Data Science Term Project

권 영 근 교수님
20172058 김 상 익
20172104 이 대 성
20172099 여 지 원

울산대학교
UNIVERSITY OF ULSAN

도입

HOME Advantage

우리는 좋아하는 스포츠팀의 경기 결과를 예측 할 때가 있습니다.
그럴 때 고려하는 요인이 여러 가지가 있지만, 사람들은 대게 상대 팀이 어느 팀인지와 홈경기인지 원정 경기인지를 가장 먼저 고려해 볼 것입니다.

여기서 저희는 경기장이 어디인지가 왜 그렇게 큰 고려사항이 되는가에 대해 의문점을 가지게 되었습니다. 단순히 Home Advantage 이라는 이점이 절대적으로 작용한다는 불문율을 만들어 내는 것이 성급한 일반화의 오류가 아닐까? 하는 의문점이 들었습니다.

프로 스포츠 경기에서는 양 팀 간의 실력 차이, 즉 객관적인 전력 차이로 경기 결과가 결정되는 요인이 크다고 생각되는데 홈과 원정의 차이에 따라 경기 결과가 달라지진 않을 것이라고 생각을 했습니다.
이러한 의문점을 데이터 분석을 통해 실제 경기 사례들과 수치화된 데이터를 통해 분석을 하면 좀 더 명확히 알아 갈 수 있다고 생각합니다.

가설

저희는 경기장의 위치, 즉 홈과 원정의 여부가 경기 결과에 큰 영향을 주지 못 한다고 가설을 세우게 되었고, 이를 토대로 실제 데이터를 분석을 해보겠습니다.

- **Null hypothesis**
  - — Home팀의 평균 승률과 Away팀의 평균 승률 사이에는 유의미한 차이가 없다
- **Alternative hypothesis**
  - — Home팀의 평균 승률과 Away팀의 평균 승률 사이에는 유의미한 차이가 있다.

· Test statistic

— Home 과 Away 경기의 평균 승률의 차이

Country	League	Year	Team	Games	HomeWins	HomeDraw	HomeLoss	AwayWins	AwayDraw	AwayLoss	HomePct	AwayPct
Algeria	Algeria-Ligue-1	2010	MC Oran	34	10	5	2	0	6	11	58.82%	0.00%
Algeria	Algeria-Ligue-1	2010	NA Hussein Dey	34	3	5	9	0	5	12	17.65%	0.00%
Algeria	Algeria-Ligue-1	2011	MC El Eulma	29	9	3	2	0	6	9	64.29%	0.00%
Algeria	Algeria-Ligue-1	2011	Annaba	29	10	5	0	0	1	13	66.67%	0.00%
Algeria	Algeria-Ligue-1	2012	Khroub	30	7	5	3	0	5	10	46.67%	0.00%
Algeria	Algeria-Ligue-1	2012	NA Hussein Dey	29	5	5	4	0	5	10	35.71%	0.00%
Algeria	Algeria-Ligue-1	2012	Sa?da	30	6	6	3	0	1	14	40.00%	0.00%
Algeria	Algeria-Ligue-1	2014	A?n Fakroun	29	5	1	8	0	4	11	35.71%	0.00%
Algeria	Algeria-Ligue-1	2016	USM Blida	22	5	5	1	0	6	5	45.45%	0.00%
Algeria	Algeria-Ligue-1	2016	ASM Oran	22	5	2	4	0	1	10	45.45%	0.00%

... (6534 rows omitted)

· Data Column별 설명

— **Country Column** : 해당 국가

— **League** : 진행 리그 명

— **Year** : 리그가 진행된 년도

— **Team** : 리그에 참가한 팀명

— **Games** : 해당 Team이 리그에서 진행한 총 경기 수

— **HomeWins** : 해당 Team이 홈에서 이긴 경기 수

— **HomeDraw** : 해당 Team이 홈에서 비긴 경기 수

— **HomeLoss** : 해당 Team이 홈에서 진 경기 수

— **AwayWins** : 해당 Team이 원정에서 이긴 경기 수

— **AwayDraw** : 해당 Team이 원정에서 비긴 경기 수

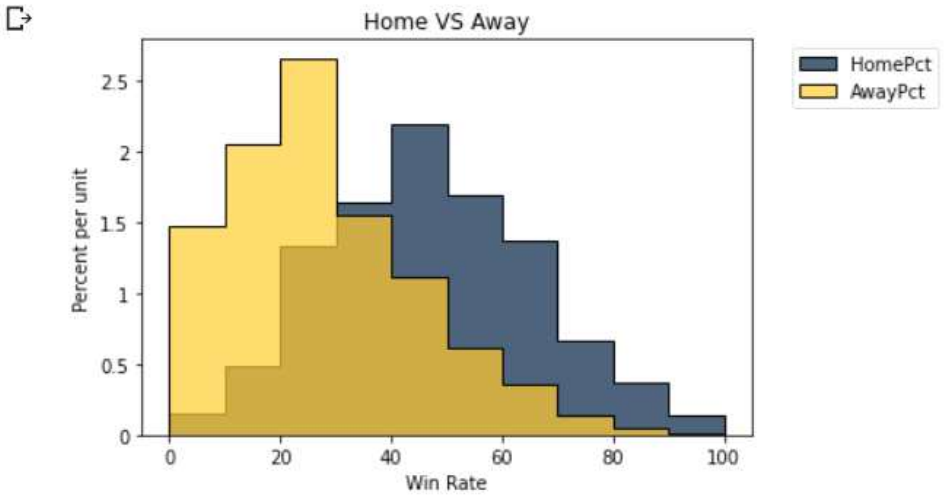
— **AwayLoss** : 해당 Team이 원정에서 진 경기 수

— **HomePct** : 해당 Team의 홈에서 승률

— **AwayPct** : 해당 Team의 어웨이에서 승률

저희가 활용할 원본 데이터는 국가, 년도, 리그별로 진행된 축구 경기 결과를 담고 있습니다. 이 데이터에 홈과 원정의 승률을 의미하는 HomePct column과 AwayPct column을 추가하였습니다. 또한, 데이터의 신뢰성을 위하여 Games column에서 경기 수가 20경기 미만인 row는 삭제하여 총 6,544개의 row를 활용하여 분석하였습니다.

• Win Rate Of Home vs Away

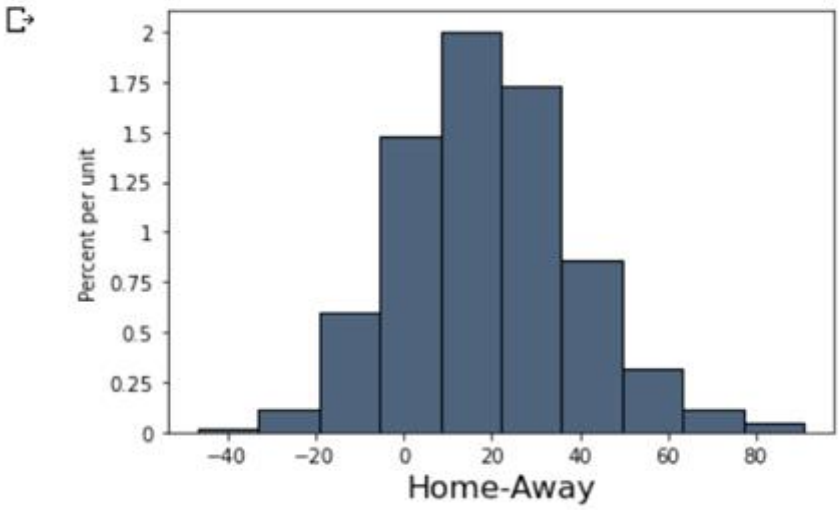


데이터 분석에서 가장 먼저 홈경기와 원정경기의 승률을 비교해보기 위해서 저희가 활용할 soceer Table에서 HomePct column과 AwayPct column을 뽑아내서 Overlaid 하게 Histogram 으로 시각화 하였습니다.

그래프의 x축은 승률을 의미합니다. 홈경기에서의 승률을 나타내는 파란색 Histogram과 원정경기에서의 승률을 나타내는 노란색 Histogram을 비교해 보았습니다.

원정경기에서의 승률은 조금 더 0 에 가까운 쪽으로 치우쳐져 있습니다. 이를 미루어 보았을 때 원정경기에서의 승률보다 홈경기 에서의 승률이 더 높다는 것을 확인할 수 있습니다.

- **Home – Away**



이번에는 홈경기 승률에서 원정경기 승률을 뺀 수치를 계산하여 Histogram으로 표현하였습니다. x축은 [홈경기 승률 – 원정경기 승률]을 의미합니다. 대부분의 수치가 0 이상으로 치우쳐 있으므로 홈경기의 승률이 원정경기의 승률보다 대체로 높다는 것을 다시 한 번 확인할 수 있습니다.

하지만 이 수치가 정말 의미 있는 값인지 알아보는 과정이 필요하다는 판단하에 저희는 홈경기와 원정경기의 승률을 무작위 shuffling 하여 관측값과 비교하는 과정을 실행하였습니다.

- **New Table**

shuffling을 위하여 홈경기와 원정경기의 승률을 담은 새로운 Table을 생성하였습니다. 홈경기의 승률과 원정경기의 승률을 합쳐서 만든 테이블이기 때문에 기존 데이터값의 2배의 row를 확인할 수 있습니다.

- **Stadium column**

  — 해당 승률의 홈, 원정여부

- **WinRate column**

  — 승률

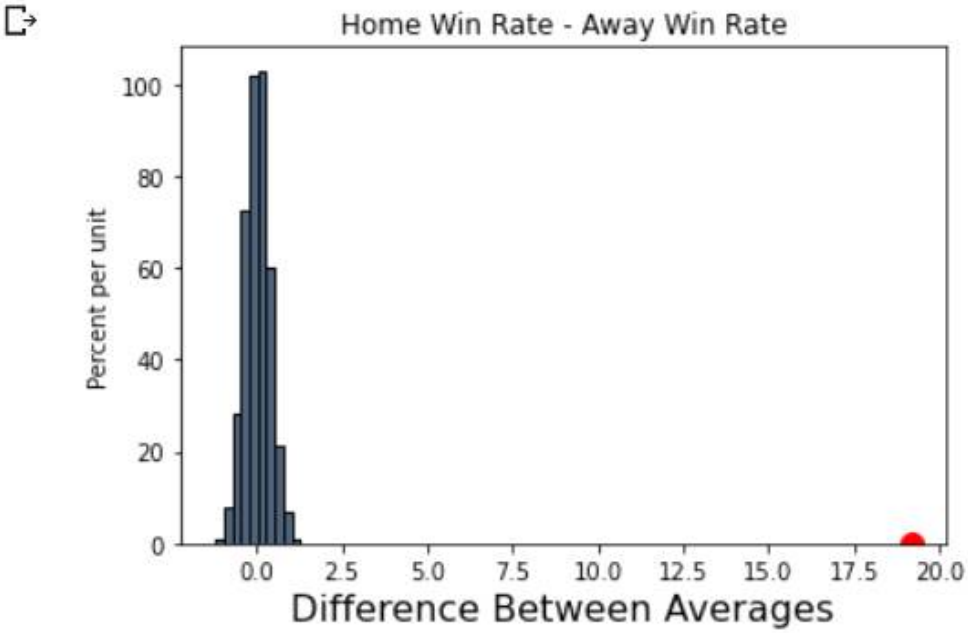
Stadium	WinRate
Home	58.8235
Home	17.6471
Home	64.2857
Home	66.6667
Home	46.6667
Home	35.7143
Home	40
Home	35.7143
Home	45.4545
Home	45.4545

... (13078 rows omitted)

Stadium	WinRate mean
Away	27.2962
Home	46.4616

19.165442874555374

데이터를 홈, 원정 정보를 담은 Stadium column으로 grouping 하여 평균값을 구하였고 [홈경기 승률의 평균–원정경기 승률 평균]의 값을 관측 값으로 설정하였습니다. 이를 토대로 홈경기의 승률과 원정경기의 승률을 shuffling 하여 홈경기 승률의 평균–원정경기 승률의 평균을 구하는 작업을 5000번 반복하여 이 값에 대한 Histogram을 그려 표현해 보겠습니다.



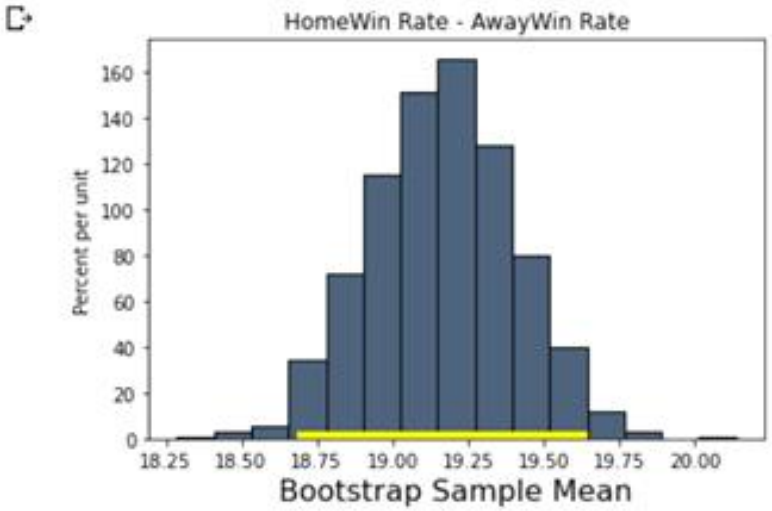
위 히스토그램의 x축은 홈경기의 승률과 원정경기의 승률을 random shuffling 하여 만들어진 shuffled data의 [홈경기의 평균 승률 – 원정경기의 평균 승률]을 나타냅니다. 우측에 있는 붉은색 반원은 앞서 구한 관측값을 표시한 것입니다.

히스토그램 x축의 대부분은 0에 밀집해서 분포한 것에 비해 관측값은 크게 차이가 나는 것을 확인할 수 있습니다.
또한 관측값 보다 크거나 같은 값이 없으므로 p-value = 0입니다.
따라서 대부분의 팀이 홈경기에서 승률이 더 높았다는 점은 Random 한 값이 아니라는 것을 알 수 있습니다.

이 분석을 통하여 홈경기의 승률이 원정경기의 승률과 유의미한 차이가 없을 것으로 추측하였던 null hypothesis보다 alternative hypothesis쪽에 힘이 실리게 됩니다.

하지만 이렇게 높은 수치의 의미있는 관측값은 저희의 데이터에서만 두드러지는 값일 수 있으므로 모집단을 분석해볼 필요가 있습니다.

전 세계의 모든 축구 경기 결과를 담은 데이터를 분석할 수 없으므로 저희는 bootstrap방법을 이용하여 모집단의 데이터를 분석해보겠습니다.



랜덤샘플을 뽑아 (홈경기 평균 승률 – 원정경기 평균 승률)을 구하는 작업을 5000번 실시하였을 때의 히스토그램입니다.
그래프 하단의 노란색 선으로 95%의 신뢰구간을 표시하였습니다.

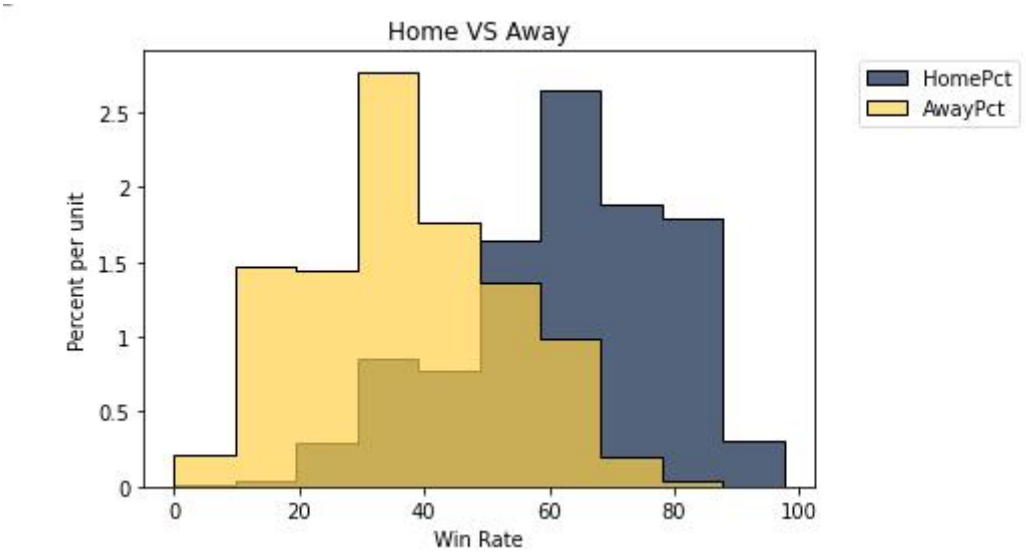
```
array([18.71136023, 19.62008697])
```

위 값은 모집단을 추측해보았을 때 95%의 신뢰구간입니다. 높은 확률로 모집단의 관측값도 이 범위 안에 있을 것으로 추측됩니다. 물론 저희가 사용한 데이터의 관측값도 모집단의 신뢰구간 안에 포함됩니다.
bootstrap 방법을 이용한 분석 이후에, 홈팀의 평균 승률과 원정팀의 평균 승률 간에는 유의미한 차이가 있다는 것이 더욱 확실해졌습니다.

• Other Sports?

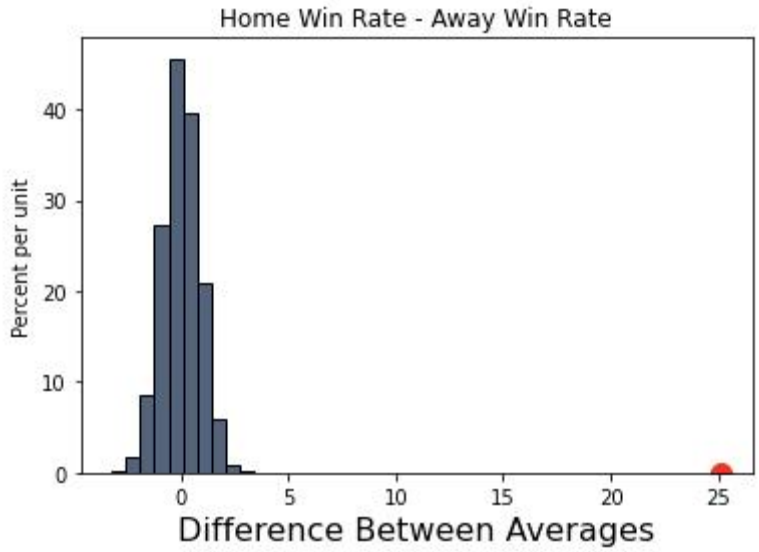
또한 저희는 이런 현상이 비단 축구경기에서 뿐만 아니라 다른 스포츠 종목에서도 이러한 차이가 날 것 이라고 예상하여 NBA에 대한 데이터들도 앞서 분석한 Soccer 데이터와 같은 방법으로 분석 해보았습니다.

• Win Rate Of Home vs Away



NBA table은 Soccer table과 달리 Draw column이 없어 Soccer table보다 좀 더 극단적인 Histogram이 나왔습니다. 이 또한 우연에 의한 승률 차이인지 확인하기 위해 Soccer table을 분석한 것과 같이 shuffling을 통하여 관측값과 비교해보았습니다.

- **Shuffling**



p-value를 계산하였을 때 0이 나왔으며 Soccer data와 같이 이러한 Home과 Away에서의 승률 차이는 절대 우연에 의한 것이 아님을 알 수 있었습니다.

결론

데이터를 다양한 방법으로 분석해본 결과 최종적으로 앞서 저희가 세운 null hypothesis를 reject하게 되었고 홈팀의 평균 승률과 원정팀의 평균 승률 사이에는 유의미한 차이가 있다는 alternative hypothesis가 합당하다는 결론을 내었습니다.

따라서 홈에서 경기를 치르는 것이 원정경기보다 평균으로 승률이 더 높다는 것을 알게 되었습니다. 하지만 홈에서 승률이 0%인 팀이 존재하고 원정에서 승률이 100%인 팀이 존재하는 것으로 보았을 때 홈에서 경기를 치르는 것이 유리하긴 하나, 스포츠의 객관적인 전력 차가 매우 크다면 이러한 홈의 이점만으로는 극복하기 힘들다고 볼 수 있습니다.

그러나 이러한 극단적인 전력 차의 경우 또한 포함한 데이터에서 수치상 유의미한 차이를 보였기 때문에 축구경기에서 홈에서 경기를 치르는 것이 경기결과에 상당히 큰 영향을 끼친다는 것은 확실합니다.

논의사항

홈의 승률이 높은 이유에 대해 논의해 보았을 때, 원정의 이동 거리 피로도와 경기장 피치의 익숙함 등의 요소가 있을 것이라고 생각하였습니다.

미비점

오직 홈에서의 승리만이 유의미한 결과가 아니라 약팀 입장에서는 비기는 것 또한 홈의 이점이 작용한 것이라고 볼 수 있는데
객관적으로 약팀과 강팀을 판단하는 기준이 모호하기 때문에 승률만 계산하였습니다.
비기게 되어도 승점은 나오기 때문에 획득한 승점을 기준으로 분석하면 우리가 분석 한 것보다는 더 큰 차이가 보였을 것이라고 예측됩니다.