ㅋ2023학년도 자연과학대학 학부생 연구 인턴십 결과보고서

【유형 Ⅰ(하계( )/동계(√)), 유형 Ⅱ(상반기( )/하반기( ))】

타과제 참여여부 (23.12.1.-24.2.29.) *참여율은 작성 불필요	N
KRI 번호	-

○ **연구주제(제목): 상대를 고려한 야구지표**

○ **제출자(소속, 학번, 이름): 통계학과 2018-11319 정진우**

○ **지도교수: 장원철 (인)**

==========

1. 서론

야구 경기는 투수와 타자의 맞대결의 반복으로 진행되며, 각 맞대결의 결과가 주자의 이동으로 경기가 끝나는 매우 특수한 상황을 제외한 거의 모든 경우에서 안타, 삼진, 홈런 등의 범주형 자료로 주어지게 된다. 이때 각 범주의 경우 투수의 승리 혹은 타자의 승리가 되는지의 여부가 뚜렷하게 구분이 되며 각 범주마다 실제 시즌을 진행하며 얻은 관측치의 횟수를 세어 그 값을 이용하여 선수들을 평가하는 지표인 타율, 출루율, 장타율 및 각종 세부 지표를 많이 생산해낸다. 이러한 지표는 선수를 평가하는데 가장 널리 사용된다. 이러한 방식의 지표들은 현대적인 측정도구 없이도 기록관이 기록하는 결과 값의 범주만으로 지표를 만들기 때문에 수집이 가장 용이하며, 이에 따라 가장 원시적인 지표로써 예전부터 사용되었다. 또한 현대에도 이 값들을 적절히 연산하는 방식을 통해 FIP, WRC 와 같이 선수를 평가하는 지표로 활용되고 있다.

본 연구에서는 각 범주의 횟수를 측정한 결과 값을 사용할 때 선수들이 지표를 얻을 때 만난 선수들의 실력을 고려해주어 선수들의 실제 능력에 더 가깝게 나타내는 새로운 지표를 제시하고자 한다. 특정 선수에 대하여 잘하는 상대를 만날 경우 좋은 결과를 낼 확률이 줄어들게 되고 반대의 경우 좋은 결과를 낼 확률이 올라간다는 점에서 해당 선수가 기록을 내는 과정에서 만난 상대들의 실력과 밀접한 관련을 갖고 있다. 반면에 현재 지표에 사용되는 각 범주의 관측 횟수의 경우 잘하는 상대/잘하지 못하는 상대를 얼마나 만났는지 반영되지 않고 있다. 즉 강한 상대를 많이 만난 경우 지표는 해당 선수의 본 실력보다 과소평가되어 있을 것이고, 약한 상대를 많이 만난 경우에 지표가 해당 선수의 본 실력보다 과대평가되어 있을 것이다.

만약 맞대결을 하게 되는 상대가 결정되는 방식이 시즌이 치러지는 동안 확률이 변함없는 다항분포에서 추출되어 결정되고, 시행횟수가 충분히 많다면 큰 수의 법칙에 의해 관측값이 실제 실력을 충분히 잘 표현해 줄 것은 자명하다. 하지만 실제 야구 시즌이 진행되는 과정을 살펴보면 (1) 메이저리그에서는 팀이 속한 리그와 지구에 따라 상대 팀을 만나는 경기 횟수가 다르게 설정되어 있고 (2) KBO 리그와 같은 경우에서 리그가 진행되는 중에 국제대회가 열려 리그의 최상급 선수들이 경기에 참여하지 않는 기간이 발생할 수 있으며 (3) 실력이 부족한 백업선수라 승부가 기울어진 상황에 주로 출전하여 그만큼 상대하는 선수들의 실력이 낮은 선수가 많은 등의 이유로 동일 분포에 대한 조건이 쉽게 깨지는 것을 볼 수 있다. 또한 이와 더불어 한 경기 내에서 같은 투수와 타자의 맞대결이 여러 번 일어나는 것이 잦은 점에서 다음 번 만나는 타석의 투수는 그 전번에 만난 투수일 확률이 매우 높은 종속적인 특성에 의해 관측 값이 실제 실력을 표현하

는 값에 빠르게 수렴하는 것을 방해하고 있다.

이에 따라 현실에서 관측되는 선수들의 각 범주에 속하는 값들은 해당 선수가 만난 맞상대의 실력에 의해 편향되어 있고 이 편향에 의해 각 선수들의 실제 실력을 왜곡하여 관측된 값으로 볼 수 있다. 이에 본 연구에서 이러한 편향이 얼마나 존재하는지 확인하고, 보정하는 방법을 고안하여 최종적으로 맞상대의 실력을 고려한 지표를 만들어보고자 한다.

2. 통계 모형 및 이론적 배경

2.1 통계 모형

m명의 투수 $A_1, A_2, \cdots, A_m$에 대하여 각각의 잠재변수 $\alpha_1, \alpha_2, \cdots, \alpha_m$

n명의 타자 $B_1, B_2, \cdots, B_n$에 대하여 각각의 잠재변수 $\beta_1, \beta_2, \cdots, \beta_n$

투수 A_i와 타자 B_j가 맞붙었을 때 사건 발생 확률 $P(A_i, B_j) = f(\alpha_i, \beta_j)$이라 하고

(f의 후보)

- Weighted Mean

  $f(\alpha_i, \beta_j) = w \times \alpha_i + (1-w) \times \beta_j$

- Inverse-Logit of Weighted Mean

  $$f(\alpha_i, \beta_j) = \frac{\exp(w \times \alpha_i + (1-w) \times \beta_j)}{1 + \exp(w \times \alpha_i + (1-w) \times \beta_j)}$$

(모수 추정)

위 모형에서 모수는 $\alpha_1, \alpha_2, \cdots, \alpha_m, \beta_1, \beta_2, \cdots, \beta_n, w$가 있으며,

실제로 투수 A_i와 타자 B_j가 맞붙었을 때 사건 발생 횟수와 사건이 발생하지 않은 횟수를 각각 x_{ij}, y_{ij}라고 할 때

$(Loss) = -\sum_{i=1}^{m}\sum_{j=1}^{n}[x_{ij}\log(P(A_i, B_j) + y_{ij}\log(1 - P(A_i, B_j)]$를 최소화하는 모수를 경사하강법을 이용하여 추정

(보정 지표)

- 보정 사건 발생 확률

  $A_1, A_2, \cdots, A_m$의 상대한 타자 수 $k_1, k_2, \cdots, k_m$

  $B_1, B_2, \cdots, B_n$의 상대한 투수 수 $l_1, l_2, \cdots, l_n$ 일 때

  $$(\text{투수 } A_i\text{의 보정 확률}) = \frac{\sum_{j=1}^{n} P(A_i, B_j) l_j}{\sum_{j=1}^{n} l_j}$$

  $$(\text{타자 } B_j\text{의 보정 확률}) = \frac{\sum_{i=1}^{m} P(A_i, B_j) k_i}{\sum_{i=1}^{m} k_i}$$

- 보정 사건 발생 수

  (투수 A_i의 보정 사건 발생 수)=(투수 A_i의 보정 확률)$\times k_i$

  (타자 B_j의 보정 사건 발생 수)=(타자 B_j의 보정 확률)$\times l_j$

2.2 모형의 이론적 배경

문제의 기본적인 상황은 관측된 투수-타자 간의 맞대결 결과를 이용하여 리그의 평균적인 상

대를 만났을 때 기대되는 특정 사건의 발생 확률 및 발생 횟수를 예측하는 것이다. 이때 일반화를 위하여 사건이라고 지칭한 것은 안타를 치는 경우, 출루를 하는 경우 또는 홈런을 치는 경우 등이 사용될 것이며, 안타를 치는 경우를 생각한다면 $P(A_i,B_j)$는 투수 A_i와 타자 B_j가 맞붙었을 때의 타율/피안타율의 추정값이 될 것이며 모수 추정시에 사용되는 x_{ij},y_{ij}의 경우 각각 안타 수와 타수 수에서 안타 수를 뺀 값이 될 것이다. 이 결과로 얻은 보정 확률은 타율/피안타율의 보정값을 의미하여 보정 사건 발생 수는 안타수/피안타수의 보정값을 의미하게 될 것이다. 같은 방식으로 적절한 x_{ij},y_{ij}설정으로 출루율/피출루율/출루 횟수/피출루 횟수, 홈런 확률/피홈런 확률/홈런수/피홈런수의 보정값을 구하는 모델이 된다.

모델에서 기본적으로 가정하는 것은 각 선수들의 실제 능력을 지시하는 잠재 변수가 있으며 이 값을 이용하여 임의의 투수와 타자가 맞붙었을 때 사건이 발생할 확률을 예측하게 된다. 각 선수들에게 잠재 변수를 각각 할당한 것은 타자 입장에서 생각했을 때 얻은 지표를 맞상대한 투수들의 실력만큼 보정하는 것이 목적인데 투수들의 사건 발생 확률을 곧바로 사용하기에는 해당 값 역시 만난 타자들의 실력에 의해 편향된 값이므로 적절하지 않다고 판단하였기에, 투수와 타자 모두 실제 실력을 의미하는 잠재 변수를 할당하였다.

이때 잠재 변수로부터 사건이 발생할 확률을 예측하는 함수 f는 다양한 형태로 생각해 볼 수 있을 텐데, 이번 연구에서 사용하기 적절하다고 판단한 함수는 Weighted Mean과 Inverse-Logit of Weighted Mean이다. 기본적으로 맞대결 시의 사건 발생 확률은 타자의 타율과 투수의 피안타율의 평균정도 될 것이라는 추측으로부터 다양한 형태의 평균 함수를 고려해보았다. 이때 투수와 타자 중 한 쪽의 잠재 변수가 더 큰 영향을 줄 수도 있을 것이라는 추측에 의해 weight를 사용하는 가중평균을 고려하였다. 첫 번째 후보로 Weighted (Arithmetic) Mean을 고려한 후, 다음 후보로 Weighted Geometric Mean을 고려하였는데, 이 경우
$f(\alpha_i,\beta_j)=\exp(w\times\alpha_i+(1-w)\times\beta_j)$의 형태가 된다. 이 식을 다시 쓰면
$\log(f(\alpha_i,\beta_j))=w\times\alpha_i+(1-w)\times\beta_j$의 형태가 되는데 f의 결과값이 확률임을 고려하면 좌변은 확률의 log값을 의미하게 된다. 이 값이 큰 의미를 갖지 못할 것이라고 판단하여 이보다는 확률값을 실수 전체범위로 확장시켜주는 함수인 logit함수를 이용하여
$logit(f(\alpha_i,\beta_j))=w\times\alpha_i+(1-w)\times\beta_j$의 꼴을 나타내는 Inverse-Logit of Weighted Mean를 두 번째 후보로 고려하였다.

모수를 추정하기 위하여 사용된 Loss 함수는 실제 관측된 데이터로부터 얻을 수 있는 로그가능도 함수의 음수를 의미하며 잠재변수들로부터 얻은 사건 발생 확률의 추정값들을 이용하여 계산하게 된다. 이 값의 경우 같은 관측 데이터들을 이용해 계산된 값이므로 모수를 추정함과 동시에 같은 모수 수일 경우 더 작은 Loss 값을 갖는 경우로 f를 선택하도록 모형을 결정하는데 사용가능하다. 해당 수식은 두 선수가 맞붙었을 때의 확률을 이용하는 것이 아닌 맞붙은 횟수와 사건이 발생한 횟수를 이용하는 것이기 때문에 많이 만나 데이터가 많은 경우 이 점이 Loss에 비중있게 반영되는 장점을 확인할 수 있다. 이 Loss 함수는 모수에 대해 복잡하고 비선형적인 형태 때문에 최대가능도 추정량을 closed form으로 구하기 어려워 경사하강법을 이용하여 Loss 값을 최소화하는 모수를 추정하고자 하였다.

위 방식으로 추정한 잠재변수들을 생각해보면 이 값이 선수의 능력을 추정하여 보여주고 있고 1차원이기 때문에 두 선수간의 비교도 가능하지만 이 값들의 범위를 모형 구성단계에서는 전혀 추측할 수 없기 때문에 어떤 값이 좋은 지 알 수 없어 사람들에게 익숙한 스케일의 값으로 변환해주는 것이 필요하다. 이를 위하여 타자의 경우 각투수를 만났을 때의 사건 발생 확률 추정값을 각 투수가 등장할 비율을 이용하여 가중 평균을 구함으로써 확률의 스케일로 변환해주었다. 마찬

가지로 투수의 경우는 각 타자를 만났을 때의 사건 발생 확률 추정값을 각 타자가 등장할 비율로 가중평균을 구하였다. 또한 이 값과 타자/투수 본인의 출전 횟수를 곱해 사건이 발생할 횟수를 추정하였다.

3. 실제 예제 적용

3.1 데이터셋

위 모델을 직접 적합해보기 위하여 사용한 데이터는 Bill Petti가 개발한 R package인 baseballr을 이용하여 최근의 메이저리그 기록을 다운로드하여 사용하였다. 위 package는 FanGraph.com, Baseball-Reference.com, baseballsavant.mlb.com 등의 웹사이트로부터 다양한 데이터를 크롤링하고 값을 이용하여 다양한 지표를 계산해주는 기능이 포함된 package로 이번 연구에 필요한 투수-타자 별 맞대결 결과 데이터 역시 제공하기에 해당 package를 사용하여 데이터를 준비하였다. 사용된 데이터는 2023년도 MLB 정규시즌 데이터의 전체 혹은 일부를 사용하였고 이번 예제에서 사용한 사건은 '안타'이며, 각 타자와 투수들의 타율/피안타율 및 각 타자와 투수간의 맞대결 시의 안타를 친 횟수와 치지 못한 횟수를 계산하여 준비하였다. 위 데이터를 준비하는 코드는 부록의 github repository 링크의 download_data.R 파일에서 확인할 수 있다.

3.2 적합

Loss 값을 최소화하는 모수를 추정하기 위한 경사하강법을 적용하기 위하여 python 모듈인 pytorch를 사용하였다. pytorch.nn.module 객체의 forward 함수에서 Loss를 구하는 연산을 할 경우 객체의 backward 함수가 자동으로 gradient를 구하고 미리 설정한 optimizer를 이용하여(이 예제 구현에서 optimizer의 hyper parameter로 Adam optimizer과 0.001의 learning rate가 이용됨) 적절하게 모수 추정값들을 업데이트해주는 등의 구조를 갖고 있어 경사하강법을 구현을 쉽게 할 수 있다. 적합에 관 한 코드는 부록의 github repository 링크의 fitting_weighted_mean.py, fitting_inv-logit_weighted_mean.py 파일에서 확인할 수 있다.

3.3 결과

3.3.1 f 간의 비교

위에서 제시한 두 가지 f의 후보지를 이용하여 23년 MLB 내셔널리그 팀 간의 경기를 모은 데이터, 23년 MLB 아메리칸리그 팀간의 경기를 모은 데이터를 이용해 모형을 적합하고 Loss를 계산해본 결과 각각의 데이터 셋에서 Weighted Mean의 경우 31605, 31374의 값을 가졌고, Inverse-Logit of Weighted Mean의 경우 31613, 31378의 값에서 수렴하는 모습을 보여주었다.

위 결과에 따르면 두 경우 모두 f가 Weighted Mean일 때 더 낮은 Loss를 갖는다는 점에서 이 경우에는 Weighted Mean 방식을 이용하는 것을 추천하고자 한다. 물론 이는 리그는 다르지만 둘 모두 비슷한 값을 갖고 있는 타율을 이용하여 추정한 값이라는 점에서 비슷한 경향성을 띄고 있다고 볼 수 있기에 이 결과를 절대적으로 신뢰하기는 실험의 양이 부족하다고 볼 수 있다. 하지만 우선 이번 연구에서는 위 결과를 토대로 타율에 대해서는 Weighted Mean 방식으로 적합한 결과를 이용하여 예제를 분석해보고자 한다.

3.3.2 Weighted Mean으로 구한 타율 적합 결과

MLB 내셔널 리그 팀간의 경기를 모은 데이터로 Weighted Mean을 이용하여 적합하였고, 그

결과 잠재 변수의 값이 가장 큰 선수들과 가장 작은 선수들의 잠재 변수 값, 이 값으로부터 계산된 각 선수들이 맞대결 했을 때의 타율/피안타율, 조정 타율/피안타율은 [표 1], [표 2]와 같다 (기준 150타수 이상). 이때 Weighted Mean에 사용되는 Weight 값은 0.4631이다.

이름	해당	α	β	조정 타율/피안타율
Devin Williams	투수 조정 피안타율 1위	0.0005		0.1293
Ronel Blanco	투수 조정 피안타율 최하위	0.4322		0.3610
Luis Arraez	타자 조정 타율 1위		0.5054	0.3614
Josh Donaldson	타자 조정 안타율 최하위		0.0586	0.1540

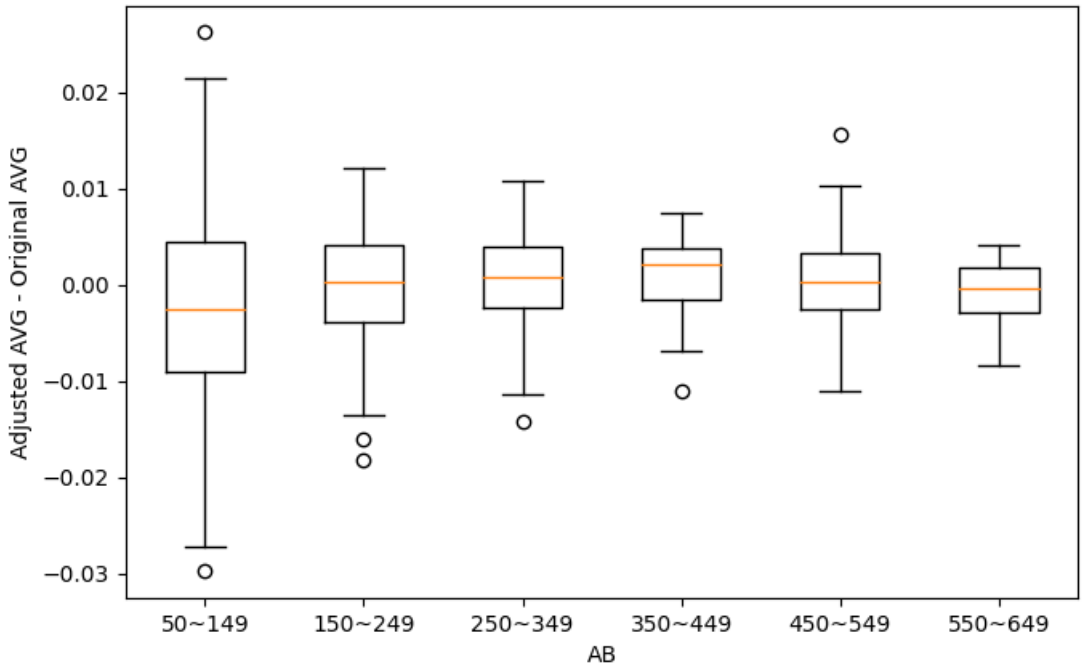
[표 1] 타자, 투수별 최고/최저 조정 타율/피안타율 기록 선수

투수\타자 맞대결 추정 타율	Luis Arraez	Josh Donaldson
Devin Williams	0.2716	0.0317
Ronel Blanco	0.4715	0.2316

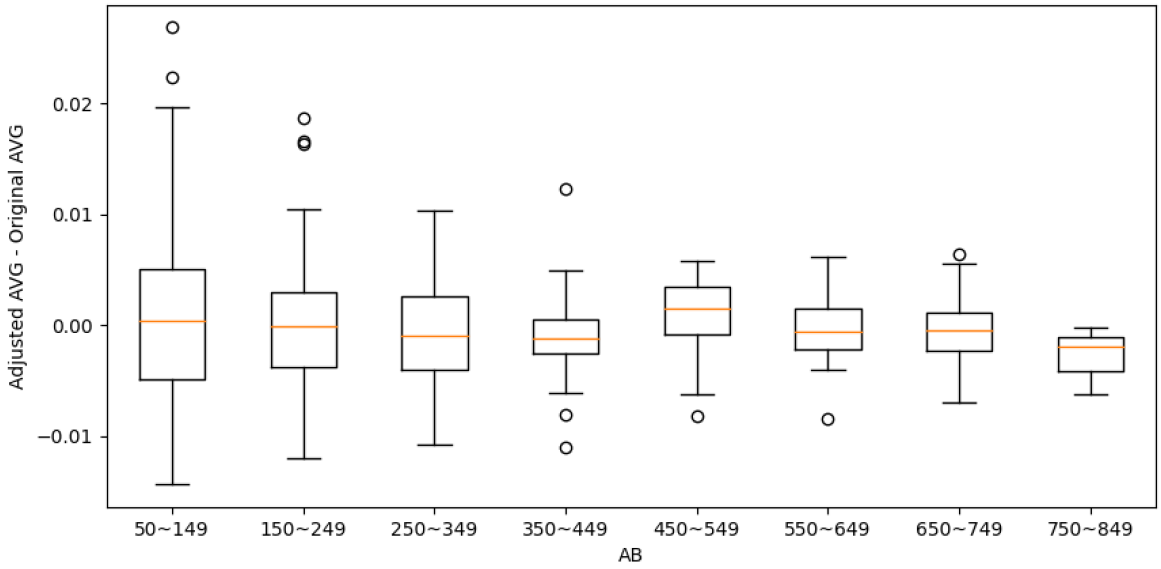
[표 2] 선수 간의 맞대결 타율/피안타율 추정값

위 적합 결과의 일부로부터 Weight는 0.5와 가까운 값을 보여주었고, 상식적으로 문제 없는 값을 보여주는 것을 확인할 수 있다.

이때 타수 수와 조정 타율간의 관계를 파악해보고자 50타수부터 100타수씩 구간을 만들어 각 구간에서 기존 타율에서 조정 타율을 뺀 값의 분포가 어떠한지 확인하였다. 그 결과는 타자의 경우 [그림 1], 투수의 경우 [그림 2]와 같다.



[그림 1] 타자의 타수 수에 대한 (기존 타율)-(조정 타율) 의 분포



[그림 2] 투수의 상대 타수 수에 대한 (기존 타율)-(조정 타율) 의 분포

해당 boxplot에서 볼 수 있는 점은 타자와 투수 모두 타수의 수가 적을 경우 더 넓게 분포되어 있고 타수가 많을 경우 더 좁게 분포되어 있는 경향성이 있기는 하나 300타수가 넘어가는 선수들의 경우에도 1푼이 넘는 차이를 갖는 경우를 확인할 수 있었다.

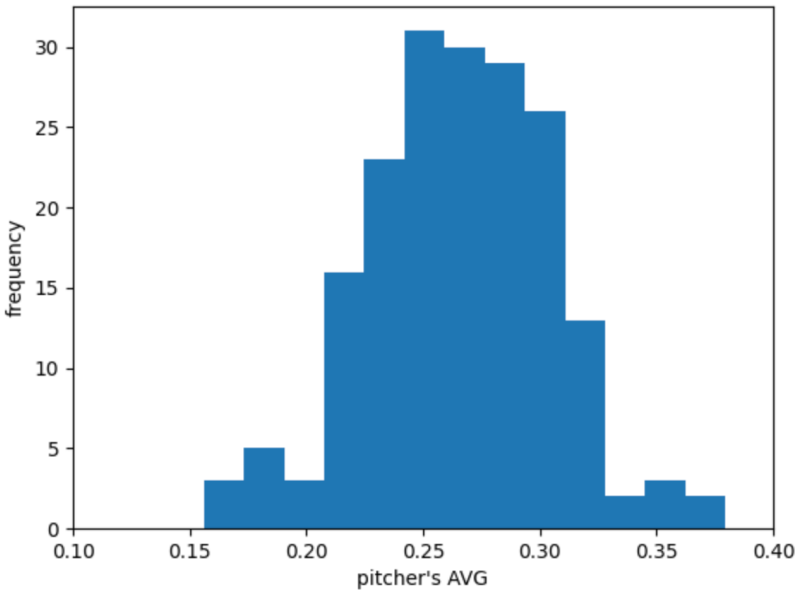
다음으로 위 적합 결과에서 가장 큰 차이를 보여주었던 선수들의 특징을 확인해보았다. 역시 타수 수가 너무 적은 선수들을 제외하고 150타수 이상의 기록을 가진 선수 중 차이가 절댓값이 가장 큰 양의 차이와 음의 차이를 보인 선수를 타자를 각각 선정하여 그들이 시즌 중에 맞대결한 선수들의 분포를 비교해보고자 하였다. 참고로 150타수 정도의 값을 가진 선수는 KBO 리그의 타자들의 타수 수 순위 120위권 정도 되며 이는 곧 팀의 3번째 백업 선수 정도의 출전 수를 보였다고 볼 수 있다.

결과를 얻으면 타자의 경우 'Brandon Drury'의 경우 -0.0157의 차이를 보여 조정타율이 기존 타율보다 낮게 나온 모습을 볼 수 있었고 이는 이 선수의 성적이 과대평가 되어있다고 볼 수 있다. 반대로 'Pavin Smith'의 경우 0.0181의 차이를 보여 해당 선수의 과소평가 되어있다고 평가할 수 있다. 두 선수가 올 시즌 데이터를 얻게 되는 과정에서 만난 투수들의 실제 피안타율을 이용하여 히스토그램은 [그림 3], [그림 4]와 같다.

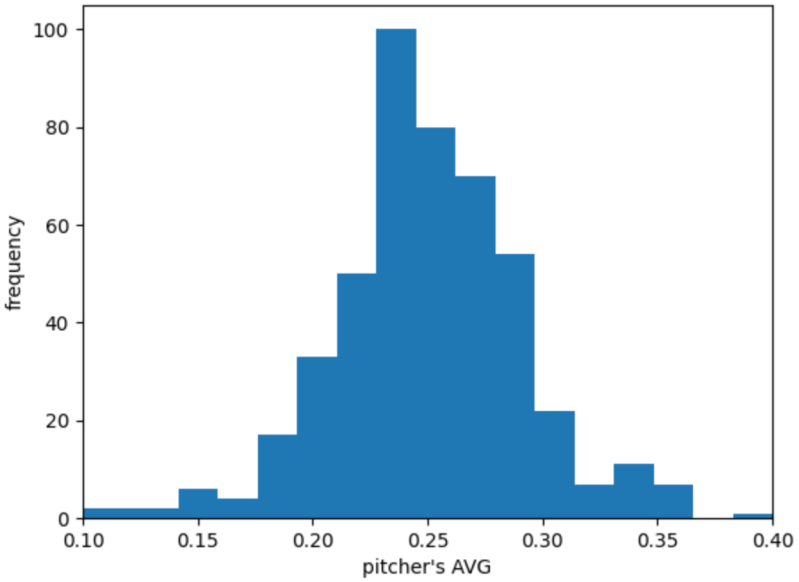
위 결과를 보면 'Brandon Drury'의 경우가 'Pavin Smith'에 비해 육안으로 관찰 될 정도로 피안타율이 높은 선수들을 만나고 있었음을 확인할 수 있다. 이는 곧 통계 모형에서 의도하고 있는 보정을 지표에서 잘 하고 있음을 확인할 수 있다.

3.3에 해당하는 모든 분석의 코드는 부록의 github repository 링크의 inference.ipynb에서 확인할 수 있다.

4. 결론 및 제언



[그림 3] Brandom Drury 선수가 만난 투수의 피안타율 분포



[그림 4] Pavin Smith 선수가 만난 투수의 피안타율 분포

이번 연구를 통해 상대를 고려한 지표를 제안하였고, 이를 위한 통계 모형을 구축하고 예제 데이터를 이용하여 적합을 통해 해당 모형이 설계했던 의도를 잘 반영하고 있으며 적지 않은 차이의 보정값을 보여주고 있음을 확인할 수 있었다. 특히 이 모형은 기존에 사용되고 있던 확률 형태의 지표(타율, 출루율, 도루저지율 등) 및 횟수 형태의 지표(1루타, 2루타, 3루타, 홈런, 볼넷, 도루성공 등)의 형태로 변환한다는 점에서 기존의 위 값들을 이용해서 계산되고 있던 가공된 지표들(FIP, wOBA, WRC+, WAR)에도 반영할 수 있다는 점에 강점이 있다는 점에서 기존의 분석 방식을 그대로 사용하며 보정을 할 수 있다는 점이 의미가 있다고 생각한다.

이번 연구에서의 예제에서는 타수의 수가 수백 정도의 값을 갖고, 사건 발생 확률이 대부분의 선수가 2할~3할 근처에서 형성되는 비교적 이상적인 경우에서의 적합을 하였는데, 빈도가 0.5에 가까운 빈도를 보여주는 사건들이나 빈도가 훨씬 적어 0.1 미만으로 발생하는 사건들을 처리하는 모형에서도 잘 적합되는지 확인하고 다양한 환경에서의 적절한 f를 추정할 수 있다면 더 강건한 모형이 될 것이라고 생각하며, 이번 연구는 상대를 고려한 지표의 개념과 Baseline 역할을 해줄

수 있는 모형을 제시한 데 의의가 있다고 생각한다.
부가적으로 야구 경기뿐만 아니라 구조가 비대칭적인 다른 모든 스포츠 혹은 게임에서의 평가를 위한 지표로써 사용될 여지가 있다고 생각되어 다양한 분야의 데이터에도 유사한 방식으로 확장하는 것 역시 고려할 가치가 있다고 생각한다.

[부록]

연구에 사용된 코드 Github 페이지 -
https://github.com/JinooJung/adversary-conidered-sabermetrics-stats