한 국 가 정 과 학 회 지 3(2) : 90~100, 2000
Korean J. of Human Ecology 3(2) : 90~100, 2000

■연구논문■

가정학 연구에서 활용되는 관능검사에 대한 신뢰도 평가방법에 관한 연구

위 은 하·박 정 수*

전남대학교 가정교육학과
*전남대학교 통계학과

Improvement of Reliability in the Sensory Evaluation for Home Economics Research

Eun-Hah Wee · Jeong-Soo Park*

Department of Home Economics Education, Chonnam National University
**Department of Statistics, Chonnam National University*

Abstract

Statistical reliability in the sensory evaluation measures the agreement of judgements for several observers. Kendall's coefficient of concordance for multiple rankings, and coefficient of consistence and coefficient of agreement for paired comparisons are frequently used as the standard reliability measures. The main idea and the computational formula of these coefficients in various situations are explained in detail with some real(cloth designing) examples. Moreover a user-friendly computer program called as MultPair is introduced. We expect that this expository paper and the computer program give a practical help to the researchers who use the sensory evaluation techniques.

Key words: Sensory evaluation, Kendall's coefficient of concordance, Multiple rankings.

†Corresponding author : Dept. of Home Economics Education Chonnam National Univ.
300 Yongbong-dong, Puk-gu, Kwangju, 500-757, Korea
Tel : 062-530-1317, Fax : 062-530-1349
E-mail : weh@chonnam.ac.kr

Ⅰ. 서 론

관능검사를 이용한 평가 또는 연구는 가정과학뿐 아니라 체육분야, 예술분야, 소비자 행동과학 및 농학 등 다양한 분야에 응용되고 있다. 이 관능검사에는 필수적으로 검사 대상물과 이를 평가하는 검사자가 존재한다. 관능검사의 성패는 실험방법과 검사자에 좌우된다고 할 수 있다. 효과적인 실험방법과 숙련된 검사자들의 평가는 실험비용을 줄일 뿐 아니라 관능검사 결과 자체의 신뢰도를 향상시킨다. 따라서 관능검사는 인간의 감각에 의하여 평가하는 것으로 다른 주관적 변인에 영향을 적게 받는 전문가에 의해 이루어지는 것이 바람직한 질적평가라고 볼 수 있다. 관능검사에 관한 전반적인 논의는 김광옥 外 3인(1993) 또는 Meilgaard, Civille and Carr(1990) 등을 참조하기 바란다.

만약 한 대상물에 대해서 검사자들 간에 매우 상이한 평가를 내렸다면 그 대상물은 관능적으로 어떤 일치된 평가를 받기 어려운 대상물이라고 볼 수 있다. 이러한 상이한 평가를 바탕으로 하여 어떤 결론을 유추하는 것은 진실로부터 멀리 떨어진 결론에 도달할 수 있는 위험이 있다. 따라서 검사자들 간에 평가가 대체적으로 일치되는 경우에만 연구가 진행되어야 한다.

검사자들 간에 평가가 얼마나 일치되는가를 재는 척도가 관능검사에서의 통계적 신뢰도 계수이다. 이 신뢰도 계수는 관능검사의 방법에 따라 다르게 정의되고 계산된다. 이들은 다점순위법에서의 켄달의 일치성계수(coefficient of concordance, W), 일대일 비교에서의 일관성계수(coefficient of consistence, c) 및 일치성계수 (coefficient of agreement, u)이다.

국내의 가정과학 연구 논문집에 자주 관능검사에 의한 연구가 보고되고 있으나 신뢰도에 대한 논의가 없는 논문이 쉽게 발견된다(예를들어, 임희정, 1996; 장경아, 1985; 박채련, 1997). 설문지를 이용한 조사 연구에는 대부분 크론바하의 알파를 제시하여 신뢰도를 검증하고 있는데 비하여, 관능검사의 연구에서는 신뢰도 제시가 미비한 상황이다. 또한 신뢰도에 관한 논의가 있다 하더라도 자세한 설명이 없어서 독자들이 비슷한 연구에서 신뢰도를 구할 수 없다. 실제로 김태경, 이경희, 박정순(1990)은 신뢰도의 계산 공식들을 제시하고 있지만 이해하기 힘들다[1]. 아울러 신뢰도의 공식을 잘 이해하고 있다 하더라도 반복되는 신뢰도의 계산을 손과 계산기(table calculator)로 해내기란 어려운 일이다. 이상과 같이 관능검사에서의 신뢰도가 제시되지 않는 이유는, 첫째 위에서 언급한 세 가지의 신뢰도 계수에 대한 이해가 부족하다. 이렇게 이해가 부족한 이유는 관능검사에서의 신뢰도 계수에 대하여 가정과학 연구자들이 쉽게 이해할 수 있는 문헌이나 정보가 부족했기 때문일 것이다. 둘째 이유는 계산이 용이하지 않았기 때문이다. 대부분 위의 신뢰도 계수를 여러 번에 걸쳐서 구해야 되는데, 기존의 통계 패키지(SPSS, SAS, BMDP 등)에는 이들을 직접 계산할 수 있는 프로시져가 없다.

따라서 본 연구에서는 통계학 전공자가 아닌 연구자들도 이해할 수 있도록 다점순위법과 일대일비교법에서의 신뢰도 계수들의 공식과 개념을 쉽고 자세하게 설명하고자 한다. 그리고 다음으로 대형 관능검사에서도 손쉽게 사용할 수 있는 MultPair 라고 불리는 신뢰도 계산을 위한 컴퓨터 프로그램을 실제 적용 예를 통해 쉽게 이해하도록 소개하고자 한다.

본 연구의 구성은 다음과 같다. 2절에서는 다점순위법에서의 신뢰도 계수(켄달의 일치성 계수)의 개념과 계산공식을 주로 설명, 그리고 실제 적용의 예와 함께 다루었으며, 추가적으로 분산분석과 다중비교와 같은 통계적 분석이 언급되어졌다. 3절에서는 일대일 비교에서의 신뢰도 계수(일관성계수 및 일치성계수)의 개념과 계산 공식이 실제 적용 예와 함께 설명되어있다. 마지막으로 4절에서는 요약 및 토의가 주어졌다. 한편 본 報文에 기술된 신뢰도의 개념 및 계산 공식은 주로 Kendall and Gibbons (1990)에 기초하고 있음을 밝혀둔다.

Ⅱ. 다점순위법에서의 신뢰도 계수

본 절에서는 다점순위법에서의 신뢰도인 켄달의 일치성계수의 개념과 계산공식을 자세히 설명한다. 또한 계산된 계수의 통계적 유의성과 분산분석 및 다중비교에 관해서도 기술한다. 이의 계산은 MultPair 내에서 multi 라는 이름의 실행 화일에서 가능하다.

1) 그들은 주로 佐藤信(1985)의 서적을 통하여 신뢰도의 개념을 이해하고 계산해 온 것으로 보인다.

2.1 기본개념과 계산공식

m명의 검사자가 각각 n개의 대상물에 대해서 1에서 n까지 순위를 매기는 경우를 생각해 보자. 예를 들어 아래와 같은 〈표 2.1〉을 얻었다고 하자. 높은 순위를 받을수록 좋게 평가된 대상물이라고 하자. 낮은 순위를 받을수록 좋게 평가된 대상물이라고 하더라도 본 논문에서 다루는 신뢰도 값 및 통계적 분석의 결과에는 변함이 없다. 이때 우리의 관심은 4명의 검사자 간에 얼마나 일치된 평가를 내렸는가이다. 검사자들 간의 평가의 일치성의 한 척도로서 켄달(Kendall)의 일치성 계수(coefficient of concordance)를 정의한다. 이 계수는 각 대상물의 순위합의 편차 제곱합에 기초하고 있다. 순위의 총 합계는 84이므로 각 대상물의 평균 순위합은 84/6=14이다. 각 대상물의 순위합에 기초한 편차제곱합 S는 다음과 같이 계산한다.

$$S=(15-14)^2+(11-14)^2+(10-14)^2+$$
$$(19-14)^2+(12-14)^2+(17-14)^2=64.$$

〈표 2.1〉 5개의 검사물에 대한 4명의 검사자의 다점순위의 결과 예

검사자 / 대상물	A	B	C	D	순위합
1	5	2	4	4	15
2	4	3	1	3	11
3	1	1	6	2	10
4	6	5	3	5	19
5	3	6	2	1	12
6	2	4	5	6	17
					84

만약 순위가 검사자들 간에 아무런 일치성도 없이 랜덤하게 매겨졌다면, 각 대상물의 순위합은 모두 14 에 가까운 값이 될 것이고, 따라서 S는 0에 가까운 값이 될 것이다. 반면 검사자들 간에 모두 일치된 순위를 주었다면 가장 좋은 대상물은 모두 6점을 받아서 순위합이 24가 되고, 가장 나쁜 대상물은 모두 1점을 받아서 순위합이 4가 된다. 그러면

$$S=(4-14)^2+(8-14)^2+\cdots+(24-14)^2=280$$

이 된다. 따라서 이 S 값이 클수록 검사자들 간에 일치도가 높은 것이며 이 값이 작을수록 검사자들 간에 일치도가 낮은 것임을 알 수 있다. 이를 표준화해 주기 위해서 S가 취할 수 있는 최대값으로 나눈 값을 켄달의 일치성 계수 W 로 정의한다. 즉 $W = S/\max S$ 인데, 위의 예에서는 64/280=0.229가 된다.

이제 일반적인 경우로 확장하여 생각해 보자. m명의 검사자가 n개의 대상물을 다점순위법로 평가할 때, R_1, R_2, …, R_n을 각 대상물의 순위합이라고 하자. 그런데 평균순위는 m(n+1)/2 이므로,

$$S = \sum_{i=1}^{n}\left[R_i-\frac{m(n+1)}{2}\right]^2 = [\sum_{i=1}^{n}R_i^2]-\frac{nm^2(n+1)^2}{4}$$

가 된다. 또한 S가 취할 수 있는 최대값은(완벽한 일치가 일어난 경우로서)

$$\max S=\frac{m^2(n^3-n)}{12}$$

이다. 따라서 켄달의 일치성 계수(coefficient of concordance) W는

$$W = \frac{12S}{m^2(n^3-n)}$$

으로 구해진다. 즉 W가 1에 가까울수록 검사자들 간에 높은 일치성을 나타내고, W가 0에 가까울수록 검사자들 간에 매우 낮은 일치성을 나타낸다. 프로그램 MultPair 내에 multi.exe라는 실행화일이 W를 계산한다. 또한 multi 프로그램에서는 각 검사자를 한 명씩 빼고, 그 때마다 (m-1)명에 대한 W값을 계산하여, 한 명도 빼지 않은 상태의 즉, m 명에 대한 W 값과 비교함으로써, 각 검사자의 일치성계수 변화의 역할을 알아보았다. 만약 i 번째 사람을 뺐을 때 W가 증가했다면, 그 i 번째 검사자는 전체 일치성을 감소시키는 역할을 함을 알 수 있다.

2.2 일치성 계수의 통계적 유의성

일단 구해진 일치성계수에 대하여 다음과 같은 통계적 가설검정을 생각해 볼 수 있다.

$$H_0 : W = 0,\ H_1 : W \neq 0.$$

즉 귀무가설은 순위가 검사자들 간에 아무런 일치성도 없이 랜덤하게 매겨졌다는 것이고, 대립가설은 검사자들 간에 아무런 일치성이 없는 것은 아니다는 가설이다.

이를 위하여 m과 n이 작은 경우 즉, $n=3$, $m=2$ 부터 10까지, 또는 $n=4$, $m=2$ 부터 6까지, 또는 $n=5$, $m=3$ 경우에는 정확한 확률이론에 의해 가설검정이 가능하고, Kendall and Gibbons(1990)의 〈표 5〉로 부터 기각값을 구할 수 있다.

한편 m과 n이 상당히 클 때 핏셔(Fisher)의 z-분포를 이용할 수도 있고[2], 다음과 같은 카이제곱 분포를 이용할 수 있다[3]. 이 경우 검정통계량은

$$\chi_0^2 = m(n-1)W$$

이다. 이 통계량은 자유도 v=n-1 을 갖는 카이제곱 분포를 따른다. 따라서 유의수준 α에서의 기각영역은 $\chi_0^2 > \chi^2(v;\alpha)$이다. 여기서 $\chi^2(v;\alpha)$는 자유도 v를 갖는 카이제곱 분포의 상위 α 임계값을 말한다. 또한 유의확률인 p-값은 $p = P(\chi_v^2 > \chi_0^2)$으로 계산되어 진다. 여기서 χ_v^2는 자유도 v를 갖는 카이제곱 확률변수를 표시한다. multi 프로그램은 χ_0^2 값과 유의확률(p-값)을 계산해 준다. 만약 계산된 유의확률이 유의수준(보통 5% 또는 1%, 즉 0.05 또는 0.01) 보다 작으면 그 유의수준에서 귀무가설을 기각하고, 반대로 유의확률이 유의수준보다 크면 귀무가설을 채택한다.

2.3 프리드만 검정과 다중비교

여기서 한가지 언급할 것은 위의 검정통계량 χ_0^2 가 프리드만(Friedman)의 이원 분산분석 검정통계량과 일치한다는 점이다. 즉 m명의 검사자가 블록이 되고, n개의 대상물이 수준이 되어 n개 대상물들 간에 평균순위에 차이가 있는가 없는가를 검정하는 임의화된 완비블럭계획(randomized complete block design)에 대한 프리드만의 검정법과 일치한다.

우리는 multi 프로그램에서 위에서 언급한 유의성 검정에 그치지 않고, 프리드만 검정에 기초한 다음과 같은 다중비교를 실시하였다. 즉 모든 가능한 2개의 대상물들 간에 평균 순위에 차이가 있는가를 검정하기 위하여, 실험오차율(experiment-wise error rate)을 α로 했을 때,

$$|R_i - R_j| \geq z\sqrt{\frac{mn(n+1)}{6}}$$

가 만족되면 i번째 대상물과 j 번째 대상물은 평균순위에서 유의한 차이가 있다고 말한다(Daniel, 1978, p.231). 이때 z는 표준정규분포에서의 상위 $\alpha/n(n-1)$ 임계값(또는 퍼센타일)을 표시한다.

2.4 적용 사례: 중년여성의 의장효과

중년여성의 불균형적인 체형을 보다 균형있게 조형화시킬 수 있는 의장효과를 살펴보기 위하여 의복형태의 면분할에 의한 의장효과를 다점순위법과 일대일비교평가에 의한 관능평가로 알아보았다(일대일비교 평가의 적용사례는 3.6절에 기술됨). 의장효과의 평가는 유행이나 개인적인 취향 등의 변인에 의한 영향을 억제하고 보다 고차원적인 디자인감각으로 이루어져야하는 질적평가를 의도하여 관능평가단은 박사과정 이상의 의복디자인을 전공하는 전문평가인 7명으로 구성되었다.

〈표 2.2〉 다점순위법에서 검사자가 사용한 평가표

평가항목 \ 평가디자인	(A)-1	(A)-2	(A)-3	(A)-4
1. 키가 커 보이며 날씬해 보인다				
2. 어깨가 넓어 보인다				
3. 가슴 부위가 좁아 보인다				
4. 허리가 가늘어 보인다				
5. 상하의 면적분배가 적절하다				
6. 상의의 전체적 면적분배가 적절하다				
7. V-네클라인의 크기가 적절하다				
8. 전체적으로 조화되어 보인다.				

2) Kendall and Gibbons(1990)의 p.122 와 〈표 7〉.
3) 실제로 multi 프로그램에는 핏셔의 z값이 계산되지만 유의확률은 제시되지 않고 있고, 카이제곱 분포를 이용하는 방법이 사용되었다.

블라우스+슬랙스+자켓으로 조합된 의복을 자켓의 길이(5종), 칼라의 유무(2종), V-zone의 깊이(2종), 허리둘레의 맞음새(2종)를 변인으로 40(5×2×2×2)개를 디자인하여 제작하였다. 제작된 실험디자인은 평균체형의 중년여성에게 착장시킨 후 사진 촬영하여 평가용 자극물로 만들어졌다. 평가는 평가용 사진을 보면서 이루어졌다.

실험디자인의 변인 중 허리 맞음새(2종류), V-네클라인의 깊이(2종류)가 다른 4종류의 디자인을 1단위로 하여 10단위(A, B, C, D, E, F, G, H, I, J)로 나누었으며 평가항목에 따라 한 번에 순위를 매기는 다점순위법(multiple rankings)으로 평가하였다. 즉, 평가내용에 가장 가까운 것에는 4점을 매기고, 가장 먼 것에는 1점을 매겨 실험디자인의 순위를 매겼다. 이때 각 검사자가 사용한 평가표가 〈표 2.2〉에 주어졌다.

디자인 단위 A에 대해 다점순위법에 의한 검사결과에 바탕하여, 실제로 multi 프로그램을 수행하기 위한 입력표는 〈표 2.3〉과 같다.

〈표 2.3〉 다점순위에 대해 multi 프로그램 수행을 위한 입력표

실험디자인(A), 평가항목(8)

실험디자인 \ 검사자	1	2	3	4	5	6	7
(A)-1	3	2	3	3	3	4	3
(A)-2	4	4	4	4	4	3	4
(A)-3	2	1	1	2	1	2	2
(A)-4	1	3	2	1	2	1	1

다음은 위의 〈표 2.3〉의 자료에 대한 프로그램 multi의 출력결과이다. 각 대상물(디자인)의 순위 합계, 평균 및 순위 제곱합(S)과 켄달의 일치성계수(W), 유의성 검정의 결과 및 프리드만의 검정과 다중비교, 그리고 각 검사자를 한 명씩 빼고 그 때마다의 W값을 계산하였다. 다중비교의 결과에서 *는 α=0.05에서, **는 α = 0.01 에서 유의함을 표시한다.

```
Sample=a, Item=8
   sum mean  1  2  3  4  5  6  7 ← 검사자 번호
1  21 3.00  3  2  3  3  3  4  3
2  27 3.86  4  4  4  4  4  3  4
3  11 1.57  2  1  1  2  1  2  2
4  11 1.57  1  3  2  1  2  1  1

S=187.0 Kendalls W= .763  ← 순위 제곱합과 켄달의 일치성계수
Chi-square=16.03 df=3.0 Approx. P= .0011 ←유의성 검정
For better P-value, Fisher z=1.481 v1=2.71 v2=16.29
Friedman test(H0:random ranks) T=16.03 Approx. P= .0011

== Multiple comparison based on Friedman test ==
  Comparison  Dist.  Signif.
    1- 2      6.0
    1- 3     10.0
    1- 4     10.0
    2- 3     16.0     **
    2- 4     16.0     **
    3- 4      .0

== Change of coeff. due to deletion of each judge ==
  Deleted judge = 1   Kendalls W = .744
  Deleted judge = 2   Kendalls W = .856
  Deleted judge = 3   Kendalls W = .744
  Deleted judge = 4   Kendalls W = .744
  Deleted judge = 5   Kendalls W = .744
  Deleted judge = 6   Kendalls W = .811
  Deleted judge = 7   Kendalls W = .744
```

결과적으로 2번과 3번 디자인간에, 그리고 2번과 4번 디자인간에 유의적인 차이가 있음이 알 수 있다. 또한 2번 검사자가 다른 검사자들과는 다른 평가를 자주 내리고 있음을 알 수 있다.

Ⅲ. 일대일 비교

일대일 비교 (paired comparisons) 에는 두 가지의 신뢰도 계수를 구하게 되는데, 먼저 일관성 계수(c)는 각 검사자 개인이 얼마나 일관되게 평가를 하고 있는 가를 재는 것이고, 일치성 계수(u)는 여러 명의 검사자들이

얼마나 일치된 평가를 내리고 있는가를 재는 척도이다. 본 절에서는 먼저 일관성 계수의 개념과 계산공식을 설명한 다음 일치성계수의 개념과 계산공식을 기술한다. 이의 계산은 MultPair 내에서 pair 라는 이름의 실행 화일에서 가능하다. 그리고 실제 적용사례(3.6의 예)를 통하여 컴퓨터 프로그램 pair 의 효용성을 보인다.

3.1 일관성 계수의 기본개념

n개의 대상물이 있다고 가정하자. 일대일로 비교를 하려면 총 n(n-1)/2개의 짝이 관측자 또는 검사자(observer 또는 judge)에게 한 번에 한 짝씩 보여진다. 그래서 각 짝에서 선호도가 기록된다. 만약 B 보다 A를 더 선호하면 A->B 또는 B<-A 로 쓰기로 하자. 예를 들어서 설명해 보자.

예 3.1: 한 개(dog)의 음식 선호도 실험을 하기 위하여, 6개의 서로 다른 음식을 준비하였다. 이 음식에 A부터 F까지 부호를 붙여서 일대일 비교를 한다면 총 6(6-1)/2 =15개의 가능한 짝을 구성할 수 있다. 이들 짝을 차례로 개에게 제시하여 선호도를 적어서 만든 결과가 〈표 3.1〉에 나타나 있다.

〈표 3.1〉 한 개의 6가지 음식에 대한 선호도표 (preference table)

	A	B	C	D	E	F
A	-	1	1	0	1	1
B	0	-	0	1	1	0
C	0	1	-	1	1	1
D	1	0	0	-	0	0
E	0	0	0	1	-	1
F	0	1	0	1	0	-

이 표에서 A행의 B열에 1의 의미는 A->B 이다. 즉 왼쪽(행)을 기준 대상물로 보고 윗쪽(열)을 비교 대상물로 했을 때 기준이 비교보다 좋다면 1이고 나쁘다면 0으로 표시한다. A와 B를 바꾼 위치, 즉 B행의 A열은 당연히 0가 된다. 따라서 위의 표의 1행으로 부터 A->B, A->C, A<-D, A->E, A->F 를 알 수 있다.

만약 한 관측자가 3개의 대상물 A, B, C에 대해서 A->B->C->A 또는 A<-B<-C<-A 와 같은 선호를 보였다면, 이것은 일관성이 없는 선호임을 알 수 있다. 이것은 A가 B보다 좋고 B가 C보다 좋은데도 C가 A보다 좋다고 평가한 것이므로 일종의 모순(circular triad)이 일어난 것이다. 이와 같은 모순의 선호가 많이 일어나면 일어날수록 그 관측자에게는 일관성(consistence)이 떨어지게 된다. 이 점에 착안하여 일관성 계수(coefficient of consistence)가 유도된다[4].

3.2 일관성 계수의 계산 공식

만약 대상물의 수 n이 홀수이면 최대의 모순의 수는 $(n^3-n)/24$ 이고, 만약 n이 짝수이면 최대의 모순의 수는 $(n^3-4n)/24$ 라고 한다(Kendall and Gibbons, 1990). 물론 최소의 모순의 수는 0이다. 이제 한 검사자의 선호의 일관성을 재는 일관성 계수(coefficient of consistence)는 다음과 같이 정의된다. 이것은 2. 에서의 일치성 계수의 유도에서와 마찬가지로,

$$c = 1 - \frac{\text{관측된 모순의 수}}{\text{최대 모순의 수}}$$

로 정의된다. 이를 구체적으로 나타내면 다음과 같다.

$$c = 1 - \frac{24d}{n^3-n}, \quad n \text{ 홀수}$$

$$= 1 - \frac{24d}{n^3-4n}, \quad n \text{ 짝수}$$

여기서 d는 관측된 모순의 수이다. c는 0과 1사이의 값을 갖는다. 만약 c=1이면 모순이 하나도 없는 것이고, 또 모순이 하나도 없으면 c=1이 된다. 반대로 모든 선호에 다 모순이 생겨서 모순이 최대로 많아지면 c=0이 되고, 또 c=0이면 모든 선호에 다 모순이 일어났음을 뜻한다.

일관성계수를 계산하기 위해서 실제로 모순의 수를 선호도표나 그림으로부터 세기란 쉽지 않다. 그런데 그것을 낱낱이 세지 않고도 간단히 계산할 수가 있다. 〈표 3.1〉에서 각 행에 대한 합계를 각각

4) 佐藤信(1985)과 김태경(1990)은 일관성 대신 일의성 계수라는 용어를 사용하고 있다.

$a_1, a_2, \ldots, a_n$ 이라고 하자. 그러면

$$d = \frac{n(n-1)(2n-1)}{12} - \frac{1}{2}\sum_{i=1}^{m} a_i^2$$

로 간단히 계산된다. 위의 〈예 3.1〉에서는 5개의 모순이 발견되었고, 최대 가능한 모순의 수는 8개이므로, c= 0.375 이다.

3.3 일관성 유무에 대한 가설검정

위의 일관성 계수가 0이라고 할 수 있는지에 대한 가설검정을 하고자 한다. 이때 영가설은 H_0: c=0이고 대립가설은 H_1: c>0이다. 이를 위해서 c의 정확한 분포를 구하여 가설 검정을 수행할 수도 있지만, 여기서는 카이제곱 분포를 이용한다. 즉 검정통계량은

$$\chi_0^2 = \frac{8}{n-4}\left[\frac{n(n-1)(n-2)}{24} - d + 0.5\right] + v,$$

이고, 여기서 $v = n(n-1)(n-2)/(n-4)^2$ 이다. 위의 검정통계량은 근사적으로 자유도가 v인 카이제곱 분포를 따른다. 따라서 유의수준 α에서 기각영역은 $\chi_0^2 > \chi^2(v;\alpha)$이다. 즉 위의 조건이 만족되면 유의수준 α에서 영가설을 기각하고 그렇지 않으면 기각 못한다. 여기서 $\chi^2(v;\alpha)$는 자유도 v를 갖는 카이제곱분포의 상위 α 임계값(또는 퍼센타일)이다.

한편 주어진 자료로부터 유의확률(p-값)은 $p = P(\chi_v^2 > \chi_0^2)$로서 구해진다. 여기서 χ_v^2는 자유도 v를 갖는 카이제곱 확률변수를 표시한다. pair 프로그램은 χ_0^2 값과 유의확률 (p-값)을 계산해 준다. 계산된 유의확률이 유의수준[5]보다 작으면 그 유의수준에서 영가설을 기각하고, 반대로 유의확률이 유의수준보다 크면 영가설을 채택한다. 3.6의 적용사례의 경우 χ_0^2=27.33, 자유도 v=20.0 이고, 유의확률은 0.126이다. 따라서 유의수준 5%에서 검사자 5는 유의확률이 0.05보다 크기 때문에 영가설을 채택하게된다.

5) 보통 5%, 즉 0.05.

3.4 일치성 계수의 기본 개념과 계산공식

m명의 관측자가 n개의 대상물에 대해 일대일 비교를 통하여 각각 〈표 3.1〉과 같은 선호도표를 제공했다고 하자. 우리의 관심은 이제 이들 m명의 관측자들간에 얼마나 일치된 선호를 보이고 있는가 이다. 이를 알기 위해서 이들 m명이 만든 m개의 선호도표를 모두 합하게 된다. 그래서 선호도표상의 각 칸의 1의 수를 합하여 새로운 선호도표를 만든다〈표 3.2〉. 따라서 각 칸은 0부터 m까지의 값을 가질 수 있다. 만약 m명의 관측자들 간에 완벽한 선호도의 일치가 이루어졌다면, 전체 m^2개의 칸 중에서 $m(m-1)/2$개의 칸에서 m이라는 값이 있게 되고 나머지 칸에는 모두 0이 있게 된다. 이러한 완벽한 일치는 각 관측자에게서 일관성이 없더라도 일어날 수 있다. 또한 각 관측자들이 모두 각각의 일관성을 가진다 하더라도 완벽한 일치가 안 일어날 수도 있다.

여기서 일관성은 각 관측자 개인 내의 문제이고, 일치성은 m명의 관측자 모두를 하나로 보고 적용되는 문제이다. 따라서 일관성 계수는 각 관측자에게서 구해지지만 아래에서 계산되는 일치성 계수는 m명의 관측자에 대해서 한번 구해진다.

〈표 3.2〉에서와 같이 합해진 선호도표에서 i번째 행과 j번째 열에 해당되는 값을 γ_{ij} 라고 하자. 또

$$S = \frac{1}{2}\left[\sum_{i=1}^{m}\sum_{j=1}^{m}\gamma_{ij}(\gamma_{ij}-1)\right]$$

$$= \frac{1}{2}\left(\sum_{i=1}^{m}\sum_{j=1}^{m}\gamma_{ij}^2 - \sum_{i=1}^{m}\sum_{j=1}^{m}\gamma_{ij}\right)$$

라고 하자. 이 S는 m명의 관측자들 간의 일치하는 선호의 수의 합계를 나타낸다. 그러면 일치도 계수(coefficient of agreement) u는 다음과 같이 정의된다.

$$u = \frac{8S}{m(m-1)n(n-1)} - 1.$$

〈표 3.2〉 10개의 대상물(A부터 J까지)에 대해 7명의 관측자의 합해진 선호도표(3.6절의 적용사례의 결과)

	A	B	C	D	E	F	G	H	I	J	합계
A	-	0	0	0	6	4	2	2	0	0	14
B	7	-	1	0	7	7	2	2	2	2	30
C	7	6	-	5	7	7	0	7	1	0	40
D	7	7	2	-	6	5	1	0	2	1	31
E	1	0	0	1	-	5	2	0	2	1	12
F	3	0	0	2	2	-	2	2	2	1	14
G	5	5	7	6	5	5	-	5	1	1	40
H	5	5	0	7	7	5	2	-	1	2	34
I	7	5	6	5	5	5	6	6	-	5	50
J	7	5	7	6	6	6	6	5	2	-	50

만약 완벽한 일치가 일어나면 u=1이다(u의 최대값). u값이 1에서 떨어질수록 관측자들 간에 일치도가 감소함을 뜻한다. u의 최소값은 m이 짝수이면 -1/(m-1)이고, m이 홀수이면 -1/m이다. m=2인 경우에만 u의 최소값은 -1이고, 그 외에는 -1보다 큰 값이 된다.

위의 〈표 3.2〉로부터 계산된 S=687이고, 일치도 계수는 u=0.454이다. 한편 〈표 3.2〉에서 각 행에 대한 합계값이 클수록 그 대상물에 대한 선호가 많음을 나타낸다. 위의 예에서는 I와 J는 좋은 평가를 받은 반면 A, E, F는 상대적으로 안 좋은 평가를 받은 것이다.

3.5 일치도 계수에 대한 가설 검정

관측자들 간에 전혀 일치도가 없다고 할 수 있는지를 검정하고자 한다. 즉

영가설 H_0: 검사자들 간의 평가가 일치하지 않는다.

대립가설 H_1: 검사자들 간의 평가에 일치한다.

이 경우 다음과 같은 검정통계량을 이용한다(Kendall and Gibbons, 1990).

$$\chi_0^2 = \frac{4}{m-2}\left[S-1-\frac{n(n-1)}{4}\frac{m(m-1)}{2}\frac{m-3}{m-2}\right]$$

이 통계량은 자유도 v 를 가지는 근사적 카이제곱 분포를 따른다. 이때 자유도는

$$v = \frac{n(n-1)}{2}\frac{m(m-1)}{(m-2)^2}$$

이다. 유의확률(p-value)은 $P(\chi_v^2 > \chi_0^2)$으로 구해진다. 만약 유의확률이 유의수준 보통, 0.05 보다 작으면 검사자들 사이에 일치성이 전혀 없는 것은 아니라고 할 수 있다. 위의 〈표 3.2〉의 u=0.454에 대한 가설검정에서 χ_0^2 =246.4, 자유도 v=75.6이고, 유의확률은 0.000이다. 따라서 유의수준 1%에서 검사자들 사이에 약간 일치된(u=0.454) 판정을 내리고 있음을 알 수 있다.

3.6 일대일 비교의 적용사례: 중년여성의 의장효과

다점순위법에 의한 평가결과를 기본으로 평균순위가 가장 높은 실험디자인을 각 단위에서 1개씩 선택하여 일대일비교평가(paired comparison)를 실행하였다. 7명의 검사자가 평가에 참여했다. 각 단위에서 선택된 실험디자인은 A-2, B-4, C-2, D-2, E-2, F-2, G-2, H-1, I-3, J-3였다. 10개의 디자인에 대해 기준 실험디자인을 왼쪽에 두고 오른쪽의 비교 실험디자인이 어떻게 보이는가를 아래의 〈표 3.3〉의 평가표를 이용하여 평가하도록 하였다.

〈표 3.3〉 일대일비교에서 검사자가 사용하는 평가표

기준 샘플명: 비교샘플명:

평 가 내 용	그렇다	그렇치 않다
1. 키가 커 보이며 날씬해 보인다		
2. 어깨부위가 넓어 보인다		
3. 가슴 부위가 좁아 보인다		
4. 허리가 가늘어 보인다		
5. 상·하의 면적배분이적절하다		
6. 상의의 전체적인 면적배분이 적절하다.		
7. V-네클라인의 크기가 적절하다		
8. 전체적으로 조화되어 보인다.		

컴퓨터 프로그램 pair의 입력을 위해서 선호도표를 〈표 3.4〉와 같이 만들었다. 여기서 기준이 비교보다 좋은 경우는 1을 쓰고, 그렇지 않은 경우는 0을 쓰도록 하였다. 이것을 7개 항목에 대해 실시하였다. 실제로 한 평가항목에 대해, 기준샘플에 비하여 비교샘플이 평가내용에 더 가까울 때는 〈표 3.3〉의 "그렇다"라는 란에 표기하고 〈표 3.4〉에는 0으로 입력하였다. 반대로 기준

샘플이 평가내용에 더 가까울 때는 〈표 3.3〉의 "그렇치 않다"라는 란에 표기하고 〈표 3.4〉에는 1을 입력하였다. 이를 바탕으로 하여 한 평가자가 얼마나 일관성있게 평가하였는가를 알아보는 일관성 계수를 7명의 검사자에 대해, 그리고 7개 항목 모두에 대해 프로그램 pair를 이용하여 계산하고 유의성 검정을 실시하였다.

〈표 3.4〉 프로그램 pair실행을 위한 일대일 비교의 입력표

기준샘플(A) 비교샘플(E)

평가항목 \ 평가자	1	2	3	4	5	6	7
항목 1	0	0	0	0	0	0	1
항목 2	0	0	0	0	1	0	0
항목 3	1	1	1	1	1	1	0
항목 4	1	1	1	1	1	1	1
항목 5	0	0	0	0	0	0	1
항목 6	0	0	1	0	0	0	0
항목 7	1	1	1	0	0	1	1

다음은 특히 위의 〈표 3.2〉의 자료에 대한 프로그램 pair 의 출력결과이다. 각 검사자에 대해 모순의 개수(d)와 일관성 계수 c를 계산하고 그 유위성 검정(p-값)이 주어졌다. 또한 항목 1번에 대해 7명의 검사자의 평가의 일치성 계수 u와 그 유의성 검정 결과가 주어졌다. 그리고 각 검사자를 한 명씩 빼고 그 때마다의 u값이 계산되었다.

Item Number = 1

Result (Pref. Table) for Judge = 1
d= .0 Coeff. Consistence (c) = 1.000
Test H0:Coeff=0: Chi-square = 60.67 DF = 20.00
P-value = .0000

Result (Pref. Table) for Judge = 2
d= 4.0 Coeff. Consistence (c) = .900
Test H0:Coeff=0: Chi-square = 55.33 DF = 20.00
P-value = .0000

Result (Pref. Table) for Judge = 3
d= .0 Coeff. Consistence (c) = 1.000
Test H0:Coeff=0: Chi-square = 60.67 DF = 20.00
P-value = .0000

Result (Pref. Table) for Judge = 4
d= 2.0 Coeff. Consistence (c) = .950
Test H0:Coeff=0: Chi-square = 58.00 DF = 20.00
P-value = .0000

Result (Pref. Table) for Judge = 5
d= 25.0 Coeff. Consistence (c) = .375
Test H0:Coeff=0: Chi-square = 27.33 DF = 20.00
P-value = .1261

Result (Pref. Table) for Judge = 6
d= 8.0 Coeff. Consistence (c) = .800
Test H0:Coeff=0: Chi-square = 50.00 DF = 20.00
P-value = .0002

Result (Pref. Table) for Judge = 7
d= 20.0 Coeff. Consistence (c) = .500
Test H0:Coeff=0: Chi-square = 34.00 DF = 20.00
P-value = .0261

Preferences Table by 7 Judges for Item Number 1

〈선호도표 생략 (〈표 3.2〉와 같음)〉

S = 687.0 Coeff. Agreement(u) = .454
Test H0:Coeff=min: Chi-square = 246.40 DF = 75.60 P-value = .0000

= Change of coeff.(u) due to deletion of each judge =
Deleted Judge = 1 u = .384 p-value = .0000
Deleted Judge = 2 u = .419 p-value = .0000
Deleted Judge = 3 u = .384 p-value = .0000
Deleted Judge = 4 u = .393 p-value = .0000
Deleted Judge = 5 u = .609 p-value = .0000
Deleted Judge = 6 u = .437 p-value = .0000
Deleted Judge = 7 u = .553 p-value = .0000

위의 결과를 보면 5번 검사자는 일관성이 없게 나타났다(c=0.375, p-값=0.1261). 또한 일치성 계수에서도 5번 검사자를 제외시켰을 때 일치성 계수가 가장 크게 증가하였다. 이것은 5번 검사자 스스로의 평가에서도 일관성이 없을 뿐 만 아니라 다른 검사자와도 상이한

평가를 내리고 있음을 알 수 있다. 다른 평가항목에서도 비슷한 양상이 나타나는 가를 보기 위하여, 모든 항목에 대해서 프로그램 pair 를 이용하여 비슷한 일을 수행했다. 〈표 3.5〉는 검사자 7명 모두일 경우와 5번 검사자를 제외한 6명일 경우의 일치도 u값이다.

〈표 3.5〉 5번 검사자를 제외한 경우의 일치성 계수 u의 비교

u / 평가항목	검사자 전부인 경우	5번 검사자를 제외한 경우
	일치도 계수 u	일치도 계수 u
항목 1	0.454	0.609
항목 2	0.628	0.612
항목 3	0.814	0.825
항목 4	0.687	0.695
항목 5	0.674	0.748
항목 6	0.623	0.692
항목 7	0.522	0.567

이러한 분석에 바탕하여 결과적으로, 검사자 5는 일관성도 낮을 뿐 만 아니라 다른 평가자들 간의 일치도를 떨어뜨리는 결과를 가져오는 것으로 판단하여, 검사자 5를 제외한 결과를 의장효과평가의 결과로 사용하기로 하였다.

다음 보고에는 평가해야 될 대상물은 매우 많은데 비하여 숙련된 평가자 수가 상대적으로 작을 때, 또는 식품관능 검사에서와 같이 관능기관의 피로에 의해서 한 검사자가 한번에 여러 가지 대상물에 대해 순위를 주기 어려운 경우에 적용할 수 있는 불완전 순위(incomplete rankings)와 균형 불완비 블록 계획(balanced incomplete block design)을 소개하고자한다. 예를 들어 일곱 가지의 서로 다른 아이스크림 종류를 7명에게 각각 주어서 다점순위법에 의해서 순위를 매긴다고 해보자. 맛을 보는 기능을 갖는 혀가 처음 몇 개를 맛본 뒤에는 그 판별력을 잃기 때문에 한 검사자가 7개의 아이스크림에 대해 적절한 순위를 매기기를 기대하기는 어렵다. 따라서 한 검사자가 7개의 아이스크림 모두를 맛보지 않고, 적절히 뽑힌 3개의 아이스크림만 맛을 보아서 1, 2, 3으로 순위를 매기기로 하고, 각 검사자가 각기 다른 3개의 아이스크림에 대해 같은 작업을 하도록 한다. 이때 각 아이스크림이 반드시 같은 횟수만큼 검사되어야 한다. 이 경우에 신뢰도의 계산 및 유의성 검정에 대한 논의는 다음 논문에서 다룰 계획이다. 불완전 순위방법은 실험의 횟수를 줄이는 측면에서 효율적이고 비용을 아끼는 장점도 있지만, 실제로 관능 기관의 피로에 의해 생기는 측정오차를 상당히 줄일 수 있기 때문에 매우 유용한 방법이라고 생각된다.

Ⅳ. 요약 및 토의

관능검사에서는 여러 명의 검사자가 대상물에 대하여 관능적 판단에 기준하여 평가하는데, 이러한 관능적 평가가 검사자들 간에 얼마나 일치되는가를 재는 척도가 관능검사에서의 통계적 신뢰도이다.

본 연구에서 살펴본 신뢰도 계수로는 다점순위법에서의 켄달의 일치성계수, 일대일 비교에서의 일관성계수 및 일치성계수이다. 이들 신뢰도 계수들의 공식과 개념을 쉽게 설명하고, 이들의 유의성 검정에 대해 논의하였으며, 실제 적용 사례를 보였다. 그리고 특히 대형의 관능검사에서도 신뢰도 계수를 손쉽게 계산할 수 있도록 MultPair 라는 컴퓨터 프로그램을 소개하였다. MultPair에는 신뢰도 계산 뿐 만 아니라 분산분석과 다중비교와 같은 통계적 분석이 가능하도록 하였다. 또한 MultPair에는 검사자를 돌아가면서 한 명씩 제외시켰을 경우에 신뢰도 계수에 어떤 변화가 생기는 가를 관찰할 수 있도록 하였으므로, 검사자 개개인의 특성 내지는 공헌도를 파악할 수 있다. 실제 적용 예로는 중년여성의 의복의 의장효과에 관한 관능검사에 기초한 의류학 연구 사례(김옥진, 1997)가 다뤄졌다.

끝으로 본 연구에서는 다점순위법에서 2개 이상의 대상물에 대해 동점(tie)을 주었을 경우에 대하여는 다루지 못하였다. 이런 동점 상황은 특히 다점순위법에서 보다는 관능평가를 5점 또는 7점척도로 점수를 주었을 때 자주 발생한다. 이 경우에는 켄달의 일치도 계수의 계산공식이 달라지게 되므로 지속적으로 연구되어야 될 부분으로 생각된다.

본 연구와 MultPair 프로그램이 향후 관능검사를 이용한 연구에 도움이 되기를 바라며 MultPair가 필요한 독자는 필자에 연락하여 얻을 수 있다.

참 고 문 헌

김광옥, 김상숙, 성내경, 이영춘(1993), **관능검사 방법 및 응용**, 신광출판사, 서울

김옥진(1997), 중년여성의 체형에 적합한 의복형태의 면분할 및 배치방안에 관한 연구, **한국의류학회지** 21(7): 1173-1183

김태경, 이경희, 박정순(1990), 노년기여성의 배면만곡도 감소효과를 위한 의복디자인 연구, **한국의류학회지** 14(3): 183-195

박채련(1997), 체형에 따른 선의 시각적효과에 관한 연구 I, **대한가정학회지** 35(1): 307-318

위은하(1999), 중년여성의 체형에 적합한 시각효과를 위한 의복형태연구, 전남대학교, 박사학위논문

위은하, 김옥진(1999), 테일러드 수트의 형태구성요인의 조합에 따른 시각효과, **한국가정과학회지**, 2(1): 99-109

이경희, 박정순, 김태경(1990), 의복 디자인 선에 따른 시각적 효과에 관한 연구, **대한가정학회지**, 28(4): 314-323

정영아(1999), 원피스드레스형 임부복의 형태구성요인의 조합에 따른 시각효과, **한국가정과학회지**, 3(2): 49-62.

佐藤信(1985), **統計的官能檢査**, 日科技連, 東京.

日科技連官能檢査委員會(1992), **新版官能檢査ハンドブク**, 日科技連出版社, 東京.

Daniel, W.(1978), *Applied nonparametric statistics*, Houghton Mifflin Co. Boston.

Kendall, M., Gibbons, J.D.(1990), *Rank correlation methods*, 5th Edition, Edward Arnold, London.

Meilgaard, M., Civille, G.V., Carr, B.T.(1990), *Sensory evaluation techniques*, 2nd Edition, CRC Press Inc., Boca Raton, FL.