

04

교차분석 실습

[적합성 검정]

#적합성 검정에서는 귀무가설과 대립가설을 잘 세우는 것이

중요합니다.

00:20

4.1.3.2 적합성 검정

- 실형에서 얻어진 관측값들이 예상한 이론적 분포를 따르는지 여부를 검정하는 방법
- 관측값들이 어떠한 이론적 분포를 따르고 있는지 관찰할 수 있음
- 모집단 분포에 대한 가정이 옳게 됐는지 관측 자료와 비교하여 검정하는 것
- 검정통계량
 - 카이제곱 통계량 값이 큰 경우 : 관찰도수와 기대도수의 차이가 크고 적합도가 낮음 (알지한다고 볼 수 없음)
 - 카이제곱 통계량 값이 작은 경우 : 관찰도수와 기대도수의 차이가 작고 적합도가 높음 (알지한다고 볼 수 있음)

```
scipy.stats.chisquare(f_obs, f_exp=None)
```

- f_obs : 각 범주에서 관찰된 빈도
- f_exp : 각 카테고리의 예상 빈도

[예제]

titanic데이터에서 sex 변수에 대한 분할표를 생성하고 아래의 가설에 대한 적합도 검정을 수행하세요.

- 귀무가설 : 전체 응답자 중 남자의 비율이 50%, 여자의 비율이 50%이다
- 대립가설 : 전체 응답자 중 남자의 비율이 50%, 여자의 비율이 50%이라고 할 수 없다

```
In [27]: 1 import pandas as pd
2 # 데이터 불러오기
3 df_t = pd.read_csv("../data/titanic.csv")
4 # titanic 데이터의 구조 확인
5 df_t.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 11 columns):
 #   Column             Non-Null Count  Dtype
 #   --   --
```

[독립성 검정]

#독립성 검정은 교차표를 만드는 방법만 알면 쉽게 진행할 수

있습니다.

03:40

```
Power_divergenceResult(statistic=77.63075196408629, pvalue=1.2422096313910336e-18)
p-value가 0.0으로 유의수준 0.05보다 작으므로 귀무가설을 기각하고 대립가설을 지지함
```

4.1.3.3 독립성 검정

- 모집단이 두 개의 변수 A, B에 의해 범주화되었을 때, 이 두 변수들 사이의 관계가 독립인지 아닌지 검정하는 것
- 교차표를 활용함

```
scipy.stats.chi2_contingency(observed)
```

observed : 교차표 (각 범주에서의 발생 횟수 표)

[예제]

titanic 데이터에서 좌석등급(class)과 생존 여부(survived)가 서로 독립인지 확인하기 위해 분할표를 생성하고, 아래 가설에 대한 독립성 검정을 수행하라

- 귀무가설 : class와 survived는 독립이다
- 대립가설 : class와 survived는 독립이 아니다

```
In [21]: 1 # 변수의 분할표 생성
2
3 table = pd.crosstab(df_t[['class']], df_t['survived'])
4 table
```

```
Out[21]:
```

survived	0	1
class		
First	80	136
Second	97	87
Third	372	119

[동질성 검정]

#독립성 검정과 분석 방법은 완전히 똑같습니다!

06:45

```
File Edit View Insert Cell Kernel Navigate Widgets Help Trusted Python 3 (pykernel)
Run Code
p-value가 유의수준 0.05보다 작으므로 귀무가설을 기각
좌석 등급과 생존은 독립이 아니라고 할 수 있다.

4.1.3.4 동질성 검정
• 모집단이 임의의 변수에 따라 R개의 속성으로 범주화되었을 때,
R개의 부분 모집단에서 추출한 표본이 C개의 범주화된 집단의 분포가 서로 동일한지 검정
• 교차표를 활용하며, 계산법과 검증법은 모두 독립성 검정과 같은 방법으로 진행됨

• 귀무가설: TV프로그램의 선호도는 성별에 관계없이 동일하다
• 대립가설: TV프로그램의 선호도는 성별에 관계없이 동일하지 않다.

• 귀무가설: class의 분포는 survived에 관계없이 동일하다
• 대립가설: class의 분포는 survived에 관계없이 동일하지 않다.

In [22]: 1 chi, p, df, expect = chi2_contingency(table)
2 result = "chi2 : {}, p-value : {}, df : {}".format(chi, p, df)
3 print(result)
4 #자유도(df, degree of freedom)는 (3-1)*(2-1)로 2 입니다
chi2 : 102.88898875696056, p-value : 4.549251711298739e-23, df : 2

In [11]: 1 print(expect)
[[133.09090909 82.90909091]
 [113.37373737 70.62626263]
 [302.53535354 188.46464646]]

In [12]: 1 table
Out[12]:
```