11장 통계적 가설검정

: 유의확률 (p값)을 이용하여 가설을 검정하는 방법

In [1]:

```
import numpy as np
import pandas as pd
from scipy import stats # 통계함수
%precision 3
np.random.seed(1111)
df = pd.read_csv("../data/ch11_potato.csv")
sample=np.array(df['무게'])
sample
```

Out[1]:

```
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 [2]:

```
s_mean=np.mean(sample)
s_mean
```

Out[2]:

128.451

In [3]:

```
np.var(sample) # 분산(표준편차 제곱) 구하는 함수
```

Out[3]:

14.735

t검정(집단간 차이분석)

1. 단일표본 t-검정

- 대립가설 : 감자튀김의 모집단 평균은 130g보다 작다. 14개의 표본평균과 모평균에 차이가 유의미하다
- 귀무가설 : 모평균은 130g이다.

In [4]:

```
# ttest_1samp()함수는 단일표본 t-검정 함수.
# 130은 모평균
# t(통계량), p(유의확률)을/를 반환함

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

Out[4]:

(-1.455, 0.169)

(결론): p값이 0.169, 유의수준(0.05) 이상으로 귀무가설(두 집단 간의 관련성 x) 채택!

- 대립가설 기각 됨, 즉 감자튀김의 모평균은 130이다

2. 대응표본 t-검정

- 근력운동이 집중력을 향상시키는 효과가 있는지 여부를 알고 싶어 실험을 함
- 20명의 친구들 근력운동 전에 집중력테스트를 하고, 근력운동 후에 집중력 테스트를 한 점수 분석
- 귀무가설: 근력운동은 집중력에 영향을 미치지 않는다. 근력운동을 하든 하지않든 집중력 테스트 점수에는 차이가 없다
- 대립가설: 근력운동은 집중력에 영향을미친다

In [5]:

```
training_rel = pd.read_csv('../data/ch11_training_rel.csv')
print(training_rel.shape) # 행과 열을 할수있음
training_rel.head() # 5개만 출력
```

(20, 2)

Out[5]:

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

In [6]:

```
# ttest_rel() 함수는 대응표본 t-검정 함수,, 전 - 후 = 값을 생략한것
t, p = stats.ttest_rel(training_rel['후'], training_rel['전'])
p
```

Out[6]:

0.040

(결론) : p값이 0.04은 유의확률 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

- 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 [7]:

```
training_ind=pd.read_csv("../data/ch11_training_ind.csv")
print(training_ind.shape)
training_ind.head()
```

(20, 2)

Out[7]:

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

In [8]:

```
# 기술통계량

training_ind.describe()
```

Out[8]:

	A	B
count	20.000000	20.000000
mean	48.750000	52.050000
std	6.711145	5.020746
min	37.000000	41.000000
25%	44.750000	49.000000
50%	48.500000	52.000000
75%	53.000000	54.250000
max	64.000000	64.000000

In [9]:

```
# Levene의 등분산 검정 scipy.stats.levene() 함수를 사용
# p-value가 유의수준(0.05)보다 작으면, 두 표본의 분산이 서로 다르다.
# p값이 유의수준이상이면, 두 표본의 분산이 같다고 분석함.
A=np.array(training_ind['A'])
B=np.array(training_ind['B'])

stats.ttest_rel(A,B)
```

Out[9]:

```
TtestResult(statistic=-1.9554736733098574, pvalue=0.06539643502362615, df=19)
```

In [10]:

```
# p값이 0.16로 유의수준 0.05 이상으므로 두 표본의 분산은 같다고 분석되 귀무가설이 채택됨
stats.levene(A,B,center='mean') # center값을 평균? 아니면 중앙값으로 잡는다
```

Out[10]:

```
LeveneResult(statistic=2.04785481661503, pvalue=0.16059385895425907)
```

In [11]:

```
# 독립표본 t-검정 수행
# 기본적으로 equal_var=True로 설정되어 있음. 디폴트.
# 즉, 두 표본의 분산이 동일하다고 가정하고 t-검정 수행
t,p=stats.ttest_ind(A,B)
t,p
```

Out[11]:

```
(-1.761, 0.086)
```

In [12]:

```
# equal_var=False를 지정하면 웰치의 방법으로 독립표본 t-검정이 수행됨.
# 웰치의 t검정은 두 표본의 분산이 다르다는 것을 가정하고 검정을 수행한다.
# 등분산을 가정하지 않고 p값을 구함
t,p = stats.ttest_ind(A, B, equal_var= False)
p
```

Out[12]:

0.087

(결론)

- 독립표본 t-검증 수행 결과, p값이 0.086으로 유의수준(0.05) 이상이므로 귀무가설이 채택됨. 대립가설 기각.
- 따라서, A학급(근력운동 한 집단)과 B학급(근력운동 안한 집단)의 집중력 평균점수에 유의한 차이가 있다고 말할 수 없다. 즉, 차이가 없다.

카이제곱검정(교차분석) : 범주형 자료들 간의 차이 분석

- 내보낸 광고와 상품 구입유무가 기록되어 있음.
- 광고A와 광고B를 내보냈을 때 구입비율에 유의한 차이가 있는지를 검정하기.
- 귀무가설 : 차이가 없다.
- 대립가설 : 차이가 있다.

In [13]:

```
ad_df = pd.read_csv("../data/ch11_ad.csv")
print(ad_df.shape)
ad_df.head()
```

(1000, 2)

Out[13]:

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

In [14]:

```
# pd.crosstab()는 교차 분할 표(cross-tabulation table)를 생성하는 함수.
ad_cross = pd.crosstab(ad_df['광고'],ad_df['구입'])
# 광고(행), 구입(열),, 빈도수가 계산됨
ad_cross
```

Out[14]:

구입	하지 않았다	했다
광고		
A	351	49
B	549	51

In [15]:

```
# stats.chi2_contingency()는 카이제곱 검정 함수.
# 인수로 고차집계표를 전달하고, correction을 False로 설정
# 이 함수의 반환값은 검정통계량, p값, 자유도, 기대도수가 된다.

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

Out[15]:

```
(3.750, 0.053, 1)
```

In [16]:

```
ef # 기대도수
```

Out[16]:

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

(결론)

- p값이 0.05이상이므로 귀무가설 채택되고,
- 광고A와 광고B에 유의한 차이가 인정되지 않는다는 결론을 내린다.