

순서 범주형 자료해석법의 비교 연구

- A Study on Comparison with the Methods of Ordered Categorical Data of Analysis -

김 홍 준*
Kim, Hong Jun
송 서 일**
Song, Suh Ill

Abstract

This paper deals with a comparison between Taguchi's accumulation analysis method and Nair test on the ordered categorical data from an industrial experiment for quality improvement. a result of Taguchi's accumulation analysis method is shown to have reasonable power for detecting location effects, while Nair test identifies the location and dispersion effects separately.

Accordingly, Taguchi's accumulation analysis needs to develop methods for detecting dispersion effects as well as location effects.

In addition this paper recommends models for analyzing ordered categorical data, for examples, the cumulative logit model, mean response model etc. Successively simple, reasonable methods should be introduced more likely to be used by the practitioners.

I 서 론

어떤 공업실험에서 특성치가 3조이상으로 분류되면서 분류한 조에 어떤 순서가 있는 경우, 즉, 상·중·하 또는 수·우·미 등으로 순서화되어지는 자료를 순서 범주형 자료라 한다[1]. 이러한 순서 범주형 자료의 해석 방법으로 실무에 널리 적용하고 있는 다구찌에 의한 방법과 이에 대한 대안으로 Nair 검정법으로 나누어 볼 수 있다. 먼저 다구찌방법으로는 누적법과 점수법으로 구분할 수 있는데, 누적법은 분류된 조에 도수를 누적하여, 이 누적 도수에 대해 분산분석을 하는 것이고, 점수법은 분류된 조에 점수(가중치)를 부여하여 해석하는 방법이다[9]. 그리고 Nair 검정법은 위치 및 산포도 효과의 유의성을 동시에 검출할 수 있는 점수법이다.

따라서 본 연구에서는 상기와 같은 순서 범주형 자료의 해석 방법에 관해 동일한 사례를 통하여 다구찌의 누적법과 Nair검정법을 적용시켜 그 결과를 비교 분석하고자 한다.

II 순서범주형 자료 해석 방법

2.1 다구찌의 누적법

순서범주형 자료를 분석하기 위해서 다구찌(1974)는 피어슨의 χ^2 검정의 대안으로 누적분석을 제안했다. 이 기법을 제안하게 된 동기는 양적 변수에 대한 분산분석의 유사성을 보여주는 것이다. 범주형 자료에 대해서는 Light 와 Margolin(1971)는 변동의 적절한 측도를 나타내는 CATANOVA 라고 부르는 방법을 제시하였다. 그러나 이러한 방법과는 달리 다구찌는 순서 범주형자료에 관한 해석법으로 누적도수에 대해 분산분석을 하였다. 우선, 누적법을 설명하기

* 대구산업 전문대학 산업안전과

**동아대학교 산업공학과 교수

위하여 I 수준의 인자 A로 구성되는 1인자 실험을 고려한다. 각 수준의 관측수는 동일한 n으로 가정하고 관측은 K순서 범주의 한가지로 분류되며, i수준의 범주K에서의 관측도수를 Y_{ik} 로 나타낸다. ($K=1, \dots, K ; i=1, \dots, I$) 누적도수는 $C_{ik} = \sum_{j=1}^k Y_{ij}$ 로 나타내고 A인자의 제곱합과 오차의 제곱합을 식(2.1) 및 식(2.2)로 표시된다.

$$SS_A = n \sum_{k=1}^{K-1} \sum_{i=1}^I [C_{ik} - C_{.k}] / [C_{.k}(n - C_{.k})] \tag{2.1}$$

$$SS_e = n \sum_{k=1}^{K-1} \sum_{i=1}^I [C_{ik}(n - C_{ik})] / [C_{.k}(n - C_{.k})] \tag{2.2}$$

그리고 인자 A 효과에 대한 검정은 식(2.3)의 통계량을 사용하고 있다[7]다시 말해서 누적법이란 누적 도수자료로부터 분산분석을 하는 것인데, 이것은 이치자료(binary data)로 간주하여 자료처리하는 것과 동일하게 되며, 각 범주에 대한 가중치는 식(2.4)와 같다

$$F_A = \frac{MS_A}{MS_e} \tag{2.3}$$

$$W_i = \frac{1}{cum_i / cum \times (1 - cum_i / cum)} = \frac{cum^2}{cum_i(cum - cum_i)} \tag{2.4}$$

여기서 cum은 i범주까지의 총누적도수이며, cum_i 는 i범주의 누적도수이다.

$cum_i / cum = p_i$ 라 두면, i번째 범주에 대한 가중치는

$$W_i = \frac{1}{p_i(1 - p_i)} \tag{2.5}$$

로 된다. 먼저 첫 번째 범주의 전변동을 구해보면

$$\begin{aligned} \text{첫 번째 범주의 전변동}(S_{T_1}) &= \text{이치자료의 총제곱합} - CF \tag{2.6} \\ &= cum_1 - cum_1^2 / cum \\ &= cum_1 / cum (cum - cum_1) \\ &= cum \cdot cum_1 / cum (1 - cum_1 / cum) \\ &= cum p_I (1 - p_I) \end{aligned}$$

이 된다. 두번째 범주에 대해서도 같은 방법으로 전변동(ST_2)을 구해보면 $cum \times p_{II} (1 - p_{II})$

이 된다. 가령 3범주로 구성된 전변동(S_T)은 첫번째 범주와 두번째 범주의 변동에 각각 가중치 W_1, W_2 를 곱하여 합한 것을 전변동으로 구한다.

$$\begin{aligned} S_T &= cum \times p_I (1 - p_I) \times W_1 + cum \times p_{II} (1 - p_{II}) \times W_2 \\ &= cum + cum \\ &= 2cum \end{aligned} \tag{2.7}$$

즉,

$$S_T = (\text{자료의 총수}) \times (\text{해석하고 있는 범주의 수}) \tag{2.8}$$

그리고 S_A 도 다음과 같이 구하게 된다.

$$S_A = (\text{첫 번째 범주의 } S_A) \times W_1 + (\text{두 번째 범주의 } S_A) \times W_2 \tag{2.9}$$

따라서 누적법에서 가중치 W_i 를 곱하여 각 범주의 전변동을 동일하게 하여, 전변동(S_T)에

대한 각 범주의 기여정도를 같게 하는 것이다[9].

2.2 Nair 검정법

전통적인 실험에서는 분산분석에 의해 목적 특성의 크기에 영향을 주는 인자에 착안하여 최적 조건(망대, 망목, 망소특성을 주는 조건)의 선택과 그 조건하에서 기대되는 목적 특성치를 추정하는 것에 주안점을 두고 왔다. 특히 모수 모델을 전제로 하는 경우, 실험인자의 수준에 의한 모평균의 상이(대소, 위치)만을 문제로 하고, 분산은 수준에 관계없이 일정하다고 가정하고 있다. 그것은 순서분류 척도를 이용한 계수치 실험에 해당된다고 하면, 도수분포의 중심 위치가 실험 인자의 수준에 의해 어떻게 변화되는지의 문제로 되어, 순서 분류에 의한 실험에서도 관심의 대상이 될 수 있다.

어떤 인자가 수준에 의해 이러한 '위치 벗어남'을 발생시키지 않는 효과를 위치 효과(location effect)라 부르지만, 순서 분류의 분할표의 해석에서는 이 위치효과만을 생각하는 것은 불충분한 경우라고 생각된다.

예를들면 [표1]의 경우와 같이 A₁의 수준과 A₂ 수준이 상이하며, A₁의 도수가 3계급의 순서분류로 거의 평균에 광범위하게 산포 되어 있는 것에 비해, A₂에서는 2급품에 집중되어 있다.

[표1] 수준에 따른 품질 분류

수준 \ 품질	1급품	2급품	3급품	계
A ₁	7	7	6	20
A ₂	1	18	1	20

이와 같은 '산포도의 대소'를 발생하는 효과를 위치효과와 구별하여, 산포도효과(dispersion effect)라 부른다. [표1]의 예에서 A₁ 보다 A₂가 좋다고 하면 그것은 위치효과는 아니고 산포도 효과로 판단하게 된다[3].

이와 같이 Nair(1986)는 위치와 산포를 검출해내는 보다 단순화 시키는 누적분석법의 대안으로서 2가지 성분의 분리 통계량인 SS(ℓ), SS(d)사용할 것을 제시 하였다[7].

Nair(1986, 1987)[7.8]는 위치 및 산포도효과의 유의성을 개별로 검출하는 점수법을 제시하고 있다. 위치효과에 대해서는 Kruskal-Wallis 검정[5]과 동일한 방법을 사용, 어떤 범주에 해당하는 관측치는 전부 동순위[4]를 할당한다. 산포도 효과에 관해서는 Ridit·점수의 2차 형식으로서 직교하는 점수를 부여한다.

따라서 비교할 인자의 수준 $i = 1, 2, \dots, I$ 가 분할표의 I 개의 행을, 목적특성의 순서분류 $j = 1, 2, \dots, J$ 가 분할표의 J 개의 열을 구성하며, 도수 x_{ij} 가 $I \times J$ 의 분할표에 교차 분류되어 진다. 도수의 행계, 열계 및 총계를

$$T_{i.} = \sum_{j=1}^J x_{ij} \quad T_{.j} = \sum_{i=1}^I x_{ij} \quad T = \sum_{i=1}^I \sum_{j=1}^J x_{ij} = \sum_{i=1}^I T_{i.} = \sum_{j=1}^J T_{.j}$$

로 한다. Nair에 의한 위치효과와 산포도 효과의 검정은 다음과 같은 순서로 실시된다.

[순서1] 응답(열) 범주에 해당하는 비율

$$q_j = T_{.j} / T, \quad j=1, 2, \dots, J \tag{3.1}$$

을 산출한다.

[순서2] 범주 j의 평균순위

$$\tau_j = \sum_{k=1}^{i-1} q_k + \frac{q_j}{2}, \quad j=1, 2, \dots, J \quad (3.2)$$

을 구한다.

[순서3] 위치 점수

$$l_j = \frac{\tilde{\tau}_j}{\left[\sum_{k=1}^J q_k \tilde{\tau}_k^2 \right]^{\frac{1}{2}}}, \quad j=1, 2, \dots, J \quad (3.3)$$

및 산포도 점수

$$d_j = \frac{e_j}{\left[\sum_{k=1}^J q_k e_k^2 \right]^{\frac{1}{2}}}, \quad j=1, 2, \dots, J \quad (3.4)$$

을 산출한다.¹⁾

식(3.3)과 식(3.4)에서

$$\tilde{\tau}_i = \tau_i - 0.5, \quad e_j = l_j \left(l_j - \sum_{k=1}^J q_k l_k^3 \right) - 1$$

이다. 다만, 위치점수와 산포도 점수와의 사이에는 다음의 관계가 성립된다.

$$\sum_{k=1}^J q_k l_k = \sum_{k=1}^J q_k d_k = \sum_{k=1}^J q_k l_k d_k = 0 \quad \sum_{k=1}^J q_k l_k^2 = \sum_{k=1}^J q_k d_k^2 = 1$$

[순서4] 위치의 기대 관측치

$$L_i = \sum_{j=1}^J l_j x_{ij}, \quad i=1, 2, \dots, I \quad (3.5)$$

및 산포도의 기대 관측치

$$D_i = \sum_{j=1}^J d_j x_{ij}, \quad i=1, 2, \dots, I \quad (3.6)$$

을 구한다.

주1) 비모수 검정에 있어서 위치차가 없는 경우 분산차를 해보는 검정으로서 Mood검정, 즉, 범주 j의 점수로써

$$S_j = \left(j - \frac{I+1}{2} \right)^2$$

을 부여한다. 이것은 j에 관한 2차 형식이 된다.

[순서5] 위치효과의 제곱합

$$SS(D) = \sum_{i=1}^I L_i^2 / T_i \quad (3.7)$$

및 산포도 효과의 제곱합

$$SS(d) = \sum_{i=1}^k D_i^2 / T_i \tag{3.8}$$

을 산출한다²⁾

인자의 위치 및 산포도 효과가 없다는 것을 귀무가설 하에서 $SS(i)$, $SS(d)$ 는 모두 근사적으로 자유도는 (수준수-1)의 χ^2 의 분포에 따른다.

[순서6] $SS(i) > \chi_{\alpha}^2(\alpha)$ 면, 유의수준 α 로 수준간의 위치효과에 차가 있다.

$SS(d) > \chi_{\alpha}^2(\alpha)$ 면, 유의수준 α 로 수준간의 산포도 효과에 차가 있다.

이러한 점수법을 사용하면, 유의한 인자의 최적 수준을 용이하게 결정할 수 있다.

산포도 효과에 대해서, 중앙에 집중하고 있는 정도를 원한다면, 각 수준의 기대관측치를 T_i 로 나눈 양 D_i/T_i 가 최소치를 갖는 수준이 최적이다. 위치효과에 관해서는 순서응답의 최초 범주가 최량이면, 각 수준의 기대관측치의 평균 L_i/T_i 가 최소치를 갖는 수준이 최적이다.

III 비교분석 및 고찰

순서범주형 자료해석을 위한 다구찌의 누적법과 Nair 검정법을 비교분석 및 고찰하기 위해서 식품 가공 실험 사례를 통해 알아보기로 한다. 어떤 식품을 가공할 때에 그 식품에 산재되어 있는 냄새를 없애고 싶다. 그것을 위해 효과가 있다고 생각되는 첨가물 A, B, C(2수준 3인자)를 취해 주효과와 교호작용(A×B, B×C)을 조사 하기 위하여 직교 배열표 L_8 에 할당해서, 8종류의 시료를 만들었다. 얻어진 시료를 5인의 검사원이 없애려는 냄새의 강도에 관해 '약' '중' '강'의 3조로 분류해서 [표2]와 같은 결과를 얻었다[3].

[표2] L_8 에 의한 인자할당과 실험 결과

列番 No.	1	2	3	4	5	6	7	약	중	강
1	1	1	1	1	1	1	1	1	1	3
2	1	1	1	2	2	2	2	0	1	4
3	1	2	2	1	1	2	2	3	1	1
4	1	2	2	2	2	1	1	3	1	1
5	2	1	2	1	2	1	2	1	4	0
6	2	1	2	2	1	2	1	2	1	2
7	2	2	1	1	2	2	1	3	2	0
8	2	2	1	2	1	1	2	1	4	0
	A	B	A×B	C	e_1	B×C	e_1	14	15	11

주2) 누적 χ^2 통계량 χ^{*2} 은 T_i 가 동등한 경우, 위치효과 및 산포도효과에 대응하는 검정통계량을 $\frac{1}{2}$ 및 $\frac{1}{6}$ 으로 가중치를 부여한 성분으로 분해시킨다.

3.1 다구찌의 누적법

[표2]의 자료를 이용하며 다구찌의 누적법을 적용한 분산분석 결과는 [표3]과 같다.

[표3] 분산 분석표

VS	SS	df	ms	F ₀	(F ₀)
A	6.144	2	3.072	3.528	3.626**
B	10.100	2	5.050	5.800	5.961**
C	1.568	2	0.784	0.900	0.926
A×B	2.887	2	1.443	1.658	1.704
B×C	1.568	2	0.784	0.900	
e ₁	2.009	4	0.502	0.577	
e ₂	55.724	64	0.871		
e	59.301	70			
계	80	78			

[표3]의 분산분석 결과로 A,B 인자가 고도로 유의한 것으로 나타났다. 이것은 A,B인자 모두 범세의 강약에 매우 유의적임을 알수 있다.

이와 같이 계수 분류치가 3조 이상으로 분류되어져 있는 경우에 대부분 누적법을 사용하여 분산 분석을 실시하고 있다. 참고로 누적법을 사용하지 않고, 각조에 가중치를 두어 SN비를 구하여 분석하는 방법도 있다.

3.2 Nair 검정법

[표2]의 자료를 이용하여 Nair 검정법을 적용해 보면 우선 인자의 효과의 유의성을 검정하기 위해 [표4]와 같이 인자 A의 주변표를 작성하여 검정을 실시하면 다음과 같다.

[표4] 인자A의 주변표

첨가물A \ 범세의 강도	범세의 강도			계
	약	중	강	
1	7	4	9	20
2	7	11	2	20
계	14	15	11	40

[순서1] 응답범주에 해당하는 비율은

$$q_1 = \frac{14}{40} = 0.35, \quad q_2 = \frac{15}{40} = 0.375, \quad q_3 = \frac{11}{40} = 0.275$$

로 계산된다.

[순서2] 범주j의 평균 순위는

$$\begin{aligned} \tau_1 &= q_1/2 = 0.1750 \\ \tau_2 &= q_1 + q_2/2 = 0.5375 \\ \tau_3 &= q_1 + q_2 + q_3/2 = 0.8625 \end{aligned}$$

로 구해진다.

[순서3] 위치점수 및 산포도 점수

$$\begin{cases} l_1 = \frac{\tilde{\tau}_j}{[\sum_{k=1}^3 q_k \tilde{\tau}_k^2]^{1/2}} \\ = -1.1977 \\ l_2 = 0.1382 \\ l_3 = 1.3359 \end{cases} \quad \begin{cases} d_1 = 0.6501 \\ d_2 = -1.2836 \\ d_3 = 0.9229 \end{cases}$$

을 얻는다.

[순서4] 위치 및 산포도의 기대 관측치는

$$\begin{cases} L_1 = 4.192 \\ L_2 = -4.192 \end{cases} \quad \begin{cases} D_1 = 7.723 \\ D_2 = -7.723 \end{cases}$$

로 계산된다.

[순서5] 위치의 제곱합 및 산포도의 제곱합을 구하기 위해 [표5]와 같이 보조표를 작성하여.

$SS(l) = 1.757$, $SS(d) = 5.964$ 을 얻는다.

[표5] 인자 A에 대한 $SS(l)$ 및 $SS(d)$ 의 계산을 위한 보조

	약 1	중 2	강 3	T_i	L_i	D_i
1	7	4	9	20	4.192	7.723
2	7	11	2	20	-4.192	-7.723
q_j	0.3500	0.3750	0.2750			
τ_j	0.1750	0.5375	0.8625			
$\tilde{\tau}_j$	-0.3250	0.0375	0.3625			
$q_j \tilde{\tau}_j^2$	0.0370	0.0005	0.0361			
l_j	-1.1977	0.1382	1.3359			
e_j	0.5007	-0.9885	0.7108			
d_j	0.6501	-1.2836	0.9229			

$$\rightarrow [\sum_{k=1}^3 q_k \tilde{\tau}_k^2]^{1/2} = 0.2713$$

[순서6] $\chi_1^2(.05) = 3.84$ 로부터

$$SS (l) < x_1^2 (.05)$$

$$SS (d) < x_1^2 (.05)$$

이기 때문에, 인자 A에 관해서는 위치효과는 유의 하지 않고 산포도 효과는 유의하게 한다. 인자 B, A×B, C, B×C에 관해서 같은 방법으로 계산할 결과는 [표6]과 같다.

[표6] Nair 검정의 결과

인 자	위치효과	산포도효과	자유도	
A	1.757	5.964*	1	$x_1^2 (.05) = 3.84$
B	6.723**	0.369	1	
A×B	1.997	0.031	1	$x_1^2 (.01) = 6.63$
C	0.981	0.189	1	
B×C	0.021	2.749	1	

[표6]의 Nair검정결과를 요약하면 산포에 관해서는 A의 효과, 위치에 관해서는 B의 효과가 있다. [표6]에 나타난 결과에서와 같이 유의한 인자 A, B에 관해서 산포도 및 위치효과에 대한 기대 관측치의 평균을 계산한 결과는 [표7]과 같다.

[표7] 인자 A, B에 관한 기대관측치의 평균

水準	D/T _i .	水準	L/T _i .
A ₁	1.545	B ₁	1.640
A ₂	-1.545	B ₂	-1.640

[표6]에서와 같이 인자 B는 냄새의 강약에 관계가 있고, B의 2 수준은 1 수준보다 냄새를 평균적으로 약하게 하는 효과를 갖는다.

반면 인자 A는 냄새에 관한 도수분포의 산포도에 관계한다. 냄새의 정도의 도수는 A의 제 1 수준에서 7, 4, 9로 넓게 산포되어 있는 반면 A의 2수준에서는 도수가 7, 11, 2로 '중'에 집중되어 있어 산포도는 작다. 상기 예의 두 분석법의 결과에서 알수 있듯이 Hamada와 Wu는 다구찌의 누적법의 장점은 단순성과 ANOVA와의 유사성이며, 제곱합의 독립성상실로 인한 허위 인자를 검출하는 잠재력을 갖는 결점이 있다고 말한다[2].

IV 결 론

본 연구의 실험 결과에서도 알수 있듯이, 다구찌의 누적법은 산포효과를 검출하는데는 Nair 검정법에 비해 뒤떨어짐을 알 수 있다. 이러한 다구찌의 산포효과를 검출하지 못하는 결함을 제거하기 위해서는 다구찌의 누적 분석 검정 통계량을 2가지 성분으로 분해하여 위치 및 산포 효과를 검정 해야하는데 그 계산 과정에서 고유치를 계산 해야 하는 불편함이 발생한다. 그래서 위치 및 산포효과로 분해된 다구찌의 누적분석 통계량과 매우 근접한 통계량을 갖고 쉽게 계산 할 수 있는 Nair의 점수법을 제시하여 사례에 적용 해본 결과, 다구찌의 누적 분석에서는 A와 B인자가 고도로 유의하였고, Nair 검정에서는 산포도 효과는 A인자, 위치 효과는 B인자

가 유의 함을 보여 주었다. 그러나 순서 범주형자료의 해석방법에 관해 위치효과와 산포효과를 동시에 검출할 수 있는 대안으로서 Nair검정법 이외에 McCullagh에 의한 누적로지트모델과 Haber 등에 의한 평균반응모델 등이 있다. 이와 같이 실무자에 의해 유용하게 사용되어질수 있는 단순하면서도 합리적인 방법들이 계속해서 개발되어 나와 그 유효성에 관해 비교분석을 실시함으로써 보다 주어진 실험에 대한 정확한 정보를 얻을 수 있다.

참 고 문 헌

1. 박성현, (1990), 응용실험계획법, 영지문화사, pp.57~66, pp.242~246.
2. M. Hamada and C. F. J. Wu (1990) : "A Critical Look at Accumulation Analysis and Related Methods", *Technometrics*, VOL. 32, NO2, pp.119~130.
3. 辻谷將明, 楠正, 松本哲天, 和田武夫 (1992) : "計數値の 解析" 品質管理, VOL. 43, NO5, pp.66~71
4. Bross, I. D. J.(1958) : "How to use Ridit Analysis", *Biometrics*, , 14, pp.18~38.
5. Lehmann, E.L. (1975) : Nonparametrics : Statistical Methods Based on Ranks, San Francisco, Holden-Day, pp 303~311
6. Mood, A. M (1954) : "on the asymptotic efficiency of certain Nonparametric Two sample Test", *Annals of Mathematical Statistics*, 25, pp.514~522
7. Nair, V. N. (1986) : "Testing in Industrial Experiments with Ordered Categorical Data", *Technometrics*, 28, pp.283~311.
8. Nair, V. N. (1987) : "Chi - Squared - Type Tests for Ordered Alternatives in Contingency Tables", *Journal of the American Statistical Association*, 82, pp.283~291 .
9. Genichi Taguchi (1987), System of Experimental Design, Kraus International Publications, New York, pp 78 ~ 115