

## 적합도 검정 : 관측도수와 기대도수의 차이가 통계적으로 유의미하다고 할 수 있는가

- 관측도수 : 실제 관측 된 값
- 기대도수 : 알려진 모집단에 의해 예상되는 값

```
import pandas as pd

#데이터명 = pd.read_csv('파일경로/파일명.csv')
hos = pd.read_csv('HOS.csv', sep=',', encoding='CP949')
hos
```

	world	range	role	size	speed	level	HP	attack	strength	agility	...	I1	I2	I3	I4	I5	P1	P2	P3	R1	R2
0	1	1	1	2	1	3	6710	169.26	24	100	...	2	2	5	2	1	6	10	8	3	61
1	3	1	2	2	2	3	5129	230.00	40	84	...	4	2	4	4	2	7	5	6	18	21
2	1	1	1	2	2	3	6064	214.23	44	100	...	4	3	5	3	2	5	5	9	6	5
3	3	1	2	2	2	3	5520	213.75	56	84	...	4	3	4	3	3	6	6	8	11	15
4	2	1	6	1	1	2	1503	81.51	80	20	...	1	5	1	2	5	3	7	1	83	80
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
78	1	1	2	1	2	3	5873	261.80	60	92	...	3	4	4	3	3	5	4	10	8	9
79	1	2	6	1	1	2	3070	155.76	76	36	...	2	5	1	1	5	5	7	5	69	75
80	3	2	5	1	2	3	4316	145.00	80	44	...	4	5	1	3	4	8	4	6	32	40
81	1	1	2	2	2	1	5702	199.29	52	88	...	4	3	4	4	3	5	3	7	10	5
82	5	2	5	1	1	3	3344	231.60	76	52	...	3	5	2	4	4	9	5	7	57	40

83 rows x 32 columns

- range는 2개의 범주형 자료
- role은 6개의 범주형 자료

```
import numpy as np
from scipy import stats # 이 두가지 패키지 먼저 실행

X = pd.crosstab(hos.range, hos.role, margins=True)
# crosstab : 교차표 불러오는 함수
# range, role : 2개의 범주형 변수
# margins = True : 행과 열의 합계를 나타냄
X
```

[빈도교차표 출력]

role	1	2	3	4	5	6	All
range							
1	9	13	9	0	3	5	39
2	1	4	6	10	12	11	44
All	10	17	15	10	15	16	83

role에 대한 관측도수와 기대도수의 차이에 대해 유의미한지 알아보자

```
Ob= x.values[1,:6]
Pr = np.array([0.1, 0.25, 0.15, 0.1, 0.15, 0.25])
# role 변수의 6개 집단에 대한 기대도수에 따른 비율 입력

n = x.values[1,6]
E = n*Pr
stats.chisquare(Ob, E)
```

```
Out[3]: Power_divergenceResult(statistic=18.68181818181818, pvalue=0.0022027974453811805)
```