이기적 빅데이터분석기사 실기 PYTHON

작업형 제 3 유형 추가문제+해설

1. 아래의 데이터는 11명 학생들의 수학 점수를 모아둔 리스트이다.

```
data = [ 75, 82, 80, 76, 84, 81, 79, 80, 78, 83, 74 ]
```

모집단의 평균값이 80일 때 일표본 t-검정(one-sample t-test)을 시행하여

평균값이 유의한지 확인하려 한다.

- 1-a) 표본 평균값을 구하시오. (반올림하여 소수 셋째자리까지 계산)

```
import numpy as np
m = round(np.mean(data),3)
print(m)                                                    # 79.273
```

- 1-b) 위의 가설을 검정하기 위한 검정 통계량을 구하시오. (반올림하여 소수 둘째자리까지 계산)

```
# 모집단 평균값
pop_mean = 80

# 일표본 t-검정
from scipy.stats import ttest_1samp
t_statistic, p_value = ttest_1samp(data, pop_mean)

t_statistic_round = round(t_statistic,2)
print(t_statistic_round)                                    # -0.74
```

- 1-c) 위의 통계량에 대한 p-값을 구하고(반올림하여 소수 넷째자리까지 계산),
  유의수준 0.05 하에서 가설검정의 결과를 (채택/기각) 중 하나를 선택하시오.

```
p_value_round = round(p_value,4)
print(p_value_round)                                        # 0.4762
# p_value 는 0.4762 로 0.05 보다 크므로 귀무가설을 채택한다.
# 즉 표본 학생들의 점수 평균은 80 이라 보기 유의하다.
```

2. p2.csv 파일에는 95명의 키에 대한 데이터가 있다.

평균키는 165cm라 판단할 수 있는지 귀무가설과 대립가설을 설정한 후

유의수준 5%로 검정하고자 한다.

- 2-a) 정규성 여부를 확인하고자 한다. Shapiro 검정 통계량을 구하시오.
  (반올림하여 소수 셋째자리까지 계산)

```
from scipy.stats import shapiro
import pandas as pd
p2 = pd.read_csv('p2.csv')

static, pvalue = shapiro(p2)
static_round = round(static,3)
print(static_round)                                    # 0.989
```

- 2-b) Shapiro 검정 p-값을 구하고(반올림하여 소수 셋째자리까지 계산) 정규성 여부에 대해
  유의수준 0.05 로 검정하시오.

```
pvalue_round = round(pvalue,3)
print(pvalue_round)                                    # 0.638
# 0.638 로 0.05 이상이므로 귀무가설 기각하지 않음, 데이터는 정규성을 만족
```

- 2-c) 데이터의 평균키는 165cm 라고 할 수 있는지 일표본 t-검정을 시행하고
  검정량을 소숫점 이하 3 자리로 구하시오.

```
from scipy.stats import ttest_1samp
static, pvalue = ttest_1samp(p2['height'],165)
static_round = round(static,3)
print(static_round)                                    # 3.225
```

3. p3.csv 파일에는 A, B 두 학급 각 학생들의 중간고사 평균 점수들이 저장되어 있다.

두 학급의 시험 평균(비모수검정의 경우 중위값)은 동일하다 말할 수 있는지 확인하려 한다.

- 3-a) 학급(class)별 각각 정규성을 만족하는지 여부를 Shapiro 검정을 통해 확인하시오.

```
import pandas as pd
from scipy.stats import shapiro
p3 = pd.read_csv('p3.csv')

a = p3[p3['class'] =='A']
b = p3[p3['class'] =='B']
s_a,p_a = shapiro(a.data)
s_b,p_b = shapiro(b.data)
print(p_a, p_b)                                    # 0.37967…  0.67936…
# 두 p-value 모두 0.05 보다 크므로 0.05 유의수준 하에서 정규성을 만족한다.
```

- 3-b) A, B 학급의 데이터는 등분산을 가지는지 Levene 검정을 통해 확인하시오.

```
from scipy.stats import levene
s, p = levene(a.data,b.data)
print(p)
# 0.1130290…으로 귀무가설을 기각하지 못한다. 따라서 등분산을 가진다.
```

- 3-c) 두 학급의 평균 점수는 동일하다고 할 수 있는지 독립표본 t-검정을 시행하여 분석하시오.

```
from scipy.stats import ttest_ind
print(ttest_ind(a.data,b.data,equal_var=True))   # pvalue=0.00619…
# 등분산이기 때문에 equal_var=True 파라미터를 주고 ttest_ind 모듈을 이용한다.
# p-value 는 0.006 이므로 귀무가설(각 그룹의 평균값은 동일하다)을 기각한다.
```

4. p4.csv 파일에는 신약을 특정 질병 집단에 투약 전후 측정한 혈류량이 저장되어 있다. 이 데이터를 바탕으로 투약 전후의 변화가 있는지 검정하고자 한다.

- 4-a) 대응표본 t-검정(paired t-test) 통계량을 구하시오. (반올림하여 소수 셋째자리까지 계산)

```
import pandas as pd
p4 = pd.read_csv('p4.csv')

from scipy.stats import ttest_rel
before = p4['before']
after = p4['after']

s,p = ttest_rel(before,after)

s_round = round(s,3)
print(s_round)                                          # -2.725
```

- 4-b) p-값을 구하고(반올림하여 소수 셋째자리까지 계산) 귀무가설 기각 여부를 판단하시오.

```
p_round = round(p,3)
print(p_round)                                          # 0.008
# 0.008 로 0.05 하에서 귀무가설을 기각한다. 따라서 투약 전후 변화가 존재한다.
```

5. 물고기 급속성장 먹이 실험을 진행하여 물고기가 먹이를 먹은 후의 몸무게를 측정하였다.

실험군 10마리의 몸무게 측정값(g)은 다음과 같다.

```
data = [ 2.1, 1.9, 1.8, 2.0, 2.2, 2.3, 1.9, 1.7, 2.4, 2.5 ]
```

이 실험 결과를 토대로 모집단의 평균값이 2.0g일 때, 얻은 표본의 평균값이 통계적으로 유의한지 일표본 t-검정을 통해 확인하고자 한다.

- 5-a) 실험에서 얻은 표본의 평균값을 구하시오. (반올림하여 소수 둘째자리까지 계산)

```
import numpy as np
m = round(np.mean(data),2)
print(m)                                                    # 2.08
```

- 5-b) 위의 가설을 검정하기 위한 검정 통계량을 구하시오. (반올림하여 소수 셋째자리까지 계산)

```
# 모집단 평균값
pop_mean = 2.0

# 일표본 t-검정
from scipy.stats import ttest_1samp
t_statistic, p_value = ttest_1samp(data, pop_mean)

t_statistic_round = round(t_statistic,3)
print(t_statistic_round)                                    # 0.952
```

- 5-c) 위의 통계량에 대한 p-값을 구하고(반올림하여 소수 셋째자리까지 계산),
  유의수준 0.05 하에서 가설검정의 결과를 (채택/기각) 중 하나를 선택하시오.

```
p_value_round = round(p_value,3)
print(p_value_round)                                        # 0.366
# p_value 는 0.366 으로 0.05 보다 크므로 귀무가설을 채택한다.
# 즉 표본 물고기들의 몸무게 평균은 2.0 이라 보기 유의하다.
```

6. 다이어트약의 체중 감량 효과를 실험하기 위해 정해진 기간 복용 후 체중을 측정하였다.

```
before = [ 60, 65, 70, 75, 80 ]     # 다이어트약 복용 전 체중
after = [ 58, 62, 68, 72, 78 ]      # 다이어트약 복용 후 체중
```

대응표본 t-검정을 통해 체중 변화가 유의미하게 있는지 확인하려고 한다.

- 6-a) 표준 오차를 구하시오. (반올림하여 소수 셋째자리까지 계산)

```
# 표본 평균 차이 계산
import numpy as np
diff = np.array(before) - np.array(after)
mean_diff = np.mean(diff)

# 표준 오차 계산
se = round(np.std(diff, ddof=1) / np.sqrt(len(diff)),3)
print(se)                                                    # 0.245
```

- 6-b) 위의 가설을 검정하기 위한 검정 통계량을 구하시오. (반올림하여 소수 셋째자리까지 계산)

```
t_value = round(mean_diff/se,3)
print(t_value)                                               # 9.796
```

- 6-c) 위의 통계량에 대한 p-값을 구하고(반올림하여 소수 셋째자리까지 계산),
  유의수준 0.05 하에서 가설검정의 결과를 (채택/기각) 중 하나를 선택하시오.

```
# p-value 계산
from scipy.stats import t
p_value = round(2 * (1 - t.cdf(abs(t_value), df=len(diff)-1)),3)
print(p_value)                                               # 0.001
# p_value 는 0.05 이하이므로 귀무가설을 기각한다.
# 즉 체중 변화가 유의미하게 발생하였다.
```

7.

주사위 숫자	숫자 등장 횟수
1	24
2	20
3	28
4	22
5	28
6	22

왼쪽의 표는 144회 주사위를 던졌을 때, 각 숫자별로 나온 횟수를 나타낸다. 이 데이터가 주사위의 분포에서 나올 가능성이 있는지 검정하고자 한다.

- 7-a) 데이터프레임을 만들고 숫자 등장 횟수의 평균을 구하시오.

```
import pandas as pd
df = pd.DataFrame({'주사위 숫자':[1,2,3,4,5,6],
                   '숫자 등장 횟수':[24,20,28,22,28,22]})
m = df['숫자 등장 횟수'].mean()
print(m)                                                    # 24.0
```

- 7-b) 카이제곱 검정으로 적합도 검정을 하려한다.
  검정 통계량을 구하시오. (반올림하여 소수 셋째자리까지 계산)

```
from scipy.stats import chisquare
df['expected'] = (df['숫자 등장 횟수'].sum()/6).astype('int')
s,t = chisquare(df['숫자 등장 횟수'],df.expected)

round_s = round(s,3)
print(round_s)                                              # 2.333
```

- 7-c) 카이제곱 검정 적합도 검정의 p-값을 구하고(반올림하여 소수 셋째자리까지 계산),
  유의수준 0.05 하에서 분포 적합 여부(채택/기각)를 선택하시오.

```
round_t = round(t,3)
print(round_t)                                              # 0.801
# p_value 는 0.8 로 유의수준 0.05 이하에서 귀무가설을 기각할 수 없다.
# 즉 주사위의 분포에서 나올 수 있는 데이터이다.
```

8. 우리나라의 오른손잡이, 왼손잡이 비율을 0.8 : 0.2라고 한다.

100명의 표본을 뽑았을 때 오른손잡이와 왼손잡이 인원이 각 70명과 30명이었다.

이 비율이 유효한지 카이제곱 독립성 검정을 통해 알아보고자 한다.

- 8-a) 검정 통계량을 구하시오. (반올림하여 소수 둘째자리까지 계산)

```
# 데이터 입력
import numpy as np
observed = np.array([70,30])

# 예상 비율 계산
total = observed.sum()
expected = np.array([total*0.8, total*0.2])

# 카이제곱 통계량 계산
chi_squared = round(((observed - expected)**2 / expected).sum(),2)
print(chi_squared)                                          # 6.25
```

- 8-b) p-값을 구하고(반올림하여 소수 셋째자리까지 계산) 귀무가설 기각 여부를 판단하시오.

```
# 자유도 계산
df = len(observed) - 1

# p-value 계산
from scipy.stats import chi2
p_value = round(1 - chi2.cdf(chi_squared, df=df),3)
print(p_value)                                              # 0.012
# 0.012 로 0.05 하에서 귀무가설을 기각한다.
# 따라서 데이터 분포는 알려진 비율과 다르다.
```

9. p9.csv 파일에는 A, B 두 공장에서 각각 생산한 제품들의 불량을 판정한 데이터가 있다. 두 공장에서 생산된 제품의 불량률이 동일한지를 카이제곱 검정으로 분석하고자 한다.

- 9-a) A 공장의 불량률(불량품/생산량*100)에서 B 공장 불량률을 뺀 결과를 구하시오. (반올림하여 소수 셋째자리까지 계산)

```
import pandas as pd
df = pd.read_csv('p9.csv')
a_ratio = df[(df.생산공장명=='A') & (df.상태=='불량')].shape[0] /
          df[(df.생산공장명=='A')].shape[0] * 100
b_ratio = df[(df.생산공장명=='B') & (df.상태=='불량')].shape[0] /
          df[(df.생산공장명=='B')].shape[0] * 100

delratio = round(a_ratio - b_ratio,3)
print(delratio)                                          # 1.535
```

- 9-b) 카이제곱 검정 통계량을 구하시오. (반올림하여 소수 셋째자리까지 계산)

```
import numpy as np
from scipy.stats import chi2_contingency
obs = pd.crosstab(df.생산공장명,df.상태).values
chi2, p, dof, expected = chi2_contingency(obs)

round_chi2 = round(chi2,3)
print(round_chi2)                                        # 0.204
```

- 9-c) 카이제곱 검정 p-값을 구하시오. (반올림하여 소수 셋째자리까지 계산)

```
round_p = round(p,3)
print(round_p)                                           # 0.652
# p-value 가 0.05 보다 크므로 귀무가설을 기각할 수 없다.
# 따라서 A, B 두 공장에서 생산된 제품의 불량률은 동일하다.
```

10. 한 기계 부품의 rpm 수치를 두 가지 상황에서 측정하고 p10.csv에 저장하였다.

b 상황이 a 상황보다 rpm 값이 높다고 할 수 있는지 검정하려 한다.

- 10-a) a, b 상황이 각각 정규성을 가지는지 Shapiro 검정을 통해 확인하시오.

```
import pandas as pd
from scipy.stats import shapiro
df = pd.read_csv('p10.csv')

a = df[df['group']=='a'].rpm
b = df[df['group']=='b'].rpm

print(shapiro(a))                                    # pvalue=0.8884…
print(shapiro(b))                                    # pvalue=0.5505…
# 두 p-value 모두 0.05 보다 크므로 귀무가설을 기각할 수 없어 정규성을 만족한다.
```

- 10-b) a, b 상황은 등분산을 가지는지 Levene 검정을 통해 확인하시오.

```
from scipy.stats import levene
print(levene(a,b))                                   # pvalue=0.7959…
# 0.79 로 귀무가설을 기각하지 못한다. 따라서 등분산을 가진다.
```

- 10-c) 대응표본 t-검정을 통해 b 가 a 보다 크다 할 수 있는지 검정하시오.

```
from scipy.stats import ttest_rel
print(ttest_rel(b,a,alternative='greater'))       # pvalue=0.0306…
# p-value 는 0.03 이므로 귀무가설(b 는 a 와 차이가 없다)을 기각한다.
# b 가 a 보다 높다고 말할 수 있다.
```