The Korean Journal of Applied Statistics (2014) DOI: http://dx.doi.org/10.5351/KJAS.2014.27.6.855
27(6), 855–863

History and Future of Bayesian Statistics

Jaeyong Lee[a] · Kyoungjae Lee[a,1] · Youngseon Lee[a]

[a]Department of Statistics, Seoul National University

(Received October 27, 2014; Revised December 1, 2014; Accepted December 5, 2014)

Abstract

The recent computational revolution of Bayesian statistics has expanded use of the Bayesian statistics significantly; however, Bayesian statistics face a new set of challenges in the era of information technology. We survey the history of Bayesian statistics briefly and its expansion in the modern times. We then take a prospective future view of statistics and list challenges that the statistics community faces.

Keywords: Bayesian statistics, Thomas Bayes, the future of statistics.

1. 서론

Bayes는 1761년 4월 17일에 사망하였다. 태어난 날은 확실하지 않지만 묘비에 59세의 나이로 사망했다고 적혀있어서 1701년이나 1702년일 것으로 추측된다. Bayes 사망 후, 1763년에 그의 친구인 Price는 Bayes가 Hume의 책에 대한 반발로 썼던 논문을 발표했다 (Bayes, 1763). 이 논문에서 Bayes는 이항분포의 확률에 대한 추론을 논의하고 논의 과정에서 베이즈 정리를 이끌어냈다.

2014년은 Bayes가 태어난지 312년 내지는 313년이 되고, Bayes의 논문이 발표된지 251주년이 되는 해이다. 2013년은 베이즈 정리가 발표된지 250년이 되는 해로 많은 행사들이 열렸다. 국제 베이지안 학회(International Society for Bayesian Analysis, ISBA)에서는 *Bayes 250*이라는 이름으로 듀크대학(Duke University)에서 학회를 개최했고 베이지안의 250년 역사와 미래를 살펴보았다. 응용통계연구에서도 250년 베이지안 통계의 역사를 살펴보고 미래를 조망해보는 특집호를 기획한다는 얘기를 편집장님께 듣고 매우 기쁘게 생각하였다. 이번 특집호를 계기로 한국에서 베이지안 통계의 연구와 이에 대한 관심이 더 고조되기를 바라는 마음이다.

베이지안 통계는 20세기 초반의 암흑기를 지나 20세기 후반부터 21세기가 들어서면서 점점 그 영향력을 더해가고 있다. 통계저널 뿐만 아니라 의학, 생물학, 기상학 등 여러 응용분야에서도 베이지안 통계학의 영향력은 더 커지고 있다. 반면에 통계학 전반에 새로운 도전도 생기고 있다. 자료의 대용량화와 새로운 형태의 자료의 등장, 통계로 풀고자하는 문제의 복잡성, 통계학의 보편화, 경쟁분야와 경쟁직업의 대두 등이다. 베이지안 통계학은 통계학에 대두되는 도전을 같이 풀어 나가야한다. 우리는 이 논문에서 베이지안 통계학의 과거를 살펴보고, 이를 바탕으로 미래를 조망해보는 기회를 가지려 한다.

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MEST)(No. 2011-0030811).

[1]Corresponding author: Department of Statistics, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 151-747, Korea. E-mail: leekjstat@gmail.com

이후의 논문 구성은 다음과 같다. 2절에서는 베이지안 통계의 역사를 살펴본다. 3절에서는 베이지안 통계학의 현재 위치에 대해 알아본다. 4절에서는 현재에 나타나고 있는 현상을 기반으로 미래를 조망해 본다. 그리고 5절은 결론으로 이루어져 있다.

2. 베이지안 통계의 역사

1748년에 Hume은 *인간 이해에 대한 질문(Philosophical essays concerning human understanding)*이라는 제목의 책을 출간하고, 기독교의 근본적인 믿음에 대한 공격을 했다 (Hume, 1748). Hume은 모든 것은 경험에 의해서만 배울 수 있다고 믿었으며, 전통적인 믿음, 도덕, 인과론을 의심하였다. 당시에 신은 첫 번째 원인(First cause)으로 여겨졌기 때문에 Hume의 주장은 큰 파장을 몰고 왔다. Hume은 세계의 정교한 디자인이 창조자의 존재를 증명하지 않는다고 주장하였다. Hume의 책은 수학적이거나 과학적이지는 않았지만, 수학자나 과학자들 사이에 깊은 파장을 야기했다. 왜냐하면, 당시의 수학자와 과학자들은 자연법칙들의 존재가 첫 번째 원인, 즉 신의 존재를 증명한다고 믿었기 때문이다.

당시 턴브리지웰스(Turnbridge Wells)의 목사로 있었던 Bayes는 Hume의 책에 대한 반응으로 역확률(Inverse Probability), 즉 어떤 사건의 원인에 대한 확률에 대해 관심을 갖고, 결과로부터 원인을 밝혀낼 수 있다는 것을 증명하려고 하였다. Bayes의 동기는 신의 존재를 수학으로 증명하고자 했던 것으로 보인다 (McGrayne, 2011). Bayes는 사후에 발표된 논문에서 이항분포의 모수에 대해 추론하는 문제를 다루고 있었다 (Bayes, 1763). 이 논문에는 균등분포를 사전분포로 사용하는 것에 대한 논리적 정당성을 관측치의 주변분포를 구해서 보이는 과정이 나와있었고, 유명한 베이즈 정리가 실려있었다.

모형을 이용한 자료적합을 처음 시작한, 현대적인 의미의 첫 번째 통계학자라고 여겨지는 프랑스의 수학자 Laplace는 1774년 그의 논문에서 이항분포의 모수를 추론하는 문제에 대해서 베이즈 정리를 재발견하였다 (Laplace, 1774). 이후 이 정리를 의학, 천문학, 법학 등의 다양한 응용분야에 적용했다. 또한 그는 현대 통계학에서의 켤레사전분포(Conjugate prior) 개념과 번스타인-폰 미시스 정리(Bernstein – von Mises Theorem), 즉, *자료가 많아질 때 서로 다른 사전분포를 이용하더라도 사후분포는 합의를 향해 간다*, 와 같은 개념에 대해서 소개하였다. Laplace는 불충분 이유의 원리(Principle of insufficient reason 혹은 Principle of indifference)를 이용하여 균등분포를 사전분포로 사용하는 것에 대한 논리적 근거를 마련하였다. Laplace의 베이지안 통계학에 대한 접근법은 이후 베이지안들에게 뿌리깊은 영향을 주어 오랫동안 균등분포만이 추론을 위한 사전분포로 이용되기도 하였다.

1920년 이후, 역확률의 문제들은 Fisher, Neyman, 그리고 Pearson이 새로운 패러다임의 통계학을 개발하면서 사람들의 관심으로부터 완전히 벗어났다. 통계적 문제들은 이들이 개발한 방법론과 접근방식으로 대체되었다. 이 방법들은 빈도론적 통계(혹은 빈도통계; Frequentist statistics)를 기반으로 한 것이었다. 확률의 빈도론적 해석은 동일한 실험이 무한번 반복되는 것을 상정하고, 확률을 사건의 빈도로 정의하는 것이다. 빈도론적 통계는 확률의 빈도론적 해석을 바탕으로 통계적 추론을 하는 통계적 방법을 의미한다. 빈도통계학이 주류로 자리잡은 이 시기에는, 베이지안의 관점의 통계학은 주류 통계학계의 관심에서 벗어나있었고 빈도론자들에 의해 전적으로 부정되었다.

그럼에도 불구하고, 베이지안 방법은 세계 2차대전 시기에 Turing 등에 의해 암호를 해독하는 등 실질적인 문제의 해결에 이용되기도 하였다. 그러나 이러한 사실이 대중에게 알려질 기회는 매우 적었다 (McGrayne, 2011).

통계학계에서 빈도통계학을 기반으로한 방법론들이 우세했기 때문에 베이지안 통계는 암흑기를 겪고 있었다. 그러나 이러한 암흑기 속에서도 Laplace의 아이디어는 조금씩 발전해 나가고 있었다. 베이지안 통계는 2개의 서로 다른 방향으로 발전되어 갔는데, 이들을 각각 객관적 베이지안(Objective

Bayesian)과 주관적 베이지안(Subjective Bayesian)이라고 부른다. 주관적인 베이지안과 객관적인 베이지안은 모두 확률이라는 것을 믿음의 정도로 표현한다는데 공통점이 있다. 그러나 객관적 베이지안들은 추론에 있어서 개인적인 기여를 최소화하기를 원했다. 사전분포는 합리적 믿음의 정도를 나타내고 사후분포는 자료로부터 발생하는 불확실성을 의미하는 것이라 주장했다. 반면 주관적 베이지안은 사전분포를 주관적 믿음으로 생각하며, 사후분포는 자료를 통해서 변하는 믿음의 정도로 생각했다.

주관적 베이지안은 객관적 베이지안보다 먼저 나타났다. 1920년대 초, Keynes는 논리학을 확률의 영역으로 확장하고자 하였다. Keynes는 이성적인 사람은 확률의 값을 선험적으로 알 수 있다고 생각했다. 이에 대해, Ramsey는 Keynes의 이러한 생각을 비판하고 모든 이성적인 사람들이 자기자신의 고유한 주관적인 확률값을 갖는 것이 자연스럽다고 주장하였다 (Gilles, 2000).

이와는 독립적으로, 1930년 경 이탈리아의 De Finetti는 교환가능성(Exchangeability)의 개념을 소개하고 사전분포의 역할을 수학적으로 설명하면서 주관적 확률을 기존과는 전혀 다른 방식으로 정의했다 (De Finetti, 1938). 그러나 De Finetti에 의해 제안된 확률의 개념은 이후 Savage에 의해서 다시 차용되기 전까지 약 20여년 동안 사장되어 있었다.

주관적 베이지안에서의 확률은 Savage에 의해서 점차 발전하게 되었다. 그는 베이지안의 확률을 공리를 바탕으로 하는 완성된 체계로 만들기를 원했다. 이 기간은 Savage가 De Finetti의 연구를 발견한 시기이기도 했다. Savage는 확률을 수학적으로 증명하는 작업을 계속하면서 베이지안 방법론의 논리적인 기초를 닦았으며, 베이지안 통계에 대한 이론을 더욱 견고하게 만들었다 (McGrayne, 2011; Savage, 1954).

반면, 객관적 베이지안 통계는 영국의 Jeffreys에 의해서 발전되었다. Jeffreys는 1919년의 논문에서, '합리적인 믿음의 정도(Degree of reasonable belief)'로 확률을 정의했다 (Wrinch와 Jeffreys, 1919). 이는 주관적 확률을 정의한 Ramsay와 한편으로는 비슷하지만 '주관적'이 아닌 '합리적' 믿음을 다룬다는데서 객관적 베이지안의 시초가 되었다. Jeffreys는 1939년 출간된 그의 책에서 객관적 베이지안의 확률 이론을 위한 공리적 체계를 구축하는 것에 노력을 기울였다 (Jeffreys, 1939). Jeffreys는 객관적인 사전분포를 유도하기 위해 피셔정보를 이용했다. 그의 책에서 다루었던 예제들은 현재까지 객관적 베이지안 통계 분야에서 널리 인용되는 모범적인 예가 되고있다.

이후 객관적 베이지안은 다양한 접근방식을 통해 독자적으로 발전하며 퍼져나갔다. Jaynes는 사전분포를 구성하는데 최대 엔트로피(Maximum Entropy)라는 개념을 사용했고, 이는 이후 객관적 베이지안의 방법론, 특히 이산적 문제(Discrete problem)를 다루는데 중요한 역할을 하게 되었다. Lindley는 1965년 그의 저서를 통해 객관적 베이지안 통계를 널리 알리는데 일조했다 (Lindley, 1965). 1979년 Bernardo는 레퍼런스 분석(Reference analysis)이라는 방법을 소개했는데 이는 객관적 베이지안 분석을 위해 일반적으로 적용가능한 체계로 알려져있다 (Bernardo, 1979).

베이지안의 이론적 발전에도 불구하고, 그 당시에는 베이지안 방법론을 이용하기 위해서 복잡한 계산작업을 수행했어야 했기 때문에 실제로 이를 이용해서 현실적인 문제들을 다루기는 매우 어려웠다. 그러나 베이지안의 논리정연함은 많은 사람들에게 매력적이었으며, 따라서 베이지안 통계를 현실에서 적용하기 위한 노력은 계속되었다. 하버드 경영대학원의 Raiffa와 Schlaifer도 그 중 하나였다 (McGrayne, 2011). 이들은 불확실성 하에서의 의사결정에 대해 연구하면서 경영학에 있어서의 중요한 문제들이 빈도통계학적 접근방법으로는 해결할 수 없다고 생각했다. 주관적/객관적 베이지안의 이론과는 독자적으로 그들은 베이지안 결정이론이라는 것을 고안했다. 이러한 연구를 통해 베이지안의 실생활에서의 응용은 점차 관심을 받게 되었다.

1990년에는 베이지안 계산에 있어 혁명적인 발전이 이루어졌다. Gelfand와 Smith는 깁스 샘플링

(Gibbs Sampling) 방법이 사후분포에서 샘플을 추출하는데 이용될 수 있다는 것을 보였다 (Gelfand와 Smith, 1990). 이는 현대 베이지안 통계 계산 방법의 핵심인 마코프 체인 몬테 카를로(Markov chain Monte Carlo, MCMC)방법의 시작이 되었다. 마코프 체인 몬테 카를로 방법의 발명으로 베이지안들은 빈도론자들 보다 더 많고 복잡한 문제들을 쉽게 풀 수 있게 되었다. 이는 이후 베이지안 통계의 새로운 전성기를 맞는데 기폭제가 되었다. 1989년, BUGS(Bayesian inference Using Gibbs Sampler)라는 베이지안 추론을 위한 통계 패키지가 개발되었고, 최근에는 R을 통해 이용할 수 있는 stan(http://mc-stan.org), JAGS(Just Another Gibbs Sampler, http://mcmc-jags.sourceforge.net)와 같은 통계 패키지들이 등장하였다. 이러한 다양한 패키지의 등장으로, 누구나 베이지안의 방법론을 쉽게 이용할 수 있는 환경이 마련되었다.

3. 베이지안 통계의 현재

마코프 체인 몬테 카를로를 비롯한 방법론과 BUGS 등의 베이지안 통계패키지의 발전 덕분에, 베이지안 통계는 현재 많은 분야에서 활발하게 사용되는 추세이다.

Ryan과 Woodall은 가장 많이 인용된 통계학 논문에 대한 조사를 진행하였다 (Ryan과 Woodall, 2005). 이들은 ISI(Institute for Scientific Information)에서 제공하는 웹 데이터베이스를 이용하여 가장 많이 인용된 통계학 논문 리스트를 발표하였다. 조사 대상이 된 논문들은 새로운 통계 방법론을 제시하거나, 기존의 통계 방법론을 수정하거나 또는 기존의 통계 방법론을 참신한 방식으로 중요한 과학 문제에 적용한 논문들이었으며, 때문에 선정 기준에는 주관적인 요소가 개입되기는 하였다. 이들은 논문 리스트를 두 가지로 나누어서 제시하였는데, 첫째는 기존의 모든 논문들 중 가장 많이 인용된 25개의 논문 리스트였고, 둘째는 1993년이나 그 이후에 출판된 논문들 중 가장 많이 인용된 15개의 논문 리스트였다.

모든 논문들을 대상으로 조사한 25개 논문 리스트에 포함되어 있는 베이지안 논문은 총 3편으로 그 비율이 12%였다. 현재까지의 모든 논문의 총 인용수 기준이므로 오래 전에 출판된 논문일수록 상대적으로 유리할 수밖에 없었다는 점을 고려할 때, 이 결과는 과거를 나타낸다고 할 수 있다. 반면, 1993년이나 그 이후에 출판된 15개 논문 리스트에서는 베이지안 논문은 무려 5편으로, 약 33%의 비율로 나타났다. 앞의 결과와 비교할 때 베이지안 논문의 중요도는 약 세 배가 증가한 것을 알 수 있고, 이것은 최근들어 베이지안의 활용도가 증가하고 있다는 것을 시사한다.

베이지안이 점점 더 활발하게 연구된다는 또다른 증거로, 4대 통계저널인 JASA(Journal of the American Statistical Association), JRSSB(Journal of the Royal Statistical Society: Series B), Biometrika, 그리고 The Annals of Statistics에서 년도별 베이지안 논문의 비율을 조사하였다. 베이지안 논문의 기준은 키워드에 베이즈 또는 베이지안이 포함된 논문들로 하였고, 1990년부터 2014년 현재까지 출판된 논문들을 대상으로 하였다. 단, JRSSB 경우 1997년부터 2014년까지의 논문들을 조사하였다.

Figure 3.1을 보면 공통적으로 모든 저널들에서 베이지안 논문의 비율이 증가하는 추세라는 것을 확인할 수 있다. 더욱이 키워드에 베이즈 또는 베이지안이 포함된 논문만을 대상으로 한 결과이기 때문에, 요즘 새롭게 추가되고 있는 베이지안의 키워드를 추가한다면 증가 추세가 더욱 뚜렷할 것이라 기대된다.

베이지안의 논문 비율이 증가한다는 것은 베이지안이 활용되는 분야가 점점 다양해지고 있다는 것과도 연관이 있다. 실제로 Berger는 베이지안 통계가 응용되고 있는 분야들과 해당 논문들을 간략하게 열거했는데, 그의 논문에 열거된 23가지의 분야는 다음과 같다 (Berger, 2000): 생물통계(Biostatistics), 인과(Causality), 분류(Classification), 분할표(Contingency table), 의사 결정 이론(Decision theory), 디자인(Design), 경험적 베이즈(Empirical Bayes), 교환가능성(Exchangeability), 유한 집단 표본추

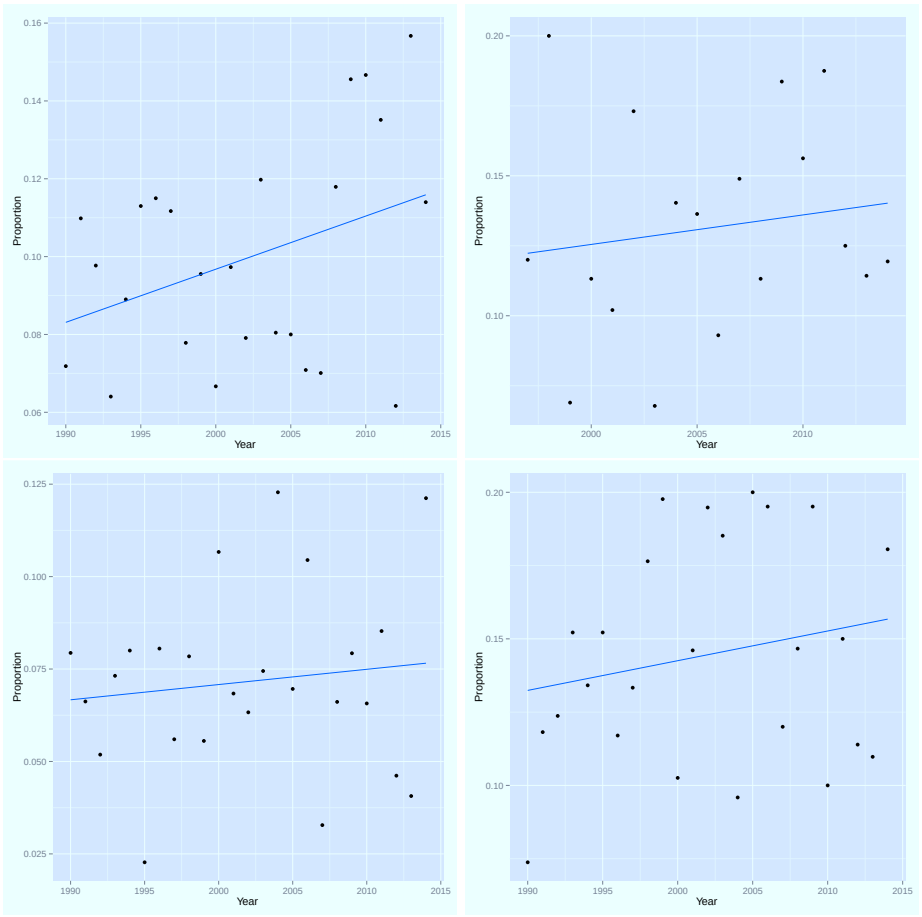

Figure 3.1. The yearly trend of Bayesian articles in JASA(left top), JRSSB(right top), Biometrika(right bottom), The Annals of Statistics(left bottom). The black points are data, and the blue lines are the results of the linear regression fittings.

출(Finite population sampling), 일반화 선형모형(Generalized linear models), 그래프 모형(Graphical model), 계층 모형(Hierarchical model), 이미지 처리(Image processing), 정보(Information), 결측자료(Missing data), 비모수와 함수 추정(Nonparametrics and function estimation), 순서형 자료(Ordinal data), 예측 추론과 모형 평균(Predictive inference and model averaging), 신뢰도와 생존 분석(Reliability and survival analysis), 순차 분석(Sequential analysis), 신호 처리(Signal processing), 공간 통계(Spatial statistics), 검정, 모형선택과 변수선택(Testing, model selection, and variable selection). 이 외에도 특성 모형(Feature model), 미분 방정식(Differential equation), 동적 모형(Dynamic model) 등 베이지안은 매우 폭넓은 분야에서 응용되고 있다.

4. 통계의 미래에 대한 조망

베이지안 통계의 미래는 통계학의 미래와 떼어 놓고 생각할 수 없다. 따라서 베이지안의 통계의 미래를

통계학의 미래에 대해 논하면서 같이 생각해 보려고 한다.

미래에 대한 예측은 불가능에 가까울 만큼 어려운 일이다. 새로운 발견이나 예기치 않은 일들로 미래는 우리가 전혀 예측할 수 없는 방향으로 나갈 수 있다. Berger는 자신이 1980년대에 베이지안 통계학의 미래에 대해서 예측을 했다면, 베이지안 통계학의 역사에서 매우 중요한 사건인 마코프 체인 몬테 카를로의 출현을 완벽하게 놓쳤을 것이라고 얘기했다 (Berger, 2000). 통계학자들의 용어로 얘기하면 우리의 현실은 정상성(Stationarity)을 갖지 않다고 얘기할 수 있을 것이다. 따라서 우리가 하는 미래의 예측은 우리의 현실이 정상성에서 크게 벗어나지 않는다는 가정하에서 현재에 벌어지는 일들로 미래에 대한 추측을 해보는 것이다.

미래 통계학의 특징으로 다음의 다섯 가지를 생각해 본다. 첫 번째, 자료의 대용량화이다. 자료가 대용량화 된다는 얘기는 이미 많은 글에서 찾아볼 수 있다. 구글, 페이스북, 네이버 등의 정보(Information Technology)회사들은 매일 발생하는 어마어마한 양의 자료를 스토리지에 쌓아놓고 필요한 정보를 뽑아내고 있다. 인간의 유전자 자료, 뇌과학 관련 자료, 기상학 자료 등은 상상을 초월할 정도의 크기를 자랑한다. 몇 테라 혹은 페타 바이트가 넘는 엄청난 양의 자료에서 통계적 정보를 실시간으로 추출해 내는 것이 현재 통계학자들에게 대두된 문제이다. 현재의 통계 계산 속도, 특히 베이지안 통계 계산 속도는 우리가 원하는 계산 속도에 훨씬 미치지 못한다. 과거에는 컴퓨터의 계산 속도가 일 이년만 기다리면 배로 빨라지곤 했지만 이제는 더 이상 컴퓨터의 계산 속도 개선을 기대할 수 없다. 하지만 자료의 양은 컴퓨터 계산 속도를 기다려 주지 않고 기하급수적으로 커지고 있다. 이제 통계 계산 속도의 향상을 더 이상 컴퓨터공학자들에게만 맡겨 놓을 수는 없게 되었다. 통계 계산 속도의 향상에 대해서 통계학자들이 나서야 할 때가 되었다. 우리는 새로운 통계 알고리듬을 개발하고, 필요하다면 기존의 추론의 패러다임을 바꾸어서 대용량 자료의 시대에 맞는 새로운 통계적 추론을 방법을 만들어야 한다. 베이지안 통계는 1990년에 베이지안 계산의 혁명을 보았다. 마코프 체인 몬테 카를로의 출현은 기존에 이론으로만 머물던 베이지안 통계를 응용통계의 최첨단으로 이끌었다. 20여년이 지난 지금 마코프 체인 몬테 카를로는 여전히 베이지안 응용통계의 가장 중요한 계산 방법이지만 현대 통계의 흐름인 자료의 대용량화에 전혀 어울리지 않는다. 베이지안 통계는 자료의 대용량화에 맞는 새로운 계산 방법을 고안해 내야 한다. 이와 같은 계산 방법에 대한 연구는 벌써 진행 중이다 (Kleiner 등, 2014; Minsker 등, 2014; Zhu와 Dunson, 2013). 또한 베이지안 모형에 GPU(Graphics Processing Unit)를 이용한 병렬계산을 적용하는 등의 연구도 진행되고 있다 (Suchard 등, 2010).

두 번째, 통계학으로 답을 얻는 문제의 복잡성이다. Efron은 통계학의 역사를 다음과 같이 간단하게 구분하였다 (Efron, 2009).

- 19세기: 대용량 자료, 간단한 질문.
- 20세기: 소용량 자료, 간단한 질문.
- 21세기: 대용량 자료, 복잡한 질문.

19세기에는 천문학과 공공 통계에서 쏟아지는 대용량 자료에서 간단한 질문을 답하는 것이 통계학의 주된 임무였다면, 20세기는 소용량의 자료에서 주어진 간단한 질문에 대해 가능한 모든 정보를 얻어내는 것이 그 주된 임무였다. 따라서, 통계 이론은 효율성(Efficiency), 최소최대수렴속도(Minimax convergence rate) 등을 다루었다. 그러나, 현대의 통계학에 주어지는 문제는 대용량의 자료와 함께 훨씬 더 많고 복잡한 질문의 해답이 통계학에 요구된다. 그 대표적인 예가 큰 변수의 개수, 작은 자료의 개수(large p, small n) 문제이다. 여기서 작은 자료의 개수라고 하지만 실제로 자료의 개수가 작다는 의미는 아니다. 단지 변수의 개수가 자료의 개수 비해 훨씬 크다는 말이다. 변수의 개수가 커지면서 새로운 문제가 야기된다. 1,000개의 관측치를 분석하여 5개의 주어진 변수 중 어떤 변수들이 중요한 변수들

인지 알아내는 것과 10,000개의 관측치와 100,000개의 변수를 가진 자료에서 어떤 변수들이 중요한 변수들인지 알아내는 것은 완전히 다른 문제이고 통계학의 이론과 실제에 새로운 도전을 가져다 준다. 베이지안 통계학도 이러한 문제에 베이지안적인 참신한 아이디어로 대처를 해야할 것이다.

세 번째, 새로운 데이터 형태의 출현이다. 20세기에는 실수와 벡터가 통계학자들이 분석해야 할 자료의 형태였다. 이제는 훨씬 다양한 형태의 자료가 주어지고 있다. 요즘의 자료는 함수, 행렬, 이미지, 문서 등 매우 다양한 형태를 띄고 있다. 지금까지는 새로운 자료의 형태에 주의를 크게 기울이지 않았지만 앞으로는 많은 주의를 기울여야 할 것이다. 베이지안 모형의 강점인 계층모형(Hierarchical model)을 이용하면, 이렇듯 복잡한 형태를 띄는 자료를 다룰 수 있다. 복잡한 자료의 형태에서 주변의 정보를 이용하여 추론은 하는 것은 베이지안 통계 이외의 방법으로는 매우 쉽지 않은 일이다.

네 번째, 통계학의 보편화이다. 다가올 미래의 세계에서 통계학이 매우 중요할 것이라는 것은 의심의 여지가 없다. 세계는 중요한 결정을 내릴 때 좀 더 정확한 증거를 원하고, 이는 곧 증거와 불확실성의 수치화를 의미한다. 통계학은 미래의 인류가 가져야할 중요한 교양 중의 하나가 될 것이다. 중요한 결정을 내려야하는 사람들에게 통계학은 필수적인 덕목이 되고 웬만한 통계적인 내용은 상식적으로 알게될 것이다. 이러한 통계학의 보편화 현상이 통계학자들에게 도움이 될까? 이 문제에 대한 답은 그렇게 간단하지 않다고 생각한다. 과거에는 간단한 적합도 검정을 수행하는데도 통계학자를 찾았다면 이제는 많은 사람들이, 얼마 전까지 일부 통계학자들의 전유물이었던 비모수 함수 추정을 R 패키지를 통해 수행할 수 있게 되었다. 많은 사람들이 상식으로 통계학을 알기 때문에 일반인들이 통계학자를 찾을 때는 매우 어려운 문제를 들고 통계학자를 찾게 되고 통계학자는 단순한 통계적 지식을 넘어서는 능력을 보여줘야 할 것이다. 비슷한 예는 수학에서도 찾아볼 수 있다. 현대의 사회에서 수학이 중요하지 않다고 할 사람은 아무도 없을 것이다. 그러나 수학자의 탁월한 수학 능력이 필요해서 수학자를 찾는 경우는 예외적인 경우를 제외하고는 별로 없을 것이다. 이미 수학을 필요로 하는 많은 사람들이 수학을 이미 자신이 필요한 만큼 잘하고 있다. 예를 들면, 물리학자, 기상학자, 천문학자, 통계학자, 기계공학자 등이 자신들이 필요한 만큼 수학을 할 수 있고, 그들이 풀기 힘들어 하는 수학문제는 수학자들에게도 이미 어려운 문제가 되었다. 사회에서 교육 이외의 분야에서 전문수학자의 직함을 갖고 활동하는 사람은 이미 거의 없다. 수학이 그런 것처럼 앞으로는 통계학자가 할 일이 통계학을 필요로 하는 일반인들에 대한 교육과 새로운 통계 방법론을 개발하는 일로 특화되는 경향을 보일 것이다.

다섯 번째로 데이터 사이언티스트(Data scientist)의 등장이다. 데이터 사이언티스트의 정의는 확실하지 않지만, 통계학자와 컴퓨터공학자의 중간 정도 인 것 같다. 이는 Wills의 데이터 사이언티스트의 정의, “데이터 사이언티스트란 통계학자들보다는 계산을 잘하고, 컴퓨터공학자들 보다 자료분석을 잘하는 사람을 말한다” 에서 잘 나타나 있다고 생각한다 (Bühlmann 등, 2014). 데이터 사이언티스트의 등장은 통계학자들에게 새로운 도전을 가져다 준다. 아직 한국에서는 확연하게 나타나는 현상은 아니지만, 구글이나 페이스북에서는 통계학 전공자 보다 데이터 사이언티스트를 선호한다고 한다. 데이터 사이언티스트의 미래가 어떻게 될지는 아무도 모른다. 몇몇 통계학자들이 말하는 것처럼 통계학은 100년이 넘은 용어이고 데이터 사이언티스트는 데이터 마이너(Data miner), 6-시그마 블랙벨트처럼 잠시 나타났다가 사라지는 말일 수도 있다. 그러나 한 가지 중요한 것은 통계학이 이러한 사회의 변화를 무시하면 안되고 우리의 교육 시스템에 이를 잘 반영해야 한다는 것이다. 즉, 다음 세대의 통계학 전공자들을 현재 데이터 사이언티스트들에게 요구되는 통계학과 컴퓨터공학의 역량을 동시에 갖는 그런 인재들로 키워내야 한다는 것이다.

5. 결론

앞에서 베이지안의 역사에 대해 간략하게 알아보고, 통계학의 미래를 조망해보면서 베이지안 통계의 미

래도 살펴보았다. 최초의 통계는 베이지안 통계였다. 이후 빈도론적 통계의 등장으로 20세기 초반에 암흑기를 맞았지만, 베이지안 계산의 혁명으로 새로운 전성기를 맞고 있다. 베이지안 통계의 개념은 이제 통계학의 테두리를 벗어나서 전 학문분야로 퍼져나가고 있다. 한국의 통계학자들이 베이지안 통계의 전성기에 중요한 역할을 하기를 바라면서 글을 마친다.

References

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. by the late Rev. Mr. Bayes, FRS. communicated by Mr. Price, in a letter to John Canton, AMFRS, *Philosophical Transactions (1683–1775)*, 370–418.

Berger, J. O. (2000). Bayesian analysis: A look at today and thoughts of tomorrow, *Journal of the American Statistical Association*, **95**, 1269–1276.

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference, *Journal of the Royal Statistical Society: Series B(Statistical Methodological)*, 113–147.

Bühlmann, P., Carroll, R., Murphy, S., Roberts, G., Scott, M., Távare, S., Triggs, C., Wang, J. L., Wasserstein, R. L., Madigan, D., Bartlett, P. and Zuma, K. (2014). Statistics and science: A report of the london workshop on the future of the statistical sciences.

De Finetti, B. (1938). Sur la condition d'equivalence partielle, *Actualites Scientifiques et Industrielles*, 739.

Efron, B. (2009). The future of statistics, *Amstat News*, **363**, 47–50.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association*, **85**, 398–409.

Gillies, D. (2000). *Philosophical Theories of Probability*, Psychology Press.

Hume, D. (1748). *Philosophical Essays Concerning Human Understanding*, Andrew Millar.

Jeffreys, H. (1939). *Theory of Probability*.

Kleiner, A., Talwalkar, A., Sarkar, P. and Jordan, M. I. (2014). A scalable bootstrap for massive data, *Journal of the Royal Statistical Society: Series B(Statistical Methodology)*, **76**, 795–816.

Laplace, P. (1774). Memoire sur la probabilite des causes par les evenements, *l'Academie Royale des Sciences*, **6**, 621–656.

Lindley, D. V. (1965). *Introduction to probability and statistics from Bayesian viewpoint, Part 2 inference*, CUP Archive.

McGrayne, S. B. (2011). *The Theory that Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, & Emerged Triumphant from Two Centuries of Controversy*, Yale University Press.

Minsker, S., Srivastava, S., Lin, L. and Dunson, D. B. (2014). Scalable and robust Bayesian inference via the median posterior, In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 1656–1664.

Ryan, T. P. and Woodall, W. H. (2005). The most-cited statistical papers, *Journal of Applied Statistics*, **32**, 461–474.

Savage, L. J. (1954). *The Foundations of Statistics*, John Wiley and Sons, Inc.

Suchard, M. A., Wang, Q., Chan, C., Frelinger, J., Cron, A. and West, M. (2010). Understanding GPU programming for statistical computation: Studies in massively parallel massive mixtures, *Journal of Computational and Graphical Statistics*, **19**, 419–438.

Wrinch, D. and Jeffreys, H. (1919). On some aspects of the theory of probability, *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, **38**, 715–731.

Zhu, B. and Dunson, D. B. (2013). Locally adaptive Bayes nonparametric regression via nested Gaussian processes, *Journal of the American Statistical Association*, **108**, 1445–1456.

베이지안 통계의 역사와 미래에 대한 조망

이재용[a] · 이경재[a,1] · 이영선[a]

[a]서울대학교 통계학과

(2014년 10월 27일 접수, 2014년 12월 1일 수정, 2014년 12월 5일 채택)

요 약

최근 계산 기술의 진보로 인하여, 베이지안 통계는 급속도로 확산되어 가고 있다. 그러나, 정보화 시대에 들어서면서 베이지안 통계를 비롯한 통계학은 새로운 문제들에 직면하게 되었다. 이 논문에서는 베이지안 통계의 역사를 간단히 살펴보고, 베이지안 통계의 현재의 영향력에 대해서 알아본다. 그리고 통계학의 미래와 통계학계가 직면한 도전과제들에 대하여 생각해 볼 것이다.

주요용어: 베이지안 통계, 토마스 베이즈, 통계의 미래.

[1]교신저자: (151-747) 서울시 관악구 관악로 1, 서울대학교 통계학과. E-mail: leekjstat@gmail.com