한국통계학회논문집 제3권 1호
1996년 4월 pp. 169-177

몬테칼로 베이지안 분석과 응용 사례[1)]

강승호[2)], 박태성[3)]

요 약

본 논문에서는 한 유명 농구선수의 과거의 연도별 평균득점과 평균 야투율을 기초로 앞으로의 경기에 대한 평균득점과 평균야투율을 추정하기 위해 몬테칼로 베이지안 분석법 중의 하나인 Sampling-Important-Resampling (SIR) 알고리즘을 이용하였다. 즉 과거의 자료로부터 평균득점과 평균야투율에 대한 사전밀도함수를 설정하고 SIR 알고리즘을 이용하여 사후밀도함수를 구한 후에 이를 기초로 베이지안 추론을 하였다.

1. 서 론

1994년 11월이 되면 이번으로 10년째를 맞이하는 94-95 농구대잔치가 시작된다. 아직 시즌은 시작되지 않았지만 어떤 한 농구 팬이 기아의 허재 선수의 94-95 농구대잔치 시즌 동안 평균 득점에 관심이 있다고 가정하자. <표 1>에서는 허재선수의 과거 농구대잔치 동안의 평균 득점, 총득점, 게임수가 제시되어 있다.

<표 1> 허재선수의 연도별 평균득점

대 회 명	평균득점	총 득점	게임수
84-85 농구대잔치	23.20	441	19
85-86 농구대잔치	20.46	266	13
86-87 농구대잔치	24.13	555	23
88-89 농구대잔치	25.25	505	20
89-90 농구대잔치	26.73	508	19
90-91 농구대잔치	25.40	508	20
91-92 농구대잔치	22.33	469	21
92-93 농구대잔치	26.69	614	23
93-94 농구대잔치	30.00	510	17
합 계	25.08	4390	175

* 자료 SOURCE : 대한농구협회

1) 본 논문은 한국외국어대학교 1995년도 교내 학술 연구비 지원에 의한 결과임.
2) (449-791) 경기도 용인군 모현면 왕산리 한국외국어대학교 통계학과.
3) (449-791) 경기도 용인군 모현면 왕산리 한국외국어대학교 통계학과 조교수.

이제 한 농구 팬이 과거 자료를 바탕으로 94-95 농구대잔치 대회에서 허재선수의 평균득점을 예상하고자 한다. 가장 단순한 방법인 선형회귀모형을 이용해서 평균득점을 추정해 보면, 94-95 년도 농구대잔치에서의 기대되는 평균득점은 약 28점이고 또한 표준편차는 0.24 이고, 95% 신뢰구간은 (19.1945 , 36.3655)으로 추정할 수 있다. 그러나 이 방법은 이 농구 팬의 허재선수에 대한 개인적인 평가 및 신뢰를 반영하지 못한다. 만약 이 농구 팬이 '허재선수의 실력이 과거보다 많이 향상되었다'는 강한 확신을 갖고 있다면 이를 어떻게 추정에 반영할 것인가? 본 논문에서는 사전정보를 추정에 응용할 수 있는 베이지안 방법을 이용하여 위의 자료를 분석하였다.

베이지안 분석 방법은 기존의 고전적인 분석 방법에 비해 유용한 사전 정보의 사용을 가능하게 할 뿐 아니라 표본의 크기가 작을 때에 상대적으로 더 신뢰성 있는 분석을 할 수 있는 장점을 가지고 있다. 베이지안 추론의 주된 목적은 관심 있는 모수의 사후밀도함수(posterior density function)를 사전밀도분포(prior density function)와 우도함수(likelihood function)로부터 구하는 것이다. 이 추론 과정은 대부분의 경우에 복잡한 적분 계산을 요구하고 있기 때문에 베이지안 분석 방법은 우도함수와 사전밀도함수가 짝관계(conjugacy)가 성립하는 몇몇 경우에 대해서만 국한적으로 사용되어 왔으며 실제 자료의 분석에는 많은 어려움이 있었다. 그러나 Albert(1993)가 지적한대로 최근 들어 컴퓨터의 발달로 사후밀도함수를 수치적으로 구할 수 있는 몬테칼로 방법이 널리 사용됨에 따라 베이지안의 분석 방법은 여러 분야에서 사용되게 되었다.

몬테칼로 방법이란 사전에 모수(parameter)에 임의의 값을 설정하고 관측 안되는 변수(unobservable variables)의 값에 대해서는 일정한 밀도함수를 가정한 후 이 밀도함수로부터 관측값(random number)을 생성시킨 다음 이 값을 가지고 다시 거꾸로 모수를 추정하는 과정을 반복하여 모수에 대한 속성을 사후적으로 구하는 방법이다. 특히 이러한 몬테칼로 방법을 통해 모수에 대한 단편적인 추론에서 보다 나아가 모수의 사후밀도함수에 대한 전체적인 분포를 추정하는 것이 가능하게 되었다.

본 논문에서는 Albert(1993)의 연구 결과를 바탕으로 몬테칼로 방법 중에서 하나인 Rubin (1987) 의 Sampling-Important-Resampling (SIR) 방법을 소개하고 허재선수의 자료를 분석해서 SIR방법을 예시하였다. 2절에서는 SIR 방법에 대하여 간단하게 설명하였으며, 3절에서는 <표 1>의 자료로부터 평균득점에 대한 사전밀도함수를 설정하고 이로부터 SIR 알고리즘을 이용하여 사후밀도함수를 구하여 베이지안 추론을 기술하였다. 아울러 평균득점외에 평균야투율에 대한 분석도 포함하였다. 마지막으로 4절에서는 이제까지 기술된 내용을 정리하고 소개한 방법의 이점을 살펴보았다.

2. SIR 연산법

SIR 방법은 Rubin (1987)이 제안한 방법으로 확률분포 $g(\theta)$로부터 표본의 값을 생성시키는데 관심이 있으나 직접 $g(\theta)$로부터 표본을 생성시키는 것이 어려운 경우에 $g(\theta)$에 근사하는 다른 함수 $h(\theta)$를 이용하여 $g(\theta)$의 표본 값을 생성시키는 방법이다. 물론 $g(\theta)$ 보다 $h(\theta)$로부터 표본을 생성시키는 것이 쉬운 경우에 사용할 수 있다.

좀 더 구체적으로, $g(\theta)$로부터 m개의 표본을 생성시키고자 할 때 어떻게 SIR 알고리즘을 적용하는 지 살펴보자. 먼저 $h(\theta)$로부터 표본 $\{\theta_1, \cdots, \theta_m\}$ 을 생성한 후 각 θ_i에 대하여 가중치 $w(\theta_i) = g(\theta_i)/h(\theta_i)$ 를 계산한다. 다음으로 표본 $\{\theta_1, \cdots, \theta_m\}$ 로부터 θ_i가 뽑힐 확률이 $w(\theta_i)$에 비례한다고 가정한 후 m개의 표본을 복원추출하여 새로운 표본 $\{\theta_1^*, \cdots, \theta_m^*\}$ 를 얻는다. 이 과정을 계속 반복하여 얻은 새로운 표본 $\{\theta_1^*, \cdots, \theta_m^*\}$ 는 근사적으로 $g(\theta)$의 밀도함수를 갖는 분포에서 생성된 것으로 간주할 수 있다. 표본의 크기 m이 큰 경우에는 표본 $\{\theta_1^*, \cdots, \theta_m^*\}$ 의 분포는 $g(\theta)$의 분포와 근사적으로 같은 분포를 갖고 이 근사는 표본의 크기 m이 커질수록 정확하게 된다. 좀 더 자세한 설명과 이론은 Efron (1982) 과 Gelfand 와 Smith (1990)를 참조하기 바란다.

그러면 이제 SIR 알고리즘이 베이지안 추론에 어떻게 적용될 수 있는 지 알아보도록 하자. $\pi(\theta)$는 사전밀도함수이고 $L(\theta)$는 우도함수라고 하자. 그러면 사전밀도함수와 우도함수로부터 사후밀도함수는 특정의 상수 k에 대해

$$\pi(\theta|\text{data}) = k\,\pi(\theta)\,L(\theta)$$

가 성립한다. 여기서 $h(\theta) = \pi(\theta)$이고 $g(\theta) = \pi(\theta|\text{data})$로 놓으면 가중치는

$$w(\theta) = g(\theta)/h(\theta) = \pi(\theta|\text{data})/\pi(\theta) = k\,L(\theta). \qquad (1)$$

이 경우에 관심 있는 사후밀도함수 $\pi(\theta|\text{data})$로부터는 표본을 생성시키기가 어렵고 사전밀도함수 $\pi(\theta)$로부터는 표본을 생성시키기가 상대적으로 쉬운 경우를 생각해 보자. 먼저 사전밀도함수 $\pi(\theta)$로부터 표본 $\{\theta_1, \cdots, \theta_m\}$을 생성한 후 각 θ_i에 대하여 가중치를 (1)로부터 구한다. 다음으로 표본 $\{\theta_1, \cdots, \theta_m\}$로부터 θ_i가 뽑힐 확률이 $w(\theta_i)$에 비례한다고 가정한 후 m개의 표본을 복원추출하여 새로운 표본 $\{\theta_1^*, \cdots, \theta_m^*\}$를 얻는다. 이 과정을 계속 반복하여 얻은 새로운 표본 $\{\theta_1^*, \cdots, \theta_m^*\}$는 근사적으로 사후밀도함수 $\pi(\theta|\text{data})$의 밀도함수를 갖는 분포에서 생성된 것으로 간주할 수 있다. 보다 자세한 설명은 3장에서 예제와 함께 설명하겠다.

3. 농구 선수의 관한 예제

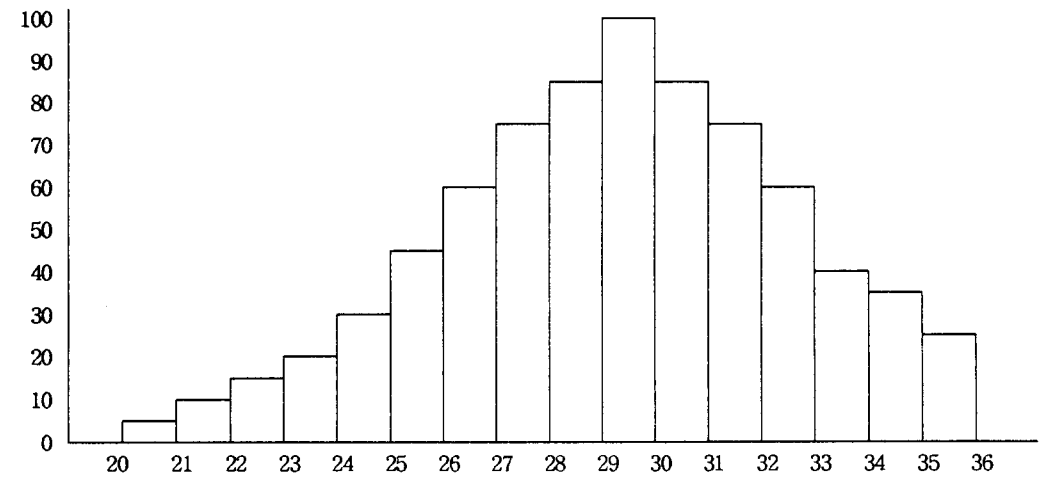
3.1 평균득점

이제 한 농구 팬이 <표 1>에 있는 허재선수의 자료를 바탕으로 94-95 농구대잔치 대회에서 허재선수의 평균득점을 예상하고자 한다. 우선 <표 1>에서 과거의 평균득점보다 93-94년도의 기

록이 월등히 높음을 알 수 있고, 또한 그의 실력이 원숙해 졌다는 소식을 대중매체를 통해서 들을 수 있었기 때문에 이 농구 팬은 94-95 시즌에서는 과거 기록들보다는 향상되리라는 기대를 하고 있다.

위의 정보를 바탕으로 평균득점에 대한 사전확률분포를 설정할 수 있다. 여기서는 Berger(1985)가 제시한 확률 분포 히스토그램의 근사 방법을 사용하겠다. 확률 분포 히스토그램의 근사 방법은 표본의 확률 분포 히스토그램을 이용하여 모수의 확률 분포를 근사적으로 구하는 방법이다.

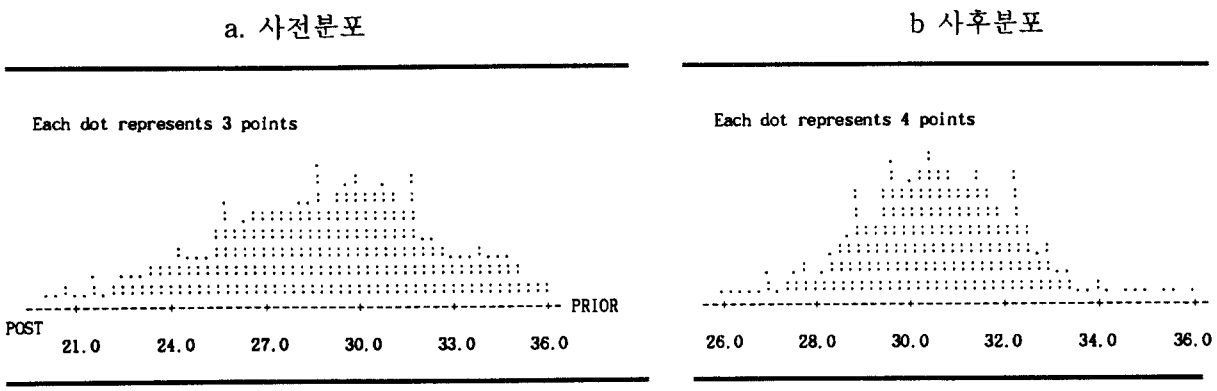
이 농구 팬은 허재선수의 평균득점에 대하여 득점 가능한 범위를 몇 개의 소구간으로 나눈 다음, 주관적으로 각각의 소구간에 확률값을 기입하였다. 이 농구 팬이 확신하는 평균득점의 범위는 (20,36) 이고 이 범위는 1점 간격으로 16개의 소구간으로 나누어 놓는다. <그림 1>에서는 사전확률분포 히스토그램이 그려져 있다. 그림에서 29점의 구간이 가장 높은 사전확률값을 갖는 것을 알 수 있다. 이는 과거의 평균 기록(25.08) 보다 높은 값을 갖는다. 즉 허재선수의 평균득점이 다소 증가되리라는 기대를 반영한 것이다.



<그림 1> 평균득점에 대한 사전확률분포 히스토그램

이 농구 팬이 허재선수의 최근 경기를 관찰하여 평균 30득점과 표준편차 1.731을 얻었다고 가정하면, 이를 바탕으로 우도함수를 설정할 수 있다. 즉 평균득점은 정규분포를 따르고 평균은 30이고 표준편차는 1.731이라고 가정하자.

이제 2장에서 설명한 SIR 연산법을 사용하여 평균득점의 사후확률분포를 SIR 연산법을 이용하여 얻을 수 있다. <그림 2>는 표본 1000개의 사전확률분포와 사후확률분포를 나타낸다. 또한 <표 2>는 사후확률분포의 분석 결과들이다. 결론적으로 이 농구 팬이 구한 허재선수의 94-95 시즌 동안의 평균득점은 29점이고 표준편차는 1.520 , 95% 신뢰구간은 (26.021 , 31.979) 이다.



<그림 2> 평균득점에 대한 사전분포와 사후분포의 표본분포

<표 2> 사전분포와 사후분포의 단순통계량

	PRIOR	POSTERIOR
N	1000	1000
MEAN	29.241	29.950
MEDIAN	29.462	29.902
TRMEAN	29.297	29.968
STDEV	3.203	1.520
SEMEAN	0.101	0.048
MIN	20.229	25.435
MAX	35.945	34.448
Q1	27.046	29.064
Q3	31.477	31.034

SIR 방법의 결과와 앞에서 구한 선형회귀분석의 결과를 비교해 보면, 평균득점의 추정량의 값은 서로 비슷하나 95% 신뢰구간에서는 많은 차이를 보이고 있다. 즉, 베이지안 신뢰구간이 선형회귀모형을 가정하여 얻은 신뢰구간 보다 좁게 되어 결과적으로 베이지안 방법이 좀 더 정확한 추정을 하였다고 볼 수 있다.

3.2 야투율

이제 이 농구 팬이 허재선수의 94-95 시즌에서의 평균득점 외에 야투율(=성공수/투사수)의 추정에도 관심이 있다고 하자. 먼저 허재선수의 과거의 야투율을 2점 야투율과 3점 야투율로 나누어

조사한 결과가 <표 3>과 <표 4>에 정리되어 있다.

2점 야투율은 대략적으로 48% 에서 59% 까지 기록했으며 최근에 이르러서는 60% 에 가까운 높은 성공률을 보이고 있다. 3점 야투율 역시 35% 에서 43% 까지의 기록을 가지며 최근에서는 40%에 가까운 성공률을 보이고 있다. 이 기록에서 알 수 있듯이 허재선수의 야투율은 평균득점과 마찬가지로 향상되고 있고, 94-95 시즌에서도 좋은 성적을 얻으리라 기대하고 있다.

3점 야투율 자료를 보면 86년도 시즌에서 가장 좋은 성적인 43.5% 를 기록했다. 그 후 89년도에서도 40% 가 넘는 좋은 기록을 보였지만, 90년대에 이르러서는 성공률은 다소 줄어듦을 보인다. 하지만 2점 야투율은 90년대에 이르러 향상됨을 보이는데, 이는 80년대에는 3점 슛을 주력한 반면에 90년대에는 2점 슛에 더욱 노력한 것으로 추측된다. 따라서 2점 야투율과 3점 야투율에 관한 이 자료와 언론에서 보도된 허재선수의 실력 향상에 관한 정보 및 각 개인의 기대와 신뢰를 바탕으로 사전확률분포함수를 설정할 수 있다. 사전확률분포함수의 설정은 평균득점과 같이 Berger(1985)의 확률분포 히스토그램 근사방법을 이용하여 구하였다.

<표 3> 허재선수의 연도별 평균 3점 야투율

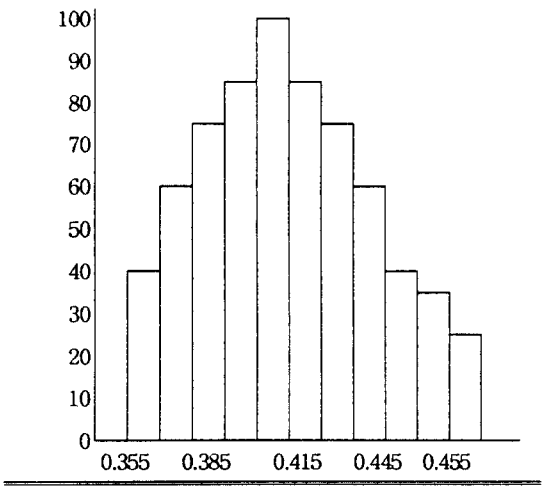
대 회 명	평균 성공률	성공횟수	총횟수
84-85 농구대잔치			
85-86 농구대잔치	36.6	22	60
86-87 농구대잔치	43.5	91	209
88-89 농구대잔치	35.5	59	166
89-90 농구대잔치	41.1	77	187
90-91 농구대잔치	35.8	69	194
91-82 농구대잔치	35.8	56	156
92-93 농구대잔치	38.2	83	217
93-94 농구대잔치	38.3	76	198
합 계	38.4	533	1387

<표 4> 허재선수의 연도별 평균 2점 야투율

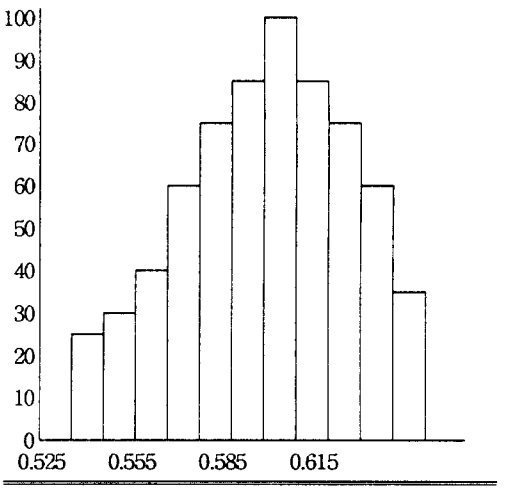
대 회 명	평균 성공률	성공횟수	총횟수
84-85 농구대잔치	50.3	199	395
85-86 농구대잔치	48.0	75	156
86-87 농구대잔치	51.0	117	229
88-89 농구대잔치	58.7	121	206
89-90 농구대잔치	51.5	102	198
90-91 농구대잔치	55.4	122	220
91-82 농구대잔치	51.5	144	221
92-93 농구대잔치	59.4	148	249
93-94 농구대잔치	58.1	96	165
합 계	55.1	1124	2039

* 자료 SOURCE : 대한농구협회

<그림 3>과 <그림 4>는 2점과 3점의 야투율에 대한 사전확률분포 히스토그램을 나타낸다. 3점 야투율에 대한 히스토그램은 구간 (0.355, 0.455) 까지의 분포로 나타내었고 2점 야투율에 대한 구간은 (0.525, 0.625)의 분포로 나타내었다. 두 분포에서 가장 높은 사전확률값을 값은 것은 각각 0.395 와 0.585 의 값이다. 이 역시 각각의 야투율에서 가장 신뢰되는 값이라고 말할 수 있다. <그림 3>은 0.395에서 다소 오른쪽으로 치우쳐 있음을 알 수 있고, <그림 4>의 경우는 0.585에서 다소 왼쪽으로 치우쳐 있음을 알 수 있다. 이는 앞서 말했듯이, 이 농구 팬이 허재선수의 야투율은 각각 0.395 와 0.585에서 가장 높은 기대를 하지만, 3점 야투율의 경우는 좀더 높은 성공률을, 2점 야투율은 좀더 낮은 성공률을 기록하리라는 기대를 나타낸 것이다.

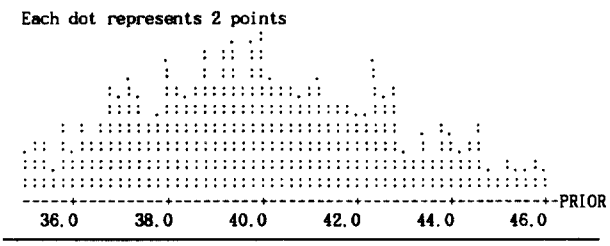


<그림 3> 3점 야투율에 대한 사전확률 히스토그램

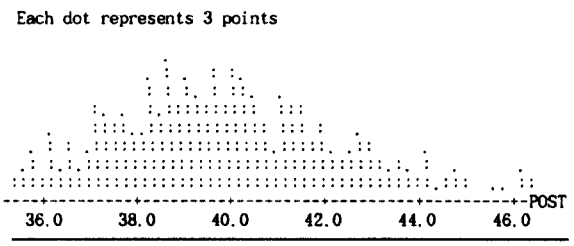


<그림 4> 2점 야투율에 대한 사전확률 히스토그램

a.사전분포

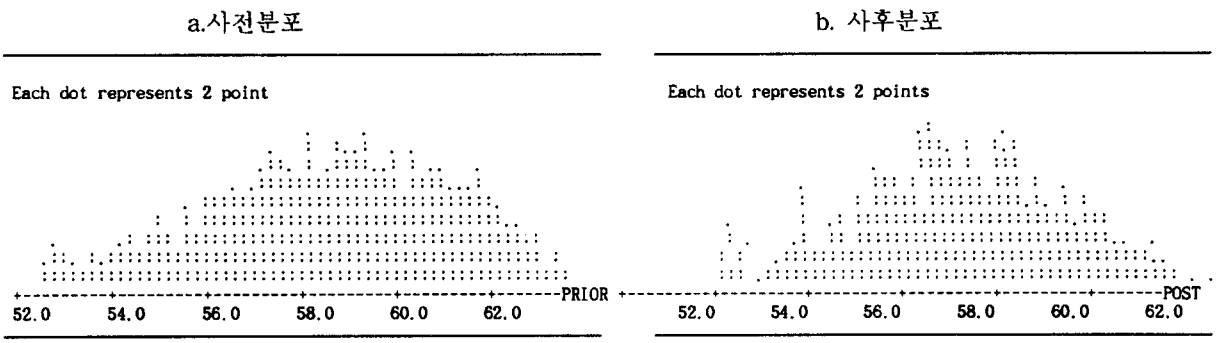


b. 사후분포



<그림 5> 3점 야투율에 대한 사전분포와 사후분포의 표본분포

이제 이 두 사전확률 히스토그램으로부터 1000개의 표본을 뽑아 사전확률분포를 구한 그림이 <그림 5.a>와 <그림 6.a>에 나타나 있다. 다음으로 우도함수를 설정하기 위해 평균득점과 마찬가지로 이 농구 팬이 허재선수의 최근 경기를 관찰한 결과 3점 야투율은 100개의 슛 중에서 38개를 성공시켜 38%를 기록했고, 2점 야투율은 100개의 슛 중에서 54개를 성공시켜 54%를 기록했다고 가정하자. 이 경우 성공한 슛의 수는 이항분포를 따르지만 표본수가 충분히 크고 야투율이 0과 1에 가깝지 않으므로 정규분포로 근사하여 계산하였다. 다음으로 SIR 연산법을 이용하여 근사적으로 사후확률분포로부터 1000개의 표본을 생성시켜 구한 분포가 <그림 5.b>와 <그림 6.b>에 나타나 있다. 이 표본들을 기초로 <표 5>와 <표 6>에서는 사전확률분포와 사후확률분포의 기초통계량이 제시되어 있다.



<그림 6> 2점 야투율에 대한 사전분포와 사후분포의 표본분포

<표 5> 3점 야투율에 대한
사전분포와 사후분포의 단순통계량

	PRIOR	POSTERIOR
N	1000	1000
MEAN	39.996	39.491
MEDIAN	39.817	39.375
TRMEAN	39.958	39.433
STDEV	2.6250	2.2670
SEMEAN	0.0830	0.0720
MIN	35.007	35.007
MAX	45.986	45.986
Q1	38.064	37.918
Q3	41.957	40.990

<표 6> 2점 야투율에 대한
사전분포와 사후분포의 단순통계량

	PRIOR	POSTERIOR
N	1000	1000
MEAN	58.030	57.189
MEDIAN	58.170	57.236
TRMEAN	58.087	57.199
STDEV	2.5360	2.4180
SEMEAN	0.0800	0.0760
MIN	52.027	52.027
MAX	62.939	62.927
Q1	56.279	55.537
Q3	59.999	58.840

이상의 결과를 정리하면 이 농구 팬은 94-95년도 시즌에서의 허재선수의 3점 야투율을 39.491% 로 추정하였고 이에 대한 표준편차는 2.267 이고 95% 신뢰구간은 (35.05%, 43.94%)로 추정하였다. 또한 94-95년도 시즌에서의 허재선수의 2점 야투율을 57.189%로 추정하였고 이에 대한 표준편차는 2.418 이며 95% 신뢰구간은 (52.45% , 61.93%)로 추정하였다. 이 추정값들은 과거 허재선수의 평균야투율보다 조금 큰 값을 갖는다. 즉 이 결과는 허재선수의 실력이 과거보다 많이 향상되었다는 이 농구 팬의 신뢰가 반영되었음을 보여주는 것이다.

4. 결 론

베이지안 방법의 가장 큰 특징은 개인의 신뢰를 추론에 반영할 수 있다는 점이다. 즉 3절에서 설정한 허재선수의 평균득점과 야투율에 대한 사전확률분포는 한 농구 팬이 평가하는 허재선수의

실력에 대한 주관적인 신뢰를 나타내고 있다. 이와는 달리 최근 언론을 통해서 허재선수의 실력이 원숙해졌다는 소식을 바탕으로 허재선수의 실력을 더 뛰어나게 평가하는 팬은 <그림 1>보다 좀 더 오른쪽으로 치우치고 왼쪽으로 긴 꼬리를 갖는 모양의 사전분포를 설정할 것이다. 또한 최근에 허재선수가 슬럼프에 빠졌다고 믿는 팬은 허재선수의 실력을 과거보다 더 좋지 못할 것으로 평가하여 <그림 1> 보다 좀 더 왼쪽으로 치우치고 오른쪽으로 긴 꼬리를 갖는 모양의 사전분포를 설정할 것이다. 이와 같이 허재선수에 대한 능력 평가가 개인마다 다르더라도 3절에서 소개한 SIR 방법을 사용하여 각 개인 신뢰를 반영하는 베이지안 추정량을 쉽게 구할 수 있다.

베이지안 방법은 일상생활에서 접하는 응용된 문제를 해결하는데 좋은 방법으로 인식되고 있음에도 불구하고 이론적 배경의 어려움과 사전밀도함수의 설정, 사전밀도함수와 우도함수의 결합의 어려움 등으로 실질적으로는 널리 사용되지 못해온 실정이었다. 그러나 본 논문에서 소개한 바와 같이 SIR 연산법은 예제와 같은 실제적인 문제를 분석하는데 유용하게 사용될 수 있고 통계학 이론에 대한 지식이 부족한 사람도 쉽게 사용할 수 있을 뿐더러 이에 필요한 컴퓨터 프로그램도 쉽게 작성할 수 있으리라 생각된다.

참고문헌

[1] Albert, J. H. (1993). Teaching Bayesian Using Sampling Methods and MINITAB, *The American Statistician*, Vol. 47, 182-191.

[2] Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, New York: Springer-Verlag.

[3] Efrond, B. (1982). *The Jacknife, the Bootstrap, and Other Resampling Plans.* National Science Foundation - Conference Board of the Mathematical Sciences Monograph 38, Philadelphia: SIAM.

[4] Gelfand, A. E. and Smith, A. F. N. (1990). Sampling-based Approaches to Calculating Marganl Densities, *Journal of the American statistical Association*, Vol. 85, 398-409.

[5] Rubin, D. B. (1987). A Noniterative Sampling / Importance Resampling Alternative to the data Augmentation Algorithm for Creating a Few Imputation Are Modest ; The SIR Algorithm, *Journal of the American statistical Association*, Vol. 82, 543- 546.