11장 통계적 가설검정

: 유의확률(p값)을 이용하여 가설을 검정하는 방법

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

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

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

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

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

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

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

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

(결론) : p값이 0.169 이므로 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 [ ]:
```
training_rel = pd.read_csv('../data/ch11_training_rel.csv')
print(training_rel.shape)
training_rel.head()
```

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

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

(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

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

In [ ]:
```
# 기술통계량

training_ind.describe()
```

In [ ]:
```
# 두 표본의 분산이 동일한지 아닌지를 먼저 판단하기
# Levene의 등분산 검정. scipy.stats.levene() 함수를 사용
# p-value가 유의수준(0.05)보다 작으면, 두 표본의 분산이 서로 다르다고 분석함.
# p값이 유의수준이상이면, 두 표본의 분산이 같다고 분석함.

A = np.array(training_ind['A'])  # 표본은 1차원이어야 하므로 넘파이배열로 변환.
B = np.array(training_ind['B'])

stats.levene(A, B, center='mean') # p값이 0.16로 유의수준 0.05 이상이므로 두 표본의
```

In [ ]:
```
# 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
```

In [ ]:
```
# equal_var=False를 지정하면 웰치의 방법으로 독립표본 t-검정이 수행됨.
# 웰치의 t검정은 두 표본의 분산이 다르다는 것을 가정하고 검정을 수행한다.

t, p = stats.ttest_ind(A, B, equal_var=False)
p
```

(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.

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

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

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

In [ ]:
```
# 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
```

In [ ]:
```
ef
```

(결론)

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

In [ ]:
```

```