제11장 범주형 자료분석

1. 서론

(1) 범주형 자료(categorical data): 관찰값이 범주형 관찰도수로 분류되는 자료

① 질적 특성에 따라 정의되는 경우
- 종교: 가톨릭, 기독교, 불교, 기타 - 직업만족도: 매우 만족, 보통, 불만족
- 냉장고의 성능: 매우 양호, 양호, 약간 결함, 불량

② 계량적 척도에 따라 정의되는 경우
- 근로자 연봉 총액: 높음, 보통, 낮음
- 연간 총강수량: 매우 많음, 보통, 적음

[예제 11.1] 여러 개 범주로 분류되는 단일표본자료

<멘델의 유전법칙> 연두색, 노랑색 완두콩의 교배실험, $n=1{,}301$

연두색 : 노랑색 : 줄무늬 연두색 : 줄무늬 노랑색 = 9 : 3 : 3 : 1

<표 11-1> 완두콩 교배실험 자료의 분류

종류	연두색	노랑색	줄무늬연두색	줄무늬노랑색	총계
도수	773	231	238	59	1,301

네 가지 유전형이 나올 확률: $p_1,\ p_2,\ p_3,\ p_4$

멘델의 유전법칙: [통계적 가설] $H_0 : p_1 = \frac{9}{16}, p_2 = \frac{3}{16}, p_3 = \frac{3}{16}, p_4 = \frac{1}{16}$

관찰도수(observed frequency)가 귀무가설에 위배되는지를 검토

[예제 11.2] 독립표본들로 구성된 범주형 자료

비타민 C 가 감기예방에 효과가 있는지, 279명의 스키선수를 대상으로, 두 개의 집단인 대조집단(control group, 비타민을 복용하지 않는 집단)과 복용집단(experimental group, 비타민을 복용한 집단)으로 나누어 실험

<표 11-2> 비타민 C 복용에 따른 감기 감염 여부

	감염	정상	총계
대조집단	31	109	140
실험집단	17	122	139
총계	48	231	279

<표 11-2a> 상대도수

	감염	정상	총계
대조집단	$0.22\left(=\frac{31}{140}\right)$	$0.78\left(=\frac{109}{140}\right)$	1
실험집단	$0.12\left(=\frac{17}{139}\right)$	$0.88\left(=\frac{122}{139}\right)$	1

<표 11-2b> 모비율 [H_0 하에서]

	감염(p_1)	정상(p_2)	총계
대조집단	p_{11}	p_{12}	1
	$\|$	$\|$	
실험집단	p_{21}	p_{22}	1

H_0 : 두 집단 간의 차이가 없다. $\Rightarrow$ $H_0 : p_{11} = p_{21},\ \ p_{12} = p_{22}$

[예제 11.3] 두 가지 특성치에 따라 동시에 분류되는 단일표본자료

<표 11-3> 콜레스테롤과 혈압에 따른 분류(심혈관관상동맥 환자 92명 대상)

콜레스테롤	혈압				총계
	127 미만	127~146	146~166	166 이상	
220 미만	5	5	3	7	20
220 이상	15	23	17	17	72
총계	20	28	20	24	92

주어진 관찰도수를 바탕으로 콜레스테롤과 혈압이 어떤 연관성을 갖는지?

H_0 : 행에 따라 열의 반응확률이 '차이가 없다' $\Rightarrow$ 두 특성이 서로 독립

$\Rightarrow$ 콜레스테롤 수준과 혈압은 서로 무관하여 독립

(2) 교차분류표(cross-classified table) [분할표(contingency table)]

① 관찰값이 두 개 또는 그 이상의 특성에 따라 분류되는 도수분포표

② $r \times c$ 분할표 (r: 분할표의 행 수, c: 분할표의 열 수)

2. 피어슨의 카이제곱 적합도검정

(1) 적합도검정 (goodness-of-fit test)

: 귀무가설에 의해 언급된 모형이 자료에 적합한지 가설검정하는 절차

(2) 적합도검정에 대한 일반적 논의

- 표본크기 n인 확률표본이 k개의 범주$(1, 2, \cdots, k)$로 분류된다
- 칸 도수: $n_1, n_2, \cdots, n_k$
- 칸 확률: $p_1, p_{2,} \ldots, p_k$

(3) 피어슨의 카이제곱 적합도검정 (n이 큰 경우)

① 어떤 정해진 값 $p_{10}, p_{20,} \ldots, p_{k0}$에 대해 다음 귀무가설에 관심이 있다고 하자

$H_0 : p_1 = p_{10},\ p_2 = p_{20},\ \cdots,\ p_k = p_{k0}$ 단, $p_{10} + p_{20} + \cdots + p_{k0} = 1$

<표 11-4> 적합도검정의 기본구조

관찰도수(O)	n_1	n_2	$\cdots$	n_k
H_0 조건에서 칸 확률	p_{10}	p_{20}	$\cdots$	p_{k0}
기대도수(E)	np_{10}	np_{20}	$\cdots$	np_{k0}

② 검정통계량: $\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_{i0})^2}{np_{i0}} = \sum_{\text{모든칸}} \frac{(O-E)^2}{E}$, 자유도=$k-1$ [칸 개수-1]

※ 관찰도수와 귀무가설 하의 기대도수 간의 전체적인 불일치의 크기

③ 기각역: $\chi^2 \geq \chi_\alpha^2$, 자유도=$k-1$ [칸 개수-1]

[예제 11.4] [예제 11.1]에 주어진 자료가 멘델의 유전법칙을 따르는지 피어슨 적합도검정을 통해 살펴보자. 단, 유의수준은 $\alpha = 0.05$로 간주한다.

✿풀이❀

$d.f. = 3$일 때, $\chi^2_{0.05} = 7.81 \rightarrow R: \chi^2 \geq 7.81$

검정통계량 $\chi^2 = 9.26$이 기각역에 포함되므로 귀무가설을 기각할 수 있다. 따라서 <표 11-1>의 자료는 멘델의 유전법칙을 따른다고 볼 수 없다.

<표 11-5> <표 11-1> 자료에 대한 χ^2 검정통계량의 계산

칸	연두색	노랑색	줄무늬 연두색	줄무늬 노랑색	총계
관찰도수(O)	773	231	238	59	1301
H_0 조건에서 칸 확률	9/16	3/16	3/16	1/16	1.0
기대도수(E)	731.9	243.9	243.9	81.3	1301
$\frac{(O-E)^2}{E}$	2.31	0.68	0.14	6.12	$\chi^2 = 9.26$ $d.f. = 3$

(4) 피어슨 카이제곱 통계량의 기본성질

① 가법성

서로 독립인 표본들로부터 계산된 χ^2 통계량의 합도 χ^2 통계량이 되며, 자유도는 각 표본의 자유도의 합과 같다.

② 모수가 추정된 경우의 자유도

칸 확률이 귀무가설 H_0에 의해 완전히 규명되지 않을 때는 기대도수를 계산하기 위해서 추가적으로 모수를 추정해야 한다. 이 경우 χ^2 통계량의 자유도는 추정된 모수의 개수만큼 감소한다. 즉,

χ^2 통계량의 자유도 = (칸 개수) - 1 - (추정된 모수 개수)

3. 분할표에 대한 독립성검정

(1) 분할표에 대한 독립성검정

[예제 11.5] [예제 11.2]의 자료에서 비타민 C의 복용여부에 따라 감기에 감염되는 확률이 서로 동일한지 가설검정하시오.

✿풀이❃

<표 11-6a> <표 11-2>의 자료에 대한 관찰 및 기대도수

	감기		총계
	감염$\left(\hat{p}_1 = \frac{48}{279}\right)$	정상$\left(\hat{p}_2 = \frac{231}{279}\right)$	
대조집단(1)	31 $(24.1 = 140\hat{p}_1)$	109 $(115.9 = 140\hat{p}_2)$	140
실험집단(2)	17 $(23.9 = 139\hat{p}_1)$	122 $(115.1 = 139\hat{p}_2)$	139
총계	48	231	279

- 두 모집단: 대조집단, 실험집단 - 반응(2개의 범주): 감기감염, 정상
- 행 합계 140과 139는 사전에 결정된 표본의 크기

① H_0 : 두 집단 간 반응확률에 차이가 없다(동질성).

$\Rightarrow$ $H_0 : p_{11} = p_{21},\ \ p_{12} = p_{22}$

<표 11-6b> <표 11-2>의 자료에 대한 측도 $\frac{(O-E)^2}{E}$의 값

	감기		
	감염	정상	
대조집단	1.976	0.411	
실험집단	1.992	0.414	
			$\chi^2 = 4.811$ $d.f. = 1$

② $d.f. = 1$일 때, $\chi^2_{0.05} = 3.84 \rightarrow R : \chi^2 \geq 3.84$

③ 검정통계량 $\chi^2 = 4.811$이 기각역에 포함되므로 귀무가설을 기각할 수 있다. 즉 주어진 자료에 의하면 비타민 C는 감기의 감염여부에 유의한 효과가 있는 것으로 판단된다.

※ 유의한 차가 생긴 요인

감기 '감염'의 범주가 χ^2값에 큰 영향을 주고 있다. 상대도수를 구하면, 대조집단은 31/140(22.1%), 실험집단은 17/139(12.2%)로, 비타민을 복용하지 않은 대조집단이 감기에 더 많이 감염되었음을 알 수 있다. 따라서 비타민 C가 감기예방에 효과에 있다고 볼 수 있다.

(2) $r \times c$ 분할표에 대한 χ^2검정

① r개 모집단으로부터 서로 독립인 표본이 c개 반응범주로 분류되는 경우

② 칸 기대도수: 행 합계와 열 합계를 곱한 후 이를 총도수로 나눈다.

③ $r \times c$ 분할표의 자유도$= r(c-1)-(c-1)=(r-1)(c-1)$

(3) 분할표의 행 합계와 열 합계가 미리 고정되지 않은 좀 더 일반적인 형태의 분할표에서 행과 열 변수 간의 독립성 검정

[예제 11.6] 앞 [예제 11.3]에 주어진 콜레스테롤과 혈압 간의 연관성을 χ^2 독립성 검정을 통해 분석하시오.

✿풀이✾

① p_{ij}: 칸 (i,j)의 확률

② 행 및 열의 주변확률

칸 확률: p_{11} =P(콜레스테롤 220미만이고 혈압 127 미만)

행 주변확률: $p_{1.}$ =P(콜레스테롤 220미만)

열 주변확률: $p_{.1}$ =P(혈압 127미만)

③ 독립사상의 곱 사상의 확률은 각각 확률을 곱하여 구할 수 있다.

④ 독립성가설의 조건에서 $p_{11} = p_{1.}p_{.1}$, $p_{12} = p_{1.}p_{.2}$ 등의 관계식이 성립한다.

⑤ 독립성의 귀무가설: H_0: 칸 확률은 이에 대응되는 주변확률의 곱과 같다.

⑥ 통계분석 결과(pp. 277~278, R code)

- 검정통계량: $\chi^2 = 1.6851$ $d.f. = 3$
- $p-value = 0.6402$
- 유의수준 5%에서 귀무가설을 기각할 수 없다.

→ 콜레스테롤과 혈압은 서로 연관성이 있다고 볼 수 없다.

<표 11-3> 콜레스테롤과 혈압에 따른 분류(심혈관관상동맥 환자 92명 대상)

콜레스테롤	혈압				총계
	127 미만	127~146	146~166	166 이상	
220 미만	5	5	3	7	20
220 이상	15	23	17	17	72
총계	20	28	20	24	92

콜레스테롤	혈압				총계
	127 미만	127~146	146~166	166 이상	
220 미만	$p_{11}=\frac{5}{92}$	$p_{12}=\frac{5}{92}$	$p_{13}=\frac{3}{92}$	$p_{14}=\frac{7}{92}$	$p_{1.}=\frac{20}{92}$
220 이상	$p_{21}=\frac{15}{92}$	$p_{22}=\frac{23}{92}$	$p_{23}=\frac{17}{92}$	$p_{24}=\frac{17}{92}$	$p_{2.}=\frac{72}{92}$
총계	$p_{.1}=\frac{20}{92}$	$p_{.2}=\frac{28}{92}$	$p_{.3}=\frac{20}{92}$	$p_{.4}=\frac{24}{92}$	1

H_0: 칸 확률은 이에 대응되는 주변확률의 곱과 같다.

$\Rightarrow$ $p_{11}=p_{1.}p_{.1}$, $p_{12}=p_{1.}p_{.2}$, $p_{13}=p_{1.}p_{.3}$, $p_{14}=p_{1.}p_{.4}$

$p_{21}=p_{2.}p_{.1}$, $p_{22}=p_{2.}p_{.2}$, $p_{23}=p_{2.}p_{.3}$, $p_{24}=p_{2.}p_{.4}$

(4) 정리: 분할표의 χ^2 독립성 검정

① 귀무가설: 행 변수와 열 변수는 서로 독립이다.

② 검정통계량: $\chi^2=\sum_{\text{모든칸}}\frac{(O-E)^2}{E}$, 단 O: 관찰도수, E=(행합계×열합계)/총합계

③ 자유도: (행 개수-1)×(열 개수-1)

④ 기각역: $\chi^2 \geq \chi^2_\alpha$