한국통계학회논문집 제3권 1호
1996년 4월 pp. 131-143

Neyman 최적배분의 공분산 행렬에 근거한 다변량 절충배분

김 호 일[1)]

요 약

다변량 층화임의추출에서 한 변수의 Neyman 최적배분은 다른 변수에 대한 층화분산을 최소화시키지 못하는 결과를 초래할 수도 있다. 따라서 다변량 자료의 경우 '최적'배분 대신에 '절충'배분이 도입되어 왔다. 이 연구에서는 각 변수 별 Neyman 최적배분에 근거해서 얻은 층화표본평균벡터의 공분산행렬에 가장 잘 적합되는 층별로 동일한 크기의 절충배분을 찾고자 한다. 이에 적절한 기준 다섯가지를 제시하고 예를 통해 비교, 분석하였다.

1. 서론

Neyman 최적배분은 층화임의추출에서 주어진 비용 하에 대상 변수에 대한 층화분산을 최소로 하는 층별 표본추출단위의 배분을 뜻한다. 그러나 대부분의 조사가 실제로 단 하나의 변수만을 대상으로 하기보다는 둘 이상의 변수를 고려하는 경우인고로, 어느 한 변수에 대한 Neyman 최적배분은 다른 변수의 층화분산을 최소화시키지 못하는 결과를 초래할 수도 있을 것이다. 따라서 주어진 비용 하에서 다변량 자료의 경우 '최적'배분 대신 '절충'배분의 개념이 도입되어 왔다. 다변량 층화평균벡터의 변화의 추이의 대상은 공분산행렬이 될 것이고 이 행렬이 최적배분을 위한 분석의 기준이 될 것이다. 따라서 이 공분산행렬을 일차원화한 변화의 추이의 측도가 필요하고, 이를 어떻게 정의하느냐에 따라 여러 가지 유형의 방법들이 제안될 수 있다.

다변량 절충배분은 1940년대 Snedecor 와 King(1942)과 Mahalanobis(1944) 등에 의해 "상대효율"개념을 이용한 방법이 고려되었고, 그 후 Dalenius(1957) 등에 의해 후속적인 연구가 있었다. 이들은 주로 각 분산에 대한 제약이 주어질 경우 조사비용을 최소로 하는 방법을 선형프로그램을 통해 연구, 개발하였다. 여기에서는 제약식의 유형에 따라 여러 가지 방법이 제안되었는데 그 중 중요한 것은 Dalenius (1953,1957), Yates(1960), Kokan(1963), Hartley(1965), Kokan & Khan(1967), Chatterjee(1968, 1972), Huddleston, Claypool 과 Hocking(1970), Bethel(1985,1989), Chromy(1987) 등이다. 그리고 Dalenius(1953, 1957)를 비롯한 Chakravarti(1955), Ghosh(1958), Yates(1960), Aoyama(1963), Folks 와 Antle(1965), Hartley(1965), Kish(1972) 등은 고정된 표본 또는 고정된 비용 하에 분산이나 일반화분산을 최소로 하는 경우를 고려하였다. 한편

1) (430-714) 경기도 안양시 만안구 안양 5동 안양대학교 통계학과 전임강사

Schuenemeyer(1975)는 층별 공분산행렬의 최대고유값을 일변량 Neyman 최적배분 공식에 그 층에 분산의 비례하는 양으로 간주하여 처리하는 방법을 제안하였다. 각 층별로 Neyman 최적배분의 단순한 평균은 각 변수에 대해 동일한 수준의 비중을 고려한 것이지만, Kim 과 Kim(1994)은 각 변수들 간에 서로 다른 가중치를 줌으로서 다른 유형의 절충배분을 제시한 바 있다.7

기존의 연구 결과(Ghosh(1958), Huddleston, Claypool 과 Hocking(1970), Sukhatme 와 Sukhatme (1970), Schuenemeyer(1975), Benn 과 Burmister(1978), Kish(1988), Kim(1993))에 따르면 각 층별로 단순히 Neyman 최적배분들의 평균을 취한 절충배분의 경우가 많은 예에서 적절하다는 사실이 보고되고 있다. 이 연구에서는 이 점에 착안하여 일차원적 Neyman 최적배분이 가지는 개개 변수별 최적성을 기초로 절충배분의 관점에서 이 정보를 다변량의 경우로 확장시키는 접근방법을 고려하고 있다. 그것은 각 변수별 Neyman 최적배분에 근거해서 얻은 층화평균의 공분산행렬을 가장 잘 적합시키는 표본크기로서, 이는 각 변수별 서로 다른 배분이 아닌 모든 변수에 대해 층별로 동일한 크기의 절충배분을 찾고자 하는 것이다.

2. Neyman 최적배분에 의한 공분산행렬의 적합

p-변량 층화임의추출에서 h층($h=1,\cdots,L$)의 모집단 크기는 N_h ($\sum_{h=1}^{L} N_h = N$)이고, j번째 변수 ($j=1,\cdots,p$)에 대한 Neyman 최적배분을 $n_{1j}^{o}, n_{2j}^{o}, \cdots, n_{hj}^{o}, \cdots, n_{Lj}^{o}$ ($\sum_{h=1}^{L} n_{hj}^{o} = n$)이라 하자. 이와 같은 각 변수별 Neyman 최적배분의 값들로 이루어진 크기 $L\times p$의 행렬을 Neyman 최적배분행렬이라 하고 이를 아래 식(2.1)과 같이 N^{o}로 나타내자.

$$N^{o} = \{n_{hj}^{o}\},\ h=1,2,\cdots,L;\ j=1,2,\cdots,p \qquad (2.1)$$

단 $n_{hj}^{o} = \dfrac{W_h S_{hj}}{\sum_{h=1}^{L} W_h S_{hj}} n$, S_{hj}은 h층에서의 j번째 변수의 모집단 표준편차이고 $W_h = \dfrac{N_h}{N}$.

N^{o}로부터 우리는 각 변수별 층화표본평균 $\bar{y}_{jst}$, $j=1,2,\cdots,p$와 그의 분산 $V(\bar{y}_{jst})$, $j=1,2,\cdots,p$를 얻을 수 있고, 이는 Neyman 최적배분에 기초하므로 각각의 변수에 대해서는 최소분산을 가진다. 즉, 행렬 N^{o} 는 각 변수별 추정량 측면에서 최적배분을 나타낸다.

이 장에서는 행렬 N^{o}에 기초한 층화표본평균벡터 $\bar{\mathbf{y}}_{st} = (\bar{y}_{1st}, \bar{y}_{2st}, \cdots, \bar{y}_{jst}, \cdots, \bar{y}_{pst})'$의 공분산행렬을 재구성하여 이를 층화표본평균벡터의 '적절한' 공분산행렬로 간주할 때, 이 재구성된 공분산행렬에 대하여 어떤 기준에 가장 근접한 공분산행렬을 생성시키는 층별 공통의 표본크기를 구하는 절충방안을 모색하고자 한다.

우선 행렬 N°를 근거로 하는 $\bar{\mathbf{y}}_{st}$의 재구성된 공분산행렬, $\Sigma_{st}^{\circ}=\mathrm{Cov}(\bar{\mathbf{y}}_{st})$는 다음 방법에 의하여 얻어진다. 여기서 각 변수별 분산은 Neyman 최적배분의 분산을, 그리고 서로 다른 변수들 간의 공분산은 행렬 N°에서 서로 다른 두 열에 있는 원소 중 작은 값을 해당하는 층의 크기로 하여 계산된 것이다. 이와 같은 방법에 따라 얻어진 크기 $p\times p$의 Σ_{st}°는 다음과 같이 쓸 수 있다.

$$\Sigma_{st}^{\circ} = \{\sigma_{ij}^{\circ}\} \tag{2.2}$$

단, $\sigma_{ij}^{\circ}=\mathrm{Cov}(\bar{y}_{ist},\bar{y}_{jst})=\sum_{h=1}^{L}W_h^2\rho_{hij}S_{hi}S_{hj}\left(\frac{1}{n_{h:ij}^{\circ}}-\frac{1}{N_h}\right)$, $\quad i,j=1,\cdots,p$

ρ_{hij}는 h층에서 (i, j)변수간의 모집단 상관계수, $n_{h:ij}^{\circ}=\min(n_{hi}^{\circ},\ n_{hj}^{\circ})$

여기서 두 층화표본평균 사이의 공분산 σ_{ij}°는 $n_{h:ij}^{\circ}$를 사용하고 있다. 이는 두 변수 i와 j에 대한 Neyman 최적배분 중 적은 쪽의 값으로서 이 방법은 곧 고려되는 두 변수에 대해 짝지은 제거(pairwise deletion)방법에 따라 공분산을 구한다는 것을 의미한다. 이와 같이 짝지은 제거 방법을 통하여 얻은 전체 공분산행렬은 양정치성(positive definiteness)이 보장되지 못한다는 것이 알려져 있다.

크기 $L\times p$인 Neyman 최적배분 행렬 N°는 각 층에 대해 변수별 서로 다른 표본단위의 추출을 의미하고 있어, 다변량 절층배분이라는 관점에서 N°를 요약한 형식의 크기 $L\times 1$인 열 벡터 $\mathbf{n}^{*}=(n_1,n_2,\cdots,n_h,\cdots,n_L)'$을 필요로 한다. 실제 각 변수별 Neyman 최적배분들의 평균 또한 크기 $L\times 1$벡터, $\bar{\mathbf{n}}=\frac{N^{\circ}\mathbf{1}}{p}$, $1'=(1,\cdots,1)$로서 N°의 단순한 선형결합으로 모든 변수별 Neyman 최적배분에 $1/p$를 가중치로 하는 것을 의미한다. 여기서는 N°에 기초하여 재구성한 Σ_{st}°를 $\bar{\mathbf{y}}_{st}$에 대한 적절한 공분산행렬로 간주할 때 이를 가장 잘 적합시키는 열 벡터 $\mathbf{n}^{*}$를 구하여 이를 기초로 하는 다변량 절층배분을 고려한다. 이에 따라 $\mathbf{n}^{*}$에 기초한 층화표본평균벡터의 공분산행렬을 $\Sigma_{st}(\mathbf{n}^{*})=\{\sigma_{ij}(\mathbf{n}^{*})\}$로 표기할 때 이는 $\Sigma_{st}(\mathbf{n}^{*})$와 Σ_{st}°의 차이를 가장 적게 하는 $\mathbf{n}^{*}$를 구하는 문제로 귀착된다. 다음 장에서 이 차이를 최적화하는 몇 가지 기준을 고려한다.

3 최적적합을 위한 판정 기준들

앞에서의 방법에 따라 재구성된 $\bar{\mathbf{y}}_{st}$의 공분산행렬 Σ_{st}°는 행렬 N°를 기초로 하고 있으므로, 이는 각 변수별 최적배분에 의한 서로 다른 표본의 크기에 근거하여 얻어진 것이다. 이와 같은

Σ_{st}^{o}를 층화표본평균벡터에 대한 '적절한' 공분산행렬로 가정할 때 모든 변수에 대해 동일한 층화표본 $\mathbf{n}^{*}=(n_1, n_2, \cdots, n_h, \cdots, n_L)'$을 기초로 한 공분산행렬 $\Sigma_{st}(\mathbf{n}^{*})$과 비교하여 두 행렬 Σ_{st}^{o}, $\Sigma_{st}(\mathbf{n}^{*})$의 차이를 가장 작게 하는 $\mathbf{n}^{*}$을 찾아내는 것이 여기서의 주목적이다. 이를 위해 두 행렬의 차이 정도를 재는 척도로서 다섯가지 판정기준을 고려하고 이를 최적화시켜 절충배분을 얻는 과정을 살펴본다.

3.1 비가중 최소제곱(Unweighted Least Squares) 기준

이 기준은 두 행렬 Σ_{st}^{o}, $\Sigma_{st}(\mathbf{n}^{*})$의 차이를 대응원소들의 차이의 제곱합으로 보는 것으로 $\sum_{h=1}^{L} n_h = n$ 인 조건하에서 다음 식을 최소로 하는 $\mathbf{n}^{*}=(n_1, n_2, \cdots, n_h, \cdots, n_L)'$을 구하여 이를 다변량 절충배분으로 간주하는 것이다.

$$\Psi_1(\mathbf{n}^{*}) = \mathrm{tr}[\Sigma_{st}^{o}-\Sigma_{st}(\mathbf{n}^{*})]^{2} = \sum_{i=1}^{p}\sum_{j=1}^{p}[\sigma_{ij}^{o}-\sigma_{ij}(\mathbf{n}^{*})]^{2} \qquad (3.3)$$

3.2 절대차이(Absolute Difference) 기준

이는 두 행렬의 원소별 차이의 절대값의 합을 최소로 하는 $\mathbf{n}^{*}$를 찾는 방법으로 이 경우에 대한 목적함수는 다음과 같다.

$$\Psi_2(\mathbf{n}^{*})=\sum_{i=1}^{p}\sum_{j=1}^{p}|\sigma_{ij}^{o}-\sigma_{ij}(\mathbf{n}^{*})| \qquad (3.4)$$

3.3 최대최소(Max-min) 기준

이는 가능한 비영(non-null)상수벡터 $\mathbf{a}$, $(\mathbf{a}\neq\mathbf{0})$와 층화표본평균벡터와의 가중결합 $\mathbf{a}'\,\bar{\mathbf{y}}_{st}$의 분산을 Σ_{st}^{o}와 $\Sigma_{st}(\mathbf{n}^{*})$의 입장에서 대비시키는 아래와 같은 목적함수 Ψ_3을 고려한다.

$$\Psi_3(\mathbf{a}, \mathbf{n}^{*})=\frac{\mathbf{a}'[\Sigma_{st}^{o}-\Sigma_{st}(\mathbf{n}^{*})]\ \mathbf{a}}{\mathbf{a}'\Sigma_{st}(\mathbf{n}^{*})\ \mathbf{a}} \qquad (3.5)$$

이때 층화절충배분의 관점에서 이 함수의 최적화는 Ψ_3을 모든 가능한 $\mathbf{a}$에 걸쳐 최소화하고, 그 결과를 $\mathbf{n}^*$에 대해 최대화하는 최대최소(Max-min)의 전략을 생각할 수 있을 것이다. 우선 $\mathbf{0}$이 아닌 모든 가능한 $\mathbf{a}$에 대해 Ψ_3의 최소는 $(\Sigma_{st}^{-1}(\mathbf{n}^*)\Sigma_{st}^{o})$의 최소고유값에 대응된다. 따라서 이 기준에 따른 배분은 모든 가능한 $(\Sigma_{st}^{-1}(\mathbf{n}^*)\Sigma_{st}^{o})$의 최소고유값들 중 그 값을 최대화하는 $\mathbf{n}^*$를 찾는 방법이 된다

3.4 우도비(Likelihood Ratio) 기준

다음과 같이 일반화 우도비의 성질에 기초한 기준을 고려할 수 있다. 우선 확률벡터 X가 p-변량 정규분포 $N_p(\mu,\Sigma)$를 따른다고 할 때 다음과 같은 구형(sphericity)가설 $H_o: \Sigma = I$를 고려해 보자. 이 때 크기 N의 표본 $x_1, x_2, \cdots, x_N$에서 얻어지는 수정된 우도비 검정통계량 $\lambda^* = e^{\frac{1}{2}pn}(|S|e^{-tr(A)})^{\frac{1}{2}n}$, 단 $S = \frac{1}{n}A$, $A = \sum_{\alpha=1}^{N}(x_\alpha-\bar{x})(x_\alpha-\bar{x})'$, $n=N-1$로 주어진다. (Sugiura 와 Nagao (1968)). 우선 $\Sigma_{st}^{*} = \Sigma_{st}^{o-1/2}\Sigma_{st}(\mathbf{n}^*)\Sigma_{st}^{o-1/2}$라 할 때 $\Sigma_{st}^{*} = I$는 곧 Σ_{st}^{o}와 $\Sigma_{st}(\mathbf{n}^*)$의 완전한 일치를 의미한다. 따라서 다음과 같은 목적함수 Ψ_4를 고려할 때

$$\Psi_4(\mathbf{n}^*) = |\Sigma_{st}^{*}|e^{-tr(\Sigma_{st}^{*})} \tag{3.6}$$

이 함수의 값을 최대로 하는 $\mathbf{n}^*$은 곧 위 행렬이 우도비의 관점에서 Σ_{st}^{*}가 I에 가장 근접하는 $\mathbf{n}^*$이 되는 것이다.

3.5 합교(Union-Intersection) 기준

3.4가 우도원리에 따른 검정통계량에 기초한 방법이라면, 여기에서 고려하는 기준은 동일한 가설에 합교원리를 적용했을 때의 검정통계량에 해당된다. 즉, 어떤 확률벡터의 분산행렬이 주어진 Σ^o와 같다는 귀무가설 $H_o: \Sigma=\Sigma^o$에 대한 합교 검정통계량은 $\Sigma_o^{-1}S$의 최대, 최소고유값의 함수가 된다. 이 가설은 $H_o: \Sigma_{st}^{o-1/2}\Sigma_{st}(\mathbf{n}^*)\Sigma_{st}^{o-1/2} = I$와 같은 구형가설로 고려할 수 있다. 이 내용을 3.4에서와 같이 다변량 절충배분의 경우에 적용하면 다음과 같다. 즉, 귀무가설은 곧 $\Sigma_{st}^{o-1/2}\Sigma_{st}(\mathbf{n}^*)\Sigma_{st}^{o-1/2}$의 최대, 최소고유값이 모두 1일 때를 의미하므로 Σ_{st}^{o}가 주어진 경우 $H_o: \Sigma_{st}^{o-1/2}\Sigma_{st}\Sigma_{st}^{o-1/2} = I$와 동일하므로 이 행렬의 최대, 최소고유값, 즉 $Ch_{max}(\Sigma_{st}^{*})$, $Ch_{min}(\Sigma_{st}^{*})$이 지나치게 극단적이지 않을 경우에 해당하는 $\mathbf{n}^*$를 구하면 될 것이다. 따라서 이 기

준은 $[\ Ch_{max}(\Sigma_{st}^{*}) - Ch_{min}(\Sigma_{st}^{*})]$ 를 최소로 하는 $\mathbf{n}^{*}$을 구하고자 한다.

앞에서 고려한 다섯가지 기준에 따른 목적함수 $\Psi(\mathbf{n}^{*})$에는 모두 n_h가 포함되어 있으므로 이를 최적화하기 위해서는 반복적인 절차가 필요하게 된다. 이 연구에서는 편의상 n_h에 대한 초기값으로 Neyman 최적배분의 평균값을 사용하였다.

4. 사례 연구

Ghosh(1958), Huddleston, Claypool 과 Hocking(1970), Sukhatme 와 Sukhatme (1970), Schuenemeyer(1975), Benn 과 Burmeister(1978), Kish(1988), Kim(1993), Kim 과 Kim(1994)는 주로 기존에 제안된 여러 가지 다변량 절충배분방법에 따른 배분값의 각 층화표본평균벡터의 총분산(분산들의 합)을 기준하여 비교하였다. 이 연구에서는 기존에 제안된 방법과 이 논문에서 고려하는 방법들을 비교하기 위해 주어진 절충배분에 따라 층화표본 평균벡터의 공분산행렬로 부터 총분산 및 일반화분산의 값을 통해 비교하고, 상관계수에 따른 배분값의 변화의 추이를 알아본다. 비교를 위해 Schuenemeyer(1975)에 의해 제공된 예를 각 다변량 최적배분방법에 고려하였다. 구체적으로 다음의 11가지 다변량 절충배분방법에 따른 결과를 비교, 평가한다.

1. 비례배분 (비례:A)
2. 각 변수별 Neyman 최적배분(변수1:B, 변수2:C, 변수:D)
3. Neyman 최적배분의 평균(평균:E)
4. 상대 효율적 접근방법(상대효율:F) [Dalenius(1957), Chatterjee(1967)]
5. 변수간의 상관계수을 고려, 또는 무시한 일반화분산의 최소(Ghosh/w:G, Ghosh/o:H) [Ghosh(1965)]
6. 최대고유값을 이용한 방법(고유값:I) [Schuenemeyer(1975)]
7. 비가중 최소제곱 기준에 의한 방법(최소제곱:J)
8. 절대차이 기준에 의한 방법(절대차이:K)
9. 최대최소 기준에 의한 방법(최대최소:L)
10. 우도비 기준에 의한 방법(우도비:M)
11. 합교 기준에 의한 방법(합교:N)

4.1 Schuenemeyer의 자료

다음 자료는 Schuenemeyer(1975)가 고려한 인공자료로 배분 가능한 총 표본은 $n=600$이고, 층별 모집단 크기 N_h와 공분산행렬 S_h는 다음과 같다.

$N_1=800$ $N_2=1000$ $N_3=1200$

$$S_1 = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 9 & 0 \\ 0 & 0 & 100 \end{bmatrix} \qquad S_2 = \begin{bmatrix} 16.00 & 2.77 & 7.35 \\ 2.77 & 1.92 & 2.55 \\ 7.35 & 2.55 & 13.5 \end{bmatrix} \qquad S_2 = \begin{bmatrix} 36.00 & 0.00 & 29.70 \\ 0.00 & 49.00 & 34.65 \\ 29.70 & 34.64 & 50.00 \end{bmatrix}$$

또한 층별, 변수간 상관계수가 다음과 같은 경우를 고려하였다.

$$\rho_{1ij} = 0.0, (i, j = 1,2,3) \qquad \rho_{2ij} = 0.5 (i, j = 1,2,3) \qquad \rho_{312} = 0.0$$
$$\rho_{313} = 0.7$$
$$\rho_{323} = 0.7$$

<표4.1> 각 방법들에 의해 배분된 표본의 수

층 \ 방법	비례	변수별 최적배분			평균	상대 효율	Ghosh: 상관	Ghosh: 무상관
		변수1	변수2	변수3				
1	160	75	118	238	144	151	168	143
2	200	187	68	109	121	125	129	121
3	240	338	414	253	335	324	301	336

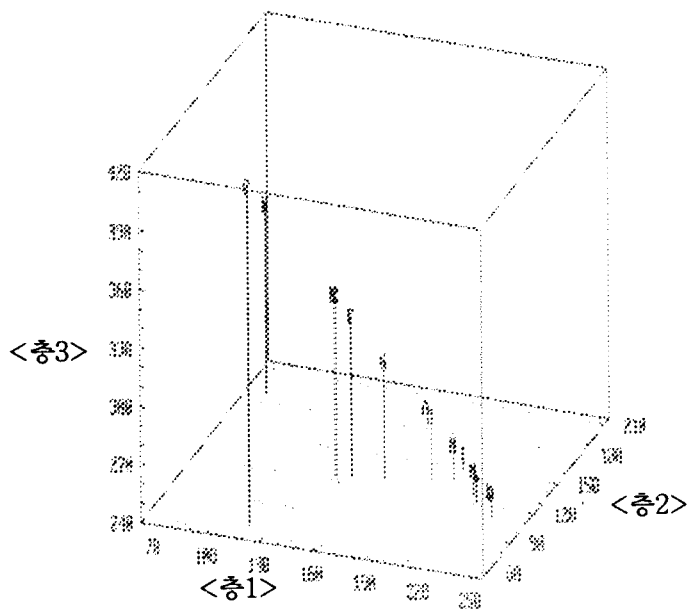
층 \ 방법	고유값	최소 제곱	절대 차이	최대 최소	우도비	합교
1	204	192	228	224	204	220
2	148	135	128	120	136	126
3	248	273	244	256	260	254

위의 자료에 대해 얻어지는 절층배분의 결과는 <표4.1>과 같다. 우선 "변수별 최적배분"란에서 변수1의한 배분방법이 변수2, 변수3에 의한 배분방법과는 상당히 다르다는 것을 알 수 있다. 이것은 한 변수의 최적배분이 다른 변수에서의 최적배분과는 상호부합하지 않음을 말해주고 있다. 우선 Schuenemeyer자료는 층1의 경우 변수들이 서로 독립적이나 변수3이 다른 변수에 비해 매우 큰 분산을 가지고 있으며, 이는 층2, 층3의 경우 변수별 분산의 차이가 그리 크지 않다거나, 변수들 간의 상관관계로 인해 상호 변화의 추이에 관한 정보를 공유한다는 점에서 크게 구별된다고 할 수 있다. 이런 특징들로부터 층1은 다른 층에 비해 상대적으로 큰 변화의 추이를 가진다고 하겠다. 기존에 제안된 절층배분(상대효율, Ghosh의 경우)결과는 층1, 층2에 비해 층3에 큰 값을 배분하고 있다. 이에 대해 Schuememeyer가 제안한 고유값 기준이나 이 연구에서 제안한 기준(비가중 최소제곱 기준, 절대차이 기준, 최대최소 기준, 우도비 기준, 합교 기준)들에 따른 결과는 층1에 비중을 둠으로써 층2의 배분은 변화가 없는 한편 층3의 배분값은 상대적으로 감소하는 결과를 보여주고 있다. 또한 층별 다변량 변화의 추이를 나타내는 다른 측도로

$$\mathrm{Tr}(S_1) = 113 \qquad \mathrm{Tr}(S_2) = 31.42 \qquad \mathrm{Tr}(S_3) = 135$$

$$|S_1| = 3600 \qquad |S_2| = 207.2 \qquad |S_3| = 1755.8$$

들을 볼 때 층1과 층3간의 차이가 그리 큰 차이가 나지 않음을 알 수 있어 이에 따른 층별 배분도 크지 않을 것이 기대된다.



<그림4.1> 각 방법들에 의해 배분된 표본의 수

아래 <그림4.1>은 각 방법에 의한 결과인 <표4.1>의 값을 도형으로 표현한 것이다. 이 그림에서 영문자 A, B, C...는 앞의 여러 방법에서 제시한 것들이다. 이 그림을 통해 알 수 있는 것은 각 변수별 Neyman 최적배분의 평균(E)과 상관을 고려하지 않은 Ghosh 방법(H)이 거의 같고 이 논문에서 제안한 다섯가지 방법(J,K,L,M,N)은 아래쪽으로 몰려 거의 비슷한 값을 가짐을 알 수 있다.

<표4.2>는 각 방법에 의한 층화평균벡터의 변수별 분산과 총분산 그리고 일반화분산을 나타낸 것이다. 일반적으로 배분방법에 대한 평가는 총분산과 일반화분산을 통해 이루어진다. 여기서는 전반적인 상황으로 보아 상관관계를 무시한 Ghosh의 방법과 비가중 최소제곱 기준이 어느 정도 작은 분산을 나타내었다. 이 연구에서 제안한 다섯가지 방법의 총분산은 절대차이 기준을 제외한 나머지 방법들은 비교적 총분산 면에서 작은 값을 나타내었다. 그 중에서도 비가중 최소제곱 기준에 의한 방법은 가장 작은 값을 나타내었는데 이는 특이하게도 상관관계을 무시한 Ghosh의 방법 다음으로 좋은 결과를 나타내었다. 또 Schuenemeyer의 자료에 대해 Schuenemeyer가 제안한 방법인 고유값 기준에 의한 방법보다도 절대차이 기준을 제외한 다른 방법이 모두 더 좋은 결과를 제공하고 있다. 또한 여기서도 각 변수 별 Neyman 최적배분의 평균에 의한 방법이 어느 정도의 최소의 분산을 나타내었지만 각기 다른 절충배분과 비교한다면 그렇게 고무적인 것은 못됨을 알 수 있다.

<표4.2> 각 방법에 의한 층화표본평균벡터의 변수별 분산, 총분산 그리고 일반화분산

분산 \ 방법	비례	변수별 최적배분			평균	상대 효율	Ghosh: 상관	Ghosh: 무상관
		변수1	변수2	변수3				
1	0.0277	0.0234	0.0352	0.0333	0.0269	0.0270	0.0269	0.0276
2	0.0302	0.0254	0.0200	0.0282	0.0221	0.0226	0.0221	0.0238
3	0.0682	0.1095	0.0843	0.0582	0.0686	0.0666	0.0689	0.0632
총분산	0.1262	0.1582	0.1396	0.1197	0.1176	0.1162	0.1179	0.1146
일반화분산	0.3786	0.5770	0.5136	0.3389	0.3230	0.3148	0.3244	0.3055

분산 \ 방법	고유값	최소 제곱	절대 차이	최대 최소	우도비	합교
1	0.0297	0.0288	0.0318	0.0317	0.0297	0.0311
2	0.0278	0.0261	0.0291	0.0277	0.0273	0.0279
3	0.0602	0.0604	0.0586	0.0584	0.0396	0.0587
총분산	0.1186	0.1153	0.1196	0.1178	0.1166	0.1177
일반화분산	0.3222	0.3072	0.3293	0.3234	0.3131	0.3209

<표4.3> 상관계수 변화에 따른 각각의 절충배분

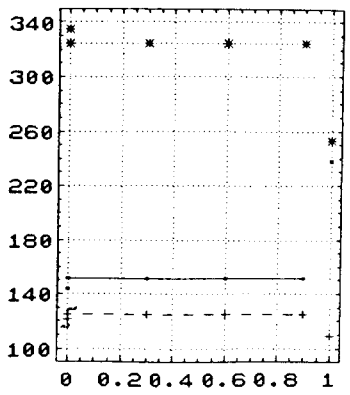
방 법	층	상 관 계 수			
		0.0	0.3	0.6	0.9
상대 효율 기준	1	151	151	151	151
	2	125	125	125	125
	3	324	324	324	324
Ghosh 방법 상관	1	143	143	140	141
	2	121	121	119	160
	3	336	336	341	229
Ghosh 방법 무상관	1	143	143	143	143
	2	121	121	121	121
	3	336	336	336	336
최대 고유값 기준	1	224	215	182	127
	2	141	148	163	191
	3	235	237	255	382
최소 제곱 기준	1	177	182	208	238
	2	124	125	148	148
	3	299	293	244	214

방 법	층	상 관 계 수			
		0.0	0.3	0.6	0.9
절대 차이 기준	1	182	213	237	238
	2	120	121	120	119
	3	298	266	243	243
최대 최소 기준	1	152	212	236	1)
	2	129	117	134	
	3	319	271	230	
우도비 기준	1	151	156	176	75
	2	126	126	121	69
	3	323	318	303	456
합교 기준	1	151	210	238	209
	2	129	118	125	108
	3	320	272	237	283

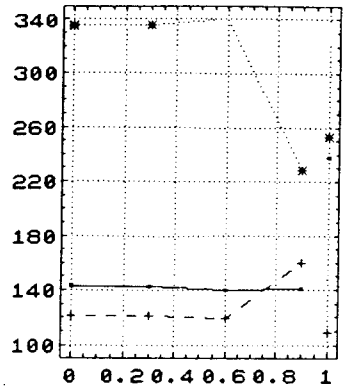
1) 최대최소 기준에서 공백은 계산 과정에서 수렴하지 않는 경우이다.

4.2 층별 상관구조의 변화에 따른 절충배분의 비교

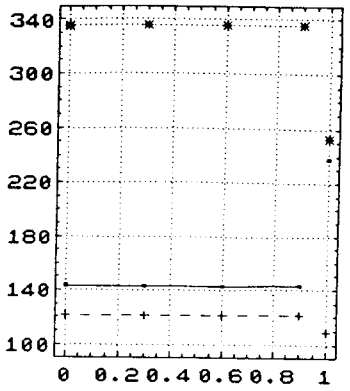
Schuenemeyer(1975)가 고려한 자료는 각 층간에 변수별 상관계수가 단 하나의 고정된 값이나



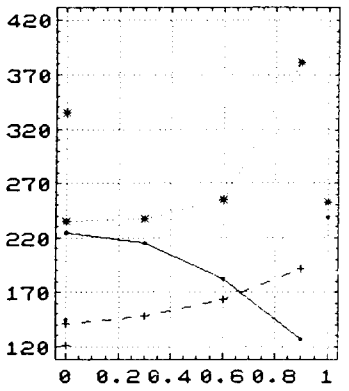
1) 상대효율 기준



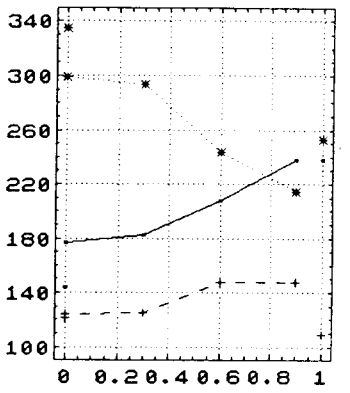
2) Ghosh 기준 (상관)



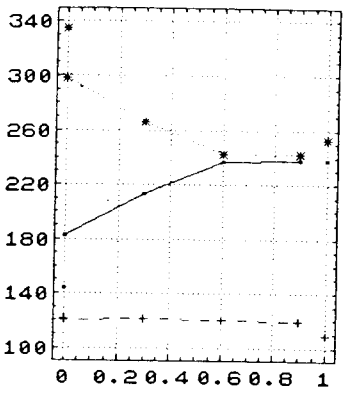
3) Ghosh 기준(무상관)



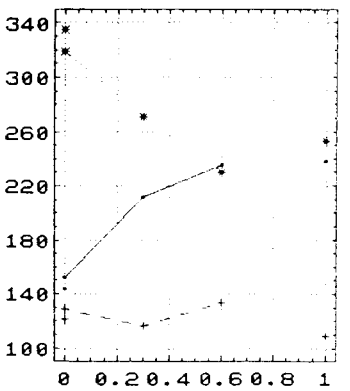
4) 고유값 기준



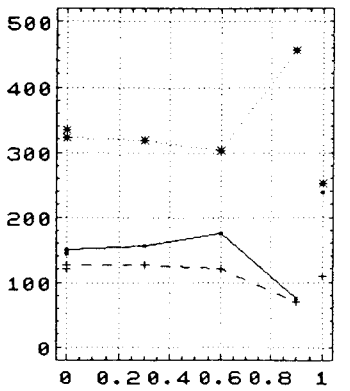
5) 비가중최소제곱 기준



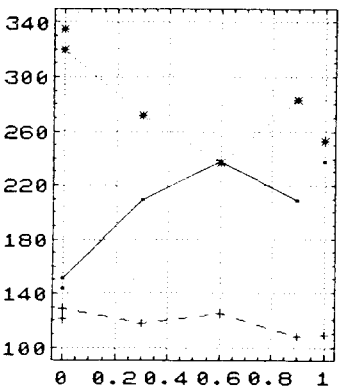
6) 절대차이 기준



7) 최대최소 기준



8) 우도비 기준



9) 합교 기준

1) +(- - -) : 층1, ·(----) : 층2, *(·····) : 층3
2) X축은 상관계수,Y축은 배분값을 가리킨다.

<그림.4.2> 각 상관계수 변화에 따른 방법 및 절충배분

이 절에서는 상관계수가 각 층에 관계없이 일률적으로 $\rho_{hij}=0.0$, $\rho_{hij}=0.3$, $\rho_{hij}=0.6$, $\rho_{hij}=0.9$ ($h=1,\cdots,3$, $i,j=1,2,3$)로 변화시켰을 때 다변량 절충배분의 추이를 알아본다.

<표4.3>은 상관계수가 커짐에 따라 절충배분의 경향을 나타내고 있다. 먼저 상관계수가 0일 때는 대부분의 방법에 따른 결과가 각 변수별 Neyman 최적배분의 평균배분과 유사하다는 것을 알 수 있고 상관관계가 커짐에 따라 배분 결과는 변수3의 경우[<표4.1>의 "변수별 최적배분"에서 '변수3']에 가까워지는 경향이 있음을 알 수 있다. 이는 상관관계가 낮을수록 각 배분값이 Neyman 최적배분의 평균값에 가깝다는 것과 상관관계가 높을수록 비교적 각 층별로 큰 분산을 지닌 변수3의 배분값으로 옮겨지는 경향이 있는 것을 알 수 있다. 이 표를 근거로 한 그림 <그림4.2>와 연관시켜 보면 특히 비가중 최소제곱 기준과 최대최소 기준, 합교 기준은 상관계수가 클수록 분산이 큰 변수(변수3)로 접근하는 경향이 있는 반면 우도비 기준은 상관계수에 대해 덜 민감하면서 상관계수가 0.9일 때는 특이한 배분을 낳고 있다. 그림에서 볼 수 있는 것은 상관계수가 없는 상대효율적 접근 방법, 변수간의 상관계수를 무시한 Ghosh 방법(일반화분산의 최소) 등은 상관계수에 변함이 없고, 보편적으로 절대차이 기준, 최대최소 기준, 우도비 기준, 합교 기준에 의한 방법 등은 한 층의 변화가 층2를 기준으로 하여 상관계수 0.5를 중심으로 볼록하게 변하고 반대로 나머지 층들 중 하나는 오목하게 변하는 것이 특징이다. 물론 상관계수가 음수의 값을 취할 경우 배분결과에는 변화가 있으나 방법에 따른 전반적인 경향은 유사함이 발견되었다.

5. 결론

이 논문에서는 각 변수별 Neyman 최적배분에 근거해서 얻은 층화평균의 공분산행렬을 가장 잘 적합시키는 적합 기준으로서 5가지의 유형을 제시하여 모든 변수에 대해 층별로 동일한 표본의 크기를 구하였다. 이에 대해 비가중 최소제곱 기준 등이 적절한 배분을 제공해 주고 있으며, 특히 상관계수를 고려할 때 상관계수가 0에 가까울 경우에 모든 일변량 Neyman 최적배분의 평균에 가까운 결과를 나타내었고, 상관계수가 커질수록 큰 분산을 가진 변수에 기초한 일변량 Neyman 최적배분방법의 결과로 귀착되어, 상관계수가 극단적인 값을 취하지 않고 중간 정도일 경우는 각 방법들을 비교, 분석하여 각 변수들의 총분산을 최소로 하는 방법들을 선택할 수 있다.

참 고 문 헌

[1] Aoyama, M.(1963). Stratification Random Sampling with Optimum Allocation for Multivariate Population, *Annals of the Institute of Statistical Mathemathics*, 14, 251-258.

[2] Benn,J. R. and Burmeister, L. F.(1978). A Review of Optimal Allocation for Multivariate Purpose Surveys, *Biometrical Journal*, 20, 1-14.

[3] Bethel, J. W.(1985). An Optimum Allocation Algorithm for Multivariate Surveys, *Proceedings of the Social Statistics Section, American Statistical Association*, 209-212.

[4] Bethel, J. W.(1989). Sampling Allocation in Multivariate Surveys, *Survey Methodology*, 47-57.

[5] Chakravarti, I. H.(1955). On the Problem of Planning a Multistage Survey for Multiple Correlated Characters, *Sankhya*, 14, 211-216.

[6] Chatterjee, S.(1968). Multivariate Stratified Survey, *Journal of American Statistical Association*, 63, 530-534.

[7] Chatterjee, S.(1972). A Study of Optimum Allocation in Multivariate Stratified Survey, *Skandinavisk Actuarietidskrift*, 55, 73-80.

[8] Chromy, J. R.(1987). Design Optimization with Multiple Objects, *Proceedings of the Social Statistics Section, American Statistical Association*, 194-199.

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

[10]. Dalenius, T.(1957). *Sampling in Sweden*, Almqvist and Wicksell, Stockholm.

[11] Folks, J. L. and Antle, C. E.(1965). A Optimum Allocation of Sampling Units to Strata When There are R Responses of Interest. *Journal of the American Statistical Association*, 59, 225-233.

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

[13] Hartley, H. O.(1965). Multiple Purpose Optimum Allocation in Stratified Sampling, *Proceedings of the Social Statistics Section, American Statistical Association*, 59, 258-261.

[14] Huddleston, H. F., Claypoll, P. L. and Hocking, R. R.(1970). Optimum Sample Allocation to Strata using Convex Programming, *App. Stat*, 19, 273-278.

[15] Kim, H.(1993). A Study on the Compromising Allocation in Multivariate Stratified Random Sampling. unpublished Ph.D. Dissertation, Korea University.

[16] Kim, K. and Kim, H.(1994) On Compromise Allocation in Multivariate Stratified Random Sampling, *Proceedings of the Eight Japan and Korea Conference of Statistics*, 19-24.

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

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

[19] Kokan, A. R. and Khan, S.(1967). Optimum Allocation in Multivariate Surveys : An Analystical Solution, *Journal of the Royal Statistical Society Series B*, 29, 115-125.

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

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

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

[23] Sugiura, N. and Nagao, H.(1968). Unbiasedness of same Test Criteria for the Equality of one or two Covariance Matrices, *Annals of Mathematical Statistics*, 39, 1686-1692.

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

[25] Yates, R.(1960). *Sampling Methods for Censuses and Surveys*, Charles Griffin and Company, Limited, London, 3th, edition.