응용통계연구 제 19권 3호, 2006년, pp. 447–460

그래프 모형을 이용한 지수분포 모수들의 기하평균 비교에 관한 연구

김대황[1] 김혜중[2]

요 약

본 연구에서는 확률모형의 모수로부터 얻어지는 여러 형태의 함수간의 크기를 다중비교 하는 방법을 제안하고자 한다. 이 방법은 비교대상인 모수 함수 간의 선호확률을 베이지안 방법으로 추정하고, 이들로부터 얻어지는 선호행렬을 이용한 새로운 다중비교법이다. 이러한 방법의 제안에 필요한 이론과 비교기준을 고안하였으며, 응용 예로 제안된 방법을 s개의 독립인 지수분포 모수의 기하평균 크기 비교에 적용하였다.

주요용어: 선호행렬, 다중비교, 기하평균, 지수분포, 그래프 모형

1. 서론

다중비교(multiple comparison)란 가설검정에서 K개 모집단의 평균들이 동일하다는 가설 $H_0 : \mu_1 = \cdots = \mu_K$이 기각되었을 때, 평균들 간의 차를 심도있게 평가하는 통계적 방법이다. 이 문제의 해결을 위해 Fisher의 최소유의차(Least Significant Difference), Duncan의 다중범위검정(Multiple Range Test), Scheffe의 방법, Tukey의 정직유의차 검정(Honestly Significant Difference)과 같은 전통적인 다중비교 검정법(Hsu, 1996; Horchberg와 Tamhane, 1987), 베이지안 검정법(Pennello, 1997), 모수들의 동시신뢰구간 문제와 다중비교법(Bauer, 1997), 그리고 모수의 중요도 관점에서 최적의 모집단을 선택하는 방법(Bechhofer et al, 1995; Kim과 Nelson, 2001) 등 다양한 이론과 방법들이 제안되어 널리 사용되고 있다. 하지만, 이 방법들은 정규성이 가정된 확률모형 하에서 모평균들의 다중비교에 관한 연구들로서 비정규적 확률모형의 모수 또는 모수들의 여러 함수형태에 대한 다중비교가 필요한 경우에는 적용할 수 없는 문제점이 있다.

이 점에 착안하여, 본 논문에서는 정규성 가정이 완화된 일반적인 확률모형의 모수로부터 얻어지는 여러 형태의 함수들에 대한 상대적 중요도를 사용하여 다중비교하는 방법을 제안하고자 한다. 예를 들어, A, B, C 세 회사에서 생산되는 반도체 중 하나를 구매한다고 하자. 이때, 각 회사에서 생산되는 반도체의 고장율이 $\mu_{\mathbf{C}} < \mu_{\mathbf{A}} < \mu_{\mathbf{B}}$로 알려져 있다

1) (100-715) 서울시 중구 필동 동국대학교 통계학과, 강사
E-mail: daehha@dgu.edu
2) (교신저자)(100-715) 서울시 중구 필동 동국대학교 통계학과, 교수
E-mail: kim3hj@dgu.edu

면 구매자는 고장율이 낮은 회사의 순 $(C \to A \to B)$으로 구매할 것이다. 이처럼 관심모수들의 상대적 중요도에 따라 순서 (C, A, B)를 구하는 방법이 마련되면 이들에 대한 다중비교가 가능하다. 이와 관련하여, Gilbert(2003)는 정규분포 가정 하에서 이원배치 분산분석 모형에 정의된 처리효과들의 순서화에 필요한 분포도출에 대한 연구를 하였고, Davison와 Solomon(1973) 및 David(1987)는 관심모수보다는 여러 사상들의 상대적 중요도를 순서화하는 방법에 대한 연구를 하였다. 그러므로, 본 연구는 모형에 대한 정규성 가정에 구애 받지 않으며, 비교대상인 여러 모수들의 함수형태(또는 관심모수)들을 상대적 중요도에 따라 순서화시키는 방법을 제안하고자 하는 점에서 선행 연구들과 다르다. 이를 위하여, 먼저 유향그래프(Skiena 1990, p.175)에 관심모수들 간에 선호확률(preference probability)을 결합시켜 얻은 그래프 모형을 제안한다. 그리고 제안된 그래프 모형의 해밀톤 경로(Hamiltonian path)가 관심모수들의 중요도에 따라 얻어지는 순서와 동일함을 보이며, 주어진 그래프 모형으로 부터 해밀톤 경로를 간단히 찾을 수 있는 방법인 행합점수벡터법을 제안한다.

이러한 내용들에 대한 이론을 정립하기 위해, 2장에서는 그래프 모형의 기초적인 용어정리와 함께 몇 가지 정리들을 제시한다. 3장에서는 베이지안 방법에 의한 그래프 모형의 추정 및 추정된 행합점수벡터에 의한 관심모수들의 중요도순서를 추정하는 방법을 s개 지수모집단의 기하평균들의 다중비교를 통해 설명한다. 4장에서는 관심모수들의 다중비교에 제안된 그래프 모형의 유용성을 모의실험 및 실제자료 분석을 통해 보였다.

2. 그래프 모형

2.1. 그래프 이론

몇 개의 꼭지점과 그 점을 연결하는 선으로 이루어진 도형에 대한 이론을 그래프 이론이라고 한다. 특히 꼭지점들의 집합 X와 두 꼭지점을 연결하는 선들의 집합 R로 이루어진 순서쌍 (X, R)을 단순 그래프(simple graph)라고 하며, 그래프의 모든 꼭지점이 연결되어 있고 연결된 모든 선에 방향이 있는 그래프를 완전 유향그래프(complete directed graph)라 한다. 또한, 그래프 (X, R)에서 모든 꼭지점을 연결하는 선들의 궤적을 경로라 하며, 그래프상의 모든 꼭지점을 한 번만 지나는 경로를 해밀톤 경로라 부른다. 그림 2.1은 꼭지점의 수가 6인 완전 유향그래프를 나타낸 것이다.

이와 같이 정의된 완전 유향그래프를 다중비교에 적용하여 보자. K개 모집단에서 얻은 모수들의 함수인 $\delta_k (k = 1, \ldots, K)$를 다중비교 한다고 하자. 그러면 모집단을 나타내는 꼭지점의 집합 $X = \{1, \ldots, K\}$와 선들의 집합 $R = \{(i, j) | i \in X, j \in X, i \neq j\}$으로 이루어진 완전 유향그래프 (X, R)로 다중비교 상황을 나타낼 수 있다. 그리고, j번째 모집단의 모수함수 δ_j보다 i번째 모집단의 δ_i를 선호할 선호확률(preference probability)을 $\theta_{ij} = Pr(i \to j)$로 나타내자. 이때 선호확률 θ_{ij}는 $\theta_{ij} + \theta_{ji} = 1, i, j = 1, \ldots, K$ 이며 $\theta_{ii} = 1/2$로 정의된다. 이러한 성질을 만족하는 선호확률들로 이루어진 $K \times K$ 선호확률행렬(preference probability matrix) $\Theta = \{\theta_{ij}\}, i, j = 1, \ldots, K$을 가정하여 보자. 그러면 완전 유향그래프 (X, R)에 선호확률행렬 Θ를 결합시켜 새로운 그래프 (X, R, Θ)을 얻을 수 있고 이를 그래프 모형(graphical model)이라 정의하자.

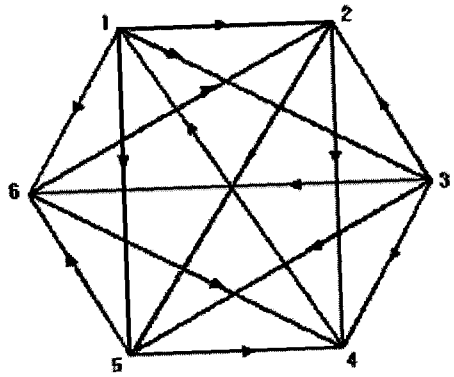

그림 2.1: 6개의 꼭지점을 가진 그래프 (X, R)

그래프 모형 $G(X, R, \Theta)$이 주어지면 이로부터 다중비교를 위한 선호순서(preference order) P를 얻을 수 있다. 만약 주어진 그래프 모형에서 모든 $i < j (i, j = 1, \ldots, K)$에 대해 i번째 모집단의 모수함수를 j번째 모집단의 모수함수보다 선호한다면 선호순서 $P = (p_1, \ldots, p_K) = \{1, \ldots, K\}$이 되며, P를 이용하여 다중비교를 행할 수 있다. 그러나 K가 큰 값을 가질 경우 주어진 그래프 모형으로부터 선호순서 P를 바로 구하는 데는 어려움이 따른다. P를 구하는 방법으로 Slater(1961)가 제안한 최근접순서(nearest adjoining order)법이 있다. $v(P)$를 그래프에서는 $p_i \to p_j$인데 도출된 선호순서 P에서는 $p_j \to p_i$로 나타난 경우의 횟수(즉, 도출된 선호순서가 그래프상의 선호관계를 위배한 횟수)라 할 때, 이 방법은 $v(P)$ 값을 최소로 하는 P를 찾는 방법이다. $v(P)$를 쉽게 계산하기 위해 그래프모형을 다음과 같이 정의되는 선호행렬(preference matrix)로 표현한다. 선호행렬은 선호확률행렬에서 (i, j)번째 원소가 $\theta_{p_i p_j} > 1/2$이면, '+', $\theta_{p_j p_i} > 1/2$이면 '−', 그리고 $\theta_{p_i p_j} = 1/2$이면 '.'의 부호를 사용하여 얻은 행렬이다.

예를 들어, 그림 2.1의 그래프 모형 $G(X, R, \Theta)$에서 모든 p_i와 p_j에 대해서 $\theta_{p_i p_j} = Pr(p_i \to p_j) = 3/4$이라 하자. 이 때 그림 2.1로부터 임의로 구한 선호순서가 $P = (1, 2, \ldots, 6)$이면 선호행렬은 다음과 같다.

$$\begin{pmatrix} \cdot & + & + & - & + & + \\ - & \cdot & - & + & + & - \\ - & + & \cdot & + & + & + \\ + & - & - & \cdot & - & - \\ - & - & - & + & \cdot & + \\ - & + & - & + & - & \cdot \end{pmatrix}.$$

이 행렬은 주대각원소를 기준으로 서로 반대의 부호를 가지므로 $v(P)$를 계산할 때 주대각원소를 기준으로 오른쪽 부분만을 고려하면 된다. 그러면 선호순서 $P = (1, 2, \ldots, 6)$가 그래프에 나타난 선호관계를 위배한 횟수는 5번이고, 이는 $-$ 부호의 수와 일치됨을 알 수 있다. 즉, $v(1, 2 \ldots, 6) = 5$이다. 이와 같은 방법을 이용하면 $P = (6, 4, 1, 3, 2, 5)$의 경우는

$v(P)=6$임을 알 수 있다.

위의 예에서 본 것과 같이, 선호확률행렬 Θ에 근거하여 비교대상인 K개 모수들의 모든 가능한 선호순서 P에 대해 $v(P)$를 계산한 후 최근접순서를 찾는 것은 복잡하고 번거러운 일이다. 이러한 문제점은 Θ에 대해 다음의 이행조건을 가정하면 쉽게 해결할 수 있다(David, 1963, p.13; Luce, 1961 참조).

약확률이행조건(C_1) : 모든 세 꼭지점 (i,j,l)에 대해

$$\theta_{ij}\geq 1/2\ ,\ \theta_{jl}\geq 1/2 \text{ 이면 } \theta_{il}\geq 1/2 \text{ 이다.} \tag{2.1}$$

강확률이행조건(C_2) : 모든 세 꼭지점 (i,j,l)에 대해

$$\theta_{ij}\geq 1/2\ ,\ \theta_{jl}\geq 1/2 \text{ 이면 } \theta_{il}\geq \max(\theta_{ij},\theta_{jl}) \text{ 이다.} \tag{2.2}$$

강확률이행조건 C_2는 다음과 같이 다시 표현할 수 있다.

모든 꼭지점 $(i,j), l=1,\ldots,K$에 대해

$$\theta_{ij}\geq 1/2 \text{ 이면 } \theta_{il}\geq \theta_{jl} \text{ 이다.} \tag{2.3}$$

그래프 모형이 확률이행조건 C_1 과 C_2을 만족하는지는 앞에서 정의한 선호행렬을 사용하여 검사할 수 있다.

다음은 그래프 모형을 이용한 다중비교에서 필요한 정리들을 소개한다. 앞에서도 언급하였듯이 해밀톤 경로는 모든 꼭지점을 한번만 지나는 경로를 의미하므로 다중비교를 위한 그래프 모형 $G(X,R,\Theta)$에 적용될 수 있다.

보조 정리. *그래프 모형 $G(X,R,\Theta)$이 확률이행조건 C_1(또는 C_2)을 만족하고 $P=(p_1,\ldots,p_K)$를 그래프 모형의 해밀톤 경로라고 가정하면 $v(P)=0$를 만족한다.*

증명. $P=(p_1,\ldots,p_K)$를 $G(X,R,\Theta)$의 해밀톤 경로라고 한다면 $\theta_{p_1p_2}>1/2$, $\theta_{p_2p_3}>1/2,\ldots,\theta_{p_{K-1}p_K}>1/2$이 성립한다. 약확률이행조건 C_1에 의해 $\theta_{p_1p_2}>1/2, \theta_{p_2,p_3}>1/2$이면 $\theta_{p_1p_3}>1/2$; $\theta_{p_1p_3}>1/2, \theta_{p_3p_4}>1/2$이면 $\theta_{p_1p_4}>1/2$; 그리고 $\theta_{p_1p_4}>1/2, \theta_{p_4p_5}>1/2$이면 $\theta_{p_1p_5}>1/2$이 성립함을 알 수 있다. 이와 같은 방법으로 모든 $i<j$에 대해 $\theta_{p_ip_j}>1/2$이 성립함을 보일 수 있으며, 이는 곧 $v(P)=0$임을 나타낸다. 이와 동일한 방법으로 강확률이행조건 C_2에 대해서도 보일 수 있다.

그래프 이론에 의하면 완전 유향그래프에는 적어도 하나의 해밀톤 경로가 존재한다. 그러므로, 보조정리는 확률이행조건 C_1(또는 C_2)을 만족하는 그래프 모형 $G(X,R,\Theta)$에는 $v(P)=0$를 만족하는 선호순서가 존재한다는 것을 의미한다.

정리 1. *$X=\{1,\ldots,K\}$인 그래프모형 $G(X,R,\Theta)$이 확률이행조건 C_1 (또는 C_2)를 만족한다면 $v(P)=0$를 만족하는 K개 꼭지점에 대한 해밀톤 경로는 오직 하나 존재한다.*

증명. 보조정리로부터 $X = \{1, \ldots, K\}$인 그래프모형 $G(X, R, \Theta)$의 해밀톤 경로는 다중비교에서 최적의 순서 $P = (p_1, \ldots, p_K)$가 됨을 알 수 있다. 여기서, 또 다른 해밀톤 경로 $P^* = (p_1^*, \ldots, p_K^*)$를 최적의 순서라고 가정하자. 이 때, 그래프 모형 $G(X, R, \Theta)$이 확률이행조건 C_1 (또는 C_2)을 만족하므로 $v(P) = v(P^*) = 0$이 성립한다. 따라서, 모든 $i < j$; $i, j = 1, \ldots, K$에 대하여, $\theta_{p_i p_j} > 1/2$와 $\theta_{p_i^* p_j^*} > 1/2$이 만족한다. 이러한 확률 관계는 모든 $i < j$에 대해서 $p_i = p_i^*$와 $p_j = p_j^*$일 때, 즉, $P = P^*$, 일 때만 성립한다. 그 이유는 $\theta_{p_i p_j} + \theta_{p_j p_i} = 1$, $\theta_{p_i^* p_j^*} + \theta_{p_j^* p_i^*} = 1$이기 때문이다.

정리 2. $X = \{1, \ldots, K\}$인 그래프 모형 $G(X, R, \Theta)$이 강확률이행조건 C_2를 만족한다면 다중비교에서 최적의 순서 P는 Θ의 행합점수(row-sum score)의 순서와 동일하다.

증명. 일반성을 잃지 않고 다중비교에서 최적의 순서를 $P = (1, 2, \ldots, K)$라고 가정하자. $G(X, R, \Theta)$이 C_2를 만족하면 다음의 관계가 성립한다. (2.3)으로부터, $l > k$ 이면 $\theta_{kl} \geq \theta_{k+1l}$, $l = k$이면 $\theta_{kl} = 1/2$이다. 그리고 (2.2)에 의해 $l < k$에 대해 $\theta_{lk+1} \geq \theta_{lk}$이 성립한다. 그리고, $\theta_{kl} = 1 - \theta_{lk}$, $\theta_{k+1l} = 1 - \theta_{lk+1}$이므로 $\theta_{kl} \geq \theta_{k+1l}$이 성립한다. Θ의 k번째와 $k+1$번째 행합점수를 분해하면 각각 다음과 같아진다.

$$\sum_{l=1}^{K} \theta_{kl} = \sum_{l<k} \theta_{kl} + \theta_{kk} + \sum_{l>k} \theta_{kl},$$

$$\sum_{l=1}^{K} \theta_{k+1l} = \sum_{l<k} \theta_{k+1l} + \theta_{k+1k} + \sum_{l>k} \theta_{k+1l}.$$

따라서, 주어진 $P = (1, 2, \ldots, K)$에 대해 $\theta_{kk} > \theta_{k+1k}$이 성립한다. 이로부터 $k = 1, \ldots, K-1$에 대해 $\sum_{l=1}^{K} \theta_{kl} > \sum_{l=1}^{K} \theta_{k+1l}$이 만족함을 알 수 있다.

2.2. 최적의 순서 기준

다중비교에서 최적의 순서를 결정하는 것은 실제로 쉬운 일이 아니다. K개 꼭지점을 가진 그래프 모형 $G(X, R, \Theta)$에서 최적의 순서를 구하는 방법은 가능한 모든 경로(또는 선호순서 P)에 대해 선호확률 Θ에 근거하여 $v(P)$ 값을 구하고, 최소의 $v(P)$ 값을 가지는 최적의 선호순서를 찾는 것이다. 또 다른 방법으로는 그래프 모형 $G(X, R, \Theta)$의 모든 해밀톤 경로에 대해 $v(P)$ 값을 계산하여 최소값을 갖는 경로를 찾는 것이다. 해밀톤 경로를 찾는 알고리즘은 (Adleman; 1994, Fu et al.; 2003)을 참고할 수 있다. 이와는 달리 정리 1과 정리 2는 새로운 방법으로 최적의 선호순서를 간단히 찾을 수 있는 이론적 바탕을 마련하고 있다.

정리 1은 그래프 모형 $G(X, R, \Theta)$의 Θ가 약확률이행조건 C_1을 만족하면, 모수에 대한 최적의 순서를 찾는 것은 모형의 유일한 해밀톤 경로를 구하는 것과 같음을 의미한다. 그러므로 그래프 모형이 약확률이행조건 C_1을 만족하고 꼭지점의 수가 7개 미만인 경우에는 Adleman(1994)의 알고리즘을 이용하여 유일한 해밀톤 경로인 최적의 순서를 찾을 수 있다. 그러나, Adleman(1994)의 방법은 확률이 결합되지 않은 그래프 (X, R)에서 모든 해밀톤 경

로 중 하나를 구하는 것이므로 이 방법을 이용한 해밀톤 경로가 $v(P)=0$를 만족하지 않을 수도 있다. 게다가, 비록 최적의 해밀톤 경로를 찾는다고 할지라도 꼭지점의 수가 7개 이상인 경우에는 알고리즘을 사용할 수 없는 문제가 발생된다. 이와는 달리 그래프 모형이 강확률이행조건 C_2를 만족한다면 정리 2를 이용하여 다중비교에서 최적의 순서를 쉽게 찾을 수 있다. 즉, 모집단 $\Pi_1,\ldots,\Pi_K$의 K개의 모수에 대한 다중비교에서 정의되는 그래프 모형인 $G(X,R,\Theta)$가 강확률조건 C_2를 만족하면 아래와 같이 정의되는 행합점수벡터 w의 크기 순이 K개 모수들의 다중비교에 필요한 최적인 순서가 된다.

$$w=\Theta\mathbf{1}. \tag{2.4}$$

여기서, $\mathbf{1}$은 모든 원소가 1인 $K\times 1$ 벡터이다.

3. 지수분포 모수들의 기하평균에 대한 다중비교

본 논문에서 제안한 그래프모형을 K개 모집단의 모수함수에 대한 다중비교에 적용하고자 한다. 여기서, 각 모집단은 s개의 독립인 지수분포로 이루어져 있으며 관심모수는 s개 모수의 기하평균 $\delta_k=(\prod_{\ell=1}^{s}(\lambda_\ell(k))^{1/s}$, $k=1,\ldots,K$,이다.

3.1. 사후선호확률

δ_k들의 다중비교를 위한 그래프 모형을 설정하기 위해 선호확률행렬 $\Theta=\{\theta_{ij}\}, i,j=1,\ldots,K$의 계산이 필요하다. 여기서, θ_{ij}는 사후분포로부터 계산되어지는 사후선호확률로 추정한다.

$X_{\ell 1}(k),\ldots,X_{\ell n_k}(k), \ell=1,\cdots,s; k=1,\cdots,K$를 평균이 $\lambda_\ell(k)$인 지수분포를 따르는 k번째 모집단에서 추출한 표본이라고 하자. 그리고, $\lambda_\ell(k)$들의 기하평균에 대한 확률대응사전분포는 Tibshirani(1989) 방법을 사용하여 다음과 같이 구했다.

$$\pi(\lambda(k))\propto\left(\prod_{\ell=1}^{s}\lambda_\ell(k)\right)^{-1}. \tag{3.1}$$

이를 사용하여 얻은 $\lambda=(\lambda(1),\ldots,\lambda(K))$의 결합사후확률분포는

$$p(\lambda(1),\ldots,\lambda(K)|data)\propto\prod_{k=1}^{K}\frac{1}{(\prod_{\ell=1}^{s}\lambda_\ell(k))^{n_k+1}}e^{-\frac{\sum_{\ell=1}^{s}\sum_{u=1}^{n_k}x_{\ell u}(k)}{\lambda_\ell(k)}} \tag{3.2}$$

이다. 여기서, $\lambda(k)=(\lambda_1(k),\ldots,\lambda_s(k))'$이다. 한편, δ_k들의 다중비교를 위한 그래프 모형 $G(X,R,\Theta)$을 설정하기 위해 δ_i가 δ_j보다 큰 것을 나타내는 선호확률 θ_{ij}를 δ_k들의 평균에 대한 상대적 위치확률을 사용하여 다음과 같이 정의하자.

$$\theta_{ij}=\frac{\pi_i}{\pi_i+\pi_j}. \tag{3.3}$$

여기서, $\delta_m = \sum_{k=1}^{K} \delta_k/K, \pi_k = E_p[I(\delta_k - \delta_m < 0)|data]$, E_p는 결합사후분포 (3.2)에 대해 기대값을 취한 것이고, $I(\cdot)$는 $(\cdot)$가 참이면 1, 아니면 0을 갖는 지시함수(indicator function)이다. 특히 π_k를 사용하여 선호확률 θ_{ij}를 정의한 이유는 θ_{ij}가 δ_i와 δ_j의 대소관계를 반영하는 확률이며, 이것으로 정의된 그래프 모형 $G(X, R, \Theta)$이 다음의 정리 3을 만족하기 때문이다. 이러한 목적에 부합하는 선호확률 식 (3.3)을 다양한 형태로 고안해 낼 수 있다. 예를 들면, δ_m 대신 δ_k들의 중위수를 사용하여 얻은 π_k를 식 (3.3)에 대입시켜 구한 선호확률도 사용가능하다.

정리 3. *식 (3.3)에 정의된 선호확률 θ_{ij}들을 원소로 하는 선호확률행렬 Θ는 강확률조건 C_2를 만족한다.*

증명. *모든 쌍 (i, j)에 대해서, $\theta_{ij} \geq 1/2$이면, 임의의 l에 대하여 $\theta_{il} \geq \theta_{jl}$을 만족함을 보이면 된다. $\theta_{ij} = \pi_i/(\pi_i + \pi_j) \geq 1/2$라고 가정하면 양의 수 ϵ에 대해 $\pi_i = \pi_j + \epsilon$이므로, $\theta_{i\ell}$과 $\theta_{j\ell}$은 각각 아래와 같이 정리할 수 있다.*

$$\theta_{i\ell} = \frac{\pi_i}{\pi_i + \pi_\ell} = \frac{\pi_j + \epsilon}{\pi_j + \epsilon + \pi_\ell} = 1 - \frac{\pi_\ell}{\pi_j + \epsilon + \pi_\ell}$$

$$\theta_{j\ell} = \frac{\pi_j}{\pi_j + \pi_\ell} = 1 - \frac{\pi_\ell}{\pi_j + \pi_\ell}.$$

따라서, $\theta_{i\ell} \geq \theta_{j\ell}$이 만족한다.

정리 3에 의해 선호확률 θ_{ij}가 강확률이행조건 C_2를 만족하므로 정리 2에 적용하면 다중비교를 위한 최적의 순서는 Θ의 행합점수의 순서와 동일하게 된다.

3.2. 선호확률 추정

사후결합확률분포인 식 (3.2) 하에서 π_k는 다음과 같이 표현된다.

$$\pi_k = E_p[I\{\delta_k - \delta_m < 0\}|data] = Pr(\delta_k - \delta_m < 0|data).$$

식 (3.2)에서 유도된 $\lambda_\ell(k)$의 주변사후확률분포는 다음과 같다.

$$p(\lambda_\ell(k)|\ data) \propto \frac{1}{\lambda_\ell(k)^{n_k+1}} e^{-\frac{\sum_{u=1}^{n_k} X_{\ell u}(k)}{\lambda_\ell(k)}}, \quad \ell = 1, 2, \ldots, s; k = 1, \ldots, K. \tag{3.4}$$

즉, $\lambda_\ell(k)^{-1} \mid data \sim Gamma(n_k,\ 1/\sum_{u=1}^{n_k} X_{\ell u}(k))$.

$\{\lambda_l^{(t)}(k), t = 1, \ldots, M; k = 1, \ldots, K; l = 1, \ldots, s\}$를 식 (3.4)에서 M번 반복 추출한 확률표본이라 하고 $\delta_k^{(t)} = \left(\prod_{\ell=1}^{s} \lambda_\ell^{(t)}(k)\right)^{1/s}$, $k = 1, \ldots, K$이면, π_k는 다음과 같이 추정된다.

$$\hat{\pi}_k = \hat{pr}(\delta_k - \delta_m < 0|data) = \frac{\sum_{t=1}^{M} I\{\delta_k^{(t)} - \delta_m^{(t)} < 0\}}{M}.$$

이와 같이 추정한 $\hat{\pi}_k$를 식 (3.3)에 대입하여 선호확률 θ_{ij} 및 선호확률행렬 Θ를 추정한 후, 식 (2.4)에 적용시켜 추정한 행합점수벡터($\hat{w}$)의 크기순에 의해 δ_k들의 다중비교 결과를 도출한다.

표 4.1: 0.05(0.95) 포함확률

k	$n_k = 20$	$n_k = 50$	$n_k = 100$
1	.040,(.953)	.042,(.943)	.047,(.959)
2	.059,(.952)	.051,(.950)	.053,(.943)
3	.054,(.952)	.060,(.945)	.044,(.955)
4	.046,(.956)	.055,(.956)	.050,(.962)
5	.052,(.940)	.047,(.957)	.058,(.952)
6	.045,(.945)	.057,(.948)	.052,(.951)

표 4.2: 추정된 행합점수벡터 $\hat{w}$

$n_k = 20$	$n_k = 50$	$n_k = 100$
3.923	4.215	4.301
3.692	4.046	4.226
3.459	3.650	3.914
3.068	2.734	3.058
2.252	2.195	1.653
1.606	1.16	0.848

4. 모의실험 및 예제

그래프이론과 선호확률행렬을 결합한 그래프 모형의 성질을 사용한 다중비교방법의 유용성을 모의실험과 예제를 사용하여 평가해 보았다.

4.1. 모의실험

k번째 모집단의 ℓ번째 변수 $X_\ell(k), \ell = 1, \ldots, s; k = 1, \ldots, K$가 평균이 $\lambda_\ell(k)$인 지수분포를 따르고 n_{ℓ_k}개의 표본이 관측되었다고 하자. 이 때, 다중비교의 대상은 각 모집단 모수의 기하평균, $\delta_k = \left(\prod_{\ell=1}^{s} \lambda_\ell(k)\right)^{1/s}$이다. 모의실험을 위해 $\lambda_\ell(k) = \ell + (k-1) \times \Delta, \ell = 1, \ldots, s; k = 1, \ldots, 6$로 가정하였다. 그리고, $n_{1k} = \cdots = n_{sk} = n_k = 20, 50, 100$인 경우에 대해서 모의실험을 실시하였다. $\Delta = 0.2$일 때 K개의 기하평균 크기의 오름차순에 대한 실제 순서는 $P = (1, \ldots, 6)$이 된다.

먼저 확률대응 사전분포가 적절한지를 전통적인 신뢰구간과 대응하는 개념인 포함확률을 통해 알 수 있다. 표 4.1은 Sun과 Ye(1995)가 제안한 알고리즘을 이용하여 계산한 것으로 n_k의 값에 상관없이 포함확률이 전통적인 신뢰구간과 아주 잘 대응 한다는 것을 알 수 있다.

3장에서 설명된 방법으로 선호확률행렬 Θ를 추정한 후 정리 2와 정리 3을 이용하여 다

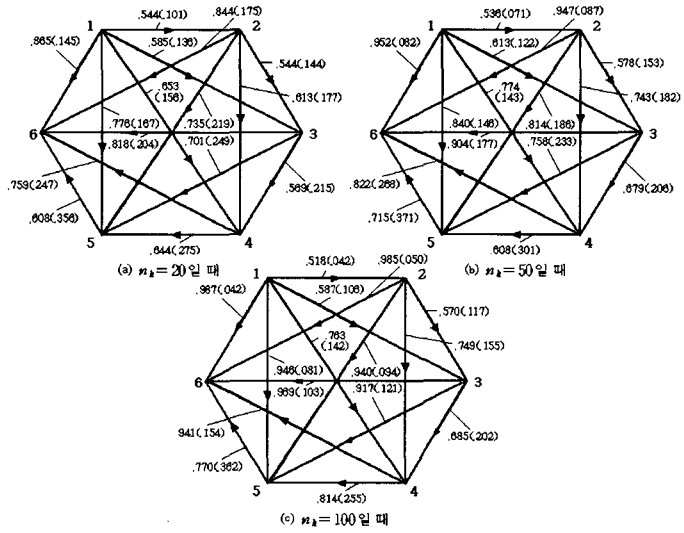

그림 4.1: $K = 6, s = 5, \Delta = 0.2$일 때 추정된 그래프 모형 $G(X, R, \hat{\Theta})$

표 4.3: 박테리아 자료(X_1 : Enterococcus Group Bacteria, X_2: E. Coli, X_3: Total Fecal Coliform) 출처 : http://wqm.igsb.uiowa.edu/ iastoret

1	X_1	10 0 0 0 50 300 36 100 10 290 30 45
	X_2	20 0 0 0 200 400 10 160 0 440 45 10
	X_3	20 0 0 0 60 300 36 200 20 450 30 45
2	X_1	130 0 40 0 55 640 110 100 100 230 50 10
	X_2	18 0 40 18 70 2200 40 150 20 400 18 27
	X_3	150 0 40 0 73 730 130 180 100 280 80 20
3	X_1	440 40 80 0 130 280 210 230 120 13000 27 10
	X_2	370 20 64 20 45 390 20 36 0 45000 27 36
	X_3	500 50 140 0 2200 480 240 250 120 25000 27 30
4	X_1	13000 2100 700 4400 210 300 2100 210 63 160 30 230
	X_2	57000 400 1500 57000 260 400 58000 100 73 120 45 200
	X_3	16000 3800 1100 7900 280 300 6700 380 140 230 30 330
5	X_1	70 40 20 2500 10 640 55 0 27 36 50 0
	X_2	40 30 0 14000 30 2200 120 0 70 0 18 0
	X_3	100 60 20 4500 18 730 73 0 36 36 80 0
6	X_1	10 0 10 4300 1300 280 190 36 230 20 27 0
	X_2	0 0 10 30000 60 390 60 0 340 0 27 0
	X_3	30 0 10 6500 1500 480 350 55 360 20 27 0

중비교를 위한 최적순서를 구하였다. 부록의 표 6.1은 $n_k = 20, 50, 100$일 때의 선호확률 $\theta_{ij} = Pr(\delta_i \to \delta_j)$를 $M = 10000$로 하여 1000번 반복추정한 것이다. 그림 4.1의 (a), (b), (c)는 각각 $n_k = 20, 50, 100$인 경우에 추정된 그래프 모형 $G(X, R, \hat{\Theta})$으로 표 6.1의 선호확률값을 사용하여 추정한 것이다. 마지막으로 표 4.2는 표 6.1을 사용하여 $n_k = 20, 50, 100$일 때의 추정된 행합점수벡터 $\hat{w}$를 나타낸 것이다. 그리고 $\hat{w}$의 행합점수를 행 번호에 의해 크기 순으로 나열하면 모든 n_k값에 대해 $(1, 2, 3, 4, 5, 6)$이 된다. 따라서, 표본의 크기에 관계없이 제안된 다중비교에서 구한 최적의 순서는 $P = (1, 2, 3, 4, 5, 6)$이며 δ_k들의 실제값의 크기 순서와 일치함을 보인다.

4.2. 박테리아 예제

미국의 EPA에서는 수중이나 공기중의 오염도를 측정하는데 일별로 관측된 여러 박테리아 수의 기하평균을 사용할 것을 제안하였다(EPA; 1986참조). 즉, 여러 박테리아 수에 대한 기하평균으로 수질을 파악하여 식용수 또는 수영장에서 사용이 가능한지를 판단하고 있다. 표 4.3은 Iowa주의 Cedar River의 6곳(1: CedarRiver Up stream of Waterloo/CedarFalls, 2: Cedar River Down stream of Waterloo, 3: Cedar River at Cedar Bluff, 4: Cedar River Down stream of Cedar Rapids, 5: Cedar River Up stream of Cedar Rapids, 6 :Cedar River near Conesville)에서 일별로 관측된 박테리아 자료이다. 이 자료를 사용하여 여섯 장소 중 어느 곳의 물을 식용으로 더 적당한 지를 상대비교 한다고 하자. 이 자료는 일 별로 관측된 박테리아 수에 관한 자료로 포아송분포를 따르는 자료로 볼 수 있다. 따라서 관측된 자료의 역수를 취하여 이들이 지수분포를 따른다고 가정하다. 그러면 $K = 6, s = 3$인 그래프 모형 하에서 $\delta_k, k = 1, \ldots, 6$들에 대해 다중비교를 시행할 수 있고, 다중비교의 결과에 의해 박테리아의 함량이 작은 순으로 여섯 군데 강의 식용성을 상대비교 할 수 있다.

표 4.4: $\hat{\Theta}$와 행합점수벡터 $\hat{w}$

$\hat{\Theta} = \{\hat{\theta}_{ij}\}$						$\hat{w} = \hat{\Theta}\mathbf{1}$
0.500	0.180	0.174	0.174	0.227	0.614	1.868
0.820	0.500	0.490	0.488	0.571	0.878	3.747
0.826	0.510	0.500	0.499	0.581	0.883	3.799
0.826	0.512	0.501	0.500	0.583	0.883	3.805
0.773	0.429	0.419	0.417	0.500	0.844	3.383
0.386	0.122	0.117	0.117	0.156	0.500	1.398

3장의 방법에 의해 추정된 선호확률 θ_{ij}의 값과 표준편차를 표 6.2에 정리하였다. 그림 4.2는 표 6.2의 추정결과를 그래프 모형으로 나타낸 것이다. 또한, 추정된 선호확률행렬 $\hat{\Theta}$과 행합점수벡터 $\hat{w}$를 표 4.4에 정리하였다. 표 4.4에서 행합점수벡터 $\hat{w}$는 (1.868, 3.747, 3.799, 3.805, 3.383, 1.398)으로 이를 내림차순으로 나열하였을 때 $4 \to 3 \to 2 \to 5 \to 1 \to 6$로서 다중비교를 위한 최적의 순서는 $P = (4, 3, 2, 5, 1, 6)$가 된다. 하지만, 이 결과는 원래 자료에 역

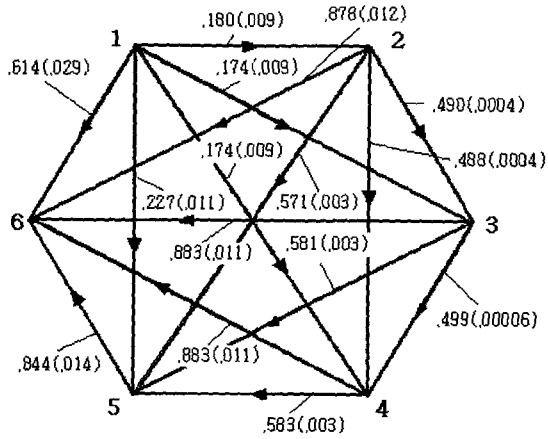

그림 4.2: 추정된 그래프 모형 $G(X, R, \hat{\Theta})$

수를 취한 결과이므로 추정된 순서를 반대로 나열하면 $P = (6, 1, 5, 2, 3, 4)$이 된다. 다중비교를 적용하면 박테리아 수가 가장 작은 곳은 Cedar River near Conesville이 되며, 가장 많은 곳은 Cedar River Down stream of Cedar Rapids로 추정할 수 있다. 즉, 식용수로 여섯 곳 중에서 식용수로 적당한 곳을 차례대로 선택한다면 Cedar River near Conesville, CedarRiver Up stream of Waterloo/CedarFalls,Cedar River Up stream of Cedar Rapids, Cedar River Down stream of Waterloo, Cedar River at Cedar Bluff, Cedar River Down stream of Cedar Rapids이 된다.

5. 결론

본 논문은 관심모수들에 대한 다중비교에 그래프 모형을 도입하여 수행하는 방법을 제안하였다. 이 방법은 정규성 가정에 위배되는 확률모형 하에서의 다중비교 뿐아니라, 정규성 가정 하에서 모수들의 복잡한 함수를 다중비교하는 데에도 적용할 수 있다는 점에서 기존의 여러 다중비교법들과 차별된다. 제안된 그래프 모형은 선호확률행렬에 수학에서의 그래프 이론을 결합시켜 얻은 모형으로 이에 대한 다양한 성질도 함께 연구되었다. 특히, 선호확률들로 이루어진 선호확률행렬이 강확률이행조건을 만족한다면 다중비교를 위한 최적의 순서는 선호확률행렬의 행합점수의 순서와 일치한다는 성질을 보임으로써, 다중비교에서 그래프 모형의 유용성을 이론적으로 보였다.

본 논문은 지수분포 모수들의 기하평균간에 다중비교에 촛점을 두고 선호확률의 베이지안 추정방법을 제시하고, 제안된 다중비교 방법의 효율성을 모의실험으로 평가하였다. 그러나, 이 방법이 다중비교 문제에서 일반적인 해가 됨을 보이기 위해서는 좀 더 폭넓은 적용 예의 개발이 필요하다. 특히, 다변량 확률모형의 공분산 비교, trace 비교, 고유근 비교 등 다차원 모수들에 대한 다중비교에 대한 방법이 없는 현 상황에서 제안된 방법은 이

러한 문제들을 해결할 수 있는 방법이 될 것으로 기대하며, 이에 대한 연구들은 본 연구 결과의 유용성을 돋보이게 할 것이다.

6. 부록

표 6.1: 선호확률의 추정값 $\hat{\theta}_{ij}$

	$n_k = 20$		$n_k = 50$		$n_k = 100$	
θ_{12}	.544	(.101)	.536	(.071)	.518	(.042)
θ_{13}	.585	(.136)	.613	(.122)	.587	(.106)
θ_{14}	.653	(.156)	.774	(.143)	.763	(.142)
θ_{15}	.776	(.167)	.840	(.146)	.946	(.081)
θ_{16}	.865	(.145)	.952	(.082)	.987	(.042)
θ_{23}	.544	(.144)	.578	(.153)	.570	(.117)
θ_{24}	.613	(.177)	.743	(.182)	.749	(.155)
θ_{25}	.735	(.219)	.814	(.186)	.940	(.094)
θ_{26}	.844	(.175)	.947	(.087)	.985	(.050)
θ_{34}	.569	(.215)	.679	(.206)	.685	(.202)
θ_{35}	.701	(.249)	.758	(.233)	.917	(.121)
θ_{36}	.818	(.204)	.904	(.177)	.969	(.103)
θ_{45}	.644	(.275)	.608	(.301)	.814	(.255)
θ_{46}	.759	(.247)	.822	(.268)	.941	(.154)
θ_{56}	.608	(.356)	.715	(.371)	.770	(.362)

표 6.2: 사후선호확률의 추정값 $\hat{\theta}_{ij}$

θ_{12}	.180(.009)	θ_{23}	.490(.0004)	θ_{35}	.581(.003)
θ_{13}	.174(.009)	θ_{24}	.488(.0004)	θ_{36}	.883(.011)
θ_{14}	.174(.009)	θ_{25}	.571(.0030)	θ_{45}	.583(.003)
θ_{15}	.227(.011)	θ_{26}	.878(.0120)	θ_{46}	.883(.011)
θ_{16}	.614(.029)	θ_{34}	.499(.0001)	θ_{56}	.844(.014)

* 괄호안의 값은 표준편차임

참고문헌

Adelman, L. M. (1994). Molecular computation of solutions. *Science*, **266**, 1021-1024.

Bauer, P. (1997). A note on multiple testing procedure in dose finding. *Biometric*, **53**, 1125-1128.

Bechhofer, R. E., Santner, T. J., and Goldsman, D. M. (1995). *Design and Analysis of Experiments for Statistical Selection, Screening, and Multiple Comparisons*, New York: Wiley.

David, H. A. (1987). Ranking from unbalanced paired-comparison data. *Biometrika*, **74**, 432-436.

Davison, R. R. and Solomon, D. L. (1973) A Bayesian approach to paired comparison experimentation. *Biometrika*, **60**, 477-487.

EPA(1986), Bacteriological Ambient Water Quality Criteria for Marine and Fresh Recreational Waters,*Ambient Water Quality Criteria for Bacteria.*

Fu, B., Beigel, R., and Zhou, F. (2003). An $O(2^n)$ volume molecular algorithm for Hamiltonian path. www.cis.temple.edu/$\sim$ beigel/papers.

Gilbert, S. (2003). Distribution of rankings for groups exhibiting heteroscedasticity and correlation. *Journal of the American Statistical Association*, **98**, 147-157.

Horchberg, Y. and Tamhane, A. C. (1987) *Multiple Comparion Procdures.* New York, Wiley.

Hsu, J. C. (1996). *Multiple Comparisons*, London: Chapman and Hall.

Iowa's STORET Water Quality Database, http://wqm.igsb.uiowa.edu/ iastoret/

Kim, S. and Nelson, B. L. (2001). A fully sequential selection procedure for indifference-zone selection in simulation. *Transactions on Modeling and Computer Simulation*, **11**, 251-273.

Pennello, G. (1997). The k-ratio multiple comparisons Bayes rule for the balanced two-way design. *Journal of the American Statistical Association*, **92**, 675-684.

Skiena, S. (1990). *Implementing Discrete Mathematics: Combinatorics and Graph Theory with Mathematica*, Reading, MA: Addison-Wesley, 1990.

Slatar, P.(1961). Inconsistencies in a schedule of paired comparisons. *Biometrika*, **48**, 303-312.

Sun, D. and Ye, K. (1995), Reference Prior Bayesian Analysis for Normal Mean Products, *Journal of the American Statistical Association*, **90,** 589-597.

Tibshirani, R. (1989). Noninformative priors for one parameter of many, *Biometrika*, **76**, 604-608.

[2006년 4월 접수, 2006년 6월 채택]

On Multiple Comparison of Geometric Means of Exponential Parameters via Graphical Model

Dae-Hwang Kim[1] Hea-Jung Kim[2]

ABSTRACT

This paper develops a multiple comparison method for finding an optimal ordering of K geometric means of exponential parameters. This is based on the paired comparison experimental arrangement whose results can naturally be represented by a completely oriented graph. Introducing posterior preference probabilities and stochastic transitivity conditions to the graph, we obtain a new graphical model that yields criteria for the optimal ordering in the multiple comparison. Necessary theories involved in the method and some computational aspects are provided. Some numerical examples are given to illustrate the efficiency of the suggested method.

Keywords: preference matrix, multiple comparison, geometric mean, exponential distribution, graphical model

1) Lecturer, Department of Statistics, Dongguk University, Seoul 100-715, Korea
E-mail: daehha@dgu.edu

2) (Corresponding author) Professor, Department of Statistics, Dongguk University, Seoul 100-715, Korea
E-mail: kim3hj@dgu.edu