분산이 미지인 정규모집단의 평균에 대한 베이즈−P* 선택방법에 관한 연구+

김 우철*, 전 종우*, 한 경수**

〈요 약〉

정규분포를 가정한 통상적인 일원배치모형에서 모평균들을 비교하는 부분집합 선택 방법으로서 베이즈−P* 선택방법을 제시하고 기존의 방법과의 관계를 알아보고, 그 운용특성에 대한 모의실험의 결과를 고찰하였다.

1. 서 론

정규분포를 가정한 일원배치모형에서 모평균 $\theta_1, \cdots, \theta_k$를 비교하는 한 방법론으로서 부분집합 선택론은 Gupta와 Panchapakesan(1979)에서 볼 수 있듯이 여러 측면에서 연구되어 왔다.

베이지안 이론에 의한 연구도 많은 학자에 의해 수행되었으며, 본 연구와 관련된 것으로는 Roth(1978), Gupta와 Yang(1985), 그리고 Berger와 Deely(1988)에 의한 연구결과가 있다.

이들은 모두 각 처리에 대해 $\theta_i = \max_{1 \le j \le k} \theta_j$일 사후확률 p_i에 주안점을 두고 연구하였다. 구체적으로 Berger와 Deely는 계층적 베이즈 모형하에서 p_i의 계산과 이에 따른 선택방법과 동일성 검정을 고려하였고, Gupta와 Yang은 2절 에서 소개될 베이즈 −P*선택방법을 제시하고 이의 최적성질을 고려하였다. 또한, Roth의 연구결과는 피듀시알 확률적 입장에서 시도된 것으로써, 이는 사전정보가 없는 경우의 사전분포에 대한 베이즈 −P*선택방법과 기존의 Gupta(1956)형 선택방법과의 관계를 규명한 것으로 요약할 수 있다.

그러나, 이들의 연구는 물론 기존의 베이지안 선택방법은 모두 공통분산인 σ^2이 기지인 경우에 국한하여 연구되었다. 이는 이론적으로는 흥미롭지만 실제 상황과는 거리가 많다고 할 수 있다. 따라서 본 연구의 동기는 공통분산이 미지인 경우에 Gupta와 Yang이 제시한 조건에 적합한 베이즈 −P*선택방법을 정의하고, 이의 운용특성에 대해 알아보고자 하는 것이었다.

+본 연구는 1988년도 한국과학재단 연구비 지원(과제번호 : 88−07−23−02)에 의한 연구결과임.
* 서울 대학교 자연과학대학 계산통계학과, 서울시 관악구 신림동 150−742
** 전북대학교 자연과학대학 통계학과, 전주시 효자동 560−240

2절에서는 베이즈 -P*선택방법을 정의하고, 3절에서는 표본크기가 같을때에 베이즈 -P* 선택방법과 Gupta형 선택방법의 관계를 알아보고, 베이즈 -P*선택방법의 운용특성을 Gupta형 선택방법과 비교하는 모의실험의 결과를 수록하였다.

2. 베이즈 -P* 선택방법

정규분포를 가정한 통상적인 일원배치모형에서 반복수가 $n_1, \cdots, n_k$인 표본평균을 $x_1, \cdots, x_k$, 공통분산의 불편추정량을 s^2이라 하면

$$x_i \sim N(\theta_i, \sigma^2/n_i) \ (i=1, \cdots, k), \ rs^2 \sim \sigma^2 x^2(r) \tag{2. 1}$$

이고 이들은 서로 독립이다. 여기에서 $r = \sum_{i=1}^{k}(n-1)$이다.

이제 $\theta_1, \cdots, \theta_k$와 σ^2의 사전분포가 $\pi(\theta, \sigma^2)$으로 주어진 경우에, 모평균 $\theta_1, \cdots, \theta_k$중에서 최대인 $\max_{1\leq j\leq k}\theta_j$에 대응하는 최우수모집단을 선택할 목적으로 $\{1, 2, \cdots, k\}$의 부분집합 $S=S(x, s^2)$를 선택한다면 Gupta와 Yang(1985)이 제시한 사후확률 $P^*(1/k<P^*<1)$조건은 다음과 같다.

$$P\{\max_{i\in S}\theta_i = \max_{1\leq j\leq k}\theta_j \mid x, s^2\} \geqq P^* \tag{2. 2}$$

이 사후확률 조건의 의미를 해석하면, 선택된 모집단 중에 최우수 모집단이 포함될 사후확률이 미리 지정된 P^*이상이도록 부분집합 s를 선택하여야 한다는 뜻이다.

사후확률 조건을 만족시키면서 선택되는 모집단의 갯수인 S의 크기 $|S|$는 작을 수록 바람직하다. 그런데 각 모집단이 최우수 모집단일 사후확률을

$$p_i = P\{\theta_i = \max_{1\leq j\leq k}\theta \mid x, s^2\} \tag{2. 3}$$

이라 하면, 이는 $\sum_{i\in S} p_i \geqq P^*$이면서 $|S|$를 최소로 하는 문제이다.

따라서 Lagrange의 방법을 이용하여

$$|S| - \lambda P\{\max_{i\in S}\theta_i = \max_{1\leq j\leq k}\theta_j \mid x, s^2\} = \sum_{i\in S}(1-\lambda p_i)$$

를 최소로 하면 된다. 즉 $p_1, \cdots, p_k$ 중에서 큰 순서로 사후확률(2. 3)을 만족시키도록 S를 정하면 된다.

이와 같이 하여 정해지는 선택방법을 Gupta와 Yang(1985)은 베이즈 $-P^*$선택방법이라고 불렀다. 즉, $p_{[1]}<\cdots<p_{[k]}$가 (2. 3)에서의 $p_1, \cdots, p_k$를 순서대로 늘어놓은 것이라면,

$$m_B = \min\{m : p_{[k]} + \cdots + p_{[k-m+1]} \geqq P^*\} \tag{2. 4}$$

를 정하여 부분집합

$$S_B = \{[k], [k-1], \cdots, [k-m_B+1]\} \tag{2. 5}$$

을 선택하는 것이 (비랜덤)베이즈 P^*선택방법이다.

이러한 베이즈 P^* 선택방법을 적용하려면 (2. 3)의 p_i를 계산해야 한다. 흔히 적용되는 공액 사전분포로서

$$\theta_i \mid \sigma^2 \sim N(\mu_i, \sigma^2/\lambda_i),\ 2\beta/\sigma^2 \sim \chi^2(2\alpha)$$

이고 σ^2이 주어진 경우에 $\theta_1, \cdots, \theta_k$가 서로 독립된 경우를 생각하면, (2. 3)의 p_i는

$$p_i = \int_0^\infty \int_{-\infty}^{\infty} \prod_{j \neq i} \Phi\left(\frac{\gamma_i}{\gamma_j} y + \frac{t_i - t_j}{\hat{\sigma}/\sqrt{\gamma_j}} u\right) d\Phi(y)\, dQ_i(u) \tag{2. 6}$$

로서 $\gamma_i = n_i + \lambda_i$, $\nu = k - r + 2\alpha$, $t_i = (\lambda_i \mu_i + n_i x_i)/(\lambda_i + n_i)$, $\nu\hat{\sigma}^2 = \sum_{i=1}^{k} \frac{(x_i - \mu_k)^2}{n_i + \lambda_i} + rs^2 + 2\beta$이고 Φ와 Q_i는 각각 $N(0, 1)$과 $\sqrt{\chi^2(\nu)/\nu}$의 분포함수이다.

따라서, (2. 6)의 p_i는 수치적분이나 몬테칼로 방법에 의해 계산할 수 있으므로 이들 $p_1, \cdots, p_k$를 계산하여 베이즈 P^* 선택방법을 적용할 수 있다.

3. 베이즈 $-P^*$ 선택방법의 성질과 운용특성

이 절에서는 표본크기가 모두 n으로서 서로 같고 사전정보가 없는 경우로써 사전 분포가 $\pi(\theta, \sigma^2) = 1/\sigma^2$일 때 베이즈 P^* 선택방법과 Gupta(1956)에 의해 제안된 부분집합

$$S_G = \{i : x_i \geqq \max_{1 \leq j \leq k} x_j - ds/\sqrt{n}\} \tag{3. 1}$$

를 선택하는 Gupta형 선택방법을 비교하도록 한다.

식 (3. 1)에서의 설계 상수인 d는

$$\int_0^\infty \int_{-\infty}^{\infty} \Phi^{k-1}(y + du)\, d\Phi(y)\, dQ_r(u) = P^* \tag{3. 2}$$

에 의해 정해지고 이는 Gupta, Panchapakesan과 Sohn(1985)의 수표에서 찾을 수 있고, 사전분포가 $\pi(\theta, \sigma^2) = 1/\sigma^2$ 로 주어지면

$$p_i \int_0^{\infty}\int_{-\infty}^{\infty} \prod_{j\neq i} \Phi\left(\frac{y-x_i}{s/\sqrt{n}}u\right) d\Phi\left(\frac{y-x_i}{s/\sqrt{n}}u\right)dQ_r(u) \tag{3.3}$$

주어짐을 알 수 있다.

Gupta형 선택방법에서 d에 대한 조건 (3. 2)는 빈도적 확률 조건인

$$P_{\theta}\{\max_{i\in S}\theta_i = \max_{1\le i\le k}\theta_i\} \geqq P^* \tag{3.4}$$

로 부터 얻어진 것이며, 이러한 제약조건 하에서 Gupta형 선택방법은 여러 의미에서 최적인 방법임이 알려져 있다(Gupta and Panchapakesan(1979)).

식 (2. 3)과 (3. 4)는 올바른 선택(correct selection, CS)인 사건 $\max_{i\in S}\theta_i = \max_{1\le i\le k}\theta_i$의 사후확률과 빈도적 확률에 대한 제약조건을 뜻하고, 각각의 제약조건하에서 최적인 S_B와 S_G를 비교하는 것은 의미가 있다.

이러한 비교의 의미로서 다음의 정리는 Roth(1978), Gupta와 Yang(1985)의 결과가 분산이 미지인 경우에도 성립함을뜻한다.

정리 1. 베이즈 $-P^*$ 선택방법에 의한 부분집합 S_B와 Gupta형 선택방법에 의한 부분집합 S_G사이에는 다음 관계가 성립한다.

$$S_B \subset S_G \tag{3.5}$$

증명 식 (3. 3)에서 알 수 있듯이 p_i와 x_i의 순서는 서로 대응하므로, x_i들을 순서대로 늘어놓은 것을 $x_{[1]}<\cdots<x_{[k]}$라고 하면 다음 방정식이 성립함을 알 수 있다.

$$\begin{aligned}
&p_{[k]}+p_{[k-1]}+\cdots+p_{[k-m+1]}\\
&= \int_0^{\infty}\int_{-\infty}^{\infty} \prod_{j=1}^{k-m} \Phi\left(\frac{y-x_{[j]}}{s/\sqrt{n}}u\right) d \prod_{i=k-m+1}^{k} \Phi\left(\frac{y-x_{[i]}}{s/\sqrt{n}}u\right)dQ_r(u)\\
&\geqq \int_0^{\infty}\int_{-\infty}^{\infty} \prod_{j=1}^{k-m} \Phi\left(\frac{y-x_{[j]}}{s/\sqrt{n}}u\right) d\Phi\left(\frac{y-x_{[k]}}{s/\sqrt{n}}u\right)dQ_r(u)\\
&\geqq \int_0^{\infty}\int_{-\infty}^{\infty} \Phi^{k-m}\left(\frac{y-x_{[k-m]}}{s/\sqrt{n}}u\right) d\Phi\left(\frac{y-x_{[k]}}{s/\sqrt{n}}u\right)dQ_r(u)\\
&\geqq \int_0^{\infty}\int_{-\infty}^{\infty} \Phi^{k-1}\left(y+\frac{x_{[k]}-x_{[k-m]}}{s/\sqrt{n}}u\right) d\Phi(y)\ dQ_r(u)
\end{aligned} \tag{3.6}$$

여기에서 첫 부등식은 $\Phi\left(\frac{y-x_{[k]}}{s/\sqrt{n}}u\right)$ 와 $\prod_{i=m+1}^{k} \Phi\left(\frac{y-x_{[i]}}{s/\sqrt{n}}u\right)$ 사이의 확률적 대소관계로부터, 다른 부등식은 $x_{[i]}$들의 대소관계에서 유도되는 것이다.

따라서 식 (2. 4), (3. 1), (3. 2), (3, 6)으로부터

$$m_B \leqq \min\{m : x_{[k]} - x_{[k-m]} \geqq ds\sqrt{n}\}$$
$$= \min\{m : x_{[k-m]} \leqq x_{[k]} - ds\sqrt{n}\}$$
$$= \max\{m : x_{[k-m-1]} > x_{[k]} - ds\sqrt{n}\}$$

임을 알 수 있고, 이로부터 식 (3. 5)가 성립함은 명백하다.

정리 1로부터 Gupta형 선택방법은 사후확률 조건인 식 (2. 2)를 만족함을 알 수 있고, 정리 1의 증명과정에서 알 수 있듯이 $k \geqq 3$이면 S_B는 S_G의 진부분 집합이 되어 베이즈 $-P^*$ 선택방법은 빈도적 확률조건인 식 (3. 4)는 만족시키지 못한다.

이제 베이즈 $-P^*$ 선택방법과 Gupta형 선택방법을 비교하기 위하여 평균의 등간격 배열인 $\theta, \theta+\triangle, \cdots, \theta+(k-1)\triangle$인 경우와 스리피지(slippage)배열인 $\theta, \theta, \cdots, \theta, \theta+\triangle$인 경우에 모의실험을 통하여 운용특성을 비교하였다.

각 경우에 올바른 선택의 빈도적 확률 PCS와 부분집합의 크기에 대한 기대값인 $E(S)$를 모의실험에서 상대돗수와 그의 합으로 추정한 결과가 표 1과 표 2에 수록되어 있다.

또한 Gupta형 선택방법에 대한 베이즈 $-P^*$ 선택방법의 상대효율로써

$$\text{REFF} = \frac{\text{PCS}_B}{\text{E}(S_B)} \Big/ \frac{\text{PCS}_G}{\text{E}(S_G)}$$

의 값을 표 1과 표 2에 수록하였다. 물론 이 상대효율이 1보다 크면 이는 베이즈 $-P^*$ 선택방법이 더욱 효율적임을 뜻한다.

이러한 모의실험은 $P^*=0.90, 0.95, 0.99$와 $k=4, 8$, $n=5, 10, 20$인 경우에 $\sqrt{n}\triangle/\sigma=0.2, 0.4, 0.6, 0.8, 1.0$인 경우에 수행하였으며 스리피지 배열의 경우에는 $\triangle=0$인 경우도 포함하였다. 이 모의실험에서 시뮬레이션의 횟수는 1,000번씩 하였으며, 난수는 IMSL의 난수생성 프로그램인 GGNML을 이용하였고, 정규분포의 누적분포는 MDNOR을 이용하고, 식 (3. 2)의 적분 계산에는 16점 Gauss-Hermite quadrature와 24점 Gauss-Laguerre quadrature를 사용하였다. 또한 Gupta형 선택방법의 상수 d는 Gupta, Pahchapakesan과 Sohn(1985)의 논문으로 부터 나온 값을 사용하였다. 이러한 모든 계산은 전북대학교의 CYBER 932-31을 사용하여 수행되었다.

이들 결과 중에서 $P^*=0.90$인 경우만을 발췌하여 표 1에는 등간격 배열에서의 결과를, 표 2에는 스리피지 배열에서의 결과를 수록하였다. $P^*=0.95, 0.99$인 경우도 표 1, 표 2의 결과와 동일한 현상을 나타내었다. 표 1, 표 2에서 하단의 숫자는 관측된 결과의 표준오차 중에서 최대값만을 수록한 것이다. 이들 표의 결과를 요약하면 다음과 같다.

1. 베이즈 $-P^*$ 선택방법은 평균이 서로 가까이 있을 때 PCS가 낮은 경향이 있고, 이는 처리의 수 k가 클 때 또한 모든 평균이 같을 때 더욱 뚜렷하게 나타난다.
2. Gupta의 선택방법은 이론적으로 보장되듯이 PCS가 지정된 P^* 이상을 유지하지만 평균이 조금만 떨어져 있으면 PCS가 지정된 값을 훨씬 초과하는 경향이 뚜렷하다.

3. 전반적으로 Gupta의 선택방법은 E(S)가 베이즈 $-P^*$ 선택방법에 비하여 크게 나타나고, 특히 k가 클 때에는 이러한 차이는 뚜렷하게 나타난다. 이러한 차이는 뚜렷하여 베이즈 $-P^*$ 선택방법이 PCS에서 떨어지는 점을 상쇄하고도 남게 되어, 상대효율 측면에서 베이즈 $-P^*$ 선택방법이 우수함을 알 수 있다.

이상에서 알 수 있듯이 비록 제한된 모의실험의 결과이지만 비교대상인 처리의 수가 크고 평균들이 조금이라도 다른 경우에는 베이즈 $-P^*$ 선택방법이 빈도적 측면에서도 Gupta의 방법에 비해 효율적이고, 베이지안 측면에서 사후확률에 대한 보장도 가능함을 알 수 있다. 따라서, 계산의 부담은 있으나 베이즈 $-P^*$ 선택방법은 다수의 처리를 비교함에 주요한 방법으로 고려되어야 한다.

참 고 문 헌

(1) Berger, J. O. and Deely, J. J.(1988). A Bayesian approach to ranking and selection of means with alternatives to AOV methodology, *Journal of the American Statistical Association,* 83, 364−373.

(2) Gupta, S. S.(1956). On a decision rule for a problem in ranking means, Ph. D. Thesis, University of North Carolina at Chapel Hill, Institute of Statistics.

(3) Gupta, S. S. and S. Panchapakesan(1979). *Multiple Desision Procedures : Theory and Mtehodology of Selecting and Ranking Populations.* John Wiley, New York.

(4) Gupta, S. S. and Yang, H. M(1985). Bayes P* subset selection procedures for best population, *Journal of Statistical Planning and Inference.* 12, 213−233.

(5) Gupta, S. S., Panchapakesan, S., and Sohn, J.(1985). On the distribution of the Studentized maximum of equally correlated normal random variable, *Communications in Statistics-Simulation and Computation* 14, 103−135.

(6) Roth, A. J.(1978). A new procedure for selecting a subset containing the best normal population, *Journal of the American Statistical Association,* 73, 613−617.

〈표 1〉 등 간격 배열의 경우

(a) k = 4, P* = 0.90

$\sqrt{n}\Delta/\sigma$	n	PCS		E(S)		REFF
		Bayes	Gupta	Bayes	Gupta	
0.2	5	.8900	.9740	2.6040	3.4110	1.1969
	10	.9440	.9900	2.4210	3.2390	1.2757
	20	.9610	.9880	2.2120	2.8460	1.2515
0.4	5	.9680	.9920	2.2220	2.9130	1.2793
	10	.9910	.9970	1.8870	2.4460	1.2884
	20	.9930	.9960	1.5840	1.8880	1.1883
0.6	5	.9780	.9920	1.8870	2.3460	1.2257
	10	.9970	1.0000	1.5580	1.9170	1.2267
	20	.9990	.9990	1.2580	1.4370	1.1423
0.8	5	.9950	.9980	1.5770	1.9350	1.2233
	10	.9970	1.0000	1.3430	1.5630	1.1603
	20	1.0000	1.0000	1.1040	1.2210	1.1060
1.0	5	.9970	.9990	1.4410	1.7040	1.1801
	10	.9990	.9990	1.1830	1.3450	1.1369
	20	1.0000	1.0000	1.0470	1.1010	1.0516
		(.0099)	(.0050)	(.0235)	(.0289)	

〈표 1〉 등 간격 배열의 경우

(b) k = 8, P* = 0.90

$\sqrt{n}\Delta/\sigma$	n	PCS		E(S)		REFF
		Bayes	Gupta	Bayes	Gupta	
0.2	5	.8890	.9810	3.2200	5.5760	1.5693
	10	.9240	.9850	2.7130	4.5200	1.5629
	20	.9680	.9970	2.2840	3.5230	1.4976
0.4	5	.9730	.9950	2.2830	3.5900	1.5377
	10	.9860	.9980	1.9350	2.7540	1.4061
	20	.9970	.9980	1.5730	1.1160	1.3439
0.6	5	.9940	1.0000	1.8730	2.6230	1.3920
	10	.9960	1.0000	1.5690	2.0450	1.2982
	20	1.0000	1,0000	1.2640	1.5750	1.2640
0.8	5	.9950	.9990	1.5920	2.1590	1.3507
	10	.9980	.9990	1.3280	1.6630	1.2510
	20	1.0000	1.0000	1.1070	1.3090	1.1825
1.0	5	.9980	1.0000	1.4500	1.8490	1.2726
	10	1.0000	1.0000	1.1680	1.4430	1.2354
	20	1.0000	1.0000	1.0330	1.1380	1.1016
		(.0099)	(.0043)	(.0320)	(.0520)	

〈표 2〉 스리피지 배열의 경우

(a) $k = 4$, $P^* = 0.90$

$\sqrt{n}\Delta/\sigma$	n	PCS		E(S)		REFF
		Bayes	Gupta	Bayes	Gupta	
0.0	5	.7160	.8970	2.8060	3.5960	1.0230
	10	.6910	.9120	2.7980	3.6140	0.9786
	20	.6950	.8980	2.7670	3.5860	1.0030
0.2	5	.8200	.9480	2.7890	3.5930	1.1143
	10	.8550	.9570	2.7150	3.4990	1.1514
	20	.9060	.9710	2.6710	3.4520	1.2059
0.4	5	.9020	.9780	2.6440	3.4160	1.1916
	10	.9450	.9910	2.5280	3.2910	1.2414
	20	.9810	.9960	2.2610	2.9630	1.2907
0.6	5	.9510	.9930	2.5240	3.2950	1.2503
	10	.9750	.9940	2.1850	2.8830	1.2942
	20	.9990	.9990	1.7810	2.3210	1.3032
0.8	5	.9750	.9940	2.2870	3.0010	1.2871
	10	.9970	.9990	1.8260	2.3820	1.3019
	20	1.0000	1.0000	1.3340	1.6700	1.2519
1.0	5	.9920	.9980	2.0680	2.7500	1.3218
	10	1.0000	1.0000	1.5050	1.9430	1.2910
	20	1.0000	1.0000	1.0840	1.2390	1.1430
		(0.0146)	(0.0096)	(0.0267)	(0.0351)	

〈표 2〉 스리피지 배열의 경우

(b) k = 8, $P^* = 0.90$

$\sqrt{n}\Delta/\sigma$	n	PCS		E(S)		REFF
		Bayes	Gupta	Bayes	Gupta	
0.0	5	.5620	.9050	4.3280	7.2250	1.0367
	10	.5450	.9070	4.3370	7.2450	1.0038
	20	.5280	.9100	4.2670	7.2010	0.9792
0.2	5	.7210	.9470	4.2660	7.1740	1.2803
	10	.7560	.9690	4.2550	7.1250	1.3064
	20	.8110	.9790	4.1560	7.0370	1.4027
0.4	5	.8220	.9780	4.0880	7.0180	1.4429
	10	.8830	.9940	3.9080	6.7340	1.5307
	20	.9530	.9980	3.4910	6.1920	1.6991
0.6	5	.8910	.9900	3.9080	6.7600	1.5568
	10	.9620	.9990	3.4770	6.1580	1.7055
	20	.9940	1.0000	2.6100	4.7810	1.8208
0.8	5	.9560	.9950	3.5500	6.2330	1.6870
	10	.9910	1.0000	2.8560	5.1340	1.7814
	20	.9990	1.0000	1.8170	3.2210	1.7709
1.0	5	.9820	.9990	3.0970	5.5570	1.7638
	10	.9980	1.0000	2.2080	3.9840	1.8007
	20	1.0000	1.0000	1.2570	1.9060	1.5163
		(.0158)	(.0093)	(.0419)	(.0698)	

A Bayes-P* Selection Procedure for Normal Means with Common Unknown Variance+

Woo-Chul Kum*, Jong Woo Jeon*, and Kyung Soo Han**

*Department of Computer Science and Statistics, Seoul National University, Seoul, Korea

**Department of Statistics, Chunbuk National University, Chonjoo, Korea

ABSTRACT

For selecting a subset of k normal populations containing the one with the largest mean, a Bayes-P* selection procedure is considered when the common variance is unknown. Performance of the Bayes-P* selection procedure is compared with a well known classical procedure through a simulation study. Some frequentist's characteristics of Bayes-P* procedure are also studied.

+ Research supported by the Korean Science and Engineering Foundation Grant 88-07-23-02.