11장 통계적 가설검정

: 유의확률(p값)을 이용하여 가설을 검정하는 방법

In [1]:
```
import numpy as np
import pandas as pd
from scipy import stats

%precision 3
np.random.seed(1111)
```

In [2]:
```
df = pd.read_csv('../data/ch11_potato.csv')
df
```

Out[2]:

	무게
0	122.02
1	131.73
2	130.60
3	131.82
4	132.05
5	126.12
6	124.43
7	132.89
8	122.79
9	129.95
10	126.14
11	134.45
12	127.64
13	125.68

In [3]:
```
sample = np.array(df['무게'])
sample
```

Out[3]:
```
array([122.02, 131.73, 130.6 , 131.82, 132.05, 126.12, 124.43, 132.89,
       122.79, 129.95, 126.14, 134.45, 127.64, 125.68])
```

In [4]:
```
s_mean = np.mean(sample)
s_mean
```

Out[4]:
```
128.4507142857143
```

In [5]:
```
np.var(sample)
```

Out[5]:
```
14.735449489795883
```

t-검정 (집단간 차이분석)

1. 단일표본 t-검정

- "ch11_potato.csv" 사용하여 분석
- 대립가설 : 감자튀김의 모집단 평균은 130g보다 작다. 14개의 표본평균과 모평균에 차이가 유의미하다.
- 귀무가설 : 모평균은 130g이다.

In [6]:
```
# ttest_1samp()함수는 단일표본 t-검정 함수.

t, p = stats.ttest_1samp(sample, 130)
t, p
```

Out[6]:
```
(-1.4551960206404198, 0.16933464230414275)
```

(결론) : p값이 0.169, 유의수준(0.05) 이상이므로 귀무가설 채택!

- 대립가설 기각 됨. 즉 감자튀김의 모평균은 130g보다 작다고 말할 수 없다.

2. 대응표본 t-검정

- "ch11_training_rel.csv"파일 사용하여 분석
- 근력운동이 집중력을 향상시키는 효과가 있는지 여부를 알고 싶어 실험을 함.
- 20명의 친구들 근력운동 전에 집중력테스트를 하고, 근력운동 후에 집중력테스트를 한 점수 분석.

- 귀무가설 : 근력운동은 집중력에 영향을 미치지 않는다. 근력운동을 하든 하지않든 집중력테스트 점수에는 차이가 없다.
- 대립가설 : 근력운동은 집중력에 양향을 미친다. 근력운동 전, 후의 집중력테스트 점수 차이는 유의미하다. 통계적으로 의미가 있다.

In [10]:
```
training_rel = pd.read_csv('../data/ch11_training_rel.csv')
print(training_rel.shape)
training_rel.head()
```

```
(20, 2)
```

Out[10]:

	전	후
0	59	41
1	52	63
2	55	68
3	61	59
4	59	84

In [13]:
```
# ttest_rel() 함수는 대응표본 t-검정 함수.

t, p = stats.ttest_rel(training_rel['후'], training_rel['전'])
p
```

Out[13]:
```
0.04004419061842953
```

(결론) : p값이 0.04은 유의수준인 0.05보다 미만이므로 귀무가설 기각 됨.

- 대립가설 채택 됨. 근력운동은 집중력에 유의미한 차이를 준다는 것. 평균의 차이는 통계적으로 의미있음.

3. 독립표본 t-검정 : 2개의 범주형 집단에 따른 연속형 자료의 평균 비교분석

- "ch11_training_ind.csv"파일로 분석
- 근력운동이 집중력을 향상시키는 효과가 있는지 여부

- 귀무가설 : A학급(근력운동을 한 집단)과 B학급(근력운동 안한 집단)의 집중력 평균점수는 차이가 없다. 근력운동은 집중력에 영향을 미치지 않는다.

- 대립가설 : A학급과 B학급의 집중력 평균 점수는 차이가 있다. 근력운동은 집중력에 영향을 미친다. 효과가 있다.

In [14]:
```
training_ind = pd.read_csv('../data/ch11_training_ind.csv')
print(training_ind.shape)
training_ind.head()
```

```
(20, 2)
```

Out[14]:

	A	B
0	47	49
1	50	52
2	37	54
3	60	48
4	39	51

In [15]:
```
t, p = stats.ttest_ind(training_ind['A'], training_ind['B'],
                       equal_var=False)
p
```

Out[15]:
```
0.08695731107259361
```

4. 카이제곱검정(교차분석) : 범주형 자료들 간의 차이 분석

- "ch11_ad.csv"파일 사용하여 분석
- 내보낸 광고와 상품 구입유무가 기록되어 있음.
- 광고A와 광고B를 내보냈을 때 구입비율에 유의한 차이가 있는지를 검정하기.

- 귀무가설 : 차이가 없다.

- 대립가설 : 차이가 있다.

In [16]:
```
ad_df = pd.read_csv('../data/ch11_ad.csv')
n = len(ad_df)
print(n)
ad_df.head()
```

1000

Out[16]:

	광고	구입
0	B	하지 않았다
1	B	하지 않았다
2	A	했다
3	A	했다
4	B	하지 않았다

In [17]:
```
# pd.crosstab()는 교차 분할 표(cross-tabulation table)를 생성하는 함수.

ad_cross = pd.crosstab(ad_df['광고'], ad_df['구입'])
ad_cross
```

Out[17]:

구입	하지 않았다	했다
광고		
A	351	49
B	549	51

In [18]:
```
#stats.chi2_contingency()는 카이제곱 검정 함수.

chi2, p, dof, ef = stats.chi2_contingency(ad_cross,
                                          correction=False)
chi2, p, dof
```

Out[18]:
```
(3.75, 0.052807511416113395, 1)
```

In [19]:
```
ef
```

Out[19]:
```
array([[360.,  40.],
       [540.,  60.]])
```

In [ ]:

In [ ]: