Journal of the Korean Data & Information Science Society
2015, **26**(1), 197–207

http://dx.doi.org/10.7465/jkdi.2015.26.1.197
한국데이터정보과학회지

한국프로야구 선수들의 타율에 기반된 타격 능력의 베이지안 추정[†]

조용주[1] · 이광호[2]

[12]영남대학교 통계학과

접수 2014년 12월 18일, 수정 2015년 1월 11일, 게재확정 2015년 1월 20일

요 약

야구경기에서 타자의 능력을 평가하는 중요 측도에는 타율, 타점, 홈런, 득점, 출루율 등이 있지만 최근에는 타자의 능력을 OPS, ISO, SECA, TA, RC, XR 등과 같은 포괄적인 지표를 사용하여 나타내는 경우가 많다. 이러한 지표들은 흔히 특정 기간 동안 얻은 데이터를 이용하여 계산된 것이기 때문에 기간에 따라 대체로 큰 편차를 보이는 경우가 많으며 특정 모수 (parameter)를 추정하는 것도 아니다. 본 연구에서는 한국프로야구 선수들의 순수한 타격 능력 (hitting ability)을 모수로 간주하여 통계적 방법으로 추정하고자한다. 타격 능력에 추정에는 타수 (at bat)가 반영된 베이지안 방법이 사용될 것이다.

주요용어: 베이지안 방법, 베타-이항 분포, 사전분포, 사후분포, 타격 능력, 타수.

1. 서 론

스포츠는 흔히 유희적인 신체활동의 의미 보다는 경쟁화된 놀이, 승리를 위한 연습의 필요에 의해 만들어 졌다는 의미에서 좁은 의미로 투쟁과 경쟁으로 정의되기도 한다 (Lee, 2000). 스포츠 중에서 특히 야구는 한국에서 남녀노소를 불문하고 가장 인기 있는 종목으로 2013년을 기준으로 한국프로야구 선수위원회의 1군에 등록되어 선수가 465여 명에 이르고 연간 관중 수도 645만여 명 (Sports2*i*, 2014)에 이르고 있다. 이는 한국프로축구의 관중수 230만여 명 (Korea Professional Football League, 2014)과 비교 할 때 그 인기도를 실감할 수 있다. 만약 아마추어 야구와 사회인 야구까지 고려한다면 야구에 직접 참여하여 경쟁하거나 즐기는 사람은 훨씬 더 많을 것으로 생각된다. 이와 같이 많은 사람들의 사랑을 받고 있는 야구는 경기 내용이 세세하게 기록되고 있기 때문에 통계분석의 대상으로 매우 훌륭한 조건을 갖추고 있다. 또한 경기의 흐름이나 결과가 우연적인 사건에 의해 바뀌거나 결정되는 경우가 많기 때문에 통계학자들이 더 많은 관심을 가지는 것 같다. 미국프로야구에 대한 통계적 분석은 James (1982)의 연구이래로 본격화 되었다. 한편 한국프로야구에 대한 통계적 연구는 Cho 등 (2003, 2004, 2005)이 BC (beane count)와 승률간의 관계를 조사하고 타자의 홈런과 4구, 투수의 피 홈런과 피 4구 등 4개의 변수를 군집 분석하여 어떠한 성격의 군집이 승률에 가장 높은 영향을 미치는지를 알아보고 승률에 대한 회귀식을 유도하였으며 WHIP (walks plus hits divided by innings pitched)이 방어율에 미치는 영향에 관한 연구를 하였다. 그 외, Lee와 Kim (2005, 2006a, 2006b), Lee와 Cho (2009), Choi와 Kim (2011), Kim (2012) 그리고 Lee (2014) 등이 활발하게 연구하고 있다.

[†] 이 논문은 2012학년도 영남대학교 학술연구 조성비에 의한 것임.
[1] (712-749) 경상북도 경산시 대학로 280, 영남대학교 통계학과, 박사수료.
[2] 교신저자: (712-749) 경상북도 경산시 대학로 280, 영남대학교 통계학과, 교수. E-mail: khlee@ynu.ac.kr

한편 타자의 능력을 측정하는 지표들에는 OPS (on base percentage plus slugging percentage), ISO (isolated power), SECA (secondary average), TA (total average)는 타자의 본인의 출루율 기반을 둔 지표라고 할 수 있는 반면 RC (runs created; James, 1985)와 XR (extrapolated runs; Furtado,1999)은 타자 본인의 출루율과 공격에서 선행 주자를 진루 시키는 능력을 고려하여 만든 지표이다. 그런데 위에서 거론된 타자의 능력에 대한 지표들은 대부분 특정 기간 (한 시즌, 전반기 시즌 혹은 최근 10게임 등) 동안 얻은 데이터로부터 측정되는 것이며 현실적으로 많이 활용되고 있긴 하지만 선수들의 컨디션이나 상대팀 혹은 상대팀의 투수에 의해 많은 영향을 받는 지표들이므로 진정한 의미에서 그 타자의 능력을 나타내기에는 다소 무리가 있다고 생각된다. 또한 동일한 지표라도 측정에 사용된 데이터를 얻은 기간이 다른 경우의 지표 값으로는 능력을 비교해서는 안 되지만 현실적으로 많이 사용되는 것 같고 그로 인하여 선수를 잘 못 인식하는 경우가 많은 것 같다. 물론 전문적인 지식을 갖춘 사람이나 전문 기관에서의 평가에는 이러한 경우가 없겠지만 많은 관중들에게는 흔할 것으로 생각된다. 예를 들어 선수 A의 OPS가 8할이고 선수 B의 OPS는 5할이라 할 때 단순히 8할과 5할의 OPS로 두 선수의 능력을 비교하거나 평가해서는 안 된다는 것이다. 최소한 두 선수의 활동기간이 비슷하거나 혹은 비슷한 타격기회가 주어졌다는 가정 하에 비교하는 것이 타당하리라 생각한다.

본 연구에서는 타자들에게 주어진 타격의 기회를 고려하여 한국프로야구 선수들의 순수한 타격 능력을 베이지안 기법으로 추정하고자 한다. 제2절에서는 2013년도 한국프로야구 선수들의 기록을 바탕으로 한국프로야구로 등록된 모든 타자들의 타수 분포를 제안하고 제3절에서는 타율에 기반된 타격 능력 추정을 위한 모형의 제안하고 타율에 따른 타격 능력의 사후 분포를 유도한다. 그리고 제4절에서 타수 분포와 따른 타격 능력을 추정하고 제5절에서는 결론 및 추후 연구에 대한 제안을 하려고 한다.

본 연구에 사용된 한국프로야구와 관련된 데이터는 2014년 한국프로야구 연감 (Sports2i, 2014)에 기록된 것이며 여러 가지 통계 계산은 마이크로소프트의 엑셀2010을 사용하였고 모의데이터의 생성은 SAS IML을 이용하였다.

2. 타수 분포

야구경기에서 타수 (at bat)나 타석수 (plate appearance)는 직접 기록되지 않는다. 단지 타자가 타석에 들어선 후부터 타석에 내려올 때까지의 여러 가지 상황에 따른 결과만 기록될 뿐이다. 타수와 타석수는 비슷하게 사용되는 용어지만 엄밀히 말하면 상당히 차이가 있다. 타석수는 타자가 타격을 위하여 타석에 들어선 횟수를 말하며 타석에 들어선 순간부터 타격 후의 결과에 이르는 모든 사건의 결과로 기록된다. 즉, 타석수에는 안타와 삼진, 범타, 4구, 사구, 희생번트, 희생플라이, 타격방해 등으로 기록되는 모든 사건의 합을 말한다. 여기서 안타란 타자가 순수한 타격의 결과로 1루 이상 진루에 성공한 타격을 말하며 범타는 안타가 아닌 순수한 타격의 결과를 말한다. 타수는 타석수에서 팀의 승리를 위한 타격 행위나 상대팀의 실책 혹은 작전으로 올바른 타격이 이루어지지 않은 결과를 제외한 타격 행위의 수를 말한다. 즉 타수는 타석수에서 희생타 (번트, 희생플라이)와 상대팀의 실책 혹은 작전 (4구, 사구, 야수의 실수)를 제외한 수치를 말한다.

야구경기에서 타자들의 여러 가지 능력 중에서 타격 능력은 흔히 타율로 추정된다. 여기에서 타율은 타자에게 주어진 총타수에 대한 안타수의 비율을 말한다. 제1장에서 언급한 바와 같이 이 타율은 많은 표본 통계량이 그러하듯 주어진 표본의 수인 총타수에 따라 상당히 큰 편차를 보인다. 예를 들어, 동일한 타자에 대한 일주일 동안의 타율이 비슷해야 마땅함에도 실제로는 타자의 컨디션과 상대팀의 투수 등에 의해 많은 편차를 보인다. 물론 타율 계산 기간을 전반기 혹은 한 시즌과 같이 길게 잡으면 단기간 보다는 편차가 적긴 하지만 그래도 상당한 편차를 보인다. Figure 2.1과 Figure 2.2는 2013년도 한국프로야구 선수들 중에서 상위권에 속하는 몇몇 선수들의 월별 타율 (Sports2i, 2014)의 변화를 보여주고 있

는데 기간별로 타율의 편차가 매우 크다는 것을 알 수 있다. 상당히 상위권에 속하는 선수들의 타율 편차가 이와 같이 큰 것을 고려하면 중위권 혹은 하위권에 속하는 타자들의 타율 변화는 더욱 심할 것으로 생각된다. 사실 상위권에 속하는 선수들은 능력 있는 선수로 인정받기 때문에 타격기회도 상당히 많겠지만 하위권에 속하는 선수들에게 주어지는 타격 기회를 고려하면 그 선수들의 타율 변화는 매우 큰 변이를 가질 것으로 예상되므로 타율로 타자들의 타격 능력을 추정하는 것은 현실적으로 많이 사용되는 측도이기는 하지만 통계적 관점에서 보면 신뢰성이 상당히 떨어진다고 볼 수 있다. 타수에게 주어진 타격기회 혹은 타수를 고려하여 타자들의 순수한 타격 능력을 추정하기 위하여 먼저 2013년 한국프로야구 선수로 등록되어 있는 타자들에 대한 타수 분포를 추정하고자 한다. 2013년에 한국프로야구에 등록된 선수로 정상적으로 리그에 참여한 타자들은 총 247명이었으며, 그 중 가장 많은 타석에 들어선 선수 573회였고 가장 적은 타석에 들어선 선수는 0회였으며 평균적으로 181.68회의 타석 (Sports2*i*, 2014)에 들어선 것으로 파악되었다.

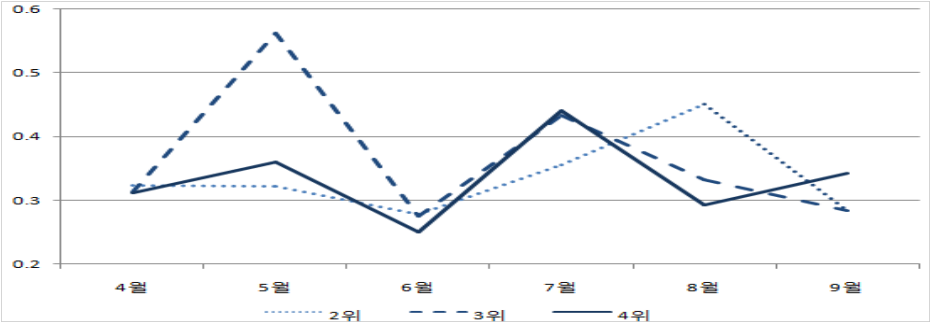

Figure 2.1 The variations of batting averages for some player per month (ranking 3, 4, 5)

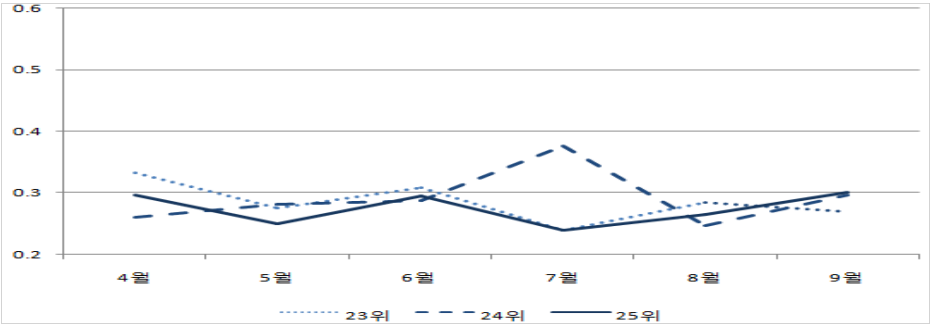

Figure 2.2 The variations of batting averages for some player per month (ranking 23, 24, 25)

본 연구에서는 타석 수 분포 보다는 타수 분포가 순수한 타격 능력을 더 잘 반영할 것으로 생각되어 리그에 정상적으로 참여한 타자들의 타수를 조사하여 Figure 2.3과 같이 히스토그램을 그렸다. 히스토그램은 타수구간의 설정에 따라 다소 상이한 형태를 보였지만 Figure 2.3은 구간의 폭이 30이 되게 그린 것이다. 2012년 시즌에 대한 히스토그램도 Figure 2.3과 거의 유사한 형태를 보였다. 본 연구에서는 Figure 2.3의 히스토그램을 이용하여 두 가지 방법으로 타수의 분포의 밀도함수를 추정하고 추정된 이 함수를 이용하여 베이지안 방법으로 타자들의 타격 능력을 추정하고자 한다. 또한 추정된 두 분포가 타격 능력 추정의 결과에 얼마나 큰 영향을 주는지도 알아보게 될 것이다.

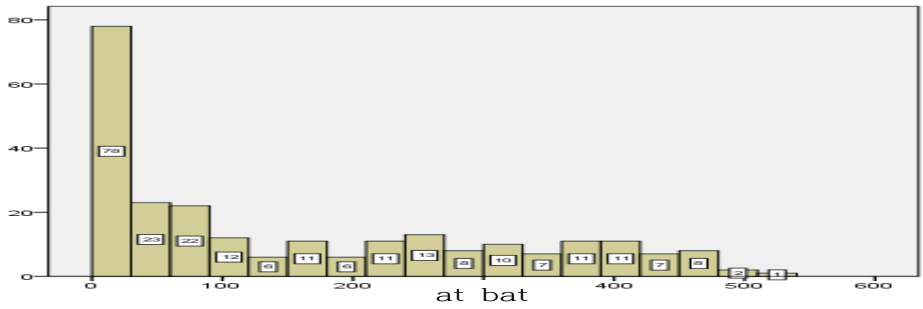

Figure 2.3 Histogram of at bat in season 2013

Figure 2.3으로부터 한국프로야구 선수들의 2013년 시즌 동안의 타수의 상대도수는 타수가 많아짐에 따라 급격히 감소하는 구간과 거의 상수에 가까운 구간 그리고 다시 급격히 감소하는 구간으로 나누어짐을 알 수 있다. 식 (2.1)은 2013년 시즌 타수의 평균 (157.02)과 규정 타석 (396.8)을 기준으로 3개의 구간으로 나누어서 Figure 2.3의 히스토그램이 보여주는 분포를 잘 반영할 수 있게 1차식 혹은 2차식으로 유도한 것이다. 이 방법은 Frey (2007)가 미국프로야구 선수들의 타수 분포에 대한 밀도함수 (probability density function; p.d.f.)를 얻기 위해 사용하였다.

$$f(x) = \begin{cases} \dfrac{1}{760}\left\{\dfrac{67}{247500}(x-150)^2 + 1\right\}, & 0 \le x \le 150, \\ \dfrac{1}{760}, & 150 \le x \le 390, \\ \dfrac{1}{760}\left(-\dfrac{1}{132}x + \dfrac{87}{22}\right), & 390 \le x \le 510. \end{cases} \tag{2.1}$$

다른 하나의 타수 분포는 회귀모형을 이용하여 추정하였다. Figure 2.3의 히스토그램을 회귀모형으로 근사화하기 위하여 히스토그램 계급의 왼쪽값과 그 계급의 상대도수를 회귀식으로 적합하여 밀도함수의 조건을 충족하도록 상숫값을 계산하여 얻은 것이 식 (2.3)이다. 회귀식의 적합에서 결정계수는 0.907이었다.

$$g(x) = \frac{1}{480814 + 69063 \times \ln 510} \times \left(9.445 + \frac{69.063}{x}\right), \ 1 \le x \le 510. \tag{2.2}$$

식 (2.1)과 식 (2.2)의 밀도함수에 대한 그래프는 Figure 2.4와 같다.

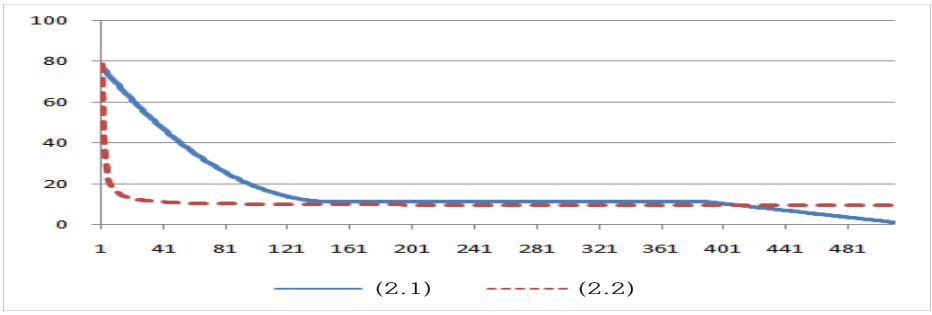

Figure 2.4 Graphs of p.d.f. (2.1) and (2.2)

3. 모형과 사후분포

타자가 한 번의 기회에서 얻을 수 있는 결과는 안타, 4구, 사구, 상대팀 수비수의 실책으로 인한 진루 등을 비롯하여 여러 가지가 있을 수 있다. 본 연구에서는 타격 능력을 추정하는 것이 목적이므로 타율 (batting average) 계산에서 배제되는 결과들 즉, 타수에서 얻을 수 있는 것은 안타를 치든가, 그렇지 않든가 두 가지 이다. 따라서 주어진 타수 $(A = a)$에서의 안타의 개수 (H)는 이항분포를 따른다고 볼 수 있다. 즉 타격 능력 $P = p$와 타수 $A = a$가 주어진다면 안타의 수는 $H \sim B(a, p)$라 할 수 있다. 이론적으로는 a번 주어지는 타수가 서로 독립적이냐 그리고 타격 능력이 상대팀의 투수에 따라 차이가 날 가능성이 충분한데도 상수로 간주 할 수 있느냐에 따라 $H \sim B(a, p)$이라는 모형이 정당하냐는 의견이 분분할 수 있지만, 사전 연구 (Frey, 2007)에서 이론적으로는 잘 설명되어 있으므로 본 연구에서는 모형으로부터 생성된 모의 데이터에 대한 타율과 타수의 산점도 (Figure 3.3)와 한국프로야구 2013년 시즌 데이터에 대한 그것 (Figure 3.1)를 비교함으로써 그 정당성을 이 절의 말미에 제시하였다. 타수 a가 주어졌을 때 타격 능력 (p)의 조건부 분포를 구하는 문제를 고려하여 보면 타격 능력도 확률이므로 p의 사전분포 (prior distribution)로 흔히 사용되는 베타분포 (beta distribution)로 가정하기로 한다. 모수 α와 β를 가지는 베타 분포에서 필요한 것은 평균과 분산을 유추하는 것인데 $B(\alpha, \beta)$ 분포의 평균과 분산은 각각 $\alpha/(\alpha+\beta)$와 $\alpha\beta/(\alpha+\beta)^2(\alpha+\beta+1)$이므로 타수 a가 주어진 경우 α와 β를 직접 추정하여 모형화하기 보다는 두 개의 매개변수 μ와 c를 추정하고자 한다. 여기서 $\mu = \alpha/(\alpha+\beta)$, $c = \alpha+\beta$이고 $B(\alpha, \beta)$분포의 분산은 $1/(4c+4)$보다 크지 않으므로 c는 분산의 크기를 제한하는 상수가 된다. 먼저 주어진 타수 a에 대응하는 μ를 추정하기 위하여 2013년 시즌의 타수와 타율의 산점도 (Figure 3.1)를 보자. 이 그림은 타수가 증가하면 타율도 어느 정도 증가하는 추세를 보여준다. 특히 약 10타수 미만인 경우의 타율은 이상치 (outlier)로 간주 할 만큼 큰 변동을 보인다.

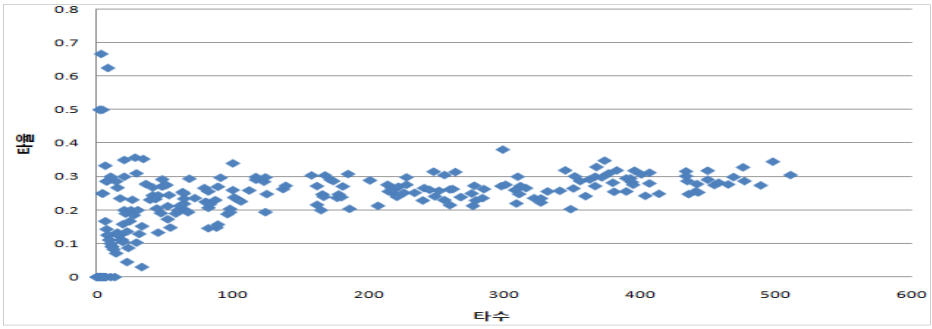

Figure 3.1 Scatter plot of batting average and at bat for season 2013 data

타율의 평균을 타격 능력 p의 추정값으로 이용할 수 있으므로 타수를 예측변수로 하고 타율을 반응변수로 하는 회귀식으로 적합해 보았다. 타율의 증가율이 다소 구간에 따라 다소 차이가 있으므로 2013년 시즌의 팀당 경기 횟수 128보다 큰 경우와 작은 경우로 구분하는 가변수 (dummy variable)를 넣어서 적합한 회귀식은 식 (3.1)과 같다. 여기서 가변수 z는 타수가 128보다 크면 0이고 128보다 크지 않으면 1의 값을 가진다.

$$\mu(a) = 0.262 + 0.0000439a + 0.001(a - 128)z. \tag{3.1}$$

한편, p에 대한 사전 분포가 베타분포로 주어졌을 때 h에 대한 예측분포는 베타-이항 분포를 따르기

때문에 2013년 시즌 타자들의 타수에 따른 분산을 계산하여 c를 추정할 수 있다.

$$Var\left(\frac{h}{a}\right) = \frac{1}{a}\mu(a) + \left(1 - \frac{1}{a}\right)\left(\frac{[c\mu(a)]^2 + \mu(a)}{c+1}\right) - [\mu(a)]^2. \tag{3.2}$$

그런데 동일한 타수를 가지는 타자들이 많으면 주어지는 타수 하나 하나에 대해 표본 분산을 계산 할 수 있지만 현실적으로는 그렇지 못하기 때문에 타수들의 값을 $[0, 40)$, $[40, 80)$, $[80, 120)$, $\cdots$, $[480, 510]$과 같은 구간으로 나누어서 각 구간에 속하는 타수에 대응되는 타율들을 표본으로 보고 표본 분산을 계산하였다. 각 구간의 중점을 수평축으로 하고 표본 표준편차 (standard deviation)를 수직축으로 하는 산점도가 Figure 3.2이다.

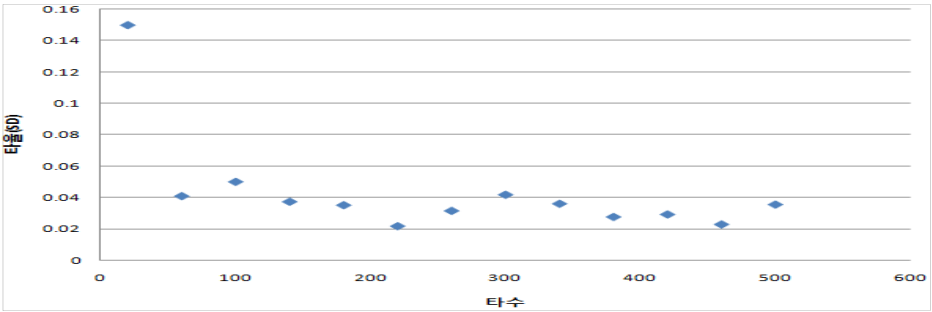

Figure 3.2 Scatter plot of standard deviation and at bat

주어진 모든 타수 (구간의 중앙값)에 계산된 표본 분산과 식 (3.2)의 분산 차의 제곱합이 최소가 되는 c값은 근사적으로 250정도가 되었다. 결국 이와 같이 얻어진 $\mu(a)$와 c에 의해 베타분포의 두 모수 α와 β의 값이 결정되므로 타수 $A = a$가 주어졌을 때 타격 능력 P의 사전분포는 Beta $(c\mu(a), c(1 - \mu(a)))$이다. 이 모형이 2013년 시즌의 데이터와 얼마만큼 적합이 잘 되느냐는 모의데이터 (simulated data)를 생성하여 타수에 따른 타율의 산점도를 그려서 확인할 수 있다. 2013년 시즌의 데이터에 기록되어 있는 모든 타수와 거기에 대응되는 안타수의 쌍이 247개 이므로 $H|_{A=a,P=p} \sim B(a,p)$이고 $P \sim B(c\mu(a),\ c(1-\mu(a)))$라는 모형으로부터 서로 독립된 247개의 (a,p,h)를 생성하여 타수 a와 타율 h/a에 대한 산점도를 그린 것이 Figure 3.3이다. 이 산점도를 2013년 시즌의 데이터에 대한 산점도 Figure 3.1와 비교해 보면 제안된 이론적인 모형이 2013년 시즌 데이터와 잘 부합되고 있음을 보여준다고 할 수 있다.

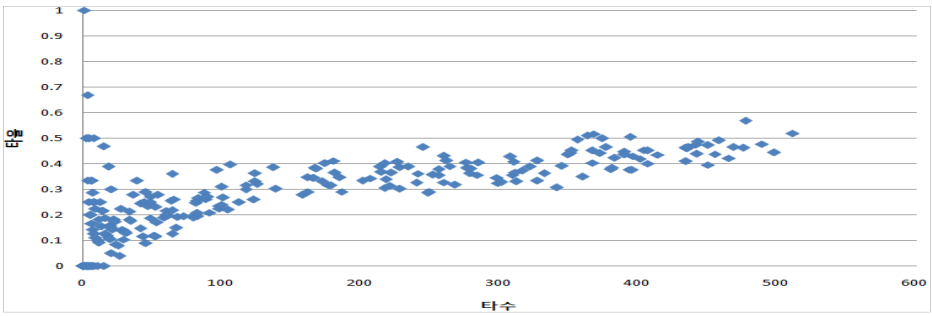

Figure 3.3 Scatter plot of batting average and at bat for simulated data

이제 타율에 따른 타격 능력의 사후분포를 구해보자. 타수가 $A = a$일 때의 타격 능력 P을 추정하기 위해서는 P의 사후분포 (posterior distribution)를 구해야 하는데 베타분포와 이항분포가 공액분포 (conjugate distribution; Berger, 1985)라는 사실로부터 P의 사후분포는 쉽게 구할 수 있다. 어떤 타자가 a번의 타석에서 h개의 안타를 쳤다고 하면, a번의 타석에 들어선 타자의 타격 능력 P에 대한 사전분포는 Beta $(c\mu(a),\ c(1-\mu(a)))$로 가정했으므로 이 타자가 a번의 타석에서 h개의 안타를 쳤으므로 P의 사후분포는 공액성에 의해 Beta $(c\mu(a)+h,\ c(1-\mu(a)+a-h))$가 된다. 이 사후분포는 사실 타수가 $A = a$, 안타의 수가 $H = h$로 주어졌을 때 타격 능력 P의 조건부분포이다.

본 연구에서는 타수와 안타 수에 대한 모든 쌍 (a, h)의 값이 주어질 때 타격 능력 P를 추정하기 보다는 타율 $B \equiv H/A$가 $b \equiv h/a$가 주어진 경우에 P를 계산하는 것이 목적이다. 이는 현실적이기는 하지만 계산적으로는 다소 복잡하게 된다. 주어진 b의 값에 대응하는 쌍 (a, h)이 유한하기는 하지만 매우 많기 때문이다. 지금 어떤 주어진 타율 b에 대한 $h/a = b$가 되는 (a, h)들의 쌍을 (a_i, h_i) $(i = 1, 2, \cdots, K)$라 하자. 여기서 $1 \le a_i \le 510$이고 K는 b의 값에 대응하는 쌍 (a, h)가 유한한 값이며 510은 2013년 시즌 한국프로야구 선수 중에서 최대 타수를 고려한 값이다. 주어진 하나의 쌍 (a_i, h_i)에 대한 타격 능력 P의 사후분포가 다음과 같은 베타분포 Beta $(c\mu(a)+h,\ c(1-\mu(a)+a-h))$가 됨을 알고 있기 때문에 타율이 $B = b$로 주어진 경우의 타격 능력 P의 분포는 식 (3.3)과 같다.

$$\sum_{i=1}^{K}\left(\frac{P(h_i \text{ hits and } a_i \text{ at bats})}{\sum_{j=1}^{K} P(h_j \text{ hits and } a_j \text{ at bats})} \times Beta(c\mu(a_i)+h_i,\ c(1-\mu(a_i))+a_i-h_i)\right). \tag{3.3}$$

식 (3.3)에서 $P(h_i \text{ hits and } a_j \text{ at bats})$의 계산은 아래에서 유도된다. 편의상 계산식에서는 h_i를 h로 a_i는 a로 표기하기로 한다.

$P(H = h,\ A = a)$는 $P(A = a)P(H = h|A = a)$이고 $p(A = a)$는 식 (2.1) 또는 식 (2.2)의 타수 분포 밀도함수를 이용하여 구할 수 있는데 그 값을 $d(a)$로 쓰기로 하고 (3.4)식과 같이 정의한다.

$$d(a) \equiv \int_{a-1}^{a} f(x)dx \tag{3.4}$$

한편 $P(H = h|A = a)$는 타수가 $A = a$로 주어진 경우의 안타 수 H의 분포는 이항 분포가 된다는 사실로부터 다음과 같이 계산된다.

$$\begin{aligned}
&P(H = h,\ A = a) = P(A = a)P(H = h|A = a)\\
&= d(a)P(H = h|h \sim Binomial(a,\ p) \text{ and } \ p \sim Beta(c\mu(a),\ c(1-\mu(a)))\\
&= d(a)\int_0^1 \left[\binom{a}{h} p^h (1-p)^{a-h}\right] \times \frac{\Gamma(c)}{\Gamma(c\mu(a))\Gamma(c(1-\mu(a)))} p^{c\mu(a)-1}(1-p)^{c(1-\mu(a))-1}dp\\
&= d(a)\left(\frac{\Gamma(a+1)\Gamma(c)\Gamma(h+c\mu(a))\Gamma(a-h+c(1-\mu(a)))}{\Gamma(h+1)\Gamma(a+1-h)\Gamma(c\mu(a))\Gamma(c(1-\mu(a)))\Gamma(a+c)}\right).
\end{aligned} \tag{3.5}$$

여기서 $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-}xdx$이다. 따라서 식 (3.3)과 식 (3.5)로부터 하나의 쌍 (a_i, h_i)에 대해 $P(h_i \text{ hits and } a_i \text{ at bats})$가 계산되므로 이들의 K개를 합하면 식 (3.3)의 분포가 계산되며 이것을 이용하여 타율이 $B = b$인 타자의 타격 능력 P를 추정하거나 예측구간 등을 계산 할 수 있다.

4. 타수 분포와 타율에 따른 타격 능력의 추정

먼저 제2절에서 제안된 2가지의 타수 분포가 타수에 따른 타격 능력에 사후평균에 얼마나 큰 영향을 미치는지를 알아보기 위하여 식 (2.1)과 식 (2.2) 그리고 식 (3.3)을 이용하여 타격 능력의 사후평균과

사후표준편차를 각각 계산하여 그림을 그렸다. 계산에서 주어지는 타율 b $(0 \leq b \leq 1)$는 연속적임에도 불구하고 1/1000이하는 반올림된 것으로 간주하였으며 또한 $\sum_{j=1}^{K} P(h_j \text{ hits and } a_j \text{ at bats})$는 1억분의 1보다 작은 값은 배제하였다. 사실 $\sum_{j=1}^{K} P(h_j \text{ hits and } a_j \text{ at bats})$는 $P(B = b)$를 가능한 모든 타수와 안타수의 쌍으로 조개어서 계산하는 것이다. Figure 4.1은 타수 분포가 식 (2.1)인 경우의 타율에 따른 타격 능력의 사후 평균과 표준편차 이고 Figure 4.2는 타수 분포가 식 (2.2)인 경우의 타율에 따른 타격 능력의 사후 평균과 표준편차를 계산하여 그린 것이다. 제시된 그림은 타수 분포에 평균과 표준편차가 비슷한 패턴을 보이고는 있지만 회귀분석 방법으로 제안된 식 (2.2)의 타수 분포에 대한 사후 표준편차가 상대적으로 그 값이 작은 쪽으로 치우친 경향을 보여 주는 것으로 보아 변이도 측면에서 다소나마 더 좋은 결과를 보여 주는 분포라 할 수 있겠다. 특히 Figure 4.1은 타격 능력에 대한 매우 높은 사후 평균을 가지는 타율은 주로 0.3과 0.45구간에 있는 경우였으며 가장 큰 사후 평균은 Table 4.1로부터 0.362임을 알 수 있다. Figure 4.2에서는 0.35와 0.45의 구간의 사후 평균이 매우 컸고 가장 큰 사후 평균은 Table 4.2로부터 0.385임을 알 수 있다. 2013년 시즌의 한국프로야구 데이터로부터 얻은 이러한 결과는 타자들이 실제로 시합에 임할 때의 마음의 자세를 엿보게 한다. 즉, 실제의 시합에서는 순수한 타격 능력 보다 더 높은 타격 능력을 보여 준다고 할 수 있겠다.

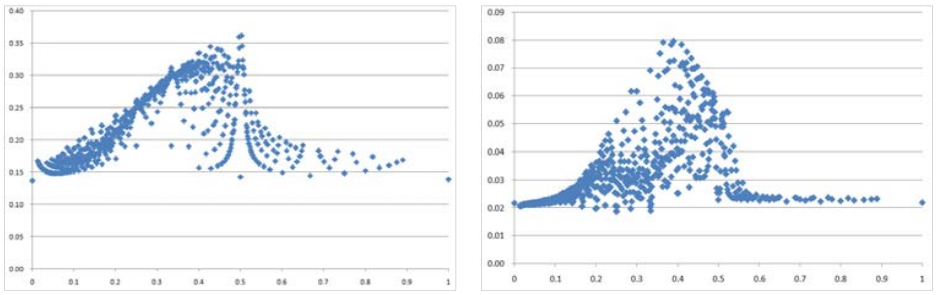

Figure 4.1 Posterior means and standard deviations by using p.d.f. (2.1), respectively

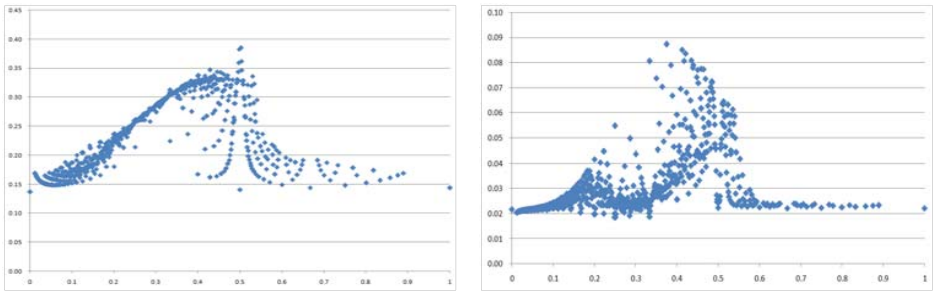

Figure 4.2 Posterior means and standard deviations by using p.d.f. (2.2), respectively

Table 4.1와 Table 4.2는 Figure 4.1과 Figure 4.2의 특성을 보다 명확하게 나타내기 위해 2013년 시즌 한국프로야구 선수의 중에서 타격 능력의 사후평균 가장 높은 상위 10개와 가장 낮은 하위 10개의 값에 대응되는 타율을 나타내었다. 타격 능력의 추정값이 높으면 실제 시합에서 타율도 거의 유사하게 높다는 것을 명확하게 알 수 있다. 반면에 타격 능력 극히 떨어지는 경우에는 타율이 극히 낮은 경우와 극히 높은 경우가 상존하고 있음을 볼 수 있는데 이것은 타격 능력이 극히 떨어지는 선수에게 주어지는 타격기회가 매우 적기 때문으로 생각된다.

Table 4.1 Posterior means and standard deviations (s.d.) for some players by using p.d.f. (2.1), respectively

b	mean	s.d.	b	mean	s.d.
0.503	0.362	0.025	0.000	0.137	0.022
0.497	0.359	0.025	1.000	0.139	0.022
0.504	0.346	0.027	0.500	0.143	0.023
0.428	0.345	0.027	0.667	0.144	0.022
0.496	0.343	0.027	0.750	0.148	0.022
0.445	0.341	0.031	0.059	0.148	0.022
0.454	0.338	0.033	0.056	0.148	0.022
0.401	0.335	0.025	0.067	0.148	0.022
0.399	0.335	0.025	0.053	0.148	0.022
0.461	0.331	0.039	0.063	0.148	0.022

Table 4.2 Posterior means and standard deviations (s.d.) for some players by using p.d.f. (2.2), respectively

b	mean	s.d.	b	mean	s.d.
0.502	0.385	0.025	0.000	0.137	0.022
0.498	0.383	0.025	0.500	0.141	0.022
0.503	0.362	0.026	1.000	0.143	0.022
0.497	0.360	0.026	0.667	0.144	0.022
0.428	0.347	0.028	0.750	0.148	0.022
0.504	0.346	0.027	0.059	0.148	0.022
0.496	0.344	0.027	0.056	0.149	0.022
0.445	0.344	0.031	0.053	0.149	0.022
0.454	0.343	0.033	0.067	0.149	0.023
0.461	0.340	0.037	0.050	0.149	0.022

5. 결론 및 제안

본 연구에서는 프로야구에서 타자들의 능력을 가장 보편적으로 나타낸다고 할 수 있는 타격 능력을 베이지안 방법으로 추정하기 위하여 타격의 결과로 나타나는 안타의 수가 타격 능력의 사전분포가 베타분포인 이항분포를 따른다는 가정과 더불어 이항분포의 시행의 수에 해당되는 타격 기회 즉, 타자에게 주워지는 타수 또한 타격 능력에 영향을 줄 것이라는 전제하에 분석하였다. 타격 능력의 사후분포를 유도함에 있어서 데이터로 주어지는 타수와 안타수를 하나하나를 고려하는 것은 복잡하고 비현실적이므로 타수와 안타 수에 의해 결정되지만 흔히 사용하는 타율만을 고려하였으며 본 연구에서 제안한 2개의 타수 분포 중에서 회귀분석 방법으로 얻은 식 (2.2)의 밀도함수가 타격 능력을 추정함에 더 신뢰성이 높다는 것을 알았으며 추정된 타격 능력에 비하여 실제 시합에서의 타율이 상당히 더 높게 나타남을 알았는데 이것은 미국프로야구에 대해 분석한 Frey 논문에서도 유사한 결과를 보여주었다. 즉, 한국프로야구 선수 중에서 순수한 타격 능력을 베이지안 방법으로 추정하였더니 타격 능력의 상위 10명과 하위 10명은 Table 4.1과 Table 4.2와 같다는 것을 알았고 특히 타수분포가 식 (2.1)과 식 (2.2)인 경우에 따라 타격 능력의 최고값은 각각 0.362와 0.385로 추정되었다.

References

Albright, S. C. (1993). A statistical analysis of hitting streaks in baseball. *Journal of the American Statistical Association*, **88**, 1175-1183.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*, 2nd Ed., Springer, New York.

Cho, Y. S. and Cho, Y. J. (2003). The research regarding a Beane Count application from Korean baseball league. *Journal of The Korean Data Analysis Society*, **5**, 649-658.

Cho, Y. S. and Cho, Y. J. (2004). Study about the influence that WHIP has on ERA in 2003 season Korean professional baseball. *Journal of The Korean Data Analysis Society*, **6**, 1415-1424.

Cho, Y. S. and Cho, Y. J. (2005). A study on OPS and runs from Korean baseball league. *Journal of The Korean Data Analysis Society*, **7**, 221-231.

Choi, Y. G. and Kim, H. M. (2011). A statistical study on Korean baseball league games. *The Korean Journal of Applied Statistics*, **24**, 915-930.

Frey. J. (2007). Is an .833 hitter better than a .338 hiter? *The American Statistician*, **61**, 105-111.

James, B. (1982, 1984, 1985, 1988). *The Bill James baseball abstract 1982 1984 1985 1988*, Ballantine Books, New York.

James, B., Zminda, D. and Munro, N. (2000). *STATS all-time major league handbook*, STATS Publishing Inc., New York.

Kim, H. J. (2012). Effects of on-base and slugging ability on run productivity in Korean professional baseball. *Journal of the Korean Data & Information Science Society*, **23**, 1065-1074.

Korea Professional Football League (2014). *2014 K LEAGUE yearbook*, Hanul, Seoul.

Lee, J. T. (2014). Measurements for hitting ability in the Korean pro-baseball. *Journal of the Korean Data & Information Science Society*, **25**, 349-356.

Lee, J. T. and Cho, H. S. (2009). Win-lose models when two teams meet using data mining in the Korean pro-baseball. *Journal of The Korean Data Analysis Society*, **11**, 3417-3426.

Lee, J. T. and Kim, Y. T. (2005). A study on runs evaluation measure for Korean pro-baseball players. *Journal of The Korean Data Analysis Society*, **7**, 2289-2302.

Lee, J. T. and Kim, Y. T. (2006a). A study on the estimation of winning percentage in Korean pro-baseball. *Journal of The Korean Data Analysis Society*, **8**, 857-869.

Lee, J. T. and Kim, Y. T. (2006b). Estimation of winning percentage in Korean pro-sports. *Journal of The Korean Data Analysis Society*, **8**, 2105-2116.

Lee, T. S. (2000). *Physical education an unabridged dictionary*, Minjunseorim, Seoul.

Journal of the Korean Data & Information Science Society
2015, **26**(1), 197–207

http://dx.doi.org/10.7465/jkdi.2015.26.1.197
한국데이터정보과학회지

Bayesian estimation of the Korea professional baseball players' hitting ability based on the batting average[†]

Yong Ju Cho[1] · Kwang Ho Lee[2]

[12]Department of Statistics, Yeungnam University

Received 18 December 2014, revised 11 January 2015, accepted 20 January 2015

Abstract

In baseball game, the hitting ability of batter is frequently assessed by a batting average, a run batted in, a home run, a run scored, an on-base percentage, etc. Recently, more comprehensive indicators such as OPS, ISO, SECA, TA, RC and XR are often used. But, these measures generally shows large deviations since they are calculated from the data for a certain period of time, and they are not an estimate of a population parameter, either. In this paper, we will presume the pure hitting ability of the korea professional baseball players as a parameter which is depend upon at bat. We will estimate the parameter by using the Bayesian method.

Keywords: Ability of hitting, at bat, Bayesian, beta-binomial distribution, posterior distribution, prior distribution.

[†] The present research was conducted by the research fund of Yeungnam University in 2012.
[1] Ph.D student, Department of Statistics, Yeungnam University, Kyungsan 712-749, Korea.
[2] Corresponding author: Professor, Department of Statistics, Yeungnam University, Kyungsan 712-749, Korea. E-mail: khlee@ynu.ac.kr