7장. 범주형자료의 분석

박명현

서울대학교 학부대학

카이제곱검정(Chi-Squared Tests)

- Three statistical techniques are presented, to analyze categorical data.
- Those tests use the χ^2 as the sampling distribution of the test statistic.

1. 적합도 검정(goodness of fit test)
   : 표본으로 얻어진 범주의 분포가 주어진 분포에 부합하는지 검정
2. 동질성 검정(test of homogeneity)
   : 두 개 이상의 집단에서의 분포가 같은지 검정
3. 독립성 검정(test of independence)
   : 범주형 변수들끼리의 독립성 여부를 검정

범주형 변수와 분할표

- 범주형 자료(categorical data)

  - 측정값이 몇 가지 범주(category) 중 하나로 기록되는 자료

  - 명목형, 순서형 자료

  예) 사람의 혈액형을 측정하는 경우 (A, B, AB, O)

  예) 직업에 대한 만족도에 대해 측정하는 경우 (매우 불만족, 다소 불만족, 다소 만족, 매우 만족)

- 분할표(contingency table)

  - 무작위로 추출한 20명에 대해 혈액형(4개의 범주)을 조사하여 표로 나타낸 자료

  - 범주형 자료의 도수분포표

	A	B	O	AB	합
인원수	6	8	4	2	20
비율	0.3	0.4	0.2	0.1	1

범주형 변수와 분할표

- 분할표(contingency table)

  - 무작위로 추출한 20명에 대해 혈액형(4개의 범주)과 성별(2개의 범주)을 조사하여 표로 나타낸 자료

	A	B	O	AB	합
F	4 (0.2)	4 (0.2)	1 (0.05)	1 (0.05)	10 [0.5]
M	2 (0.1)	4 (0.2)	3 (0.15)	1 (0.05)	10 [0.5]
합	6 [0.3]	8 [0.4]	4 [0.2]	2 [0.1]	20

  - ( )안의 숫자는 결합확률(joint probability)
  - [ ]안의 숫자는 주변확률(marginal probability)

- 위 표본결합분포로 모집단의 결합분포를 추정할 수 있다.
- 표본에서의 연관성 여부가 완전한 모집단에서의 연관성에 대한 증거가 될 수 있는가?

다항분포(multinomial distribution)

Definition (Multinomial Distribution)

하나의 시행에서 여러 범주 중 하나가 일어날 수 있을 때, 이를 독립적으로 n번 반복했을 때 각 범주의 발생 횟수를 나타내는 확률분포를 다항분포, $(X_1, X_2, \ldots, X_k) \sim Multinomial(n; p_1, p_2, \ldots, p_k)$라고 하고, 이의 확률질량함수는 다음과 같다.

$$P(X_1 = x_1, X_2 = x_2, \cdots, X_k = x_k) = \frac{n!}{x_1! x_2! \cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k},$$

- k: 범주의 개수, n: 전체 시행 횟수
- p_i: i번째 범주의 발생 확률, $p_i \geq 0$ 이며 $\sum_{i=1}^{k} p_i = 1$
- X_i: i번째 범주의 관측 횟수, $\sum_{i=1}^{k} X_i = n$
- $E(X_i) = np_i$, $Var(X_i) = np_i(1 - p_i)$

χ^2 Tests

Theorem

만약 $X_1, X_2, \ldots, X_k$ 가 다항분포 $M(n, p_1, p_2, \ldots, p_k)$를 따르고, 표본의 크기 n이 충분히 커서 $np_i \geq 5$ $(i = 1, 2, \ldots, k)$를 만족한다면,

$$\sum_{i=1}^{k} \left(\frac{(X_i - E(X_i))^2}{E(X_i)} \right) = \sum_{i=1}^{k} \left(\frac{(X_i - np_i)^2}{np_i} \right) \sim \chi^2(k-1)$$

근사적으로 자유도 $k-1$인 카이제곱분포를 따른다.

적합도 검정(goodness of fit test) -모집단의 분포가 알려진 경우-

- 주거형태, 주식의 등급, 사고발생 요일 등 하나의 범주형 변수의 경우 각 범주에 속하는 분포에 관심이 있을 수 있다.
- 과거자료 또는 추측에 의해 가정할 수 있는 분포를 실제 표본과 비교하여 가정된 분포가 맞는지 검정한다.

- 적합도 검정(GOF, goodness of fit test):

H_0: 표본은 가정된 분포를 따른다.

H_1 : 표본은 가정된 분포를 따르지 않는다.

- 가정된 분포가 적합한지 검정하는 것.

- 각 범주의 관측도수(observed frequency)와 H_0하에서의 추정 기대도수(estimated expected frequency)의 차이를 비교한다.

적합도 검정(goodness of fit test) -모집단의 분포가 알려진 경우-

예) 다음의 표는 제품 4개에 대해 알려진 시장점유율과 실제 500명을 대상으로 사용하고 있는 제품의 사용자수를 나타낸 것이다.

Table: 제품의 시장점유율

제품	A	B	C	D	합계
시장점유율	20%	40%	10%	30%	100%
표본수	110	195	47	148	500

- k: 전체 범주의 수
- p_i: 전체 k개의 범주 중 i번째 범주의 기대비율
- O_i: 실제 표본으로부터 얻어진 i번째 범주의 관측(observed)도수
- E_i: 표본수 n에 대해 i번째 범주에 기대되는(expected) 갯수
- i번째 범주의 기댓값: $E_i = n \times p_i$

적합도 검정(goodness of fit test)

- 적합도 검정의 가설:

$$H_0 : p_1 = 0.2,\ p_2 = 0.4,\ p_3 = 0.1,\ p_4 = 0.3$$

H_1: At least one p_i is not equal to its specified value

- 분포에 대한 가정: $X_1, X_2, \cdots, X_k \sim M(n, p_1, p_2, \cdots, p_k)$
- Under H_0,

  - 적합도 검정 통계량의 근사분포:

$$\chi^2 = \sum_{i=1}^{k} \left(\frac{(O_i - E_i)^2}{E_i} \right) \dot{\sim} \chi^2(k-1)$$

- 기댓값(E_i)과 관측값(O_i)이 차이가 많이나면 $(O_i - E_i)^2$ 값과 검정통계량의 관측값이 커지게 되고, H_0을 벗어난다고 할 수 있다.

적합도 검정(goodness of fit test)

- 검정통계량:

$$\chi^2 = \sum_{i=1}^{k} \left(\frac{(O_i - E_i)^2}{E_i} \right) = \sum_{i=1}^{k} \left(\frac{(O_i - np_i)^2}{np_i} \right) \dot{\sim} \chi^2(k-1)$$

이의 관측값을 χ_0^2이라고 하면 다음이 성립한다.

유의확률	유의수준 α의 기각역
$P = P(\chi^2 \geq \chi_0^2),\ \chi^2 \sim \chi^2(k-1)$	$\chi_0^2 \geq \chi_\alpha^2(k-1)$

- 이 때, 각 범주에서의 기대값 E_i이 5 이상으로 가정하고, 이를 만족하지 않으면 범주를 통합하여 검정한다.

적합도 검정(goodness of fit test)

예) (continue..) Under H_0, 기댓값과 실제 관측값의 차이를 계산하면

제품	A	B	C	D	합계
시장점유율(p_i)	20%	40%	10%	30%	100%
제품수(O_i)	110	195	47	148	500
기댓값(E_i)	100	200	50	150	500
$\frac{(O_i-E_i)^2}{E_i}$	1	0.125	0.18	0.027	1.33

- 검정통계량의 관측값 $\chi_0^2 = \sum_{i=1}^{k} \left(\frac{(O_i - np_i)^2}{np_i} \right) = 1.33$이고
- $0.5 < P(\chi^2 \geq \chi_0^2) < 0.75$이므로 H_0를 기각할 수 없다.
- 따라서 표본자료는 알려진 시장점유율을 통계적으로 부정할 만한 증거가 되지 못한다.

동질성 검정(test of homogeneity)

- 동질성 검정: 두 개 이상의 집단에서 분포가 같은지 여러 모집단을 비교, 분석하는 방법.
- r개의 모집단과 c개의 범주(category) 가정
- 동질성검정의 통계적 가설

  H_0: 여러 모집단의 분포가 동일하다.

  H_1 : 이들이 서로 다르다.

- 위 가설에 대한 검정은 카이제곱분포에 근사하여 수행.
- 각 범주의 관측도수(observed)가 귀무가설 H_0하에서의 추정 기대도수(expected)와 차이가 많이 나면 H_0를 기각.

동질성 검정(test of homogeneity)

- c개의 범주로 나뉘어진 r개의 다항모집단을 관측한 결과가 다음과 같이 $r \times c$ 분할표(contingency table)로 주어져 있다.

Table: r개의 다항모집단에 대한 관측도수(O_{ij})

	범주 1	범주 2	$\cdot$	범주 j	$\cdot$	범주 c	표본크기
모집단 1의 표본	O_{11}	O_{12}	$\cdot$	O_{1j}	$\cdot$	O_{1c}	n_1
$\cdots$	$\cdots$	$\cdot$	$\cdot$	$\cdots$	$\cdot$	$\cdots$	$\cdots$
모집단 i의 표본	O_{i1}	O_{i2}	$\cdot$	O_{ij}	$\cdot$	O_{ic}	n_i
$\cdots$	$\cdots$	$\cdots$	$\cdot$	$\cdots$	$\cdot$	$\cdots$	$\cdots$
모집단 r의 표본	O_{r1}	O_{r2}	$\cdot$	O_{rj}	$\cdot$	O_{rc}	n_r
합계	$O_{\cdot 1}$	$O_{\cdot 2}$	$\cdot$	$O_{\cdot j}$	$\cdot$	$O_{\cdot c}$	n

- O_{ij} = 표에서 (i,j)칸의 관측도수(observed frequency)(or 표본수)
- $O_{\cdot j} = \sum_{i=1}^{r} O_{ij}$ (j번째 열의 합)
- $O_{i\cdot} = \sum_{j=1}^{c} O_{ij} = n_i$ (i번째 행의 합)

동질성 검정(test of homogeneity)

Table: r개의 다항모집단에 대한 분할표 (i,j)칸에서의 모비율 p_{ij}

	범주 1	범주 2	·	범주 j	·	범주 c
모집단 1의 표본	p_{11}	p_{12}	·	p_{1j}	·	p_{1c}
모집단 2의 표본	p_{21}	p_{22}	·	p_{2j}	·	p_{2c}
...	...	·	·	...	·	...
모집단 i의 표본	p_{i1}	p_{i2}	·	p_{ij}	·	p_{ic}
...	...	...	·	...	·	...
모집단 r의 표본	p_{r1}	p_{r2}	·	p_{rj}	·	p_{rc}
추정 모비율	$\hat{p}_1$	$\hat{p}_2$	·	$\hat{p}_j$	·	$\hat{p}_c$
	$= \frac{O_{\cdot 1}}{n}$	$= \frac{O_{\cdot 2}}{n}$	·	$= \frac{O_{\cdot j}}{n}$	·	$= \frac{O_{\cdot c}}{n}$

- $p_j (= p_{1j} = \cdots = p_{rj})$ (Under H_0, 각 범주의 공통 모비율)
- $\hat{p}_j = \frac{O_{\cdot j}}{n}$ (범주 j에서 공통 모비율의 추정값)
- $\sum_{j=1}^{c} p_{ij} = 1$ (i^{th} 다항모집단의 각 범주에 대한 모비율의 합은 1)

동질성 검정(test of homogeneity)

Table: r개의 다항모집단에 대한 기대도수($\hat{E}_{ij}$)

	범주 1	범주 2	$\cdot$	범주 j	$\cdot$	범주 c	표본크기
모집단 1의 표본	$\hat{E}_{11}$	$\hat{E}_{12}$	$\cdot$	$\hat{E}_{1j}$	$\cdot$	$\hat{E}_{1c}$	n_1
모집단 2의 표본	$\hat{E}_{21}$	$\hat{E}_{22}$	$\cdot$	$\hat{E}_{2j}$	$\cdot$	$\hat{E}_{2c}$	n_2
$\cdots$	$\cdots$	$\cdot$	$\cdot$	$\cdots$	$\cdot$	$\cdots$	$\cdots$
모집단 i의 표본	$\hat{E}_{i1}$	$\hat{E}_{i2}$	$\cdot$	$\hat{E}_{ij}$	$\cdot$	$\hat{E}_{ic}$	n_i
$\cdots$	$\cdots$	$\cdots$	$\cdot$	$\cdots$	$\cdot$	$\cdots$	$\cdots$
모집단 r의 표본	$\hat{E}_{r1}$	$\hat{E}_{r2}$	$\cdot$	$\hat{E}_{rj}$	$\cdot$	$\hat{E}_{rc}$	n_r
추정 모비율	$\hat{p}_1$	$\hat{p}_2$	$\cdot$	$\hat{p}_j$	$\cdot$	$\hat{p}_c$	

- $\hat{E}_{ij}$: H_0하에서의 추정 기대도수(estimated expected frequency)
- $\hat{E}_{ij} = n_i \times \hat{p}_j$($H_0$하에서의 추정 확률)

동질성 검정(test of homogeneity)

- 동질성 검정의 가설

$$H_0 : p_{1j} = p_{2j} = \cdots = p_{rj} \ (j = 1, 2, \cdots, c), \quad H_1 : not \ H_0$$

- 분포에 대한 가정:

$$X_{i1}, X_{i2}, \cdots, X_{ic} \ \sim \ M(n_i, p_{i1}, p_{i2}, \cdots, p_{ic}), \quad i = 1, \cdots, r$$

- Under H_0, p_j의 추정값은 다음과 같이 합동표본비율로 주어지고

$$\hat{p}_j = \frac{O_{1j} + O_{2j} + \cdots + O_{rj}}{n_1 + n_2 + \cdots + n_r} = \frac{O_{\cdot j}}{n}$$

- i번째 집단에서의 추정 기대도수는 다음과 같다.$(i = 1, \cdots, r)$

$$\hat{E}_{i1} = n_i \hat{p}_1 = n_i \frac{O_{\cdot 1}}{n}, \quad \hat{E}_{i2} = n_i \hat{p}_2 = n_i \frac{O_{\cdot 2}}{n}, \quad \cdots, \quad \hat{E}_{ic} = n_i \hat{p}_c = n_i \frac{O_{\cdot c}}{n}$$

동질성 검정(test of homogeneity)

- 관측도수 O_{ij}와 추정 기대도수 $\hat{E}_{ij}$의 차이를 나타내는 다음의 카이 제곱통계량은

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

  Under H_0, 표본크기가 충분히 크면 근사적으로 다음과 같이 카이 제곱분포를 따른다고 알려져 있다.

$$\chi^2 \dot{\sim} \chi^2((r-1)(c-1))$$

- 각 범주의 관측도수 O_{ij}가 귀무가설 H_0하에서의 추정 기대도수 $\hat{E}_{ij}$와 차이가 많이 나면 χ^2 관측값도 커지게 되고 H_0를 기각하게 된다.

동질성 검정(표본크기가 큰 경우: $\hat{E}_{ij} \geq 5$)

- c개의 범주로 나뉘어지는 다항모집단 r개를 비교하기 위한 가설

$$H_0 : p_{1j} = p_{2j} = \cdots = p_{rj} \ (j = 1, 2, \cdots, c) \quad H_1 : not\ H_0$$

- 검정통계량:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}} \quad (\hat{E}_{ij} = n_i O_{\cdot j}/n)$$

이의 관측값을 χ_0^2이라고 하면 다음이 성립한다.

유의확률	유의수준 α의 기각역
$P = P(\chi^2 \geq \chi_0^2),\ \chi^2 \sim \chi^2((r-1)(c-1))$	$\chi_0^2 \geq \chi_\alpha^2((r-1)(c-1))$

동질성 검정(test of homogeneity)

예) [예제 7-7]에서 충동적 살인범과 계획적 살인범의 가석방 성공률에 차이가 있는지 유의수준 5%에서 검정하였다.

$$H_0 : p_1 = p_2, \quad H_1 : p_1 \neq p_2$$

- 동질성 검정: 충동적 살인범과 계획적 살인범의 모집단이 동일하다.

$$H_0 : p_{11} = p_{21},\ p_{12} = p_{22} \quad H_1 : not\ H_0$$

	성공(j=1)	실패(j=2)	표본크기
충동적 살인범(i=1)	O_{11} =13	O_{12} =29	$n_1 = 42$
계획적 살인범(i=2)	O_{21} =22	O_{22} =18	$n_2 = 40$
합계	35	47	$n = 82$

동질성 검정(test of homogeneity)

- Under H_0, $\hat{p}_1 = 35/82 = 0.427$, $\hat{p}_2 = 47/82 = 0.573$이므로,

  - 충동적 살인범에 대한 가석방이 성공할 $\hat{E}_{11}$와 실패할 $\hat{E}_{12}$는

$$\hat{E}_{11} = n_1\hat{p}_1 = 42 \times 35/82 = 17.93$$
$$\hat{E}_{12} = n_1\hat{p}_2 = 42 \times 47/82 = 24.07$$

  - 계획적 살인범에 대한 가석방이 성공할 $\hat{E}_{21}$와 실패할 $\hat{E}_{22}$는

$$\hat{E}_{21} = n_2\hat{p}_1 = 40 \times 35/82 = 17.08$$
$$\hat{E}_{22} = n_2\hat{p}_2 = 40 \times 47/82 = 22.92$$

- 추정 기대도수와 관측도수를 표로 정리하면 다음과 같다.

	성공	실패
충동적 살인범	$O_{11} = 13(\hat{E}_{11} = 17.93)$	$O_{12} = 29(\hat{E}_{12} = 24.07)$
계획적 살인범	$O_{21} = 22(\hat{E}_{21} = 17.08)$	$O_{22} = 18(\hat{E}_{22} = 22.92)$
합계	35	47

동질성 검정(test of homogeneity)

- 카이제곱통계량 χ^2의 관측값을 구해보자.

$$\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-\hat{E}_{ij})^2}{\hat{E}_{ij}} = \frac{(13-17.93)^2}{17.93} + \cdots + \frac{(18-22.92)^2}{22.92} = 4.84$$

- Under H_0, 검정통계량 χ^2는 근사적으로 $\chi^2(1)$을 따르고, 유의수준 5%에서 기각역은 다음과 같아서 귀무가설을 기각하게 된다.

$$\chi^2 \geq \chi^2_{0.05}(1) = 3.84$$

Z 검정통계량과의 비교($r = c = 2$)

- 그런데 카이제곱통계량 χ^2의 관측값은 다음의 예제 7.7에서의 Z 검정통계량의 제곱과 같다.

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(n_1^{-1} + n_2^{-1})}} = -2.20$$

- 즉, 다음의 관계가 일반적으로 성립하며,

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

H_0하에서 Z가 근사적으로 표준정규분포를 따르므로 그것의 제곱꼴은 자유도가 1인 카이제곱분포를 따르게 됨을 알 수 있다.

다항모집단의 동질성검정 예제

Example

액화천연가스(LNG) 저장 기지의 후보로 고려되는 세 지역의 여론을 조사하기 위해 각 지역에서 각각 400명, 350명, 350명을 랜덤추출하여 기지건설에 대한 찬반을 조사했다. 조사결과가 [표 7-5]와 같을 때, 지역에 따라 찬성률에 차이가 있다고 할 수 있는지 5% 유의수준으로 검정하여라.

표 7–5 LNG 저장 기지 건설에 대한 찬반 조사 결과

(단위 : 명)

	찬성	반대	표본 크기
지역 1	198	202	$n_1 = 400$
지역 2	140	210	$n_2 = 350$
지역 3	133	217	$n_3 = 350$
합계	471	629	$n = 1100$

sol) 세 지역의 찬성률 p_{11}, p_{21}, p_{31}에 차이가 있는지 검정하기 위해 다음과 같은 가설을 세운다.

$$H_0 : p_{11} = p_{21} = p_{31}, \qquad H_1 : not\ H_0$$

- H_0하에서 $\hat{p}_1 = 471/1100,\ \hat{p}_2 = 629/1100$이고,

- 추정 기대도수를 구해보면

$$\hat{E}_{11} = 400 \times (471/1100) = 171.27, \quad \hat{E}_{12} = 400 \times (629/1100) = 228.73$$

$$\hat{E}_{21} = 350 \times (471/1100) = 149.86, \quad \hat{E}_{22} = 350 \times (629/1100) = 200.14$$

$$\hat{E}_{31} = 350 \times (471/1100) = 149.86, \quad \hat{E}_{32} = 350 \times (629/1100) = 200.14$$

- 이로부터 카이제곱통계량의 관측값을 계산해보자.

$$\sum_{i=1}^{3}\sum_{j=1}^{2}\frac{(O_{ij}-\hat{E}_{ij})^2}{\hat{E}_{ij}} = \frac{(198-171.27)^2}{171.27}+\cdots+\frac{(217-200.14)^2}{200.14} = 11.75$$

- P값은 다음과 같고 카이제곱분포표로부터 유의수준 5% 이하인 것을 알 수 있다.

$$0.0025 < P = P(\chi^2 \geq 11.75) < 0.005, \quad \chi^2 \sim \chi^2(2)$$

- 즉 지역에 따라 LNG기지 건설에 대한 찬성률에 차이가 있음을 뚜렷이 나타낸다고 할 수 있다.

다항모집단의 동질성검정 예제

Example

공단에 인접한 세 지역에서 공해를 느끼는 정도가 지역에 따라 차이가 있는지를 확인하기 위해, 각 지역에서 97명, 95명, 99명을 랜덤추출하여 산업공해로 인한 악취를 느끼는 빈도를 조사한 결과가 다음 표와 같다. 표에서 범주는 다음과 같이 정의된다.

* 범주 1: 매일 * 범주 2: 적어도 일주일에 한 번 * 범주 3: 적어도 한달에 한 번

* 범주 4: 한 달에 한 번보다 적게 * 범주 5: 전혀 악취를 느끼지 않음

이 자료를 이용하여 지역에 따라 공해를 느끼는 정도가 다르다고 할 수 있는지를 유의수준 1%에서 검정하여라.

표 7–7 지역별 공해 인식 빈도

(단위 : 명)

	범주 1	범주 2	범주 3	범주 4	범주 5	표본 크기
지역 1	20	28	23	14	12	$n_1 = 97$
지역 2	14	34	21	14	12	$n_2 = 95$
지역 3	4	12	10	20	53	$n_3 = 99$
합계	38	74	54	48	77	$n = 291$

- 지역 i의 각 범주 j에 대한 모비율을 p_{ij} $(i = 1, 2, 3,\ j = 1, 2, 3, 4, 5)$이라고 하면, 가설은 다음과 같다.

$$H_0 : p_{1j} = p_{2j} = p_{3j} (j = 1, 2, \cdots, 5) \quad H_1 : not\ H_0$$

- H_0하에서 공통인 각 범주의 모비율에 대한 추정값은

$$\hat{p}_1 = 38/291, \hat{p}_2 = 74/291, \hat{p}_3 = 54/291, \hat{p}_4 = 48/291, \hat{p}_5 = 77/291$$

- 이를 이용하여, H_0하에서의 추정 기대도수는 다음과 같고, 아래 표에 그 값이 정리가 되어 있다.

$$\hat{E}_{1j} = 97\hat{p}_j, \hat{E}_{2j} = 95\hat{p}_j, \hat{E}_{3j} = 99\hat{p}_j \quad (j = 1, 2, \cdots, 5)$$

	범주 1	범주 2	범주 3	범주 4	범주 5
지역 1	12.67	24.67	18.00	16.00	25.67
지역 2	12.41	24.16	17.63	15.67	25.14
지역 3	12.93	25.18	18.37	16.33	26.20

- 카이제곱통계량의 관측값을 계산해보자.

$$\chi_0^2 = \sum_{i=1}^{3}\sum_{j=1}^{5} \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}} = \frac{(20 - 12.67)^2}{12.67} + \cdots + \frac{(53 - 26.20)^2}{26.20}$$

$$= 71.36$$

- 모든 기대도수가 5 이상이므로 $\chi^2 \sim \chi^2((3-1)(5-1))$

- 유의수준 1%의 기각역은 카이제곱분포표로부터 다음과 같고

$$\chi_0^2 \geq \chi_{0.01}^2(8) = 20.09$$

관측값이 기각역에 속하므로, 지역에 따라 공해를 느끼는 정도에 차이가 있다고 유의수준 1%에서 말할 수 있다.

범주형 자료에 의한 독립성 검정

- 한 모집단의 각 개체에 대하여 두가지 특성 A, B를 관측
  예) 혈액형과 성별, 학년과 학점 등 두 개의 범주형 변수를 관측
  $\Longrightarrow$ 이들 특성 간의 관련성이 주된 관심.
- 독립성 검정: 한개의 모집단에서 각 개체에 대해 두 특성 A, B를 관측한 경우 이들 특성이 서로 독립(or 관련성)인지 검정하는 방법
  - 사건 A와 사건 B가 독립이면 다음이 성립

$$P(A \cap B) = P(A)P(B)$$

- 동질성 검정: r개의 모집단에서 하나의 범주형 특성을 관측한 경우, 여러 모집단 별 범주형 특성의 분포가 같은지 검정하는 방법

범주형 자료에 의한 독립성 검정

- 특성 A, B가 각각 r개와 c개의 범주로 나뉘어지는 경우에 두 특성에 대한 관측도수는 다음의 $r \times c$분할표로 주어진다.

Table: 두 특성 A, B의 관측도수(O_{ij})

	B_1	B_2	$\cdots$	B_j	$\cdots$	B_c	합계
A_1	O_{11}	O_{12}	$\cdots$	O_{1j}	$\cdots$	O_{1c}	$O_{1\cdot}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
A_i	O_{i1}	O_{i2}	$\cdots$	O_{ij}	$\cdots$	O_{ic}	$O_{i\cdot}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
A_r	O_{r1}	O_{r2}	$\cdots$	O_{rj}	$\cdots$	O_{rc}	$O_{r\cdot}$
합계	$O_{\cdot 1}$	$O_{\cdot 2}$	$\cdots$	$O_{\cdot j}$	$\cdots$	$O_{\cdot c}$	n

독립성 검정(test of independence)

Table: 두 특성 A, B의 모비율(p_{ij})

	B_1	B_2	$\cdots$	B_j	$\cdots$	B_c	합계
A_1	p_{11}	p_{12}	$\cdots$	p_{1j}	$\cdots$	p_{1c}	$p_{1\cdot}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
A_i	p_{i1}	p_{i2}	$\cdots$	p_{ij}	$\cdots$	p_{ic}	$p_{i\cdot}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
A_r	p_{r1}	p_{r2}	$\cdots$	p_{rj}	$\cdots$	p_{rc}	$p_{r\cdot}$
합계	$p_{\cdot 1}$	$p_{\cdot 2}$	$\cdots$	$p_{\cdot j}$	$\cdots$	$p_{\cdot c}$	1

- 두 특성 A, B가 서로 독립인 경우: $p_{ij} = p_{i\cdot}p_{\cdot j}$
- 분할표 (i, j)셀에서의 추정된 모비율

$$\hat{p}_{ij} = \hat{p}_{i\cdot}\hat{p}_{\cdot j} = (O_{i\cdot}/n)(O_{\cdot j}/n)$$

독립성 검정(test of independence)

Table: 두 특성 A, B의 기대도수($\hat{E}_{ij}$)

	B_1	B_2	$\cdots$	B_j	$\cdots$	B_c	합계
A_1	$\hat{E}_{11}$	$\hat{E}_{12}$	$\cdots$	$\hat{E}_{1j}$	$\cdots$	$\hat{E}_{1c}$	$O_{1\cdot}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
A_i	$\hat{E}_{i1}$	$\hat{E}_{i2}$	$\cdots$	$\hat{E}_{ij}$	$\cdots$	$\hat{E}_{ic}$	$O_{i\cdot}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
A_r	$\hat{E}_{r1}$	$\hat{E}_{r2}$	$\cdots$	$\hat{E}_{rj}$	$\cdots$	$\hat{E}_{rc}$	$O_{r\cdot}$
합계	$O_{\cdot 1}$	$O_{\cdot 2}$	$\cdots$	$O_{\cdot j}$	$\cdots$	$O_{\cdot c}$	n

- 독립일 때 (i,j)셀에서의 추정 기대도수

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

범주형 자료에 의한 독립성 검정

- O_{ij}와 $\hat{E}_{ij}$를 이용하여 다음의 카이제곱통계량을 구할 수 있고

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

- H_0가 사실일 때 표본크기가 충분히 크면 근사적으로 다음과 같이 자유도가 $(r-1)(c-1)$인 카이제곱분포를 따른다고 알려져 있다.

$$\chi^2 \dot{\sim} \chi^2((r-1)(c-1))$$

- 동질성 검정과 독립성 검정

  : 두 검정은 검정통계량의 형태와 검정방법은 동일하지만, 가설 및 표본의 추출 형태가 서로 다르므로 문제의 뜻에 맞게 표본을 추출하고 그에 맞는 추론을 하여야 한다.

두 특성의 독립성 검정(표본크기가 큰 경우: $\hat{E}_{ij} \geq 5$)

- 독립성 검정에 대한 가설:

$$H_0 : p_{ij} = p_{i\cdot}p_{\cdot j} \quad (i = 1, \cdots, r)(j = 1, \cdots, c), \quad H_1 : not\ H_0$$

- 분포에 대한 가정:

$$X_{11}, \cdots, X_{rc} \sim M(n, p_{11}, \cdots, p_{rc})$$

- 검정통계량:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}} \quad (\hat{E}_{ij} = O_{i\cdot}O_{\cdot j}/n)$$

- 검정통계량의 관측값을 χ_0^2이라고 하면 다음이 성립한다.

유의확률

$$P = P(\chi^2 \geq \chi_0^2),\ \chi^2 \sim \chi^2((r-1)(c-1))$$

유의수준 α의 기각역

$$\chi_0^2 \geq \chi_\alpha^2((r-1)(c-1))$$

두 특성의 독립성 검정 예제

Example

대도시 근교에서 출퇴근하며 혼자 승용차를 이용하는 사람들 중 250명을 랜덤추출하여 이들을 승용차 크기와 통근 거리로 분류한 결과가 [표 7-10]과 같다. 이 자료를 이용하여 승용차 크기와 통근 거리 사이 관계가 있는지를 유의수준 5%에서 검정하여라.

표 7-10 승용차 크기와 통근 거리

(단위 : 대)

	15km 미만	15km 이상 30km 미만	30km 이상	합계
경승용차	6	27	19	52
소형승용차	8	36	17	61
중형승용차	21	45	33	99
대형승용차	14	18	6	38
합계	49	126	75	250

sol) 표에서 (i,j)칸에 대응하는 모비율을 p_{ij}라고 하면, 가설은 다음과 같다.

$$H_0 : p_{ij} = p_{i\cdot}p_{\cdot j} \quad (i = 1, \cdots, 4)(j = 1, 2, 3), \quad H_1 : not\ H_0$$

- H_0하에서 각 칸의 추정 기대도수를 구하면

$$\hat{E}_{11} = \frac{52 \cdot 49}{250} = 10.19, \ \hat{E}_{12} = \frac{52 \cdot 126}{250} = 26.21, \ \hat{E}_{13} = \frac{52 \cdot 75}{250} = 15.60$$

$$\cdots \qquad \cdots \qquad \cdots$$

$$\hat{E}_{41} = \frac{38 \cdot 49}{250} = 7.45, \quad \hat{E}_{42} = \frac{38 \cdot 126}{250} = 19.15, \ \hat{E}_{43} = \frac{38 \cdot 75}{250} = 11.40$$

- 다음으로 카이제곱통계량의 관측값을 계산해보자.

$$\chi_0^2 = \sum_{i=1}^{4}\sum_{j=1}^{3} \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}} = \frac{(6 - 10.19)^2}{10.19} + \cdots + \frac{(6 - 11.40)^2}{11.40}$$

$$= 14.15$$

- 모든 기대도수가 5 이상이므로 $\chi^2 \sim \chi^2((4-1)(3-1))$

- 유의수준 5%의 기각역은 카이제곱분포표로부터 다음과 같고

$$\chi_0^2 \geq \chi_{0.05}^2(6) = 12.59$$

관측값이 기각역에 속하므로, 유의수준 5%에서 승용차의 크기와 통근 거리 사이에는 관계가 있다고 말할 수 있다.