베이지안을 이용한 이항, 포아송, 정규 분포 데이터 예측

김재훈[1] 이항분포, 프로그래밍, PPT 작성, 보고서 작성

나상우[2] 포아송분포, 프로그래밍, PPT 작성, 보고서 작성

백대환[3] 정규분포, 프로그래밍, PPT 작성, 보고서 작성

요약

손흥민 선수가 올 시즌 득점왕을 차지하여 단순히 슈팅횟수가 많아서인지 골 결정력이 높아서인지 확인하여 분석을 진행을 하였다. 또한 19-20, 20-21시즌 성공과 실패 횟수를 이용하여 21-22예측분포를 구해 실제 21-22이항분포와 비교해본다. 이를 통해 미래에 선수 득점률을 예측하여 구단과 축구통계에 이바지할 것을 기대한다.

코로나 시기에 따른 인구 유동량감소로 인해 교통사고량이 줄어들 것으로 예상이 된다. 따라서 실제 감소했는지 알아보기 위해 2015-2021년 데이터를 이용했고, 사전분포 우도함수, 사후분포를 이용해 베이지안 예측을 진행했다. 분석 결과 유동량감소를 확인했다. 차후 코로나가 완화됨에 따라 유동량 증가로 교통사고에 유의해야 할 것으로 보인다.

정규분포를 따르는 분기별 하의 수치 데이터 집단을 사용하여 다음의 분기에 하의 수치를 예측하는 베이지안 추론을 하였다. 본 논문에서는, 정규사전분포와 관측데이터의 정규분포를 이용하여 정규사후분포와 예측분포를 구하였다. 분석하는 과정을 통해 보급분야에 기여가 되는 결론을 도출해냈다.

주요용어 : 이항분포, 포아송분포, 정규분포, 베이지안, 사전분포, 사후분포, 우도함수.

1. 서론

1.1 이항분포

2022년 프리미어리그에서 대한민국 축구 선수가 아시아 선수 최초로 득점왕을 차지하여 국내 축구 팬들을 물론 해외축구 팬들에게도 깊은 인상을 남긴 선수였다. 리그 마지

[1]30019 세종시 세종로 2511, 고려대학교 공공정책대학 경제통계학부 빅데이터전공 학사과정, E-mail : hun02034@korea.ac.kr

[2]30019 세종시 세종로 2511, 고려대학교 공공정책대학 경제통계학부 빅데이터전공 학사과정, E-mail : ra0622@korea.ac.kr

[3]30019 세종시 세종로 2511, 고려대학교 공공정책대학 경제통계학부 빅데이터전공 학사과정, E-mail : qoreoqhks1@korea.ac.kr

막 경기에 멀티 골을 득점하면서 살라와 공동으로 21-22시즌 총 23골을 기록하여 공동 득점왕을 차지하였다. 더 놀라운 것은 페널티킥 없이 오직 필드골로만 득점한 것이다. 이는 손흥민 선수가 주어진 슈팅 기회가 많아서 득점을 많이 한 것인지 득점률이 높아서인지 알기 위해 분석을 하였다. 이에 비교 케인 선수로 하였다. 득점률은 좋은 퀄리티의 패스에도 영향을 받고 상대 팀에도 영향을 받는다 하지만 두 선수는 같은 팀이기 때문에 다른 팀의 선수들을 비교하는 것 보다 덜 영향을 받을 거로 생각하였다. 또한, 케인은 작년에 득점왕을 받았기 때문에 두 선수의 득점률을 비교하는 것은 의미가 있다. 확률모형으로는 20-21시즌 슈팅성공과 실패 이항분포를 사용하였고, 사전분포로는 바로 확률모형 전년도 19-20시즌 슈팅성공 실패 데이터로 삼았다. 이를 사용하여 21-22 예측분포를 만들기 위해서이다. 또한, 21-22시즌 실제 이항분포와 비교해보았다. 득점률은 공격수 능력을 평가하는 중요한 요소 중 하나인데 분석을 통해 다음 시즌 득점률을 어느 정도 예측할 수 있다면, 팀에 필요한 선수와 방출할 선수를 선택하는 지표 중 하나로 사용될 것이다.

1.2 포아송분포

코로나가 퍼지기 시작하면서 유동인구량이 줄었다. 따라서 유동량이 줄어든 만큼 교통사고가 줄어들 것으로 예상이 된다. 유동량이 제일 많은 서울시를 기준으로 하여 코로나 시기 교통사고 감소량을 알아보고자 한다. 한 연구에 따르면[2] 코로나 시기 타 지역대비 유동량이 증가한 송파구가 있고, 타지역 대비 감소한 성북구가 있다. 따라서 두 구를 기준으로 연구를 진행하고자 한다. 연구를 하기 위해 사전분포를 가정했다. Jeffrey 사전정보를 가정했다. 코로나 시기에 대한 사전분포를 가정할 근거가 없다. 따라서 우도 함수가 0보다 유의하게 큰 부분에서의 변화만 의미가 있게 만들어 극적인 변화를 보여주는 Jeffrey 사전정보를 사용했다. 또한 현재 구한 데이터가 포아송분포가 되므로 이를 통해 우도함수를 구할수 있다. 이 연구를 진행함으로써 차후 코로나가 풀릴 시기 교통사고 증감을 예상하여 교통사고에 대해 미리 대비할 수 있을 것으로 기대한다.

1.3 정규분포

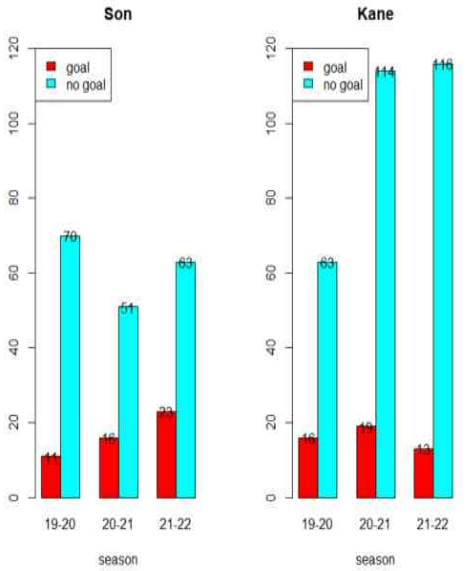
과거 자료 및 현재 자료를 가지고 베이지안 추론 방법을 이용하여 추정하고 분석을 한 후에 서로 그 결과를 비교하여 판단을 해보는 과정을 거쳤다. 본 논문에서는 하의 수치 1분기 데이터의 분포인 정규분포를 사전분포로 정의를 하였고, 2분기 데이터도 정규분포 자료이고 이 데이터도 사용하여 분석을 실시했다. 사후분포 구하는 식을 이용하여 사후분포를 구하였고, 사후분포를 이용해 예측분포 및 최대사후구간(격자점, 사분위수)를 구하였고, 고전적 신뢰구간과 비교를 해보았다. 예측분포를 이용해 결론을 도출해 내면서 논문을 마무리한다.

2. 데이터

2.1 이항 데이터

fbref.com라는 사이트에서 축구통계사이트에서 손흥민 선수와 케인 선수 19-20시즌,

20-21시즌, 21-22시즌 데이터를 참고하여 슈팅 시 골 넣은 횟수, 실패횟수를 얻어 이항데이터를 얻었다. 또한 득점률을 계산하였다.



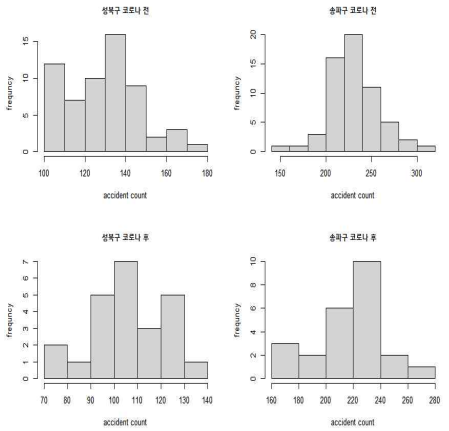
슈팅성공률 (필드골)	Son	Kane
19-20	13.58%	20.25%
20-21	23.88%	14.28%
21-22	26.74%	10.07%

<그림. 1> 각 시즌 별 데이터를 막대 그래프와 표로 표현

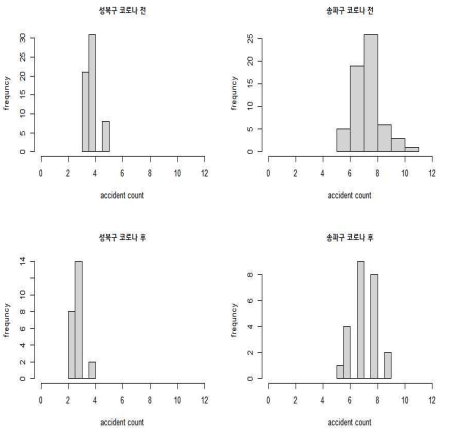
<그림. 1>을 통해 Kane선수가 Son선수에 비해 슈팅횟수가 많은 것을 알 수 있다. Son선수는 득점률이 매 시즌 증가함을 볼 수 있고, Kane선수는 매 시즌 득점률이 낮아지는 것을 볼 수 있다.

2.2 포아송 데이터

'TASS 교통사고 분석시스템'에서 구했으며, 성북구, 송파구 2015년-2021년 월별 교통사고를 구하였다. 구별 코로나 전후로 하여 60개, 24개가 된다. 기존 데이터는 <그림. 2>에서와 같이 월별 교통사고이기 때문에 수치가 크다. 따라서 분석의 편의를 위해 단위별로 수치를 변환시켜줬으며, 25단위 구간으로 끊어 최솟값 구간을 1로, 최댓값을 11로 나눴다. <그림. 3>는 수치 변환 후 그래프이다. 성북구에서 눈에 띄는 교통사고량 차이가 보이는 반면 송파구에서는 격차가 미미한 그래프를 보여준다. 분석에 앞서 sung_before, sung_after, song_before, song_after 변수를 지정해줬다. 각각 성북구 코로나 전, 후 교통사고량, 송파구 코로나 전, 후 교통사고량이다.



<그림. 2> 데이터 변환 전 교통사고 그래프



<그림. 3> 데이터 변환 후 교통사고 그래프

2.3 정규 데이터

데이터는 공공데이터 포털에서 가져온 국방부_공군_신체측정정보(1분기, 2분기)를 이용하였다. 이 데이터 안에서 배꼽 수준 허리둘레, 엉덩이둘레를 평균을 낸 하의 수치 데이터를 사용하였다. 1분기, 2분기 데이터가 Shapiro-test를 하였을 때, p-value가 0.05를 넘지 않아 정규성을 가지지 못했고, 정규성을 가지기 위한 과정으로 log변환(data <- log(data+1))을 진행하였다. 진행 후 Shapiro-test결과 두 데이터 셋의 p-value 모두 0.05를 넘겼다. 따라서 log변환 된 데이터를 이용해 분석을 진행하였다. 분석을 진행하고 최종적으로 보일 때는 지수변환(exp)를 이용해 값을 나타냈다.

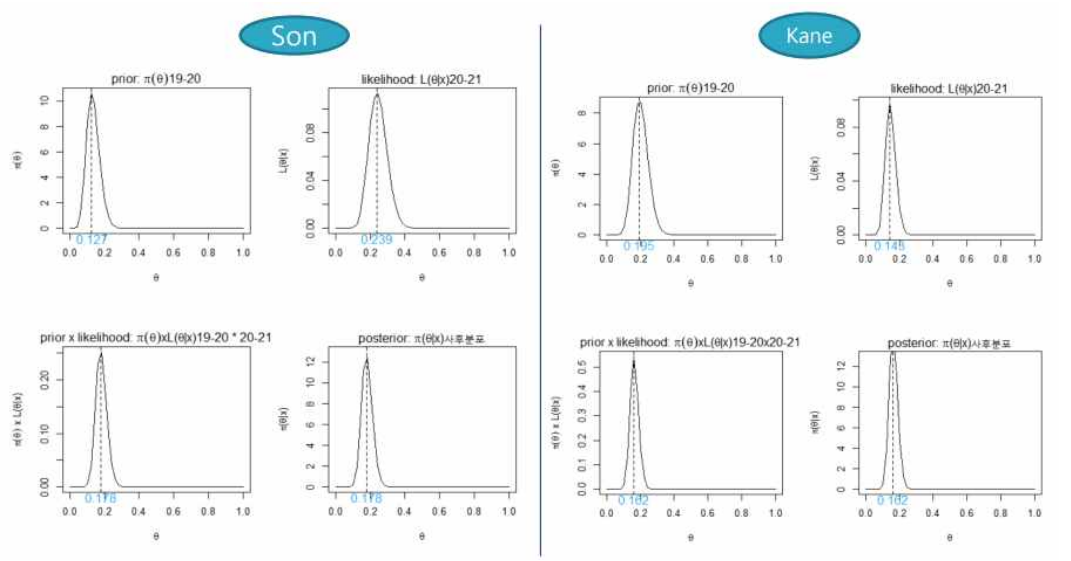
3. 분석 모형

3.1 이항분포에 대한 베이지안 추론

19-20 [식1], [식2]를 이용하여 슈팅성공과 실패횟수를 이용하여 사전분포는 Beta분포를 만들었다. 20-21 슈팅성공과 총 슈팅횟수, 득점률을 이용하여 우도함수를 만들었다. 사후분포는 19-20시즌, 20-21시즌 총 슈팅성공횟수와 실패횟수를 이용하여 만들었다.

$$\mathrm{Beta}(x;a,b)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1} \quad \text{<식. 1>}$$

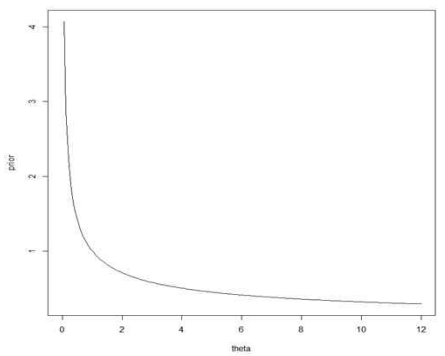
$$P(X=r) = {}_nC_r\, p^r\, q^{n-r}$$ <식. 2>



<그림. 4> 사전분포, 우도함수, 사전분포*우도함수, 사후분포

<그림. 4> 파란색 수치는 최빈값으로 확률분포가 가장 커지는 위치를 의미한다.
사전분포에 영향으로 그래프가 변하였음을 최빈값을 통하여 알 수 있다.

3.2 포아송분포에 대한 베이지안 추론



<그림. 6> 제프리 사전분포 그래프

분석 모형 설명 순서는 사전분포, 우도 함수, 사후분포 순서로 진행된다. 우선 사전분포로 분석을 위해 Jeffrey 사전정보를 가정했다. 코로나 시기에 대한 사전분포를 가정할 근거가 없다. 따라서 우도 함수가 0보다 유의하게 큰 부분에서의 변화만 의미가 있게 만들어 극적인 변화를 보여주는 Jeffrey 사전정보를 사용했다. <그림. 6>는 Jeffrey 사전분포의 그래프이다. 사후분포에서 Gamma(1/2, 0)과 같은 형태를 보여준다.

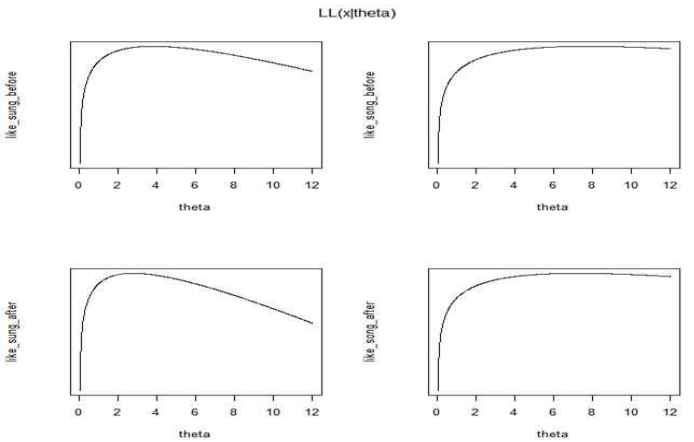
다음으로 우도 함수이다. [식3]와 같은 형태로 나오기 때문에 우도 함수가 발산하는 문제가 발생한다. 따라서 로그 우도 함수를 사용

$$\theta^{\Sigma x_i} e^{-n\theta}$$ <식. 3>

한다.

<그림. 7>는 로그 우도 함수 그래프이다. 왼쪽 성북구 그래프는 오른쪽 송파구 그래프와 달리 전후 차이가 눈에 띄게 보인다.

각 우도 함수의 MLE는 포아송분포의 평균값이다. <표. 1>는 그에 대한 값이다.



<그림. 7> 로그 우도함수 그래프(왼쪽 성북구, 오른쪽 송파구 아래, 위 전, 아래 후)

<표. 1> MLE

sung_before	3.783
sung_after	2.75
song_before	7.783
song_after	7.25

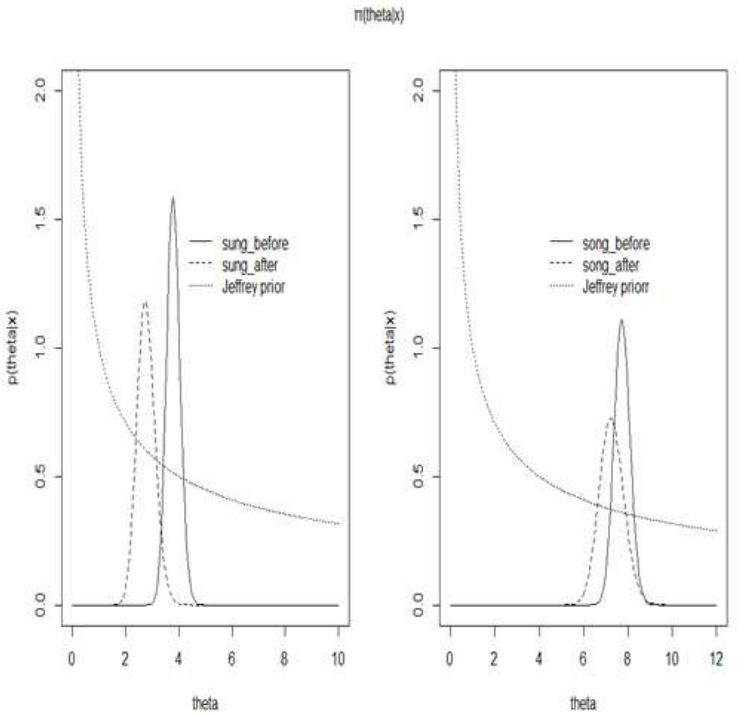
<표. 2> 사후분포(감마분포)

sung_before	Gamma(0.5+227, 60)
sung_after	Gamma(0.5+66, 24)
song_before	Gamma(0.5+464, 60)
song_after	Gamma(0.5+174, 24)

포아송 데이터의 사후분포는 감마 분포를 따른다. <표 2.>는 사후분포 형태이다. 이를 활용한 그래프는 <그림. 8>을 통해 나타난다. 왼쪽은 성북구의 사후 그래프, 오른쪽은 송파구의 사후 그래프이다. 제프리 사전정보가 성북구에서 3~5, 2~4 부분에서, 송파구에서 7~9, 6~8 구간에 유의미하게 영향을 미치는 것을 확인 할 수 있다. 그 영향으로 각 구 코로나 전후 사후분포와 포아송분포의 최빈값 차이는 <표. 3>에서 나타나듯 사후분포에서 더 극명하게 나타난다.

<표. 3> 최빈값

	사후분포 최빈값 차이	포아송분포 최빈값차이	두 분포의 차이
성북구	1.046	1.033	+0.013
송파구	0.496	0.483	+0.013



<그림. 8> 사후분포, 사전분포

3.3 정규분포에 대한 베이지안 추론

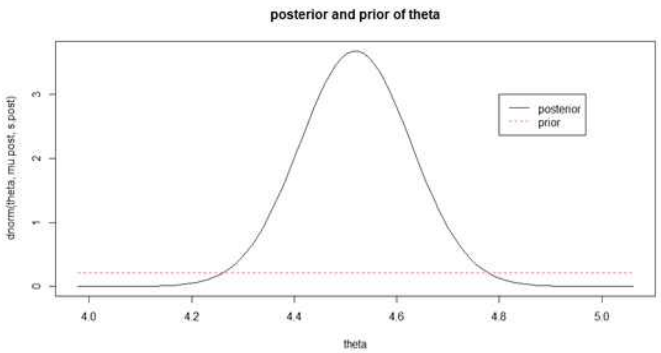
사후분포와 예측분포를 구하기 위해서 사전분포로는 1분기 데이터의 분포를 이용하였다. 1분기 데이터의 분포는 정규분포를 따른다(θ ~ N(μ_0, σ_0^2)). 2분기 데이터는 추가된 관측 데이터(2분기데이터)로 330개의 데이터가 있고, 마찬가지로 정규분포를 따른다((θ ~ N($\overline{x}$, σ^2)). 위 두 데이터에서 나온 수치로 아래의 식에 대입해 사후분포를 구한다(N(μ_n, σ_n^2)).

$$w_n = \left(1 + \frac{\sigma^2}{n\sigma_0^2}\right)^{-1} = \frac{\frac{1}{\sigma^2/n}}{\frac{1}{\sigma^2/n} + \frac{1}{\sigma_0^2}}$$ <식. 3>

$$\mu_n = \frac{\bar{x} + \mu_0\left(\frac{\sigma^2}{n\sigma_0^2}\right)}{1 + \frac{\sigma^2}{n\sigma_0^2}} = w_n\bar{x} + (1 - w_n)\mu_0 \quad \text{<식. 4>}$$

$$\sigma_n^2 = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1} = W_n \cdot \frac{\sigma^2}{n} \quad \text{<식. 5>}$$

사전분포의 분산이 표본평균 $\bar{x}$의 분산(N으로 나눈 값)에 비해 상당히 큰 값으로 사전분포의 영향이 미미해진다. 그러므로 사전밀도함수는 거의 균일분포에 가까워지게 된다. <그림. 9>는 사전분포(빨간색), 사후분포(검은색) 그래프로 나타내 보았다.



<그림. 9>

구해진 사후분포의 평균과 분산을 [식6]에 대입하여 예측분포를 구하는 과정을 진행했다.

$$\begin{aligned} E(X_{n+1}|x_1,\dots,x_n) &= \mu_n = E(\theta|x_1,\dots,x_n) \\ Var(X_{n+1}|x_1,\dots,x_n) &= \sigma^2 + \sigma_n^2 = Var(X_{n+1}|\theta) + Var(\theta|x_1,\dots,x_n) \\ &\geq Var(\theta|x_1,\dots,x_n) \end{aligned} \quad \text{<식. 6>}$$

구해진 예측분포(log된 값)을 실제 수치로 보여주기 위해 지수변환(exp)를 해주었다. 이후, 사후분포에 대해 격자점, 사후분위수 이용한 최대사후구간을 구했고, 고전적 신뢰구간과 비교하였다.

4. 분석 결과

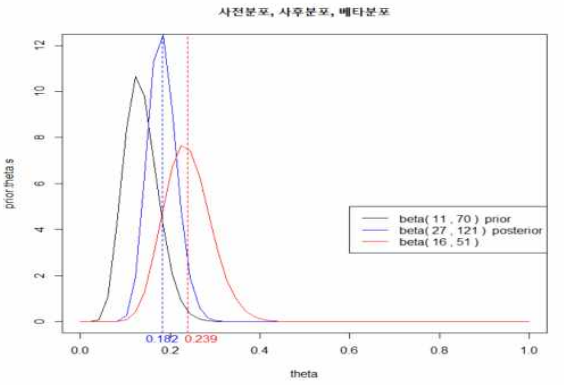
4.1 이항분포에 대한 베이지안 추론

사후분포에 대한, 베이지안 추정치와 MLE 고전 추정치를 [식7]을 이용하여 나타내 보았다.

$$E(\theta|x) = \frac{a+x}{a+b+n}$$
$$Var(\theta|x) = \frac{(a+x)(b+n-x)}{(a+b+n+1)(a+b+n)^2}$$

<식. 7>

(1)Son



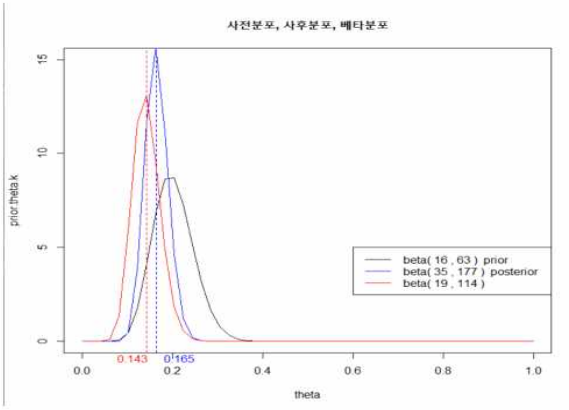
검정-사전 베타분포(19-20)
빨강-베타분포(20-21)
파랑-사후분포

	베이지안	MLE
평균	0.1824	0.2388
분산	0.001	0.0026

<그림. 10> 손흥민 선수에 대한 사후분포

손흥민 선수는 낮은 사후분포 평균에 대한 영향으로 베인지안 추정치평균이 고전적 추정치 평균보다 낮게 나왔다.

(2)Kane



검정-사전 베타분포(19-20)
빨강-베타분포(20-21)
파랑-사후분포

	베이지안	MLE
평균	0.165	0.1428
분산	0.0006	0.0009

<그림. 11> 케인 선수에 대한 사후분포

케인 선수는 높은 사전분포 평균에 대한 영향으로 베이지안 추정치 평균값이 고전적 추정치 평

균값 보다 높게 나왔다.

슈팅100번에 대한 21-22예측분포와 21-22실제 이항분포를 비교해보았다. 예측분포는 식[7]를 사용하여 나타냈고, 이항분포는 식[8]를 이용하여 나타냈다.

$$
\begin{aligned}
f(z|x_1,\cdots,x_n) &= \int f(z|\theta,x_1,\cdots,x_n)\pi(\theta|x_1,\cdots,x_n)d\theta \\
&= \int \binom{m}{z}\theta^z(1-\theta)^{m-z}\frac{\Gamma(a+b+n)}{\Gamma(a+x)\Gamma(b+n-x)}\theta^{a+x-1}(1-\theta)^{b+n-x-1}d\theta \\
&= \binom{m}{z}\frac{\Gamma(a+b+n)}{\Gamma(a+x)\Gamma(b+n-x)}\int \theta^{a+x+z-1}(1-\theta)^{b+n-x+m+z-1}d\theta \\
&= \binom{m}{z}\frac{\Gamma(a+b+n)}{\Gamma(a+x)\Gamma(b+n-x)}\frac{\Gamma(a+x+z)\Gamma(b+n-x+m-z)}{\Gamma(a+b+n+m)} \\
&= \binom{m}{z}\frac{Be(a+x+z,b+n-x+m-z)}{Be(a+x,b+n-x)}
\end{aligned}
$$

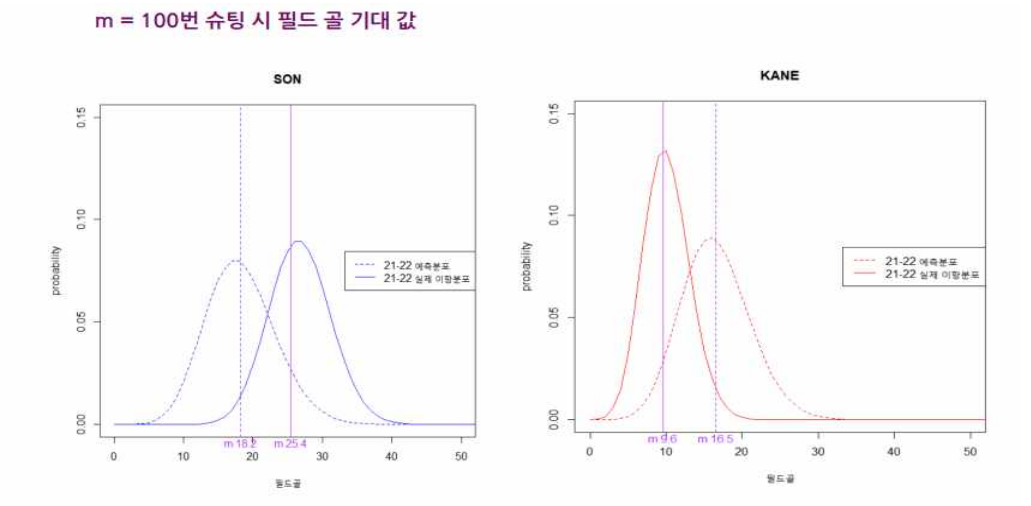
<식. 7>

성공확률

$$p(x) = \binom{n}{x}p^x(1-p)^{n-x}$$

실패확률

<식. 8>



<그림. 12> 21-22 예측분포와 21-22 실제 이항분포

<그림. 12>에서 점선은 21-22예측 분포이고, 선은 21-22이항분포를 나타낸다. 세로축은 100번 슈팅에 대한 평균 골 기대 값을 의미한다. 손흥민 선수는 21-22시즌때 예측보다 더 많은 골을 넣었고, 케인 선수는 예측보다 더 적은 골을 넣은 것을 알 수 있다. 슈팅100번 하였을 때 베이지안 예측치와 고전적 예측치 평균은 <그림. 13> 에서 나타난 평균값에 총 슈팅횟수 100을 곱한 값이다.

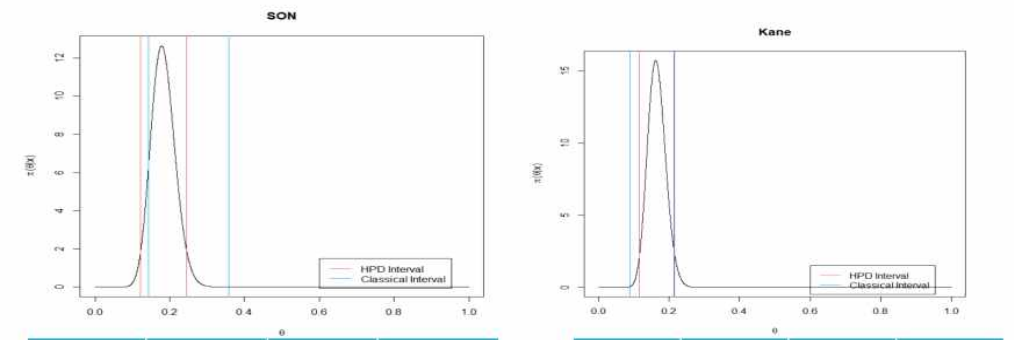
SON

골 기대 값	베이지안	MLE
평균	18.2	23.8
분산	24.8251	18.1777

Kane

골 기대 값	베이지안	MLE
평균	16.5	14.2
분산	20.19	12.24

<그림. 13> 베이지안 예측치와 고전적 예측치

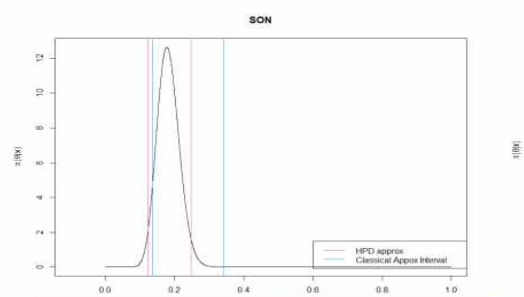


Son	시작	끝	신뢰구간길이
격자점	0.1221	0.2452	0.1231
고전적	0.143	0.3586	0.2155

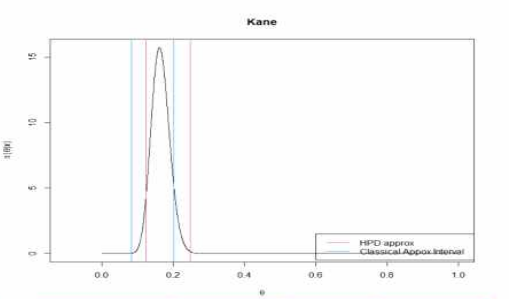
Kane	시작	끝	신뢰구간길이
격자점	0.1164	0.2155	0.0991
고전적	0.088	0.214	0.1258

<그림. 14>격자점 신뢰구간과 고전적 신뢰구간

<그림. 14>는 격자점 방법을 이용하여 95% 신뢰구간을 구하였고 R 패키지에 binom이 제공하는 confint 함수를 사용하여 95% 신뢰구간을 나타낸 그림이다. 고전적 신뢰구간의 길이가 격자점 신뢰구간 길이에 비해서 긴 것을 알 수 있다.



Son	시작	끝	신뢰구간길이
분위수	0.1245	0.2483	0.1236
정규분포	0.1367	0.3408	0.2041



Kane	시작	끝	신뢰구간길이
분위수	0.1183	0.2178	0.0994
정규분포	0.0833	0.2023	0.1189

<그림. 15> 분위수 신뢰구간과 고전적 신뢰구간

$$p \pm 1.96\sqrt{p(1-p)/n}$$ <식. 9>

<그림. 15>는 Beta인 사후분포에서 2.5%와 97.5% 분위수의 구간을 이용하여 베이지안 최대 사후구간을 구하였고, 이항분포에 대한 [식 9]를 정규분포 근사를 하여 95% 신뢰구간을 구하였다. 고전적 신뢰구간의 길이가 분위수를 이용한 신뢰구간 길이에 비교해서 긴 것을 알 수 있다. 베이지안 신뢰 구간길이가 짧은 이유는 베타 사전분포를 이용하여 극단적인 값이 관측되는 경우의 자료를 희석하기 때문이다.

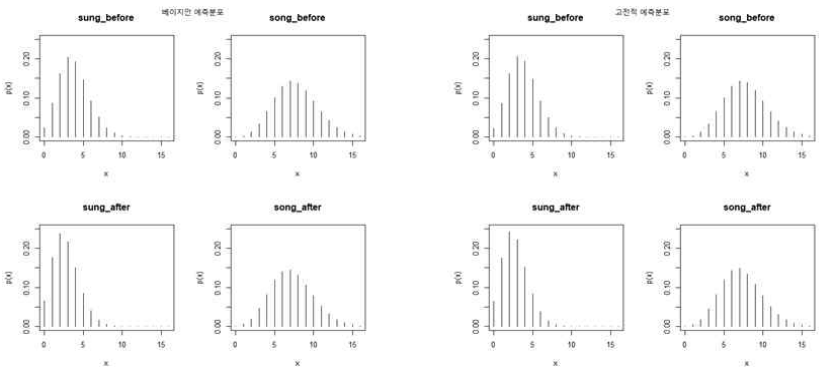
4.2 포아송분포에 대한 베이지안 추론

3장에서 분석 모형을 세웠으므로 분석 결과를 비교한다. <표. 4>에서 사후분포, 고전적 추정치 비교를 했다. 앞서 최빈값에서 볼 수 있듯이, 사후분포의 최빈값 차이가 더 컸다. 따라서 사후분포의 추정치가 더 높을 것을 예측할 수 있고, 실제로 더 커졌다. 사전분포로 감마분포의 알파 값이 더 커졌기 때문이다.

<표. 4> 사후분포, 고전적 추정치 비교

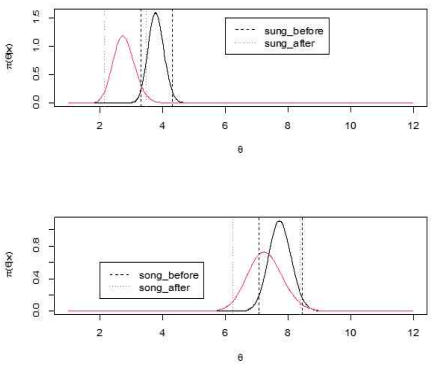
	사후분포 평균	고전적 평균	사후분포 분산	고전적 분산
sung_before	3.792	3.783	0.063	0.063
sung_after	2.771	2.75	0.115	0.114
song_before	7.742	7.733	0.129	0.129
song_after	7.271	7.25	0.303	0.302

예측분포 그래프를 그려 베이지안 예측과 고전적 예측을 비교했다. <그림. 16>에서 볼 수 있듯이 두 그래프의 차이는 미비하다.

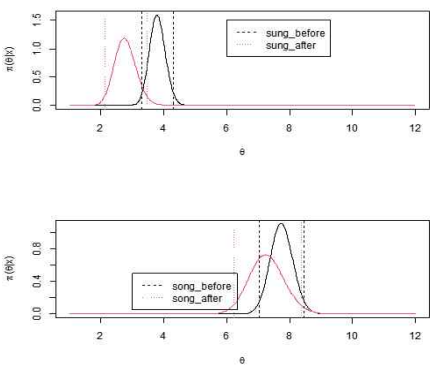


<그림. 16> 베이지안, 고전적 예측분포

<그림. 17>과 <그림. 18>은 격자점과 사후 분위수를 이용한 95% 최대사후구간 그래프이다. 성북구에서 코로나 전후 교통사고량이 크게 차이가 난다. <표. 5>를 통해 격자검과 사후 분위수 구간이 오른쪽으로 이동한 것을 볼 수있다. 사전분포의 영향으로 오른쪽으로 이동했다. 신뢰구간의 길이 차이는 격자검, 사후 분위수, 고전적 모두 비슷하게 나타난다.



<그림. 17> 격자점

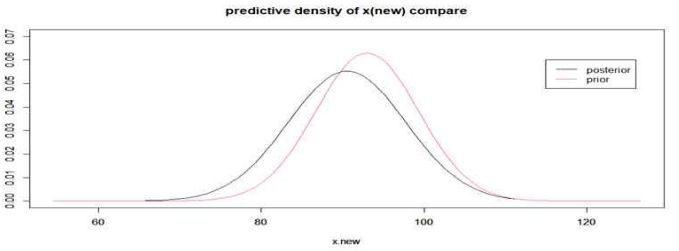


<그림. 18> 사후분위수

<표. 5> 신뢰구간

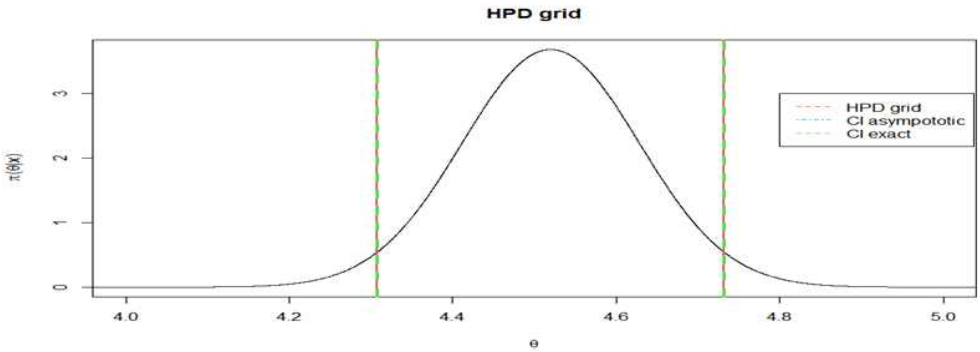
	격자점	사후분위수	고전적	격자점	사후분위수	고전적
sung_before	3.305, 4.288	3.315, 4.3	3.291, 4.276	0.983	0.985	0.984
sung_after	2.12, 3.445	2.145, 3.48	2.087, 3.413	1.325	1.33	1.327
song_before	7.05, 8.447	7.053, 8.461	7.03, 8.437	1.407	1.408	1.407
song_after	6.214, 8.359	6.232, 8.389	6.173, 8.327	2.153	2.156	2.155

4.3 정규분포에 대한 베이지안 추론



<그림. 19> 사전분포, 예측분포 그래프

<그림. 19>는 사전분포와 예측분포의 그래프이다(빨간색 예측분포, 검은색 사전분포). 사전분포와 관측데이터 330개를 이용해 모형을 만들어 사후분포와 예측분포(베이지안 추정치)를 예측하였다. 예측분포의 평균값은 사전분포보다 작게 나타나고, 분산 값은 더 크게 나타난다. 그 이유는 추가된 관측 데이터(2분기) 330개의 분포가 사전분포보다 평균값이 더 작게 나타났고, 분산 값이 더 크게 나타났기 때문이다. 여기서 예측분포는 3분기를 예측한 값이고 3분기에 들어오는 군인의



<그림. 20>

하의수치는 더 작을 것이라고 판단이 가능하다.

<그림. 20>은 격자점과 사후 분위수를 이용한 최대사후구간과 고전적 신뢰구간을 비교한 그래프이다. 빨간색, 파란색, 초록색 점선이 겹쳐있는 것을 볼 수 있다. 모두 95%신뢰구간을 기준으로 구하였고, 최대사후구간을 구한 값이랑 고전적 신뢰구간을 구한 값이 거의 같은 값으로 나타났다. 사후분포가 정규분포를 따르고 좌우대칭이기 때문에 위와 같은 결과가 나타났다.

5. 결론

손흥민 선수는 케인 선수에 비해 슈팅 정확도가 확실히 높은 것을 알 수 있다. 베이지안 예측분포와 실제 이항분포와 비교하였을 때 추정치 값이 어느 정도 비슷하다는 것을 알 수 있다. 따라서 이러한 이항데이터를 이용하여 다음시즌 선수 득점률과 골 기대 값을 예측할 수 있으므로 각 팀의 구단의

원하는 인재상을 영입할 수 있고, 현대 축구 통계에 하나의 지표로 이용이 된다. 본 연구에 부족한점은 예측분포와 실제 이항분포를 비교할 때 t검정을 하여 유의성을 판단하여 객관적이지 않다는 것이다.

포아송 데이터로 교통사고 데이터를 이용했다. 제프리 사전분포를 이용해 우도 함수가 0보다 유의하게 큰 부분에서의 변화만 의미가 있게 만드는 것을 사후분포, 베이지안 추론, 베이지안 예측치를 통해 확인했으며, 격자점을 이용한 신뢰구간이 고전적 신뢰구간보다 더 작은 것을 확인하며 분석을 마쳤다. 앞으로 T-test를 이용하여 실제로 차이가 있는지 검정을 진행해 볼 것이며, 2022년 데이터가 쌓이면 실제값이랑 예측값이 일치하는지 비교해 볼 것이다.

유동량이 타 지역 대비 줄어든 성북구가 코로나 전후 교통사고량 차이가 심했던 것을 보아 결국 유동량이 교통사고에 영향을 미치는 것을 확인할 수 있다. 따라서 앞으로 코로나가 완화됨에 따라 유동량 증가로, 교통사고에 유의해야 할 것으로 보인다.

1분기 데이터(사전분포)와 2분기 데이터(관측데이터)를 이용하여 3분기 예측분포(베이지안 추정치)를 예측해본 결과, 1분기에 비해 더 작아진 값으로 나오는 것을 볼 수 있다. 이는 관측데이터(2분기 데이터)가 더 작은 값을 가지고 있었기 때문이다. 결과적으로, 1분기, 2분기 데이터를 이용해 3분기에는 하의 수치가 더 작게 예측이 됐고, 3분기에 하의를 보급할 때 2분기보다 더 작은 사이즈의 하의를 보급해야 할 것이라고 판단이 가능하다.

참고문헌

[1]최영규. (2011). 축구경기에서 슈팅 방법에 따른 득점 성공률에 관한 연구 = A Study about the Goal Rate of Success of Shooting Method in Soccer Game 국내석사학위논문 원광대학교 교육대학원. p59

[2]나형선, 김지우, 안진현, 전대성, 임동혁. (2021). 코로나-19 전후에 따른 서울시 유동인구, 카드소비 데이터 관계분석. 한국정보처리학회 학술대회논문집. vol.28, no.4, 301.

[3]박주원. (2009). 혼합 정규분포의 베이지안 ROC 곡선 추정. Bayesian ROC curve estimation with a normal mixture distribution.