제 15 장 도수분석

제1절 독립성검정 (Test of Independence)

관찰자료를 두 가지 분류기준으로 나누었을 때 분류기준이 된 변수들이 서로 독립적인가를 알아보기 위해 χ^2검정을 이용한다.

[표 15-1] 세 도시에 사는 사람들이 선호하는 월간지 종류 - Contingency Table

도시	월간지A	월간지B	월간지C	합계
서울	150	115	135	400
부산	80	75	145	300
광주	60	140	100	300
합계	290	330	380	1,000

[표 15-2] 기대도수 $(f_e)_{ij} = \dfrac{(i\text{번째 행합계})(j\text{번째 열합계})}{\text{총합계}}$

도시	월간지A	월간지B	월간지C	합계
서울	116	132	152	400
부산	87	99	114	300
광주	87	99	114	300
합계	290	330	380	1,000

[표 15-3] 관찰도수 f_o와 기대도수 f_e

도시	월간지A		월간지B		월간지C	
	f_o	f_e	f_o	f_e	f_o	f_e
서울	150	116	115	132	135	152
부산	80	87	75	99	145	114
광주	60	87	140	99	100	114

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

$(f_o - f_e)^2 / f_e$ Table

9.97	2.19	1.90
0.56	5.82	8.43
8.38	16.98	1.72

합계 = 55.95

① H_0: 월간지 구독은 도시 종류와 독립적이다.

H_A: 월간지 구독은 도시 종류와 독립적이 아니다.

② Test Statistic: $\chi^2_{(r-1)(c-1)} = \sum \frac{(f_o - f_e)^2}{f_e}$

③ For α, Rejection Region: $\chi^2 > \chi^2_{\alpha,\ (r-1)(c-1)}$

[표 15-1]의 예

$\alpha = 0.05,\ \chi^2_{0.05,\ (3-1)(3-1)} = \chi^2_{0.05,\ 4} = 9.4877$

Test Statistic의 값이 55.95이므로, H_0를 기각한다.

Yates Correction (예츠수정)

2×2 분할표인 경우, χ^2의 자유도가 1이 된다. → 오차가 상당히 커진다.

$$\chi^2 = \sum \frac{\left(| f_o - f_e | - \frac{1}{2} \right)^2}{f_e}$$

제2절 적합도검정 (Goodness-of-Fit Test)

가정한 확률분포가 타당한 지의 여부를 검정

1. 포아송분포에 대한 적합도검정

① H_0: 병원에 오는 환자의 확률분포는 $\lambda = 0.8$인 포아송분포이다.

H_A: 병원에 오는 환자의 확률분포는 $\lambda = 0.8$인 포아송분포가 아니다.

② Test Statistic: $\chi^2_{k-m-1} = \chi^2_{k-2} \sum \frac{(f_o - f_e)^2}{f_e}$

여기서 k : 계급의 수,

m : 모수 추정 위해 사용된 표본 추정치의 수

(m=1)

③ For α, Rejection Region: $\chi^2 > \chi^2_{\alpha,\, k-2}$

환자 수 X	관찰도수 f_o	포아송확률	기대도수 f_e	$\frac{(f_o - f_e)^2}{f_e}$
0	36	0.4493	44.93	1.78
1	40	0.3595	35.95	0.46
2	19	0.1438	14.38	1.49
3 이상	5	0.0474	4.74	0.01
합계	100	1.0000	100.00	3.73

예) $k = 4$, $\alpha = 0.05$, $\chi^2_{0.05,\, 2} = 5.9915$, $\chi^2 = 3.73$으로 $\chi^2_{0.05,\, 2}$ 보다 작다.

환자의 확률분포가 포아송분포가 아니라고 기각할 수 없다.

2. 정규분포에 대한 적합도검정

① H_0: 점수는 정규분포이다.

H_A: 점수는 정규분포가 아니다.

② Test Statistic: $\chi^2_{k-m-1} = \chi^2_{k-3} = \sum \frac{(f_o - f_e)^2}{f_e}$

여기서 k : 계급의 수,

m : 모수 추정 위해 사용된 표본 추정치의 수

(평균과 분산을 추정하므로 m=2)

③ For α, Rejection Region: $\chi^2 > \chi^2_{\alpha,\, k-3}$

점수	관찰도수	$P(X)$	f_e	adj f_e	$\frac{(f_o - f_e)^2}{f_e}$
20~30	2	0.0089	0.89		
30~40	3	0.0403	4.03		
40~50	10	0.1251	12.51	17.43	0.34
50~60	23	0.2379	23.79	23.79	0.03
60~70	36	0.2770	27.70	27.70	2.49
70~80	14	0.1975	19.75	19.75	1.67
80~90	9	0.0862	8.62	10.92	0.11
90~100	3	0.0230	2.30		
합계	100				4.63

예) k = 5 (기대도수가 5 미만인 계급은 합한 결과, 계급의 수가 8에서 5로 축소되었음), α = 0.05

$\chi^2_{0.05,\,2}$ = 5.9915, χ^2 = 4.63으로 $\chi^2_{0.05,\,2}$ 보다 작다.

점수 확률분포는 정규분포가 아니라고 기각할 수 없다.

제3절 동일성검정

자동차	남자	여자	합계
A	50	60	110
B	46	54	100
C	25	29	54
합계	121	143	264

H_0: $\pi_M = \pi_F$ (자동차 A)

$\pi_M = \pi_F$ (자동차 B)

$\pi_M = \pi_F$ (자동차 C)

H_1: 위의 모든 등식이 반드시 성립하지는 않는다.

자동차	남자		여자	
	f_o	f_c	f_o	f_c
A	50	55	60	55
B	46	50	54	50
C	25	27	29	27

Test Statistic: $\chi^2_{(r-1)(c-1)}$

Rejection Region: $\chi^2 > \chi^2_{\alpha,(r-1)(c-1)} = \chi^2_{0.05,\,2} = 5.99$

The value of the Test Statistic:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

$$= \frac{(50-55)^2}{55} + \frac{(60-55)^2}{55} + \frac{(46-50)^2}{50} + \frac{(54-50)^2}{50} + \frac{(25-27)^2}{27} + \frac{(29-27)^2}{27}$$

$$= 1.845$$

Conclusion: Do not reject H_0.