工業經營學會誌 第 20卷 第 42輯 1997年 5月 1

경험로지트변환과 Freeman-Tukey형 역정현 변환에 의한 계수치 자료의 해석

- Analysis of binary data by empirical logit transformation and the type of Freeman-Tukey inverse sine transformation -

김 홍준 *
Kim Hong Jun
채 규용 **
Chae Gyoo Yong
이 상용 ***
Yi Sang Yong

Abstract

In case of analysis of discrete data, it shows by way of example orthogonal array experiment for o,1 data. This paper introduced expirical logit transformation and the type of Freeman-Tukey inverse sine transformation. As the result of analysis of variance, empirical logit transformation turned out a mistake in application but it is possible for graphical analysis by normal probability paper.

Ⅰ. 서론

1.1 0,1 자료의 해석법

불량률 P의 모집단에서 크기가 n인 랜덤 샘플 Y_1 Y_2 …, Y_n을 취한 경우

* 대구산업전문대학 산업안전과
** 건국대학교 산업공학과 박사과정
*** 건국대학교 산업공학과

$$Y_i = \begin{cases} 0: \text{양품} \\ 1: \text{불량품} \end{cases} \tag{1.1}$$

으로 표시한다. 이 경우, n개의 샘플 중에 포함되어 있는 불량품의 개수 $Y = Y_1 + Y_2 + \cdots + Y_n$ 을 이항분포에 따른다고 가정한다. 먼저 0,1 자료의 해석에 사용되고 있는 종래의 일반적인 방법을 정리해보면 다음과 같다.

(1) (0,1)법에 의한 계량치 해석

양품을 0, 불량품을 1의 계량치를 보고 분산분석을 실시한다. 이방법의 결점은 분산 p(1-p)가 p에 의존해서 변함에도 불구하고, 분산을 일정하다고 보는 것이다. p의 범위가 0.2 ~0.8이면 분산의 변화는 작지만, p가 0혹은 1에 근접하면 크게변화한다. 이 점을 고려하여 분산을 일정하게 하는 것이 역정현 변환이다[6].

(2)역정현 변환에 의한 불량률 해석

$p = Y/n$에 역정현 변환 $\theta = \sin^{-1}\sqrt{P}$ 를 적용한다. n이 충분히 큰 경우(실용적으로는 10이상), θ은 근사적으로 평균 $E[\theta] = \sin^{-1}\sqrt{p}$, 분산 $Var[\theta] = 1/(4n)$에 따르는 것을 이용한다. 따라서 분산은 p에 관계없이 항상 일정하게 된다.

(3)로지트변환에 의한 불량률 해석

불량률이 크게 산포하는 경우에는, 표본 불량률 p에 로지트변환(다구찌의 오메가 변환 이라고도 한다).

$$\theta = \log\{p/(1-p)\} \tag{1.2}$$

을 적용한다. 이항분포의 평균과 분산에 있어 n=1인 경우를 고려하여, 이항분포모집단에서의 $Y_1, Y_2, \cdots Y_n$을 $N(P, P(1-P))$에서의 랜덤 샘플로 간주한다. 평균 $E[Y_i] = P$ 와 분산 $Var[Y_i] = P(1-P)$와의 비는

$$\frac{(\text{평균})^2}{\text{분산}} = \frac{p^2}{p(1-p)} = \frac{p}{1-p} \tag{1.3}$$

로 된다. (1.2)와 (1.3)식에서

$$10\log\left(\frac{p}{1-p}\right) = 10\log\left\{\frac{\text{평균}}{\text{분산}}\right\} \tag{1.4}$$

의 관계가 성립한다[2]. (1.2)식의 로지트변환을 실시하면 가법성이 높아지고, 역정현과 비교하면 불량률이 크게 산포한 경우에도 적용할 수가 있다.

한편 다구찌 방법에서는 Y_i 의 값이 작은 것이 좋으므로 망소특성으로 생각하고, SN공식을 사용하여

$SN = -10\log\left[\frac{1}{n}\sum_{i=1}^{n} Y_i^2\right]$ 을 최대로 하는 조건을 찾는 것과 일치하게 된다.

$$SN = -10\log\left[\frac{1}{n}\sum_{i=1}^{n} Y_i^2\right] \quad = 10\log\left[\frac{\sum Y_i}{n}\right] \quad = -10\log(p) = 10\log\left(\frac{1}{p}\right)$$

이 된다. 따라서 p는 시료의 불합격률로서 이 p값이 작으면 작을수록 SN 값은 커지게 된다. 그리고

$$Y_i = \begin{cases} 0: \text{불량품} \\ 1: \text{양품} \end{cases} \tag{1.1}$$

으로 표시하는 0-1 자료에 대해서는 SN비를 (1.6)식과 같이 정의하여 사용한다.

$$
\begin{aligned}
SN &= 10\log\left[\frac{\frac{1}{n}S_m}{V}\right] \\
&= 10\log\left[\frac{S_m}{nV}\right] = 10\log\left[\frac{S_m}{S_e}\right] \\
&= 10\log\left[\frac{np^2}{np(1-p)}\right] = 10\log\left[\frac{p}{1-p}\right] \\
&= -10\log\left[\frac{1}{p}-1\right]
\end{aligned}
\tag{1.6}
$$

(1.7)식은 (1.4)식과 같은 동일한 내용의 오메가 변환으로 된다. 이 경우 시료 합격률 p가 증가할수록 SN값은 커지며, SN값이 클수록 좋다[1].

종래 계수치 해석은 분할표에 의한 독립성 검정과 로지스틱 회귀분석의 회귀계수의 유의성 검정 등 유의수준을 설정한 검정에 주안점을 두어왔다. 본 논문에서는 그래프 분석을 용이하게 할 수 있어 가시적으로 쉽게 이해할 수 있는 경험 로지트 변환과 종래의 역정현 변환에서 n이 일정한 경우를 고려한 Freeman- Tukey 형 역정현 변환을 제시하여 이 3가지 분석을 비교 하고자 한다.

II. 경험 로지트 변환과 Freeman-Tukey형 역정현 변환

2.1 0,1 자료의 직교배열 실험

어떤 회사의 용접 공정에서 용접불량의 재작업률을 개선시키기 위해서 요인 A(처리시간), 요인 B(처리전류), 요인 C(전류제어 패턴)모두 2 수준으로하여 L_8 직교표에 의한 실험의 결과이다. 교호작용으로서는, A×B, A×C, B×C를 고려하였다. 기술적으로는, 주효과 C 및 교호작용효과 A×B, B×C 유의차는 적다고 생각되지만 그것을 확인하기 위하여 직교표에 할당해서 얻어진 것이 표 1과 같다[5].

표 1. 요인의 할당과 자료

요인열 / No	A	B	A×B	C	A×C	B×C	e	불량개수	생산대수
	1	2	3	4	5	6	7		
1	1	1	1	1	1	1	1	11	370
2	1	1	1	-1	-1	-1	-1	12	384
3	1	-1	-1	1	1	-1	-1	1	380
4	1	-1	-1	-1	-1	1	1	1	382
5	-1	1	-1	1	-1	1	-1	14	376
6	-1	1	-1	-1	1	-1	1	16	370
7	-1	-1	1	1	-1	-1	1	0	376
8	-1	-1	1	-1	1	1	-1	2	370

2.2 경험로지트 변환

모불량률 P_i 인 모집단에서 크기 n인 샘플을 추출할 때, 그 중에 포함되어 있는 불량품의 개수를 Y_i 라한다.

Y_i는 0,1,2,⋯, n의 값을 취한다. Y_i가 이항분포에 따른다면 그 평균과 분산은

$$\begin{cases} E[Y_i] = nP_i \\ Var[Y_i] = nP_i(1-P_i) \end{cases} \tag{2.1}$$

이다. 표1에서, 실험번호 i 에서 생산대수 n_i 중에 Y_i개의 불량이 포함되어 있다. 이 경우(1.4)식에 대응하는 로짓은

$$\log\left(\frac{p_i}{1-p_i}\right) = \log\left(\frac{Y_i}{n_i - Y_i}\right) \tag{2.2}$$

로 된다. 그러나 표1의 실험번호 7과 표3의 실험번호 2에서는 불량개수 $Y_7=0$, $Y_2=0$이기 때문에 (2.2)식의 값을 구하지 못한다. 그래서 (2.2)식의 분모, 분자에 0.5을 더하면

$$\theta_i = \ln\left(\frac{Y_i + 0.5}{n_i - Y_i + 0.5}\right) \tag{2.3}$$

으로 된다[3][4]. 이것을 '경험 로지트 변환'이라 부른다. 표 2와 4에 (2.3)식 값을 나타내고 있다.

표2 로지트 변환치

요인 / 열 / No	A	B	A×B	C	A×C	B×C	e	θ_i
	1	2	3	4	5	6	7	
1	1	1	1	1	1	1	1	-3.442
2	1	1	1	-1	-1	-1	-1	-3.395
3	1	-1	-1	1	1	-1	-1	-5.533
4	1	-1	-1	-1	-1	1	1	-5.539
5	-1	1	-1	1	-1	1	-1	-3.219
6	-1	1	-1	-1	1	-1	1	-3.067
7	-1	-1	1	1	-1	-1	1	-6.624
8	-1	-1	1	-1	1	1	-1	-4.993

θ_i값 자체는 통계량으로써 표 2와 표 4를 L_8 직교표 자료로써 구하는 간편법에서의 자료 구조식은 각각 다음과 같다.

(1) $\theta_i = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\beta\gamma)_{jk} + (\alpha\gamma)_{ik} + \varepsilon$ (2.4)

(2) $\theta_i = \mu + a + b + c + d + e + (ab) + \varepsilon$

(2.4)식의 (1)에서 μ는 일반평균이며, $\alpha_i, \beta_j, \gamma_k$는 각각 요인 A,B,C의 주효과, 나머지 항들은 차례로 A×B 교호작요효과, B×C 교호작용효과, A×C 교호작용효과이다.
오차항 ε는

$$\varepsilon \sim N(0, \sigma_e^2) \tag{2.5}$$

로 가정한다. 이 경우 요인 A의 평방합 S_A 는

$$S_A = \left(\frac{A_1\text{에서의자료의합} - A_2\text{에서의자료의합}}{\text{자료총수}} \right) \tag{2.6}$$

이고 자유도는 1이 된다.
그리고 표3은 샘플의 크기는 일정하지만 극히작고 ($n=6$), 불량률이 크게 산포하는 경우의 L_8 직교표에 할당한 실험의 예이다.

표3 할당과 자료

열번 / No	1 A	2 B	3 A×B	4 C	5 D	6 E	7 e	불량개수 Y_i ($n=6$)
1	1	1	1	1	1	1	1	4
2	1	1	1	-1	-1	-1	-1	0
3	1	-1	-1	1	1	-1	-1	0
4	1	-1	-1	-1	-1	1	1	1
5	-1	1	-1	1	-1	1	-1	6
6	-1	1	-1	-1	1	-1	1	1
7	-1	-1	1	1	-1	-1	1	0
8	-1	-1	1	-1	1	1	-1	1

2.3 Freeman-Tukey형 역정현 변환

샘플불량률 $P_i = Y_i/n$에 있어 역정현 변환

$$\eta_i = \sin^{-1}\sqrt{Y_i/n} \tag{2.7}$$

을 실시할 때 n이 충분히 클 때 η_i는 근사적으로 평균 $E[\eta_i] = \sin^{-1}\sqrt{P_i}$, 분산 $Var[\eta_i] = 1/4(n)$ 의 정규분포에 따른다. (이때 단위는 rad으로 표시된다.)
따라서 분산은 p에 관계없이 샘플의 크기 n에 의해 변화 하지만 n이 일정한 경우, 분산이 안정하게 하는 것을 이용하는 것이 역정현 변환법이다. 경험로지트 변환과 같은 방법으로 (2.7)을 구하기 위해 Y_i 와 n에 관해 수정한

$$\eta_i = \{\sin^{-1}\sqrt{Y_i/(n+1)} + \sin^{-1}\sqrt{(Y_i+1)/(n+1)}\}/2 \tag{2.8}$$

을 이용한다.
$\eta_i^* \sim N(\sin^{-1}\sqrt{P_i}, 1/\{4(n+1)\}$로 된다. 이 변환은 'Freeman -Tukey형 역정현 변환'이라고 부른다.[7]

표4 변환치의 보조표

열번 / No	1	2	3	4	5	6	7	불량개수 Y_i $(n=6)$	불량률 P_i	경험로지트 변환 θ_i	종래의 역정현 변환		Freeman-Tukey 형
	A	B	A×B	C	D	E	e				η_i	η_i^2	η_i^*
1	1	1	1	1	1	1	1	4	0.667	0.588	0.950	0.904	0.927
2	1	1	1	-1	-1	-1	-1	0	0	-2.565	0	0	0.194
3	1	-1	-1	1	1	-1	-1	0	0	-2.565	0	0	0.194
4	1	-1	-1	-1	-1	1	1	1	0.167	-1.299	0.421	0.177	0.476
5	-1	1	-1	1	-1	1	-1	6	1	2.565	1.317	1.734	1.234
6	-1	1	-1	-1	1	-1	1	1	0.167	-1.299	0.421	0.177	0.476
7	-1	-1	1	1	-1	-1	1	0	0	-2.565	0	0	0.194
8	-1	-1	1	-1	1	1	-1	1	0.167	-1.299	0.421	0.177	0.476

2.4 자료해석

표3의 자료해석을 다음과 같이 경험로지트 변환, 종래의 역정현 변환, Freeman-Tukey형 역정현변환의 3가지를 제시하고 정리한 것이 표4이다.

$$\text{계산예}\begin{cases} ①\theta_1 = \ln\{(4+0.5)/(6-4+0.5)\} = 0.588 \\ ②\eta_1 = \sin^{-1}\sqrt{4/6} = 0.950 \\ ③\eta_1^* = \{\sin^{-1}\sqrt{4/7} + \sin^{-1}\sqrt{5/7}\}/2 = 0.927 \end{cases}$$

②의 경우에 있어, 일반적인 분산분석을 단계별로 실시하면 다음과 같다.(나머지 경우도 동일한 순서로 행한다.)

(1) 순서 1 : 수정항 CT와 총제곱합 S의 계산

$$CT = \left(\sum_{i=1}^{8}\eta_i\right)^2/8 = 3.53^2/8 = 1.56$$

$$S = \sum\eta_i^2 - CT = 3.169 - 1.56 = 1.61$$

(2) 순서 2 : 각 열에 있어 수준 2와 수준1에 대하여 η_i의 합을 구하여 각 제곱합을

$$\frac{[\,(\text{수준2의 }\eta_i\text{의합}) - (\text{수준1의 }\eta_i\text{의합})]^2}{8}$$

으로 구해보면 $S_A=0.0776, S_B=0.4260, S_C=0.1260, S_D=0.0036, S_E=0.9032, S_{A\times B}=0.0776$

(3) 순서 3 : 오차 평균 제곱의 계산

오차 평균 제곱$= 1/(4n) = 0.0417$

(4) 순서 4 : 분산분석표 작성

상기와 같이 작성된 3가지 변환의 분산분석표는 표 5와 같다.

표 5 분산분석표

	경험로지트변환			종래의 역정현 변환			Freeman-Tukey 형 역정현변환		
요인	자유도	평균제곱	F_0	자유도	평균제곱	F_0	자유도	평균제곱	F_0
A	1	1.314	20.80	1	0.0776	1.86	1	0.0434	1.22
B	1	6.155	97.40	1	0.4260	10.22**	1	0.2779	7.78**
C	1	2.514	39.79	1	0.1260	3.02	1	0.1074	3.01
D	1	0.063	1.00	1	0.0036	<1	1	0.0008	<1
E	1	11.398	180.37*	1	0.9032	21.66**	1	0.5279	14.79**
A×B	1	1.315	20.80	1	0.0776	1.86	1	0.0434	1.22
오차	1	0.063	-	∞	0.0417 (1/24)	-	∞	0.0357 (1/28)	-
F=(1,1;0.05)=161.0				F=(1,∞;0.01)=6.63 ; F(1,∞ ; 0.05) =3.84					

Ⅲ. 결론

계수치 자료의 해석에 관해 (0,1)법은 이항분포의 정규근사를 이용하지만, 샘플크기 n과 모불량률 P에 관해, $nP \geq 5$ 또는 $n(1-P) \geq 5$인 경우만 그 근사가 성립한다. 요인실험에 있어 분할표에 의한 해석을 할 수 있지만 요인이 한 개이면 유효하지만 2개이상이 되면 보다 정확한 해석이 필요해진다. 로지트변환의 경우 불량률이 0.2~0.8범위이면 가법성은 높아진다. 그러나 불량률이 크게 산포되면 적용하는 것이 문제이기 때문에 경험 로지트변환에 의한 그래프 해석을 적용할 수가 있다. 그런데 그래프 해석은 요인효과의 정확한 유의성 검정은 될수없지만, 탐색적인 자료로서 효과의 유의성을 대략적으로 얻을수 있는 간편한 방법이 될 수있다. 즉 경험로지트변환을 적용하면 실무에서 대비를 이용한 비교적 용이하게 활용될 수 있는 정규 확률지에 의한 그래프 해석을 실시함으로써 ① 요인효과의 가법성 ②표준화대비의 정규성을 활용할 수 있다.

표5에서 알수있듯이 이항분포의 정규근사를 좋게 하기 위해서 사용되어지는 종래의 역정현 변환 및 Freeman-Tukey형의 역정현 변환을 실시한 경우 요인 B와 E가 고도로 유의하다. 그러나 경험로지트변환의 경우 요인 E만이 유의하다. 이것은 표 4의 샘플 불량률 P_i를 보면 P_i가 0에서 1까지 크게 산포해있어 경험로지트변환치를 계량치(분산일정)로 해석하면 잘못된 결과를 유도하고 있음을 말해준다. 그러나 본 논문에서는 일반실무에서 경험로지트 변환, 역정현 변환 및 Freeman-Tukey형 역정현 변환의 편리성과 사용방법 즉 실용성 측면에서 명확하게 제시하고 있지 않다. 따라서 요인의 유의성 검정을 위해 시뮬레이션에 의한 성능 평가의 문제와 3가지 변환방법별 그 실용성 규명을 위해 샘플의 크기를 고려한 유효성 평가가 해결해야 할 과제이다.

참 고 문 헌

[1] 박성현, 응용실험계획법, 영지문화사 , pp. 237~239 1990.

[2] 田口玄一, 품질설계를 위한 실험계획법, 일본규격협회, 1988.

[3] Cox D.R. & Lauh,E.,"A note on the graphical analysis of multidimensional contingency tables ," *Technometrics*, 9 , pp. 481~488 1967.

[4] Cox D.R. & Snell, E.J., *The Analysis of binary data* , 2nd edition, chapman & Hall, 1989.

[5] 辻俗將明 & 天坂格郎 , "계수치 자료의 해석," 품질관리, Vol. 44, NO. 4, pp. 61-63, 1993.

[6] 竹內啓 & 藤野和建, 이항분포와 포아송분포, 동경대학출판회, 1981.