통계학실험(005)
과제 5 (6장)
2022-18608 문현호

1
예제 1. (textbooks.txt) 주어진 자료는 UCLA 내의 서점과 Amazon.com에서 판매되는 교재들의 가격에 대한 자료이다. 2010년 봄학기에 개설된 UCLA의 강의 중에서 73개를 선택하여 각 강의에 쓰이는 교재의 온라인 판매 가격(amazNew)과 오프라인의 판매 가격(uclaNew)을 조사하였다. 교재의 판매 가격은 판매 장소(온라인 또는 오프라인)에 따라 차이가 난다고 볼 수 있는가? 적절한 가설을 세우고 유의수준 5%에서 이를 검정하시오.

```
textbooks = pd.read_csv("./data/textbooks.txt", sep=" ")

price = textbooks[["uclaNew", "amazNew"]]
print(price.describe())
result = ttest_rel(price["uclaNew"], price["amazNew"])
print(result)

'''
            uclaNew       amazNew
count     73.000000     73.000000
mean      72.221918     59.460274
std       59.659128     48.995571
min       10.500000      8.600000
25%       24.700000     20.210000
50%       43.560000     34.950000
75%      116.000000     88.090000
max      214.500000    176.000000
Ttest_relResult(statistic=7.648771112479753, pvalue=6.927581126065491e-11)

'''
```

- UCLA 서점에서의 가격은 평균 72.22달러, 표준편차 59.66달러였던 반면, 아마존에서의 가격은 평균 59.46달러, 표준편차 49.00달러로 나타났다.
- 귀무가설로 두 서점에서의 가격이 같다, 대립가설로 두 서점에서의 가격이 다르다, 라고 가정한 다음, 두 값을 연관 T test로 분석했다. 그 결과 p value < 0.05로, 두 값은 유의수준 5%에서 서로 다르다는 사실을 확인할 수 있었다.

2

예제 2. (run10samp.txt) 주어진 자료는 2012년 Washington, DC에서 열렸던 Cherry Blossom 10 mile run 경기에서 완주를 한 선수 100명의 자료이다. 주요 변수에 대한 설명은 다음과 같다.

변수명	설명
time	10마일 달리기 완주 기록(분)
age	선수 나이
gender	성별 (M=남성, F=여성)
sale	출신지역 (또는 국가)

```
run10sample = pd.read_csv("./data/run10samp.txt", sep=" ")

print(run10sample.describe(include="all"))
```

```
'''
                time         age gender state
count     100.000000  100.000000    100   100
unique           NaN         NaN      2    11
top              NaN         NaN      F    VA
freq             NaN         NaN     55    39
mean       95.614600   35.050000    NaN   NaN
std        15.776914    8.972883    NaN   NaN
min        55.310000   20.000000    NaN   NaN
25%        85.925000   29.000000    NaN   NaN
50%        95.460000   32.500000    NaN   NaN
75%       103.995000   38.250000    NaN   NaN
max       141.940000   67.000000    NaN   NaN

'''
```

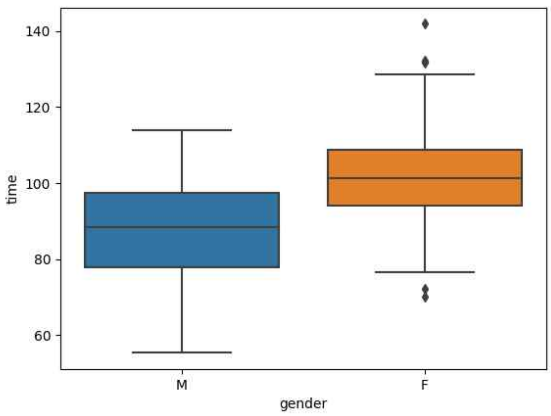
- 우선 해당 파일을 불러오는 코드를 작성했다. 이 파일은 아래 추가 과제들에서 이용된다.
- 100개의 데이터에 대해, 시간은 평균 95.61분, 나이는 평균 35.05세, 성별은 두 종류로, 여성이 55로 남성 45보다 많았다. 출신 지역은 11 종류의 값이 있었고, 대부분 VA 지역에서 온 것으로 나타났다.

add 1
예제1. 선수의 성별별로 완주시간의 분포를 상자그림을 그려 비교해보고, 남자 선수의 완주시간이 여자 선수의 완주시간보다 빠른지 검정해보시오.

```
male = run10sample[run10sample["gender"] == "M"]["time"]
female = run10sample[run10sample["gender"] == "F"]["time"]
gender_run = pd.DataFrame({'male_time': male, "female_time": female})
print(gender_run.describe())
sns.boxplot(data=run10sample, x="gender", y="time")
plt.show()

result = ttest_ind(male, female)
print(result)
```

```
'''
          male_time   female_time
count     45.000000     55.000000
mean      87.645333    102.134909
std       12.532218     15.236103
min       55.310000     70.120000
25%       77.860000     94.105000
50%       88.310000    101.320000
75%       97.340000    108.705000
max      114.060000    141.940000
```




```
Ttest_indResult(statistic=-5.117305648181875, pvalue=1.5438819783437747e-06)

'''
```

- 남자 선수의 완주시간은 독립 T 검정으로 확인된 결과, p<0.05로, 5% 유의수준에서 여자 선수의 완주시간과 달랐다. 상자 그림에서 보이는 완주시간의 분포는 남자가 더 작은 쪽에 있었고, 평균값도 그러했다.

add 2
예제2. pandas의 groupby함수를 이용하여 출신지역별 완주시간의 평균과 분산을 비교해보자. 분산이 계산되지 않는 지역이 있다면, 왜 계산되지 않는지에 대해 서술하시오.

```
by_state = run10sample.groupby(["state"])
by_state = by_state['time']
print(by_state.describe())

'''
         count          mean          std  ...       50%         75%       max
state                                      ...
DC        26.0     96.321923    16.419822  ...    96.715   106.4725    128.51
IN         1.0    141.940000          NaN  ...   141.940   141.9400    141.94
LA         1.0    117.020000          NaN  ...   117.020   117.0200    117.02
MD        18.0     90.436667    18.105047  ...    88.025   101.0075    132.14
ME         1.0     96.390000          NaN  ...    96.390    96.3900     96.39
MI         1.0     97.970000          NaN  ...    97.970    97.9700     97.97
NC         1.0     68.460000          NaN  ...    68.460    68.4600     68.46
NJ         4.0     98.947500    22.919945  ...    92.690   102.8525    131.84
NY         6.0     98.190000    15.657488  ...    98.065   104.8000    120.19
PA         2.0     93.765000     9.411591  ...    93.765    97.0925    100.42
VA        39.0     95.768974    12.034769  ...    97.340   102.5550    131.99

[11 rows x 8 columns]

'''
```

- 출신 지역별 완주 시간 분포는 표와 같다. 평균이 가장 작은 곳은 NC였고, 다음은 MD였으며, 평균이 가장 큰 곳은 IN이였고, 다음은 LA였다. 표준편차가 가장 큰 곳은 NJ, 가장 작은 곳은 PA였다. 다만, 데이터의 수 자체가 작으므로, 그 표준편차의 의미가 크지 않다.
- 분산은 위 표준편차의 제곱 값과 같으며, 따로 확인하지는 않았다. NaN(Not a Number) 에러가 뜨는 값은 그 데이터의 수가 1개뿐이기 때문에, 분산을 계산할 수 없기 때문이다.

add 3
예제3. DC와 MD 지역 선수들의 완주시간의 등분산성을 검정해보세요.

```
DC_time = run10sample[run10sample['state'] == 'DC']['time']
MD_time = run10sample[run10sample['state'] == 'MD']['time']

DC_N = DC_time.count() - 1
MD_N = MD_time.count() - 1

DC_S = float(DC_time.var()) / DC_N
MD_S = float(MD_time.var()) / MD_N

F = DC_S / MD_S

print(2 * min(f.cdf(F, dfn=DC_N, dfd=MD_N), 1 - f.cdf(F, dfn=DC_N, dfd=MD_N)))

'''
0.18195542549788948

'''
```

- 등분산성을 검정하기 위해, 각 분산과 자유도를 구하고, F 분포를 이용하여 양측검정한 결과, p > 0.05로, 유의수준 0.05에서 두 분산이 다르다는 대립가설이 기각되었다.
- 따라서 두 지역 선수들의 완주시간은 등분산성을 가진다.

add 4
예제4. 시작 값(x)와 끝 값(y)를 매개 변수로 받아 x와 y 사이의 모든 자연수의 합을 반환값으로 주는 함수(f)를 작성하시오. 단, x와 y는 모두 자연수이고 함수의 이름은 자유롭게 하세요.
(f(1,10)의 값이 55가 나와야 합니다.)

```
def add(x: int, y: int):
    ans = 0
    for i in range(x, y + 1):
        ans += i
    return ans


print(add(1, 10))

'''
55

'''
```

- 함수 정의 방법은 위와 같다. 입력값에 형식을 지정하지 않을 수도 있지만, 자연수 입력을 강제하기 위해 int 타입을 지정하였다.