Seoul National University

제 7장. 이산자료의 분석

모비율의 추정

- 표본비율의 표본분포

  : 무한모집단에서 어느 특정 속성의 비율이 p 이고, 크기 n의 랜덤 표본에서 그 속성을 갖는 것의 개수를 X라고 할 때, 표본비율 $\hat{p} = X/n$ 의 표본분포는 다음과 같다

$$\hat{p} \dot{\sim} N\left(p, \frac{p(1-p)}{n}\right)$$

  이는 $np > 5,\ n(1-p) > 5$ 라는 조건을 요구한다

- 모비율에 대한 $(1-\alpha)100\%$ 신뢰구간 (단, $n\hat{p} > 5,\ n(1-\hat{p}) > 5$)

$$\left[\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

- $(1-\alpha)100\%$ 오차한계를 d 이하로 만족하는 표본의 크기 (if $p = p^*$ is given)

$$n \geq p^*(1-p^*)\left(\frac{z_{\alpha/2}}{d}\right)^2$$

  : 모비율이 주어지지 않은 경우에는 $p^* = 1/2$

모비율의 검정

- 모비율의 검정 : 표본의 크기가 큰 경우, $H_0 : p = p_0$ 에 대한 검정통계량

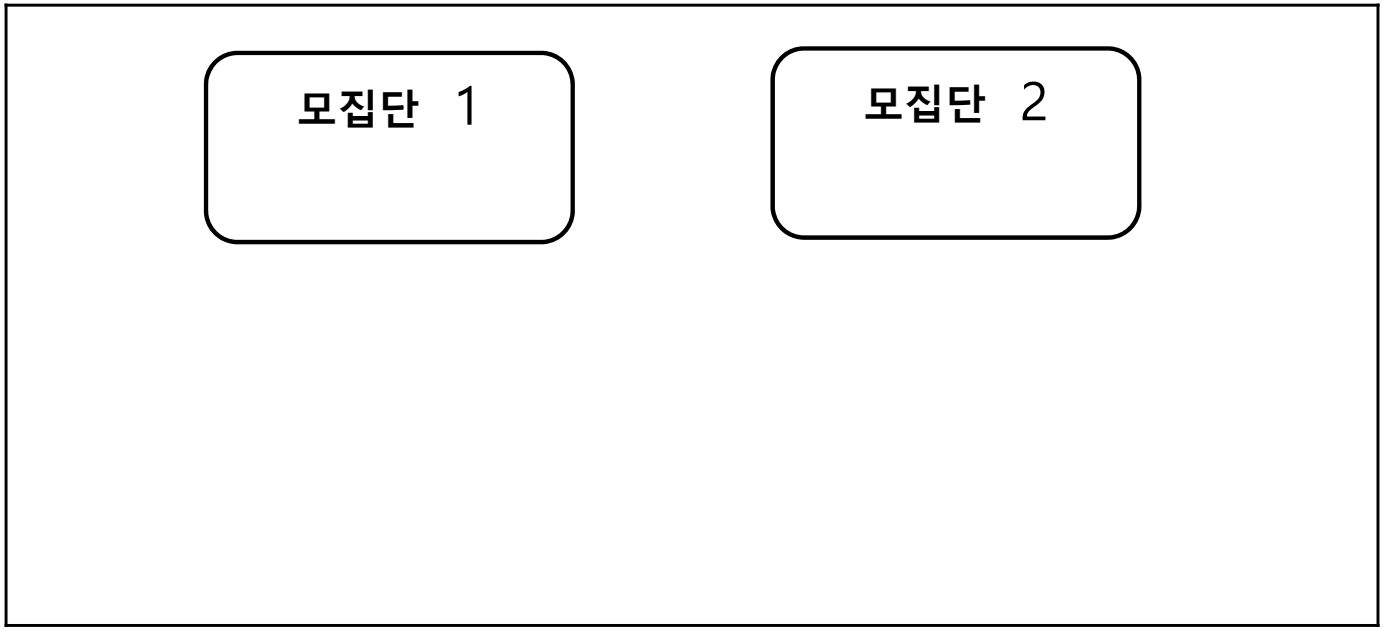
$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \quad \dot{\sim} \; N(0,1)$$

단, $np_0 > 5,\ n(1-p_0) > 5.$

- 예제 7.4 : 올해 추석에 귀향하게 될 가구의 비율을 알아보기 위해 500가구를 조사한 결과, 그 중 79가구가 귀향하려고 한다. 과거 추석 때의 귀향률이 20%였다고 할 때, 올해 귀향률이 감소 했다고 할 수 있는지 유의수준 1%에서 검정하여라.

두 모비율의 비교

- 예 : 두 개의 공장에서 생산된 제품의 불량률에는 차이가 있는가?



- 관심모수 :

두 모비율의 비교

- $\hat{p}_1 - \hat{p}_2$ 의 기대값과 분산

$$E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$$

$$V(\hat{p}_1 - \hat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

- $\hat{p}_1 - \hat{p}_2$ 의 표본분포

  : $n_1 p_1 \geq 5,\ n_1(1-p_1) \geq 5,\ n_2 p_2 \geq 5,\ n_2(1-p_2) \geq 5,$ 를 만족할 때,

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2,\ \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$$

- 두 모비율의 차이에 대한 $100(1-\alpha)\%$ 신뢰구간

$$\left[(\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}},\ (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right]$$

두 모비율의 비교

- 검정통계량
  - CASE I. $H_0 : p_1 - p_2 = 0$

$$Z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

이 때, $\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{X_1 + X_2}{n_1 + n_2}$: **합동 표본비율 (pooled proportion)**

  - CASE II. $H_0 : p_1 - p_2 = D \quad (D \neq 0)$

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - D}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \sim N(0,1)$$

두 모비율의 비교

- 예제 7.7 : 다음의 자료는 충동적 살인범과 계획적 살인범의 교화에 차이가 있는가를 알아 보기 위한 것이다. 일정기간 복역 후에 가석방 된 충동적 살인범과 계획적 살인범 중에서 각각 42명과 40명을 랜덤추출하여, 가석방이 성공적인 경우(재범이 없는 경우)와 실패한 경우의 도수를 관측한 결과가 다음과 같다. 살인범의 유형에 따라 가석방의 성공률에 차이가 있는 가를 유의수준 5%에서 검정하여라

	성공	실패	표본크기
충동적 살인범	13	29	42
계획적 살인범	22	18	40
합계	35	47	82

범주형 자료에 의한 여러 모집단의 비교

- 각 모집단이 두 가지 이상의 서로 다른 속성을 갖는 개체들로 나뉘는 경우

- H_0 : 여러 모집단이 동일하다 vs H_1 :이들이 서로 다르다
  ⇒ 각 속성의 관측도수(observed frequency)와 기대도수(expected frequency)를 비교하는 것으로 검정 가능

- 기대도수 (expected frequency)
  : 귀무가설 H_0 하에서의 추정 기대도수
  : (표본크기) × (H_0 하에서의 추정 확률)

- 귀무가설 하에서 각 범주별로 계산된 기대도수와 실제의 관측도수의 차이를 이용하여 검정통계량을 정의.

적합도 검정

- 예 : 다음은 4개의 제품에 대해 알려진 시장 점유율과 500명을 대상으로 실제 사용하는 제품을 조사한 결과이다. 알려진 시장 점유율이 여전히 유효하다고 말 할 수 있는지 유의수준 5%에서 이를 검정하시오.

제품	A	B	C	D
시장 점유율	20%	40%	10%	30%
사용자 수	110	195	47	148

여러 개의 이항 모집단의 비교

- $(r \times 2)$ 분할표 (contingency table)

	성공	실패	표본 크기
모집단 1에서의 표본	O_{11}	O_{12}	n_1
모집단 2에서의 표본	O_{21}	O_{22}	n_2
...	...	...	...
모집단 r 에서의 표본	O_{r1}	O_{r2}	n_r
합계	$O_{\cdot 1}$	$O_{\cdot 2}$	n

- 이항 모집단의 성공률을 비교하기 위한 가설

$$H_0 : p_1 = p_2 = \ldots = p_r \qquad vs \qquad H_1 : \text{Not } H_0$$

- 검정 통계량

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{2} \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}} \sim \chi^2(r-1)$$

where $\hat{E}_{ij} = n_i(O_{\cdot j} / n)$

여러 개의 이항 모집단의 비교

- 예제 7.8 : 액화천연가스(LNG)의 저장기지의 후보지로 고려되고 있는 세 지역의 여론을 알아보기 위하여, 세 지역에서 각각 400명, 350명, 350명을 랜덤 추출하여 기지의 건설에 대한 찬성 여부를 물은 결과가 다음과 같다. 지역에 따라 찬성률에 차이가 있다고 할 수 있는가?

	찬성	반대	표본 크기
지역 1	198	202	400
지역 2	140	210	350
지역 3	133	217	350
합계	471	629	1100

여러 개의 이항 모집단의 비교

- 예제 7.7 : 다음의 자료는 충동적 살인범과 계획적 살인범의 교화에 차이가 있는가를 알아 보기 위한 것이다. 일정기간 복역 후에 가석방 된 충동적 살인범과 계획적 살인범 중에서 각각 42명과 40명을 랜덤추출하여, 가석방이 성공적인 경우(재범이 없는 경우)와 실패한 경우의 도수를 관측한 결과가 다음과 같다. 살인범의 유형에 따라 가석방의 성공률에 차이가 있는 가를 유의수준 5%에서 검정하여라

	성공	실패	표본크기
충동적 살인범	13	29	42
계획적 살인범	22	18	40
합계	35	47	82

여러 개의 다항 모집단에 대한 비교

- $(r \times c)$ 분할표

	범주 1	범주 2	...	범주 c	표본 크기
모집단 1에서의 표본	O_{11}	O_{12}	...	O_{1c}	n_1
모집단 2에서의 표본	O_{21}	O_{22}	...	O_{2c}	n_2
...	...	...	...	...	...
모집단 r 에서의 표본	O_{r1}	O_{r2}	...	O_{rc}	n_r
합계	$O_{\cdot 1}$	$O_{\cdot 2}$	...	$O_{\cdot c}$	n

- 다항 모집단을 비교하기 위한 가설

$$H_0 : p_{1j} = p_{2j} = \ldots = p_{rj} \ (j = 1, 2, \ldots, c) \qquad vs \qquad H_1 : \text{Not } H_0$$

- 검정 통계량

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}} \sim \chi^2 \left((r-1)(c-1) \right)$$

where $\hat{E}_{ij} = n_i (O_{\cdot j} / n)$

여러 개의 다항 모집단에 대한 비교

- 예제 7.9 : 공단에 인접한 세 지역에서 공해를 느끼는 정도가 지역에 따라 차이가 있는가를 알아보고자 하여 조사한 결과가 다음과 같다. 범주 1은 매일, 점주 2는 적어도 일주일에 한번, 범주 3은 적어도 한달에 한번, 범주 4는 한달에 한번보다는 적게, 범주 5는 전혀 악취를 느끼지 않는 경우를 뜻한다. 이 자료로부터 지역에 따라 공해를 느끼는 정도가 다르다고 할 수 있는가를 유의수준 1%에서 검정하여라.

	범주 1	범주 2	범주 3	범주 4	범주 5	표본 크기
지역 1	20	28	23	14	12	97
지역 2	14	34	21	14	12	95
지역 3	4	12	10	20	53	99
합계	38	74	54	48	77	291

범주형 자료에 의한 독립성 검정

- 한 모집단의 각 개체에 대하여 두 가지 특성을 관측하고, 이들 각 특성을 여러 개의 범주로 나눌 수 있을 때, 이들 특성의 관련성 여부를 검정하는 방법

- 두 특성에 대한 모비율과 관측도수

	B_1	B_2	...	B_c	합계
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}	$P_{r\cdot}$
합계	$P_{\cdot 1}$	$P_{\cdot 2}$	...	$P_{\cdot c}$	

	B_1	B_2	...	B_c	합계
A_1	O_{11}	O_{12}	...	O_{1c}	$O_{1\cdot}$
A_2	O_{21}	O_{22}	...	O_{2c}	$O_{2\cdot}$
⋮	⋮	⋮	...	⋮	⋮
A_r	O_{r1}	O_{r2}	...	O_{rc}	$O_{r\cdot}$
합계	$O_{\cdot 1}$	$O_{\cdot 2}$	...	$O_{\cdot c}$	

- 가설 : $H_0 : p_{ij} = p_{i\cdot} p_{\cdot j} \ (i = 1,...,r)(j = 1,..,c) \quad vs \quad H_1 : \text{Not } H_0$
- 검정통계량

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}} \sim \chi^2\left((r-1)(c-1)\right)$$

where $\hat{E}_{ij} = n\hat{p}_{ij} = O_{i\cdot} O_{\cdot j} / n$

범주형 자료에 의한 독립성 검정

- 예제 7.10 : 대도시의 근교에서 출퇴근하며 혼자서만 승용차를 이용하는 사람 들 중에서 250명을 랜덤추출한 결과가 다음과 같다. 이 자료에 의하면 승용차의 크기와 통근 거리 사이에 관계가 있다고 할 수 있는지를 유의수준 5%에서 검정하여라.

	15km 미만	15km 이상 30km 미만	30km 이상	합계
경차	6	27	19	52
소형차	8	36	17	61
중형차	21	45	33	99
대형차	14	18	6	38
합계	49	126	75	250