cases} 1 & Q_{h,T,k} \text{ 가 } \theta_{h,T,\alpha}\text{의 구성원일 경우} \\ 0 & \text{그렇지 않을 경우} \end{cases} \tag{1}$$

여기서 $Q_{h,T,k}$는 $\pi_h(h = 1, \cdots, L)$에서의 $T(T = \tau, \tilde{t}, \hat{t})$방법에 따른 $k(k = 1, ..., M_h)$번째 추출단위이고, $I_{h,T,\alpha}$는 $(\alpha = 1, ..., N)$는 T의 조사방법에 따른 h 번째 층에서 α번째 조사단위이다. 또 $\theta_{h,T,\alpha}$는 $T(= \tau, \tilde{t}, \hat{t})$ 방법에 의해 $I_{h,T,\alpha}$를 보고하는 π_h내의 추출단위의 네트워크를 가르킨다.

따라서 h층에서 T방법에 의한 특성치의 평균은 다음과 같이된다.

$$\overline{Y}_{h,T} = \frac{1}{M_h} \sum_{k=1}^{M_h} Y_{h,T,k} \tag{2}$$

$$\begin{aligned} \text{단, } Y_{h,T,k} &= \sum_{\alpha=1}^{N} \frac{1}{s_{h,T,\alpha}} \delta_{h,T,k,\alpha} \\ &= T \text{ 방법에 의한 } Q_{h,T,i,k}\text{에 의해 보고되는 개인의 가중치} \\ s_{h,T,\alpha} &= \sum_{k=1}^{M_h} \delta_{h,T,k,\alpha} \\ &= T \text{ 방법에 의한 } I_{h,T,i,\alpha}\text{를 보고하는 추출단위의 수} \end{aligned}$$

$m = \sum_{h=1}^{L} m_h$인 추출단위의 층화표본이 비복원으로 추출되었다 하고 T방법에 의한 모집단에 있는 조사단위 N의 추정량은 $N'_{st,T}$은 다음과 같다.

$$N'_{st,T} = \sum_{h=1}^{L} N'_h, T = \sum_{h=1}^{L} \frac{M_h}{m_h} \sum_{k_j}^{m_h} Y_{h,T,k_j} \tag{3}$$

여기서 k_j를 π_h에서 추출된 표본의 순서라 하자. 그러나 $T = \tau, \tilde{t}$의 경우에 각 층에 대한 평균과 총계의 추정량은 불편추정량이나 $T = \hat{t}$의 경우에는 반드시 불편추정량이 되는 것은 아니다. 이는 $N'_{h,\hat{t}}$이 다른 층과 관련되어 있기 때문이다.

또 기존조사, $T = \tau$의 경우는 $s_{h,T,\alpha} = 1$인 경우이다. 그러나 중복조사, $T = \tilde{t}, \hat{t}$ 의 경우는 $s_{h,T,\alpha} \geq 1$의 경우이므로 표본을 통해 보고되는 또 다른 정보가 더 필요하다.

2.2 중복수 추출의 총계 분산

방법 $T(T = \tilde{t}, \hat{t})$에서 변수 i의 중복수 추출의 총계에 대한 분산은 기존의 층화추출의 분산과 같다. 따라서 $M_h - 1 = M_h$라 하면 총계에 대한 분산은 다음과 같다.

$$\begin{aligned} \sigma_T^2(N) &= Var(N'_{st,T}) \\ &= \sum_{h=1}^{L} Var(N'_{h,T}) \\ &= \sum_{h=1}^{L} M_h^2 Var(Y_h, T)(\frac{1}{M_h} - \frac{1}{m_h}) \end{aligned} \tag{4}$$

단, $Var(Y_{h,T}) = \frac{1}{M_h}\sum_{k=1}^{M_h}(Y_{h,T,k} - \overline{Y}_{h,T})^2$이고 $Var(Y_{h,T})$에 대한 추정량은 $\widehat{Var}(Y_{h,T}) = \frac{1}{m_h}\sum_{k=1}^{m_h}(Y_{h,T,k} - \bar{Y}_{h,T})^2$이다.

2.3 비례배분의 경우

표본의 층화 추출단위가 $m = \sum_{h=1}^{L} m_h$와 $m_h = \frac{M_h}{M}m$과 같이 비례배분이 되었다면 모집단 총계 N의 불편추정량 $N'_{st,T}$에 대한 비례배분의 분산은 다음과 같다.

$$Var(N'_{prop,T}) = \frac{M-m}{m}\sum_{h=1}^{L} M_h Var(Y_{h,T}) \tag{5}$$

2.4 사전에 최적배분의 경우

일변량의 층화임의추출의 경우 최적배분은 잘 알려진대로 주어진 비용 하에서 변수의 층화평균의 분산을 최소로 하는 Neyman 최적배분(Neyman allocation)이다. 이 경우는 층별로 독립되어 있다는 가정하에 최적배분이 계산되어졌으므로 $T = \hat{t}$ 경우에는 해당이 되지 않는다. 따라서 표본의 층화 추출단위가 $m_T = \sum_{h=1}^{L} m_{h,T}$가 $m_{h,T} = \frac{M_h\sqrt{Var(Y_{h,T})}}{\sum_{h=1}^{L} M_h\sqrt{Var(Y_{h,T})}}m$과 같이 최적배분이 되었다면 i 변수에 대한 N의 불편추정량 $N'_{st,T}(T = \tau, \tilde{t})$에 대한 최적배분의 분산은 다음과 같다.

$$Var(N'_{opt,T}) = M^2 \frac{\left(\sum_{h=1}^{L} M_h \sqrt{Var(Y_{h,T})}\right)^2}{m} - \frac{\sum_{h=1}^{L} M_h^2 Var(Y_{h,T})}{M} \quad (6)$$

3. 중복수가 있는 다변량 층화표본조사의 배분

3.1 Neyman 최적배분의 공분산행렬

중복수가 있는 다변량 층화임의추출의 배분문제도 기존의 다변량 층화임의추출의 배분문제와 연관시켜 볼 필요가 있다. 먼저 기존 다변량의 층화임의추출의 배분 문제를 생각해보면 다음과 같다. 다변량의 경우 대부분의 조사가 둘 이상의 변수를 고려하는 경우인데, 이와 같은 경우 어느 한 변수의 층화평균의 분산을 줄이는 배분방법을 사용하면 그 배분방법은 다른 변수의 층화평균의 분산을 증가시키는 결과를 초래할 수도 있기 때문에 일변량 최적배분의 사용은 적합하지 않다. 기존의 변수가 둘 이상인 경우의 1940년대 이후 여러학자들에 의해 다양한 형태의 최적배분방법들이 제시되어 왔다. 다변량 층화임의추출에서 절충배분의 가장 단순한 형태로 각 변수별 Neyman 최적배분의 평균을 생각할 수 있는데, 이 Neyman 평균배분은 기존의 다변량 최적배분에 관한 연구에서 많이 고려되었고, 일반적으로 제안된 대부분의 방법에 비해 결코 뒤떨어지지 않는 방법으로 인정되어 왔다(Ghosh(1958), Huddleston,Claypool 과 Hocking (1970), Sukhatme와 Sukhatme (1970), Schuenemeyer (1975), Benn 과 Burmeister(1978), Kish(1988)). 그러나 Neyman 최적배분의 평균은 사실 다변량 절충배분에서 각 변수가 동일한 수준의 영향을 가지는 상황을 설정하고 있다는 점에서 변수들간의 관계가 비교적 독립일 경우 Neyman 평균배분은 그의 적절성을 지니게 되지만, 변수간의 적당한 수준이상의 상관관계가 존재할 경우 이 상관관계 를 반영하는 절차를 고려하는 것도 한 방법이 될 수 있다. 이 장에서는 일변량 Neyman 최적배분의 적절성을 일차적으로 포용한다는 입장에서 Neyman 최적배분 행렬 Σ_{st}로 부터 출발하여 행렬 Σ_{st}의 각 행의 단순평균 대신 각 열(변수)에 적절한 가중치를 부여하는 일종의 Neyman 가중평균을 고려하고자 한다. 먼저 기존의 추출방법이나 중복수가 있는 경우의 추출방법 $T(T = \tau, \tilde{t})$에 대한 $i(i = 1, 2, \cdots, p)$번째 변수의 중복수 추출의 총계의 분산은 식(4)과 동일하여 다음과 같다.

$$\begin{aligned} \sigma^2_{T,i}(N) = \sigma_{T,i,i}(N) &= Var(N'_{st,T,i}) \\ &= \sum_{h=1}^{L} Var(N'_{h,T,i}) \end{aligned} \quad (7)$$

$$= \sum_{h=1}^{L} M_h^2 Var(Y_{h,T,i})(\frac{1}{M_h} - \frac{1}{m_h})$$

또한 총계에 대한 i번째 변수와 i'번째 변수간의 공분산은 다음과 같이 얻어진다.

$$\begin{aligned} \sigma_{T,i,i'}(N) &= Cov(N'_{st,T,i}, N'_{st,T,i'}) \qquad (i \neq i') \\ &= \sum_{h=1}^{L} Cov(N'_{h,T,i}, N'_{h,T,i'}) \\ &= \sum_{h=1}^{L} M_h^2 \rho_{h,T,i,i'} \sqrt{Var(Y_{h,T,i})} \sqrt{Var(Y_{h,T,i'})} (\frac{1}{M_h} - \frac{1}{m_h}) \end{aligned} \tag{8}$$

여기서 분산,$Var(Y_{h,T,i})$은 $Var(Y_{h,T,i}) = \frac{1}{M_h}\sum_{k=1}^{M_h}(Y_{h,T,i,k} - \overline{Y}_{h,T,i})^2$이고 $\rho_{h,T,i,i'}$은 h층에서 T방법에 의한 i번째 변수와 i'번째 변수의 상관계수이다. 따라서 T방법에 의한 Neyman 최적배분의 행렬 $\Sigma^O_{st,T}$는 다음과 같이 정의할 수 있다.

$$\Sigma^O_{st,T} = \begin{pmatrix} \sigma^O_{T,1,1}(N) & \cdots\cdots & \sigma^O_{T,1,p}(N) \\ \ddots & & \cdot \\ \cdot & \sigma^O_{T,i,i}(N) & \cdot \\ \cdot & \ddots & \cdot \\ \sigma^O_{T,p,1}(N) & \cdots\cdots & \sigma^O_{T,p,p}(N) \end{pmatrix} \tag{9}$$

즉, 중복수 추출에 있어서 Neyman 최적배분의 행렬 Σ^O_{st}이란 각 층별로 Neyman 최적배분$(m_{h,T,i}(h = 1, 2, \cdots, L, T = \tau, \tilde{t},\ i = 1, 2, \cdots, p))$ 평균으로 이루어진 행렬을 말한다.

단

$$\sigma^O_{T,i,i}(N) = \sum_{h=1}^{L} M_h^2 Var(Y_{h,T,i})(\frac{1}{M_h} - \frac{1}{\overline{m}_{h,T}})$$

$$\sigma^O_{T,i,i'}(N) = \sum_{h=1}^{L} M_h^2 \rho_{h,T,i,i'} \sqrt{Var(Y_{h,T,i})} \sqrt{Var(Y_{h,T,i'})} (\frac{1}{M_h} - \frac{1}{\overline{m}_{h,T}})$$

이다.

여기서

$$\overline{m}_{h,T} = \frac{\sum_{i=1}^{p} m_{h,T,i}}{p}$$

이다.

위의 $\overline{m}_{h,T}$는 각 변수별로 동일한 가중치를 준 것이라고 볼 수 있다. 그러나 각 변수별로 Neyman 최적배분의 가중치가 다른 선형결합을 고려해 볼 수 있다. 따라서 이 논문에서 고려하고 있는 것은 각 변수별로 Neyman 최적배분의 가중치가 다른 선형결합을 이용한 다변량 층화임의추출의 공분산행렬에 기초한다. 여기에서 가중치의 원소들의 합은 1이 된다. 즉 $\mathbf{1}'D_\omega\mathbf{1} = 1(\mathbf{1}' = (1,1,\cdots,1))$, $D_\omega = diag(\omega_1,\ldots,\omega_p)$이다. 여기서 $diag(\cdot)$는 정방행렬의 대각원소를 가르킨다. 이 행렬은 절충배분의 최적성을 재는 여러 가지 척도에 따라 여러 가지 값을 가지게 될 것이다. 이와 같은 행렬이 $m^{\omega}_{h,T}\left(=\sum_{i=1}^{p}\omega_i m_{h,T,i}/p\right)$에 기초하여 다변량 층화임의추출의 공분산행렬을 구하면 다음과 같다.

$$\Sigma^{\omega}_{st,T} = \begin{pmatrix} \sigma^{\omega}_{T,1,1}(N) & \cdots\cdots & \sigma^{\omega}_{T,1,p}(N) \\ \cdots & & \cdots \\ \cdots & \sigma^{\omega}_{T,i,i}(N) & \cdots \\ \cdots & & \cdots \\ \sigma^{\omega}_{T,p,1}(N) & \cdots\cdots & \sigma^{\omega}_{T,p,p}(N) \end{pmatrix} \tag{10}$$

단,

$$\sigma^{\omega}_{T,i,i}(N) = \sum_{h=1}^{L} M_h^2 Var(Y_{h,T,i})\left(\frac{1}{M_h} - \frac{1}{m^{\omega}_{h,T}}\right)$$

$$\sigma^{\omega}_{T,i,i'}(N) = \sum_{h=1}^{L} M_h^2 \rho_{h,T,i,i'}\sqrt{Var(Y_{h,T,i})}\sqrt{Var(Y_{h,T,i'})}\left(\frac{1}{M_h} - \frac{1}{m^{\omega}_{h,T}}\right)$$

이다.

여기서

$$\sigma^{\omega}_{h,T} = \frac{\sum_{i=1}^{p}\omega_i m_{h,T,i}}{p}$$

이다.

3.2 척도의 기준

여기에서 가중치가 결정되면 다변량 층화임의추출의 의한 공분산행렬의 다변량 시스템 변이를 재는 두가지 전형적인 방법으로서 공분산행렬 Σ_{st}^{ω}의Trace와 일반화 분산을 최소화하는 방법을 고려할 수 있겠다. Trace는 고유값들의 합으로 일반화 분산은 고유값들의 곱으로 표현할 수 있으므로 일반화 분산의 경우 다른 모든 고유값이 적당히 크다고 하더라도 작은 몇개의 고유값에 의해 그 값이 매우 민감하게 움직인다는 단점이 있어 통상적으로 Trace가 전체의 변이의 측도로 많이 이용된다.

1. Trace의 기준
   $tr[\sum_{st}^{\omega}]$는 곧 각 층화표본평균의 분산의 합으로서 다음과 같이 표현할 수 있다.

$$\sum_{i=1}^{p}\sigma_{ii}^{\omega}=\sum_{i=1}^{p}V^{\omega}(N'_{st,T,i})=\sum_{i=1}^{p}\sum_{h=1}^{L}M_h^2Var(Y_{h,T,i})(\frac{1}{M_h}-\frac{1}{m_{h,T}^{\omega}})$$

   여기서 $\sum_{i=1}^{p}\omega_i=1$인 조건하에서 윗식을 최소로 하는 $\omega_i(i=1,2,\cdots,p)$를 찾는 것이 목적이다. 이 경우에 $\omega_i(i=1,2,\cdots,p)$가 식에 포함되어 있으므로 반복을 통해 구하여야 한다. 또한 변수가 많은 경우에는 그 만큼 시간이 많이 걸린다.

2. 일반화분산 기준
   이 기준에서는 $\sum_{j=1}^{p}\omega_j=1$인 조건하에서 가중치가 있는 다변량 층화임의추출의 공분산행렬의 일반화 분산인 $|\sum_{st}^{\omega}|$를 최소로하는 ω_i를 구하고자 한다. 이 경우에도 $\omega_i(i=1,2,\cdots,p)$가 $|\sum_{st}^{\omega}|$에 포함되어 있으므로 반복을 통해 구하여야 한다.

4. 결론

중복수가 있는 다변량 층화추출의 배분의 경우에도 기존의 다변량 층화추출에서의 배분문제와 동일하게 다룰 수 있다. 본 논문에서는 Neyman 최적배분의 평균은 단순히 각 층에 대한 배분에서 동일한 수준의 가중값를 줌으로서 같은 수준의 영향력을 지니고 있는 것으로 생각하여 변수들간의 상관관계의 크기에 따라 서로 다른 가중치을 주어야 한다는 관점에서 일차원 Neyman 최적배분의 가중평균을 고려하였다. 그러나 논문에서는 층간이 독립인 경우에만 가정한 것으로 층간에도 서로 중복수가 있는 경우도 연구되어야 할 것이다.

참고문헌

1. Birnbaum, Z. W. and Sirken, M. G.(1965). Design of sample surveys to estimate the prevalence of rare disease: Three Unbiased Estimates, *Vital and Health Statistics, Series 2*, No. 11. Washington: Government Printing Offices.

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

3. Dalenius, T.(1953). Multivariate Sampling Problem, *Skandinavisk Actuarietidskrift*, 36, 92-102.

4. Faulkenberry, G. D. and Garoui, A.(1991). Estimating a population total using a area frames, *Journal of the American Statistical Association*, 86, 445-449.

5. Huddleston, H.F., Claypoll, P.L. and Hocking, R.R.(1970). Optimum Sample Allocation to Strata using Convex Programming, *Applied Statistics*, 19, 273-278.

6. Ghosh, S.P.(1958). A Note on the Stratified Random Sampling with Multiple Characters, *Calcutta Statistical Association Bulletin*, 8, 81-90.

7. Kalton, G. and Anderson, D. W(1986). Sampling rare populations, *Journal of the Royal Statistical Society, Series A*, 149, 69-82.

8. Kish, L.(1976). Optima and Proxima in Linear Sample Designs, *Journal of the Royal Statistical Society, Series A*, 113, 80-95.

9. Kokan, A.R.(1963). Optimum Allocation in Multivariate Survey, *Journal Of the Royal Statistical Society, Series A*, 126, 557-565.

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

11. Nathan, G. (1976). An empirical study of reponse and sampling errors for multiplicity estimates with different counting rules.*Journal of the American Statistical Association*, 71, 808-815.

12. Mahalanobis, P.C.(1944). On Large-Scale Sample Survey, *Philosophcal Transaction of the Royal Society of London, Series B*, 231, 329-451.

13. Schuenemeyer, J.H.(1975). *Maximum Eccentricity as a Union- Intersection Test in Multivariate Analysis*, Ph.D. Geogia University.

14. Snedecor, G. and King, A.J.(1942). Recent Development in Sampling for Agricultural Statistics, *Journal of the American Statistical Association*, 95-102.

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

16. Sirken, M. G(1972). Stratified sample surveys with multiplicity, *Journal of the American Statistical Association*, 67, 224-227.

17. Sirken, M. G. and Levy, P. S.(1974). Multiplicity estimation of proportiona based on ratio of random variables, *Journal of the American Statistical Association*, 69, 68-73.

18. Sudman, S., Sirken, M. G. and Cowan, C. D.(1988). Sampling rare and elusive populations, *Science*, 240, 991-996.

19. Sukhatme, P.V and Sukhatme, B.V.(1970). *Sampling Theory of Surveys with Applications* : Food and Agriculture Organization, Rome, 2nd edition.

A Study on the Multivariate Stratified Random Sampling with Multiplicity [3]

Hoil Kim [4]

요약

A counting rule that allows an element to be linked to more than one enumeration unit is called a multiplicity counting rule. Sample designs that use multiplicity counting rules are called network samples. Defining a network to be a set of observation units with a given linkage pattern, a network may be linked with more than one selection unit, and a single selection unit may be linked with more than one network. This paper considers allocation for multivariate stratified random sampling with multiplicity.

Key words and Phrases: Multiplicity, Multivariate stratified sampling, Neyman allocation

[3]This paper was supported by Anyang University Research Fund 1997

[4]Assistant Professor, Department of Information Statistics, Anyang University, Kyunggi-do, Manangu, 430-714, Korea