통계학 개론

제9장 범주형 데이터의 분석

9.1 분할표

범주형 데이터의 분석(categorical data analysis): 데이터를 특정 변수가 갖는 속성에 의해 몇 개의 군으로 나누고, 이것들이 독립적인지 데이터들이 이론적 분포와 일치하는가 등을 분석하는 것

❖ 분할표(contingency table): 변수의 속성에 따라 분류된 전체 데이터의 빈도표

분할표에서의 통계적 검정은
 하나 이상 변수의 속성에 의해 분류된 관찰수: 관측도수 O_i
 어떤 이론적 분포의 가설하에서 기대되는 도수: 기대도수 E_i
→ 이 둘의 차이를 계산하여 실시하게 된다.

첫째, 두 변수가 있을 때 두 변수가 서로 독립인지 아닌지에 대해 검정을 실시
→ 독립성 검정
둘째, 한 변수의 표본분포가 어떤 이론분포와 일치하는지를 검토
→ 적합도 검정

검정통계량은 다음과 같다.

$$\begin{aligned}\chi^2 &= \frac{(O_1-E_1)^2}{E_1} + \frac{(O_1-E_1)^2}{E_1} + \cdots + \frac{(O_1-E_1)^2}{E_1} \\ &= \sum_i^k \frac{(O_i-E_i)^2}{E_i}\end{aligned}$$

이 통계량은 χ^2_{k-1} 분포를 따른다.

9.2 독립성 검정

 한 모집단에서 변수 A의 속성이 r개이고, 변수 B의 속성이 c개인 r×c 분할표에서 각 속성이 나올 확률이 p_{ij}라 하자. 이것을 표로 표시하면 다음과 같다.

❖ 변수 A와 B에 관한 분할표의 칸별 확률

구분		변수 B				행의 합
		B_1	B_2	…	B_c	
변수 A	A_1	P_{11}	P_{12}	…	P_{1c}	$p_{1\cdot}$
	A_2	P_{21}	P_{22}	…	P_{2c}	$p_{2\cdot}$
	⋮	⋮	⋮	⋮	⋮	⋮
	A_r	P_{r1}	P_{r2}	…	P_{rc}	$p_{r\cdot}$
열의 합		$P_{\cdot 1}$	$P_{\cdot 2}$	…	$p_{\cdot c}$	1

속성 A_i와 B_j가 독립이라면 다음 수식이 만족되어야 한다.

$$P(A_i \cap B_j) = P(A_i) \cdot P(B_j)$$

$$p_{ij} = p_i \cdot p_j$$

독립성 검정: 분할표에서 변수 A와 B가 위의 성질을 만족하는지 검정하는 것

H0: 변수 A와 B는 독립이다. (모든 i,j에 대하여 $p_{ij}=p_i \cdot p_j$)

H1: 변수 A와 B는 독립이 아니다.(관련이 있다.)

❖ 변수 A와 B에 관한 관찰도수 분할표

구분		변수 B				행의 합
		B_1	B_2	…	B_c	
변수 A	A_1	O_{11}	O_{12}	…	O_{1c}	$T_{1\cdot}$
	A_2	O_{21}	O_{22}	…	O_{2c}	$T_{2\cdot}$
	⋮	⋮	⋮	⋮	⋮	⋮
	A_r	O_{r1}	O_{r2}	…	O_{rc}	$T_{r\cdot}$
열의 합		$T_{\cdot 1}$	$T_{\cdot 2}$	…	$T_{\cdot c}$	n

❖ 변수 A와 B에 관한 기대도수 분할표

구분		변수 B				행의 합
		B_1	B_2	⋯	B_c	
변수 A	A_1	E_{11}	E_{12}	⋯	E_{1c}	$E_{1\cdot}$
	A_2	E_{21}	E_{22}	⋯	E_{2c}	$E_{2\cdot}$
	⋮	⋮	⋮	⋮	⋮	⋮
	A_r	E_{r1}	E_{r2}	⋯	E_{rc}	$E_{r\cdot}$
열의 합		$E_{\cdot 1}$	$E_{\cdot 2}$	⋯	$E_{\cdot c}$	1

$$E_{ij} = n \times \left(\frac{T_{i\cdot}}{n}\right) \times \left(\frac{T_{\cdot j}}{n}\right)$$

가설을 검정하는 통계량은 O_{ij}와 E_{ij}의 차이에 근거하여 구한다.

$$\chi^2_{obs} = \sum_{i=1}^{r}\sum_{j-1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

이 통계량은 근사적으로 자유도가 (r-1)(c-1)인 χ^2분포를 따른다.

❖ 독립성 검정

H_0: 변수 A와 변수 B는 독립이다.

H_1: 변수 A와 변수 B는 관련이 있다.

검정기준: 유의수준이 α일때, $\chi^2_{obj} > \chi^2_{(r-1)(c-1),\alpha}$ 이면 H_0를 기각

※ 주의: 독립성 검정에서 χ^2분포를 이용하려면 모든 기대도수가 적어도 5이상이어야 한다. 5보다 작은 기대도수는 인접구간을 합쳐서 분석하는 것이 바람직하다.

9.3 적합도 검정

적합도 검정(goodness of fit): 관찰도수가 정규분포 또는 이항분포 등의 이론분포와 일치하는가를 검정하는 것

❖ 적합도 검정

H_0: $(p_1, p_2, \cdots, p_k) = p(p_{1O}, p_{2O}, \cdots, P_{kO})$

H_1: 적어도 하나의 p_i는 가정된 p_{iO}와 다른다.

선택기준

관찰된 도수가 $(O_1, O_2, \cdots, O_k)$일 때 기대도수가

$(E_1, E_2, \cdots, E_k) = (np_{1O}, np_{2O}, \cdots, np_{KO})$

이므로 유의수준이 α일 때 선택기준은 다음과 같다.

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} > \chi^2_{k-1,\alpha} \text{이면 } H_0 \text{기각}$$

k는 변수값의 개수

※ 주의: 독립성 검정에서 χ^2분포를 이용하려면 모든 기대도수가 적어도 5이상이어야 한다. 5보다 작은 기대도수는 인접구간을 합쳐서 분석하는 것이 바람직하다.