범주형 자료분석

서론

- 범주형 자료: 관측치들이 몇 개의 범주로 분류되고 각 범주의 도수로 자료가 주어지는 것을 말한다.

ex) 사람들의 종교별 분류, 직업의 만족도에 의한 분류 또는 나무들의 유전자 형태에 의한 분류 등

예제1 – 적합도 검정(Goodness-of-fit Test)

- (몇 개의 범주로 분류된 한 표본의 문제) 어떤 나무의 자가수정의 결과로 나올 수 있는 유전자의 형태는 세 종류로 분류된다고 한다. 이 세 종류를 각각 A, B, C 라 할 때 생물학에서의 한 이론에 의하면 비율이 1:2:1 로 나타난다고 하자. 이를 입증하기 위하여 자가수정의 결과로 생겨난 나무 100그루를 유전자의 형태별로 분류하여 다음과 같은 표를 얻었다.

- 적합도 검정(Goodness-of-fit Test)

$$H_0: p_A = \frac{1}{4},\ p_B = \frac{1}{2},\ p_C = \frac{1}{4}$$

유전자 형태	A	B	C	합계
관측 도수	18	55	27	100

예제2 – 동질성 검정(Homogeneity Test)

- (몇 개의 범주로 분류된 두 개의 독립인 표본의 문제) 두가지 식이요법 A와 B의 효과를 비교하기 위해서 150명의 환자를 대상으로 조사를 실시하였다. 임의로 추출된 80명에게는 식이요법 A를 적용하고, 나머지 70명에게는 식이요법 B를 적용한 후 얼마간의 시간이 흐른 후에 각 환자의 건강상태에 따라 다음의 표에서와 같이 세 가지 범주로 분류하였다.

- 동질성 검정(Homogeneity Test)

H_0: $p_{A1} = p_{B1},\ p_{A2} = p_{B2},\ p_{A3} = p_{B3}$

	양호	보통	불량	표본의 크기
식이요법 A	37 (p_{A1})	24 (p_{A2})	19 (p_{A3})	80
식이요법 B	17 (p_{B1})	33 (p_{B2})	20 (p_{B3})	70
합계	54	57	39	150

예제3 – 독립성 검정(Independence Test)

- (한 표본을 두 가지 특성에 따라 분류하는 문제) 텔레비전에서 방영되는 오락물에 대한 사람들의 의견이 성별과 어떤 관계가 있는지 조사하기 위해서 1250명의 사람을 임의 추출하여 성별과 오락물 방영에 대한 의견으로 다음 표와 같이 분류하였다.

- 독립성 검정(Independence Test)

H_0: 오락물 방영에 대한 의견은 개인의 성별과 무관하다.(독립이다.)

성별	오락물 방영			합계
	너무 많다	적당하다	너무 적다	
남자	378	237	26	641
여자	388	196	25	609
합계	766	433	51	1250

분할표(Contingency Table)

	양호	보통	불량	표본의 크기
식이요법 **A**	37 (p_{A1})	24 (p_{A2})	19 (p_{A3})	80
식이요법 **B**	17 (p_{B1})	33 (p_{B2})	20 (p_{B3})	70
합계	54	57	39	150

성별	오락물 방영			합계
	너무 많다	적당하다	너무 적다	
남자	378	237	26	641
여자	388	196	25	609
합계	766	433	51	1250

피어슨의 χ^2(카이제곱) 적합도 검정

- 각 범주의 확률에 대한 가설 $H_0: p_1 = p_{10}, p_2 = p_{20}, \cdots, p_k = p_{k0}$ 를 검정하기 위한 검정통계량은 다음과 같다.

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \sum_{i=1}^{k} \frac{(n_i - np_{i0})^2}{np_{i0}} \quad \textbf{where } O = n_i\text{: 관측도수}, E = np_{i0}\text{: 기대도수}$$

귀무가설이 맞을 때 검정통계량의 분포는 표본의 크기가 클 때 자유도가 $k-1$ 인 χ^2 분포를 따른다. 따라서 유의수준 α 에 대한 기각역은

$$R: \chi^2 \geq \chi^2{}_{\alpha}(k-1)$$

이다.

χ^2 통계량의 특성

- 독립인 표본으로부터 계산된 χ^2 통계량들을 더하면 그 합도 χ^2 통계량이 되고 자유도는 각 자유도의 합과 같다. 즉, $\chi^2(k_1), \chi^2(k_2), \cdots, \chi^2(k_r)$ 이 서로 독립인 r 개의 χ^2 통계량이라면

$$\chi^2(k_1) + \chi^2(k_2) + \cdots + \chi^2(k_r) \sim \chi^2(k_1 + k_2 + \cdots + k_r)$$

이다.

- χ^2 통계량을 계산하는데 만약 모수의 추정치를 사용하였다면 그 통계량의 자유도는 추정한 모수의 개수만큼 감소하게 된다. 즉,

χ^2의 자유도 = (모수를 알고 있을 경우의 자유도) – (추정된 모수의 수)

분할표에서의 χ^2 동질성 검정

- 귀무가설: 각각의 범주에 대한 비율이 모든 모집단에 대하여 동일하다.

검정통계량: $\chi^2 = \sum_{칸} \frac{(O-E)^2}{E}$

O: 관측도수

E: 기대도수 $= \frac{\text{행의 합} \times \text{열의 합}}{\text{전체 합}}$

$$\text{자유도} = \text{행의 수} \times (\text{열의 수} - 1) - (\text{열의 수} - 1)$$
$$= (\text{행의 수} - 1) \times (\text{열의 수} - 1)$$

기각역 $R: \chi^2 \geq \chi^2_{\alpha}$

독립성 검정

- 동질성 검정과 검정통계량, 지유도, 기각역이 모두 동일하고 단지 검정하고자 하는 가설이 달라서 검정 결과에 대한 해석이 달라진다.

- 칸의 추정기대도수 $= \dfrac{\text{칸이 속한 열의 합} \times \text{칸이 속한 행의 합}}{\text{전체 합}}$

	너무 많다	적당하다	너무 적다	합계
남자	p_{M1}	p_{M2}	p_{M3}	p_M
여자	p_{F1}	p_{F2}	p_{F3}	p_F
합계	p_1	p_2	p_3	1

$p_{M1} = p_M p_1, p_{M2} = p_M p_2$ 등이 성립해야 함.

독립성 검정

- 자유도 = (행의 수 × 열의 수 – 1) – (행의 수 – 1) – (열의 수 – 1)
  = (행의 수 × 열의 수) – 행의 수 – 열의 수 + 1
  = (행의 수 – 1)(열의 수 – 1)