Communications of the Korean Statistical Society Vol. 15, No. 4, 2008, pp. 605–614

주변동질성검정법을 이용한 종속된 두 일치도의 비교

오명식[1)]

요 약

종속된 두 개의 일치도를 비교하는 간단한 검정법을 제시하였다. Oh (2008)에 의해 연구된 우도비 검정은 순위 제약하의 검정기법을 사용함으로서 통계량의 계산이나 유의확률을 구하기가 까다롭다. 본 논문에서는 기존의 주변동질성(marginal Homogeneity)에 관한 검정법 즉 Bhapkar 혹은 Stuart-Maxwell 검정을 이용할 수 있는 검정법을 제시하였다. 제시된 검정법을 2008년 세계피겨스케이팅선수권대회의 여자싱글부분의 심판자료를 분석하였다.

주요용어: 일치도; Bhapkar검정; Stuart-Maxwell 검정; 일반화 McNemar검정; 주변동질성.

1. 서론

일치도(Agreement)는 연관성(Association)에 비해 비교적 적은 관심을 받은 분야이기는 하지만 의학, 스포츠 등 많은 분야에서 아주 활발하게 사용되고 있다. 어떠한 질병에 대한 의사들간의 정확한 진단 특히 병의 진행정도 혹은 유형 등에 대한 일치가 환자의 치료에 있어서는 매우 중요한 요소로 작용한다. 한편 심판에 의해 점수가 부여되는 스포츠 종목 예를 들면 체조, 피겨스케이팅, 다이빙 등에서는 가끔 심판들간의 점수차이가 심해 자칫 편파판정 등의 시비가 끊이지 않고 있다. 따라서 이러한 분야에 있어서의 일치도의 측정은 매우 중요한 통계적 방법으로 자리 잡고 있다.

잘 알려진 바와 같이 일치도는 Cohen (1966)의 카파(kappa)가 가장 오래된 일치도의 측정방법이다. Fleiss와 Cohen (1973)의 가중카파(weighted kappa)는 순서형범주가 사용되는 경우에 일치도로 사용된다. 이 두 가지의 카파통계량은 대표본하에서 정규분포를 따름이 알려져 있어 독립성이 유지되는 경우에는 두 개의 일치도의 비교는 전통적인 이표본 평균차 검정을 사용할 수 있다. 그러나 많은 경우 독립성이 유지되지 않기에 이러한 검정은 사용할 수가 없다. 예를 들자면 특정 질병을 가진 환자들을 두 그룹의 의사들에게 진찰하게 하고 질병의 진행정도를 판단하게 하자. 이렇게 얻어진 두 그룹간의 일치도는 확연하게 독립이 아니다. 따라서 이표본 평균차 검정으로는 이를 검정해 낼 수 없다.

이러한 어려움을 해결하고자 McKenzie *et al.* (1996)와 Donner *et al.* (2000) 등은 이항척도(binary scale)가 사용되는 경우 두 개의 종속된 일치도를 비교검정하는 방법을 고

1) (608-738) 부산시 남구 우암1동 55-1, 부산외국어대학교 데이터경영학과, 교수.
E-mail: moh@pufs.ac.kr

안하였다. 한편 Oh (2008)은 주변동질성(marginal homogeneity)과 주변분포의 확률적순위(marginal stochastic ordering)를 이용하여 일반적인 척도를 사용하여 얻어진 일치도를 비교하는 우도비검정을 고안하였다. 이 검정법의 고안으로 종속된 일치도의 차이에 대한 단측 검정을 시행할 수 있게 되었지만 검정을 실제로 사용하는 데 있어 몇가지 여러운 점이 있다. 첫째, 검정통계량을 구하기 위해서는 quadratic programming 등의 사용과 같은 복잡한 계산 절차를 거쳐여야 하며 둘째, 범주의 크기가 큰 경우 검정통계량의 분포를 구하기 어려워 근사적인 방법을 사용하여야 한다. 물론 좋은 검정 결과를 얻기 위해서라면 이러한 어려움을 감내해야 하지만 일부 사용자에게는 이 검정법을 기피하는 한 이유가 될 수 있을 것이다. 따라서 본 논문에서는 주변동질성에 대한 기존의 비교적 사용하기 손쉬운 검정방법을 이용하는 검정법에 대하여 논의하기로 한다. 이렇게 제시된 검정방법을 이용하여 2008년 세계피겨스케이팅선수권대회의 심판점수에 관한 자료를 분석한다.

2. 주변동질성과 일치도

두 평가자에 대한 일치도는 카파와 같은 통계량을 쓰기도 하지만 주변동질성을 이용하여 검증하기도 한다. 이에 관하여서는 Agresti (2002) 등의 참고문헌을 참조하기 바란다. 이러한 검증방법은 두 그룹간의 일치도의 비교에 원용하여 다음과 같이 사용될 수 있다. 하나의 평가대상(subject)에 대하여 두 그룹으로 나뉜 여러명의 평가자(raters)가 평가를 한다고 가정하고 그 결과를 $\mathbf{X}_i = (\mathbf{X}_{1i}, \mathbf{X}_{2i}) = (X_{11i}, \ldots, X_{1k_1 i}, X_{21i}, \ldots, X_{2k_2 i})$, $j = 1, 2, \ldots, k_i$, $i = 1, 2, \ldots, n$라 하자. 이때 $X_{\ell ji}$는 i번째 평가대상에 대한 ℓ번째 그룹의 j번째 평가자의 평가결과이다. 단 각 평가자는 동일한 점수체계를 사용한다고 가정한다. 여기서 다음과 같이 평가대상내 일치도(within-subject agreement)를 다음과 같은 함수를 도입하여 사용한다. 다음과 같은 음이 아닌 값을 갖는 두 개의 함수 D_1과 D_2를 생각해 보자.

(1) $D_i : \mathbf{R}^{k_i} \to \mathbf{R}^+ \cup \{0\}$, 단 $i = 1, 2$,

(2) $D_i(\mathbf{x}) = D_i(\mathbf{x}_\gamma)$, 단 $\mathbf{x}_\gamma$는 $\mathbf{x}$의 임의의 순열,

(3) $D_i((a, a, \ldots, a)) = 0$, 단 $a > 0$,

(4) $D_i(\cdot)$는 $\mathbf{R}^{k_i}$에서 정의되는 볼록함수(convex function)이고,

(5) $\{y_i \in \mathbf{R}^+ \cup \{0\} : \ y_i = D_1(\mathbf{X}_{1i})\} = \{y_i \in \mathbf{R}^+ \cup \{0\} : \ y_i = D_2(\mathbf{X}_{2i})\}$,

여기서 함수 D_1과 D_2는 산포(dispersion)를 측정하는 도구이기도 하다. 예를 들어 다음과 같은 함수를 사용하여 설명해 보자.

$$D_1(\mathbf{X}_{1i}) = \max_{j=1,\ldots,k_1} \{X_{1ji}\} - \min_{j=1,\ldots,k_1} \{X_{1ji}\},$$
$$D_2(\mathbf{X}_{2i}) = \max_{j=1,\ldots,k_2} \{X_{2ji}\} - \min_{j=1,\ldots,k_2} \{X_{2ji}\}. \tag{2.1}$$

즉 (2.1)의 D_1과 D_2는 각각 범위를 나타낸다.

D_1과 D_2을 이용하여 두 개의 일치도의 비교하는 과정을 간단하게 설명한다. 모든 t, 단 $t > 0$, 에 대하여 다음의 부등식이 성립한다고 하자.

$$\frac{\#\{\ell : D_1(\mathbf{X}_{1\ell}) \le t,\ \ell = 1, \ldots, n\}}{n} \ge \frac{\#\{\ell : D_2(\mathbf{X}_{2\ell}) \le t,\ \ell = 1, \ldots, n\}}{n}. \tag{2.2}$$

이는 D_1과 관련된 첫번째 그룹의 일치도가 D_2와 관련된 두번째 그룹의 일치도보다는 높다는 것을 의미한다. 한편 모든 t에 대하여 등호가 성립한다면 두 그룹의 일치도는 동등하다고 말할 수 있게 된다. 즉 D_1과 D_2의 값들로 분할표를 만들었을 때 주변동질성(marginal homogeneity)이 성립하면 두 그룹의 일치도가 동일하다고 말할 수 있다. 아울러 (2.2)가 성립 즉 주변확률들간에 확률적순위(stochastic ordering)가 성립한다면 두 개의 일치도간에 순위가 있다는 것을 의미한다. 여기서 주의해야 할 점은 일치도가 같다는 것은 통상 일치도를 측정하기 위해 사용하는 카파 혹은 가중카파 등이 같은 값을 갖는다는 것과는 관련이 없다.

이러한 사실에 입각하여 Oh (2008)는 두 일치도를 비교하는 우도비 검정법을 제시하였다. 앞서 지적한 바와 같이 주변확률간에 확률적순위가 있다는 대립가설에 대한 주변동질성에 대한 우도비 검정은 매우 복잡한 계산과정을 거쳐야 한다. 따라서 우리는 단측검정 대신 대립가설을 임의의 가설 즉 귀무가설이 성립하지 않는다는 가설을 설정하여 기존에 개발되어 있는 비교적 간단히 사용 가능한 통계패키지 예를 들어 SAS 등으로 처리할 수 있는 검정법에 대하여 알아보기로 한다.

3. 일치도 비교에 관한 검정법

2절에서 설명한 평가대상내 평가자간의 일치도는 음이 아닌 임의의 값을 갖도록 정의되어 있지만 여기서는 한 분할표의 주변동질성을 이용하는 방법을 설명하기 위하여 D_1과 D_2이 가질 수 있는 값의 집합을 $\{s_1, \ldots, s_k\}$이라 하자. 그리고 $i, j = 1, \ldots, k$에 대하여

$$p_{ij} = \frac{\#\{\ell : D_1(\mathbf{X}_{1\ell}) = s_i,\ D_2(\mathbf{X}_{2\ell}) = s_j,\ \ell = 1, \ldots, n\}}{n},$$

$\hat{p}_{ij}$는 p_{ij}의 관측치라고 하고 $n_{ij} = n\hat{p}_{ij}$으로 하자. 여기서 다음의 가설을 고려해 보자.

$$H_0 : p_{i+} = p_{+i}, \quad \text{단 } i = 1, \ldots, k$$
$$H_1 : p_{i+} \ne p_{+i}, \quad \text{적어도 하나 이상의 } i\text{에 대하여 성립}$$

참고로 대립가설을 아래와 같이 설정하는 경우 우도비검정법은 Oh (2008)의 연구결과를 참조하기 바란다.

$$\sum_{i=1}^{j} p_{i+} \ge \sum_{i=1}^{j} p_{+i}, \quad \text{단 } j = 1, \ldots, k-1, \text{ 적어도 하나 이상의 } j\text{에 대하여 부등호가 성립.}$$

여기서 현재까지 연구된 주변동질성에 대한 몇가지 검정을 간단히 살펴보기로 한다. 주변동질성에 대한 검정으로는 본래 이항척도를 사용하는 경우 즉 2×2 분할표에 대한

McNemar검정이 가장 널리 쓰인다. 그러나 이는 이항척도로 제한되어 있기에 이를 $I \times I$ 분할표($I > 2$)에 사용할 수 있는 일반화 McNemar검정이 개발되어 있다. 이는 Stuart-Maxwell 검정과 동일하다. Stuart-Maxwell 검정을 간단하게 살펴 보기로 한다. 귀무가설 H_0는 행렬과 벡터를 사용하여 다음과 같이 표현될 수 있다. 먼저 c_{ij}를 다음과 같이 정의하자.

$$c_{ij} = I_{\{k\times(j-1)+1,\ldots,k\times j\}}(i) - I_{\{mod(i-1,k)+1\}}(j),$$

여기서 $I_A(\cdot)$는 인덱스함수이며, $\mathrm{mod}(i,\, k)$는 i를 k로 나눈 나머지를 나타낸다. 이를 이용하여 다음의 행렬을 정의하자.

$$C = \{c_{ij}\}_{k^2\times k}, \quad i = 1,\ldots,k^2,\ j = 1,\ldots,k.$$

분할표를 $\mathbf{p} = (p_{11},\ldots,p_{1k},\ldots,p_{kk})'$로 표시하면 주변동질성을 나타내는 귀무가설은

$$C'\mathbf{p} = \mathbf{0}$$

와 같이 표현될 수 있다. 그러나 이는 중복된 제약을 포함하고 있으므로 다음과 같이 중복된 제약을 배제하여야 한다. 여기서 $i \le j$이면 1이고 그 이외엔 0이 되게 b_{ij}를 정의하고 이를 이용하여 다음의 행렬을 생각하자.

$$B = \{b_{ij}\}_{k\times(k-1)}, \quad i = 1,\ldots,k,\ j = 1,\ldots,k-1,$$

그러면 주변동질성의 귀무가설은 다음과 같다.

$$B'C'\mathbf{p} = \mathbf{0}. \tag{3.1}$$

Stuart (1955)는 다음의 통계량이 근사적으로 자유도가 $k-1$인 카이제곱분포를 따름을 보였다.

$$Z_0 = n\hat{\mathbf{p}}'CB(B'C'W^{-1}CB)^{-1}B'C'\hat{\mathbf{p}},$$

단 $W^{-1}_{k^2\times k^2} = \mathrm{diag}\{1/\hat{p}_{ij}\} - \hat{\mathbf{p}}\hat{\mathbf{p}}'$.

Bhapkar (1966)는 W^{-1}를 조금 다르게 정의하여 검정통계량을 제시하였다. Z_0와 Z_1를 각각 Stuart-Maxwell 검정통계량, Bhapkar 검정통계량이라 하면 $Z_1 = Z_0/(1 - Z_0/n)$의 관계가 성립한다. 물론 두 검정통계량은 표본크기가 커지면 그 값이 같아지고 아울러 근사적으로 자유도가 $k-1$인 카이제곱분포를 따른다. 일반적으로 Bhapkar 검정이 좀더 검정력이 크다고 알려져 있다.

한편 두 가지의 검정통계량의 계산은 비교적 작은 k값에 대하여도 복잡한 편이다. 다행스럽게 이 두 가지 검정통계량은 SAS의 PROC CATMOD를 이용하여 값을 얻을 수 있다. PROC CATMOD의 사용이 일반적으로 쉽지는 않지만 몇가지의 방법이 잘 알려져 있어 비교적 손쉽게 계산할 수 있다. 이를 위해 Agresti (2002), Schuster와 von Eye (1998)를 참조하기 바란다. 또한 Uebersax (2006)가 작성하여 website에 공개한 간단한 프로그램 mh.exe를 이용할 수도 있다. 최근 일반화 McNemar검정은 Sun과 Yang (2008)에 의해

SAS macro를 이용하여 작성되어 공개되었다. 이는 SAS Grobal Forum 2008을 위한 website에서 구할 수 있다. 다만 결측치를 갖고 표본크기가 작은 경우 프로그램이 실행되지 못한다. Lang (2007)은 다항 및 포아송 동질성 모형(Multinomial-Poisson Homogeneous Model)을 위한 R code mph.fit를 작성하였다. 이를 이용하여 계산이 가능하다. 참고로 여기에서 언급한 검정법들의 사용예는 해당 논문에 자세하게 언급되어 있어 그 사용례의 언급은 생략한다.

반면 주변동질성은 로그선형모형으로는 표현될 수 없지만 유사대칭성(quasi symmetry)모형을 사용할 수 있다. Caussinus (1966)는 대칭성을 만족하면 동질성과 유사대칭성이 동시에 만족됨을 보였다. 이의 역도 성립하는데 이에 대한 증명은 Agresti (2002)를 참조하기 바란다. 따라서 이를 이용하면 주변동질성을 검정할 수 있다. 즉 유사대칭성의 조건하에서 대칭성의 검정은 주변동질성의 검정과 동일하다.

참고로 2절에서 잠시 언급한 대립가설을 주변확률들간의 확률적순위를 가정하는 경우의 검정통계량은 다음과 같다.

$$n \min_{\alpha \le \mathbf{0}} \left\{ [(B'C'W^{-1}CB)^{-1}B'C'\hat{\mathbf{p}} - \alpha]'(B'C'W^{-1}CB)[(B'C'W^{-1}CB)^{-1}B'C'\hat{\mathbf{p}} - \alpha] \right\},$$

단 $\alpha' = (\alpha_1, \ldots, \alpha_k)$이며 벡터간의 연산과 비교는 성분끼리 하는 것으로 가정한다. 위의 통계량은 quadratic prgramming을 이용하여 계산할 수 있다. 여기서는 자세한 언급은 생략하고 관심있는 자는 Oh (2008)를 참조하기 바란다.

다음절에서는 앞서 언급한 Bhapkar검정을 이용하여 2008년 세계피겨스케이팅선수권대회의 여자부 피겨스케이팅 점수 자료를 분석한다.

4. 2008년 세계피겨스케이팅선수권대회 자료분석

피겨스케이팅은 네가지 종목 즉 남녀싱글, 페어 그리고 아이스댄싱 등 네 종목으로 진행된다. 세계선수권대회에는 포함되지 않은 싱크로나이즈드 스케이팅은 여기서는 제외한다. 이중 아이스댄싱을 제외하고 세 종목은 쇼트프로그램과 프리스케이팅의 두 경기의 점수의 합산으로 순위가 결정되다. 참고로 아이스댄싱은 종목 프리스케이팅, 오리지날댄스 그리고 규정댄스(Compulsory dance)의 세 경기로 구성되어 있다. 여기서 우리는 여자싱글 부분의 점수에 대하여 분석하기로 한다.

먼저 새로운 심판제도에 대하여 간략하게 알아보기로 한다. 구 심판제도는 6.0점 척도를 사용하는 데 쇼트프로그램과 프리스케이팅의 각각 기술적표현과 예술적표현의 점수를 합하여 그 순위를 정하였다. 이 심사제도는 구체적인 기술의 사용 등에 대한 점수를 공개하지 않음으로 끊임없는 편파판정의 시비를 불러 왔다. 특히 2002년 솔트레이크 동계올릭픽 페어부문에서 추문이 일어나 2002년 국제빙상연맹은 새로운 심판제도의 도입하기로 하고 2004/2005년 시즌부터 이 새로운 심사제도를 모든 국제시합에 적용하기 시작하였다. 그림 4.1은 2008년도 세계선수권대회여자싱글에 참가한 김연아선수의 프리스케이팅 점수이다. 성적은 12명의 심판이 두 부분으로 작성되는데 하나는 구사한 기술에 대한 점수와 프로그램구성요소별 점수이다. 이렇게 작성된 점수는 컴퓨터에 의해 무작위로 9명이 선발

ISU World Figure Skating Championships 2008

LADIES FREE SKATING JUDGES DETAILS PER SKATER

Rank	Name	NOC Code	Total Segment Score =	Total Element Score +	Total Program Component Score (factored) +	Total Deductions -
1	Yu-Na KIM	KOR	123.38	64.82	58.56	0.00

#	Executed Elements	Info	Base Value	GOE	The Judges Panel (in random order)											Scores of Panel
1	3F+3T		9.50	1.86	1	2	2	2	1	2	2	2	2	2	3	0	11.36
2	2A		3.50	0.57	1	0	1	0	0	1	0	2	1	1	2	1	4.07
3	FSSp4		3.00	0.50	1	1	0	1	1	1	1	2	2	1	1	0	3.50
4	3Lz+2T+2Lo		8.80	0.71	1	1	1	0	0	0	1	1	0	1	2	-1	9.51
5	SpSq4		3.40	1.00	1	1	1	0	1	1	1	1	1	1	2	1	4.40
6	2A+3T		8.25 x	0.43	0	0	0	1	0	0	1	0	1	0	2	-1	8.68
7	FCCoSp4		3.50	0.64	1	1	0	0	1	1	2	2	2	2	2	0	4.14
8	1Lz		0.66 x	0.03	0	1	-1	0	-1	0	0	0	0	0	1	0	0.69
9	3S		4.95 x	-0.86	-1	-1	-1	-1	-1	0	0	-1	-1	-1	0	-1	4.09
10	CoSp3		2.50	0.14	0	0	0	0	0	0	0	1	1	1	1	0	2.64
11	SlSt3		3.10	0.29	1	1	0	0	0	1	1	1	0	1	1	1	3.39
12	2A		3.85 x	0.71	1	1	1	0	0	2	1	1	0	1	1	0	4.56
13	CCoSp4		3.50	0.29	0	0	1	0	0	1	1	1	1	1	1	0	3.79
			58.51														64.82
	Program Components			Factor													
	Skating Skills			1.60	7.50	7.75	7.25	7.25	7.25	7.25	8.50	7.25	8.00	7.25	8.00	7.50	7.57
	Transition / Linking Footwork			1.60	6.75	7.50	6.00	7.00	6.75	6.25	7.00	7.00	7.25	6.75	8.00	7.00	7.00
	Performance / Execution			1.60	7.25	7.50	6.50	7.25	6.75	7.50	8.25	7.25	8.25	7.00	8.25	7.50	7.46
	Choreography / Composition			1.60	7.25	7.25	6.75	6.75	7.00	7.25	8.00	7.00	7.75	7.00	7.75	7.00	7.25
	Interpretation			1.60	7.25	7.25	6.50	7.00	6.75	7.50	8.25	7.00	7.75	7.25	8.00	7.25	7.32
	Judges Total Program Component Score (factored)																58.56
	Deductions:																0.00

* Invalid element e Jump take off with wrong edge < Downgraded jump x Credit for highlight distribution, jump element multiplied by 1.1

그림 4.1: **김연아선수의 2008년 세계선수권대회 여자싱글 프리스케이팅 성적**

되고 최고점과 최저점을 제외한 7명의 점수의 평균이 해당 선수의 최종점수가 된다. 자세한 점수부여는 국제빙상연맹(ISU)의 규정집을 참고하기 바란다.

두 부분의 점수 가운데 프로그램구성요소별 점수는 구 심판제도와 유사하다. 즉 이 부분은 심판들간의 일치도가 중요하다고 할 수 있다. 여기서 우리는 각 프로그램 구성 요소별로 쇼트프로그램과 프리스케이팅의 점수들의 일치도를 알아보려고 한다. 프로그램의 다섯가지 구성요소는 스케이팅기술(Skating Skill), 트랜지션(Transitions), 수행(Performance/Execution), 구성/안무(Choreography), 연출/해석(Interpretation) 등 모두 다섯가지이다. 심판은 10점 만점으로 0.25점 간격으로 점수를 부여한다. 2008년도의 출전선수는 53명으로 이중 23명이 프리스케이팅에 진출하였다.

분석을 위하여 선택한 평가대상내 일치도는 먼저 2절의 예에서 사용한 식 (2.1)을 사용한다. D_1과 D_2의 값에 따라 범주의 값을 정하였는데 심판들이 0.25간격으로 점수를 부여하기에 0.25당 1을 부여하였다. 예를 들어 2이면 2×0.25 즉 0.5점의 차이를 말한다. 김연아선수의 프리스케이팅의 스케이팅스킬의 최고점과 최저점의 차이가 $8.5 - 7.25 = 1.25$이므로 속하는 범주의 값은 5가 된다. 표 4.1은 프리스케이팅에 진출한 23명의 선수들의 스케이팅스킬에 대한 쇼트프로그램과 프리스케이팅의 평가대상내 일치도를 구해 분할표로 작성한 것이다. 나머지 네 개의 구성요소에 대하여서도 같은 작업을 통하여 분할표를 얻었다. 이렇게 얻어진 다섯개의 분할표상에서 주변동질성의 검정을 실시하였다. 검정통계량의 계산은 Uebersax (2006)가 작성한 프로그램(mh.exe)를 이용하여 계산하였으며 표 4.2는 스케이팅스킬 구성요소에 대한 결과이다. 비교를 위하여 단측검정 즉 주변확률간

표 4.1: 쇼트프로그램과 프리스케이팅의 스케이팅스킬 구성요소에 대한 범위를 사용한 평가대상자내 일치도의 분할표

		프리스케이팅								
		2	3	4	5	6	7	8	계	누계
쇼트프로그램	2					1		1	2	2
	3		1		3				4	6
	4		2		6	1	1	1	11	17
	5		1	1					2	19
	6			3					3	22
	7					1			1	23
	8								0	23
	계	0	4	4	9	3	1	2	23	
	누계	0	4	8	17	20	21	23		

표 4.2: 쇼트프로그램과 프리스케이팅의 구성요소별 주변동질성 검정통계량 - 범위에 의한 일치도

구성요소	Bhapkar	(p-value)	Stuart -Maxwell	(p-value)	df	LRT	(p-value)
스케이팅기술	13.835	(0.0315)	8.639	(0.1949)	6	8.636	(0.2951)
트랜지션	8.594	(0.1977)	6.256	(0.3951)	6	5.922	(0.1311)
수행	13.530	(0.0354)	8.519	(0.2025)	6	8.515	(0.0796)
구성/안무	4.978	(0.5466)	4.093	(0.6642)	6	4.090	(0.3400)
연출/해석	21.957	(0.0012)	11.233	(0.0814)	6	11.230	(0.0283)

의 확률적순위가 존재한다는 대립가설에 대한 우도비검정의 경우 검정통계량의 값과 유의확률도 계산하였다. 참고로 우도비검정의 유의확률은 등가중치의 레벨확률(equal weight level probability)을 사용하였다. 자세한 내용은 Robertson *et al.* (1988)을 참조하기 바란다.

일반적으로 Bhapkar검정이 검정력이 높다고 알려져 있는 것과 같이 스케이팅기술, 수행 그리고 연출/해석의 세가지 구성요소에서 유의수준을 0.05로 할 때 주변동질성이 기각되었다. 반면 Stuart-Maxwell 검정의 경우는 다섯가지 구성요소 모두에서 주변동질성이 기각되지 못하였다. 표 4.1에서 누계를 비교해 보면 프리스케이팅보다 쇼트프로그램이 좀 더 일치한다고 말할 수 있다. 즉 주변확률들간에 확률적순위가 정확하게 관찰된다. 그러나 검정결과는 통계적으로는 유의하지 않다. 이러한 현상은 트랜지션을 제외하고는 똑같은 현상이 관찰되지만 연출/해석을 제외하곤 통계적으로는 유의하지 않다.

다음은 평가대상내 일치도를 표준편차로 정의한 경우를 고려해 보자. 즉

$$D_1(\mathbf{X}_{1i}) = \sqrt{\frac{\sum_{j=1}^{k_1} X_{1ji}^2 - \left(\sum_{j=1}^{k_1} X_{1ji}\right)^2 / k_1}{k_1}}, \quad D_2(\mathbf{X}_{2i}) = \sqrt{\frac{\sum_{j=1}^{k_2} X_{2ji}^2 - \left(\sum_{j=1}^{k_2} X_{2ji}\right)^2 / k_2}{k_2}}.$$

표 4.3: 쇼트프로그램과 프리스케이팅의 구성요소별 주변동질성 검정통계량—표준편차에 의한 일치도

구성요소	Bhapkar	(p-value)	Stuart -Maxwell	(p-value)	df	LRT	(p-value)
스케이팅기술	3.450	(0.6310)	3.000	(0.7000)	5	4.449	(0.0998)
트랜지션	9.530	(0.0490)	6.738	(0.1500)	4	5.737	(0.2296)
수행	9.825	(0.0430)	6.884	(0.1420)	4	6.884	(0.1043)
구성/안무	4.652	(0.1990)	3.870	(0.2760)	3	2.069	(0.4510)
연출/해석	19.277	(0.0020)	10.487	(0.0630)	5	10.484	(0.0498)

표 4.3은 D_1과 D_2를 계산하여 이를 0.1단위로 범주화하여 주변동질성을 검정한 결과이다. 대부분의 경우는 범위를 사용했을 때와는 크게 다른 점을 찾을 수 없지만 스케이팅기술과 트랜지션 등 두 가지의 구성요소에 대한 Bhapkar검정의 결과가 표 4.2에서 볼 수 있는 것과는 크게 상반된다.

표 4.2과 4.3에 본 바와 같이 평가대상내 일치도 즉 D_1과 D_2를 어떻게 정하느냐에 따라 결과가 달라짐을 볼 수 있다. 본 논문에서 평가대상내 일치도를 어떤 것으로 정하느냐는 문제는 논의대상에서 제외를 하지만 적정한 평가대상내 일치도를 찾는 것도 중요한 문제가 될 것이다.

참고문헌

Agresti, A. (2002). *Categorical Data Analysis*, John Wiley & Sons, New York.

Bhapkar, V. P. (1966). A note on the evidence of two test criteria for hypotheses in categorical data, *Journal of the American Statistical Association*, **61**, 228–235.

Caussinus, H. (1966). Contribution à l'analyse statistique des tableaux de corrélation, *Annales de la faculté des sciences de Toulouse*, **29**, 77–182.

Cohen, J. (1966). A coefficient of agreement for nominal scales, *Educational and Psychological Measurement*, **20**, 37–46.

Donner, A., Shoukri, M. M., Klar, N. and Bartfay, E. (2000). Testing the equality of two dependent kappa statistics, *Statistics in Medicine*, **19**, 373–387.

Fleiss, J. L. and Cohen, J. (1973). The equivalence of weighted kappa and the intraclss correlation coefficient as measures of reliability, *Educational Psychology Measurement*, **33**, 613–619.

Lang, J. B. (2007). "Maximum Likelihood Fitting of Multinomial-Poisson Homogeneous (MPH) Models for Contingency Tables using MPH.FIT." online html document, "Available from: http://www.stat.uiowa.edu/ jblang/mph.fitting/mph.fit.documentation.htm"

McKenzie, D. P., MacKinnon, A. J., Peladeau, N., Onghena, P., Bruce, P. C., Clarke, D. M., Harrigan, S. and McGorry, P. D. (1996). Comparing correlated kappas by resampling: Is one level of agreement significantly different from another?, *Journal of Psychiatric Research*, **30**, 483–492.

Oh, M. (2008). Inference on measurements of agreement using marginal association, *Journal of the Korean Statistical Society, To Appear.*

Robertson, T., Wright, F. T. and Dykstra, R. L. (1988). *Order Restricted Statistical Inference*, John Wiley & Sons, Chichester.

Schuster, C. and von Eye, A. (1998). Modeling turn-over tables using the GSK approach, *Methods of Psychological Research Online*, **3**, 39–53, Available from: http://www.pabst-publicatioers.de/mpr/.

Stuart, A. A. (1955). A test for homogeneity of the marginal distributions in a two-way classification, *Biometrika*, **42**, 412–416.

Sun, X. and Yang, Z. (2008). Generalized McNemar's test for homogeneity of the marginal distributions, *SAS Grobal Forum 2008*, Paper 382–2008.

Uebersax, J. S. (2006). *User Guide for the MH Program (Version 1.21), Computer program documentation*, Available from: http://ourworld.compuserve.com/homepages /jsuebesax/mh.htm

[2008년 6월 접수, 2008년 7월 채택]

Comparison of Two Dependent Agreements Using Test of Marginal Homogeneity

Myongsik Oh[1)]

Abstract

Oh (2008) has proposed the one-sided likelihood ratio test of the equality of two agreement measures. However the use of this test may be limited since the computations of test statistic and critical value are not easy. We propose a test for comparing two dependent agreements using some well known tests for marginal homogeneity, for instance, Bhapkar test, Stuart-Maxwell test. Data obtained from 2008 world figure skating championship ladies single is analyzed for illustration purposes.

Keywords: Agreement; marginal homogeneity; Bhapkar test; Stuar-Maxwell test; generalized McNemar test.

1) Professor, Department of Data Management, Pusan University of Foreign Studies, 55-1 Uam-Dong, Nam-Gu, Busan 608-738, Korea. E-mail: moh@pufs.ac.kr