Statistics 2

Bayesian Inference.

2014110374 정상만

7.1 Statistical Inference

Example 7.1.1) Lifetimes of Electronic Components

NOTE

* $i.i.d$ (independent identically distributed) : 각 확률변수는 독립이고, 각각이 동일한 분포를 가짐
* **확률변수(random variable)** : 표본공간 S를 정의역으로 갖는 실함수 (real-valued function)

* **감마분포(Gamma Distribution)**

실수 α, $\beta > 0$ 에 대하여, 연속확률변수 X가 다음을 만족하면 모수 α, β를 갖는 감마분포를 따른다.

$$f(x|\alpha,\beta)=\begin{cases}\frac{\beta^{\alpha}}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x} & x>0\\ 0 & x\le 0\end{cases},\ \Gamma(\alpha)=\int_0^{\infty}x^{\alpha-1}e^{-x}dx\ ,\ \int f(x|\alpha,\beta)dx=1$$

* **지수분포(Exponential Distribution)**

감마분포에서, $\alpha = 1$인 경우 확률변수 X 의 p.d.f 가 다음을 만족한다.

$$f(x|\beta)=\begin{cases}\beta e^{-\beta x} & x>0\\ 0 & x\le 0\end{cases},\ E(X)=\frac{1}{\beta}\ ,\ Var(X)=\frac{1}{\beta^2}$$

* **확률적으로 수렴(Convergence in Probability)**

확률변수들로 이루어진 수열 $Z_1, Z_2, \cdots\cdots$가 확률적으로 b 로 수렴한다.

$\Leftrightarrow\ \forall \epsilon > 0,\ \lim_{n\to\infty}\Pr(|Z_n - b| < \epsilon) = 1\ \Leftrightarrow\ Z_n \xrightarrow{p} b,$: Z_n converges to b in probability.

* **큰 수의 법칙(The Law of Large Number)**

$X_1, \cdots\cdots, X_n$ 이 $\sigma^2 < \infty$, 모평균을 μ 로 갖는 분포로부터 무작위 표본을 생성한다고 가정하자.

$\overline{X}_n$ 을 표본평균이라고 하면, $\overline{X}_n \xrightarrow{p} \mu.$ 를 만족한다.

($\mu < \infty$, $\sigma^2 = \infty$ 인 경우에도 성립한다.)

$X_1, X_2, \cdots\cdots$ 를 연간 전자부품 수명에 관한 수열이라고 하자. 회사가 불량률 θ를 알고 있다고 하면 이는 $i.i.d$ 를 만족하는 확률변수로써 모수(parameter) θ를 갖는 지수분포를 따른다.

또한 $X_1, X_2, \cdots\cdots$인 확률변수 중 $X_1, \cdots\cdots, X_m$ 관측수명값을 알고 있다고 가정하자, 모수 θ를 포함하는 지수적 확률변수의 평균(mean), 즉 평균수명은 $E(X)=\frac{1}{\theta}$ 인 관계를 갖는 것을 상기하면 이에 역수를 취한 θ는 불량률을 의미한다. 앞서 언급했듯이, 만약 모수 θ를 알고 있다고 가정하면, $X_1, X_2, \cdots\cdots$ 는 $i.i.d$ 인 확률변수이므로, 이 경우, 큰 수의 법칙(The Law of Large Number)에 의해 평균 $\frac{\sum_{i=1}^{n} X_i}{n}$ 는 확률적으로

$\frac{1}{\theta}$ 에 수렴하게 되고, 큰 수의 법칙의 따름정리로써, $\frac{n}{\sum_{i=1}^{n} X_i}$ 는 확률적으로 θ로 수렴하게 된다.

이는 θ가 각 지수적인 결과 값들로 구성된 수명에 관한 수열로써의 함수이므로, 확률변수로 다룰 수 있기 때문이다.

이제, 데이터를 관찰하기 전에, 회사가 불량률이 확률적으로 0.5 / year 로 근사하지만 정확한 값을 모른다고 가정하자. θ를 매개변수들을 1, 2 로 갖는 감마분포의 확률변수로 모델링한다고 하고, $X_1, X_2, \cdots\cdots$를 $i.i.d$ 인, 조건부로 θ를 갖는 지수적 확률변수들로 모델링한다고 하자. 우리는 표본 $X_1, \cdots\cdots, X_m$을 검증하는 것으로부터 θ를 좀 더 알아보길 원한다. 그러나 이 θ를 자세하게 알 순 없는데, 이는 전체의 무한수열 $X_1, X_2, \cdots\cdots$을 관찰하는 것이 필요하기 때문이다. 이런 이유로, θ는 오직 가설적으로 관측가능하다.

▲

Definition 7.1.1) Statistical Model

Statistical model $\Leftrightarrow (S, P)$ when $P = \{P_\theta \mid \theta \in \Theta\}$, Θ : the parameters of the model,
S : sample space/the set of possible observations, P : a set of probability distributions on S.

NOTE

A parameterization is generally required to have distinct parameter values give rise to distinct distributions, i.e. $P_{\theta_1} = P_{\theta_2} \Rightarrow \theta_1 = \theta_2$ must hold (in other words, it must be injective). A parameterization that meets the requirement is said to be **identifiable**.

Definition 7.1.2) Statistical Inference

1. 통계적 모형의 모든, 또는 어떤 부분에 대한 확률적인 진술과정.
2. 표본자료에서 구한 통계량에 의거하여 모집단의 모수에 대한 확률적 의사결정을 하는 과정.
3. 검정통계량의 확률분포에서 검정통계량의 값을 발견할 확률에 의거하여 귀무가설에 대한 거부 여부를 결정하는 과정

Definition 7.1.3) Parameter / Parameter space

* 통계적 추론의 문제에서, random variables of interest(관측 or 가설적 관측 가능한 확률변수) 의 결합 확률분포를 결정하는 특성들의 조합 또는 특성을 그 분포의 **parameter(모수, 매개변수)** 라고 한다.
* parameter θ 또는 vector of parameters $(\theta_1, \cdots, \theta_k)$의 모든 가능한 값들의 집합 Ω을 **parameter space(모수공간)** 라고 한다.

* Note) 일반적으로, $\Omega = \{(0, \infty)\}$ 이다.

Example 7.1.2) A Clinical Trial

NOTE

* **베르누이 분포(Bernoulli Distribution)**

확률변수 X 가 다음을 만족하면 모수 p $(0 \le p \le 1)$를 갖는 베르누이 분포를 따른다.

X 가 오직 0과 1을 갖고 확률이 $\Pr(X=1)=p$, $\Pr(X=0)=1-p$ 이며

$$\text{p.f of } X = f(x \mid p) = \begin{cases} p^x(1-p)^{1-x} & x=0,1 \\ 0 & otherwise \end{cases}$$

이 때, 기댓값은 $E(X)=p$, 분산은 $Var(X)=p(1-p)$ 이다.

* **이항 분포(Binomial Distribution)**

확률변수 X 가 다음을 만족하면 모수 $n \in Z_+$, p $(0 \le p \le 1)$를 갖는 이항분포를 따른다.

X 가 이산분포(Discrete Distribution)를 갖고, p.f of X 는

$$f(x \mid n,p) = \begin{cases} \binom{n}{x} p^x(1-p)^{n-x} & x=0,1,\cdots,n \\ 0 & otherwise \end{cases}$$

이 때, 기댓값은 $E(X)=np$, 분산은 $Var(X)=np(1-p)$ 이다.

환자에 대한 확률변수 X_i , $i=1,\cdots,40$ 이 주어졌다 가정하자. $\forall i \in Z_+$에 대하여 $X_i=1$ if a patient recovers 또는 $X_i=0$ if a patient relapses 일 때, 이는 $i.i.d$ 인 베르누이 확률을 가지므로 이산형이고, 모수가 p $(0 \le p \le 1)$ 인 베르누이 분포를 따른다. 이 경우 $p \in [0,1]$이고 $[0,1]$은 모수공간이 된다.

큰 수의 법칙에 따르면, 모수 p는 $n \to \infty$ 일 때, $X = \sum_{i=1}^{40} X_i$ 에 대하여 $\overline{X} = \frac{X}{40} \approx p$ 임을 암시한다.

▲

Definition 7.1.4) Statistic

* 관측 가능한 확률변수들을 $X_1,\cdots,X_n$ 이라 하고, r : n개의 실변수를 갖는 임의의 실함수라고 하자.
그러면, 확률변수 $T=r(X_1,\cdots,X_n)$ 을 통계량(Statistic) 이라고 한다.

* 모집단의 모수를 추정하기 위하여 표본에서 계산한 추정량의 값.
추정량은 통계량을 계산하는 규칙이며, 통계량은 추정량의 계산규칙에 의거하여 표본자료에서 생성된 값.
통계량은 표본에 따라 값이 다른 확률변수로서 확률분포를 가지며, 통계량은 모수에 의존하지 않으나,
통계량의 분포는 모수에 의존한다.

7.2 Prior and Posterior Distributions

Introduction)
임의의 데이터를 관측하기 전의 모수의 분포 : **Prior distribution (사전분포)**
관측된 데이터를 조건으로 하는 모수의 조건부 분포 : **Posterior distribution (사후분포)**
모수를 조건부로 하는 데이터의 조건부 p.f / p.d.f : **Likelihood function (우도함수)**

Example 7.2.1) Lifetimes of Electronic Components
지난 예제에서, 전기부품의 수명 $X_1, X_2, \cdots$ 을 모수를 θ로 갖는 $i.i.d$인 지수분포를 따르는 확률변수들로 가정하고, 이 때, θ는 부품의 불량률(failure rate)로 정의하였다. 큰 수의 법칙에 따라 $n/\sum_{i=1}^{n} X_i$ 는 확률적으로 θ로 수렴하므로, 결과적으로 우리는 θ가 모수로 $\alpha = 1$, $\beta = 2$를 갖는 감마분포를 따른다고 말할 수 있었다. 이러한 θ의 분포는 어떠한 전기 부품의 수명에 대한 관측을 하지 않고 정하였으므로, 이런 이유에서 이 분포는 사전분포(Prior distribution) 라고 부른다.

> Definition 7.2.1) Prior Distribution / p.f / p.d.f
> * 모수를 θ로 갖는 어떤 통계적 모형을 가정하자. 이 때 θ를 확률변수로 간주한다면,(θ를 모르므로) θ가 아닌 다른 확률변수(r.v of interest)를 관측하기 전에 채택한 분포를 **사전분포(Prior-Distribution)** 라 한다.

Note) 모수공간이 셀 수 있다면(countable), 사전분포는 이산형(discrete) 이고, 그것의 p.f를
prior p.f. of θ 라고 한다.
사전 분포가 연속(continuous)분포이면, 그것의 p.d.f. 는 **prior p.d.f. of θ** 라고 한다.
모수 θ에 대한 함수로써 prior p.f / p.d.f. $\Leftrightarrow \xi(\theta)$ 로 표기한다.

Example 7.2.4) Lifetimes of Fluorescent Lamps
특정한 종류의 형광등의 시간당 수명에 대한 관측값이 모수 θ를 갖는 지수분포를 따른다고 가정하자. 또한 모수 θ의 정확한 값은 알려지지 않고, 선험적으로 사전분포가 평균 0.0002, 표준편차가 0.0001인 감마분포를 따른다고 가정하자. 그러면 이제 prior p.d.f. of θ를 결정해야 한다.

우선 감마분포는 $E(X) = \frac{\alpha}{\beta}$, $Var(X) = \frac{\alpha}{\beta^2}$ 임을 상기하면, $E(X) = 0.0002, Var(X) = 0.0001$ 이므로 방정식을 풀면 $\alpha = 4, \beta = 20000$ 을 얻는다. 따라서 감마분포의 모수 θ에 관한 p.d.f. 는 다음과 같다.

$$\xi(\theta) = \begin{cases} \frac{(20000)^4}{3!}\theta^3 e^{-20000\theta} & \theta > 0 \\ 0 & \theta \le 0 \end{cases}$$

▲

> NOTE
> * Notation) $\vec{X} = (X_1, \cdots, X_m)$, $\vec{x} = (x_1, \cdots, x_m)$, $X_1 \mapsto x_i$ ⋯ if $S \to R$ 일 때
> $$f_m(\vec{x}\,|\theta) = f(x_1, \cdots, x_m|\theta) = f(x_1|\theta)\cdots f(x_m|\theta)$$

§. Sensitivity Analysis and Improper Priors

Example 2.3.7) A Clinical Trial

이전 예제들을 참조하자. 환자들이 처방 후 비재발(no relapse) / 재발(relapse) 두 가지 경우로 나뉠 때

$$E_i : i \text{ 번째 환자가 성공할(no relapse) 사건}$$

$$B_j : \text{모든 가능한 환자들 중 성공할 환자의 비율이 } p=\frac{(j-1)}{10} \text{ , } j=1,\cdots,11 \text{ 인 사건}$$

으로 두자. 만약 B_j 가 발생했음을 안다면 $E_1, E_2, \cdots$가 독립적이라고 할 수 있다. 즉, 우리는 그 환자들 값을 각 B_j로 주어지는 조건부독립으로 생각할 수 있고, $\Pr(E_i \mid B_j)=\frac{j-1}{10}$, $\forall\, i,j$ 로 잡을 수 있다.

또한 우리는 trial을 시행하기 이전에(prior) 모든 j에 대하여 $\Pr(B_j)=\frac{1}{11}$ (Prior Probabilities) 임을 가정할 것이다. 이제 우리는 각 환자들이 trial을 마친 후의 사건들 B_j에 대한 사후(posterior)확률들을 계산해서 p에 대해 우리가 알 수 있는 것을 표현하고자 한다.

예를 들면, 첫 번째 환자를 고려하자. 이전 예제들에서, $\Pr(E_1)=1/2$ 이므로, 만약 E_1이 발생하면, 베이즈 정리(Bayes' theorem)를 적용하여 다음을 얻을 수 있다.

$$\Pr(B_j \mid E_1)=\frac{\Pr(E_1 \mid B_j)\Pr(B_j)}{1/2}=\frac{2(j-1)}{10\times 11}=\frac{j-1}{55} \quad \text{(Posterior Probabilities)}$$

그러면 예상했듯이, 하나의 성공확률을 관측한 후에, p값이 크면 사후확률이 사전확률보다 높고 p값이 작으면 사후확률이 사전확률보다 작다는 것을 알 수 있다.

예를 들면, E_1이 이미 발생했으므로, $p=0$인 경우는 배제된다. 따라서 $\Pr(B_1 \mid E_1)=0$. 마찬가지로, $\Pr(B_2 \mid E_1)=0.0182$ 일 때 사전 값(prior value)은 $1/11=0.0909$ 이므로 p가 작을 때 사후확률이 사전확률보다 작음을 알 수 있고, $\Pr(B_{11} \mid E_1)=0.1818$ 의 경우도 마찬가지로 0.0909보다 훨씬 큰 값을 갖는다는 것을 확인할 수 있다.

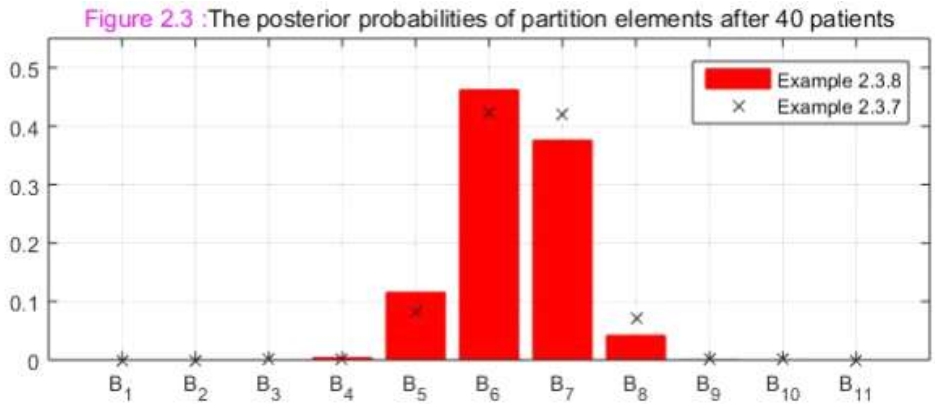
이제 우리는 사후확률이 각 환자가 관측된 후에 어떻게 행동하는지 확인해 보아야 한다.
그러나 Table 2.1(교재 참조) imipramine 열에 관측된 40명의 환자들에만 초점을 둘 것이다.
A가 22명이 성공하고 18명이 실패한 것으로 관측된 사건이라 하자. 40명 중 22명이 성공했고, 각 사건 B_j가 독립이므로

$$\Pr(A \mid B_j)=\binom{40}{22}([j-1]/10)^{22}(1-[j-1]/10)^{18}, \quad \forall\, j.$$

그러면 베이즈 정리(Bayes' theorem)에 의해

$$\Pr(B_j \mid A)=\frac{\frac{1}{11}\binom{40}{22}([j-1]/10)^{22}(1-[j-1]/10)^{18}}{\sum_{i=1}^{11}\frac{1}{11}\binom{40}{22}([i-1]/10)^{22}(1-[i-1]/10)^{18}} \ .$$

이제 $\Pr(B_j|A)$를 구하여 40명의 환자들에 대한 11개 사건의 사후확률의 그래프를 그려보면 다음과 같다.



또한, 우리는 다음 환자가 성공할 사건에 대한 사전확률과 40명 이후의 사후확률을 계산할 수 있다. 사전확률로는 $\Pr(E_{41}) = \Pr(E_1) = 1/2$ 로 두고, 40명의 환자들 이후의 관측 값에 대한, 즉 사후확률은 전확률의 법칙(The Law of Total Probability) 의 조건부 버전을 이용하면 다음과 같다.

$$\Pr(E_{41}|A) = \sum_{j=1}^{11} \Pr(E_{41}|B_j \cap A)\Pr(B_j|A)$$

여기서 B_j가 주어질 때 E_i는 조건부독립이므로, $\Pr(E_{41}|B_j \cap A) = \Pr(E_{41}|B_j) = (j-1)/10$ 이고, 이를 반영하면 구하고자 하는 41번째 환자의 성공확률은 0.5476 임을 구할 수 있다. 이는 관찰된 성공 횟수와 매우 근사하다.

▲

Example 2.3.8) The Effect of Prior Probabilities

예제 2.3.7의 Clinical Trial 사례에서, 또 다른 연구자는 성공확률과 p값에 대해 다른 사전의견(prior opinion)을 갖고 있다고 가정하자. 이 연구자는 다음과 같은 사전확률을 제시한다.

Event	B_1	B_2	B_3	B_4	B_5	B_6	B_7	B_8	B_9	B_{10}	B_{11}
p	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Prior prob.	0.00	0.19	0.19	0.17	0.14	0.11	0.09	0.06	0.04	0.01	0.00

사후 확률은 베이즈 정리를 이용하여 다시 계산하고, 그 값을 그래프에 막대로 표시하고, 2.3.7 예제의 사후 확률은 그래프 상에 X 표시로 표시하였다. (Example 2.3.7, Figure 2.3 참조.)
두 경우, 사전확률간의 차이가 큼에도 불구하고 사후확률은 근접하게 나타나는 것을 볼 수 있다.
즉 40명 이상의 환자 관측값을 갖고 있어야 사전확률과 큰 관계없이 적정한 사후확률을 얻을 수 있음을 의미하게 된다. (사전확률 - train data / 사후확률 - test(검증) data 개념과 유사하게 생각해보자.)
만약 두 경우 모두 관측 값이 적으면, 사후 확률에 대한 두 경우 사이의 차이가 커진다. 이는 관측된 사건들이 더 적은 정보를 제공하기 때문이다.

▲

NOTE

* **베이즈 정리(Bayes' Theorem)**

사건 $B_1,\cdots,B_k$ 가 $\Pr(B_j)>0$ for $j=1,\cdots,k$ 을 만족하는 표본공간 S 의 분할(partition)을 형성하고 사건 A가 $\Pr(A)>0$ 을 만족하면, $j=1,\cdots,k$ 에 대하여

$$\Pr(B_i|A)=\frac{\Pr(B_i)\Pr(A|B_i)}{\sum_{j=1}^{k}\Pr(B_j)\Pr(A|B_j)}$$

* **조건부 독립(Conditional independence)**

사건 $A_1,\cdots,A_k$ 가 B 에 대하여 **조건부 독립(Conditional independent)** 일 필요충분조건은 임의의 집합족 $A_{i_1},\cdots,A_{i_j}$, $j=2,3,\cdots,k$ 에 대하여

$$\Pr(A_{i_1}\cap\cdots\cap A_{i_j}|B)=\Pr(A_{i_1}|B)\cdots\Pr(A_{i_j}|B)$$

* **전확률의 법칙(Law of total probability)**

사건 $B_1,\cdots,B_k$ 가 $\Pr(B_j)>0$ for $j=1,\cdots,k$ 을 만족하는 표본공간 S 의 분할(partition)을 형성하면 임의의 사건 $A\subset S$ 에 대하여,

$$\Pr(A)=\sum_{j=1}^{k}\Pr(B_j)\Pr(A|B_j)$$

* **조건부 확률(Conditional Probability)**

The conditional probability of the event A given that the event B has occurred if and only if

$$\Pr(A|B)=\frac{\Pr(A\cap B)}{\Pr(B)}\ \text{ if }\ \Pr(B)>0$$

* **민감도 분석(Sensitivity Analysis)**

여러 다른 Prior distribution 과 Posterior distribution 을 비교함으로써 우리가 중요시하는 문제에 대한 답을 제시할 수 있는 가장 효율적인 Prior distribution 이 무엇인지 확인하는 과정을 **민감도 분석(Sensitivity-Analysis)** 라고 한다.

*** Matlab Code in Example 2.3.7 ***

```
clear all
clc

% Bayes probability in example 2.3.7
C=nchoosek(40,22);
i=1:11;
a=(1/11)*C*((i-1)/10).^22.*(1-(i-1)/10).^18;
A=sum(a);
B=(1/11)*C*((i-1)/10).^22.*(1-(i-1)/10).^18;
Pr=B/A; %% bayes prob.

%Bayes probability in example 2.3.8
a1_1=(0)*C*((1-1)/10).^22.*(1-(1-1)/10).^18;
a1_2=(0.19)*C*((2-1)/10).^22.*(1-(2-1)/10).^18;
a1_3=(0.19)*C*((3-1)/10).^22.*(1-(3-1)/10).^18;
a1_4=(0.17)*C*((4-1)/10).^22.*(1-(4-1)/10).^18;
a1_5=(0.14)*C*((5-1)/10).^22.*(1-(5-1)/10).^18;
a1_6=(0.11)*C*((6-1)/10).^22.*(1-(6-1)/10).^18;
a1_7=(0.09)*C*((7-1)/10).^22.*(1-(7-1)/10).^18;
a1_8=(0.06)*C*((8-1)/10).^22.*(1-(8-1)/10).^18;
a1_9=(0.04)*C*((9-1)/10).^22.*(1-(9-1)/10).^18;
a1_10=(0.01)*C*((10-1)/10).^22.*(1-(10-1)/10).^18;
a1_11=(0.00)*C*((11-1)/10).^22.*(1-(11-1)/10).^18;
a1=[a1_1 a1_2 a1_3 a1_4 a1_5 a1_6 a1_7 a1_8 a1_9 a1_10 a1_11];
sum_a1=sum(a1);
Pr_1=a1/sum_a1; %% bayes prob.

% Graph
bar(Pr_1,'r','EdgeColor','r')
ylim([0 0.55]);
ax=gca;
ax.YTick=[0:0.1:0.5];
ax.XTickLabel={'B_{1}','B_{2}','B_{3}','B_{4}','B_{5}','B_{6}','B_{7}','B_{8}','B_{9}','B
_{10}','B_{11}'};
title(['\color{magenta}Figure 2.3 :' '\color{black}The posterior probabilities of
partition elements after 40 patients'])
grid on
hold on
plot(Pr,'xk')
legend({'Example 2.3.8','Example 2.3.7'})
hold off
```

Example 7.2.5) Lifetimes of Fluorescent Lamps

Example 7.2.4에서, 모수 θ를 갖는 n개의 형광등의 수명에 관한 모음에서 사전분포를 지수분포로 채택하였다. 이제 우리는 관측된 데이터를 고려하여 θ에 관한 분포를 어떻게 바꿀 것인지 주목하고자 한다.

▲

Definition 7.2.2) Posterior Distribution / p.f / p.d.f

* 관측된 확률변수 $X_1,\cdots,X_n$ 를 갖는 모수 θ를 찾는 통계적 추론 문제를 고려하자. $X_1,\cdots,X_n$ 를 조건부로 갖는 θ에 관한 조건부 분포를 θ에 관한 **사후분포(Posterior Distribution)**라 한다.

* $X_1=x_1,\cdots,X_n=x_n$ 을 조건부로 하는 θ에 관한 p.f. / p.d.f.를 **posterior p.f / p.d.f. of θ** 라 하고 $\xi(\theta|x_1,\cdots,x_n)$ 로 표기한다.

Theorem 7.2.1) 확률변수 $X_1,\cdots,X_n$ 가 $f(x|\theta)$ 를 p.f. / p.d.f. 로 갖는 확률표본을 구성한다고 가정하자. 그리고 모수 θ의 값은 알려져 있지 않고, prior p.f. / p.d.f. of θ가 $\xi(\theta)$라 가정하자. 그러면 posterior p.f / p.d.f. of θ는 다음과 같다.

$$\xi(\theta|\vec{x})=\frac{f(x_1|\theta)\cdots f(x_n|\theta)\xi(\theta)}{g_n(\vec{x})} \quad \text{for } \theta\in\Omega \text{ , where } g_n : \text{marginal joint p.d.f./p.f. of } X_1,\cdots,X_n.$$

Example 7.2.6) Lifetimes of Fluorescent Lamps

Example 7.2.4, 7.2.5에서 가정했던 대로, 특정한 타입의 형광등 수명이 모수 θ를 갖는 지수분포를 따르고, θ에 대한 사전분포는 prior p.d.f. 가 $\xi(\theta)=\begin{cases}\frac{(20000)^4}{3!}\theta^3e^{-20000\theta} & \theta>0\\ 0 & \theta\le 0\end{cases}$ 인 특정한 감마분포를 따른다고 가정하자.

또한, 이러한 유형의 n개의 형광등의 확률 표본으로써 수명 $X_1,\cdots,X_n$ 을 관측하였다고 가정하자. 이제 우리는 $X_1=x_1,\cdots,X_n=x_n$ 을 조건부로 하는 θ에 관한 사후(posterior) p.d.f.를 결정하고자 한다.

우선, 각 관측값 X_i는 지수분포를 따르므로, 각 X_i 의 p.d.f.는 다음과 같다.

$$f(x|\theta)=\begin{cases}\theta e^{-\theta x} & x>0\\ 0 & otherwise\end{cases}$$

그러면 $X_1,\cdots,X_n$ 에 대한 joint p.d.f.는 $x_i>0,\ i=1,\cdots,n$ 에 대하여 다음과 같이 구할 수 있다.

$$f_n(\vec{x}|\theta)=\prod_{i=1}^{n}\theta e^{-\theta x_i}=\theta^n e^{-\theta y}, \quad \text{where } y=\sum_{i=1}^{n}x_i.$$

이제 $\theta>0$ 일 때, 앞서 가정한 사전분포를 곱하면 $f(\vec{x}|\theta)\xi(\theta)=\theta^{n+3}e^{-(y+20000)\theta}$ 임을 얻는다.

베이즈 정리를 이용하여 얻은 Theorem 7.2.1 에 의하면, posterior p.d.f. of θ 를 구하기 위하여 $g_n(\vec{x})$를 구해야 하고, 이는 주변결합확률밀도함수(marginal joint p.d.f)이므로,

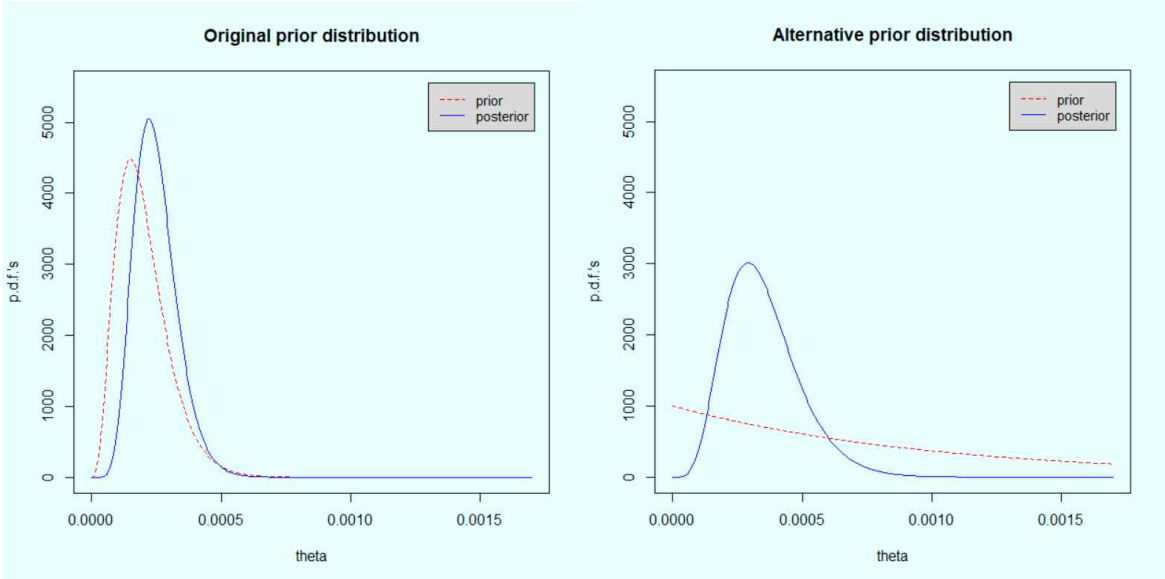
$$g_n(\vec{x}) = \int_0^\infty \theta^{n+3} e^{-(y+20000)\theta} d\theta = \frac{\Gamma(n+4)}{(y+20000)^{n+4}} \quad (\because \text{ [note] } \forall\, \alpha > 0,\ \beta > 0,\ \int_0^\infty x^{\alpha-1} e^{\beta x} dx = \frac{\Gamma(\alpha)}{\beta^\alpha})$$

이고, 최종적으로 Theorem 7.2.1을 적용하면 다음과 같은 posterior p.d.f.를 얻는다.

$$\xi(\theta \mid \vec{x}) = \frac{\theta^{n+3} e^{-(y+20000)\theta}}{\frac{\Gamma(n+4)}{(y+20000)^{n+4}}} = \frac{(y+20000)^{n+4}}{\Gamma(n+4)} e^{-(y+20000)\theta} \ , \text{ for } \theta > 0.$$

이제 구한 posterior p.d.f.를 일반적인 감마분포의 p.d.f.와 비교하면, 이는 파라미터를 $n+4$, $y+20000$ 으로 갖는 감마 분포를 따른다는 것을 쉽게 확인할 수 있다. 따라서 최종적으로 θ의 사후분포는 감마분포가 된다.

이제 이해를 돕기 위해, 특정한 예를 들면, 각 수명(lifetimes)/시간(hour) : 2911, 3403, 3237, 3509, 3118 5개 ($n=5$)를 관찰했다고 하자. 그러면 $y=16178$ 이고, θ에 관한 사후분포는 모수가 9, 36178 인 감마분포를 따른다.



좌측 그래프를 보면, 사전분포와 사후분포로부터 θ의 분포가 어느 정도 조정되었음을 쉽게 확인할 수 있다.
만약 사전분포의 모수 값을 1, 1000으로 주고, 사후분포의 모수 값을 6, 17178로 주어(기존 5개의 관측값을 그대로 활용한다) 그래프를 그려본다면 우측 그래프와 같은 형태가 나타나는데, 이는 좌측 그래프와는 확연히 다른 두 분포의 차이가 나는 것을 볼 수 있다.

여기서 알 수 있는 것은, 이러한 작은 data set에서, 사전분포의 선택에 따라 각기 크고 작은 차이를 보일 수 있다는 것이다.

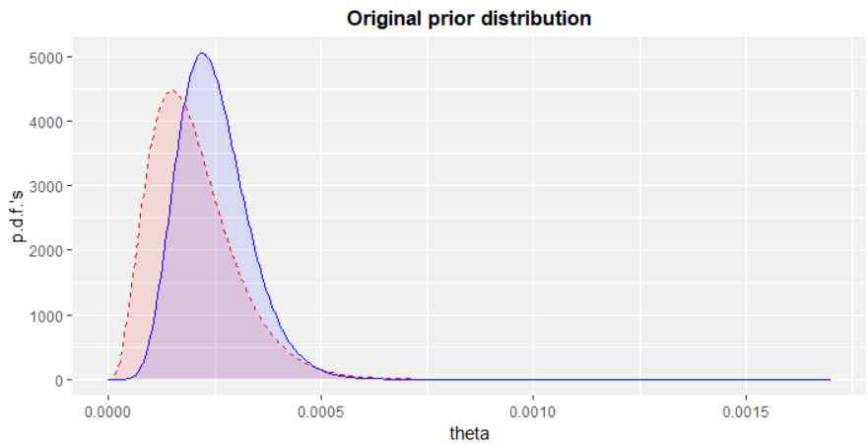
▲

*** R Code in Example 7.2.6 ***

```
theta <- seq(0, 0.0017, length = 500)

par(bg="light cyan")
plot(theta, dgamma(theta,1,1000),
     main="Alternative prior distribution",
     xlab="theta",
     ylab="p.d.f.'s",
     type="l",
     col="Red",
     lty=2,
     ylim=c(0,5500))
points(theta, dgamma(theta,6,17178),type="l",col="Blue",ylim=c(0,5500))
legend(x=0.0013,y=5550,legend=c("prior","posterior"),
       col=c("red","blue"),lty=c(2,1),bg="light grey",cex=0.9)
```

*** R Code with 'ggplot2' package in Example 7.2.6 ***

```
library(ggplot2)

theta <- seq(0, 0.0017, length = 500)
prior<-dgamma(theta,4,20000)
posterior<-dgamma(theta,9,36178)

pdfframe_1<-data.frame(theta,prior,posterior)

graph<-ggplot(data=pdfframe_1)+
       geom_line(aes(x=theta,y=prior), color="red", linetype=2,cex=0.4)+
       geom_line(aes(x=theta,y=posterior), color="blue", linetype=1,cex=0.4)+
       geom_area(aes(x=theta,y=prior), fill="red", alpha=0.1) +
       geom_area(aes(x=theta,y=posterior), fill="blue", alpha=0.1)

graph + xlab('theta') + ylab("p.d.f.'s") + ggtitle('Original prior distribution')+
theme(plot.title = element_text(hjust = 0.5,lineheight=.8,face="bold"))
```



Definition 7.2.3) Likelihood Function

When the joint p.d.f. / joint p.f $f_n(\vec{x}|\theta)$ of the observations in a random sample is regarded as a funtion of θ for given values of $x_1,\cdots,x_n$, it is called the **Likelihood Function(우도함수).**

Example 7.2.8) Lifetimes of Fluorescent Lamps

앞선 예제 7.2.6에서, θ에 조건부인, 형광등의 수명은 모수를 θ로 갖는 독립적인 지수적인 확률 변수로 가정하였다. 또한 5개의 형광등의 수명을 관측하였고, θ에 관한 사후분포는 모수를 9, 36178 로 갖는 감마분포임을 확인하였다.

이제 위의 가정을 전제로 우리는 다음 형광등의 수명 X_6을 예측하고자 한다.

우선, X_6의 5개의 수명 관측값을 조건으로 하는 conditional p.d.f.는 $\xi(\theta|\vec{x})f(x_6|\theta)$를 θ에 관해 적분하는 것과 같으므로, 이를 구하기 위해 먼저 θ의 Posterior p.d.f.를 구하면 $\xi(\theta|\vec{x}) = 2.633\times 10^{36}\theta^8 e^{-36178\theta}$, $\theta > 0$ 이다. 이제 이를 적분하면 다음과 같은 결과를 얻는다.

$$f(x_6|\vec{x}) = \int_0^\infty 2.633\times 10^{36}\theta^8 e^{-36178\theta}\theta e^{-x_6\theta}\,d\theta = 2.633\times 10^{36}\frac{\Gamma(10)}{(x_6+36178)^{10}} = \frac{9.555\times 10^{41}}{(x_6+36178)^{10}}$$

우린 이제 이 p.d.f.를 주어진 관측 수명값에 관한 X_6의 분포의 계산을 하기 위해 이용할 수 있다.
예를 들면, 6번째 형광등의 수명시간이 3000시간 이상일 확률을 구하려면 다음과 같이 하면 된다.

$$\Pr(X_6 > 3000|\vec{x}) = \int_{3000}^\infty \frac{9.555\times 10^{41}}{(x_6+36178)^{10}}\,dx_6 = \frac{9.555\times 10^{41}}{9\times 39178^9} = 0.4882$$

마지막으로, 예제 7.2.6에서 시작했던 민감도 분석(sensitivity analysis)를 계속하고자 한다.
만약 다음 형광등의 수명이 적어도 3000시간 이상일 확률을 아는 것이 중요하다고 하면, 우리는 우리가 이전에 행했던 계산에서 사전분포의 선택이 얼마나 영향을 미치는지 볼 수 있다. 우선 예제 7.2.6의 두 번째 사전분포(모수값을 1, 1000로 갖는 감마분포)를 이용하자. 이 때 사후분포의 모수 값은 6, 17178 이었음을 상기하자.
앞서 구했던 Posterior p.d.f.와 마찬가지 방법으로, 새로 바뀐 모수 값에 대한 Posterior p.d.f.와 그 확률을 구하면

$$f(x_6|\vec{x}) = \frac{1.542\times 10^{26}}{(x_6+17178)^7}, \quad \text{for } x_6 > 0 \ ,$$

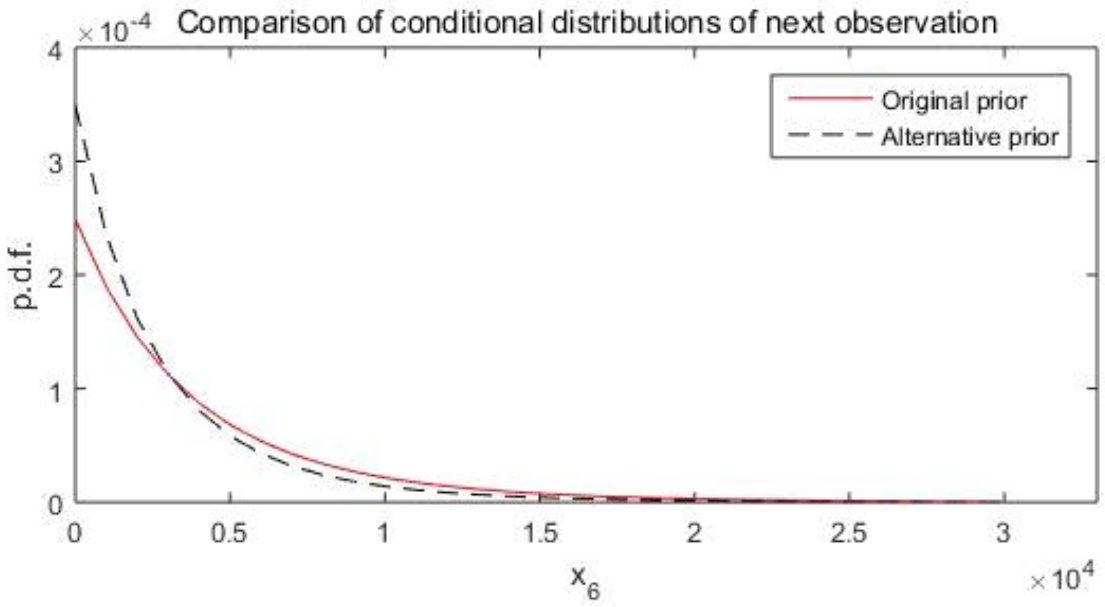
$$\Pr(X_6 > 3000|\vec{x}) = \int_{3000}^\infty \frac{1.542\times 10^{26}}{(x_6+17178)^7}\,dx_6 = 0.3807 \ .$$

예제 7.2.6에서 언급했듯, 다른 사전분포의 선택은 우리가 하고자 하는 추론에 있어 고려할만한 차이를 만든다.
만약 확률 $\Pr(X_6 > 3000|\vec{x})$에 대한 두 분포가 더욱 근사하는 것이 중요하다면, 더 큰 표본이 필요하다.

▲

*** Matlab Code in Example 7.2.8 ***

```
clear all
clc

% observed data
x6=0:1000:30000;

% posterior p.d.f. of original
pri_f=(9.555*10^41)./(x6+36178).^10;
% posterior p.d.f. of alternative
pos_f=(1.542*10^26)./(x6+17178).^7;

% Graph of two possible conditional p.d.f.'s
plot(x6,pri_f,'r')
xlim([0 33000])
ylim([0 0.00040])
hold on
plot(x6,pos_f,'--k')
xlabel('x_6')
ylabel('p.d.f.')
title('Comparison of conditional distributions of next observation')
legend('Original prior','Alternative prior')
hold off
```



§. Additional Topics of Likelihood Function

식 $\xi(\theta|\vec{x}) = \frac{f(x_1|\theta)\cdots f(x_n|\theta)\xi(\theta)}{g_n(\vec{x})}$ for $\theta \in \Omega$ 의 분모 $g_n(\vec{x})$ 는 가능한 모든 θ 값에 대한 분자의 적분과 같으므로, 비록 $g_n(\vec{x})$ 의 적분값이 관측 값 $x_1,\cdots,x_n$ 에 의존한다고 해도, 적분값은 θ에 의존하지 않는다. 또한 이는 위 식의 우변이 θ의 p.d.f. 로 간주되므로, 상수로 다루어질 수 있다. 따라서 식 $\xi(\theta|\vec{x}) = \frac{f(x_1|\theta)\cdots f(x_n|\theta)\xi(\theta)}{g_n(\vec{x})}$ 은 다음과 같은 관계로 대체할 수 있다.

$$\xi(\theta|\vec{x}) \propto f_n(\vec{x}|\theta)\xi(\theta) \quad , \ \xi(\theta) : \text{prior p.d.f. of } \theta, \ f_n(\vec{x}|\theta) : \text{likelihood function.}$$

이 때, 비례기호(Proportionality symbol) $\propto$ 는 좌변이 $x_1,\cdots,x_n$에 의존하고, θ에 의존하지 않는 constant factor를 제외한다면 우변과 같음을 의미한다. 위 식의 두 변이 등식이 성립하도록 하는 적절한 constant factor는 $\int_\Omega \xi(\theta|\vec{x})d\theta = 1$ 인 사실을 이용하면 손쉽게 결정된다. 이는 $\xi(\theta|\vec{x})$ 이 θ의 p.d.f. 이기 때문이다.

NOTE

* **베타 분포(Beta distribution)**

$\alpha, \beta > 0$ 이고 X : 확률변수라 하자. X의 p.d.f. 가

$$f(x|\alpha,\beta) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} & 0 < x < 1 \\ 0 & otherwise \end{cases}$$

이면 확률변수 X는 **모수를 α, β로 하는 베타분포(Beta distribution)**를 가진다고 한다.
이 때, $\alpha = 1$, $\beta = 1$ 인 경우 [0,1] 에서의 **균등분포(Uniform distribution)**가 된다.

* Sequential Observations and Prediction

$n-1$ 개의 관측값이 주어질 때, X_n에 대한 조건부 p.f / p.d.f. 로써, $\xi(\theta|\vec{x}) \propto f(x_n|\theta)\xi(\theta|x_1,\cdots,x_{n-1})$ 인 관계를 가질 때, 비례상수(Proportionality constant) 가 1이면 다음이 성립한다.

$$f(x_n|x_1,\cdots,x_{n-1}) = \int f(x_n|\theta)\xi(\theta|x_1,\cdots,x_{n-1})d\theta$$

이는 수열적으로 정의되는 n 번째 관측값(observation)을 예측(prediction) 하기 위한 연속분포의 함수이다.

Example 7.2.7) Proportion of Defective Items

앞선 예제 7.2.3 의 가정을 다시 이용하자. 대형 제비뽑기에서 결함 품목의 비율 θ가 알려져있지 않고, θ에 대한 사전분포는 구간 $[0,1]$에서의 균등분포(uniform distribution) $\xi(\theta)=\begin{cases}1 & 0<\theta<1\\ 0 & otherwise\end{cases}$ 이다.

그리고 n개의 품목에 대한 확률표본은 뽑기로부터 취해지는 값이고, $X_i=1$ if i : defective / $X_i=0$ otherwise 라고 가정하자. 그러면 $X_1,\cdots,X_n$ 모수 θ를 갖는 베르누이 시행이 된다. 따라서 베르누이 분포의 확률함수(p.f.)의 정의에 의해 각 관측값 X_i에 대한 p.f. of θ는 다음과 같다.

$$f(x|\theta)=\begin{cases}\theta^x(1-\theta)^{1-x} & x=0,1\\ 0 & otherwise\end{cases}$$

또한, $f(x|\theta)$를 이용하면, $y=x_1+\cdots+x_n$ 일 때 $X_1,\cdots,X_n$의 결합확률함수(joint p.f.)를 다음과 같이 구할 수 있다.

$$f_n(\vec{x}|\theta)=\theta^y(1-\theta)^{n-y} \quad , \ x_i=0 \text{ or } 1 \text{ for } i=1,\cdots,n$$

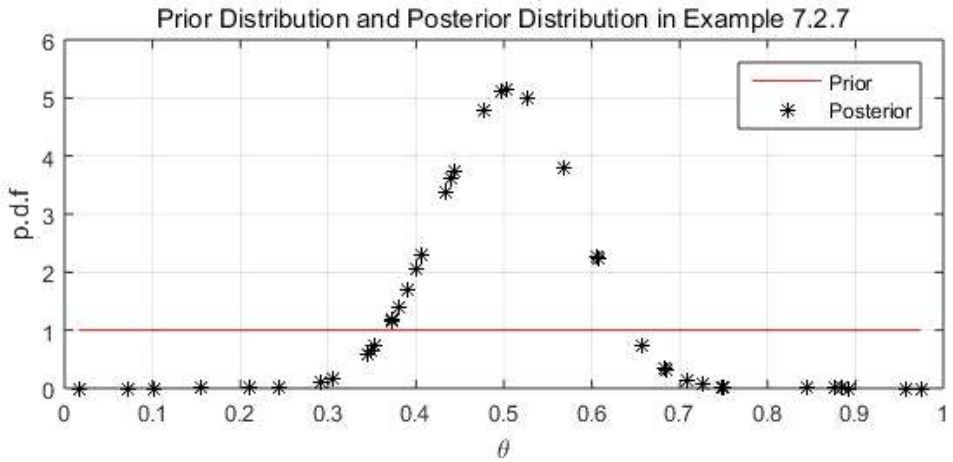
$\xi(\theta)$를 앞서 가정했으므로, 이를 이용하여 $f_n(\vec{x}|\theta)\xi(\theta)=\theta^y(1-\theta)^{n-y}$ for $0<\theta<1$ 임을 얻는다. 여기서 $f_n(\vec{x}|\theta)$는 constant factor를 제외하면 베타 분포의 꼴을 가지므로, 결국 $f_n(\vec{x}|\theta)$ 는 모수 $\alpha=y+1$, $\beta=n-y+1$ 을 갖는 베타분포의 p.d.f. 임을 알 수 있다.

$\xi(\theta|\vec{x}) \propto f_n(\vec{x}|\theta)\xi(\theta)$ 인 관계를 생각하면, 사후분포 $\xi(\theta|\vec{x})$ 는 우변에 비례하므로, $\xi(\theta|\vec{x})$ 또한 반드시 베타분포를 갖는 p.d.f.로 결정된다. 따라서 $0<\theta<1$ 일 때 사후분포는 다음과 같다.

$$\xi(\theta|\vec{x})=\frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)}\theta^y(1-\theta)^{n-y}$$

이 예제에서, 사후분포를 구하기 위해 통계량 $Y=X_1+\cdots+X_n$ 을 이용하기 때문에, 사후분포에 기초를 둔 어떤 추론에도 사용가능하다.

▲

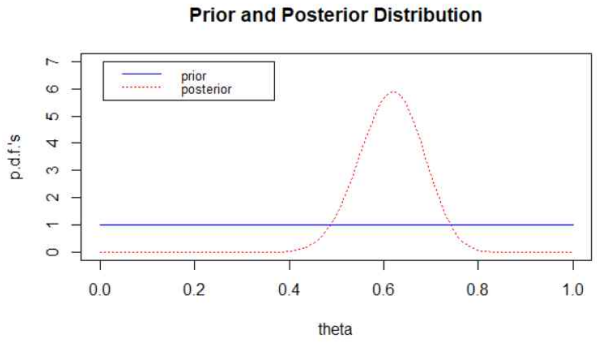


*** Matlab Code in Example 7.2.7 ***

```
clear all
clc

% Observations and some factors
x=randi([0 1],1,40); % random variables (succ or fail)
theta=rand(1,40); % Suppose that theta is unknown parameter
n=length(x);
y=sum(theta);
a=1; b=1;

% Gamma function for prior (integral : numerical integration)
g=@(x,alpha) (x.^(alpha-1)).*exp(-x);
g_pri1=integral(@(x)g(x,a),0,inf);
g_pri2=integral(@(x)g(x,a+b),0,inf);

% Prior distribution as the uniform distribution
pri_f=(g_pri2/(g_pri1^2))*(theta.^(a-1)).*(1-theta).^(b-1); % beta dist. when a=1,b=1

% Gamma function for posterior (integral : numerical integration)
g_pro1=integral(@(x)g(x,y+1),0,inf);
g_pro2=integral(@(x)g(x,n-y+1),0,inf);
g_pro3=integral(@(x)g(x,n+2),0,inf);

% Posterior distribution as the beta distribution
pos_f=(g_pro3/(g_pro1*g_pro2))*(theta.^y).*(1-theta).^(n-y);

% Graph of two distributions : Prior vs Posterior
plot(theta,pri_f,'-r') % prior
hold on
plot(theta,pos_f,'*k') % posterior
title('Prior Distribution and Posterior Distribution in Example 7.2.7')
xlabel('\theta')
ylabel('p.d.f')
legend('Prior','Posterior')
hold off
grid on
```

*** R Code in Example 7.2.7 ***

```
theta<-seq(0,1,length.out=100)
plot(theta,dbeta(theta,32,20),
        main="Prior and Posterior Distribution",
        xlab="theta",
        ylab="p.d.f.'s",
        ylim=c(0,7),
        type="l",col="red",lty=3)
lines(theta,dunif(theta,0,1),type="l",col="blue",lty=1)
legend(x=0.007,y=7,legend=c("prior","posterior"),
                col=c("blue","red"),lty=c(1,3),cex=0.8)
```



7.3 Conjugate Prior Distributions

§. Sampling from a Bernoulli Distribution

Example 7.3.1) A Clinical Trial

예제 5.8.5를 참조하면(page 330), 모든 가능한 환자들 중 successful outcome의 비율 P 는 베타 분포를 따르는 집합으로부터 택한 어떤 분포의 확률변수였다. 이러한 선택은 관측된 데이터를 가지는 P의 조건부 분포 계산을 매우 단순하게 만들어주었다. 사실, 데이터가 주어질 때 P의 이러한 조건부 분포는 베타분포 집합의 또다른 구성원이었음을 암시한다.

▲

Theorem 7.3.1)

$X_1,\cdots,X_n$ 이 알려지지 않은 모수 θ $(0<\theta<1)$를 갖는 베르누이 분포로부터 확률표본을 생성하고,
θ의 사전분포는 모수를 $\alpha>0,\beta>0$ 로 갖는 베타분포를 따른다고 가정하자. 그러면

$X_i=x_i,\ i=1,\cdots,n$ 이 주어진 θ의 사후분포는 모수 $\alpha+\sum_{i=1}^{n}x_i$ 와 $\beta+n-\sum_{i=1}^{n}x_i$ 를 갖는 베타분포를 따른다.

Definition 7.3.1) Conjugate Family / Hyperparameters

Let $X_1,X_2,\cdots$: conditionally $i.i.d$ given θ with common p.f. / p.d.f. $f(x|\theta)$.

Let Ψ : a family of possible distributions over the parameter space Ω.

* Ψ 로부터 택한 사전분포 ζ , 관측한 관찰 값 $\vec{X}=(X_1,\cdots,X_n)$, 관찰 값에 대한 결과값 $\vec{x}=(x_1,\cdots,x_n)$ 에 상관없이 사후분포 $\xi(\theta|\vec{x})$ 가 Ψ의 원소이면, Ψ를 분포가 $f(x|\theta)$인 표본들의 사전분포들의 **Conjugate family(켤레 집합족)** 라고 말한다.

note) 위와 같은 집합족 Ψ는 분포들 $f(x|\theta)$로부터 **표본추출에 닫혀있다(closed under sampling)**고도 한다.

* Ψ의 분포들이 다른 매개변수들에 의해 매개화(parametrized) 되었다면, 사전분포에 대응하는 매개변수들을 **prior hyperparameters** 라고 하고, 사후분포에 대응하는 매개변수들을 **posterior hyperparameters** 라고 한다.

Example 7.3.2) The Variance of the Posterior Beta Distribution.

대형 선박의 선적 결함 품목의 비율 θ가 알려지지 않고, θ의 사전분포가 구간 [0,1]에서 균등분포(uniform distribution)을 따르며, 품목들은 θ에 관한 사후분포의 분산(variance)이 0.01 이하의 값으로 줄여질 때 까지 선적(shipment)에서 무작위로 선택하여 검열되었다고 하자. 우리는 표본 추출과정이 끝나기 전에 결함 / 정상 품목들이 포함된 total number를 결정해야만 한다.

Section 5.8.에서 서술했듯이, [0,1]에서의 균등분포는 모수를 1, 1로 갖는 베타분포를 따른다. 그러므로, y개의 결함 품목들과 z개의 정상 품목들이 포함된 후, θ에 관한 사후분포는 모수 $\alpha = y+1$, $\beta = z+1$을 갖는 베타분포를 따를 것이다.

이는 정리 5.8.3에서, 모수 α와 β를 갖는 베타분포의 분산이 $\alpha\beta/[(\alpha+\beta)^2(\alpha+\beta+1)]$ 임을 확인할 수 있고, 따라서 θ의 사후분포의 분산 V는 다음과 같이 얻어진다.

$$V = \frac{(y+1)(z+1)}{(y+z+2)^2(y+z+3)}$$

표본추출은 결함 품목 y의 수와 정상 품목 z의 수가 $V \le 0.01$ 이 될 때 중지하도록 한다. 그러면 이는, (Exercise 2를 참조) 22개 이상의 품목을 택할 필요는 없지만, 적어도 7개의 품목은 택해야 한다는 결과를 확인할 수 있다.

▲

Example 7.3.3) Glove Use by Nurses.

어떤 연구기관에서 23명의 간호사들에 대해 위생장갑 착용의 중요성에 대한 교육 프로그램 수강 전과 후에 대한 조사를 실시했다. 연구기관은 체액에 접촉할 수도 있는 진료 절차동안 장갑 착용 여부에 대해 기록하였는데, 교육 프로그램 수강 전에는 51개의 진료과정 동안 간호사들을 관찰한 결과, 13명만 장갑을 착용했음을 확인했다.

이제 θ를 간호사들이 교육 프로그램 수강 후 2달 간 장갑을 착용할 확률이라고 하자. 우리는 프로그램 수강 전에 관측된 비율인 13/51 과 θ가 어떻게 비교될 것인지에 초점을 두고자 한다.

우선, 사전분포의 선택이 θ에 관한 사후분포에 얼마나 민감할지 결정하기 위해 θ에 관한 두 개의 다른 사전분포를 생각하자. 첫 번째 사전분포는 구간 [0,1]에서의 균등분포(모수를 1, 1로 갖는 베타분포와 동일)로 하고, 두 번째 사전분포는 모수를 13과 38로 갖는 베타분포로 정하자. 이 사전분포는 첫 번째 분포보다 훨씬 작은 분산을 갖고 평균을 13/51 로 갖는다.
이 때, 두 번째 사전분포를 택하는 연구자는 교육 프로그램이 주목할 만한 효과를 주지 못할 것이라 굳게 주장한다.

교육 프로그램 수강 이후 2달간, 56개의 진료동안 간호사들 중 50명이 장갑을 착용했음이 관찰되었다. 그러면 첫 번째 사전분포에 기초한 θ에 관한 사후분포는 모수를 $1+50=51$, $1+6=7$로 갖는 베타분포를 따르게 된다. 특히, θ의 사후분포의 평균은 $51/(51+7)=0.88$이고, $\theta > 2\times 13/51$ 이 본질적으로 1인 사후확률을 갖게 된다.
두 번째 사전분포에 기초한 사후분포는 모수를 $13+50=63$, $38+6=44$로 갖는 베타분포를 따르게 된다.
이 때 사후분포의 평균은 0.59이고, $\theta > 2\times 13/51$ 일 사후확률은 0.95이다.
결과적으로, 처음에는 누군가가 프로그램에 회의적이었다고 해도, 교육 프로그램은 매우 효과적이었음을 보여준다. 프로그램을 받기 전보다 받은 후에 간호사들이 적어도 두 번은 장갑을 착용할 비율이 아주 높아졌음을 확인할 수 있다.

▲

§. Sampling from a Poisson Distribution

Example 7.3.4) Customer Arrivals

어떤 점포의 점주가 고객 방문율을 알려지지 않은 시간 당 비율 θ를 갖는 Poisson process 로써 모델링하였다. 이 때, 점주는 사전분포를 모수 3과 2를 갖는 감마분포로 결정하였다. X 가 특정한 한 시간 동안 방문한 고객들의 수라고 가정하자. 만약 $X=3$ 으로 관찰되었다고 하면, 점주는 θ의 사후분포를 업데이트하고 싶다.

▲

note) 표본들이 포아송 분포를 따를 때, 감마분포의 집합족(family)이 사전분포들의 conjugate family 가 된다. 이러한 관계는 Theorem 7.3.2에서 확인할 수 있다.

Theorem 7.3.2)

$X_1,\cdots,X_n$ 이 평균이 $\theta>0$인 알려지지 않은 θ를 갖는 포아송 분포로부터 생성된 확률 표본이고, θ의 사전분포가 모수 $\alpha>0,\beta>0$를 갖는 감마분포라고 하자. 그러면

$X_i=x_i,\ i=1,\cdots,n$ 이 주어진 θ의 사후분포는 모수를 $\alpha+\sum_{i=1}^{n}x_i$ 와 $\beta+n$을 갖는 감마분포이다.

Example 7.3.5) Customer Arrivals

앞선 예제 7.3.4에서, Theorem 7.3.2를 적용하면, $n=1,\alpha=3,\beta=2,x_1=3$ 이므로, $X=3$ 으로 주어질 때 θ의 사후분포는 모수 6과 3을 갖는 감마분포가 된다.

▲

Example 7.3.6) The Variance of the Posterior Gamma Distribution

평균 θ가 알려져 있지 않고 θ에 관한 p.d.f.가 다음과 같은 포아송 분포를 생각하자.

$$\xi(\theta)=\begin{cases}2e^{-2\theta} & \theta>0\\ 0 & \theta\le 0\end{cases}$$

또한, 관측 값들은 θ에 대한 사후분포의 분산이 0.01 이하로 줄어들 때까지 주어진 포아송 분포로부터 무작위로 뽑는다고 가정하자. 그러면 우린, 관측 값들의 수를 표본 추출이 멈춰지기 전에 결정해야만 한다.

주어진 prior p.d.f. $\xi(\theta)$가 prior hyperparameters를 $\alpha=1,\beta=2$로 갖는 감마분포의 p.d.f.이므로, n개의 관측 값의 결과로 $x_1,\cdots,x_n$와 그 합인 $y=\sum_{i=1}^{n}x_i$를 얻은 후에, θ에 관한 사후분포는 posterior hyperparameters를 $y+1,n+2$로 갖는 감마분포가 된다. 정리 5.4.2에서 감마분포의 분산이 α/β^2 임을 확인해보면, θ에 관한 사후분포의 분산 V는 다음과 같이 구할 수 있다.

$$V=\frac{y+1}{(n+2)^2}$$

표본추출(Sampling)은 관측된 값 $x_1,\cdots,x_n$의 수열이 $V\le 0.01$이 될 때 중지하도록 한다. 예제 7.3.2와는 달리, 여기엔 얼마나 큰 n이 필요한지에 대한 균일한 경계가 없다. 이는 y가 n이 얼마인지 상관없이 임의로 큰 값이 될 수 있기 때문이다. 분명하게 알 수 있는 것은, $V\le 0.01$ 이기 전에 n은 적어도 $n=8$ 의 관측 값들을 취한다는 것이다.

▲

§. Sampling from a Normal Distribution

Example 7.3.7) Automobile Emissions

예제 5.6.1(page302)에서, 특정 질소 산화물인 자동차 배기가스의 표본추출에 대해 다시 생각해보자.
데이터를 관찰하기에 앞서, 엔지니어가 각 배기가스들의 측정값이 평균을 θ로 갖고, 표준편차를 0.5로 갖는 정규분포를 따르며, θ는 알려져 있지 않다고 믿고 있음을 가정하자. 이러한 θ에 관한 엔지니어의 불확실성은 평균을 2.0, 표준편차를 1.0으로 갖는 또 다른 정규분포에 의해 나타낼 수 있다. Fig.5.1.을 참조한 후에, 어떻게 이 엔지니어가 자신의 θ에 관한 불확실성을 나타낼 수 있을 것인가?

▲

note) 표본들이 알려져 있지 않은 평균 θ와 알려져 있는 분산 σ^2를 갖는 정규분포로부터 취해질 때, 정규분포의 집합족(family)은 그 스스로 다음 정리 7.3.3에서 확인하게 될, 사전분포로써의 conjugate family가 된다.

Theorem 7.3.3)

$X_1,\cdots,X_n$이 알려져 있지 않은 평균 θ와 알려져 있는 분산 $\sigma^2>0$를 갖는 정규분포로부터 확률 표본을 생성하고,
θ에 관한 사전분포가 $N(\mu_0,v_0^2)$ 이면 $X_i=x_i,\ i=1,\cdots,n$ 이 주어진 θ의 사후분포는 정규분포이고, 평균 μ_1, 분산 v_1^2 을 다음과 같이 갖는다.

$$\mu_1=\frac{\sigma^2\mu_0+nv_0^2\overline{x_n}}{\sigma^2+nv_0^2}\ ,\ v_1^2=\frac{\sigma^2v_0^2}{\sigma^2+nv_0^2}$$

Example 7.3.8) Automobile Emissions

이제 정리 7.3.3을 이용하면, 우리는 예제 7.3.7에 대한 답변을 할 수 있다. 정리 7.3.3의 식을 이용하면, $n=46,\sigma^2=(0.5)^2=0.25,\mu_0=2,v^2=1.0$ 임을 얻는다. 46개의 측정값들에 대한 평균은 $\overline{x_n}=1.329$로 얻어진다.
위의 구한 값들을 이용하면, θ에 관한 사후분포는 다음의 평균과 분산을 갖는 정규분포를 따르게 된다.

$$\mu_1=\frac{0.25\times2+46\times1\times1.329}{0.25+46\times1}=1.333,\ v_1^2=\frac{0.25\times1}{0.25+46\times1}=0.0054$$

▲

Example 7.3.9) The Variance of the Posterior Normal Distribution.

관측 값들이 평균 θ, 분산 1 로 주어지는 정규분포로부터 무작위 추출된다고 하자. 이 때, θ는 알려져 있지 않다. θ에 대한 사전분포가 분산을 4로 갖는 정규분포라고 가정하자. 관측 값들은 사후분포의 분산이 0.01 이하로 줄어들 때까지 계속하여 추출하기로 하자. 그러면 우린 이제 관측 값들의 수를 샘플링이 멈춰지기 전에 결정해야만 한다.
정리 7.3.3을 이용하자. n개의 관측 값들이 뽑히고 난 후, θ의 사후분포의 분산 v_1^2은 다음과 같이 구할 수 있다.

$$v_1^2=\frac{4}{4n+1}$$

그러면, 분산이 $v_1\le0.01$ 을 만족할 필요충분조건은 $n\ge99.75$ 임을 얻는다. 이런 이유로, $v_1\le0.01$을 만족하려면 적어도 100개의 관찰 값들이 취해져야만 한다.

▲

Example 7.3.10) Calorie Counts on Food Labels.

4명의 연구자가 전국적으로 20개의 수입 식품들을 뽑아 식품 라벨에 적힌 그램(g)당 칼로리 성분과 실험실에서 결정되는 칼로리 성분을 비교하였다. Figure 7.4 는 관측된 실험실에서의 칼로리 측정값과 식품 라벨에 표기된 칼로리 성분과의 비율 차이에 대한 히스토그램을 나타낸다. 우리는 평균 θ, 분산 100을 갖는 정규분포의 모수 θ가 주어질 때 조건부 분포의 차이를 모델링한다고 가정하자. (물론, 이 예제에서는 분산은 알려져 있다고 가정하고, Section 8.6에서 평균과 분산 두 경우 다 알려지지 않은 확률변수로 취급하는 문제에 대해 다룰 것이다.) 우리는 평균이 0이고 분산이 60인 정규분포를 사전분포로 사용할 것이다. 데이터 $\vec{X}$ 는 평균이 0.125인 Fig.7.4의 20개의 differences의 모음으로 구성한다. 그러면 θ의 사후분포는 다음과 같은 평균과 분산을 갖는 정규분포가 된다.

$$\mu_1 = \frac{100\times 0 + 20\times 60\times 0.125}{100+20\times 60} = 0.1154 \ , \ v_1^2 = \frac{100\times 60}{100+20\times 60} = 4.62 \ .$$

예를 들면, 우리는 식품을 포장하는 관리자가 시스템적으로 식품의 칼로리를 적어도 1% 줄여서 포장하는지에 관심이 있다고 하자. 이러한 관심은 $\theta > 1$ 인 것과 대응한다. Theorem 5.6.6을 이용하면, 다음을 얻는다.

$$\Pr(\theta > 1|\vec{x}) = 1 - \Phi\left(\frac{1-0.1154}{\sqrt{4.62}}\right) = 1 - \Phi(1.12) = 0.3403$$

이 결과는 무시할 수 없지만, 압도적이진 않은 결과를 나타내므로, 포장 관리자들이 식품 라벨의 칼로리를 적어도 1% 줄일 수 있는 가능성이 있다고 판단된다.

▲

*** Matlab Code in Example 7.3.10 ***

```
clear all
clc

X=[-29 -15 -13 -8 -5 -3 -1 -7 -6 1 3 1 4 2 5
5 7 13 19];
hist(X,5)
axis([-31,21,0,9])
title(['\color{red}Figure 7.4 : '
'\color{black}Histogram of percentage
differences in Example 7.3.10'])
h = findobj(gca,'Type','patch');
h.FaceColor = [1 0.7 0.7];
h.EdgeColor = 'w';
xlabel('\itLaboratory calories - label
calories')
ylabel('\itNumber of foods')
```

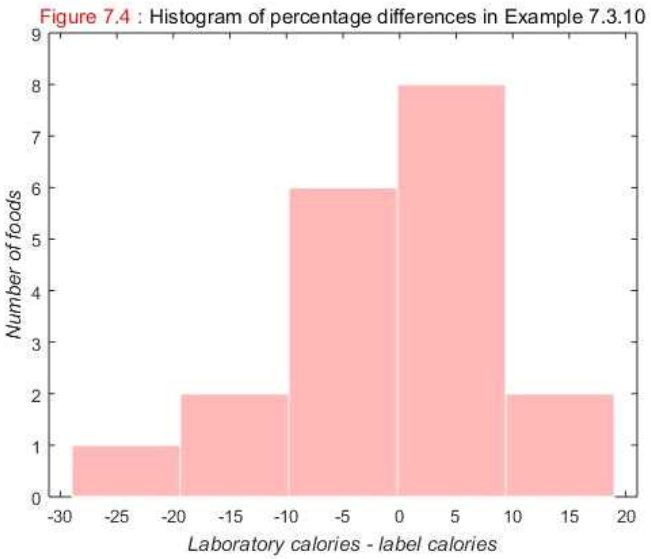


Figure 7.4. Histogram of percentage differences between observed and advertised calories in Example 7.3.10.

§. Sampling from an Exponential Distribution

Example 7.3.11) Lifetimes of Electronic Compoenets.

예제 7.2.1에서, 3개의 수명 값 $X_1 = 3, X_2 = 1.5, X_3 = 2.1$ 을 관측하였다. 이는 주어진 θ에 대하여 $i.i.d$ 한 지수적 확률변수로써 모형화 하였다. 또한 우리의 θ에 관한 사전분포는 모수를 1과 2로 갖는 감마분포였음을 상기하자. 그러면 관측된 세 개의 수명값이 주어질 때 θ의 사후분포는 무엇인가?

▲

note) 알려지지 않은 모수 θ의 값에 대한 지수분포로부터 표본추출을 할 때, 감마분포들의 집합족은 다음 정리에 보여지는 것과 같이 사전분포의 conjugate family 로 사용된다.

Theorem 7.3.4)

$X_1, \cdots, X_n$이 알려져 있지 않은 모수 $\theta > 0$ 를 갖는 지수분포로부터 확률 표본을 생성하고, θ에 관한 사전분포가 모수 $\alpha, \beta > 0$ 를 갖는 감마분포이면 $X_i = x_i,\ i = 1, \cdots, n$ 이 주어진 θ의 사후분포는 모수를 $\alpha + n$ 과 $\beta + \sum_{i=1}^{n} x_i$ 로 갖는 감마분포가 된다.

Example 7.3.12) Lifetimes of Electronic Compoenets.

예제 7.3.11에서, 사후분포를 찾기 위해 Theorem7.3.4를 적용할 수 있다. Theorem7.3.4의 내용과 증명을 이용하면, $n = 3, \alpha = 1, \beta = 2$ 임을 얻고, $y = \sum_{i=1}^{n} x_i = 3 + 1.5 + 2.1 = 6.6$ 임을 얻을 수 있다.

따라서 θ에 관한 사후분포는 모수를 $\alpha = 1 + 3 = 4$ 와 $\beta = 2 + 6.6 = 8.6$ 으로 갖는 감마분포가 된다.

▲

note) Theorem 7.3.4가 예제 7.2.6의 사후분포를 도출하는 과정을 매우 짧아지게 해준다는 점을 알아두자.

§. Improper Prior Distribution

Example 7.3.13) A Clinical Trial.

이 예제에서 우리가 설명하는 것은, 모수를 θ로 갖는 베르누이 분포로부터 θ를 조건부로 $i.i.d$ 한 표본을 포함하는 모든 예제에 적용될 수 있다.

예제 2.1.4의 imipramine group 의 피험자들을 생각하자. 여기서 imipramine을 처방받은 모든 환자들 사이의 성공에 대한 비율 P를 이 chapter에서의 일반적인 표기법인 θ로 표기하도록 하자. 또한 θ는 모수 α, β를 갖는 general conjugate prior 인 베타분포를 가진다고 하자.
imipramine group에서, $n=40$의 환자들이 있고, 그 중 22명이 성공하였다. 그러면 정리 7.3.1에 의해 θ의 사후분포는 모수 $\alpha+22, \beta+18$을 갖는 베타분포가 된다. 사후분포의 평균(기댓값)은 $\frac{\alpha+22}{\alpha+\beta+40}$ 임을 기억하자.

만약 α, β가 작으면, 사후분포의 평균은 정확히 관측된 성공의 비율인 22/40에 가까워진다. 사실, 실제 베타분포와는 일치하지 않게 되지만, $\alpha=\beta=0$ 로 두면 사후 평균은 정확히 22/40 이 된다. 그러나 우린 α, β가 0에 가까워질 때 어떤 일이 일어나는 지 살펴보기로 한다.

(constant factor를 무시한) 베타 p.d.f.는 $\theta^{\alpha-1}(1-\theta)^{\beta-1}$ 으로 구했었다.
또한 우린 $\alpha=\beta=0$으로 두고 $\xi(\theta) \propto \theta^{-1}(1-\theta)^{-1}$ 을 θ의 prior p.d.f. 로 가정할 수 있다. 우도함수(likelihood function)는 $f_{40}(\vec{x}|\theta)=\binom{40}{22}\theta^{22}(1-\theta)^{18}$ 로 구할 수 있다. 우린 constant factor $\binom{40}{22}$ 를 무시하여, 다음의 결과를 구할 수 있다.

$$\xi(\theta|\vec{x}) \propto \theta^{21}(1-\theta)^{17} \text{ , for } 0<\theta<1$$

이것은 constant factor를 제외하면 모수 22와 18을 갖는 베타분포의 p.d.f.와 동일함을 쉽게 인식된다.
따라서, 만약 우리가 prior hyperparameters로 0과 0을 갖는 improper한 "beta distribution" prior를 사용한다면, hyper parameters 로 22와 18을 갖는 θ에 관한 베타 사후분포를 얻을 수 있다. Theorem 7.3.1은 부적절한 이전의 경우에서도 정확한 사후분포를 산출한다는 것에 주목하자.

▲

Definition 7.3.2) Improper Prior

ζ 를 음이 아닌 통계적 모델의 모수공간을 포함하는 정의역을 갖는 함수라고 하고, $\int \zeta(\theta)d\theta=\infty$ 라 가정하자.
만약 이러한 $\zeta(\theta)$를 θ의 사전 p.d.f. 로 가장할 수 있다면, **θ에 대한 improper prior**를 사용하고 있다고 한다.

Example 7.3.14) Prussian Army Deaths.

보르트키에비치는 20년동안 총 280명인 14개 군 부대에서 말의 발길질에 의해 사망한 프로이센 군인들의 수를 세었다. 280개의 카운팅은 다음과 같은 값들을 가진다. : 144카운트=0, 91카운트=1, 32카운트=2, 11카운트=3, 2카운트=4. 어떤 부대도 임의의 1년동안 말발길질에 의해 사망한 수가 4명을 넘지 않았다.

우리는 280개의 카운트를 평균이 θ으로 모수 θ에 조건적인 포아송 확률변수 $X_1, \cdots, X_{280}$를 확률표본으로 하는 모델을 모형화한다고 가정하자. 켤레사전분포는 사전 초모수 α, β를 갖는 감마분포 집합족의 원소가 된다. 정리 7.3.2는 θ에 대한 사후분포가 사후 초모수 $\alpha+196, \beta+280(\because \Sigma 280 counts = 196)$을 갖는 감마분포가 됨을 보여준다. α 또는 β가 매우 큼에도 불구하고, 사후 감마분포는 사후 초모수로 196, 280을 갖는 감마분포와 거의 같다. 이 사후분포는 사전 초모수로 0과0을 갖는 켤레사전분포를 이용한 결과로 볼 수 있다.

constant factor를 무시하고, 모수 α, β를 갖는 감마분포의 p.d.f. 는 $\theta^{\alpha-1}e^{\beta\theta}$ for $\theta > 0$ 로 구할 수 있다. 이 때, 만약 $\alpha = 0, \beta = 0$로 둔다면, 우리는 improper prior p.d.f. $\zeta(\theta)\theta^{-1}$ for $\theta > 0$을 얻는다. 만약 이를 사전 p.d.f.로 가장하고, 확률변수에서의 베이즈 정리를 적용하게 된다면, 다음과 같은 결과를 얻을 수 있다.

$$\zeta(\theta \mid \vec{x}) \propto \theta^{195} e^{-280\theta} \text{ for } \theta > 0$$

이것은 모수를 196과 280으로 갖는 감마분포의 p.d.f. 가 됨을 쉽게 받아들일 수 있다. 이 예제에서의 이러한 결과는 우리가 데이터를 포아송분포로 모형화하는 모든 경우에 적용할 수 있다. 초모수를 0,0으로 갖는 improper "gamma distribution" 은 정리 7.3.2에서 이용될 수 있고, 그 결과는 여전히 성립한다.

▲

Example 7.3.15) Failure Times of Ball Bearings.

예제 5.6.9로부터, 평균 θ, 분산 0.25를 갖는 정규 확률변수 $X_1, \cdots, X_{23}$ 인 볼 베어링의 failure times 에대한 23개의 logarithms을 모형화한다고 가정하자. θ에 대한 켤레사전분포는 평균 μ_0, 분산 v_0^2를 갖는 정규분포가 될 것이다. 23개의 log-failure times 의 평균은 4.15 이므로, θ에 대한 사후분포는 다음과 같은 평균과 분산을 갖는 정규분포가 될 것이다.

$$\mu_1 = \frac{0.25\mu_0 + 23 \times 4.15 v_0^2}{0.25 + 23 v_0^2} \quad , \quad v_1^2 = \frac{0.25 v_0^2}{0.25 + 23 v_0^2}$$

만약 μ_1과 v_1^2에 대한 저 식에서 $v_0^2 \to \infty$라고 하면, 우리는 $\mu_1 \to 4.15, v_1^2 \to \frac{0.25}{23}$ 임을 얻는다. θ의 사전분포에서 무한 값을 갖는 분산은 θ가 똑같이 실수축 어디에나 있다고 할 수 있다. 이와 같은 일은 데이터 $X_1, \cdots, X_n$를 평균 θ와 알려진 분산 σ^2을 조건으로 하는 정규분포로부터의 확률표본으로 모형화하는 모든 예제에 대해 발생할 수 있다.

만약 우리가 분산이 ∞인(사전 평균은 문제가 되지 않는다) improper "normal distribution" 사전분포를 사용한다면, 정리 7.3.3의 계산은 평균 $\overline{x_n}$과 분산 σ^2/n을 갖는 정규분포인 사후분포를 산출할 것이다. 이 경우 그러한 improper prior p.d.f.는 $\zeta(\theta)$가 상수와 같아진다.

이 예제에서, 정의 7.3.2 를 켤레사전분포를 좀 더 편리한 초모수 : $u_0 = \frac{1}{v_0^2}$, $t_0 = \frac{\mu_0}{v_0^2}$ 의 관점에서 설명했다.

이러한 초모수들의 관점에서, 물론 사후분포는 $u_1 = u_0 + \frac{n}{0.25}, t_1 = \frac{\mu_1}{v_1^2} = t_0 + 23 \times \frac{4.15}{0.25}$ 를 갖게 된다. 각 u_1, t_1은 통계량이 더해진 사전 초모수와 대응하는 형태를 갖는다. improper prior가 $u_1 = 0, t_1 = 0$ 일 때, 마찬가지로 $\zeta(\theta)$는 상수와 같아진다.

▲

Example 7.3.16) Very Rare Events.

예제 5.4.7에서, 우리는 일반적으로 매우 낮은 농도에서 일어나는 cryptosporidium이란 식수오염물질에 대해 논하였다. 수도청에서 리터 당 oocysts의 비율 θ 를 갖는 포아송 프로세스로, 상수도관의 cryptosporidium 의 oocysts (접합자낭)을 모형화한다고 가정하자. 그들은 θ에 대해 알기 위해 물에서 25리터의 표본을 정한다.

또한 그들은 p.d.f.를 θ^{-1}로 갖는 improper gamma prior를 사용한다고 가정하자(이는 예제 7.3.14 와 같은 improper prior 임을 상기하자). 만약 25리터의 표본이 oocysts를 포함하지 않는다면, 수도청은 모수를 0과 5를 갖는 감마분포로 θ에 대한 실질적이 아닌 사후분포 유도하게 될 것이다.
얼마나 많은 양의 리터를 표본추출 하는지에 상관없이, 적어도 하나의 oocyst가 관측될 때까지 사후분포는 실제 분포가 될 수 없다.

이처럼 희귀한 사건에 대해 샘플링을 할 때, 사후분포에 기초한 추론을 만들 수 있기 위한 적절한 사전분포의 구성은 사전정보를 수량화하는 것이 강제된다.

▲

7.4 Bayes Estimators

Example 7.4.1) Calorie Counts on Food Labels.

예제 7.3.10에서, 우리가 θ에 관한 사후분포를 구하였고, θ는 홍보한 칼로리 지수와 측정된 칼로리 지수 간의 백분율 차이의 평균이었다. 소비자 집단은 θ의 전체적인 분포를 지정하지 않고, 추정치 θ가 한자리 수로 보고되길 바랄 것이다.(즉, 홍보 칼로리와 측정된 칼로리의 차이가 거의 없기를 원한다.) 그렇다면, 일반적으로 어떻게 이러한 한자리 수의 추정치를 얻는가는 앞으로 소개 될 7.4장에서 제시할 수 있을 것이다.

▲

Definition 7.4.1) Estimator / Estimate

$X_1,...,X_n$이 실직선 Ω의 부분집합에서 취해지는 모수 θ에 의해 정해지는 결합분포의 관측가능한 데이터라 하자. 그러면 θ의 **Estimator** 는 실함수 $\delta(X_1,...,X_n)$이고, $X_1=x_1,...,X_n=x_n$으로 관측될 때, $\delta(x_1,...,x_n)$를 θ의 **Estimate** 라 한다.

note) Definition 7.4.1에서, estimator와 estimate를 구분하였는데, 이는 estimator $\delta(X_1,...,X_n)$가 확률변수 $X_1,...,X_n$에 관한 함수로, estimator 그 자체로 확률변수이며 확률분포 또한 $X_1,...,X_n$의 결합분포로부터 유도되기 때문이다. 한편으로, estimate는 estimator의 특정한 값 $\delta(x_1,...,x_n)$으로, 특정한 관측값 $x_1,...,x_n$에 의해 결정된다. 만약 벡터 표기법으로 $\vec{X}=(X_1,...,X_n)$, $\vec{x}=(x_1,...,x_n)$으로 두면, estimator는 확률벡터 $\vec{X}$에 대한 함수 $\delta(\vec{X})$가 되며, 마찬가지의 원리로 estimate도 정해진다. 이는 이제부터 표기를 간단히 하기 위해 δ로 사용하기로 하자.

Example 7.4.2) Calorie Counts on Food Labels.

예제 7.4.1에서, 소비자 집단이 참값인 평균차이 θ에 대한 estimate $\delta(\vec{x})$가 잘 추정되지 않을 것이라고 생각할 수도 있다. 이상적으로, 그들은 이러한 부정적 영향을 estimate $\delta(\vec{x})$와 θ에 대한 함수로써 정량화하길 원한다. 만약 정량화된다면 그들은 그들이 구한 추정치 $\delta(\vec{x})$에 대한 다양한 결과들에 대한 혼란을 줄일 수 있을 것이다.

▲

note) 좋은 estimator δ를 구하기 위해 우선적으로 요구되는 것은 θ의 estimate가 θ의 실제 값과 얼마나 가까운지에 달려있다. 바꾸어 말하면, 좋은 estimator는 오차 $\delta(\vec{X})-\theta$가 0에 가까워질 가능성이 높은 하나를 택하는 것이다. 우리는 $\theta\in\Omega$의 각각 가능한 값들과 가능한 estimate a를 가정할 것이다. 또한 estmate a와 모수 θ에 대한 손실을 측정하는 값인 $L(\theta,a)$을 정의할 것이다. 이는 a와 θ 사이의 거리가 클수록, $L(\theta,a)$의 값이 커지게 된다.

Definition 7.4.2) Loss Function

손실함수(Loss function)는 두 변수 $\theta\in\Omega, a\in R$에 대한 실함수(real-valued function)을 말한다.
이 때, θ는 모수(parameter), a가 estimate인 경우를 일컫는다.

*** 손실함수의 기대손실(Expected loss) : $E[L(\theta,a)]=\int_{\Omega} L(\theta,a)\xi(\theta)d\theta.$

이 때, 통계학자들은 위의 기대손실이 최소가 되게하는 estimate a를 택하기를 원한다고 가정한다.

이제 통계학자들이 θ를 추정하기 전에 확률벡터 $\vec{X}$의 관측값 $\vec{x}$를 관측할 수 있다고 가정하고, $\xi(\theta|\vec{x})$를 $\theta\in\Omega$의 사후 p.d.f. 라고 하자(이산형인 모수의 경우도 마찬가지의 흐름을 따른다). 각 estimate a에 대하여, 기대손실은

$$E[L(\theta,a)]=\int_{\Omega} L(\theta,a)\xi(\theta)d\theta.$$

이 때 기대손실이 최소가 되도록 하는 a를 택하고자 한다. 이제 각 가능한 $\vec{x}$에 대하여, $\delta^*(\vec{x})$ 를 기대손실이 최소가 되게 하는 estimate a의 값으로 두자. 그러면 함수 $\delta^*(\vec{X})$는 θ의 estimaotr가 된다. 이제 정의를 살펴보자.

Definition 7.4.3) Bayes Estimator / Estimate

$L(\theta,a)$를 손실함수라고 하자. 그리고 $\vec{X}$의 각 가능한 $\vec{x}$에 대하여, $\delta^*(\vec{x})$를 $E[L(\theta,a)|\vec{x}]$가 최소가 되게 하는 a의 값이라 하자. 그러면 δ^*를 θ의 **Bayes estimator**라고 한다. 또한 $\vec{X}=\vec{x}$로 관측될 때, $\delta^*(\vec{x})$를 θ의 **Bayes estimate**라고 한다.

이러한 Bayes estimator를 다르게 표현하면, $E[L(\theta,a)|\vec{x}]=\min_{\forall a}E[L(\theta,a)|\vec{x}]$ 이다.

§. Different Loss Functions

이제, 다양한 손실함수에 대하여 살펴보고자 한다. 추정에 관한 문제에서 가장 흔히 쓰이는 손실함수는 제곱오차 손실함수이다. 이를 다음과 같이 정의한다.

Definition 7.4.4) Squared Error Loss Function : $L(\theta,a)=(\theta-a)^2$

Corollary 7.4.1) θ가 real-valued parameter, 손실함수가 $L(\theta,a)=(\theta-a)^2$, θ의 사후평균(기대값) $E(\theta|\vec{X})$이 유한(finite)하면 θ의 Bayes estimator 는 $\delta^*(\vec{X})=E(\theta|\vec{X})$이다.

Example 7.4.3) Estimating the Parameter of a Bernoulli Distribution.

확률 표본 $X_1,...,X_n$이 모수가 θ인 베르누이 분포로부터 취해진다고 하자. 이 때 θ는 알려지지 않고 반드시 추정되어야 하는 모수이다. 또한 θ의 사전분포는 모수가 $\alpha>0,\beta>0$인 베타분포라고 하자. 이제 손실함수를 제곱오차 손실함수 $L(\theta,a)=(\theta-a)^2$, $0<\theta<1, 0<a<1$ 로 사용한다고 가정하자. 우리는 θ의 Bayes estimator를 결정하고자 한다.

관측값 $x_1,...,x_n$에 대하여, $y=\sum_{i=1}^{n}x_i$라고 두자. 그러면 이는 θ의 사후분포가 모수 $\alpha_1=\alpha+y, \beta_1=\beta+n-y$를 갖는 베타분포를 따른다는 정리 7.3.1의 내용을 적용할 수 있다. 베타분포의 기댓값(평균)은 $\alpha_1/(\alpha_1+\beta_1)$이므로, θ의 사후분포는 $(\alpha+y)/(\alpha+\beta+n)$이다. 그러면 이제 Bayes estimate $\delta(\vec{x})$는 따름정리 7.4.1에 의하면 관측된 벡터 $\vec{x}$의 값들과 같아지게 된다. 그러므로 Bayes estimator $\delta^*(\vec{X})=(\alpha+\Sigma_{i=1}^{n}X_i)/(\alpha+\beta+n)$ 으로 구해진다.

▲

Example 7.4.4) Estimating the Mean of a Normal Distribution.

확률 표본 $X_1,...,X_n$이 평균이 θ인 정규분포로부터 취해진다고 하자. 이 때 θ는 알려지지 않고 반드시 추정되어야 하는 모수이며 분산 σ^2는 알려져 있다고 하자. 또한 θ의 사전분포는 모수가 μ_0, v_0^2인 정규분포라고 하자. 이제 손실함수를 제곱오차 손실함수 $L(\theta,a)=(\theta-a)^2$, $-\infty<\theta<\infty, -\infty<a<\infty$ 로 사용한다고 가정하자. 우리는 θ의 Bayes estimator를 결정하고자 한다.

정리 7.3.3.에 의하면, 앞선 예제와 마찬가지로, 모든 관측 가능한 값 $x_1,...,x_n$에 대하여, θ의 사후분포는 평균이 $\mu_1=\frac{\sigma^2\mu_0+nv_0^2\overline{x}_n}{\sigma^2+nv_0^2}$ 인 정규분포를 따른다. 그러므로, Bayes estimator $\delta^*(\overrightarrow{X})$는 다음과 같이 구해진다.

$$\delta^*(\overrightarrow{X})=\frac{\sigma^2\mu_0+nv_0^2\overline{X}_n}{\sigma^2+nv_0^2}$$

θ에 관한 사후분산은 이 계산에 포함되지 않는다.

▲

Definition 7.4.5) Absolute Error Loss Function : $L(\theta,a)=|\theta-a|$

Corollary 7.4.1) θ가 real-valued parameter, 손실함수가 $L(\theta,a)=|\theta-a|$이면 θ의 Bayes estimator $\delta^*(\overrightarrow{X})$는 θ에 관한 사후분포의 중앙값(median)과 같다.

Example 7.4.5) Estimating the Parameter of a Bernoulli Distribution.

앞선 예제 7.4.3의 가정을 다시 이용하자. 이 때, 손실함수를 절대오차 손실함수 $L(\theta,a)=|\theta-a|$로 사용하자. 따름정리 7.4.1에 의하면, Bayes estimator $\delta^*(\overrightarrow{X})$는 사후분포가 모수 $\alpha_1=\alpha+y, \beta_1=\beta+n-y$를 갖는 베타분포의 중앙값(median)과 같게 된다. 이 때, 중앙값에 관한 명료한 수식이 없으므로 이는 수치적인 근사로 결정된다. 이는 통계 프로그래밍을 통하여 임의의 베타분포의 중앙값을 구할 수 있다.

▲

Example 7.4.6) Estimating the Mean of a Normal Distribution.

마찬가지로, 앞선 예제 7.4.4의 가정을 다시 이용하자. 이 때, 손실함수를 절대오차 손실함수 $L(\theta,a)=|\theta-a|$로 사용하자. 따름정리 7.4.1에 의하면, Bayes estimator $\delta^*(\overrightarrow{X})$는 사후분포가 모수 θ를 갖는 정규분포의 중앙값과 같다. 하지만 각 정규분포의 중앙값과 평균은 같으므로, estimate $\delta^*(\overrightarrow{x})$또한 사후분포의 평균과 같다. 그러므로, 절대오차 손실함수인 경우의 Bayes estimator는 제곱오차 손실함수를 사용했을 때의 Bayes estimator와 같아지게 된다. 이 때의 Bayes estimator 는 다음과 같이 구해진다.

$$\delta^*(\overrightarrow{X})=\frac{\sigma^2\mu_0+nv_0^2\overline{X}_n}{\sigma^2+nv_0^2}$$

▲

**이외에도 여러 손실함수의 형태들이 존재하지만, 그러한 손실함수들은 특정한 추정 문제에서와 연관이 있으므로, 이 장에서는 제곱오차와 절대오차 손실함수에 대해서만 다룬다.

Definition 7.4.6) Consistent Estimator

estimator들의 수열이 $n\to\infty$일 때 알려지지 않은 모수로 확률적으로 수렴하면 그 수열을 **consistent sequence of estimators** 라 한다.

note) 예를 들어, $X_1,...,X_n$이 주어진 θ에 대한 베르누이 분포의 확률 표본이라 하자. 또한 θ의 켤레(conjugate) 사전분포를 이용하자. 그러면 θ가 그러한 켤레분포의 확률 표본에 의해 취해지는 평균이므로, 큰 수의 법칙에 의하면 $n\to\infty$일 때 $\overline{X}_n$ 은 확률적으로 θ에 수렴한다. 그러면 $\delta^*(\vec{X})$와 $\overline{X}_n$ 사이의 차이가 $n\to\infty$일 때 확률적으로 0으로 수렴하므로, 결국 $\delta^*(\vec{X})$ 가 $n\to\infty$일 때 알려지지 않은 값 θ로 확률적으로 수렴한다고 결론지을 수 있다.

Definition 7.4.7) Estimator / Estimate (in general)

$X_1,...,X_n$이 k차원 공간 Ω의 부분집합에서 취해지는 모수 θ에 의해 정해지는 결합분포의 관측가능한 데이터라 하자. 또한 h를 d차원 공간 Ω의 함수라고 하고, $\psi=h(\theta)$로 정의하자. 그러면 ψ의 **Estimator** 는 d차원 함수 $\delta(X_1,...,X_n)$이고, $X_1=x_1,...,X_n=x_n$으로 관측될 때, $\delta(x_1,...,x_n)$를 ψ의 **Estimate** 라 한다.

Example 7.4.7) Lifetimes of Electronic Components.

1846년에 5732명의 스코틀랜드 군인들의 가슴둘레 측정(inches)에 대한 약간의 오차가 있는 데이터가 보고되었다. 이 데이터는 1817년 의료 저널에서 먼저 보고되었고, Stigler(1986)에 의해 논의되었다. Figure 7.6에서는 이 데이터의 히스토그램을 보여준다. 이제 우리가 개개인의 가슴둘레 측정값을 모수 θ와 분산이 4로 주어진 정규확률변수의 확률 표본으로써 모형화한다고 가정하자. 가슴둘레 측정값의 평균은 $\overline{x}_n=39.85$이다. 만약 θ가 모수가 μ_0, v_0^2인 정규 사전분포를 갖는다면, 정리 7.3.3의 θ의 사후분포에 관한 식에서 정규적이고, 평균과 분산은 다음과 같이 구할 수 있게 된다.

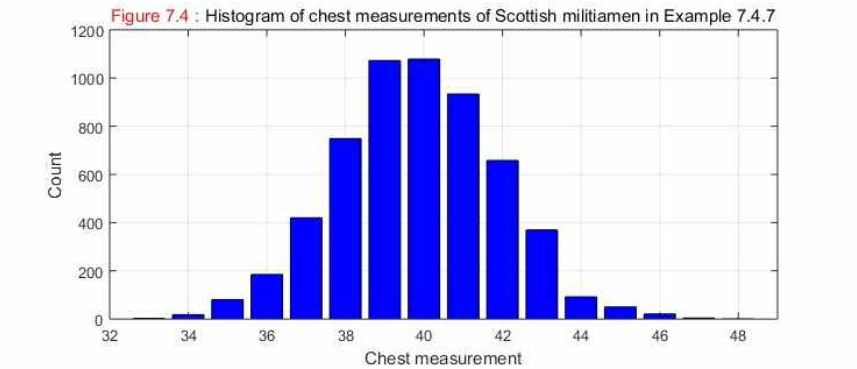
$$\mu_1=\frac{4\mu_0+5732\times v_0^2\times 39.85}{4+5732\times v_0^2}\ ,\ v_1^2=\frac{4v_0^2}{4+5732v_0^2}$$

그러면 Bayes estimate는 $\delta(\vec{x})=\mu_1$이 된다. 여기서 μ_0이 굉장히 커지거나 v_0^2이 매우 작아진다고 해도, 우리는 μ_1이 거의 39.85와 같고 v_1^2이 거의 4/5732와 같아진다. 만약 θ에 관한 prior p.d.f.가 임의의 연속함수로써 $\theta=39.85$에 양의 자리 근처이고 θ가 39.85와 멀지 않을 때 극단적으로 커지지 않는다면 그에 대응하는 posterior p.d.f.는 평균과 분산이 각각 39.85, 4/5732인 정규분포의 p.d.f.와 매우 가깝게 된다. 이런 이유로, 그러한 사후분포의 평균과 중앙값은 사전 분포와 상관없이 거의 $\overline{x}_n$에 가깝게 된다.

▲

*** Matlab Code in Example 7.4.7 ***

```
clear all
clc

% data
count=[3 18 81 185 420 749 1073 1079 934 658 370 92 50 21 4 1];
chest=33:48;

% histogram
bar(chest,count,'b');
title(['\color{red}Figure 7.4 : ' '\color{black}Histogram of chest measurements of Scottish militiamen in Example 7.4.7'])
xlabel('Chest measurement');
ylabel('Count');
xlim([32 49]);
grid on
```



Example 7.4.8) Lifetimes of Electronic Components.

예제 7.3.12에서, estimate $\psi = 1/\theta$(가전 품목의 불량인 경우의 평균 시간)를 구하길 원한다고 가정하자.

θ에 관한 사후분포는 모수를 4와 8.6으로 갖는 감마분포이다. 만약 여기서 제곱오차 손실함수 $L(\theta, a) = (\psi - a)^2$를 이용한다면, 정리 4.7.3에 의하여 Bayes estimate는 ψ에 관한 사후분포의 평균(기대값)이 된다. 즉,

$$\begin{aligned}
\delta^*(\vec{x}) &= E(\psi|\vec{x}) = E\left(\frac{1}{\theta}|\vec{x}\right) \\
&= \int_0^\infty \frac{1}{\theta}\xi(\theta|\vec{x})d\theta \\
&= \int_0^\infty \frac{1}{\theta}\frac{8.6^4}{6}\theta^3 e^{-8.6\theta}d\theta \\
&= \frac{8.6^4}{6}\int_0^\infty \theta^2 e^{-8.6\theta}d\theta \\
&= \frac{8.6^4}{6}\frac{2}{8.6^3} = 2.867
\end{aligned}$$

여기서 마지막 등식은 정리 5.7.3으로부터 유도된다. $1/\theta$의 평균은 $1/E(\theta|\vec{x}) = 8.6/4 = 2.15$보다 약간 더 높다.

▲

7.5 Maximum Likelihood Estimators

Example 7.5.1) Lifetimes of Electronic Components.

예제 7.3.11의 데이터인 전자부품의 수명을 포함하는 데이터를 관측한다고 가정하자. 그러면 처음 사전분포와 손실함수 없이 불량률(failure rate) θ를 추정할 수 있는 방법이 있는가?

▲

Definition 7.5.1) Likelihood Function

확률표본들의 관측값들의 결합 p.d.f. $f_n(\vec{x}|\theta)$가 주어진 $x_1,...,x_n$에 대한 θ의 함수로 간주될 때 $f_n(\vec{x}|\theta)$를 **우도함수/가능도함수(Likelihood function)**이라 한다.

Definition 7.5.2) Maximum Likelihood Estimator / Estimate

각각의 모든 가능한 관측 벡터 $\vec{x}$에 대하여, $\delta(\vec{x}) \in \Omega$를 우도함수 $f_n(\vec{x}|\theta)$가 최대가 되게 하는 $\theta \in \Omega$의 값이라고 하자. 그리고 $\hat{\theta} = \delta(\vec{X})$를 이러한 관점으로 정의된 θ의 estimator라고 하자. 그러면 estimator $\hat{\theta}$를 Maximum Likelihood Estimator라 한다. $\vec{X} = \vec{x}$로 관측되었을 때, 그 값인 $\delta(\vec{x})$를 Maximum Likelihood Estimate라 한다.

Example 7.5.2) Lifetimes of Electronic Components.

예제 7.3.11에서, 관측된 데이터는 $X_1 = 3,\ X_2 = 1,5,\ X_3 = 2.1$이었다. 확률변수들은 크기가 3인 확률표본으로써 모수 θ를 갖는 지수분포로부터 모형화 하였다. 여기서 우도함수는 다음과 같다.

$$f_3(\vec{x}|\theta) = \theta^3 \exp(-6.6\theta),\ \theta > 0,\ \vec{x} = (2, 1.5, 2.1).$$

우도함수 $f_3(\vec{x}|\theta)$를 최대가 되게 하는 θ값을 찾는 것은, 로그함수가 증가함수임을 생각하면, $\log f_3(\vec{x}|\theta)$를 최대가 되게 하는 θ값을 찾는 것과 같아진다. 따라서 다음의 식을 최대화하는 θ값을 찾음으로써 M.L.E를 손쉽게 결정할 수 있다.

$$L(\theta) = \log f_3(\vec{x}|\theta) = 3\log(\theta) - 6.6\theta$$

$L(\theta)$의 도함수를 구하여 $\frac{dL(\theta)}{d\theta} = 0$ 으로 두면(일계도함수판정), θ에 대한 임계값은 $\theta = 3/6.6 = 0.455$로 구해지고, $\frac{d^2L(\theta)}{d\theta^2} = 0$ 으로 두고 θ에 관하여 풀면(이계도함수판정), 방정식을 만족하는 θ는 $\theta < 0$이므로 임계값이 곧 최대값이 된다. 따라서 Maximum Likelihood Estimate는 0.455로 구해진다.

▲

Example 7.5.3) Test for a Disease.

누군가가 길을 걷다가 공공보건소에서 특정 질병에 대한 무료 의학진단을 실시하는 것을 보았다고 하자. 그 의학 진단은 다음과 같은 관점에서 90% 신뢰할 수 있다.

만약 누군가 질병이 있으면, 양성일 진단결과가 나올 확률이 0.9인 반면에, 만약 질병이 없으면 양성일 진단결과가 나올 확률이 0.1 이다. 이러한 의학진단과 같은 예시는 예제 2.3.1에서 다루었다. 우리는 X를 그러한 진단에 대한 결과를 나탄다고 하고, $X=1$이 양성 결과, $X=0$이 양성이 아닌 결과를 의미한다고 하자. 모수 공간은 $\Omega=\{0.1,0.9\}$으로, $\theta=0.1$이면 질병에 걸리지 않았다는 진단 결과를, $\theta=0.9$는 질병에 걸렸다는 진단 결과를 의미한다고 하자. 이러한 모수공간은 주어진 θ에 대하여 X가 모수 θ를 갖는 베르누이 분포를 따른다고 할 수 있다. 그러므로 우도함수는 다음과 같다.

$$f(x|\theta)=\theta^x(1-\theta)^{1-x}$$

이 때, $x=0$으로 관측된 경우, $f(0|\theta)=\begin{cases}0.9 & \text{if } \theta=0.1\\ 0.1 & \text{if } \theta=0.9\end{cases}$ 이고, 여기서 $\theta=0.1$ 인 경우 우도함수가 최대가 됨을 쉽게 알 수 있다. 만약 $x=1$으로 관측된 경우, $f(1|\theta)=\begin{cases}0.1 & \text{if } \theta=0.1\\ 0.9 & \text{if } \theta=0.9\end{cases}$ 이고, 여기서 $\theta=0.9$인 경우 우도함수가 최대가 됨을 쉽게 알 수 있다. 이런 이유로, 우리는 다음과 같은 M.L.E를 얻는다.

$$\hat{\theta}=\begin{cases}0.1 & \text{if } \theta=0.1\\ 0.9 & \text{if } \theta=0.9\end{cases}$$

▲

Example 7.5.4) Sampling from a Bernoulli Distribution.

확률변수 $X_1,...,X_n$이 알려지지 않은 모수 $\theta(0\le\theta\le1)$를 갖는 베르누이 분포로부터 확률표본을 형성한다고 하자. 각각의 x_i가 0 또는 1을 갖는 모든 관측값 $x_1,...,x_n$에 대하여, 우도함수는 다음과 같이 구해진다.

$$f_n(\vec{x}|\theta)=\prod_{i=1}^{n}\theta^{x_i}(1-\theta)^{1-x_i}$$

직접 우도함수 $f_n(\vec{x}|\theta)$를 최대화하는 것 대신, 이를 앞선 예제와 마찬가지로 최대화하기 쉽도록 로그를 취해준다.

$$\begin{aligned}L(\theta)=\log f_n(\vec{x}|\theta)&=\sum_{i=1}^{n}\left[x_i\log\theta+(1-x_i)\log(1-\theta)\right]\\&=\left(\sum_{i=1}^{n}x_i\right)\log\theta+\left(n-\sum_{i=1}^{n}x_i\right)\log(1-\theta).\end{aligned}$$

이제 앞선 예제와 마찬가지로 일계도함수 판정을 수행한다. 만약 $\Sigma_{i=1}^{n}x_i\notin\{0,n\}$ 이면, 일계도함수가 0이 되는 값은 $\theta=\bar{x}_n$이고, 이는 이계도함수 판전에 의하여 결국 $L(\theta)$과 우도함수를 최대화하는 값이 된다. 만약 $\Sigma_{i=1}^{n}x_i=0$ 이면 모든 θ에 대하여 $L(\theta)$가 감소함수가 되고, 따라서 $L(\theta)$가 $\theta=0$에서 최대가 된다. 마찬가지로 $\Sigma_{i=1}^{n}x_i=n$인 경우도 증가함수가 되므로 $\theta=1$에서 최대가 된다. 즉, $\Sigma_{i=1}^{n}x_i\in\{0,n\}$인 경우 모두 우도함수의 최대값이 $\theta=\bar{x}_n$에서 일어난다. 그러므로 θ의 M.L.E 는 $\hat{\theta}=\bar{X}_n$로 구해진다.

▲

Example 7.5.5) Sampling from a Normal Distribution with Unknown Mean.

확률변수 $X_1,...,X_n$이 알려지지 않은 평균 μ와 알려진 분산 σ^2을 갖는 정규분포로부터 확률표본을 생성한다고 하자. 모든 관측값 $x_1,...,x_n$에 대하여, 우도함수 $f_n(\vec{x}|\mu)$는 다음과 같이 구해진다.

$$f_n(\vec{x}|\mu)=\frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right]$$

이는 다음의 식을 최소화하는 μ값을 찾으면 우도함수 $f_n(\vec{x}|\mu)$를 최대화할 수 있게 된다.

$$Q(\mu)=\sum_{i=1}^{n}(x_i-\mu)^2=\sum_{i=1}^{n}x_i^2-2\mu\sum_{i=1}^{n}x_i+n\mu^2$$

이제 일계도, 이계도함수 판정을 이용하여 방정식 Q가 최소가 되게 하는 μ을 구하면 $\mu=\overline{x}_n$ 임을 얻고, 따라서 μ에 관한 M.L.E는 $\hat{\mu}=\overline{X}_n$으로 구해진다.

▲

Example 7.5.6) Sampling from a Normal Distribution with Unknown Mean and Variance.

예제 7.5.5와 마찬가지로 가정하고, 분산 σ^2또한 알려져있지 않다고 가정하자. 그러면 구하고자 하는 모수는 $\theta=(\mu,\sigma^2)$이 된다. 모든 관측값 $x_1,...,x_n$에 대하여, 우도함수 $f_n(\vec{x}|\mu,\sigma^2)$도 이전 예제와 마찬가지로 다음과 같다.

$$f_n(\vec{x}|\mu)=\frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right]$$

이 함수는 반드시 $-\infty<\mu<\infty$와 $\sigma^2>0$ 의 모든 가능한 값들에 대해 최대가 될 것이다. 직접적으로 이러한 우도함수를 최대화하는 것 대신에, 로그를 취하여 좀 더 쉽게 최대화 하는 방법을 생각하자. 우리는 다음을 얻는다.

$$\begin{aligned}L(\theta)&=\log f_n(\vec{x}|\mu,\sigma^2)\\&=-\frac{n}{2}\log(2\pi)-\frac{n}{2}\log\sigma^2-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\end{aligned}$$

이제 다음 세 단계의 과정을 통하여 $L(\theta)$가 최대가 되게 하는 $\theta=(\mu,\sigma^2)$를 찾도록 한다.

(1) 각 고정된 σ^2에 대하여, $L(\theta)$의 우변을 최대화하는 $\hat{\mu}(\sigma^2)$를 찾는다.

(2) $L(\theta')$, $\theta'=(\hat{\mu}(\sigma^2),\sigma^2)$를 최대화하는 $\widehat{\sigma^2}$를 찾는다.

(3) 관측값들의 확률 벡터가 되는 θ의 M.L.E는 $(\hat{\mu}(\sigma^2),\sigma^2)$이 된다.

(1)번의 경우는 예제 7.5.5에서 이미 확인하였으며, $\hat{\mu}(\sigma^2)=\overline{x}_n$ 임을 얻었다. (2)번의 경우에서는, $\theta'=(\overline{x}_n,\sigma^2)$로 두고, 다음을 최대화(maximize)한다.

$$L(\theta)=-\frac{n}{2}\log(2\pi)-\frac{n}{2}\log\sigma^2-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\overline{x}_n)^2$$

이는 σ^2에 관해 미분하여 0으로 두고 σ^2에 대해 풀면 최대화할 수 있다. 그 도함수는 다음과 같이 구한다.

$$\frac{d}{d\sigma^2}L(\theta') = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$$

이를 0으로 두고 풀면

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$$

임을 얻는다. $L(\theta')$에 대한 2계도함수는 앞서 구한 σ^2에서 음수가 되므로, 우리는 이 σ^2 값이 최대값임을 확인할 수 있다. 그러므로 $\theta = (\mu, \sigma^2)$의 M.L.E 는 다음과 같이 얻을 수 있다.

$$\hat{\theta} = (\hat{\mu}, \hat{\sigma^2}) = \left(\bar{X}_n, \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right)$$

위의 M.L.E의 첫 번째 좌표는 데이터의 표본평균(sample mean)이라 말한다. 마찬가지로, 두 번째 좌표는 표본분산(sample variance)라고 말한다. 이는 표본분산의 관측값들이 표본의 각 관찰값 $x_1, ..., x_n$의 확률이 $\frac{1}{n}$로 배정되는 분포의 분산이 됨을 확인할 수 있다.

▲

Example 7.5.7) Sampling from a Uniform Distribution.

$X_1, ..., X_n$이 구간 $[0, \theta]$에서의 균등분포로부터 확률표본을 생성한다고 하자. 이때 모수 $\theta > 0$은 알려져있지 않다. 각 관측값의 p.d.f. $f(x|\theta)$는 다음과 같은 형태를 갖는다.

$$f(x|\theta) = \begin{cases} \frac{1}{\theta} & \text{for } 0 \le x \le \theta \\ 0 & otherwise \end{cases}$$

그러면, $X_1, ..., X_n$의 joint p.d.f. $f_n(\vec{x}|\theta)$는 다음과 같은 형태를 갖는다.

$$f_n(\vec{x}|\theta) = \begin{cases} \frac{1}{\theta^n} & \text{for } 0 \le x_i \le \theta \ (i = 1, ..., n) \\ 0 & otherwise \end{cases}$$

위의 joint p.d.f.로부터, θ의 M.L.E는 반드시 $\theta \ge x_i$ for $i = 1, ..., n$인 θ의 값으로써 i중 $1/\theta^n$을 최대화하는 값을 찾는 것임을 알 수 있다. $1/\theta^n$는 θ에 대한 감소함수이므로, estimate는 $\theta \ge x_i$ for $i = 1, ..., n$를 만족하는 가장 작은 θ값을 갖는다. 이러한 값은 $\theta = \max\{x_1, ..., x_n\}$이므로, θ에 대한 M.L.E는 $\hat{\theta} = \max\{X_1, ..., X_n\}$ 이 된다.

▲

7.6 Properties of Maximum Likelihood Estimators

§. Invariance

Example 7.6.1) Lifetimes of Electronic Components.

예제 7.1.1에서, 모수 θ는 전자 부품의 불량률(failure rate)로 해석하였다. 또한 예제 7.4.8에서 우리는, Bayes estimate를 $\psi = 1/\theta$ (평균수명)으로 구하였다. 그러면 이러한 ψ의 M.L.E 계산을 위한 대응방법이 존재하는가?

▲

note) 우선 분포에 대한 모수에 대해 다음과 같이 표현하도록 하자. 모수 θ에 관한 p.d.f. $f(x|\theta)$를 다음과 같은 새로운 모수의 관점으로 표현한다. : $\psi = g(\theta)$ s.t. g is one-to-one function of θ. 여기서 우리는 θ의 M.L.E와 ψ의 M.L.E의 관계가 있는지 살펴보고자 한다.

Theorem 7.6.1) Invariance Property of M.L.E.'s

만약 $\hat{\theta}$가 θ의 최대우도추정량(MLE)이고 g가 일대일 함수이면, $g(\hat{\theta})$는 $g(\theta)$의 최대우도추정량이다.

Example 7.6.2) Lifetimes of Electronic Components.

정리 7.6.1에 따르면, ψ의 M.L.E는 θ의 M.L.E를 분모로 갖는 $1/\hat{\theta}$이 된다. 예제 7.5.2에서, 우리는 θ의 M.L.E를 $\hat{\theta} = 0.455$으로 계산하였다. 그러므로 $\hat{\psi}$는 $1/0.455 = 2.2$로 계산된다. 이는 예제 7.4.8에서 제곱오차 손실함수를 이용하여 구한 Bayes estimate 2.867보다 조금 작은 수치이다.

▲

note) 이러한 invariance 성질은 일대일함수가 아닌 경우에도 확장될 수 있다. 예를 들면, 우리가 평균과 분산이 알려지지 않은 정규분포에서 평균 μ를 추정하고 싶다고 가정하자. 그러면 μ는 일대일이 아닌 모수 $\theta = (\mu, \sigma^2)$의 함수이다. 이 경우, 우리가 추정하고 싶은 함수는 $g(\theta) = \mu$가 된다. 이러한 사례에서, 일대일 조건이 필요치 않은 모수 θ에 대한 함수의 M.L.E를 정의하기 위한 많은 방법이 있다. 그 중 대표적으로 많이 쓰는 방법을 제시하고자 한다.

Definition 7.6.1) M.L.E. of a Function.

$g(\theta)$를 모수에 대한 임의의 함수라 하고, $G \subset \Omega$를 함수 g의 상(image)라고 하자. 각 $t \in G$에 대하여, $G_t = \{\theta : g(\theta) = t\}$로 정의하고, $L^*(t) = \max_{\theta \in G_t} \log f_n(\vec{x}|\theta)$로 정의한다.

그러면, 최종적으로 $\hat{t}$에 대한 $g(\theta)$의 M.L.E.는 다음과 같이 정의된다.

$$L^*(\hat{t}) = \max_{t \in G} L^*(t)$$

Theorem 7.6.2) $\hat{\theta}$를 θ의 M.L.E.라하고, $g(\theta)$를 θ의 함수라 하자. 그러면 $g(\theta)$의 M.L.E.는 $g(\hat{\theta})$이다.

Example 7.6.3) Estimating the Standard Deviation and the Second Moment.

$X_1,...,X_n$을 알려지지 않은 평균 μ와 분산 σ^2를 갖는 정규분포로부터 확률 표본을 생성한다고 하자. 우리는 표준편차 σ의 M.L.E.와 정규분포의 2차 적률(second moment) $E(X^2)$를 결정하려고 한다. 우선 우리는 예제 7.5.6에서 $\theta=(\mu,\sigma^2)$의 M.L.E.가 $\hat{\theta}=(\hat{\mu},\widehat{\sigma^2})$임을 구하였다. invariance 성질로부터, 우리는 표준편차의 M.L.E., $\hat{\sigma}$가 그저 표본분산의 제곱근으로 결론지을 수 있다는 것을 알고 있다. 따라서 이는 $\hat{\sigma}=(\widehat{\sigma^2})^{1/2}$이다. 또한 $E(X^2)=\sigma^2+\mu^2$이므로, $E(X^2)$의 M.L.E.는 $\widehat{\sigma^2}+\hat{\mu}^2$이다.

▲

§. Consistency

textbook의 428page를 참고하라.

§. Numerical Computation

많은 문제에서, 주어진 모수 θ에 대한 유일한 M.L.E. $\hat{\theta}$가 존재하지만, 이러한 M.L.E.는 표본에서의 관측값들에 대한 함수의 가까운 형태로써 표현될 수 없는 경우가 있다. 이러한 문제의 경우, 관측값들에 대한 집합이 주어질 때, 수치적인 계산에 의해 $\hat{\theta}$의 값을 결정할 필요가 있다. 우리는 이러한 상황을 설명하기 위해 다음과 같은 두 가지 예제를 제시한다.

Example 7.6.4) Sampling from a Gamma Distribution.

$X_1,...,X_n$가 다음과 같은 p.d.f.를 갖는 감마분포로부터 확률표본을 생성한다고 가정하자.

$$f(x|\alpha)=\frac{1}{\Gamma(\alpha)}x^{\alpha-1}e^{-x} \text{ for } x>0,\ \alpha>0 \text{ is unknown.}$$

그러면 우도함수(likelihood function)는 다음과 같다.

$$f_n(\vec{x}|\alpha)=\frac{1}{\Gamma^n(\alpha)}\left(\prod_{i=1}^{n}x_i\right)^{\alpha-1}\exp\left(\sum_{i=1}^{n}x_i\right)$$

그러면 이제 α의 M.L.E.는 다음의 방정싱을 만족하는 α값을 구하면 얻을 수 있다.

$$\frac{\partial \log f_n(\vec{x}|\alpha)}{\partial\ \alpha}=0$$

위의 미분방정식을 이 예제에서 적용하면, 우리는 다음과 같은 방정식을 얻을 수 있다.

$$\frac{\Gamma'(\alpha)}{\Gamma(\alpha)}=\frac{1}{n}\sum_{i=1}^{n}\log x_i$$

여기서 $\Gamma'(\alpha)/\Gamma(\alpha)$를 디감마 함수(digamma function)이라고 하고, 이는 몇몇 수리 프로그래밍 패키지들로 계산할 수 있다. 모든 주어진 $x_1,...,x_n$에 대해, 위의 디감마 함수를 만족하는 유일한 α는 위와 같은 식 또는 디감마 함수의 수치해석적 방법으로 구할 수 있다. 이 α값이 곧 α의 M.L.E. 가 된다.

▲

Example 7.6.5) Sampling from a Cauchy Distribution.

$X_1,...,X_n$가 알려지지 않은 $\theta(-\infty < \theta < \infty)$를 다음의 p.d.f.를 갖는 중심으로 하는 코시분포로부터 확률표본을 생성한다고 가정하자.

$$f(x|\theta) = \frac{1}{\pi[1+(x-\theta)^2]} \text{ for } -\infty < x < \infty, \text{ the value of } \theta \text{ is to be estimated.}$$

그러면 여기서 우도함수는 $f_n(\vec{x}|\theta) = \frac{1}{\pi^n \Pi_{i=1}^n [1+(x_i-\theta)^2]}$ 로 구할 수 있고, 따라서 θ에 관한 M.L.E.는 다음을 최소화 하는 값이 된다.

$$\prod_{i=1}^{n} [1+(x_i-\theta)^2]$$

대부분의 $x_1,...,x_n$에 대하여, 위의 식을 최소화하는 θ값은 분명 수치적인 계산에 의해 결정될 것이다.

▲

note) 이러한 수치적 계산을 위해 사용되는 여러 수치해석적 방법 중 본 교재에서는 다음과 같은 두 가지의 수치적인 계산 방법을 제시하였다. 이는 다음에 명시된 페이지에 있는 정의와 예제를 참고하는 수준으로 읽어보길 권장한다.

Definition 7.6.2) Newton's method - 429 page (뉴턴 방법을 이용한 MLE 수치계산)
Definition 7.6.3) Method of Moments - 430 page (적률을 이용한 MLE 수치계산)

§. M.L.E.'s and Bayes Estimators

Example 7.6.11) Sampling from an Exponential Distribution.

$X_1, X_2, ...$가 모수 θ에 관한 지수분포를 *i.i.d.*하게 갖는다고 가정하자. $T_n = \sum_{i=1}^{n} X_i$ 라 두면, θ의 M.L.E.는 $\hat{\theta}_n = \frac{n}{T_n}$으로 구할 수 있다(7.5절의 연습문제 7 참고). 이는 $1/\hat{\theta}_n$이 유한 분산을 갖는 *i.i.d.*한 확률변수들의 평균이므로, 중심극한정리(the central limit theorem)에 의하면 $1/\hat{\theta}_n$의 분포는 정규분포에 가까워진다. 이 경우에 평균과 분산은, 근사 정규분포의 모수로써 각각 $1/\theta$, $1/(n\theta^2)$이 된다. 앞선 설명에서의 표기로, $V_n(\theta) = \theta^2$가 된다.

다음으로, θ에 관한 사전분포를 모수 α, β를 갖는 감마분포라고 하자. 정리 7.3.4에 의하면 θ의 사후분포는 $\alpha+n, \beta+t_n$인 모수를 갖는 감마분포가 된다. 우리는 이러한 감마분포가 근사적으로 정규분포가 된다는 것을 보임으로써 결론내릴 수 있다. 간단히 하기 위해, α가 정수라고 가정하자. 그러면 θ의 사후분포는 모수 $\beta+t_n$을 갖는 *i.i.d.*한 지수확률변수 $\alpha+n$의 합에 대한 분포와 같아진다. 그러한 합은 근사적으로 평균 $(\alpha+n)/(\beta+t_n)$, 분산 $(\alpha+n)/(\beta+t_n)^2$인 정규분포를 갖는다. 만약 α와 β 모두 작으면, 근사적 평균은 거의 $n/t_n = \hat{\theta}$이고, 근사적 분산은 거의 $n/t_n^2 = \hat{\theta}^2/n = V_n(\hat{\theta})/n$이 된다.

▲

Example 7.6.12) Sampling from an Exponential Distribution.

예제 7.3.14에서, 우리는 280개의 관측값들을 표본으로 하는, 프로이센군 병사들이 말의 발길질로 사망한(연간) 평균 사망횟수 θ의 사후분포를 구하였다. 사후분포는 모수가 196과 280을 갖는 감마분포였음을 상기하자. 예제 7.6.11의 논의를 참고하면, 이러한 감마분포는 근사적으로 모수 280을 갖는 $i.i.d.$인 196개의 지수확률변수들의 합의 분포가 된다. 그러한 합의 분포는 근사적으로 정규분포가 되며 평균을 $196/280$, 분산을 $196/280^2$으로 갖는다.

예제 7.3.14와 같은 데이터를 이용하면, 우리는 280개의 관측값들의 평균인 θ에 관한 M.L.E.를 찾을 수 있다. 중심극한정리에 의하면, 평균 θ를 갖는 280개의 $i.i.d.$한 포아송 확률변수들의 평균에 대한 분포는 근사적으로 평균 θ와 분산 $\theta/280$을 갖는 정규분포에 근사하게 된다. 따라서, 이전의 notation에 의하면 $V_n(\theta)=\theta$으로 갖는다. 그러면 앞선 관측 데이터들의 최대우도추정치(The maximum likelihood estimate)는 사후분포의 평균으로써 $\hat{\theta}=196/280$으로 갖는다. 또한 사후분포의 분산은 $V_n(\hat{\theta})/n=\hat{\theta}/280$이 된다.

▲

note) 다음의 일반적인 두 상황에서는 사후분포와 MLE의 분포가 정규분포로 근사하지 않는 상황이 발생한다.
(1) 표본의 크기가 충분히 크지 않은 경우, (2) 우도함수가 smooth(미분가능한)하지 않은 경우. 다음에 소개할 예제들은 이러한 상황에서의 예제이다.

(1) 표본의 크기가 충분히 크지 않은 경우.

Example 7.6.13) Lifetimes of Electronic Components.

예제 7.3.12에서, 우리는 표본크기가 $n=3$인 모수를 θ로 갖는 지수확률변수들을 갖고 있었다. 그리고 사후분포는 모수를 4와 8.6으로 갖는 감마분포였음을 상기하자. 그러면 M.L.E.는 $\hat{\theta}=3/(X_1+X_2+X_3)$로, 1/모수를 3과 3θ로 갖는 감마확률변수 의 분포를 갖는다. Figure 7.8은 이러한 사후 p.d.f.와 M.L.E. p.d.f.를 보여준다. M.L.E.의 관측값인 θ를, $\theta=3/6.6$으로 가정하자. 그러면 두 p.d.f.의 분포는 비슷할지라도, 조금 다르다. 또한 두 p.d.f.는 서로 비슷하지만, 여전히 사후분포의 표본평균과 분산을 갖는 정규 p.d.f.와 다르다.

▲

(2) 우도함수가 smooth 하지 않은 경우.

Example 7.6.14) Sampling from a Uniform Distribution.

예제 7.5.7에서, 우리는 구간 $[0,\theta]$에서의 균등분포로부터 표본크기가 n인 θ에 대한 M.L.E.를 구하였다. 그러한 M.L.E.는 $\hat{\theta}=\max\{X_1,...,X_n\}$으로 구하였다. 우리는 예제 3.9.6.의 결과를 이용하여 정확한 $\hat{\theta}$의 분포를 찾을 수 있다. 여기서 $Y=\hat{\theta}$의 p.d.f.는 다음과 같다.

$$g_n(y|\theta)=n[F(y|\theta)]^{n-1}f(y|\theta),$$

where $f(\bullet\,|\theta)$: p.d.f. of Uniform dist, $F(\bullet\,|\theta)$: c.d.f. of Uniform dist. (on $[0,\theta]$.)

위의 p.d.f.는 정리하면 다음과 같이 쓸 수 있다.

$$g_n(y|\theta)=n\left[\frac{y}{\theta}\right]^{n-1}\frac{1}{\theta}=n\frac{y^{n-1}}{\theta^n}\quad \text{for } 0<y<\theta$$

이러한 p.d.f.는 normal p.d.f.와 조금도 비슷하지 않다. 이는 매우 비대칭적이고 가장 큰 가능한 M.L.E값에서 최대값을 갖는다. 사실, $\hat{\theta}$에 대한 평균과 분산을 구하면 각각 다음과 같다.

$$E(\hat{\theta})=\frac{n}{n+1}\theta\ ,\ \ Var(\hat{\theta})=\frac{n}{(n+1)^2(n+2)}\theta^2$$

이 분산은 우리가 이전에 보던 근사적으로 정규적인 예제에서의 $1/n$ 대신에 $1/n^2$에 점점 가까워진다.
만약 n이 크면, θ에 대한 사후분포는 근사적으로 우도함수와 상수곱의 형태를 갖는 p.d.f.를 갖는다. 그러한 함수에 분모를 θ로 취하여 적분하면, 요구되는 상수는 다음과 같은 θ의 근사 사후 p.d.f.로 유도된다.

$$\xi(\theta|\vec{x}) \approx \begin{cases} \dfrac{(n-1)\hat{\theta}^{n-1}}{\theta^n} & \text{for } \theta > \hat{\theta} \\ 0 & otherwise \end{cases}$$

이러한 근사 사후분포에서의 평균과 분산은 각각 $(n-1)\hat{\theta}/(n-2)$와 $(n-1)\hat{\theta}^2/\left[(n-2)^2(n-3)\right]$으로 구해진다. 사후 평균은 여전히 거의 M.L.E.와 같고, 사후 분산은 M.L.E.의 분산인 $1/n^2$의 비율로 감소한다. 그러나 사후분포는 p.d.f.가 가장 작은 가능한 θ값에 대해 최대값을 가짐으로써 조금도 정규적이지 않다.

▲