

제 7장. 이산자료의 분석

7.1 모비율의 추정과 검정

한 모비율에 대한 추론 :

$X \sim B(n, p)$ 일 때 모비율 p 의 추정량: $\hat{p} = \frac{X}{n}$.

p (또는 np)에 대한 (정규근사에 의한) 추론:

$$Z = \frac{X - np}{\sqrt{np(1-p)}} = \frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}} \simeq N(0, 1)$$

단, 위의 정규근사는 $np \geq 5$ 이고 $n(1-p) \geq 5$ 일 때 성립한다.

두 모비율의 차에 대한 추론 :

• $X_1 \sim B(n_1, p_1)$, $X_2 \sim B(n_2, p_2)$ 일 때 $p_1 - p_2$ 의 추정량: $\hat{p}_1 - \hat{p}_2$.

• $p_1 - p_2$ 의 $100(1 - \alpha)\%$ (근사) 신뢰구간:

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

• $H_0 : p_1 = p_2$ 의 검정통계량 (pooled proportion 사용):

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}, \quad \hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

7.2 범주형 자료에 의한 여러 모집단의 비교

[$r \times c$ 분할표의 분석]

$A \backslash B$	B_1	B_2	\dots	B_c	계
A_1	O_{11}	O_{12}	\dots	O_{1c}	$O_{1.}$
A_2	O_{21}	O_{22}	\dots	O_{2c}	$O_{2.}$
\vdots	\vdots	\vdots		\vdots	\vdots
A_r	O_{r1}	O_{r2}	\dots	O_{rc}	$O_{r.}$
	$O_{.1}$	$O_{.2}$	\dots	$O_{.c}$	n

동질성 검정 모형: $O_{1.} = n_1, O_{2.} = n_2, \dots, O_{r.} = n_r$ 은 미리 정해진 상수. (A_1, A_2, \dots, A_r 은 부차모집단)

- $H_0 : p_{1j} = p_{2j} = \dots p_{rj} \ (j = 1, 2, \dots, c)$

독립성 검정 모형: $O_{1.}, O_{2.}, \dots, O_{r.}$ 은 확률변수.

- $H_0 : p_{ij} = p_{i.}p_{.j} \ (i = 1, 2, \dots, r) \ (j = 1, 2, \dots, c)$

검정통계량:

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

기각역: $\chi^2 \geq \chi^2_{\alpha}((r-1)(c-1))$

예 1) 성별에 따라 국어, 수학, 영어 세 과목에 대한 선호도가 다른가를 조사하고자 한다. 남학생 250명과 여학생 250명을 랜덤 추출하여 가장 좋아하는 한 과목을 택하게 하여 분류한 결과가 아래의 표와 같을 때, 남학생과 여학생에 따른 과목의 선호도가 다르다고 할 수 있는지 유의수준 5%에서 검정하여 보자.

선호도	국어	수학	영어	계
남학생	73	98	79	250
여학생	82	58	110	250
계	155	156	189	500

[풀이] 남학생이 세 과목에서 국어, 영어, 수학을 좋아할 확률을 각각 p_{11}, p_{12}, p_{13} 라 하고, 여학생에 대해서도 마찬가지로 p_{21}, p_{22}, p_{23} □ 라고 하면 검정하고자 하는 가설을 다음과 같다.

$$H_0 : (p_{11}, p_{12}, p_{13}) = (p_{21}, p_{22}, p_{23})$$

$H_1 : H_0$ 가 아니다.

분할표의 분석은 Scipy의 모듈 stats의 chi2_contingency함수를 사용한다.

```
chi2_contingency(observed)
```

모수 값:

- observed: 분할표

실행 결과:

- chi2: 검정통계량
- p: p값
- dof: 자유도
- expected: 기대 빈도

```
import numpy as np
from scipy import stats
import pandas
```

```
x = np.array([[73, 98, 79], [82, 58, 110]])
table = pandas.DataFrame(x, ['Male', 'Female'], ['Lit', 'Math', 'Eng']) # 주어진 분할표를 입력하고 확인한다.
print(table)

g, p, dof, expctd = stats.chi2_contingency(table) # 입력된 x에 대해 동질성 검정을 시행한다.
print("Chi-squared test: X-squared = %s, p-value = %s, df = %s" % (round(g,5),round(p,5),dof))
```

```
      Lit  Math  Eng
Male    73    98   79
Female  82    58  110
Chi-squared test: X-squared = 15.86365, p-value = 0.00036, df = 2
```

검정 통계량은 15.86이고 유의 확률은 약 0.0004이므로 유의수준 5%에서 귀무가설을 기각할 수 있다. 따라서 성별에 따른 과목별 선호도는 다르다고 말할 수 있다.

예 2) (survey.txt) 주어진 자료는 University of Adelaide 에서 총 237명의 학생들을 대상으로 한 조사의 결과이다. 주요 변수에 대한 설명은 다음과 같다.

- Sex: 성별(Male, Female)
- Smoke: 흡연정도 (Heavy, Regul, Occas, Never)
- Exer: 운동빈도 (Freq, Some, None)

학생들의 흡연 정도와 운동 빈도는 서로 독립이라고 말할 수 있는가? 적절한 가설과 함께 유의수준 5%에서 이를 검정하시오.

[풀이] 독립성 검정을 위한 가설은 다음과 같다.

H_0 : 흡연 정도와 운동 빈도는 서로 독립이다.

H_1 : H_0 가 아니다.

주어진 자료를 이용하여 분할표를 작성한 후 분석을 시행하도록 한다. 다음은 분할표 작성 결과와 독립성 시행 결과이다.

```
%cd D:/
```

```
D:\W
```

```
survey = pandas.read_table("survey.txt", sep = " ")
y = survey.groupby(['Smoke', 'Exer']).size()
table = y.unstack() # Smoke와 Exer 변수를 이용하여 분할표를 작성한다.
print(table)
```

```
Exer  Freq  None  Some
Smoke
Heavy    7     1     3
Never   87    18    84
Occas   12     3     4
Regul    9     1     7
```

```
g, p, dof, expctd = stats.chi2_contingency(table) # 독립성 검정을 시행한다.
print("Chi-squared test: X-squared = %s, p-value = %s, df = %s, expected value = %s" % (round(g,5),round(p,5),dof, expctd))
```

```
Chi-squared test: X-squared = 5.48855, p-value = 0.48284, df = 6, expected value = [[ 5.36016949
1.0720339  4.56779661]
[92.09745763 18.41949153 78.48305085]
[ 9.25847458  1.85169492  7.88983051]
[ 8.28389831  1.65677966  7.05932203]]
```

검정 결과 검정 통계량의 값은 5.4885이고 유의확률은 0.4828로 계산되었다. 따라서 유의수준 5%에서 귀무가설을 기각할 수 없으며 따라서 학생들의 운동 빈도와 흡연 정도는 서로 독립이라고 말할 수 있다.

<참고> survey 자료의 경우, 각 cell별의 기대빈도가 5보다 작은 것을 확인할 수 있다. 이러한 문제를 해결하기 위해서는 빈도수가 작은 범주들은 서로 병합하는 방법을 쓸 수 있다.

7.3 예제

예제1. 다음은 랜덤하게 선택된 806명의 Facebook 사용자들에 대해 Facebook의 privacy setting 기능의 사용법을 알고 있는지에 대해 조사한 결과이다. 성별과 privacy setting의 사용법 인지 여부가 관계가 있다고 할 수 있는가? 유의수준 5%에서 이를 검정하시오.

		성별		
		남	여	합계
사용법 인지여부	알고 있다	288	378	666
	잘 모른다	10	7	17
	전혀 모른다	61	62	123
	합계	359	447	806

• Survey USA, News Poll #17960, data collected February 16-17, 2011.

예제2. 심폐 소생술은 심장과 폐의 활동이 갑자기 멈추었을 때 실시하는 응급처치방법이다. 심폐 소생술 시행 시 혈액 뭉침 등을 완화하기 위해 혈액 희석제(blood thinner)를 처방하는 경우가 있는데, 만약 심폐 소생술로 인해 신체 내부 손상을 입은 경우라면 혈액 희석제는 환자에게 치명적일 수도 있다. 다음은 심폐 소생술을 받은 환자들 중 혈액 희석제를 사용하지 않은 환자들 50명과 혈액 희석제를 사용한 환자들 40명을 대상으로 그들의 생존 여부를 조사한 결과이다.

생존 여부	생존	사망	합계
혈액 희석제 사용 안함	11	39	50
혈액 희석제 사용함	14	26	40
합계	25	65	90

혈액 희석제 사용 여부에 따른 생존률은 차이가 있는가? 유의수준 5%에서 이를 검정하시오.