유럽축구리그 홈팀 승률에 코로나19가 끼친 영향분석 Based on Permutation Test(순열검정)

12182183 주형진 통계학과 선택과제 (A형)

1. 서론

본연구에서는 경기가 무관중으로 치러졌을 때 홈팀과 어웨이팀 사이에 유의한 승률 차이가 존재하는지 알아보려고 합니다. 스포츠 경기에서 홈팀이 유리하다는 건 어제 오늘 일이 아닙니다. 특히 축구에서 더 그렇습니다. 2018-2019시즌을 기준으로 유럽 5대 프로축구 리그(잉글랜드, 이탈리아, 스페인, 독일, 프랑스) 경기 결과를 살펴 보면 홈팀은 817승 427무 537패를 기록했습니다. 다른 스포츠처럼 '승리 / (승리+패)'로 계산하면 홈팀 승률은 60.3%입니다.

축구 경기에서 홈팀이 유리한 이유로 손꼽히는 것 중 하나가 '응원'입니다. 관중 응원 소리는 홈팀 선수들의 사기를 끌어올릴 뿐만 아니라 심판 판정에도 영향을 끼칩니다. 그리고 축구는 '저득점 경기'이기 때문에 다른 종목보다 심판 판정에 큰 영향을 받습니다.[1]

코로나19를 겪는 동안 산업의 많은 영역에서 큰 변화가 있었고 축구 시장 또한 여러 변화를 맞이 했습니다. 특히나 2019-2020 시즌은 좀 특이했습니다. 전 세계적인 코로나19의 확산으로 '무관중 경기'가 적지 않았습니다. 심지어 가장 큰 축구 행사 중 하나인 2020년 리스본에서 개최된 챔피언스리그 결승전 또한 무관중으로 경기가 진행되었습니다.

2019-2020시즌 유럽 5대 리그에서 펼쳐진 3450개의 경기를 데이터를 기반으로 홈팀과 어웨이팀의 승률차이와 코로나 기간동안 홈팀과 어웨이 팀의 승률차이가 변화가 있었는지를 검정하고자 합니다. 두 독립 집단의 평균의 차이를 확인하기 위해 필요한 t-검정법과 데이터를 랜덤으로 뒤섞은 후에 복제 샘플을 만들어서 재정렬하는 순열검정(Permutationm test)를 기반으로 데이터 분석을 진행합니다. 아래의 가설검정을 진행합니다.

$$H_0 : \mu_{홈팀승률} = \mu_{어웨이팀_{승률}}$$
$$H_1 : \mu_{홈팀승률} > \mu_{어웨이팀_{승률}}$$

[1] Home Advantage in Football: A Current Review of an Unsolved Puzzle
(ttps://opensportssciencesjournal.com/contents/volumes/V1/TOSSJ-1-12/TOSSJ-1-12.pdf)

2. 본론

[분석의 순서]

0. 데이터의 정의와 전처리

1. 홈경기와 원정경기 사이의 승률 차이의 유의성을 살펴본다.
   - 1-1. Box-plot과 기술통계량을 통해 홈과 원정 승률 차이를 확인한다.
   - 1-2. T-Test를 통해 유의성 가설검정을 확인한다.
   - 1-3. Permutation test를 우연성 통계적 가설검정

2. 코로나전 홈경기 승률과 코로나 이후 홈경기 승률 차이의 유의성을 살핀다.
   - 2-1. Box-plot과 기술통계량을 통해 홈과 원정 승률 차이를 확인한다.
   - 2-2. T-Test를 통해 유의성 가설검정을 확인한다.
   - 2-3. Permutation test를 우연성 통계적 가설검정

0.) 데이터의 정의와 전처리

2019-2020시즌 유럽 5대 프로축구 리그 경기 결과를 담고 있는 '19-20_uefa_big5.csv'파일을 불러옵니다.

1. 데이터 입력하기

```
import pandas as pd
df=pd.read_csv('/content/gdrive/MyDrive/19_20_uefa_big_5.csv', encoding = 'euc-kr')
df.head(5)
```

	날짜	리그	팀	상대	득점	실점	승리	장소	시기
0	2019-08-10	EPL	크리스탈 팰리스	에버턴	0	0	0	안방	BC
1	2019-08-10	EPL	왓포드 FC	브라이튼	0	3	0	안방	BC
2	2019-08-10	EPL	웨스트햄	맨시티	0	5	0	안방	BC
3	2019-08-10	EPL	AFC 본머스	셰필드	1	1	0	안방	BC
4	2019-08-10	EPL	번리	사우샘프턴	3	0	1	안방	BC

이 파일에는 이런 내용이 들어 있습니다.

```
df1 = df.groupby(['팀', '장소'], as_index = False)\
    .agg(승리 = ('승리','mean'))

df1
```

	팀	장소	승리
0	AC 밀란	방문	0.526316
1	AC 밀란	안방	0.473684
2	AFC 본머스	방문	0.210526
3	AFC 본머스	안방	0.263158
4	AS 로마	방문	0.578947
...	...	...	...
191	헤르타 BSC	안방	0.352941
192	헤타페	방문	0.315789
193	헤타페	안방	0.421053
194	호펜하임	방문	0.470588
195	호펜하임	안방	0.411765

196 rows × 3 columns

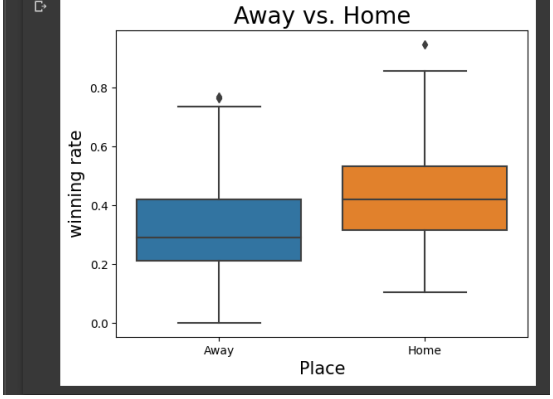
여기서 '장소' 열은 안방 경기와 방문 경기를 구분하는 역할을 합니다. 현재 데이터에는 승률 관련 정보가 없으니까 팀별로 안방 경기와 방문 경기 승률이 어떻게 되는지 계산해서 넣도록 하겠습니다. 축구에서 승률은 '승리 / 전체 경기 숫자'로 계산합니다. 따라서 현재 승리 여부가 0 아니면 1이니까 평균을 구하면 그 값이 바로 승률입니다.

1.) 홈경기와 원정경기 사이의 승률 차이의 유의성을 살펴본다.

1.1) Box-plot과 기술통계량을 통해 홈과 원정 승률 차이를 확인한다

전처리된 데이터 'df1'을 통해 안방 경기와 방문 경기 승률을 비교하는 'Boxplot'을 그려보겠습니다.

```
#PKG
import matplotlib.pyplot as plt
import seaborn as sns

# Boxplot
sns.boxplot(x = '장소', y = '승리',data = df1)
plt.title('Away vs. Home', fontsize=20)
plt.xlabel('Place', fontsize=15)
plt.ylabel('winning rate', fontsize=15)
plt.xticks([0, 1], ['Away', 'Home'])
plt.show()
```



안방(Home) 경기 승률이 높은 것처럼 보이는 건 사실입니다. 간단하게 평균을 내보면 안방 경기 승률 평균은 44.2%이고 방문 경기는 31.4%입니다.

```
[7] df1.groupby('장소', as_index = False).agg(mean = ('승리', 'mean'))
```

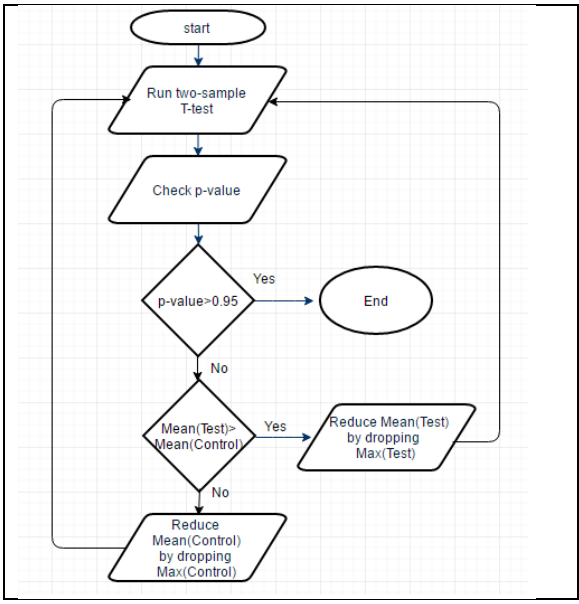
	장소	mean
0	방문	0.314397
1	안방	0.441927

그러면 12.8% 포인트 차이가 납니다. 이 차이가 통계적으로 의미가 있다고 할 수 있을까요? 혹시 이런 결과가 '어쩌다' 나타난 건 아닐까요?

1.2) T-Test를 통해 유의성 가설검정을 확인한다

$$H_0 : \mu_{홈팀승률} = \mu_{어웨이팀승률}$$

$$H_1 : \mu_{홈팀승률} > \mu_{어웨이팀승률}$$



$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{(s^2(\frac{1}{n_1} + \frac{1}{n_2}))}}$$

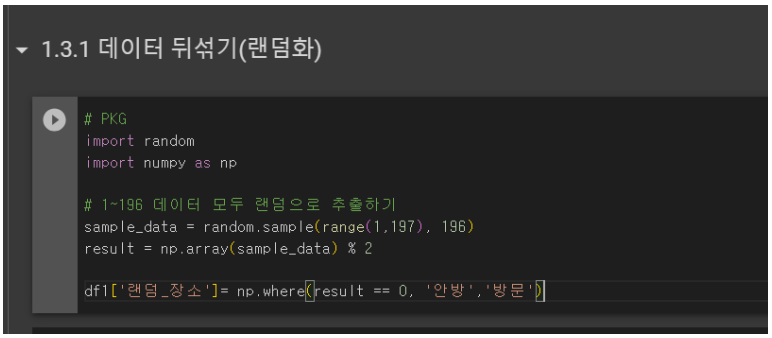
두 독립 표본의 평균 값을 비교하기 위해 t-검정법이 가장 많이 사용됩니다. 검정통계량과 p 값을 계산하고 신뢰구간을 구합니다. 만약 p-value가 0.05보다 작은 경우에 귀무가설을 기각하고 대립가설을 채택합니다.

위 가설검정을 수행했을 때 p-value는 0으로 0.05보다 작기 때문에 홈팀의 승률이 원정팀의 승률보다 유의미하게 크다고 결론 내릴 수 있습니다.

1.3) Permutation test를 우연성 통계적 가설검정

1.3.1) 데이터 뒤섞기(랜덤화)

이 결과가 어쩌다 나타난 건지 아닌지 알아보는 방법은 간단합니다. '장소'를 마구잡이로 섞어 보면 됩니다. 즉, 실제 경기 장소와 관계 없이 '이건 안방 경기, 저건 방문 경기' 라고 무작위 지정하는 겁니다. 만약 경기 장소가 승률에 영향을 끼치는 요소라면 이렇게 장소를 마음대로 배치했을 때는 현재 승률 분포를 유지 할 수 없을 것입니다.



행의 개수 196행을 1부터 196개를 무작위 배열을 한 후에 2로 나누었을 때 나머지 0과 1을 추출하였습니다. 나머지 값이 0이면 '안방', 1이면 '방문'으로 배정하였습니다. 랜덤으로 배정한 값들을 '랜덤_장소'라는 열로 생성하였습니다.

1.3.2) Boxplot & 기술통계량

이번에는 (가상) 방문 경기 승률이 더 높아 보입니다.

실제 코드를 써서 확인해보면 가상 방문 경기 승률은 40.6%, 안방 경기 승률은 35.1%입니다.

정말 경기 장소가 승률에 영향을 미쳤는지 검정(Testing)의 단계가 필요합니다.

1.3.3) 순열검정

앞선 단계에서 경기 장소를 한 차례 뒤섞었더니 방문 경기 승률이 오히려 높아지는 현상을 목격했습니다. 하지만 이 작업을 여러 번 반복하면 결과가 달라질 수 있습니다.

[가설검정]

$$H_0 : \mu_{\text{홈팀승률}} = \mu_{\text{어웨이팀승률}}$$
$$H_1 : \mu_{\text{홈팀승률}} > \mu_{\text{어웨이팀승률}}$$

위 가설을 순열검정(Permutation test)로 검정하기 위해 1000번의 1.3.1단계의 랜덤화 과정을 거칩니다. 196개의 데이터를 랜덤화 하여

$$Permute\_Statistic = \mu_{\text{랜덤 홈팀승률}} - \mu_{\text{랜덤 어웨이 팀승률}}$$

값을 얻습니다. 그리고 1000번 반복하여 1000개의 Permute_Statistic 값들을 얻습니다.

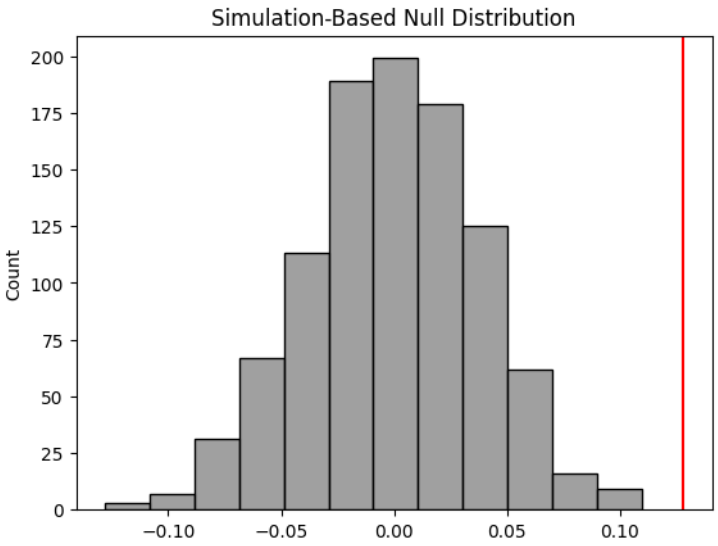
$$Permute\_Statistics = [Permute\_Statistic1, \quad Permute\_Statistic2, \quad \dots \quad , Permute\_Statistic1000]$$

즉, Permute_Staistics 값은 1000번 재표집 했을 때 안방 경기 승률 평균과 방문경기 승률이 얼마나 차이가 나는지 계산한 결과입니다.

[알고리즘]

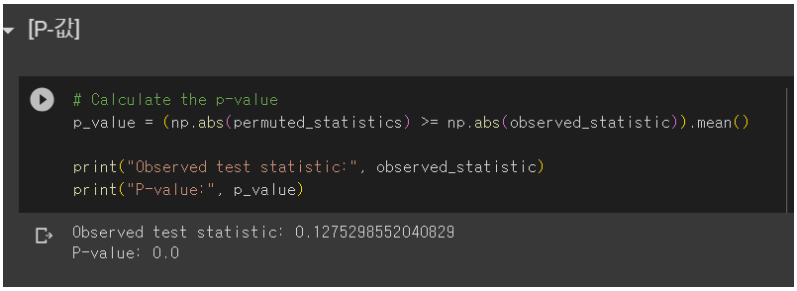
```python
import numpy as np

# Define two groups of data
away = df1[df1['장소'] == '방문']['승리']
home = df1[df1['장소'] == '안방']['승리']

# Calculate the observed test statistic
observed_statistic = np.mean(home) - np.mean(away)

# Combine the two groups
combined = np.concatenate([home, away])

# Number of permutations
num_permutations = 1000

# Initialize an array to store permuted test statistics
permuted_statistics = np.zeros(num_permutations)

# Permutation test
for i in range(num_permutations):
    # Randomly permute the combined data
    permuted_data = np.random.permutation(combined)

    # Split the permuted data into two groups
    permuted_home = permuted_data[:len(home)]
    permuted_away = permuted_data[len(home):]

    # Calculate the test statistic for the permuted data
    permuted_statistic = np.mean(permuted_home) - np.mean(permuted_away)
```

Permute_Statistics 값으로 히스토그램을 확인할 수 있습니다.



이 그래프에서 가로 축이 0이라는 건 안방 경기 승률 차이가 나지 않는다는 뜻입니다. 이럴 경우가 제일 많습니다. H_0가 '경기 장소에 관계없이 승률이 일정하다.' 였습니다. 이 가정이 맞다고 해도 안방 경기 승률과 방문 경기 승률 차이가 날 수도 있습니다. 대신 양쪽 끝으로 갈수록 안방 승률 평균이 유독 높거나 방문 경기 승률 평균이 유독 높은 사례는 점점 줄어듭니다.

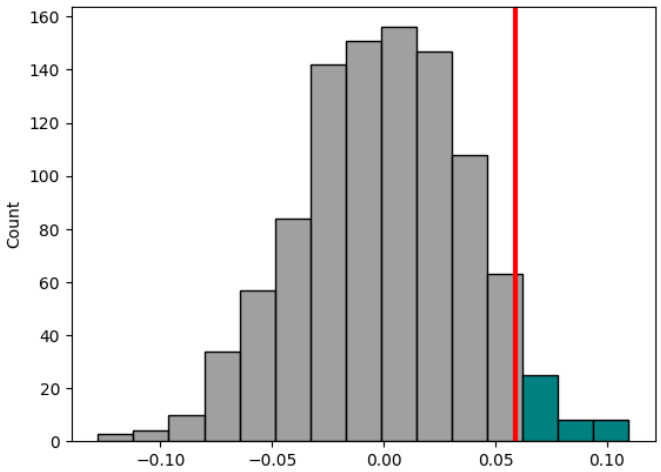
맨 처음에 (안방 경기 – 방문 경기) 승률 차이는 12.8% 포인트였습니다. 이 지점을 수직선으로 그래프에 표시하면 이렇게 됩니다.



경기 장소를 1000번 뒤섞었을 때 안방 경기 승률 평균과 방문 경기 승률 평균이 12.8% 포인트 차이가 난 적은 한 번도 없었습니다. '경기 장소와 승률 관계 없이 승률은 일정하다'는 H_0이 맞다면 이 결과는 매우 예외적입니다. 이 정도면 'H_0를 기각할 수 있는 증거가 나왔다'고 말 할 수 있습니다. 즉, 1000번의 시뮬레이션 결과 홈 경기에서 승률이 원정 경기에서 승률보다 유의미하게 높다고 평가 할 수 있습니다.

[P-값]

p-값이 0.05이하 일 때부터 홈 경기 승률과 원정 경기 승률의 차이가 있다고 할 수 있습니다. P-값이 0.05 이하가 되는 검정통계량을 찾기 위해 1000개 permuted_statistic 값의 백분율 95%의 값을 찾아야합니다. 그 결과 차이값이 5.9% 이상일 때부터 홈 경기 승률과 원정 경기 승률의 차이가 유의하다고 확인 할 수 있습니다. 즉, 1000개 중 50개 값이 0.059보다 큽니다.



히스토그램으로 확인 했을 때 검정통계량 값이 빨간 선 우측 영역(초록색)에 위치했을 때 귀무가설을 기각하고 대립가설을 채택할 수 있습니다. 반대로 빨간 선의 좌측 영역(회색)에 위치한다면 귀무가설을 기각할 수 없습니다.

2.) 코로나전 홈경기 승률과 코로나 이후 홈경기 승률 차이의 유의성을 살핀다.

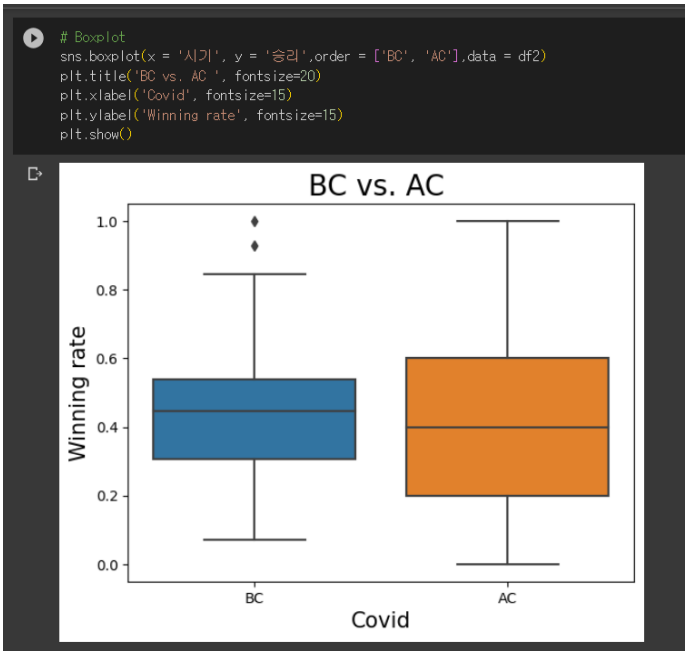
```
df2 = df[(df['장소'] == '안방') & (df['리그'] != '리그1')].groupby(['팀','시기'], as_index = False)
df2 = df2[['승리']].agg(승리 = ('승리', 'mean'))
df2
```

이 작업을 진행하기 위해 데이터 전처리가 필요합니다. 일단 안방 경기만을 골라냅니다. 그리고 프랑스 리그앙(리그1) 경기 경과는 제외해야 합니다. 나머지 4대 리그는 코로나19가 잠잠해지자 결국 무관중 상태로 모든 일정을 소화한 반면 리그앙은 서둘러 일정을 마무리했기 때문입니다.

2.1) Box-plot과 기술통계량을 통해 홈과 원정 승률 차이를 확인한다.

이 작업을 진행하기 위해서 안방
전처리된 데이터 'df2'을 통해 안방 경기와 방문 경기 승률을 비교하는 'Boxplot'을 그려보겠습니다.

코로나 전(BC) 안방 경기 승률이 코로나 전(AC) 안방 경기 승률보다 살짝 높은 것처럼 보입니다. 다만 육안으로 보았을 때 큰 차이가 있는지는 잘 모르겠습니다.




```
# Boxplot
sns.boxplot(x = '시기', y = '승리',order = ['BC', 'AC'],data = df2)
plt.title('BC vs. AC ', fontsize=20)
plt.xlabel('Covid', fontsize=15)
plt.ylabel('Winning rate', fontsize=15)
plt.show()
```

기술통계량을 비교하였을 때 BC의 평균은 0.439, AC의 평균은 0.421입니다. 코로나 이후에 1.8%포인트 하락한 사실을 알 수 있습니다.

```
[ ] df2.groupby('시기', as_index = False).agg(mean = ('승리', 'mean'))
```

	시기	mean
0	AC	0.421337
1	BC	0.438894

2.2) T-Test를 통해 유의성 가설검정을 확인한다.

$$H_0 : \mu_{AC\ 홈팀승률} = \mu_{BC\ 홈팀승률}$$
$$H_1 : \mu_{AC\ 홈팀승률} < \mu_{BC\ 홈팀승률}$$

두 독립 표본의 평균 값을 비교하기 위해 t-검정법이 가장 많이 사용됩니다. 검정통계량과 p 값을 계산하고 신뢰구간을 구합니다. 만약 p-value가 0.05보다 작은 경우에 귀무가설을 기각하고 대립가설을 채택합니다.

```
[ ]  # data
     AC = df2[df2['시기'] == 'AC']['승리']
     BC = df2[df2['시기'] == 'BC']['승리']

     # PKG
     from scipy import stats
     stat_value, p_value = stats.ttest_ind(AC, BC,
                                           alternative = 'less')
     print('P-Value :', round(p_value,3))

     P-Value : 0.321
```

위 가설검정을 수행했을 때 p-value는 0.321로 0.05보다 크기 때문에 코로나 이전 홈팀의 승률이 코로나 이후 홈팀의 승률보다 유의미하게 크다고 보기 어렵습니다.

2.3) Permutation test를 우연성 통계적 가설검정

2.3.1) 데이터 뒤섞기(랜덤화)

이 결과가 어쩌다 나타난 건지 아닌지 알아보는 방법은 간단합니다. '시기'를 마구잡이로 섞어 보면 됩니다. 즉, 실제 경기 시기와 관계 없이 '이건 AC 경기, 저건 BC 경기' 라고 무작위 지정하는 겁니다. 만약 경기 시기가 승률에 영향을 끼치는 요소라면 이렇게 시기를 마음대로 배치했을 때는 현재 승률 분포를 유지 할 수 없을 것입니다.

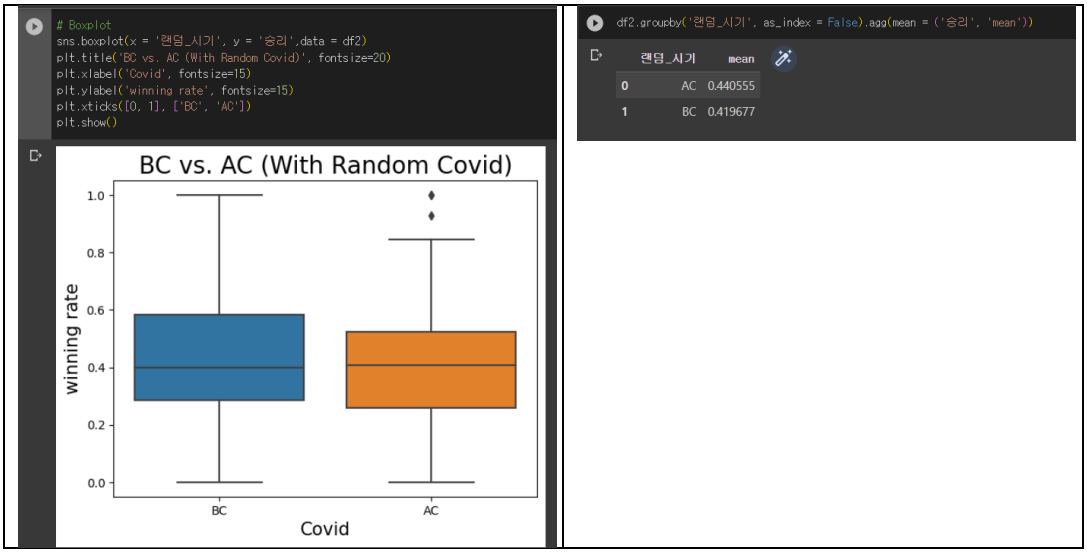
데이터 뒤섞기

```
[ ]  # PKG
     import random
     import numpy as np

     # 1~156 데이터 모두 랜덤으로 추출하기
     sample_data = random.sample(range(1,157), 156)
     result = np.array(sample_data) % 2

     df2['랜덤_시기']= np.where(result == 0, 'BC','AC')
```

행의 개수 156행을 1부터 156개를 무작위 배열을 한 후에 2로 나누었을 때 나머지 0과 1을 추출하였습니다. 나머지 값이 0이면 'BC', 1이면 'AC'로 배정하였습니다. 랜덤으로 배정한 값들을 '랜덤_장소'라는 열로 생성하였습니다.

2.3.2) Boxplot & 기술통계량

이번에는 (가상) 방문 경기 승률이 더 높아 보입니다.



실제 코드를 써서 확인해보면 가상 BC홈 경기 승률은 42%, 가상 AC홈 경기 승률은 44.1% 입니다.

정말 경기 장소가 승률에 영향을 미쳤는지 검정(Testing)의 단계가 필요합니다.

2.3.3) 순열검정

앞선 단계에서 경기 장소를 한 차례 뒤섞었더니 AC홈 경기 승률이 오히려 높아지는 현상을 목격했습니다. 하지만 이 작업을 여러 번 반복하면 결과가 달라질 수 있습니다.

[가설검정]

$$H_0 : \mu_{AC\ 홈팀승률} = \mu_{BC\ 홈팀승률}$$
$$H_1 : \mu_{AC\ 홈팀승률} < \mu_{BC\ 홈팀승률}$$

위 가설을 순열검정(Permutation test)로 검정하기 위해 1000번의 2.3.1단계의 랜덤화 과정을 거칩니다. 156개의 데이터를 랜덤화 하여

$$Permute\_Statistic = \mu_{AC\ 홈팀승률} - \mu_{BC\ 홈팀승률}$$

값을 얻습니다. 그리고 1000번 반복하여 1000개의 Permute_Statistic 값들을 얻습니다.

$$Permute\_Statistics$$
$$= [Permute\_Statistic1,\quad Permute\_Statistic2,\quad \ldots\quad , Permute\_Statistic1000]$$

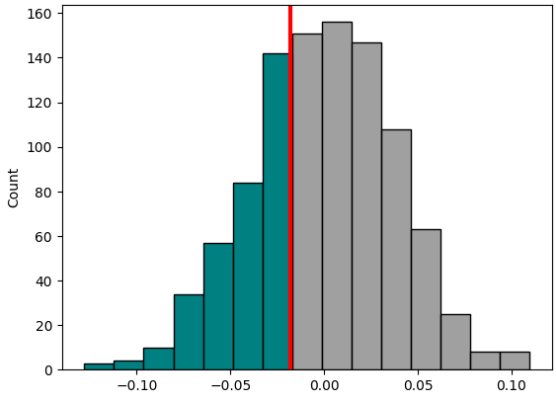
즉, Permute_Staistics 값은 1000번 재표집 했을 때 AC 홈팀 승률 평균과 BC 홈팀 승률이

얼마나 차이가 나는지 계산한 결과입니다.

[알고리즘]

```
import numpy as np

# Define two groups of data
AC = df2[df2['시기'] == 'AC']['승리']
BC = df2[df2['시기'] == 'BC']['승리']

# Calculate the observed test statistic
observed_statistic = np.mean(AC) - np.mean(BC)

# Combine the two groups
combined = np.concatenate([AC,BC])

# Number of permutations
num_permutations = 1000

# Initialize an array to store permuted test statistics
permuted_statistics = np.zeros(num_permutations)

# Permutation test
for i in range(num_permutations):
    # Randomly permute the combined data
    permuted_data = np.random.permutation(combined)

    # Split the permuted data into two groups
    permuted_AC = permuted_data[:len(AC)]
    permuted_BC = permuted_data[len(AC):]

    # Calculate the test statistic for the permuted data
    permuted_statistic = np.mean(permuted_AC) - np.mean(permuted_BC)

    # Store the permuted test statistic
    permuted_statistics[i] = permuted_statistic
```

Permute_Statistics 값으로 히스토그램을 확인할 수 있습니다.



이 그래프에서 가로 축이 0이라는 건 AC와 BC 홈경기 승률 차이가 나지 않는다는 뜻입니다. 이럴 경우가 제일 많습니다. H_0가 '코로나19 유행 전후 안방 승률에는 변화가 없다' 입니다.

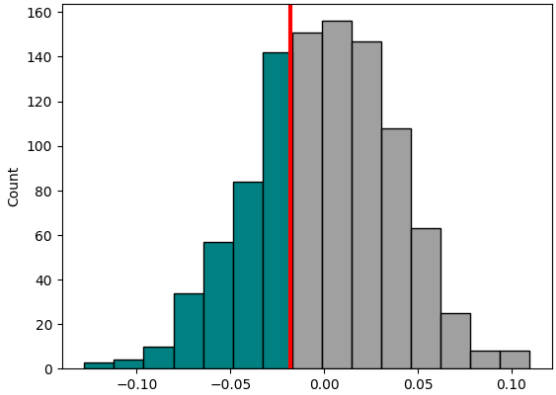
맨 처음에 (AC 홈팀승률 평균 – AC 홈팀승률 평균) 승률 차이는 -1.8% 포인트였습니다. 이 지점을 수직선으로 그래프에 표시하면 이렇게 됩니다.

경기 장소를 1000번 뒤섞었을 때 안방 경기 승률 평균과 방문 경기 승률 평균이 -1.8% 포인트 차이가 난 적은 꽤 많아 보입니다. '코로나19 유행 전후 안방 승률에는 변화가 없다'는 H_0가 맞는 걸로 예상됩니다. 이 정도면 'H_0를 채택 할 수 있는 증거가 나왔다'고 말할 수 있습니다. 즉, 1000번의 시뮬레이션 결과 AC홈 경기에서 승률이 BC홈 경기에서 승률과 유의미한 차이가 없다고 평가 할 수 있습니다.

[P-값]

p-값이 0.05이하 일 때부터 AC홈 경기 승률과 BC홈 경기 승률의 차이가 있다고 할 수 있습니다.

```
import seaborn as sns
from matplotlib.patches import Rectangle

p=sns.histplot(permuted_statistics, color='gray', bins = 15)
plt.axvline(round(observed_statistic, 3), color = 'r', linewidth = 3)

for rectangle in p.patches:
    if rectangle.get_x() <= round(observed_statistic, 3) :
        rectangle.set_facecolor('teal')

Rectangle(xy=(round(observed_statistic, 3), 0),
          width=1.28523-round(observed_statistic, 3),
          height=180, angle=0).set_facecolor('teal')
```



P-값을 구하기 위해 -0.018 이하가 되는 permuted_statistics 값은 1000개 중 346개로 p-값이 34.6%라고 할 수 있습니다. 따라서 히스토그램의 빨간선 좌측 부분인 초록색 영역인 p값이 5%보다 크기 때문에 H_0 기각할 수 없습니다. 즉 코로나가 안방경기 승률에 영향을 미쳤다고 보기는 어렵습니다.

```
[ ]  p_value = round((permuted_statistics < -0.018).mean(),3)
     print(p_value)

     0.346
```

3. 결론

지금까지 본 보고서에서 홈 경기 승률과 원정 경기의 승률 차이를 알아보고, 코로나 전후로 홈 경기의 승률 차이에 변화가 있는지 유무를 살펴보았습니다. 검정 방법으로 t-검정으로 두 집단의 평균 사이의 유의성을 비교해 보았고, Permutation-test(순열검정)을 통해 사건의 우연성을 검정하였습니다.

첫 번째 분석결과 홈 경기의 승률(44.2%)과 원정 경기 승률(31.4%) 사이에 유의미한 차이가 있는 것으로 확인되었습니다. 둘 사이에 차이는 12.8%포인트 였습니다. T-검정을 통해 두 집단 사이의 승률 차이는 유의미하고, Permutation-test으로 1000번의 시뮬레이션 알고리즘을 분석한 결과 우연적으로 일어난 사건이 아니란 것을 알았습니다.

즉, 홈에서 100번 경기를 펼쳤을 때 승리 할 기대 경기수는 44경기입니다. 반면, 원정에서 100번 경기를 펼쳤을 때 승리 할 기대 경기수는 31경기 입니다. 유럽리그 경기에서는 통상적으로 홈과 원정에서 모두 한 번씩 경기를 치르게 됩니다. 홈팀은 홈 경기장에서 응원석에서 뜨거운 응원과 열기를 받았을 때 승리할 힘을 받는다고 판단 할 수 있습니다. 따라서 홈에서 경기를 할 때일수록 홈팀은 승리하여 승점을 확실하게 확보할 필요성이 있습니다(강팀의 경우는 홈과 원정 경기 사이 승률 차이가 별로 없을 수 있지만 약팀의 경우는 차이가 클 수 있다).

두 번째 분석 결과 코로나 이전의 홈팀 승률과 코로나 이후의 홈팀 승률 차이가 없는 것으로 확인되었습니다. T-검정과 Permutation-test 모두에서 유의미한 차이가 없는 것으로 확인되었습니다.

첫 번째 분석에서 홈팀 관중의 효과로 홈 어드밴티지 효과를 받아 홈 경기 승률이 상승한다는 것을 확인했습니다. 따라서 무관중으로 치러진 홈 경기에서는 홈 팬들의 응원을 받을 수 없으니 코로나 이후의 홈 경기 승률이 유의미하게 낮을 것으로 판단하였지만 결과는 전혀 그렇지 않았습니다.

관련 기사를 찾아본 결과 응원도 큰 영향을 미치지만 팀에 친숙한 환경 같은 다른 요소가 홈 경기의 이점이 존재한다고 합니다. 결론적으로 홈에서 경기를 하였을 때 승률이 더 높은 것은 사실이지만 그 원인이 홈팀 관중의 영향으로만 판단하기 어렵다고 볼 수 있습니다. 즉, 홈 경기장의 관중의 함성이 없더라도 '홈 어드밴티지'는 존재한다는 결론을 내릴 수 있습니다.

4. 참여후기

4학년 2학기를 맞아 마지막 과목으로 <컴퓨팅사고와 데이터분석>을 수강하였습니다. 이번 강의를 수강하며 기존에 내가 알고있던 파이썬 코딩 실력에 새로운 지식과 관점들을 쌓을 수 있었습니다. 그리고 머신러닝과 AI의 알고리즘에 대한 매커니즘도 보다 쉽게 이해할 수 있었습니다.

기말대체 과제로 평소에 즐겨보는 해외축구에 통계적 기법을 적용하고 응용해보았습니다. 먼저 주어진 데이터를 분석의 목적에 맞게 전처리하는 과정을 거치고, 히스토그램과 박스플랏을 그려가며 EDA 관점으로 분석을 시작했습니다. 기술통계량을 분석하였고, 가장 보편적인 t-검정법을 적용하여 통계분석을 적용하였습니다. 그리고 응용의 영역으로 순열검정(Permutation Test)를 적용하여 우연성까지 검정을 하였습니다. 검정통계량과 P-값을 기반으로 가설을 검정하는 결과를 얻어냈습니다.

이번 기말대체 과제를 통해 평소에 관심 갖는 분야에 직접 데이터를 분석해보고 작은 분석 레포트를 만들어보는 좋은 기회였습니다. 리포트를 제작하며 Python을 통해 데이터를 정제하고 분석하는 역량을 많이 길렀습니다.

[참고문헌]

1. 친절한 스포츠데이터 with R(황인규, 2021)
2. 코로나 시대 홈 관중 함성 없어도 '홈 어드밴티지'는 있었다. 2021.04.05 https://www.dongascience.com/news.php?idx=45272