Understanding Bayesian Statistics

정윤식[1)]

요약

통계학은 불확실성(uncertainty)에 대한 연구이다. 베이지안 통계 방법은 불확실성 아래서 통계 추론과 의사 결정 모두를 위한 *완전한(complete)* 패러다임을 제공한다. 베이지안 방법론은 합리적인 초기 정보와 결합하는 것을 가능하게 만들고, 전통적인 통계적 방법론에 의하여 직면하는 많은 어려움들을 풀 수 있는 *coherent* 방법론을 제공하면서 엄격한 수학적 기본에 근거하고 있다. 베이지안 패러다임은 일반적인 용어로써 확률이란 단어의 사용을 가장 잘 어울리게 하는 *불확실성의 조건부 측도(conditional measure of uncertainty)*로써 확률의 해석에 근거한다. 관심있는 것에 대한 통계적 추론은 증거의 관점에서 그 값에 대한 불확실성의 변형으로써 묘사되며, 베이즈 정리(Bayes' theorem)는 이러한 변형이 어떻게 만들어지는 가를 자세히 설명할 수 있다. 베이지안 방법들은 전통적인 통계적 방법론에 접근할 없는 복잡하고, 다양한 구조적 문제들에 응용할 수 있다.

1. Introduction

과학적인 실험 또는 관찰의 결과들은 자료들의 일반적 형태인 $D=\{x_1, x_2, \cdots, x_n\}$로 구성된다. 이때 x_i들은 "동질적(homogeneous)"이다. 여기서 동질적이란 그들 자체의 값만이 중요하지 그들의 순서는 무시함을 의미한다. 통계적 방법들은 이러한 관측치를 생성한 과정의 본질과 동일 과정에 의하여 생성되는 미래 관측치의 행태에 대한 결론들을 유도하기 위하여 사용된다. 통계적 해석의 중요한 요소는 모수 $\omega \in \Omega$의 함수로써 관측된 자료를 생성하는 기법을 묘사한다고 가정되는 확률 모형의 설정에 있다. 따라서 이러한 유도된 통계적 결론들은 가정된 확률 모형에 완전히 의존된다.

Bernardo와 Smith(1994)는 수학의 대부분 다른 분야들과는 다르게 통계적 추론의 형식적인 방법들은 공리적 기본이 부족함을 겪는다고 서술하였다. 결론적으로, 그들의 제안된 요망들은 가끔 상호간에 모순되고, 동일 자료의 해석이 다른 직관적인 방법들에 사용될때 모순된 결과들을 유도할 수 있다. 현격한 차이로써, 통계적 추론의 베이지안 접근은 논리적 구조의 통합을 제공하고, 제안된 방법들의 상호간 일치성을 보장하는 공리적 기본에 견고하게 근거한다고 주장한다.

베이지안 통계학은 오직 확률론의 *수학*과 일반적인 용어로써 확률이란 단어의 사용을 가장 근접하게 대응하는 확률의 *설명*을 오직 요구한다. 이는 Laplace(1814), de Finetti(1974), Jeffreys(1961)등과 같은 유명한 통계학자들의 베이지안 통계학에 대한 매우 중요하고 생산적인 교재들이 "확률론(Probability Theory)"란 이름을 사용하는 것이 우연은 아니다. 베이지안 방법들은 (1) 통계적 추론을 확률론에 있는 문제로 축소하고, 그러므로 완전하게 새로운 개념들을

1) 부산대학교 자연과학대학 통계학과 yschung@pusan.ac.kr

위한 필요들을 최소화 하고, 그리고 (2) 논리적 정당성을 어떤것에 제공하거나 다른 것들의 논리적 불일치성을 증명하므로써 형식적인 통계적인 기교들중에서 구분하기 위하여 사용된다.

이러한 기본들의 주된 결론은 확률분포들의 수단에 의하여 문제에서 존재하는 모든 불확실성들을 묘사할 수학적 필요성에 있다. 실제적으로 확률 모형들에서 미지의 모수들은 그들의 값들에 대한 이용 가능한 정보를 묘사할 결합 확률 분포를 가져야만 한다. 이는 가끔 베이지안 접근의 가장 특이한 요소로써 인식된다. 또한, 전통적인 통계학과 비교하여 모수들은 베이지안 페러다임안에서는 확률변수로 인식된다. 이러한 인식은 모수를 고정된 미지로 취급하므로써 생각하는 그들의 변이로써의 묘사가 아니라 그 들의 참값에 대한 불확실성의 묘사에 있다. 적절한 사전 정보가 쉽게 이용될 수 없거나 그 정보들이 주관적이지만 "객관적인 (objective)"해석이 요구될 때 중요하고 특이한 경우들이 발생한다. 이는 적당한 참조 사후 분포를 유도하는 정보-이론개념을 사용하는 참조해석(refernce analysis)에 의하여 소개된다.

2. Foundations

베이지안 패러다임의 중요한 요소는 모든 적절한 미지의 양들을 묘사하기 위하여 확률분포를 이용하는 것이다. 이러한 분포들은 어떤 특별한 조건에 있는 사건의 발생에 대한 불확실성의 조건부 척도로써 0과 1사이의 수로 평가하고자 하는 사건의 확률을 설명한다.

2.1. 조건부 불확실성의 척도로써 확률

Lindley(1983)는 베이지안 통계학은 확률을 수단으로 불확실성에 대한 만족스러운 묘사를 하고자 하는 아이디어에 근거한다 하였다. 이를 다시 표현하면, 베이지안 통계학은 가능한 정보와 채택된 가정들이 주어진 아래 특별한 사건의 발생과 연관이 있는 불확실성의 조건부 척도로써 언어에서 단어의 의미를 갖는 확률(probability)이란 단어를 이용한다. 그러므로 $\Pr(E|C)$는 조건 C아래서 사건 E의 발생에 대한 믿음의 척도이다. 전형적인 응용들에서, 우리들은 이용 가능한 자료 D와 자료를 생성하는 기법으로 만들도록 준비한 가정 A와 이용할 수 있을 적절한 문맥상 지식 K가 주어진 조건 아래 사건 E에 관심이 있다. 그러므로 $\Pr(E|D,A,K)$는 자료 D, 가정 A와 다른 이용 가능한 지식 K아래서 사건 E의 발생에 대한 믿음의 정도로 설명할 수 있다. 다음 예들은 불확실성의 조건부 측도로써 확률을 사용하는 것을 설명한다.

Probabilistic diagnosis: 인간 집단은 특별한 바이러스에 감염되는 비율이 약 2%라 알려져 있다 하자. 여기서 V는 한사람이 바이러스를 보유할 사건이고, +는 양성반응을 나타내는 사건이라 하자. 실험실 자료에 의하여 $\Pr(+|V)=0.98$, $\Pr(+|\overline{V})=0.01$임을 알 수 있고, 이때 한사람의 검사 결과는 양성이라 하자. 그러므로 사람들은 $\Pr(V|+,A,K)$에 관심이 있다. 즉, 이는 양성 반응 결과 $D(+)$, 이 검사 결과를 수행한 확률적 기법에 대한 가정 A와 주어진 연구아래서 집단에 있는 감염 보급에 대한 이용 가능한 지식 K (여기서 $\Pr(V|K)=0.02$이다)아래서 사람이 바이러스를 보유할 확률이다. 이때, 베이즈 정리를 이용하면 $\Pr(V|+,A,K)=0.164$이다. 이 문제에서 포함된 네 가지 확률들은 모두 정확하게 같은 설명들을 갖고 있다.: 그들 모두가 불확실성의 조건부 확률들이다. 그 밖에, $\Pr(V|+,A,K)$는 양성으로 검사된 사람이 실제로 감염될 사건에 연관된 불확실성의 측도이고, 양성반응을 받은 사람들 중에서 감염됐다고 궁극적으로 판명될 사람들의 비율에 대한 추정치(약 16.4%)이다.

Estimation of a proportion : 주어진 성질을 공유하는 집단에 있는 개인들의 비율 θ를 추정하고자 하나의 조사를 수행한다. n개의 확률 표본들 중 r개가 이 성질을 갖고 있다고 알려졌다. 이때 θ의 미지 값이 놓이기로 기대될 수 있는 [0, 1]구간을 만들기 위하여 표본에서의 결과들을 이용하는데 관심이 있다; 이 정보는 $\Pr(a<\theta<b|r,n,A,K)$형태의 확률로써 제공된다. 예를 들면, 1500명중 720명이 특별한 정치 이념에 공감을 갖고 있다는 정치적 조사후 $\Pr(\theta<0.5|720,1500,A,K)=0.933$이라는 결론을 얻었다하자. 이는 이 문제에 대한 국민투표가 실시하면 부결될 확률이 약 93%임을 의미한다. 비슷하게, 100명을 검사하여 그들중 아무도 감염되지 않은 것으로 판명되는 감염에 대한 선별 검정후, $\Pr(\theta<0.01|0,100,A,K)=0.844$임을 얻을 수 있다 하자. 이는 감염될 사람의 비율이 1%보다 적을 확률은 약 84%이다.

Prediction: 하나의 실험은 잘 정의된 상황의 n번 반복중 각각에서 사건 E가 r번 일어났음을 알 수 있다. 사건 E가 i번째 반복에서 r_i 번 일어나는 것을 관측한다. 같은 상황에서 사건 E가 미래에 r번 일어나기를 예측하고자 한다. 이는 예측에 대한 문제로써 $r=0,1,\cdots$ $\Pr(r|r_1,r_2,\cdots,r_n,A,K)$과 같은 확률들의 계산을 요구한다. 예를 들면, 연속 $n=10$개월중 매달마다 고객들로부터 에어백에 대하여 전혀 불평사항을 접수되지 않았다 하자. 이때 자동차의 제어 장치를 생산하는 공장의 숙련된 기술자가 다음과 같은 것을 보고하였다.

$$\Pr(r=0|r_1=r_2=\cdots=r_n=0\ A,\ K)=0.953$$

이는 관측된 자료, 채택된 가정들과 다음달 생산에 에어백에 대한 불평이 있을 사건에 대한 문맥상 지식이 주어진 상황아래서 조건부 불확실성이 측도로 인식된다.

2.2 Statistical Inference and Decision Theory

결정 이론은 불확실성 아래서 결정론 문제를 다루는 명확한 방법론을 제공할 뿐만이 아니라 그 것의 완벽한 공리적 기본은 역시 베이지안 접근의 논리력을 위한 강력한 경우를 제공한다.

두 가지 이상의 가능한 행동(action)들이 있을 때 결정론 문제는 존재한다. A를 가능한 행동들의 집합이라 하자. 더구나, $a\in A$에 대하여 Θ_a는 a를 선택하는 결과에 영향을 주는 적절한 사건들의 집합이다. $c(a,\theta)\in C_a$, $\theta\in\Theta_a$는 사건 θ가 일어날 때 행동 a을 선택한 결과이다. $\{(\Theta_a, C_a), a\in A\}$인 쌍들의 집합은 결정론 문제의 구조를 서술한다. 일반성의 손실 없이 가능한 행동들은 상호간 배타적이다.

원론적으로 다른 집합들은 이성적인 의사 결정을 위하여 민감하게 요구되는 논리적 규칙들의 최소한의 모임을 설명하기 위하여 제안되었다. 이 들은 강력하고 직관적인 설득력을 갖는 공리들로 구성한다: 예를 들면, 선호도의 추이성(만약 주어진 C아래서 $a_1 > a_2$이고 주어진 C아래서 $a_2 > a_3$이면, 주어진 C아래서 $a_1 > a_3$이다)과 sure-thing principle(만약 주어진 C, E아래서 $a_1 > a_2$이고 주어진 C, $\overline{E}$아래서 $a_1 > a_2$이면, 주어진 C아래서 $a_1 > a_2$이다)등이 있다. 채택할 수 있는 원리들의 집합을 위한 다른 선택들이 있지만, 기본적으로 그 들 모두는 동일한 결론들에 유도된다. 즉,

(1) 결론들 중의 선호도는 그들의 호감도를 수치 값으로 지정하는 유계인 실 효용(utility)함수 $U(a,\theta)$로 측정된다.

(2) 적절한 사건의 불확실성은 결정이 이루어지는 조건들 C아래서 그들의 설명력을 묘사하는

확률분포의 집합들 $\{p(\theta|C,a),\ \theta\in\Theta_a,\ a\in A\}$을 갖고 측정된다.

(3) 이용 가능한 행동의 호감도는 그들에 대응되는 다음과 같은 기대 효용에 의해 측정된다.

$$\overline{U}(a|C)=\int_{\Theta_a}U(a,\theta)p(\theta|C,a)\,d\theta,\ \ a\in A$$

이는 가끔 다음과 같이 정의된 비음인 손실함수의 형태로 작업하는 것이 편리하다. 즉,

$$L(a,\theta)=\mathrm{Sup}_{a\in A}\{U(a,\theta)\}-U(a,\theta)$$

이는 잘못된 행동의 선택에 대한 벌점을 직접 측정하는 것이다. 이용 가능한 행동 $a\in A$들 의 상대적인 비호감도는 다음과 같은 그 들의 기대 손실에 의하여 측정된다.

$$\overline{L}(a|C)=\int_{\Theta_a}L(a,\theta)p(\theta|C,a)\,d\theta,\ \ a\in A.$$

특히, 위에 서술된 논쟁들은 모든 적절한 미지의 값들에 대한 불확실성을 수량화 할 필요성을 확립하고, 이러한 값등은 확률분포의 수학적 구조를 가져야만 한다. 이러한 확률들은 이러한 결정이 취해진 상황 C(전형적으로 어떤 적당한 실험 또는 관찰된 자료의 결과들을 포함하는 것) 아래서의 조건부이다.

2.3 Exchangeability and Representation Theorem

이용 가능한 자료는 "동질적(homogeneous)" 관측 치들의 집합 $\{x_1,x_2,\cdots,x_n\}$형태를 갖는다. 정상적으로 이를 호환성(exchangeability)이라 한다. 만약 랜덤 벡터 $\{x_1,x_2,\cdots,x_n\}$의 결합 분포가 순열에 대하여 불변이면 이를 호환 적이라 한다. 랜덤 벡터인 하나의 무한 수열 $\{x_j\}$중 모든 유한 수열들이 호환 적이면 이를 호환 적이라 한다. 1930년대에 de Finette에 의하여 소개된 호환성은 현대 통계적 사고의 중심에 있다. 사실, 일반적인 **표현이론(representation theorem)**은 만약 관측 치의 집합이 호환적 수열의 부분 집합이라 가정한다면, 이때 이 것은 어떤 확률모형 $\{p(x|\omega),\ \omega\in\Omega\},\ x\in X$에서 온 *랜덤 샘플*을 구성한다는 것이다. 더구나, 모수 ω는 관측 치들의 적당한 함수의 극한($n\to\infty$일 때)으로 정의된다. 일반적인 조건 C에 있는 ω값에 대한 이용 가능한 정보는 어떤 확률 분포 $p(\omega|C)$에 의하여 필연적으로 묘사된다. 예로써, 호환 적인 이진 랜덤 자료 $x_j\in\{0,1\}$의 수열 $\{x_1,x_2,\cdots\}$의 경우에서, de Finette의 표현 이론은 $x_1,x_2,\cdots,x_n$의 결합 분포는 다음과 같은 형태의 적분 표현을 갖는다.

$$p(x_1,x_2,\cdots,x_n|C)\ =\ \int_0^1\prod_{i=1}^n\theta^{x_i}(1-\theta)^{1-x_i}\ p(\theta|C)\,d\theta,\quad \theta=\lim_{n\to\infty}\frac{r}{n}$$

여기서 $r=\sum x_j$는 양의 값을 갖는 수이다. 이는 정확히 확률 분포 $p(\theta|C)$의 존재성이 보여진 모수 θ를 갖는 독립적인 베르누이 시행들의 결합분포이다. 더 일반적으로, 확률 벡터 $\{x_1,x_2,\cdots,x_n\}$의 수열 에 대하여, 호환성은 다음 형태의 적분 표현을 유도한다.

$$p(x_1,x_2,\cdots,x_n|C)\ =\ \int_0^1\prod_{i=1}^n p(x_i|\omega)\ p(\omega|C)\,d\omega$$

여기서 $\{p(x_i|\omega),\omega\in\Omega\}$는 확률모형을 나타내고, ω는 $n\to\infty$일 때 관측치들의 함수의 극한이고, $p(\omega|C)$는 Ω위에서의 확률분포이다. 여기서 $p(\omega|C)$는 ω의 가능한 변이를 나타내는 것이 아니고 그 것의 실제 값에 관한 불확실성의 표현이다.

적당한 조건들 아래서 호환성은 매우 일반적인 가정이고, 랜덤 샘플의 전통적인 개념의 강력

한 확장이다. 사실, 많은 통계적 해석들은 자료들이 어떤 확률 모형에서 조건부 독립 관측 치들의 랜덤 샘플이라 가정한다. 따라서, $p(x_1, x_2, \cdots, x_n | \omega) = \prod_{i=1}^{n} p(x_i | \omega)$이다. 그러므로 랜덤 샘플들은 모두 호환 적이다.

3. Bayesian Paradigm

관측된 자료 D의 통계적 해석은 관측된 자료를 생성하는 확률적 메커니즘을 나타내는 형식적인 확률모형 $\{p(D|\omega), \omega \in \Omega\}$을 제시하기 위하여 사용되는 적당한 비형식적인 묘사적 평가를 갖고 시작된다. 제 2장에서 설명된 논쟁들은 모수공간 Ω위에서 자료가 관측되기 전에 ω 값에 대한 사용 가능한 지식 K로 설명되는 *사전(prior)* 확률분포 $p(\omega|K)$를 평가할 논리적 필연성을 제시하였다. 만약 확률모형이 맞는 다면, 자료 D가 관측된 후, ω에 대한 모든 이용 가능한 정보는 베이즈(Bayes) 정리를 통하여 즉시 얻을 수 있는 확률밀도 함수 $p(\omega | D, A, K)$를 갖는 *사후(posterior)*분포에 포함된다. 즉,

$$p(\omega | D, A, K) = \frac{p(D | \omega)\, p(\omega | K)}{\int_{\Omega} p(D | \omega)\, p(\omega | K)\, d\omega}$$

여기서 A는 확률모형을 만드는 가정들을 의미한다.

Example 1.(Bayesian inference with a finite parameter space)

$p(D|\theta)$, $\theta \in \{\theta_1, \theta_2, \cdots, \theta_m\}$는 관측자료 D를 생성하는 것으로 가정된 확률 메커니즘이라 하자. 이때 베이즈 정리를 이용하면, 자료 D후 θ_i의 사후 분포는 다음과 같다.

$$\Pr(\theta_i | D) = \frac{p(D|\theta_i)\Pr(\theta_i)}{\sum_{j=1}^{m} p(D | \theta_j)\Pr(\theta_j)}$$

여기서 사전 분포함수 $p(\theta) = \{\Pr(\theta_1), \cdots, \Pr(\theta_m)\}$는 θ값들에 대한 이용 가능한 지식들을 나타내며, 각각의 $\Pr(\theta_j | D)$는 사전분포함수로 주어진 초기 지식과 자료 D로 제공된 정보아래서 θ_j가 어느 정도로 인가를 평가하는 값이다. 이러한 간단한 기술의 중요하고 자주 쓰이는 응용은 확률적 진단에 의하여 제공된다. 예를 들면, 바이러스를 찾고자 하는 특별한 검사는 실험실 연구에서 감염된 사람 중에는 98%, 비 감염자중 1%가 양성반응이 나타난다고 알려져 있다 하자. 이때, 양성반응이 나타난 사람이 감염될 사후확률은

$$\Pr(V|+) = \frac{0.98p}{0.98p + 0.01(1-p)}$$

이며, 여기서 $p = \Pr(V)$는 한사람이 감염될 사전 확률이다. 이때, 사전 확률이 영이면 사후 확률도 역시 영이다(즉, 집단이 감염에 자유롭다고 알려져 있다) 그리고 사전확률이 1이면 역시 사후확률도 1이다(이는 집단이 완전히 감염되었다 알려져 있다). 만약 감염이 희귀하다면, 랜덤하게 선택된 사람이 감염될 사후 확률은 검사의 결과가 양성반응일지라도 상대적으로 매우 낮다. 사실 $\Pr(V) = 0.002$에 대하여, $\Pr(V|+) = 0.164$이다. 이는 오직 0.2%만이 감염된 집단에서, 랜덤샘플안에서 이러한 양성반응자중 오직 16.4%가 실제적으로 감염된 것으로 판명날 것이다. 이는 대부분의 양성반응은 실제적으로 잘못된 양성반응일 수 있다.

3.1. The Learning Process

베이지안 패러다임에서, 자료로부터의 학습 과정은 원하는 사후 분포를 생성하기 위하여 이용 가능한 사전 정보와 자료로부터 제공된 정보를 종합하여 베이즈 정리를 이용하므로 써 기계적으로 수행된다. 사후분포의 계산은 베이즈 정리에 의하여 다음과 같이 쓰여진다. 즉, $p(\omega|D) \propto p(D|\omega)p(\omega)$ 이다. $\int_\Omega \pi(\omega)d\omega = \infty$을 만족하는 양의 함수 $\pi(\omega)$를 *비적절(improper) 사전함수*라 정의한다. 만약 $p(\omega)$가 비적절 사전분포 $\pi(\omega)$로 치환후 적절한 사후분포를 유도할 수 있다면 기술적으로 유효하다.

베이즈 정리는 "*사후 분포는 우도함수에 사전함수를 곱한 것에 비례한다*"라 말한다. 베이즈 정리의 결론으로 즉시 *우도 원리(likelihood principle)*가 제안된다.

자연적으로, 사전과 사후라는 용어는 자료들의 특별한 집합에 대하여 상대적이다. 만약 자료 $D = \{x_1, x_2, \cdots, x_n\}$가 축차적으로 제시된다면, 그 자료를 전체적으로 다룰 때와 축차적으로 다룰 때의 결과들은 동일하다. 즉, $i = 1, 2, \cdots, n-1$에 대하여

$$p(\omega|x_1, x_2, \cdots, x_{i+1}) \propto p(x_{i+1}|\omega)\, p(\omega|x_1, \cdots, x_i)$$

임으로, 이는 현 상태에서의 사후는 다음 단계에서는 사전으로 사용된다 것을 의미한다.

대부분의 상황들에서 사후분포는 사전분포보다 "좁다(narrower)". 그러므로 이러한 경우들에서 $p(\omega|x_1, x_2, \cdots, x_{i+1})$이 $p(\omega|x_1, \cdots, x_i)$보다 ω의 참값 주위에 몰려 있을 것이다. 그러나 이는 항상 참이 아니다. 때때로, 하나의 "놀랄만한(surprising)" 관측이 ω의 불확실성을 감소보다는 증가시킨다.

Example 2. (Inference on a binomial parameter)

만약 자료 D가 r개의 양의 시행을 포함하는 모수 θ를 갖는 n개의 베르누이 관측을 구성하고 있다면, 이때 $p(D|\theta, n) = \theta^r(1-\theta)^{n-r}$이고 $t(D) = \{r, n\}$는 충분 통계량이다. 만약 사전분포로써, $p(\theta|\alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$인 베타분포 $Be(\alpha, \beta)$를 따른다면, 이때 베이즈정리를 이용하면, θ의 사후 분포는 $p(\theta|r, n, \alpha, \beta) = Be(\theta|r+\alpha, n-r+\beta)$ 이다.

예로써, 앞에서 언급한 조사의 관점에서 국민투표에서 특별한 정치 쟁점에 대하여 찬성할 국민의 비율 θ에 대한 이용 가능한 정보는 $Be(\theta|50, 50)$이라 하자. 이는 국민투표에서 찬성과 반대가 거의 같을 것이라 볼 수 있고 찬성이 60%보다 적을 확률이 0.95를 의미한다. 1500명에 대한 표본 조사가 시행되었고, 이중 오직 720명이 찬성하였다. 이러한 결과들에 따라서, 사후분포는 $Be(\theta|730, 790)$이다. 기대하는 봐와 같이 자료의 효과는 θ에 대한 초기 불확실성을 강력하게 감소시킨다. 더욱더 명확하게, $\Pr(\theta < 0.5|720, 1500) = 0.933$이며, 이는 조사에서의 자료가 포함된 후 국민투표가 부결될 확률이 약 93%임을 의미한다.

관심 있는 벡터가 모수벡터 ω의 전체가 아니라 ω보다 가능한 적은 차원을 갖는 적당한 함수 $\theta = \theta(\omega)$인 경우를 생각한다. $\{p(D|\omega),\ \omega \in \Omega\}$는 자료 D를 생성하는 가정된 확률 모형이라 하자. $p(\omega)$는 ω의 이용 가능한 정보를 설명하는 확률분포이고, $\theta = \theta(\omega) \in \Theta$는 자료 D에 근거하여 추론되도록 요구되는 원 모수들의 함수이다. 관심 있는 벡터 θ의 유효한 결론은 사후분포함수 $p(\theta|D)$에 포함된다. 이러한 요구되는 사후분포 $p(\theta|D)$는 확률 계산의 일반적인 이용으로 구해진다. 사실 베이즈 정리에 의하여 $p(\omega|D) \propto p(D|\omega)p(\omega)$이다. 더욱이, $\lambda = \lambda(\omega)$는 $\psi = \{\theta, \lambda\}$가 ω에 일대일 대응 변환이 되게 하는 원 모수들의 다른 함수이다. 이때

$J(\omega)=(\partial\psi/\partial\omega)$를 대응되는 쟈코비안(Jacobian)행렬이다. 변수 변환 확률 기법에 의하면, ψ의 사후분포는 $p(\psi|D)=p(\theta,\lambda|D)=\left[\frac{p(\omega|D)}{|J(\omega)|}\right]_{\omega=\omega(\psi)}$ 이므로, θ의 사후분포는 ψ 분포의 주변(marginal) 함수인 $p(\theta|D)=\int_{\Lambda}p(\theta,\lambda|D)d\lambda$ 이다.

3.2 Predictive Distributions

$D=\{x_1,\cdots,x_n\}$, $x_i\in X$을 호환 적 관측 치라 하고, 자료 D를 생성했던 동일한 랜덤 매캐니즘에 의하여 생성될 미래 관측치 $x\in X$를 예측하기 위한 상황을 생각해보자. 이러한 예측 문제에 대한 해결책은 D에 의해 제공된 정보와 다른 유용한 지식이 주어졌을 때 x가 갖는 값에 대한 불확실성을 설명하는 예측 분포 $p(x|D)$에 의하여 간단히 요약 될 수 있음은 2장에서 논의하였다. 문맥상의 정보는 자료 D가 $\{p(x|\omega),\omega\in\Omega\}$족에 있는 분포의 표본공간이란 가정을 제안한다. 또한 $p(\omega)$를 ω값에 대한 정보를 설명하는 사전 분포라 하자. 이때 $p(x|\omega,D)=p(x|\omega)$이기 때문에 $p(x|D)=\int_{\Omega}p(x|\omega)p(\omega|D)d\omega$이다. 이는 D가 주어진 ω의 사후분포를 가중함수로 한 ω의 확률분포에 대한 평균이다.

Example 3. (Prediction in a Poisson process)

$D=\{r_1,\cdots,r_n\}$는 모수 λ를 갖는 포아송 분포 $Po(r|\lambda)$의 랜덤 샘플이라 하자. 이때, 우도함수는 $p(D|\lambda)\propto\lambda^{t}e^{-\lambda n}$ 이며, 여기서 $t=\sum r_i$ 이다. λ의 정보가 없으므로, Jeffrey 규칙에 의하여 비정보 사전 분포인 $p(\lambda)=\lambda^{-1/2}$을 이용한다. 베이즈 정리를 이용하면, 이에 대응되는 사후 분포는 $p(\lambda|D)=Ga(\lambda|t+1/2,\ n)$이다. 또한 이에 대응되는 예측분포는 포아송-감마 혼합 분포로

$$p(r|D)=\int_0^{\infty}Po(r|\lambda)\,Ga(\lambda|t+\frac{1}{2},n)d\lambda=\frac{n^{t+1/2}}{\Gamma(t+1/2)}\frac{1}{r!}\frac{\Gamma(r+t+1/2)}{(1+n)^{r+t+1/2}}$$

이다.

3.3 Asymptotic Behaviour

샘플 크기가 클 때 사후 분포의 형태를 생각해보자. 이 것은 적어도 두 가지 다른 이유 때문에 중요하다. 첫째는 실제 샘플이 상대적으로 클 때 점근적 결과는 유용한 일차(first-order) 근사치를 제공한다. 두번째로, 객관적인 베이지안 방법은 전형적으로 가정된 모델의 점근적 속성에 의존한다. $D=\{x_1,\cdots,x_n\}$은 $\{p(x|\omega),\omega\in\Omega\}$의 랜덤 샘플이라 하자. $n\to\infty$때 이산형 모수 ω의 사후 분포 $p(\omega|D)$가 전형적으로 ω의 참값에 확률 일로 되는 퇴화(degenerate)분포로 수렴하고 연속형 모수 ω의 사후 분포는 $1/n$로 감소하는 분산과 최대 우도 추정치 $\hat{\omega}$을 중심으로 갖는 정규분포에 수렴한다는 것을 알 수 있다.

이제 ω가 k-차(dimensional) 연속 모수인 상황을 생각해보자. 베이즈 정리를 이용하면, $p(\omega|x_1,\cdots,x_n)\propto\exp\left\{\log[p(\omega)]+\sum_{j=1}^{n}\log p(x_j|\omega)]\right\}$으로 표현된다. $\sum_j\log[p(x_j|\omega)]$을 $\hat{\omega}$에 대하여 확장하고, 정규 조건을 가정하면서, ω의 사후 분포는 대략 k 차 정규분포임을 알 수 있다. 즉,

$$p(\omega \mid x_1, \ldots, x_n) \approx N_k\{\hat{\omega}, S(D, \hat{\omega})\}, \quad S^{-1}(D, \omega) = \left(- \sum_{i=1}^{n} \frac{\partial^2 \log[p(x_i \mid \omega)]}{\partial \omega_i \partial \omega_j} \right)$$

간단하지만 다소 오차가 많은 근사식은 $p(\omega \mid \chi_1, \ldots, \chi_n) \approx N_k(\omega \mid \hat{\omega}, \ n^{-1}F^{-1}(\hat{\omega}))$ 이며, 여기서 $F(\omega)$가 피셔 정보 행렬이다. 그러므로 적당한 정규 조건하에서 샘플 크기가 증가할 때 모수 벡터 ω의 사후 확률 밀도는 분산 행렬이 n^{-1}로 감소하는 중심이 $\hat{\omega}$인 다변량 정규 밀도에 수렴한다.

4. Historical Comments

Fisher는 베이지안 접근에 호의적이지 않았으며, 가끔 매우 비판적이었다. Bayesianism "which like an impenetrable jungle arrests progress towards precision of statistical concepts"(1922, p.311)이라 표현하였다. Gill(2002)은 Fisher는 동료를 억압하고 심지어 다른 학자들을 잘못 인용하므로 써 Bayesianism과 Inverse probability를 불신하는 작업을 하였다 한다. Fisher(1935)는 일양 사전분포 없이 inverse probability를 적용할 수 있는 시도인 *신뢰 추론 (fiducial inference)*을 발전시켰으나 이 방법은 실패하였다. Efron(1998, p.105)은 이 실패를 "Fisher's biggest blunder"라 부른다.

1930년대까지 Fisher와 Neyman등에 의하여 Bayesianism은 거의 치명적인 일격을 겪었으나, 죽지는 않았다. Jeffreys(1961), Good(1950), Savage(1954, 1962), de Finetti(1972, 1974, 1975), Lindley(1961, 1965)등과 같은 학자들이 20세기 중반에 고전적 방법에 존재하는 결점들의 반응으로써 베이지안 방법에 관심들을 다시 보이기 시작하였다. 불행하게도 이러한 현대 베이지안 학자들에 의하여 전개된 많은 특성화들은 다루기 힘든 수학적 형태들로 유도되었다. 다행히도 최근에 통계 계산의 혁명으로 알려진 Markov chain Monte Carlo방법은 베이지안 해석에서 오래 끌어온 이러한 문제점들을 해결하고 있다.

참고문헌

1. Bernardo, J.M. and Smith, A.F.M.(1994) *Bayesian Theory*, Chichester: Wiley.
2. de Finetti, B.(1974) *Theory of Probability* (Vol. 1), London: John Wiley.
3. Gill, J.(2002) *Bayesian methods: A social and behavioral sciences approach*, Chapman and Hall
4. Jeffreys, H.(1961) *Theory of Probability*, Clarendon, Oxford.
5. Laplace, P.S.(1814) *Essai Philosophique sur les Probabilites*, Paris: Courcier.
6. Lindley, D.V.(1983) Theory and Practice of Bayesian Statistics, *The Statistician* 32, 1-11.
7. Lindley, D.V.(2000) The philosophy of statistics, *The Statistician* 49, 293-337.