韓國數學敎育學會誌 시리즈 E
<數學敎育 프로시딩>
1997. 7. 제6집, 159-171.

Journal of the Korea Society of Mathematical Education
Proceedings of Mathematical Education
Series E: July 1997, Vol. 6, 159-171.

이원화 평가자료의 최적평가치 산정과 평가원의 신뢰도에 관한 연구

홍 석 강 (동국대학교)

Ⅰ. 서 론

학습평가의 시행과정에서 좋은 평가도구인 질높은 시험문제 또는 학력검사를 제작하고져 할 때 미리 설정된 평가목표에 부합되는 평가문항을 선정하기 위해서는 먼저 학습내용을 평가목표의 교육목표와 행동영역에서의 3가지 영역 즉 지식, 이해력, 적응력, 분석력, 종합력, 평가력 등을 평가하는 인지적 영역과 감수, 반응, 가치화, 조직화, 인격화 등을 평가하는 정의적 영역, 또 표현력, 교육기구의 조작능력, 신체적 운용 능력 등을 평가하는 운동기능적 영역으로 분류체계화한 이원 목표 분류표와 행동평가표를 작성해야하고 그 분류표에 입각하여 작성된 검사들은 그것이 좋은 검사들이기 위해서는 반드시 문항의 내용면에서 타당도, 신뢰도, 객관도, 문항곤란도, 문항변별도 및 문항 반응 분포의 분석 등이 고려되어져야 한다. 일반적으로 그런 평가 문항들은 대개 진위형, 배합형, 선다형의 문항으로 이루어진 선택형 문항과 단답형, 완성형, 논문형 등의 정답을 써넣도록 하는 서답형 문항 등으로 분류되는데 그 가운데서 특히 진위형 문항이나 합격, 불합격의 판정, 찬성, 반대 또는 만족, 불만족 등의 의사표시와 같이 양분화된 사상중 어느 한편의 사상을 선택하고자 할 때 일어나는 시행의 결과를 이원화 평가자료(Dichotomous Rating Variable)라 하며 그것은 필답고사의 문항반응자료, 면접시험문항의 점수, 적성검사 점수의 표현등과 같이 모든 교육 평가 영역에서 자주 다루게 되는 자료들이다. 다음에는 평가에 임하는 평가원들의 평가에 대한 신뢰도에 차이가 있는 경우를 고려한 즉 그들이 주관적인 판단 기준에 따라 각각 다른 판정을 내릴 경우 그 예로써 합격점이하인 수험자들을 합격자로 선정한다던지 또는 막연히 어느 수준 이하, 수준 이상 또는 절반이상, 절반이하 등 모호한 판정이나 아니면 어느 편으로도 정하기 어려운 판정경계선(Border Line)에 있는 수험자들을 합격, 불합격 등으로 임의로 판정을 내리는 경우처럼 서로 다른 판정을 내릴 수 있으므로 각 평가원

간의 평가에 관한 신뢰도 계수를 정확히 측정하고 또 그 신뢰도 계수 차이의 크기도 최소화시킴으로써 평가에 대한 타당성과 신뢰성을 확립하는 방안을 강구해야 한다. 그렇게 함으로써 평가를 받는 수험자들 편에서는 그들의 참된 능력(True Ability)이 반영되게 하며 또 훌륭한 수험자를 선발하고자 하는 평가원들은 그들이 세운 최종 목표에 부합되는 최적의 평가를 시행할 수 있을 것이다. 결론적으로 이 논문에서는 그 이원화 평가자료를 이용한 평가 모델에서 여러 학자들이 논한 난해한 여러 경우의 결과들을 이용하는 것보다 조금 간편하고 쉽게 표현된 자료의 적합법을 다루고 그 모수들을 추정하여 그 자료의 최적평가치와 신뢰도계수를 계산하는 과정을 제시하고자하며 앞으로 교육부에서 시행예정인 유능한 교사선발을 위한 방안으로써 교원순위고사나 사립학교 교원선발고사에 합격된 예비교사들을 교육현장에 임하는 그들의 교육자적 자질과 전공분야의 능력을 테스트할 현장 평가제가 시행될 것을 예상하여 선진국에서 시행한 평가자료를 이용하여 그 자료처리과정을 예로 제시함으로써 우리 평가원들이 훌륭한 교사를 선발하는데 조금이나마 도움이 되는 연구가 되었으면 한다.

Ⅱ. 연구의 내용

Ⅱ-1. 대수선형모형(Log Linear Model)과 신뢰도 계수

어느 평가 기관에서 시행한 검사 문항이 이원화 평가자료로 표현된 표본이라 하고 평가 반응표에 제시된 시행의 결과를 x_{ij}로 표시하면 단, $i=1,\cdots,I$, 평가원수, $j=1,\cdots J$, 수험자수 x_{ij}는 $x_{ij}=1$은 합격 $x_{ij}=0$는 불합격으로 표현된다[1]. 이 때 각 평가원의 판정기준에 따라 수험자의 합격기회는 다음과 같은 고사점수(Test Score)의 함수로 표현될 수 있다.

$$p(x_{ij}=1|x_i;a,b)=g(a+bx_i) \qquad \cdots\cdots \quad (1)$$

단, a,b는 평가원의 판정계수이고 g는 모든 평가원의 판정기준에 적용되는 증가함수로서 연결함수(Link Function)로 정의함.

여기서 기울기 b를 판별계수(Discrimination Coefficient), 절편 a를 판정기준치

1] 합격점수 1 점내를 소문항으로 세분화시키고 결측치가 있는 경우에 관한 연구는 본인의 연구인 홍석강(1994)에서 논한바 있으며 그 경우는 필답고사나 면접고사 문항 또는 인성적성검사 등 여러 분야에서 활용도가 많은 편이지만 이 연구에서는 총합의 점수를 기준으로 하여 연구를 전개한다.

(Least Discrimination Value)라 하며 만일 b의 값이 크면 합격점수내의 수험자수가 많고 그 값이 작으면 그 수가 작을 것이며 현실적으로 $a=0$인 점은 평가 과정에서 거의 일어나지 않을 것으로 이해할 수 있다. 또 여기서 논하는 이원화 자료는 $(-\infty,+\infty)$ 상에서 정의된 실수치를 $(0,1)$인 치역으로의 함수화시킨 실변수 함수 g를 정의한 것이고 추정계수 a와 b로써 적합시킨 $g(\hat{a}+\hat{b}x)$도 역시 개구간 $(0,1)$ 상에서의 확률 측도로 표시되어지므로 그 확률측도의 크기들은 확률측도의 성질인 유계성과 비가법성인 성질들을 만족함을 전제로 하고 있다. 일반적으로 이원화 자료의 최적 평가치를 산정하는 방법에 관한 연구에는 Subkoviak, M.(1978)의 평균회귀 계수법(Method of Average Regression Parameter)과 McCullagh, P.와Nelder, J.A.(1996)의 일반 선형모형(Generalized Linear Model)에의한 모수추정법, Holland P.W.와 Thayer, D.T.(1993) 등이 논한 Random Coefficient 법을 이용한 적합법 등이 있으나 본 연구에서는 위의 학자들의 연구 견해와 조금 다르게 1차고사와 2차고사의 평가자료를 상호 독립인 확률변수로 간주하여 그 경우에 모수를 추정하는데 자주 이용되는 대수선형모형(Log Linear Model)으로써 판별계수를 추정하고 그 처리 과정들을 예로써 제시하고자 한다.

정의 1. 한 수험자가 평가원들이 제시한 모든 영역의 평가문항에 최종적으로 합격 또는 불합격될 x_{ij}, $i=1,\cdots,I$, $j=1,\cdots,J$는 그것의 주변밀도 x_i에 대하여 x_i를 i로 표시하면 $\log m_i=a+bi$. 단, a와 b는 미지의 계수. 를 가지며 독립 등등으로 분포하는 확률변수이다.

정의 2. 판별계수의 최우추정법

대수 선형모형의 최우추정치 $\hat{a}$ 와 $\hat{b}$는 다음과 같이 계산되어진다.

$$\sum_i m_i=\sum_i n_i=N,\quad \sum_i im_i=\sum in_i=N\bar{x}$$

표 1.

수험자 / 평가원	1	2	3	…	j	…	J	합계
1	X_{11}	X_{12}	X_{13}	…	X_{1j}	…	X_{1J}	N_1
2	X_{21}	X_{22}	X_{23}	…	X_{2j}	…	X_{2J}	N_2
3	X_{31}	X_{32}	X_{33}	…	X_{3j}	…	X_{3J}	N_3
⋮	⋮	⋮	⋮		⋮		⋮	⋮
i	X_{i1}	X_{i2}	X_{i3}	…	X_{ij}	…	X_{iJ}	N_i
⋮	⋮	⋮	⋮		⋮		⋮	⋮
I	X_{I1}	X_{I2}	X_{I3}	…	X_{Ij}	…	X_{IJ}	N_I
합계	M_1	M_2	M_3	…	M_j	…	M_J	N

단, $\widehat{m_i} = N\widehat{p_i} = \exp(\hat{a} + \hat{b}i)$

$$\theta_0{}^{2)} = \frac{\sum_i i m_{i_0}}{\sum_i m_{i_0}} \qquad \cdots\cdots \quad (2)$$

$$b_0 = \frac{\sum_i (i-\theta_0) y_{i_0} m_{i_0}}{\sum_i (i-\theta_0)^2 m_{i_0}} \qquad \cdots\cdots \quad (3)$$

$$a_0 = \log\ g_0, \qquad \cdots\cdots \quad (4)$$

$$g_0 = \frac{N}{\sum_i \exp(b_0 i)} = \exp(a_0)$$

이며, 위의 계수들은 뉴우튼 랩슨(Newton Raphson)의 반복과정을 시행하여 수렴시킨 값으로 최종해를 구한다. 다음에는 평가원들의 신뢰도에 관한 신뢰도 계수와 그 차이의 크기를 측정하는 연구로써 Maxwell, A.E.와 Pillinger, A.E.G(1968) 등이 연구한 2원 및 3원의 배치표에서 분산분석법에 의하여 구한 신뢰도 계수 r^*와 Fleiss, J.L.(1971)과 Cogner, A.J.(1980)가 제시한 다중 Kappa 통계량(Multiple Kappa Statistic) $\overline{K_m}$와 K_m'를 Gordon, R.(1984)은 2원화 자료의 경우 그 배치표에서 주변분포가 동일하면 $r^* = \overline{K_m} = r_u = K_m'$이고 주변분포가 다르면 $r_u > K_m'$, $r_u > r^* = \overline{K_m}$ 임을 증명하였다. 여기서는 신뢰도 계수의 크기에서 가장 안정적인 Gordon, R.의 신뢰

2] $_0$ 는 뉴우튼랩슨(Newton Raphson)의 반복식에 대입하는 초기치를 표시하는 점수 기호임.

도 계수를 이용하고 다음과 같이 수험자군과 평가원군의 배치모형을 가정한다.

정의 3. $x_{ij}=\pi_i+\alpha_i+n_{ij}$인 배치모형에서 평가원군인 행간 평방합을 SS_b, 행내 평방합을 SS_w, 수험자군간의 열간 평방합을 SS_r, 오차 평방합을 SS_e이라 할 때 각각의 평균자승은

$$MS_b=\frac{SS_b}{J-1},\quad MS_w=\frac{SS_w}{J(I-1)}$$

$$MS_r=\frac{SS_r}{I-1},\quad MS_e=\frac{SS_e}{(I-1)(J-1)}$$

이다.

그러면 Fleiss의 신뢰도계수 r^*와 Gordon의 계수 $\overline{K_m}$는

$$r^*=\overline{K_m}=\frac{SS_b-\frac{SS_w}{I-1}}{SS_b+SS_w}$$

이고

Maxwell과 Pillinger의 신뢰도 계수는

$$r_u=\frac{SS_b-\frac{SS_e}{I-1}}{SS_b+SS_e}$$

Cogner의 신뢰도 계수는

$$K_m'=\frac{(I-1)SS_b-SS_e}{(I-1)SS_b+(I-1)SS_e+ISS_r}$$

로 표현되고 특히 j가 충분히 크면

$$r^*=\overline{K_m}=\frac{MS_b-MS_w}{MS_b+(I-1)MS_w}\cdots(5)$$

인데, 이 가운데 r^*의 값이 가장 안정적이므로 이것으로써 평가원의 신뢰도 크기를 측정하기로 한다.

Ⅱ-2. 계산 예 및 검토

표 2는 Longford, N.T.(1995)[3]에 수록된 미국 예비교사 훈련 프로그램의 평가자료로

고사지가, 독해, 논술, 수학의 3부로 구성되어 있어 그중 논술고사 부분에서 평가원이 평가한 2원화평가 자료와 수험자들의 독해와 수학성적을 표기한 것이다.

표 2. PPST(Pre-Professional Skills Test) 1차고사 성적과 논술고사에 대한 평가자료.

	수험자																	
평가원	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	합격점수
1차고사 성 적	8	4	5	11	8	10	6	9	7	12	11	10	6	9	8	6	7	
1	1	0	0	0	1	1	1	1	1	1	1	0	1	1	1	1	0	12
2	1	1	0	1	1	1	1	0	1	1	1	1	1	1	1	1	1	15
3	1	0	0	1	1	1	0	0	1	1	1	1	1	0	0	0	0	9
4	1	0	0	1	1	1	1	0	1	1	1	0	0	1	1	0	1	11
5	1	0	0	1	1	1	0	1	1	1	1	1	1	1	1	1	0	13
6	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	16
7	1	0	0	1	1	1	0	1	1	1	1	1	1	1	1	1	0	13
8	0	0	0	0	1	1	0	0	0	1	1	1	1	1	0	1	1	9
9	1	0	0	1	1	1	0	1	1	1	1	1	1	1	1	1	0	13
10	1	0	0	1	0	1	0	1	0	1	0	1	1	1	1	0	0	9
11	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	15
12	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	15
13	1	0	1	1	0	1	0	0	0	0	0	0	0	1	1	0	0	6
14	0	0	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	13
15	1	0	0	1	0	1	1	1	1	1	0	0	1	1	1	0	1	11
16	1	0	0	0	1	1	1	1	1	1	1	1	0	1	1	1	1	13
합격점수	14	3	1	13	13	16	9	11	13	15	13	12	13	15	13	11	8	193

Longford는 이 표2의 1차 고사 성적과 논술고사의 자료들로써 상호종속인 함수관계를 추정하기 위하여 일반선형모형(GLM)을 적용하고 있으나 이 연구에서는 1차 고사 성적과 논술고사의 자료들을 독립인 변수로 분리하여 대수 선형모형에 적합시켜 계산하였다.

3] *Longford* N.T.(1995)의 p231~249에서 인용한 자료임.

[예1] 평가원과 수험자들의 평가자료에서 판별계수 $\hat{a}$, $\hat{b}$ 를 추정하기 위하여 Haberman S.H.(1978)[4]가 제시한 계산의 순서대로 표기하여 표3을 얻었다.

표 3. 대수선형 모형에 의한 판별계수의 계산

i	n_i	in_i	m_{i_0}	y_{i_0}	im_{i_0}	$(i-\theta_0)y_{i_0}m_{i_0}$	$(i-\theta_0)^2y_{i_0}m_{i_0}$	$\exp(b_0 i)$	m_{i_1}	y_{i_1}
1	12	12	12.5	2.52573	12.5	-233.881	685.973	0.99695	12.3411	2.51293
2	15	30	15.5	2.74084	31.0	-272.230	636.460	0.99390	12.3034	2.50987
3	9	27	9.5	2.25129	28.5	-115.662	277.837	0.99087	12.2658	2.50682
4	11	44	11.5	2.44235	46.0	-123.806	223.446	0.98785	12.2284	2.50376
5	13	65	13.5	2.60269	67.5	-119.743	156.792	0.98483	12.1911	2.50070
6	16	96	16.5	2.80336	99.0	-111.381	95.671	0.98182	12.1538	2.49765
7	13	91	13.5	2.60269	94.5	-49.171	26.762	0.97883	12.1167	2.49459
8	9	72	9.5	2.25129	76.0	-8.725	1.581	0.97584	12.0798	2.49153
9	13	117	13.5	2.60269	121.5	20.802	4.732	0.97286	12.0429	2.48847
10	9	90	9.5	2.25129	95.0	34.049	24.079	0.96989	12.0061	2.48542
11	15	165	15.5	2.74084	170.5	110.118	104.139	0.96693	11.9695	2.48236
12	15	180	15.5	2.74084	186.0	152.601	199.993	0.96398	11.9329	2.47930
13	6	78	6.5	1.87180	84.5	55.870	137.064	0.96104	11.8965	2.47624
14	13	182	13.5	2.60269	189.0	196.484	422.157	0.95810	11.8602	2.47319
15	11	165	11.5	2.44235	172.5	185.151	499.732	0.95518	11.8240	2.47013
16	13	208	13.5	2.60269	216.0	266.756	778.127	0.95226	11.7879	2.46707
합계	193	1622	201.0	40.07543	1690.0	-13.068	4274.545	15.59113	193.0001	39.84003

평가원군의 판별계수는 앞절의 식 (2),(3),(4)에 의하여

$\theta_0 = 8.4079601$, $\hat{b} = -0.0030572$

$\hat{a} = 2.51599$, $\hat{y} = 2.51599 - 0.0030572\hat{x}$

이고

$\hat{b}$ 의 95% 신뢰구간은

$$\hat{b} \pm Z_{0.025}S(\hat{b}),\ \text{단}\ S(\hat{b}) = \frac{1}{\sum_i (i-\bar{x})^2 \hat{m}_i} = 0.00024385$$

에 의하여 $[-0.033663,\ 0.027549]$을 얻을 수 있다. 또 수험자군의 판별계수와 신뢰구간도 같은 식에 의하여 계산되고 그것을 정리하면, 평가원군의 판별계수와 회귀방

4] Haberman, S.H.(1978)는 그의 저서 11페이지의 Table 1.3에서 7번째 열 $(I-\theta_0)Y_{i_0}m_{i_0}$ 대신 $(I-\theta_0)m_{i_0}$를 계산하여 $\hat{b}$의 계산과정에 틀린 결과를 제시하고 있음.

정식은

$\theta = 8.4079601$, $\hat{b} = -0.0030572$, $\hat{a} = 2.51599$,

$\hat{y} = 2.51599 - 0.0030572\hat{x}$

이고 수험자군의 판별계수와 회귀방정식은

$\theta = 9.4466501$, $\hat{b} = -0.0032505$, $\hat{a} = 2.40010$,

$\hat{y} = 2.40010 - 0.0032505\hat{x}$

이며 각 군의 $\hat{b}$에 대한 95% 신뢰구간은 각각 [−0.033663, 0.027549]와 [−0.025460, 0.03196]이었다.

다음에는 잔차 $d_i = n_i - m_i$의 분산

$$\hat{c}_i = \hat{m}_i[1 - \frac{m_i}{N} - \frac{(i-\bar{x})^2\hat{m}_i}{\hat{s}}], \quad \hat{s} = \sum(i-\bar{x})^2\hat{m}_i$$

로써 조정된 잔차(Adjusted Residual)

$r_i = \frac{m_i - n_i}{\sqrt{\hat{c}_i}}$의 값을 구하여 다음 표4를 얻었다.

표 4. 평가원군의 잔차 d_i, 조정된 잔차 r_i와 잔차 d_i의 분산 c_i

평가원	d_i	c_i	r_i
1	-0.34105	9.5160	-0.11056
2	2.69662	10.0052	0.85252
3	-3.26583	10.4149	-1.01197
4	-1.22838	10.7463	-0.37472
5	0.80894	11.0010	0.24389
6	3.84616	11.1803	1.15027
7	0.88326	11.2855	0.26292
8	-3.07975	11.3179	-0.91545
9	0.95712	11.2789	0.28499
10	-3.00612	11.1697	-0.89947
11	3.03053	10.9917	0.91408
12	3.06707	10.7462	0.93561
13	-5.89650	10.4343	-1.82542
14	1.13981	10.0573	0.35941
15	-0.82399	9.6164	-0.26571
16	1.21211	9.1129	0.40152

지금 이 결과들을 Longford의 결론과 비교하면 그는 총성적과 논술고사 성적간의 회귀방정식, $-2.264+0.382\hat{x}$ 에 의하여 수험자군의 불합격 판정 총성적이 $|\frac{2.264}{0.382}|=5.93$임을 지적하고 있는데 여기서도 수험자들의 판정기준치를 관찰하면 $\hat{a}=2.40010$으로 논술성적의 합격점이 2.4이하인자는 한명으로써 같은 결론임을 보이고 있으며 다음 잔차의 분석에서는 대수모형에 의한 자료의 적합이 적당하는 사실을 증명하기 위해서 두 군의 잔차가 모두 정규분포 $N(0,1)$에 수렴하는 것을 보이면 충분하므로 이것을 프로그램에서 적율(Moment)로 표기하여 다음과 같이 분포의 특성치와 분포형의 적합정도를 나타내었다.

표 5. 잔차 가 정규분포로 수렴하는 결과의 표시

Moments

	N	16	Sum Wgts	16
	Mean	0.000122	Sum	0.001949
	Std Dev	0.836168	Variance	0.699177
	Skewness	-0.62543	Kurtosis	-0.17527
	USS	10.48765	CSS	10.48765
	CV	686594.4	Std Mean	0.209042
	T:Mean=0	0.000583	Pr>:T:	0.9995
	Num ^=0	16	Num > 0	9
	M(Sign)	1	Pr>=:M:	0.8036
Variable=RI	Sgn Rank	4	Pr>=:S:	0.8603
	W:Normal	0.944044	Pr<W	0.3947

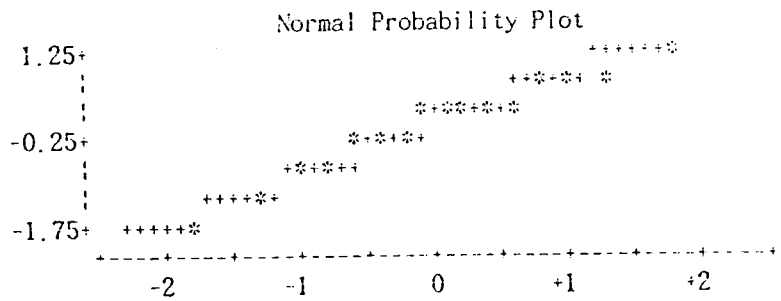

그림1. 평가원군의 잔차 r_i의 분포함수

[예2] 이 예에서는 평가원들의 신뢰도 계수를 앞절의 식(5)를 이용하여 표6과 같이 계산하였다.

표 6. 평가원 군의 평방합과 평균자승, 신뢰도 계수 r^*

SSB	SSW	SSR	SSE	MSB	MSW	MSE	r^*
16.6176	39.4375	6.64338	32.7941	1.03860	0.15466	0.13664	0.26320

여기서 $r^*=0.26320$이고 $F_1=\frac{MSb}{MS_w}=6.71537$, $F_2=\frac{MS_b}{MS_e}=7.6009$으로

$F_{0.95}(16.255)\fallingdotseq 1.75$, $F_{0.95}(16.240)\fallingdotseq 1.75$로써 F_1과 F_2는 모두 유의한 결과가 나오므로 평가원들의 판정에 차이가 있다는 결론을 얻게되고 r^*의 값의 신뢰도도 매우 낮은 결과를 보이고 있다. 그러나 Longford는 평가원들의 판정에 관한 구체적 근거를 제시함이 없이 논하고 있으므로 이 결과는 서로 상반된 결과를 나타내고 있는 것 같다.

[예3] 다음 표7 은 Longford N.T.(1995)의 자료로써 어느 대학 체육고사 점수이며 제 1부의 1차고사 성적과 제 2부의 2원화 평가자료로 표현된 것이다.

표 7.

	수험자												합격
평가원	1	2	3	4	5	6	7	8	9	10	11	12	점수
1차고사 성 적	60	56	52	49	47	40	38	37	32	30	15	14	
1	1	1	1	1	1	1	1	1	0	0	0	0	8
2	0	1	1	1	1	1	1	0	0	0	0	0	6
3	1	1	1	1	1	1	0	0	0	0	0	0	6
4	1	1	1	1	1	1	1	0	0	1	0	0	8
5	1	1	1	1	1	1	1	1	0	0	0	0	8
6	1	1	1	1	1	1	1	1	0	1	0	0	9
7	1	1	1	1	0	1	1	1	0	0	0	0	7
8	1	1	1	0	1	1	0	0	0	0	0	0	5
9	1	1	0	0	1	1	1	0	0	1	0	0	6
10	0	1	0	0	1	0	0	0	0	0	0	0	2
11	1	1	1	0	0	1	0	0	0	1	0	0	5
	9	11	9	7	9	10	7	4	0	4	0	0	70

이 자료에 대한 계산과정과 검토는 [예1]과 [예2]의 경우와 같으며 그 자료처리 결과를 평가원과 수험자군으로 구별하여 판별계수, 판별계수의 신뢰구간, 회귀방정식과 신뢰도계수 r^*순으로 기재하면 다음과 같다.

평가원군 : $\hat{a}=2.10724,\ \hat{b}=-0.044413$

$\hat{y}=2.10724-0.044413\hat{x}$

$[-0.11890,\ 0.030074]$

$r^*=0.48857$

수험자군 : $\hat{a}=2.45476,\ \hat{b}=-0.11913$

$\hat{y}=2.45476-0.11913\hat{x}$

$[-0.18903,\ -0.049228]$

또 평가원의 신뢰도에 대한 F검정의 결과는

$$F_1=\frac{MS_b}{MS_w}=11.5,\quad F_2=\frac{MS_b}{MS_e}=13.199$$

가 모두 $F_{0.95}[11,120]\fallingdotseq 1.28$, $F_{0.95}[11,165]\fallingdotseq 1.28$보다 크고 신뢰도도 조금 높은 편이나 평가원 판정에는 차이가 있다는 결론으로써 그 이유는 10번 평가원 한명이 너무 적은 합격점수를 주는 판정 결과인 것 같다.

Ⅲ. 결 론

본고에서는 앞에서 논한 여러학자들의 견해와는 달리 1차 고사성적과 2차고사 성적의 자료들을 서로 독립인 확률변수로 두었을 때 그 자료의 적합과정이 가장 적당한 대수 선형모형을 응용하여 2원화평가 자료의 최적평가치 추정과 Gordon의 신뢰도계수로 평가원의 신뢰도 계수를 계산하였는데 이 논문의 주요 연구 결과와 기대효과 및 활용방안을 요약하면 다음과 같다.

(1) 이 논문의 인용자료는 Longford N.T. (1995)에서 인용한 것으로 이것은 앞으로 시행예정인 예비교사 훈련 프로그램에서 그들의 교육자적 자질과 전공능력의 검정, 평가원들의 평가에 대한 신뢰도 등을 검토하는데 참고가 될 수 있을 것이다.

(2) Ⅱ장 Ⅱ-1에서 논한바와 같이 대수선형모형에 의한 판별계수 추정법으로써 다른 자료 적합법보다 더 간편하고 용이하게 계산하는 해석법을 제시하였다.

(3) 여러 학자들이 제시한 신뢰도 계수의 종류와 그 이용도에 대해서 요약한 결과 이원화 평가자료에서는 Gordon의 계수가 가장 안정적임을 보였다.

(4) Longford는 그의 자료에서 일반 선형 모형에 의한 자료처리결과 합격과 불합격의 판정을 총점을 기준으로 하여 5.93에서 결정되는 것으로 논하고 있으며 본 연구 결과에서도 이원화자료의 판별계수 $\hat{a}$ 가 2.4001로 그 점수 이하인 수험자는 불합격 될 수 있는 것으로써 두 결과가 모두 동일한 것으로 증명되었다.

(5) Longford는 평가원들의 판정에 대한 신뢰도가 모두 동일하다는 논평에 대하여 계산근거를 제시하지 않고 있으나 이 연구에서는 Gordon의 신뢰도 계수치와 분산분석표에 의하여 F 검정을 한 결과 평가원들의 판정에는 유의적인 차이가 있으며 신뢰도 계수치도 낮은 것으로 나타났다.

(6) 이 연구 결과의 프로그램은 면접문항의 평가, 인성적성 검사, 예체능시험점수 등의 일반적인 이원화 평가자료의 평가에 널리 활용될 수 있다.

참 고 문 헌

홍석강(1990), 신구 두고사 평가치 변환에 의한 진분포와 모수 추정에관한 연구, 한국수학교육학회지, 수학교육 제 29권, 제 2호, 79-93.

홍석강(1994), 결측치를 가진 목표지향형 평가 모델에서 수학학습능력의 평가에 관한 연구, 한국수학교육학회지, 수학교육, 제 33권, 제 2호, 167-175.

Cogner, A.J.(1980), Integration and generalization of kappa for multiple raters, Psychological Bulletin, 88, 322-328.

Fleiss, J.L.(1971), Measuring nominal scale agreement among many raters, Psychological Bulletin, 76, 378-382.

Fleiss, J.L. and Cohen, J.(1973), The equivalence of weighted kappa and intraclass correlation coefficient as measures of reliability, Educational and Psychological Measurement, 33, 613-619.

Gordon, R.(1984), On measuring agreement among several judges on the presence or absence of a trait, Educational and Psychological Measurement, 44, 247-253.

Haberman, S.J.(1978), Analysis of qualitative data, Vol.1, Academic Press, N.Y.

Holland, P.W. and Thayer, D.T.(1993), Stability of the MHD-DIF statistics across population, in Holland P.W. and Walner, H.(Eds.) Differential Item Functioning 171-196 LEA. N.J.

Longford N.T.(1995), Models for uncertainty in educational testing, Springer, N.Y.

Maxwell, A.E. and Pillinger, A.E.G.(1968), Deriving coefficients of reliability and agreement for ratings, British Journ. of Mathematical and Statistical Psychology 21, 105-116.

McCullagh, P. and Nelder, J.A.(1996), Generalized linear models (3rd Ed.), Chapman and Hall, London, UK.

Subkoviak, M. (1978), Empirical investigation of procedures for estimating reliability for mastery tests, Journ. of Educational Measurement 15, 111-116.