Historia Mathematica Vol. 13, No. 1, 2000, 99-112

베이즈와 이산형 모형을 이용한 비율에 대한 추론 교수법의 고찰

서경대학교 수리정보통계학부 **박태룡**

Abstract

In this paper we discuss the teaching methods about statistical inferences. Bayesian methods have the attractive feature that statistical conclusions can be stated using the language of subjective probability. Simple methods of teaching Bayes' rule described, and these methods are illustrated for inference and prediction problems for one proportions. Also, we discuss the advantages and disadvantages of traditional and Bayesian approachs in teaching inference

0. 서론(통계적 추론의 교수법)

본 논문은 1995년 앨버(Jim Alber)가 *Journal of Statistics Education*에 기고한 논문을 참고하여 그 내용을 소개하고 대학의 기초통계학 교육중 추정과 관련된 부분의 개념을 실제 문제를 가지고 베이지안 관점에서 접근하는 것을 소개하고자 한다. 통계교육에서 주관적 확률을 고려하는 베이지안 접근 방법을 보강하여 전적으로 자료에만 의존하는 고전적 접근 방법에서 탈피하는 시도가 이루어져야 할 것으로 믿는다. 고전적 관점에서 기초적인 통계적 추론을 가르친다는 것은 표본분포를 가르치는 것에 따른 어려움과 통계적 신뢰성에 대한 정확한 설명의 어려움 때문에 매우 어렵다. 베이지안 방법은 주관적 확률에 대한 전문어를 사용하여 통계적 결론이 논의되기 때문에 매력적인 특징을 가지고 있다. 베이즈 규칙을 가르치는 단순한 방법이 소개되고 이러한 방법들이 하나의 비율에 대한 추론 문제와 예측 문제에 대하여 소개된다. 또한, 고전적 접근 방법과 베이지안 접근 방법의 유리한 점과 불리한 점을 논의한다.

한 학기 동안 전공교양 과정으로 기초통계학을 가르치는 것을 고려해보자. 수강자들에게 통계적 추론을 소개하기 위하여 많은 통계 교육자들은 t-검정이나 분산분석과 같은 특별한 방법을 가르치는 것이 필요하지 않다고 믿는다. 좀더 강조되어야 할 것은 일반적인 추론에

대한 개념이라고 생각한다. 한 가지 중요한 개념은 모집단과 그들의 특징을 나타내는 모수, 그리고 표본과 그들의 특징을 나타내는 통계량들간의 구별이다. 추론에서 두 번째로 유념해야할 것은 통계적 신뢰성의 의미이다, 특히, 95% 신뢰구간에 대한 설명인 것이다. 수강자들은 표본 크기의 역할을 이해해야만 한다. 일반적으로 임의표본으로 취해진 많은 데이터는 모집단의 모수에 대하여 더 많은 정보를 제공한다. 또 다른 기본적 인식은 표본 추출 과정과 확률모형에 대한 가정을 토대로 통계적 결론은 종속되고 있다는 것이다.

위에서 언급한 추론 개념에 대하여 가르칠 때 고전적 접근은 확률에 대한 상대도수 인식에 기초한다는 것이다. 이런 접근 방법이 실제로 기초적인 모든 통계학 교과서에 나타나있음에도 불구하고 가르치는 것이 어려울 수 있다. 특히, 모집단의 비율을 나타내는 모수 p와 표본으로부터 계산되는 비율에 대한 통계량 $\hat{p}$ 등과 같은 모수에 대한 구별이 수강자들에게는 어려울 수 있다. 수강자들에게 표본분포에 대한 개념은 종종 정확히 알수 없는 신기한 것 일 수도 있다. 하나의 표본분포 대한 많은 논의 가운데 몇 가지를 열거해보면 임의확률표본을 취하는 것에 대한 개념, 표본으로부터 통계량을 계산하는 것, 통계량의 행태(behavior)를 이해하기 위한 반복 표본추출 등이 있다. 첨가하여, 통계적 신뢰성의 정확한 설명을 전달하는데 어려움이 있을 수도 있다. 만약, 특별한 표본으로부터 비율 p에 대한 95% 신뢰구간을 계산한다면 이러한 특별한 구간이 높은 확률을 가지고 관심의 대상이 되는 모수를 포함한다고 생각할 수 있을 것이다. 그러나 이러한 생각은 바로 잡아야 한다. 고전적 추론에서 우리는 확률구간을 모수가 포함되는 확률구간으로서만 인식하고 있다. 같은 방법으로, 하나의 p 값도 적어도 극단값을 하나의 관측되는 값으로 관측하는 표본결과의 확률 대신에 귀무가설의 확률로 잘못 소개될 수도 있다. p 값에 대한 인식이 직관적이지 않기 때문에 이러한 오류를 범하기 쉽다. 데이터가 수집되면 귀무가설을 뒷받침하는 면에서 관측된 데이터에 포함된 증거의 정도에 관심이 있다. 여기서 우리는 왜 우리가 관측한 것보다 좀더 극단적인 표본결과의 확률에 관심을 갖아야만 하는가 하는 데 관심을 가질 수 있다?

베이지안 접근 방법으로 위에서 언급한 추론 개념을 가르치는 것을 생각해보자. 추론에서, 우리는 모집단 모수의 값에 대하여 확신하지 못하고, 관측된 데이터로부터 모수에 대한 부가적인 정보를 얻는다. 베이지안 추론에서, 모수에 대한 불확실성은 주관적 확률을 이용하여 표현되고, 베이즈 규칙(Bayes' Rule)은 데이터로부터 관찰자의 주관적 인식(확률)을 개선하는 데에 대한 하나의 기법(메카니즘)이다.

고전적 통계를 가르치는데 있어서 한 가지 어려움은 통계적 '신뢰'(confidence)에 대하여 정확한 설명 또는 묘사를 전달하는 것이다. 베이지안 신뢰에 대한 진술은 모수에 대하여 관찰자의 인식이 확률언어를 사용하여 묘사되기 때문에 수강자들에게 좀더 친근할 수도 있다. 관측된 자료가 주어졌다면, 모수를 포함하는 고정된 구간의 확률에 대하여 언급할 수 있거나 통계적 가설이 사실일 확률에 대하여 언급할 수 있을 것이다. 그래서 하나의 비율 p에 대한 95% 베이지안 구간 추정치는 확률 0.95를 갖고 p를 포함하는 하나의 구간이다. 만약 관찰자가 앞면의 비율 p에 대한 확률을 갖는 동전이 있다면, 관찰자는 그 동전이 공정하다($p=0.5$)하다는 가설에 대한 확률을 고려할 수 있다.

추론을 가르친다는 관점에서 베이지안 추론의 매력적인 특징은 관측된 데이터를 기초하여 조건적으로 이루어진다는 것이다. 고전적 통계에서는 실제 관측되는 데이터와 구별되는 자료 집합들의 가능성을 반드시 고려해야만 한다. 베이지안 추론에서는 결론을 이끌어내는 데에 관계된 데이터 집합은 관찰자가 관측한 바로 그 자료 집합이다.

위에서 열거한 여러 이점에도 불구하고 추론을 가르치는데 있어서 베이지안 접근 방법은 고전적 통계 수업에서 배제되었을 수도 있는 새로운 교수법이 요구된다. 보다 엄격히 말하면, 수강자들은 확률에 대한 주관적 설명과 조건부 확률을 이해하는 것이 필요하다. 베이즈 규칙은 주어진 새로운 정보에 대하여 관찰자의 조건부 확률을 변화시키는 방법으로 반드시 교육 시켜야만 한다.

베이즈 규칙을 어떻게 가르쳐야 하는가? 비율 또는 평균에 대한 모수공간은 연속적인 값들을 취하기 때문에 표면적으로는 하나의 모수에 대하여 연속적인 값을 갖는 사전분포를 고려해야만 할 것 같다. 이 경우에 사후밀도 함수를 도출해내는 것은 계산적인 문제를 나타내는 해석적 또는 수치적 적분을 포함할 수도 있다. 그러나 베이지안 추론의 기본적 원리는 관심의 대상이 되는 모수 값들에 대하여 작은 집합이 있는 곳에서 좀더 단순한 이산형을 다룸으로써 전달될 수 있다.

본 논문의 목적은 하나의 비율에 대한 추론과 예측을 가르치기 위하여 사용될 수 있는 이산형 모수 값들을 베이즈 규칙에 어떻게 적용하는지를 소개하는 것이다. 2절에서는 비율 p에 대한 것을 알아보기 위하여 이와 관련된 간단한 표 형태에서 베이즈 규칙을 설명한다. 관찰자는 비율에 대한 그럴듯한 값들의 집합을 취하고 이항실험의 결과를 관측한 후에 비율 p 값에 대한 관찰자의 인식(확률)을 개선한다. 미래에 일어날 수 있는 이항실험의 결과를 예측하기 위한 하나의 단순한 도구(메카니즘)인 전확률의 법칙이 제공된다. 3절에서는 하나의 비율을 포함하는 문제에서 베이지안 예측과 베이지안 추론을 설명하는 예제들이 제공된다. 3.1절의 문제는 1994년 미국 프로야구 시즌동안 파업으로 인하여 정상적으로 시즌이 끝나지 않았을 경우에 윌리엄스(Matt Williams)가 쳐낼 수 있었을 홈런 수에 대한 예측 문제이다. 여기서, 가정으로 3명의 야구팬에 대하여 윌리엄스의 홈런 비율에 대하여 사전분포를 세우는 것을 설명하고 흥미로운 예측을 토대로 윌리엄스의 홈런 능력에 대하여 관찰자 의견의 효과를 알아본다. 4절에서는 베이지안 관점에서 기초통계를 교수하는데 경험을 소개하고 5절에서는 고전적인 관점과 베이지안 관점에서 기초적인 추론을 가르칠 때에 이로운 점과 이롭지 못한 점을 종합해 본다.

2. 베이즈 규칙과 전확률의 법칙

베이즈 규칙은 보통 확률론 입문서의 한 장을 차지한다. 그러나 일반적으로 보통은 통계적인 시각에서는 소개되지 않는다. 우리가 비율 p에 대하여 배우는데 관심이 있다고 가정해 보자. p의 실제값은 알려져 있지 않으나 비율에 대한 p의 그럴듯한 값들의 집합을 만들 수

있다고 가정하고 그것을 $p_1, p_2, \cdots, p_k$로 나타내자. 만약 수강자가 그럴듯한 비율값들의 집합을 만들지 못한다면 0에서 1 사이의 값들을 동등한 간격으로 비율값을 정해도 많은 문제를 대할 때 충분하다.

다음으로, 실제값에 좀더 가까울 수도 또는 멀 수도 있다는 관찰자의 믿음을 반영하는 서로 다른 비율 p 값에 대하여 확률을 할당한다. 비율 값들 $p_1, p_2, \cdots, p_k$의 각각에 대하여 이들이 확률 $P(p_1), \cdots, P(p_k)$를 갖는다고 하자. 이렇게 확률을 할당하는 것이 수강자들에게는 일반적으로 어려울 수도 있는데 이러한 이유는 확률의 크기에 대하여 고려해보는 실제경험이 거의 없기 때문이다. 그러나 어떤 수강자는 p에서 '최고의 추측'을 이끌어낼 수 있고 그 다음 가장 일어날 가능성이 높은 값에 대하여 대칭적인 분포를 만든다. 또 다른 수강자는 비율 값들 사이의 판단을 하기가 힘들다는 것을 나타내기 위하여 k개의 비율 p 값에 대하여 동등한 확률을 부여할 수도 있다. 이러한 과정에서 중요한 관점은 수강자가 특별한 응용을 위하여 비율 p의 설명에 대하여 생각해야만 한다는 것이다.

p의 값들에 대하여 좀더 알아보기 위하여 이항실험이 행해진다. 성공을 s, 실패를 f라고 가정하자. 베이즈 규칙에 의하여 p 값에 대한 개선된 확률은 다음의 곱에 비례한다.

$$P(p)\times P(s \text{ 성공}, f \text{ 실패} \mid p)$$

여기서 $P(s \text{ 성공}, f \text{ 실패} \mid p)=p^s(1-p)^f$는 비율값이 모두 p로 동일하다면 성공과 실패의 특별한 수열의 관찰에 대한 확률 또는 우도(likelihood)이다. p의 모든 값들에 대하여 개선된 확률들을 계산하기 위하여 모든 값들에 대하여 위의 곱셈을 계산하고, 이러한 곱셈들의 합을 계산하고, 다음으로 확률들을 얻기 위하여 합으로 각각의 곱셈 값을 나눈다.

추론의 두 번째 유형은 미래의 이항실험의 결과를 예측하는 것이다. 위에서와 같이 k개의 비율값에 대하여 확률들이 있다고 가정하자. 수강자가 미래의 이항실험 $(s+f)$번의 시행에서 s번의 성공을 관측할 확률에 관심 있다고 할 때에는 전확률의 법칙을 사용한다

$$\begin{aligned} P(s \text{ 성공}, f \text{ 실패}) &= \sum_{j=1}^{k} P(s \text{ 성공}, f \text{ 실패 그리고 } p=p_j) \\ &= \sum_{j=1}^{k} P(s \text{ 성공}, f \text{ 실패} \mid p=p_j) P(p_j) \\ &= \sum_{j=1}^{k} \frac{(s+f)!}{s!f!} p_j^s (1-p_j)^f P(p_j) \end{aligned}$$

위의 표현식에서, 확률 $P(p_j)$는 어떠한 데이터도 취해지기 전의 사전확률들을 나타낼 수 있고 또는 데이터가 관측된 후 사후 확률들을 나타낼 수 있다. 주어진 비율값 p에 대하여 s번의 성공 확률은 이항분포식에 의하여 주어진다.

위에 주어진 수식은 s와 f의 임의 값에 대하여 예측확률을 계산하는데 사용될 수 있다. 예측확률들의 집합 $\{P(s \text{ 성공}, f \text{ 실패})\}$은 여러 가지로 사용될 수 있다. 사전분포의 서로 다른 집합에 대한 예측 확률들을 계산함으로써 수강자는 사전확률의 선택을 평가할 수 있다. 모집단의 비율에 대하여 직접적으로 생각하는 것보다 주어진 표본크기에 대하여 앞으로 행해

질 실험에서 성공의 횟수에 대하여 생각하는 것이 더 쉬울 수도 있다. 또한, 사후분포를 기초로 한 예측확률과 사전분포를 기초로 한 예측확률을 비교함으로써 관측된 데이터로부터 알 수 있었던 것이 무엇인가를 볼 수 있다.

이러한 이산형 경우에, 아래의 표 1에서 볼 수 있듯이 베이즈 규칙이 소개된다. 이 예제에서는 비율 p 값을 동등한 간격으로 하였다. 표에서 열은 비율값, 사전확률값, 우도값, 사전확률과 우도값의 곱, 사후확률값들을 나타낸다. 'SUM'으로 나타낸 사전확률과 우도값들 곱의 합은 특별한 s번의 성공과 f번의 실패에 따르는 수열의 사전확률에 기초한 예측확률들이다.

모형(비율값)	사전확률값	우도값	곱의 값	사후확률값
0	p(0)	(0)^s(1-0)^f	p(0) (0)^s(1-0)^f	p(0) (0)^s(1-0)^f /SUM
0.1	p(0.1)	(0.1)^s(1-0.1)^f	p(0.1)(0.1)^s(1-0.1)^f	p(0.1)(0.1)^s(1-0.1)^f/SUM
0.2	p(0.2)	(0.2)^s(1-0.2)^f	p(0.2)(0.2)^s(1-0.2)^f	p(0.2)(0.2)^s(1-0.2)^f/SUM
0.3	p(0.3)	(0.3)^s(1-0.3)^f	p(0.3)(0.3)^s(1-0.3)^f	p(0.3)(0.3)^s(1-0.3)^f/SUM
0.4	p(0.4)	(0.4)^s(1-0.4)^f	p(0.4)(0.4)^s(1-0.4)^f	p(0.4)(0.4)^s(1-0.4)^f/SUM
0.5	p(0.5)	(0.5)^s(1-0.5)^f	p(0.5)(0.5)^s(1-0.5)^f	p(0.5)(0.5)^s(1-0.5)^f/SUM
0.6	p(0.6)	(0.6)^s(1-0.6)^f	p(0.6)(0.6)^s(1-0.6)^f	p(0.6)(0.6)^s(1-0.6)^f/SUM
0.7	p(0.7)	(0.7)^s(1-0.7)^f	p(0.7)(0.7)^s(1-0.7)^f	p(0.7)(0.7)^s(1-0.7)^f/SUM
0.8	p(0.8)	(0.8)^s(1-0.8)^f	p(0.8)(0.8)^s(1-0.8)^f	p(0.8)(0.8)^s(1-0.8)^f/SUM
0.9	p(0.9)	(0.9)^s(1-0.9)^f	p(0.9)(0.9)^s(1-0.9)^f	p(0.9)(0.9)^s(1-0.9)^f/SUM
1	p(1)	(1)^s(1-1)^f	p(1)(1)^s(1-1)^f	p(1)(1)^s(1-1)^f/SUM

SUM

표 1. s번의 성공과 f번의 실패를 갖는 관측된 자료와 0부터 1까지 동등한 간격 상에서 이산 사전확률을 갖는 비율에 대한 사후확률의 계산.

3. 예제

먼저, 베이지안 방법을 이용한 간단한 예제와 베이즈 정리에 대한 요약 및 몇 가지 진술을 살펴보고 1994년 미국프로 야구 홈런 기록경신에 대한 예제를 살펴보기로 하자.

예제 3.1 (불량부품)

8개의 부품이 담겨있는 하나의 상자를 공급자로부터 받았다. 과거 경험으로, 이런 모든 상자가 불량품을 하나도 포함하지 않는 경우가 70%, 하나의 불량부품을 포함하는 경우가 20%, 2개의 불량품을 포함하는 경우가 10%이었다. 따라서 8개의 부품을 포함하는 모든 상자는 0 또는 1 또는 2개의 불량부품을 가질 것이라고 가정한다. 8개의 부품을 포함하는 상자로부터 3개의 부품이 임의로 선택되고 하나의 부품이 불량품으로 발견되었다. 공급자로부터 받은 8개의 부품을 포함하는 상자가 실제적으로 2개의 불량품을 포함하고 있을 확률은 얼마인가?

(풀이)

한 상자로부터 3개의 표본을 추출한 경우 1개의 불량품이 발견되었을 때 그 상자가 실제적으로 2개의 불량품을 포함하고 있을 확률을 계산해보자.

우리가 먼저 인식해야 할 것은 한 상자 안에 있는 부품들에서 표본 크기 n개의 부품을 추출할 때 불량부품의 수 X는 근사적으로 이항분포를 따른다는 것이다(실제적으로, 초기하분포이지만 이러한 정확성은 무시한다). 따라서 이러한 확률은 $P\{X=x|\theta, n\}\equiv f(x|\theta)$ $=\binom{n}{x}\theta^x(1-\theta)^{n-x}$ 이며, 여기서 θ는 불량부품의 확률을 나타낸다.

하나의 예로 $n=3$일 때 하나의 불량부품이 출현하게 되는 경우의 우도함수(likelihood function)는 다음과 같다.

$$f(1|\theta)=\binom{3}{1}(\theta)(1-\theta)^2=3\theta(1-\theta)^2$$

8개의 부품으로 이루어진 한 상자에서 0개 또는 1개 또는 2개의 불량부품이 있을 수 있는 세 가지 가능성이 있다. 또는 θ가 불량부품에 대한 확률을 나타내기 때문에 $\theta=0$, 1/8, 2/8 이다. 과거의 경험을 토대로 사전확률(믿음, prior belief)은 다음과 같다

불량확률: θ	0	0.125	0.25
확률질량함수: $p(\theta)$	0.70	0.20	0.10

더 나아가, 3개의 θ 값에 대하여 우도함수는 다음과 같다

θ	0	0.125	0.25
$f(1|\theta)$	0	0.287	0.42

베이즈 정리에 의하여 다음을 알 수 있다

$$\begin{aligned}
&p\{\text{Box contains 2 defective parts} \mid \text{sample conains 1 defective}\}\\
&=\frac{f(1|0.25)p(0.25)}{f(1|0.25)p(0.25)+f(1|0.125)p(0.125)+f(1|0)p(0)}\\
&=\frac{(0.42)(0.10)}{(0.42)(0.10)+(0.287)(0.20)+(0)(0.70)}\\
&=0.424=42\%
\end{aligned}$$

이 문제에서 박스 안에서 2개의 불량품이 나올 사전 확률은 10%인 반면에 사후 확률은 이것의 약 4배 정도로 나타났다.

베이즈 정리(이산형 모수)(Bayes' theorem(discrete parameter))

$X_1, X_2, \cdots, X_n$을 확률질량함수 $f(X|\theta)$ 또는 밀도함수(르베그 측도를 갖는) $f(X|\theta)$를 갖는 독립적이고 일양적으로 분포된 관측 가능한 확률벡터 변수라고 하자. 즉, f는 확률변수 $\Theta=\theta$라는 조건에서 확률벡터 X에 대한 밀도 또는 질량함수를 나타낸다. 모수공간 Θ는 관측되지 않는다고 가정하고 θ는 Θ의 조건부에서 수치적 값을 나타낸다. Θ는 이산 모수 공간이고 $g(\theta)$는 모수공간의 확률질량함수를 나타낸다. 혼동을 일으키는 경우가 거의 없기 때문에 Θ는 확률변수를, θ는 확률변수의 값을 나타낸다. 이러한 기호 아래서 베이즈 정리는 주어진 $(X_1, \cdots, X_N)$에 대하여 θ의 확률질량함수가 다음과 주어진다.

$$h(\theta|x_1, x_2, \cdots, x_N)=\frac{f(x_1|\theta)\cdots f(x_N|\theta)g(\theta)}{\sum_{\theta} f(x_1|\theta)\cdots f(x_N|\theta)g(\theta)}$$

참고 사항

1) 베이즈 정리를 이용한 θ에 대한 확률질량함수에서 분모에 있는 식은 X_i들에 의존하기 때문에(θ에는 의존하지 않음) $h(\theta|x_1, \cdots, x_N) \propto L(x_1, \cdots, x_N|\theta)g(\theta)$로 다시 나타낼 수 있으며, 여기서 $\propto$는 비례를 나타내고 $L(x_1, \cdots, x_N|\theta)\equiv f(x_1|\theta)\cdots f(x_N|\theta)$는 자료와는 독립적인 주어진 모수 θ에 대한 자료들의 우도함수를 나타낸다. 우도함수는 모수 θ에 대한 함수로 간주할 때 오직 곱의 상수로 유일하다는 것을 알수 있다. 상수는 비례식에서 상수로 흡수되기 때문에 베이즈 정리에는 아무런 차이점이 없다.

2) $g(\theta)$는 모수공간 Θ에 대한 사전확률분포라고 하는데, $g(\theta)$는 현재의 실험에서 자료 X를 관측하기 전에 결정되기 때문이다. 즉, $g(\theta)$는 과거의 경험 또는 인식에 바탕을 두기 때문이다. 예를 들어, 만약 θ가 모집단의 평균치를 나타낸다면, $g(\theta)$는 이러한 평균치와 동일한 모집단을 갖는 과거의 경험에 기초한, 평균치에 대하여 관찰자가 갖는 믿음의 정도(the degree of belief)를 나타낸다. 한 가지 예로, 관측자는 모수공간 Θ의 값들에 대하여 불확실성의 정도를 Θ의 모든 가능한 값들이 동등하게 같은 정도로 발생한다는 분포로 나타낼 수도 있다는 것을 느낄지도 모른다.

3) $h(\theta|x_1, \cdots, x_N)$는 현재의 자료가 주어진 상태에서 θ의 사후확률질량함수라고 부르는데, 이것은 자료가 관측된 후에 또는 자료가 관측된 시점에서 후에 결정되기 때문이다.

4) 베이즈 정리와 동등한 명제(문장): 사후확률$\propto$우도함수$\times$사전확률

5) 베이즈 정리의 설명: 우리가 관측된 현재의 자료를 전혀 가지고 있지 않다면, 과거의 경험으로 θ에 대한 모든 판단을 해야만 한다. 즉, 사전확률질량함수 $g(\theta)$만을 사용한다. 그러나 우리가 과거 경험과 관측된 자료에 기초한 현재의 인식 정도를 모두 가졌다면 베이즈 정리에 의하여 $g(\theta)$를 개선할 수 있고 사후확률질량함수인 $h(\theta|x_1, \cdots, x_N)$을 기초로 $g(\theta)$에 대한 추론의 토대가 된다.

예제 3.2 윌리엄스가 한 시즌동안 홈런 기록을 갱신 할 수 있었을까 ?

1994년 프로야구 시즌은 특히 흥미로웠는데, 상대적으로 다른 시즌에 비하여 홈런의 수가 많았고 특별한 선수가 시즌동안 아주 출중한 타격을 기록했기 때문이다. 몇몇 선수들은 한 시즌동안 61개의 홈런을 초과하여 기록할 이유 있는 기회를 가지고 있는 것으로 나타났다. 또한, 한 선수는 평균타율이 0.400에 가까웠다. 불행히도, 야구시즌은 선수들의 파업으로 인하여 8월 11에 끝났고 팬들은 162게임 전부가 시즌동안 치루어졌다면 특별한 타율 기록에 어떤 일인가 일어났을 것이라는 아쉬움을 남겼다.

특히, 1994년 8월 11일자로 처음 445번의 타격에서 43개의 홈런을 기록했던 윌리엄스라는 선수를 생각해 보자. 야구 파업이 발생하지 않았었다면, 부상이 없고 남은 시즌동안 199번의 타격이 더 있었다고 가정하자(199번의 타격이 있었을 수 있던 것은 Cramer와 Dewan(1995)가 1994년 시즌을 시뮬레이션을 통하여 산출한 남은 시즌에서의 타격수를 사용하였다). 전체 61개의 홈런보다 더 많은 홈런을 쳐낼 수 있는 합리적인 확률과 홈런 기록을 갱신할 수 있는 합리적인 확률을 가지고 있는가? 즉, 그의 마지막 199번의 타격에서 적어도 19개의 홈런을 쳐낼 수 있었을 것인가?

이 질문에 대한 답은 시즌의 마지막 기간동안 선수가 홈런을 칠 수 있는 능력에 대한 개개인의 생각에 달려있다. 가정해 볼 수 있는 한 가지 경우는 전반부 시즌동안 기록했던 홈런 비율과 남아 있는 시즌동안 기록할 수 있는 홈런 비율이 유사할 것이라는 가정과 또는 그 선수는 처음 445타격에서 일반적으로 말하는 '최고조'의 상태였고 남은 시즌동안은 그 열기가 점점 식어갈 수 있다는 가정을 할 수도 있다.

여기서 남은 시즌동안 개별적인 타격을 한번 타격에 출장하는 동안 그 선수가 홈런을 쳐낼 수 있는 확률 p를 모수로 갖는 베르누이 시행으로 가정하자. 홈런의 비율 모수를 설명하는 p는 무엇인가? 이것은 남은 시즌동안 동일한 조건에서 선수가 많은 타격을 할 수 있다는 것이 전제된다면 그 선수의 홈런 비율이다. 가르치는데 있어서, 이러한 확률 값을 남은 시즌동안 기록한 홈런의 표본비율 $\hat{p}$와 구별하는 것이 중요하다. 또한, p의 값은 남은 시즌동안 선수가 홈런을 기록할 수 있는 능력을 나타낸다. 한 시즌 진행동안 홈런을 기록할 확률이 변할 수도 있으며, 또한 그의 경력에 비추어 홈런을 기록할 확률이 변할 수도 있다. 1994년 시즌 말에 그 선수가 홈런을 기록할 확률은 전년도 시즌동안 홈런을 기록한 확률보다는 달라졌을 수도 있다.

베르누이 모수 p에 대한 의미를 이해한 후에 1994년 시즌의 남은 기간동안 그 선수가 홈런을 기록할 수 있는 능력에 대하여 개개인의 믿음을 반영하는 그럴듯한 p 값을 토대로 하나의 확률분포를 만든다. 그 선수의 능력에 대한 개개인의 믿음이 다양하기 때문에 여기서는 3명의 야구팬 알랜, 보브, 샐리를 가정하여 문제를 고려하자. 아래에서, 우리는 그 선수의 홈런 기록 능력에 대한 3명의 야구팬의 의견을 논하고, 3명의 각 팬에 대하여 개개인의 의사와 대응되는 확률분포를 어떻게 만드는지를 설명한다.

첫 번째 야구팬인 알랜은 1994년 남은 시즌동안 윌리엄스의 홈런 기록 능력은 1994년 시

즌 초반동안 보여주었던 기록대로 최고로 측정된다고 믿는다. 사실, 그는 윌리엄스가 홈런을 기록할 확률 p는 전체 시즌을 통하여 같을 것이라고 생각한다. 덧붙여서, 그는 지난해에 홈런을 기록했던 능력이 1994년 시즌동안 홈런을 기록할 능력에 대하여 알아보는 것에 무관하다고 믿는다. 다른 말로, 알랜은 1993년과 그 이전의 홈런 기록 확률은 1994년 홈런을 기록할 확률과는 다르다고 믿는다. 이것은 야구시즌이 진행되지 않는 동안 새로운 타격 폼, 발의 벌림 정도, 또는 그 외의 부가적인 체력 훈련 등의 여러 요인에 기인한다.

알랜은 윌리엄스의 홈런확률 p는 1994년 전체 시즌을 통틀어 일정하게 유지된다고 믿기 때문에 사전 확률분포를 만들기 위하여 시즌 전반부에서 관측된 타자의 홈런 데이터를 사용할 것이다. 알랜은 비율 p의 값에 대하여 거의 알고 있지 못하므로 가능한 홈런의 비율 집합을 {0.01, 0.02. 0.03, ···, 0.20}으로 하고 이러한 집합의 각 값에 대하여 동일한 확률을 할당한다. 2절에서 소개한 방법을 이용하여 그는 1994년의 데이터-45번의 성공과 402번의 실패-를 가지고 그가 고려했던 확률들을 개선한다. 여기서 성공이라 함은 홈런을 기록할 것으로 정의한다. 3명의 야구팬에 대한 개선된 확률들이 표 2에 주어져 있다.

P	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10	0.11	0.12	0.13	0.14
알랜	0	0	0	0	0	0	0.03	0.13	0.25	0.28	0.19	0.08	0.03	0.01
보브	0	0	0	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0	0	0	0
샐리	0.06	0.06	0.06	0.12	0.19	0.19	0.12	0.06	0.06	0.06	0	0	0	0

표 2. 윌리엄스의 홈런 비율에 대한 3명 팬의 사전확률분포

두 번째 팬인 보브는 1994년 남은 시즌동안 윌리엄스가 홈런을 기록할 능력에 대하여 다른 생각을 가지고 있다. 그는 p 값을 알아보기 위하여 전년도 메이저리그 시즌동안에 윌리엄스의 타격수행 능력을 생각한다. 그래서 보브는 지난 5개년 시즌동안 홈런 통계를 조사한다. 표 3은 타격 수, 홈런 수, 관측된 홈런비율을 나타내고 있다. 2절에서 사용한 방법으로 각 시즌에 대하여 대응하는 시즌 동안의 홈런 확률을 알 수 있다. 우리는 홈런 확률에 대한 사전 확률들을 {0.01, 0.02, 0.03, ···, 0.20}의 눈금 상에서 균등하다고 가정할 수 있고 해당 시즌으로부터 관측된 홈런 데이터를 이용하여 사후확률분포를 얻는다. 표 3에서 열 '실제비율'은 적어도 0.01의 사후확률을 받을 수 있는 해당시즌의 홈런 확률 값들의 리스트이다. 1989년에, 윌리엄스는 292번의 타격에서 18개의 홈런을 기록하였고 관측된 홈런 비율은 0.062이다. 금번 시즌에서 타격 데이터는 0.04, 0.05, 0.06, 0.07, 0.08, 0.09에 대한 홈런 확률들이 일정하다.

표 3에 종합되어진 계산을 살펴본 후 보브는 1989-1993년과 1994년 시즌 전반부에 대하여 윌리엄스의 홈런 확률에 대한 몇 가지 생각을 갖게 된다. 그는 지난 몇 년간의 홈런 데이터 모두를 공동 합산하는 것이 바람직하다고 생각하지 않는다. 왜냐하면 1989부터 1994년을 통하여 윌리엄스의 홈런 기록 능력이 변했기 때문이다. 그러나 1994년 남은 시즌동안에

연도	홈런수	타격수	관측비율	실제비율
1989	18	292	0.062	{0.04, 0.05, 0.06, 0.07, 0.08, 0.09}
1990	33	617	0.053	{0.04, 0.05, 0.06, 0.07, 0.08}
1991	34	589	0.058	{0.04, 0.05, 0.06, 0.07, 0.08}
1992	20	529	0.038	{0.02, 0.03, 0.04, 0.05, 0.06}
1993	38	579	0.066	{0.05, 0.06, 0.07, 0.08, 0.09}
1994	43	445	0.097	{0.07, 0.08, 0.09, 0.10, 0.11, 0.12, 0.13}

표 3. 메이저리그에서 윌리엄스의 홈런 통계

대하여 홈런확률 p의 값은 전년도 시즌동안으로부터 그럴듯한 홈런확률 값 사이라고 생각한다. 몇 가지 상황을 고려한 후 홈런확률 값 p는 집합 {0.04, 0.05, ⋯, 0.10} 에 포함된다고 생각한다. 보브는 1992년 윌리엄스의 홈런 확률은 0.02와 같이 저조했을 수도 있었지만 1994년 남은 시즌동안에 1992년보다 더 좋은 타격을 보여줄 것이라고 생각한다. 1994년에 관측된 데이터로부터 윌리엄스의 홈런 확률은 0.13과 같이 높을 수도 있다는 것이 가능하다. 그러나 보브는 이러한 높은 확률은 지난 시즌동안의 확률 값들과 너무나 크게 향상되었다고 생각한다. 그래서 그는 1994년 시즌 마지막에 대한 윌리엄스의 홈런 확률을 0.10이 상계가 되는 값을 취한다. 보브는 p에 대한 특별한 값을 지정하는 것이 어렵기 때문에 집합 {0.04, 0.05, ⋯, 0.10} 내에 있는 각 값에 동일한 사전확률을 할당하기로 했다.

윌리엄스의 홈런에 대한 샐리의 생각은 앞서의 두 팬의 생각과는 다르다. 그녀는 1994년 전반부에서 윌리엄스가 기록한 홈런에 특히 감동을 받았었다. 그러나 그녀는 시즌 전반부 동안 윌리엄스는 '최고의 상태'였었다고 생각하고 남은 시즌동안 이러한 최고의 상태를 유지해 나가는 것이 힘들다고 생각했을 것이다. 또한, 남은 시즌동안 윌리엄스가 부수적인 많은 압박을 받았다고 생각했을 것이다. 한 시즌동안 홈런기록이 61개인 것은 30년 전에 세워진 것이고 윌리엄스가 그 기록에 접근한다면 집중적인 매체의 취재를 받게 될 것이다. 이러한 부수적인 매체의 압박은 그 시점에서 시즌동안 윌리엄스의 타격 능력에 역효과를 줄 수도 있다. (이러한 믿음에 대한 대안적 설명은 회귀효과라는 것으로 잘 알려져 있다. 이러한 배경에서, 회귀효과는 한 시즌의 전반부에서 최대의 타격수행은 그 시즌의 후반부 동안 평균 이상의 타격 수행능력을 보여줄 수 있다는 경향을 갖는 선수들에게 나타나는 현상이다.)

샐리는 p에 대한 사전분포를 어떻게 구축할 것인가? 전년도에 기록한 윌리엄스의 홈런기록을 살펴본 후 p에 대한 가능한 값들을 0.01부터 0.10까지라고 생각한다. 1994년 시즌의 전반부에서 보여주었던 홈런비율 0.097보다는 급격히 떨어질 것이라고 느낀다. 그녀는 전년 시즌동안 윌리엄스가 이루었던 것과 동일하다는 0.05와 0.06의 값 상에 가장 큰 확률을 할당한다. 비록 0.05와 0.06이 가장 그럴듯한 p의 값이지만 샐리는 윌리엄스가 홈런 기록을 경신하고 특별한 일이 일어날 것이라는 데에는 작은 확률이 있다고 생각한다. 또한, 윌리엄스가 나머지 시즌동안 특별한 슬럼프와 아주 적은 홈런을 보여줄 것이라는 데에도 작은 확률을 할당한다. 그래서 그녀는 비율 p 위에 가장 그럴듯한 값으로부터 천천히 감소하는 방향으로 이동하면서 확률들을 할당한다. 가장 극단적인 값 0.01과 0.1 상에 대한 확률은 사전확률로

0.06을 할당한다.

윌리엄스가 홈런 기록을 경신할 것이라는 예측에 대하여 윌리엄스 홈런 비율 p에 대한 서로 다른 믿음이 내포하고 있는 것은 무엇일까? 2절의 기본 수식을 이용하여 각 3명의 야구팬에 대한 홈런 수 y에 대한 예측 확률을 계산할 수 있다. 이러한 확률분포에 대한 점도가 그림 1에 보여진다. 첫 번째로 알아야 할 것은 나머지 시즌동안 윌리엄스가 기록할 홈런의 수에서 각 사전확률에 대하여 상당한 변동이 있다는 것이다. 예를 들자면, 알랜의 사전확률이 사용되었다면, 윌리엄스는 10-30개 사이의 홈런을 더 기록할 수 있을 것이다. 두 번째는, 이러한 세 명의 예측 분포는 위치와 퍼짐에 반하여 상당히 변하고, 이것은 윌리엄스의 홈런 비율에 대한 개개인의 사전 확률들이 나머지 시즌동안 윌리엄스의 홈런 기록에 대한 개개인의 예측에 큰 효과를 가질 수 있다는 것을 말한다. 특히, 우리는 윌리엄스가 홈런 기록을 경신할 확률, 즉 홈런의 수가 19개 이상이 되는 확률을 계산하는 데에 관심이 있다. 이 확률은 개개인의 예측 확률분포로부터 쉽게 계산된다. y가 19 이상이 되는 모든 값에 대한 예측확률들을 더한다. 알랜, 보브, 샐리에 대응하는 3명의 사전확률 분포에 대하여 이러한 확률은 각각 0.571, 0.205, 0.099로 주어진다. 이러한 3명의 야구팬은 윌리엄스가 홈런 기록을 경신할 것이라는 데에 매우 다른 확률들을 갖는다.

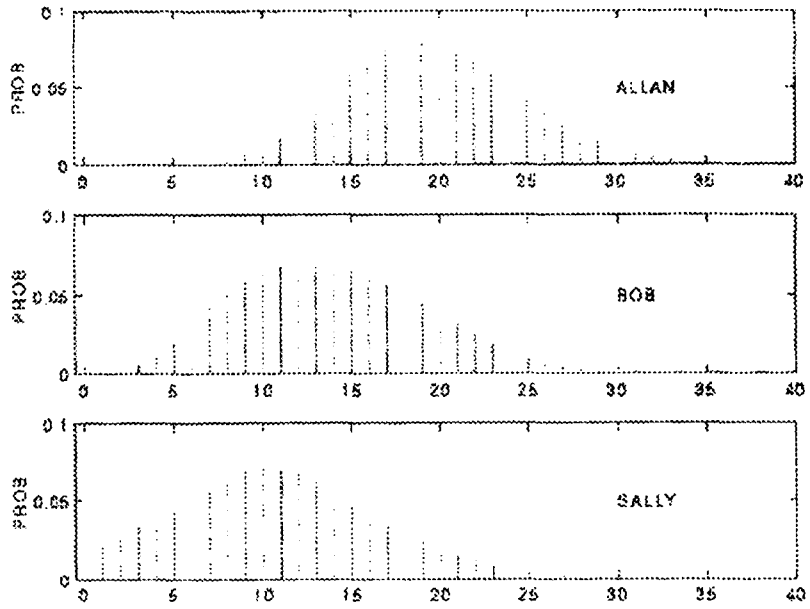

그림 1. 알랜, 보브, 샐리 등 3명 팬의 나머지 시즌에서 홈런 수에 대한 예측확률

위의 계산을 살펴본 야구팬들은 누가 그것을 믿을까 하는 의구심을 가질지도 모른다. 위의 분석을 만족하게 하는 것이 가능한가? 대답은 '아니다'이다. 이 예제는 가정에 대한 통계

적 결론에 대한 민감성을 소개한 것이다. 윌리엄스의 홈런 기록을 예측하기 위하여, 우리는 반드시 몇 가지 가정을 해야만 한다. 이러한 가정이란 그가 남은 시즌동안 홈런을 기록할 능력(확률 p에 의하여 측정된)이 전년도와 1994년 시즌 초반부에서 보여주었던 것과 어떠한 관계가 있는지에 대한 가정이다. 이러한 가정들은 마지막 결론에 결정적으로 영향을 주기 때문에 개개인의 가정에 대한 생각들이 주의 깊게 이루어져야 한다. 개인적으로, 비록 1994년에 윌리엄스의 홈런에 감명을 받았지만, 그 시즌의 홈런기록이 달성하기 위한 매우 어려운 장애라고 관망하며, 그래서 홈런 수 y가 19개 이상이 될 것이라는 사상에 상대적으로 작은 확률을 줄 것이다. 이러한 믿음에 대한 함축적 의미는 나는 나머지 시즌동안 윌리엄스의 홈런 확률 p가 적을 것이고 보브나 샐리가 할당했던 확률과 유사한 확률을 할당할 것이다. 비슷하게, 또 다른 야구팬들은 윌리엄스가 홈런 기록을 경신할 것이라는 개개인의 예측을 만들기 위하여 윌리엄스의 홈런 능력에 대한 개개인의 믿음에 대하여 생각해야만 한다.

4. 베이지안 관점에서 기초통계 교수법

앞 절에서 소개한 베이지안 방법은 주당 3시간의 기초통계 수업에서 이루어질 수 있다. 이러한 기초통계 수업 과정은 통계전공이 아닌 일반 수강자들을 위한 한 학기 과정 프로그램이다. 이 과정의 일반적인 목적은 실생활에서 통계적 타당성에 대한 사용법을 학생들에게 소개하는 것이며 수강자들은 t-검정과 같은 특별한 통계적 방법을 배울 필요가 없기 때문에 가르치는 사람은 방법에 있어서 선택의 자유를 가질 수 있다.

이러한 수업을 진행하는데 있어서 교재는 베이지안 관점에서 기초 통계학을 다룬 것이라면 어떠한 것이라도 관계는 없을 것으로 생각 하지만 1995년 Berry 교과서를 이용하여 가르친다면 좀더 효과적일 것이다(국내에 출판된 서적으로 베이지안 관점에서 기초통계학을 다룬 것은 현재까지 없는 것으로 알고 있음). 주요 주제들은 자료분석, 자료수집, 확률, 비율에 대한 추정. 단순회귀 등이다. 베이즈 규칙은 모수에 대한 추론을 하는데 사용된다. 이 과정에서는 표본분포와 구간추정 및 검정에 대한 상대도수 설명이 논의될 필요가 없다.

주당 3시간 수업에서 2시간은 강의를 하고 1시간은 컴퓨터 실습으로 이루어질 수 있으며 실습시간에 수강자들은 추론에 대한 몇 가지 문제를 받는다. 전형적인 문제에서 수강자들은 미지의 모수에 대한 사전확률분포를 세울 것을 요구받게 된다. 자료가 주어진다면 미니탭을 이용하여 사후확률을 계산하게 된다.

실험계획을 위하여 수강자들을 몇 개의 소그룹으로 나누어 각 그룹에서 관심의 대상이 되는 자료의 대상을 선택하고 전화통화나 설문지를 통하여 실제자료를 수집하며 이를 근거로 하여 각 그룹에서 가정하였던 사전확률을 개선하게 된다. 그리고 수강자들이 프로젝트를 통하여 배울 수 있었던 결과들을 종합하게 된다.

이러한 프로젝트 수업은 가장 효과적인 것이라고 할 수 있으며 수강자들은 과학적인 방법과 관심을 가졌던 문제에 대한 작업을 자신이 직접 행함으로써 그 과정이 좀더 유익하다는

것을 알게 된다. 사전분포를 구축하는 것은 자료조사를 통하여 배우고자 하는 것이 무엇인가를 알게 하는데 도움이 될 것이다. 수강자들은 어떤 면에서 사전분포를 설명하는 것에 대하여는 보수적인 경향이 있으며 사후확률에 대해서는 관측된 자료에 의하여 지배되는 경향이 있다는 것은 흥미로운 일이다.

5. 맺는 말

이 논문은 이항추정과 예측을 가르치는데 있어서 이산 사전분포를 가지고 베이지안 추론에 대한 것을 소개하였다. 여기서는 비록 하나의 비율문제에 국한하였지만 이러한 접근 방법은 다양한 추정 문제에 적용될 수 있을 것이다. 이러한 다양한 접근방법을 소개한 것으로는 Berry(1995)와 Albert(1996)를 들 수 있으며 많은 도움이 될 것이다.

베이지안과 고전적인 접근 방법을 비교함에 있어서 처음 통계수업에 대한 기본적인 목적이 무엇인가를 상기하는 것은 아주 중요한 일이다. 많은 통계 수업들이 생물학, 경영학, 교육학 등 특별한 과목을 위하여 설강되는데 이러한 수업에 있어서는 수강자들이 그들 과목을 이해하는데 특별한 방법이나 그와 대응되는 고전적 설명이 가르치는 목적이 될 수도 있기 때문에 고전적 접근 방법이 필요할 수도 있다.

또 다른 기초통계 수업은 특별한 과목을 이해하기 위하여 설강되지 않은 경우가 있으며 이러한 수업의 목적은 과학적 방법의 도구로써 통계학의 사용법을 소개하는 것이 목적이다. 본 논문은 전공교양 과목 수업에서 기본적으로 뒤따르는 통계적 추론의 기본 내용을 가르침에 있어서 고전적인 관점에서보다 베이지안 관점에서 가르치는 것이 더 나을 수도 있다는 것이다. 요점은 추론에 대한 고전적 접근 방법이 기본적으로 결함이 있다는 것은 아니다. 즉, 베이지안 접근 방법을 선호한다는 것이다. 추론의 교수법에 대하여 고전적 접근 방법과 베이지안 접근 방법에 대한 장점과 단점을 요약해 보았다.

고전적 접근 방법의 주요 장점은 유사성이며 이러한 방법은 이미 잘 알려져 있고 도수의 개념이 구간추정이나 가설검정에 공통으로 사용된다는 것이다. 고전적 방법은 일반적으로 호감이 가는 방법의 대상으로 간주된다. 그러나 표본분포에 대한 개념을 설명하는 것이 어렵기 때문에 신뢰에 대한 정확한 설명을 전달하기에 어려움이 뒤따르는 것이다. 또한, 개념보다는 추론방법에 초점을 맞춘다는 것이다. 방법에 역점을 두다보면 반복적인 표본추출에 기초하여 신뢰에 대한 기본개념을 가르치는데 어려움이 뒤따를 것이다. 개념보다는 방법을 가르치는 것이 훨씬 쉬운 것은 확실하다.

이산형 베이지안 접근 방법의 주요 특징은 사후확률분포와 예측확률분포를 계산함에 있어서 그 과정이 단순하다는 것이다. 이러한 계산이 많은 모형을 갖는 경우에는 따분할 수도 있지만 프로그램을 이용하면 자동적으로 계산이 가능하다. Albert(1996)가 사용한 미니탭의 매크로를 이용하면 아주 쉽게 문제를 해결할 수 있다. 이러한 매크로에는 'p_disc'와 'p_disc_p'가 있으며 이산형 사전확률분포를 이용하여 사후확률과 예측확률을 계산할 수 있

다.

기초통계 수업에서 베이지안 방법을 소개하는 데에는 몇 가지 기본적인 것이 있다. 첫째로, 사전확률분포를 소개함과 동시에 추론과정 속에 주관적 요소를 소개하는 것인데 학생들이 주관적 확률을 현명하게 특징지우는 것을 훈련시키기에는 어려울 수도 있다. 그러나 이러한 훈련 과정은 추론이라는 것이 결정적으로 특정한 가정에 의존한다는 중요한 내용을 가르치는 것이다. 둘째로, 베이지안 방법을 사용한다면 표본분포를 가르치는 대신 조건부 확률과 베이즈 규칙을 가르치는 것이다. 조건부 확률이 어려운 주제이지만 간단히 이해시킬 수 있다.

참고 문헌

1. Albert, J., *Bayesian Compuration Using Minitab*, Wadsworth, Belmont, 1966.
2. Albert, J., "Teaching Inference about Proportions Using Bayes and Discrete Models," *Jouranal of Statistical Education*, V.3, n.3, 1995.
3. Berger, J.O., *Statistical Decision Theory and Bayesian Analysis*, Springer, Berlin, 1985.
4. Berry, D.A., *Basic Statistics: A Bayesian Perspective*, Wadsworth, Belmont, 1995.
5. Press, S. James, *Bayesian Statistics:Princiole, Models, Application*, Wiley, 1989.