Seoul National University

- 일반통계학 -
제 7장
이산자료의 분석

7-1 모비율의 추정과 검정

1. 실업률, 불량률, 지지율 등과 같이 모집단의 특정한 속성을 갖는 개체의 비율을 뜻하는 모비율
2. X 는 특정한 속성을 갖는 것의 개수라고 할 때, $X \sim B(n, p)$이므로 $\hat{p} = X/n$ 으로 모비율 p를 추정한다.
3. $\frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \approx \mathrm{N}(0,1)$ 단, $np \geq 5$, $n(1-p) \geq 5$
4. 모비율 p 에 관한 $100(1-\alpha)\%$ 근사 신뢰구간은

$$\left(\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\quad \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \text{ 단, } n\hat{p} \geq 5,\ n(1-\hat{p}) \geq 5$$

5. 오차한계$\leq d$를 만족하는 표본의 크기 결정

$$z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} \leq d$$

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

p를 모르므로 다음을 사용하여 n을 구함

1. 사전조사 혹은 과거조사를 통한 p^*사용
2. $p = ½$를 사용 ($p(1-p)$를 최대가 되게 하는 p이므로)

7-1 모비율의 추정과 검정

(예제) 어느 지역의 구청장 선거에 갑, 을 두명의 후보가 입후보하였을 때, 이 두 후보에 대한 지지율을 알아보고자 한다. 아무런 사전 정보가 없는 상태에서, 갑 후보의 지지율에 대한 추정의 95%오차한계를 3%이내로 하려면 필요한 표본의 크기는 얼마인가?

(풀이) $n \geq 0.5(1-0.5)(z_{\frac{0.05}{2}}/0.03)^2$=1067.1 을 만족해야 하므로 1068명을 조사해야 한다.

기사 1. 리얼미터의 이번 주중집계는 4월 18일부터 20일까지 3일간 전국 19세 이상 유권자 1524명을 대상으로 전화면접(CATI) 및 자동응답(ARS) 방식으로 무선전화(59%)와 유선전화(41%) 병행 임의걸기(RDD) 방법으로 조사했다. 응답률은 4.4%(총 통화 34,757명 중 1524명 응답 완료)이다. 표본오차는 95% 신뢰수준에서 ±2.5%p이다.

: $n \geq 0.5(1-0.5)(z_{\frac{0.05}{2}}/0.025)^2$=1536

기사 2. 이번 여론조사는 지난 3일부터 5일까지 도내 각 선거구별 유권자 500명씩을 대상으로 유선전화 임의전화번호걸기(RDD)에 의한 전화면접조사방법으로 실시됐다. 각 선거구별 인구비례할당으로 표본을 추출했으며 각 선거구별로 95% 신뢰수준에 표본 오차는 ±4.4%포인트다.

: $n \geq 0.5(1-0.5)(z_{\frac{0.05}{2}}/0.044)^2$=496

7-1 모비율의 추정과 검정

- 모비율에 대한 가설검정

1. 가설

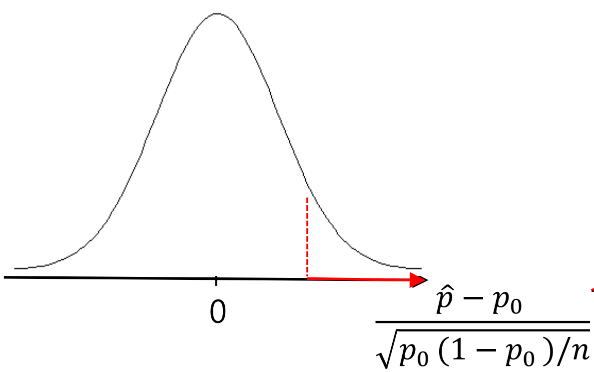
$H_0: p = p_0 \quad vs \quad H_1: p > p_0 \qquad H_0: p = p_0 \quad vs \quad H_1: p < p_0 \qquad H_0: p = p_0 \quad vs \quad H_1: p \neq p_0$

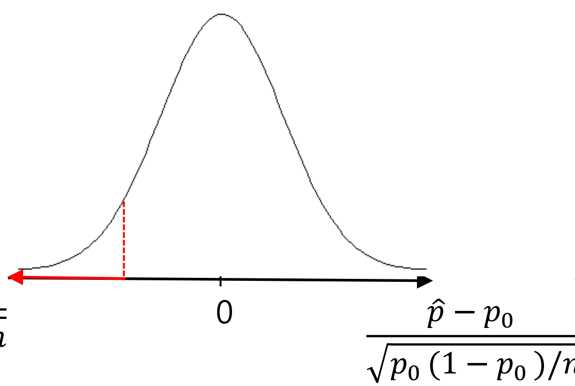
2. 검정통계량

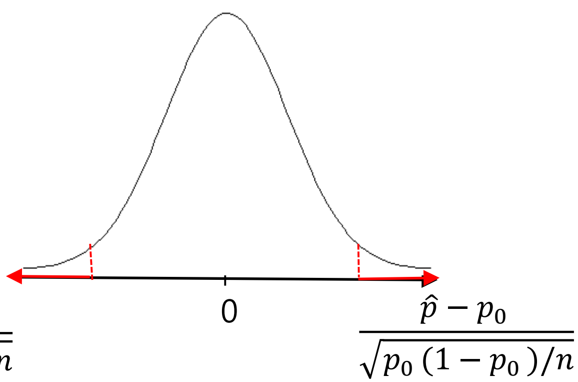
$\frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}} \approx \mathrm{N}(0,1)$, 단, $np_0 \geq 5$, $n(1-p_0) \geq 5$

3. 기각역

$H_0: p = p_0 \quad vs \quad H_1: p > p_0 \qquad H_0: p = p_0 \quad vs \quad H_1: p < p_0 \qquad H_0: p = p_0 \quad vs \quad H_1: p \neq p_0$







7-1 모비율의 추정과 검정

(예제-소표본) 어떤 절삭 공구의 제조 공정에서 기존 공구에 의한 불량률은 5%인데 저렴한 공구를 사용하면 불량률이 증가한다는 증거가 있는지를 확인하기 위하여 저렴한 공구에 의하여 20개의 제품을 시험 생산하였다. 시험 생산에서 3개의 불량품이 나온 경우에 적절한 가설을 세우고 유의수준 5%에서 검정하여라.

1. 가설 $H_0 : p = 0.05 \;\; vs \;\; H_1 : p > 0.05$
2. 유의수준 0.05
3. 불량품의 수 $X \sim B(20, 0.05)$이므로 유의확률 $P(X \geq 3) = 0.075$
4. 결론 유의수준 5%에서 저렴한 공구에 의한 불량률이 증가하는 것을 시사하는 것이라 보기 어렵다.

$np_0 = 20 * 0.05 = 1$ 이므로 이항분포의 정규근사를 이용해서 검정할 수 없다. 따라서 이항분포를 이용하여 유의확률을 구해서 검정한다.

7-1 모비율의 추정과 검정

(예제) 과거 추석때 귀향률이 20%이었다고 하자. 돌아오는 추석에 서울 거주자 중 귀향할 가구의 비율이 과거에 비하여 줄어 든다고 할 수 있는지 알아보기 위하여 500명을 단순랜덤추출하여 조사한 결과 79가구가 귀향하려 한다. 유의수준 1%에서 검정하여라.

1. 가설 $H_0: p = 0.2 \;\; vs \;\; H_1: p < 0.2$
2. 유의수준 0.01
3. 검정통계량 500*0.2=100≥ 5, 00*0.8=400 ≥ 5 이므로 $Z = \frac{\hat{p} - 0.2}{\sqrt{0.2(1-0.2)/n}} \approx \mathrm{N}(0,1)$
4. 계산 $z^* = \frac{79/500 - 0.2}{\sqrt{0.2(1-0.2)/500}}$=-2.34787
5. 검정 $-z_{0.01}$=-2.327 > z^*=-2.34787 이므로 귀무가설을 기각할 만한 증거가 된다.
6. 유의수준 1%에서 귀향률이 과거에 비하여 낮아질 유의미한 증거가 된다.

7-2 두 모비율의 비교

1. 서로 다른 두 치료방법에 의한 치료율의 비교, 두 제조과정에서의 불량률의 비교 등
2. $X_1 \sim B(n_1, p_1)$, $X_2 \sim B(n_2, p_2)$이고, 서로 독립
3. 두 모비율 $p_1 - p_2$의 차에 대한 추정량 $\widehat{p_1} - \widehat{p_2} = \frac{X_1}{n_1} - \frac{X_2}{n_2}$
4. $\mathrm{Var}(\widehat{p_1} - \widehat{p_2}) = \mathrm{Var}(\widehat{p_1}) + Var(\widehat{p_2}) = p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2$
5. $\frac{\widehat{p_1} - \widehat{p_2} - (p_1 - p_2)}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}} \approx N(0,1)$ 단, $n_1 p_1 \geq 5, n_1(1-p_1) \geq 5, n_2 p_2 \geq 5, n_2(1-p_2) \geq 5$
6. 모비율의 차 $p_1 - p_2$에 관한 $100(1-\alpha)\%$ 근사 신뢰구간은

$$\left(\widehat{p_1} - \widehat{p_2} - z_{\frac{\alpha}{2}}\sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1} + \frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}},\ \widehat{p_1} - \widehat{p_2} + z_{\frac{\alpha}{2}}\sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1} + \frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}}\right)$$

단, $n_1\widehat{p_1} \geq 5, n_1(1-\widehat{p_1}) \geq 5, n_2\widehat{p_2} \geq 5, n_2(1-\widehat{p_2}) \geq 5$

7-2 두 모비율의 비교

- 두 모비율의 비교를 위한 가설검정

1. 가설

(1)$H_0: p_1 = p_2 \;\; vs \;\; H_1: p_1 > p_2 \;\; (H_0: p_1 - p_2 = 0 \; vs \;\; H_1: p_1 - p_2 > 0)$

(2)$H_0: p_1 = p_2 \;\; vs \;\; H_1: p_1 < p_2 \;\; (H_0: p_1 - p_2 = 0 \; vs \;\; H_1: p_1 - p_2 < 0)$

(3)$H_0: p_1 = p_2 \; vs \;\; H_1: p_1 \neq p_2 \;\; (H_0: p_1 - p_2 = 0 \;\; vs \;\; H_1: p_1 - p_2 \neq 0)$

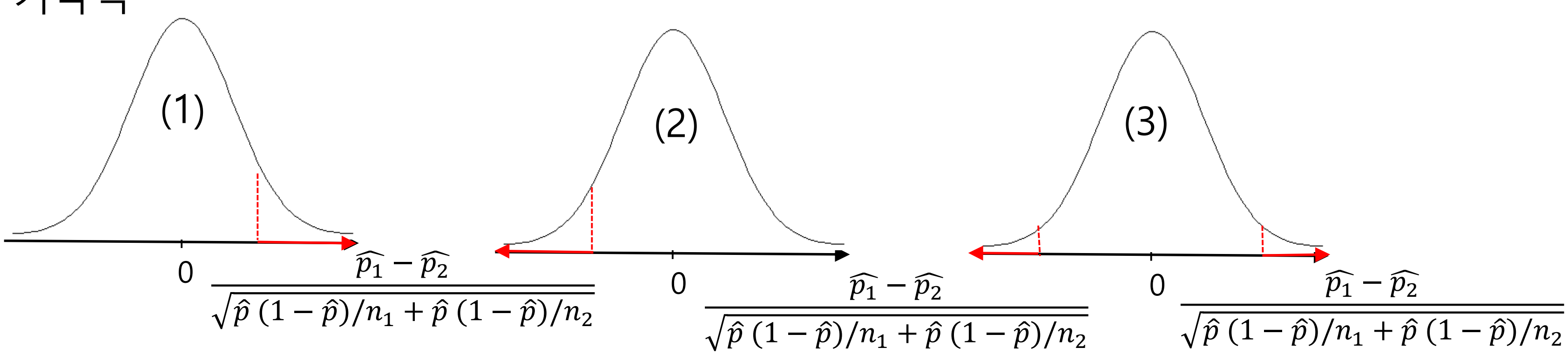
2. 검정통계량 (귀무가설 $H_0: p_1 = p_2 = p$)

$$\frac{\widehat{p_1} - \widehat{p_2}}{\sqrt{\hat{p}\,(1-\hat{p})/n_1 + \hat{p}\,(1-\hat{p})/n_2}} \approx \mathrm{N}(0,1)$$

단, $\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$: 공통모비율의 추정량=*합동표본비율*

$n_1\hat{p} > 5, n_1(1-\hat{p}) > 5, n_2\hat{p} > 5, n_2(1-\hat{p}) > 5$

3. 기각역



7-2 두 모비율의 비교

(예제) 한 범죄학 잡지에 실린 것으로서, 충동적 살인범과 계획적 살인범의 교화에 차이가 있는가를 알아 보기 위한 것이다. 일정기간 복역 후 가석방된 충동적 살인범과 계획적 살인범 중에서 각각 42명과 40명을 랜덤추출하여 가석방이 성공적인 경우, 즉 재범이 없는 경우와 실패한 경우의 도수를 관측한 결과가 아래와 같다. 살인범의 유형에 따라 가석방의 성공률에 차이가 있는가를 유의수준 5%에서 검정하여라.

	성공	실패	표본크기
충동적 살인범	13	29	n_1=42
계획적 살인범	22	18	n_2=40
합계	35	47	n=82

(풀이)

1. 가설 $H_0: p_1 = p_2 \;\; vs \;\; H_1: p_1 \neq p_2 \;\; (H_0: p_1 - p_2 = 0 \;\; vs \;\; H_1: p_1 - p_2 \neq 0)$

2. 유의수준 0.05

7-2 두 모비율의 비교

3. 검정통계량

$\hat{p} = \frac{13+22}{42+40}$=0.427 이므로 $n_1\hat{p} \geq 5, n_1(1-\hat{p}) \geq 5, n_2\hat{p} \geq 5, n_2(1-\hat{p}) \geq 5$이다. 따라서

$$Z=\frac{\widehat{p_1}-\widehat{p_2}}{\sqrt{\hat{p}\,(1-\hat{p})/n_1+\hat{p}\,(1-\hat{p})/n_2}} \approx \mathrm{N}(0,1)$$

4. 계산

$\widehat{p_1}$=13/42=0.310, $\widehat{p_2}$=22/40=0.550 이므로 $z^*=\frac{\widehat{p_1}-\widehat{p_2}}{\sqrt{\hat{p}\,(1-\hat{p})/n_1+\hat{p}\,(1-\hat{p})/n_2}} = -2.20$

5. 검정

$-z_{0.025}$=-1.96 > z^*=-2.20 이므로 귀무가설을 기각할 수 있다.

6. 결론

유의수준 5%에서 살인범의 유형에 따라 가석방의 성공률에 차이가 있다는 유의미한 증거가 된다.

7-3 범주형 자료에 의한 여러 모집단의 비교(동질성 검정)

❖ 각 모집단이 두 가지 이상의 서로 다른 속성을 갖는 개체들로 나뉘는 경우에 여러 모집단을 비교 하는 방법

- 가설 : H_0: 여러 모집단이 동일하다. vs H_1: 이들이 서로 다르다.
- 기각역 : 각 속성의 관찰 도수가 귀무가설 H_0하에서의 기대도수와 차이가 많이 나면 H_0를 기각하게 된다.

(기대도수 : 표본크기 × H_0하에서의 추정 확률)

(예제) 살인범의 유형에 따라 가석방의 성공률에 차이가 있는가?

가설: 충동적 살인범과 계획적 살인범의 모집단이 동일하다 vs 이들이 서로 다르다.

$$H_0: p_1 = p_2 \;\; vs \;\; H_1: p_1 \neq p_2$$

	성공	실패	표본크기
충동적 살인범	13(42 × 0.427=17.93)	29(42×0.573=24.07)	n_1=42
계획적 살인범	22(40×0.427=17.08)	18(40×0.573=22.92)	n_2=40

$$H_0: p_1 = p_2 = p \text{ 하에서의 추정값 } \hat{p} = \frac{13+22}{42+40} = 0.427$$

7-3 범주형 자료에 의한 여러 모집단의 비교(동질성 검정)

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

- 계산 : ${\chi_0}^2 = \frac{(13-17.93)^2}{17.93} + \cdots + \frac{(18-22.92)^2}{22.92} = 4.84$

이 값은 $z = \frac{\widehat{p_1}-\widehat{p_2}}{\sqrt{\hat{p}\,(1-\hat{p})/n_1+\hat{p}\,(1-\hat{p})/n_2}} = -2.20$의 제곱과 같다

- 검정 : $\chi^2 = 4.84 > {\chi^2}_{0.05}(1) = 3.84$ 이므로 귀무가설을 기각할만한 증거가 된다.

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

$X_1+X_2 = n$, $p_1+p_2=1$ 일 때, $\chi^2 = \frac{(X_1-np_1)^2}{np_1} + \frac{(X_2-np_2)^2}{np_2} = \frac{(X_1-np_1)^2}{np_1} + \frac{(n-X_1-n(1-p_1))^2}{n(1-p_1)}$

$$= \frac{(X_1-np_1)^2}{np_1} + \frac{(X_1-np_1)^2}{n(1-p_1)} = \frac{(X_1-np_1)^2}{np_1(1-p_1)} = \left(\frac{X_1-np_1}{\sqrt{np_1(1-p_1)}}\right)^2 \sim \chi^2(1)$$

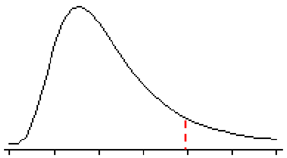
($\because$ 중심극한정리에 의해 $\frac{X_1-np_1}{\sqrt{np_1(1-p_1)}} \sim N(0,1)$ 이므로)

7-3 범주형 자료에 의한 여러 모집단의 비교(동질성 검정)

❖ r개의 이항모집단에 대한 관측 도수

	성공	실패	표본크기
모집단 1에서의 표본	O_{11}	O_{12}	n_1
모집단 2에서의 표본	O_{21}	O_{22}	n_2
… … …			
모집단 r에서의 표본	O_{r1}	O_{r2}	n_r
합계	$O_{.1}$	$O_{.2}$	n

1. $H_0: p_1 = p_2 \ldots = p_r = p \quad vs \quad H_1: H_0$가 아니다.
2. 합동표본비율 $\hat{p} = \frac{(O_{11}+O_{21}+\cdots+O_{r1})}{(n_1+n_2\ +\cdots+n_r)} = \frac{O_{.1}}{n}$
3. 기대도수 $\hat{E}_{i1} = n_i\hat{p} = n_i(\frac{O_{.1}}{n})$, $\hat{E}_{i2} = n_i(1-\hat{p}) = n_i(\frac{O_{.2}}{n})$ $(i=1,2,\ldots, r)$
4. 검정통계량 $\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{2}\frac{(O_{ij}-\hat{E}_{ij})^2}{\hat{E}_{ij}} \sim \chi^2(r\text{-}1)$ (단, H_0가 참이고 표본크기가 충분히 크면 $(\hat{E}_{ij} \geq 5)$)
5. 기각역



7-3 범주형 자료에 의한 여러 모집단의 비교(동질성 검정)

(예제) 액화천연가스(LNG)의 저장 기지의 후보지로 고려되고 있는 세 지역의 여론을 알아보기 위하여, 세 지역에서 각각 400명, 350명,350명을 랜덤추출하여 기지의 건설에 대한 찬성여부를 물은 결과가 아래와 같다. 지역에 따라 찬성률에 차이가 있다고 할 수 있는가?

	찬성	반대	표본크기
지역1	198(171.27)	202(228.73)	n_1=400
지역2	140(149.86)	210(200.14)	n_2=350
지역3	133(149.86)	217(200.14)	n_3=350
합계	471	629	n=1100

(풀이) 1. 가설 : H_0: $p_1 = p_2 = p_3$ vs H_1: H_0가 아니다.

2. 합동표본비율 $\hat{p} = \frac{(O_{11}+O_{21}+\cdots+O_{r1})}{(n_1+n_2\ +\cdots+n_r)} = \frac{O_{.1}}{n}$=471/1100

3. 카이제곱통계량 계산$\chi_0{}^2 = \sum_{i=1}^{r}\sum_{j=1}^{2}\frac{(O_{ij}-\hat{E}_{ij})^2}{\hat{E}_{ij}} = \frac{(198-171.27)^2}{171.27} + \ldots + \frac{(217-200.14)^2}{200.14}$=11.75

4. 검정 : p값$=P(\chi^2 \geq 11.75) = 0.003$, $\chi^2 \sim \chi^2$(3-1)

5. 결론 : 지역에 따라 LNG건설에 대한 찬성률에 차이가 있음을 뚜렷이 나타낸다.

7-3 범주형 자료에 의한 여러 모집단의 비교(동질성 검정)

❖ r개의 다항모집단에 대한 관측 도수

	범주1	범주2	…	범주c	표본크기
모집단 1에서의 표본	O_{11}	O_{12}	…	O_{1c}	n_1
모집단 2에서의 표본	O_{21}	O_{22}	…	O_{2c}	n_2
… … …			…		
모집단 r에서의 표본	O_{r1}	O_{r2}	…	O_{rc}	n_r
합계	$O_{.1}$	$O_{.2}$	…	$O_{.c}$	n

i번째 다항모집단의 각 범주에 대한 모비율을 $p_{i1}, p_{i2} \ldots, p_{ic}$ $(\sum_{j=1}^{c} p_{ij} = 1)$ 라고 하면 이들 다항모집단을 비교하기 위한 가설은 $H_0: p_{1j} = p_{2j} = \ldots = p_{rj}$ (j=1,2,…, c) vs H_1: H_0가 아니다.

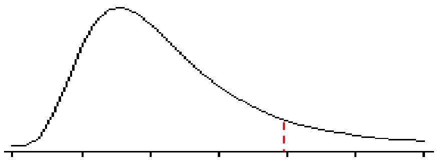
1. 합동표본비율 $\widehat{p_j} = \frac{(O_{1j}+O_{2j}+\cdots+O_{rj})}{(n_1+n_2 \ +\cdots+n_r)} = \frac{O_{.j}}{n}$

2. 기대도수 $\hat{E}_{i1} = n_i\widehat{p_1} = n_i(\frac{O_{.1}}{n})$, $\hat{E}_{i2} = n_i\widehat{p_2} = n_i(\frac{O_{.2}}{n})$ …, $\hat{E}_{i2} = n_i\widehat{p_c} = n_i(\frac{O_{.c}}{n})$ (i=1,2,…, r)

3. 검정통계량 $\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-\hat{E}_{ij})^2}{\hat{E}_{ij}} \sim \chi^2((r-1)(c-1))$ (단, H_0가 참이고 표본크기가 충분히 크면($\hat{E}_{ij} \geq 5$))

7-3 범주형 자료에 의한 여러 모집단의 비교(동질성 검정)

4. 기각역



(예제) 공단에 인접한 세 지역에서 공해를 느끼는 정도가 지역에 따라 차이가 있는가를 알아 보고자 하여 세 지역에서 각각 97명, 95명, 99명을 랜덤추출하여 산업 공해로 인한 악취를 느끼는 횟수에 대하여 조사한 결과가 아래와 같다. 범주 1은 매일, 범주 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.

	범주1	범주2	범주3	범주4	범주5	표본크기
지역1	20(12.67)	28(24.67)	23(18.00)	14(16.00)	12( )	n_1=97
지역2	14( )	34( )	21 ( )	14( )	12( )	n_2=95
지역3	4( )	12( )	10( )	20( )	53( )	n_3=99
합계	38	74	54	48	77	n=29`

7-3 범주형 자료에 의한 여러 모집단의 비교(동질성 검정)

풀이)

1. 가설 : H_0: p_{1j} = p_{2j} =…=p_{rj} (j=1,2,3,4,5) vs H_1: H_0가 아니다.

2. 각 범주의 모비율에 대한 추정값

$\widehat{p_j} = \frac{o_{.j}}{n}$ 이므로 $\widehat{p_1}$=38/291, $\widehat{p_2}$ = , $\widehat{p_3}$ = , $\widehat{p_4}$ = , $\widehat{p_5}$ =

3. H_0하에서 추정 기대도수는

$\hat{E}_{1j}$=97$\widehat{p_j}$, $\hat{E}_{2j}$=95$\widehat{p_j}$, $\hat{E}_{3j}$=99$\widehat{p_j}$ (j=1,2,3,4,5)

3. 카이제곱통계량 계산 $\chi_0{}^2=\frac{(20-12.67)^2}{12.67}+\ldots+\frac{(53-26.20)^2}{26.20}$=71.36

4. 검정 : 기각역이 $\chi_0{}^2 \geq \chi^2{}_{0.01}(8)$=20.09 인데 $\chi_0{}^2$=71.36은 기각역에 속하므로 귀무가설 기각

5. 결론 : 유의수준 1%에서 지역에 따라 공해를 느끼는 정도에 차이가 있다고 말할 수 있다.

7-4 범주형 자료에 의한 독립성의 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

(모비율)

	B_1	B_2	…	B_c	**합계**
A_1	P_{11}	P_{12}	…	P_{1c}	$P_{1.}$
A_2	P_{21}	P_{22}	…	P_{2c}	$P_{2.}$
⋮	⋮	⋮	…		
A_r	P_{r1}	P_{r2}	…	P_{rc}	$P_{r.}$
합계	$P_{.1}$	$P_{.2}$	…	$P_{.c}$	1

(관측도수)

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

1. 두 특성 A,B가 독립 - $H_0: p_{ij} = p_{i.}\,p_{.j}$ (i=1,2,…, r 이고 j=1,2,…, c) vs H_1: H_0가 아니다.

2. H_0하에서 (i,j)칸의 모비율 추정은 $\widehat{p_{ij}} = \widehat{p_{i.}}\widehat{p_{.j}} = \frac{O_{i.}}{n}\frac{O_{.j}}{n}$ (i=1,2,…, r 이고 j=1,2,…, c)

3. H_0하에서 (i,j)칸의 추정 기대도수는 $\hat{E}_{ij} = n\widehat{p_{ij}} = n\frac{O_{i.}}{n}\frac{O_{.j}}{n} = \frac{O_{i.}O_{.j}}{n}$ (i=1,2,…, r 이고 j=1,2,…, c)

4. 카이제곱통계량 $\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-\hat{E}_{ij})^2}{\hat{E}_{ij}} \sim \chi^2((r\text{-}1)(c\text{ -}1))$ (단, H_0가 참이고 표본크기가 충분히 크면 ($\hat{E}_{ij} \geq 5$))

7-4 범주형 자료에 의한 독립성의 검정

(예제) 대도시의 근교에서 출퇴근하며 혼자서만 승용차를 이용하는 사람들 중에서 250명을 랜덤 추출하여 이들을 승용차의 크기와 통근 거리에 따라 분류한 결과가 다음과 같다.

	15km미만	15km이상 30km미만	30km이상	합계
경승용차	6(10.19)	27( )	19( )	52
소형승용차	8(11.96)	36( )	17( )	61
중형승용차	21( )	45( )	33( )	99
대형승용차	14( )	18( )	6( )	38
합계	49	126	75	250

승용차의 크기와 통근 거리 사이에 관계가 있다고 할 수 있는지를 유의수준 5%에서 검정하여라.

(풀이) 1. 가설 $H_0 : p_{ij} = p_{i.}\, p_{.j}$ (i=1,2,3,4 이고 j=1,2,3) vs H_1: H_0가 아니다.

2. 추정기대도수 $\hat{E}_{ij} = \frac{O_{i.} O_{.j}}{n}$

3. 카이제곱통계량 계산 $\chi_0{}^2 =$

=14.15

7-4 범주형 자료에 의한 독립성의 검정

4. 검정 $\chi_0{}^2 = 14.15 \geq \chi^2{}_{0.05}$((4-1)(3-1))=12.59 이므로 H_0 기각

5. 결론

유의수준 5%에서 승용차의 크기와 통근 거리 사이에는 관계가 있다고 말할 수 있다.

❖ 다항모집단의 동질성 검정과 독립성 검정에서 검정통계량의 형태와 검정방법은 동일하지만, 표본의 추출과정이 다르다는 점에 유의하여 문제의 뜻에 맞게 표본을 추출하고 그에 맞는 가설을 세워 추론을 해야 한다.

7-5 적합도 검정(Goodness-of-fit test)

❖ 도수표로 나타낸 자료들이 이론적인 또는 가정된 모형과 일치하는지에 대한 검정

❖ 관찰된 도수와 이론적으로 기대하는 도수의 차이가 크면 귀무가설을 기각

(예제) 시장에 두 경쟁기업 A와 B가 있는데, 두 기업은 시장 내에서 절대적인 점유율을 갖고 있고 두 기업은 최근 공격적인 마케팅 캠페인을 실시하였다. 마케팅 캠페인 실시 이전의 기업들의 시장점유율:

Company A = 45%, Company B = 40%, Other competitors = 15%

마케팅 캠페인의 시장점유율에 대한 효과를 살펴보기 위해서, 점유율에 대한 조사를 실시하였는데, 200명의 고객이 광고가 된 상품에 대한 조사에 응하였고 그 결과

* 102명의 고객이 A사의 상품을 선호
* 82명의 고객이 B사의 상품을 선호
* 16명의 고객이 그 밖의 경쟁사의 상품을 선호

5%의 유의수준하에서 시장점유율이 마케팅 캠페인에 영향을 받았다고 할 수 있겠는가?

7-5 적합도 검정(Goodness-of-fit test)

1. 가설 H_0: p_A=0.45, p_B=0.4, p_C=0.15 vs H_1: H_0가 아니다.
2. 카이제곱 통계량 계산

$\chi_0{}^2 = \sum$(관찰도수$-$기대도수)²/ 기대도수 =

3. 검정 $\chi^2{}_{0.05}$(3-1)=5.99147, p값=$P(\chi^2 \geq 8.18)$=0.01679

$\chi_0{}^2$=8.18 > $\chi^2{}_{0.05}$(2)=5.99147이므로 귀무가설 기각 / p값이 0.05보다 희박하게 나타났으므로 귀무가설 기각할 만한 증거가 된다.

4. 결론 5%유의수준에서 시장점유율은 마케팅 캠페인에 영향을 받았다고 할 수 있다.