

NHPP에 기초한 소프트웨어 신뢰도 모형에 대한 베이지안 추론에 관한 연구

이 상 식[†]·김 회 철[†]·송 영 재^{††}

요 약

본 논문에서는 비동질 포아송 과정에 기초한 소프트웨어 오류 현상에 대한 신뢰도 모형을 고려하고 사전정보를 이용한 베이지안 추론을 시행하였다. 고장 패턴은 NHPP에 대한 강도함수와 평균값 함수로서 나타낼 수 있다. 따라서 본 논문에서는 기존의 모형인 Goel이 제시한 모형과 신뢰성 분포로 많이 사용되는 와이블 분포의 특수형태인 레일리분포와 라플라스 분포를 이용한 모형을 제시하여 베이지안 추론을 시행하고 또, 효율적 모형을 위한 모형선택으로서 편차자승합을 이용하여 비교하였다. 모수의 추정을 위해서 마코브체인 몬테카를로 기법중에 하나인 깁스샘플링을 이용한 근사추정 기법이 사용되었다. 수치적인 예에서는 실측자료인 NTDS 자료를 이용하여 모수 및 신뢰도를 추정하였고 편차자승합을 이용한 모형비교의 결과를 나열하였다.

The Bayesian Inference for Software Reliability Models Based on NHPP

Sang-Sik Lee[†] · Hee-Cheul Kim[†] · Young-jae Song^{††}

ABSTRACT

Software reliability growth models are used in testing stages of software development to model the error content and time intervals between software failures. This paper presents a stochastic model for the software failure phenomenon based on a nonhomogeneous Poisson process (NHPP) and performs Bayesian inference using prior information. The failure process is analyzed to develop a suitable mean value function for the NHPP; expressions are given for several performance measure. Actual software failure data are compared with several model on the constant reflecting the quality of testing. The performance measures and parametric inferences of the suggested models using Rayleigh distribution and Laplace distribution are discussed. The results of the suggested models are applied to real software failure data and compared with Goel model. Tools of parameter point inference and 95% credible intereval was used method of Gibbs sampling. In this paper, model selection using the sum of the squared errors was employed. The numerical example by NTDS data was illustrated.

키워드: 소프트웨어 신뢰도 모형(Software reliability growth models), 깁스샘플링(Gibbs sampling), 비동차 포아송과정(nonhomogeneous Poisson Process), 레일리분포(Rayleigh distribution), 라플라스 분포(Laplace distribution), 편차자승 합(sum of the squared errors), 소프트웨어 신뢰도(Software reliability)

1. 서 론

우리의 주변에는 복잡한 소프트웨어 시스템들로 둘러 쌓여 있으며 이러한 시스템의 혜택을 받는 일이 커짐에 따라 소프트웨어 신뢰성의 역할은 점차 커지게 되었다.

그러나 시스템이 애러가 나면 그 원인을 찾아 필요할 경우 새로운 디자인을 개발하거나 새로운 기술을 도입하게 된다. 따라서 시간이 지남에 따라 신뢰도의 증가가 기대되어

진다.

소프트웨어 테스트 단계에서 소프트웨어 오류 수와 오류 간격시간에 의해 소프트웨어 애러현상을 수리적으로 모형화하면 소프트웨어에 대한 평가를 쉽게 할 수 있으며, 신뢰도 모형에 의해 소프트웨어 오류 수, 소프트웨어 애러발생간격 시간, 소프트웨어 신뢰도 및 애러율 등의 신뢰성 평가측도들이 추정되어 예측할 수 있다.

우리가 가능한 시스템이나 소프트웨어 신뢰도에서 관측시간 $(0, t]$ 사이에 발견된 오류 수 $N(t)$ 를 모형화하는데 비동차 포아송 과정(NHPP; non-homogeneous poisson process)이 널리 사용되어 왔다. 강도함수 혹은 고장 발생률(ROCOF

[†] 정 회 원: 송호대학 정보산업계열 교수

^{††} 중신회원: 경희대학교 산업정보대학원 원장

논문접수: 2001년 6월 28일, 심사완료: 2002년 3월 22일

; rate of occurrence of failure) $\lambda(t) = dE[N(t)]/dt$ 은 t 에 대한 단조함수로 흔히 가정한다. 예를 들어 동차 포아송 과정 (HPP; homogeneous poisson process), Musa-Okumoto과정, Weibull 과정 그리고 Cox-Lewis 과정은 각각 ROCOF를 상수, 시간에 관한 부분 함수, 멱 함수, 대수선형 함수를 가정하였다[12, 13].

본 논문에서는 사전분포를 고려한 베이지안 측면에서 기존의 모형인 Goel 모형과 연관되고 신뢰도 분포로 많이 사용되는 와이블 분포의 특수형태인 레일리분포와 라블라스분포 [10, 13]를 적용한 모형을 제안하였다. 또, 효율적 모형을 위한 모형선택으로서 편차자승합(SSE)을 이용하여 비교하고자 한다[9].

베이지안 측면에서 NHPP에 대한 베이지 추정치들을 계산하기 위한 사후분포의 형태를 구하는데 깃스 표본 추출알고리즘이 제안되었다.

깃스 알고리즘은 MCMC(Markov Chain Monte Carlo) 알고리즘인데, 이것은 바람직한 사후분포로서 정상성 분포를 가지는 마코프 체인에 따라 표본들이 변화하는 알고리즘이다[4-6]. 이 마코프 체인의 추이 측정은 주로 보통 조건부 밀도함수의 곱이다. 즉, 한 변수는 소프트웨어에 남아있는 수많은 오류들이며, 다른 변수는 포아송 확률들 중에서 척도모수의 변화에 따라서 어떤 요소가 도움을 주는지를 나타낸다.

본 논문의 구성의 2장에서는 관련연구로서 소프트웨어 신뢰도에 대한 일반적 내용과 깃스 샘플링에 대하여 요약하였고, 3장에서는 Goel 모형과 제안된 모형에 대한 깃스 알고리즘을 이용한 모수 추정에 대하여 서술하였고, 4장에서는 실측자료인 Naval Tactical Data System 자료를 이용하여 각 모형에 대한 모수 추정 및 모형 비교에 대해서 설명한 수치적인 예를 나열하였고 마지막 5장에서는 결론 및 향후 연구방향을 제시하였다.

2. 관련 연구

2.1 소프트웨어 신뢰도

관측시간 $(0, t]$ 사이에서 n 번째까지 고장시점이 관찰된 고장 절단 모형일 경우 $x_i (i=1, 2, \dots, n)$ 를 여러 발생시간 간격으로 표시하고, 최종고장 시점 $s_n = \sum_{i=1}^n x_i$ 이라고 표현하면 데이터 집합 D_n 은 $\{s_1, s_2, \dots, s_n; t\}$ 으로 구성된다. 임의의 테스트 시점 t 까지의 신뢰도는 다음과 같이 알려져 있다[12, 13].

$$1 - F(t) = R(t) = \exp[-m(t)] \tag{2.1}$$

식 (2.1)을 이용하여 양변을 미분하면 $-f(t) = -m'(t) \exp$

$[-m(t)]$ 가 되고, 밀도함수는 $f(t) = \lambda(t) \exp[-m(t)]$ 으로 표현이 가능하다(단, $F(t)$ 는 누적밀도함수, $\lambda(t)$ 는 강도함수이고 $m'(t) = \lambda(t)$ 이 되고 $m(t)$ 는 평균값 함수를 $\lambda(t)$ 는 강도함수를 의미한다).

따라서 s_1 에서부터 s_n 까지 우도함수는 다음과 같다[12, 13].

$$L_{s_n} = \left[\prod_{k=1}^n \lambda(s_k) \right] \exp[-m(t)] \tag{2.2}$$

단, $t = s_n$ 이면 시간절단모형이 된다. 본 논문에서는 이러한 시간절단모형을 이용하고자 한다.

테스트에 의해 최종적으로 발견되는 총 기대 오류 수 즉, 테스트 개시 전에 소프트웨어 내에 잠재하는 총 기대 오류 수를 β_0 라고 할 때, 일반적으로는 $m(\infty) = \beta_0$ 인 것으로부터 다음과 같이 됨이 알려져 있다[13].

$$\lim_{t \rightarrow \infty} \Pr\{N(t) = n\} = \frac{\beta_0^n}{n!} e^{-\beta_0} \quad (n=0, 1, 2, \dots) \tag{2.3}$$

이러한 사실로부터 소프트웨어 신뢰도에서 관측시간 $(0, t]$ 사이에 발견된 오류 수 $N(t)$ 의 극한분포는 평균치가 β_0 인 포아송 분포에 따르는 것을 알 수 있다.

따라서 NHPP에 기초한 소프트웨어 신뢰도 모형으로부터 정량적인 신뢰성 평가척도를 도출할 수 있다.

그리고 테스트 시점 t 에서 소프트웨어 고장이 일어난다고 하는 가정 하에서 신뢰구간 $(t, t+x]$ (단, x 는 임무시간(mission time))사이 소프트웨어의 오류가 일어나지 않을 조건부 확률은 $R(x|t) = F(t+x) - F(t)$ 로 놓고 식 (2.1)과 관련하여 정리하면 다음과 같다.

$$R(x|t) = \exp[-\{m(t+x) - m(t)\}] \quad (t \geq 0, x \geq 0) \tag{2.4}$$

식 (2.4)을 소프트웨어 신뢰도(software reliability)라고 부른다[9].

2.2 깃스 샘플링

깃스 샘플링은 화상공정모형을 연구하는 Geman과 Geman (1984)의 논문을 통하여 비로소 연구가 시작되어 신경망, 전문가 시스템과 같은 척도가 큰 복잡한 모형에 적용하여 사용되어 왔다[3, 5, 6].

본 절에서는 Geman과 Geman(1984)과 Gelfand와 Smith (1990)에 의해 더욱 구체화된 깃스 샘플링을 소개하고 Gelfand와 Rubin(1992)이 제시한 다중 열을 이용한 반복적 시물레이션 기법을 요약하고자 한다.

깃스 샘플링의 가정은 추출된 표본은 서로 독립이고 충분한 수의 반복이 이루어지면 깃스 알고리즘으로부터 추출된

표본은 안정된 특정분포로 수렴한다.

확률변수 (U_1, U_2, \dots, U_p) 의 결합분포함수가 주어졌을 때 각각의 주변 확률밀도함수가 이용 가능하다고 가정하고 과산포분포에서 선택된 초기값을 $(u_1^{(0)}, u_2^{(0)}, \dots, u_p^{(0)})$ 라고 정의하면 다음과 같은 조건부 밀도함수로부터 변량이 추출된다.

$$\begin{aligned} u_1^{(1)} &\sim f(U_1 | U_2^{(0)}, U_3^{(0)}, \dots, f(U_p^{(0)}), \\ u_2^{(1)} &\sim f(U_2 | U_1^{(1)}, U_3^{(0)}, \dots, f(U_p^{(0)}), \\ &\dots\dots\dots \\ u_p^{(1)} &\sim f(U_p | U_1^{(1)}, U_2^{(1)}, \dots, f(U_{p-1}^{(1)}). \end{aligned}$$

각각의 변수들은 최신값들로 갱신되면서 p개의 변량을 발생시킬 수 있고 충분한 t번의 반복이후에 다음과 같은 변량을 얻을 수 있다.

$$(U_1^{(t)}, U_2^{(t)}, \dots, U_p^{(t)})$$

따라서, Geman과 Geman(1984)은 적당한 조건하에서 t가 충분히 크면 다음과 같이 수렴함을 제시하였다(단, \rightarrow^d 은 분포 수렴을 의미함).

$$(U_1^{(t)}, \dots, U_p^{(t)}) \rightarrow^d (U_1, U_2, \dots, U_p) \sim f(U_1, U_2, \dots, U_p) (t \rightarrow \infty)$$

그러므로 충분히 큰 t에 대해서 $u_s^{(t)}$ 는 $f(u_s)$ ($s=1, 2, \dots, p$)의 표본으로 간주될 수 있고 이런 과정을 m번 적용하면 m개의 독립 표본이 다음과 같이 얻어진다.

$$(U_{ij}^{(t)}, \dots, U_{jn}^{(t)}) (j=1, 2, \dots, m)$$

위 식에서 U_1, \dots, U_p 는 경우에 따라서 벡터가 될 수 있다.

사후 밀도함수의 값을 계산하기 위해서 본 논문은 다음과 같이 Rao-Blackwell개념을 사용하였다.

$$f(\widehat{U}_s) \approx (m)^{-1} \sum_{j=1}^m f(U_{sj} | U_{nj}^{(t)}, r \neq s).$$

이러한 깃스 추출법의 주된 어려움은 반복 정지시점 t를 결정하는 일이다. 이러한 이유와 관련하여 Gelman과 Rubin (1992)는 반복 단계에서 수렴성의 효율을 높이고 초기분포의 의존성을 작게하기 위해 각 열의 전반부 t/2번 반복을 제외하고 후반부 t/2번 반복만을 고려하는 기법을 제안하였다. 본 연구는 이 방법을 적용하였다. 그리고 대안적인 방법으로 충분한 표본을 발생하여 표본내의 분산과 표본 열 사이의 분산을 작게 하는 시점을 반복정지시점으로 간주됨을 제안하기도 하였다[5].

3. 깃스 샘플링을 이용한 NHPP 소프트웨어 신뢰도 모형에 대한 모수 추정

NHPP모형에 대한 모수추정을 위해 미지의 남아 있는 결합의 수(N)를 사용한다. 즉, N은 평균 β_0 을 가진 포아송 분포를 한다는 가정이 필요하다($\lim_{m \rightarrow \infty} m(t) = \beta_0$).

이 사실을 이용하여 데이터 증대를 적용하면 깃스 알고리즘을 이용하는데 필요한 조건부 분포를 구체화 할 수 있다 [11, 17, 18].

식 (2.2)의 우도함수에 있는 $\exp(-m(t)) = \exp[-\beta_0 F(t | \beta_1)]$ 의 형태 때문에 깃스 샘플링에 있어서 필요로 하는 주어진 β_0 와 D_i 에 대한 β_1 의 조건부 밀도함수 형태를 구체화하기 어렵다. 그러므로 모수 $\beta_0 [1 - F(t | \beta_1)]$ 을 가지는 포아송 분포를 따르는 잠재 변수 $N' = N - n$ 을 이용한다. 그러면 사후분포 $p(\beta_0, \beta_1 | D_i)$ 는 결합밀도 $p(\beta_0, \beta_1, N' | D_i)$ 을 주변 밀도화 함으로서 얻을 수 있다. 이 밀도는 다음과 같은 조건부 밀도의 반복을 이용하여 깃스 샘플링을 통하여 근사밀도함수를 구할 수 있다. 즉, $p(N' | \beta_0, \beta_1, D_i)$, $p(\beta_0 | N', \beta_1, D_i)$, $p(\beta_1 | N', \beta_0, D_i)$, 단, 이 밀도들은 구체화가 가능한 밀도라야 한다[11].

수식표현을 위해서 $X \perp Y$ 는 X와 Y가 독립을 의미하고 $N' = N - n$ 을 의미하고 $\Gamma(a, b)$ 는 평균이 a/b 인 감마분포를 나타낸다. 그리고 $P(\theta)$ 는 평균이 θ 인 포아송 분포를 의미하고, β_1 은 벡터도 될 수 있으나 본 논문에서는 잔존에러 1개당 오류 발견율을 의미하는 모수 β 을 나타낸다.

3.1 Goel 모형

Goel 모형에서 식 (2.2)을 이용하면 우도함수와 사전분포는 각각 다음과 같이 나타낼 수 있다[7, 8].

$$L_{Goel}(\beta_0, \beta | D_s) = \left(\prod_{i=1}^n \beta_0 \beta e^{-\beta s_i} \right) \exp[-\beta_0 (1 - e^{-\beta s_n})] \quad (3.1)$$

$$N' \sim P(\beta_0 e^{-\beta s_n}), \beta_0 \sim \Gamma(a, b), \beta \sim \Gamma(c, d), N' \perp \beta_0 \perp \beta \quad (3.2)$$

따라서 잠재변수 $N' = N - n$ 를 사용하고 사전분포와 우도함수를 사용하면 다음과 같은 결합사후분포를 표현할 수 있다[11, 16].

$$\begin{aligned} p(N', \beta_0, \beta | D_s) &\propto \beta_0^n \beta^n e^{-\beta \sum_{i=1}^n s_i} \exp[-\beta_0 (1 - e^{-\beta s_n})] \\ &\cdot \frac{1}{N!} \exp(-\beta_0 e^{-\beta s_n}) \cdot (\beta_0 e^{-\beta s_n})^N \beta_0^{a-1} \cdot e^{-b\beta_0} \beta^{c-1} \\ &\cdot e^{-d\beta} \propto \frac{1}{N!} \beta_0^{N+n+a-1} e^{-\beta_0(b+1)} \beta^{c-1+n} e^{-\beta(\sum_{i=1}^n s_i + d + s_n N)} \end{aligned} \quad (3.3)$$

식 (3.3)에서 베이즈 정리와 장애모수의 개념을 이용하여 깁스 샘플링 알고리즘을 이용하기 위한 각 모수의 주변 사후분포는 다음과 같다[11, 16].

$$N' | \beta_0, \beta, D_{s_n} \sim P(\beta_0 e^{-\beta s_n}); \quad (3.4.1)$$

$$\beta_0 | N', \beta, D_{s_n} \sim \Gamma(a+n+N', b+1); \quad (3.4.2)$$

$$\beta | N', \theta, D_{s_n} \sim \Gamma\left(c+n, d+s_n N' + \sum_{i=1}^n s_i\right); \quad (3.4.3)$$

<깁스 표본추출 알고리즘>

적당한 사전 분포를 사용하여 m 번 적용을 하고 t 번 반복을 한다. 사전분포는 분산이 비교적 큰 과산분포를 고려하여 $\beta \sim \Gamma(1, 0.0001)$, $\beta_0 \sim \Gamma(1, 0.0001)$ 을 선택이용하였다. Goel 모형(지수분포)를 적용하면 다음과 같은 깁스 알고리즘이 된다.

(1) 변수 (β_0, β) 의 초기값을 위의 사전분포에서 하나를 랜덤추출하여 $(\beta_0^{(0)}, \beta^{(0)})$ 를 정하고 모수추정을 위해서 식 (2.8.1), 식 (2.8.2), 식 (2.8.3)에 있는 조건부 밀도함수를 이용하자.

(1-1) $\theta = \beta^{(0)}, \beta = \beta^{(0)}$ 로 고정시켰을 경우에 N' 의 조건부 분포로부터 랜덤표본을 하나 생성시키고 이를 $N^{(1)}$ 이라 표현된다. 즉,

$$N^{(1)} \sim f(N' | \theta^{(0)}, \beta^{(0)}, D_{s_n});$$

(1-2) $\beta_0^{(1)} \sim f(\beta_0 | N^{(1)}, \beta^{(0)}, D_{s_n});$

(1-3) $\beta^{(1)} \sim f(\beta | N^{(1)}, \beta_0^{(1)}, D_{s_n});$

(2) 단계 (1-1), (1-2), (1-3)에서 고정시키는 N', β_0, β 의 값을 가장 최근에 생성된 랜덤표본의 값으로 대체하면서 단계 (1-1), (1-2), (1-3)을 충분히 큰수 t 번만큼 반복 수행한다. 이런 시행을 하여 얻게 되는 최종랜덤표본을 $(N_1^{(t)}, \theta_1^{(t)}, \beta_1^{(t)})$ 이라 한다.

(3) (1)단계와 (2)단계를 다시 $(m-1)$ 번 적용 시행하면 총 m 개의 랜덤표본

$$(N_1^{(t)}, \beta_{01}^{(t)}, \beta^{(t)}), \dots, (N_m^{(t)}, \beta_{0m}^{(t)}, \beta_m^{(t)})$$

이 얻어진다.

(4) 최종적인 결과에 의해 β 와 β_0 의 추정은 아래와 같다.

$$\hat{E}(\beta_0) = \frac{1}{m} \sum_{k=1}^m \beta_{0k}^{(t)} \quad \hat{E}(\beta) = \frac{1}{m} \sum_{k=1}^m \beta_k^{(t)} \quad (2.9)$$

신뢰도는 식 (2.4)을 연관하여 다음과 같이 나타낼 수 있다(단, x 는 임무시간)를 의미함).

$$\hat{R}(x | s_n) = \exp[-\hat{\beta}_0 \{ \exp(-\hat{\beta}_1 s_n) - \exp(-\hat{\beta}_1 (s_n + x)) \}] \quad (2.10)$$

3.2 제안 모형

기본적인 Goel[8]이 제시한 모형은 평균값 함수 $m(t) = \beta_0(1 - e^{-\beta_1 t})$ ($\beta_0 > 0, \beta_1 > 0$)이라고 알려져 있다. 이 평균값 함수를 일반화하면 다음과 같이 표현할 수 있다.

$$m(t) = \beta_0 \cdot F(t) \quad (\text{즉, } \lambda(t) = \beta_0 \cdot f(t))$$

여기서, $F(t)$ 는 분포함수이고, $f(t)$ 는 확률밀도함수를 의미한다. 즉, $F(t)$ 가 지수분포의 분포함수 $F(t) = (1 - e^{-\beta_1 t})$ 이면 Goel 모형이 된다 따라서 $F(t)$ 의 형태를 바꾸면 새로운 모형을 제안할 수 있다. 그러므로 신뢰도 분포로 많이 사용되는 와이블 분포의 특수형태인 레일리분포와 라플라스 혹은 이중 지수분포[10]를 이용하여 다음과 같은 모형을 제시한다.

3.2.1 레일리분포 모형

레일리 분포는 형상모수가 2인 와이블 분포의 특수한 경우이므로 평균값 함수는 임의의 시간 t 에 대하여 평균값 함수와 강도함수는 각각 다음과 같이 표현할 수 있다.

$$m(t) = \beta_0 [1 - e^{-\beta_1 t^2}], \quad \lambda(t) = 2\beta_0 \beta_1 t \exp(-\beta_1 t^2) \quad (\beta_0 > 0, \beta > 0)$$

따라서 우도함수와 사전분포는 식 (2.2)과 관련하면 다음과 같다.

$$L_{Ray}(\beta_0, \beta | s_1, s_2, \dots, s_n) = \left\{ \prod_{k=1}^n 2\beta_0 \beta_1 s_k \exp(-\beta_1 s_k^2) \right\} \cdot \exp[-\beta_0 (1 - e^{-\beta_1 s_n^2})] \quad (3.7)$$

$$N' \sim P(\beta_0 e^{-\beta_1 s_n^2}), \quad \beta_0 \sim \Gamma(a, b), \quad \beta \sim \Gamma(c, d),$$

$$N' \perp \beta_0 \perp \beta \quad (3.8)$$

따라서 잠재변수 $N' = N - n$ 를 사용하고 사전분포와 우도 함수를 사용하면 다음과 같은 사후결합분포는 다음과 같은 형태로 나타난다.

$$p(N', \beta_0, \beta | D_{s_n}) \propto \beta_0^a \beta^c \beta^n \left(\prod_{i=1}^n s_i \right) e^{-\beta \sum_{i=1}^n s_i} \exp[-\beta_0 (1 - e^{-\beta_1 s_n^2})] \cdot \frac{1}{N'!} \exp(-\beta_0 e^{-\beta_1 s_n^2}) \cdot (\beta_0 e^{-\beta_1 s_n^2})^N \beta_0^{a-1} \cdot e^{-b\beta_0} \beta^{c-1} \cdot e^{-d\beta} \propto \frac{1}{N'!} \beta_0^{N+n+a-1} e^{-\beta_0(b+1)} \beta^{c-1+n} e^{-\beta(\sum_{i=1}^n s_i + d + s_n^2 N)} \quad (3.9)$$

식 (3.9)에서 베이지 정리와 장애 모수의 개념을 이용하여 깃스 샘플링 알고리즘을 이용하기 위한 주변 사후분포는 다음과 같고 깃스 추출 알고리즘은 Goel모형과 유사한 방법을 이용하면 모수추정이 이루어진다.

$$N' | \beta_0, \beta, D_{s_n} \sim P(\beta_0 e^{-\beta s_n^2}); \quad (3.10.1)$$

$$\beta_0 | N', \beta, D_{s_n} \sim \Gamma(a+n+N', b+1); \quad (3.10.2)$$

$$\beta | N', \theta, D_{s_n} \sim \Gamma\left(c+n, d+s_n^2 N' + \sum_{i=1}^n s_i^2\right); \quad (3.10.3)$$

신뢰도는 식 (2.4)을 연관하여 다음과 같이 나타낼 수 있다 (단, x 는 임무시간을 의미함).

$$\hat{R}(x | s_n) = \exp[-\beta_0 \{ \exp(-\beta_1 s_n^2) - \exp(-\beta_1 (s_n+x)^2) \}] \quad (3.11)$$

3.2.2 라플라스분포 모형

라플라스 분포는 형상모수가 1/2인 와이블 분포의 특수한 경우이므로 평균값 함수는 임의의 시간 $t(=s_n)$ 에 대하여 $m(t) = \beta_0(1 - 1/2 e^{-\beta_1 s_n^2})$, ($\beta_0 > 0, \beta > 0$)으로 표현할 수 있고 강도함수 $\lambda(t) = 1/2 \beta_0 \beta e^{-\beta_1 s_n^2}$ 이 되어 식 (2.2)과 관련하면 우도함수와 사전분포는 다음과 같다.

$$L_{\text{data}}(\beta_0, \beta | D_{s_n}) = \left(\prod_{i=1}^n 1/2 \beta_0 \beta e^{-\beta_1 s_i^2} \right) \exp[-\beta_0 (1 - e^{-\beta_1 s_n^2})] \quad (3.12)$$

$$N' \sim P(\beta_0 e^{-\beta_1 s_n^2}), \beta_0 \sim \Gamma(a, b), \beta \sim \Gamma(c, d), N' \perp \beta_0 \perp \beta \quad (3.13)$$

따라서 잠재변수 $N' = N - n$ 를 사용하고 사전분포와 우도 함수를 사용하면 다음과 같은 결합사후분포 형태로 나타낸다.

$$p(N', \beta_0, \beta | D_{s_n}) \propto \beta_0^n \beta^n e^{-\beta \sum_{i=1}^n |s_i|} \exp[-\beta_0 (1 - e^{-\beta_1 s_n^2})] \\ \cdot \frac{1}{N!} \exp(-\beta_0 e^{-\beta_1 s_n^2}) \cdot (\beta_0 e^{-\beta_1 s_n^2})^N \beta_0^{a-1} \cdot e^{-b\beta_0} \beta^{c-1} \\ \cdot e^{-d\beta} \propto \frac{1}{N!} \beta_0^{N+n+a-1} e^{-\beta_0(b+1)} \beta^{c-1+n} e^{-\beta(\sum_{i=1}^n s_i + d + |s_n| N)} \quad (3.14)$$

식 (2.7)에서 베이지 정리와 장애모수의 개념을 이용하여 깃스 샘플링 알고리즘을 이용하기 위한 주변 사후분포는 다음과 같고 Goel 모형의 알고리즘과 유사한 깃스알고리즘을 시행하면 모수 추정이 이루어진다.

$$N' | \beta_0, \beta, D_{s_n} \sim P(1/2 \beta_0 e^{-\beta_1 s_n^2}); \quad (3.15.1)$$

$$\beta_0 | N', \beta, D_{s_n} \sim \Gamma(a+n+N', b+1); \quad (3.15.2)$$

$$\beta | N', \theta, D_{s_n} \sim \Gamma\left(c+n, d+|s_n| N' + \sum_{i=1}^n |s_i|\right); \quad (3.15.3)$$

신뢰도는 식 (2.4)을 연관하여 다음과 같이 나타낼 수 있다 (단, x 는 임무시간을 의미함).

$$\hat{R}(x | s_n) = \exp[-\hat{\beta}_0 \{ \exp(-\hat{\beta}_1 |s_n|) - \exp(-\hat{\beta}_1 (|s_n|+x)) \}] \quad (3.16)$$

3.3 모형의 선택과 신용구간

효율적 모형의 비교를 위해 고장번호(i)는 시점 s_i 에 관찰된 실제고장의 수 ($n_i(s_i)$)이고 $\hat{m}_i(s_i)$ 는 시점 s_i 에서 추정된 고장의 누적수를 나타내므로 실제 고장수와 추정된 고장의 수의 편차자승합을 계산해서 작은 모형이 효율적인 모형이라고 간주할 수 있다. 즉, 다음과 같이 나타낼 수 있다.

$$C_{SS\hat{E}} = \sum_{i=1}^n (n_i(s_i) - \hat{m}_i(s_i))^2 \quad (3.17)$$

단, n_i 는 $(0, s_i]$ 사이에 관찰된 오류의 수이고 $\hat{m}_i(s_i)$ 는 깃스추정법으로 구한 평균값 함수의 추정값을 의미한다[7, 8, 13].

충분한 반복수(t)를 가진 깃스 샘플링을 통해서 얻은 m 개의 표본을 $\hat{R}_1^{(t)}, \hat{R}_2^{(t)}, \dots, \hat{R}_m^{(t)}$ 으로 표시하면 이 표본들은 확률밀도함수 $f(R(t))$ 의 표본으로 간주되고, 95% 신용구간(CI; Credible interval)은 아래와 같이 됨이 알려져 있다[2].

$$[\hat{R}_{0.05m}^{(t)}, \hat{R}_{0.95m}^{(t)}]$$

단, 0.05 m 과 0.95 m 는 각각 0.05 m 번째와 0.95 m 번째의 순서 통계량을 의미하고 m 은 추출된 표본의 크기를 나타낸다.

4. 수치적인 예

고장간격 시간 x_i 에 대한 자료는 실제 관측 데이터[8]를 이용하여 모수를 추정하고 신뢰도를 구하고자 한다. 관측 데이터는 최종고장시간 $s_{25} = 250$ ($s_0 = 0$)이고 각 고장 간격시간에 대한 자료는 <표 1>에 요약하였다.

본 연구는 이 자료를 이용하여 모수추정과 신뢰도, 모형선택을 시행하고자 한다. 1990년도 초에 Hossain과 Dahiya[9]는 이 NTDS 자료를 가지고 최우추정법을 사용한 (베이저안 방법이 아닌 일반적인 방법) Goel 모형에 대한 추정치 $\hat{\beta}_0$ 는 33.99, $\hat{\beta}_1$ 는 0.00579를 추정한 바 있는데 본 연구의 추정치와 거의 같은 추세를 보이고 있다.

<표 1> NTDS 데이터($x_i = s_i - s_{i-1}$)

고장번호 (i)	고장간격시간 (x_i)	누적고장시간 (s_i)	고장번호 (i)	고장간격시간 (x_i)	누적고장시간 (s_i)
1	9	9	14	9	87
2	12	21	15	4	91
3	11	32	16	1	92
4	4	36	17	3	95
5	7	43	18	3	98
6	2	45	19	6	104
7	5	50	20	1	105
8	8	58	21	11	116
9	5	63	22	33	149
10	7	70	23	7	156
11	1	71	24	91	247
12	6	77	25	2	249
13	1	78	26	1	250

본 연구에서는 사전분포는 $\beta \sim \Gamma(1, 0.0001)$, $\beta_0 \sim \Gamma(1, 0.0001)$ 을 선택 이용하여 (3.1)절에 제시한 깃스 알고리즘을 적용하였다. IMSL¹⁾ 소프트웨어[22]를 사용하여 각 깃스 열의 전 반부 $t/2$ 번 반복을 제외하고 후반부 $t/2$ 번 반복만을 고려하는 기법 즉, 분산 분석표를 이용하는 Gelman & Rubin (1992)이 제시한 Marko Chain 방법을 이용하여 각 모형에 대한 사후밀도의 결과를 <표 2>, <표 3>, <표 4>에 나타내었다.

<표 2> Coel 모형에 대한 사후 밀도

모수	적용수(m)	반복수(t)	$\hat{E}(N')$	S.D	95% C.I
N'	500	50	7.778	5.423	(4.115, 19.011)
		70	7.879	5.310	(4.564, 18.016)
	2000	50	7.875	5.148	(4.989, 17.119)
		70	7.872	4.498	(5.126, 17.231)
모수	적용수(m)	반복수(t)	$\hat{E}(\beta_0)$	S.D	95% CI
β_0	500	50	32.237	3.987	(27.011, 39.982)
		70	31.678	3.478	(27.233, 39.764)
	2000	50	31.023	3.419	(27.652, 39.484)
		70	31.011	3.382	(28.981, 38.541)
모수	적용수(m)	반복수(t)	$\hat{E}(\beta)$	S.D	95% CI
β	500	50	0.0011	0.0028	(0.0009, 0.0211)
		70	0.0048	0.0021	(0.0015, 0.0198)
	2000	50	0.0057	0.0019	(0.0025, 0.0114)
		70	0.0059	0.0017	(0.0034, 0.0092)

<표 3> 라플라스 모형에 대한 사후 밀도

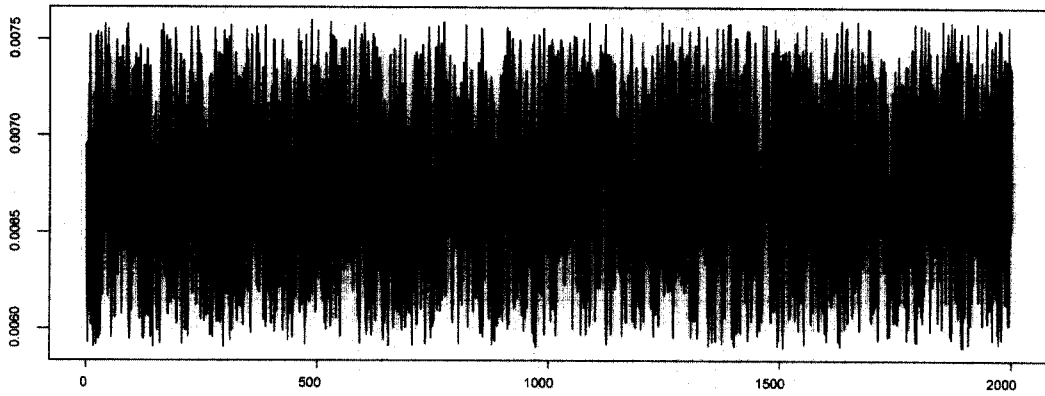
모수	적용수(m)	반복수(t)	$\hat{E}(N')$	S.D	95% C.I
N'	500	50	7.778	5.423	(4.115, 19.011)
		70	7.879	5.310	(4.564, 18.016)
	2000	50	7.875	5.148	(4.989, 17.119)
		70	7.872	4.498	(5.126, 17.231)
모수	적용수(m)	반복수(t)	$\hat{E}(\beta_0)$	S.D	95% CI
β_0	500	50	32.432	3.438	(24.012, 39.991)
		70	32.356	3.236	(24.121, 39.834)
	2000	50	32.193	3.219	(24.323, 39.574)
		70	32.034	3.182	(24.953, 39.346)
모수	적용수(m)	반복수(t)	$\hat{E}(\beta)$	S.D	95% CI
β	500	50	0.00045	0.0028	(0.00013, 0.00201)
		70	0.00046	0.0021	(0.00043, 0.00156)
	2000	50	0.00048	0.0019	(0.00037, 0.00143)
		70	0.00043	0.0017	(0.00026, 0.00141)

<표 4> 레일리 모형에 대한 사후 밀도

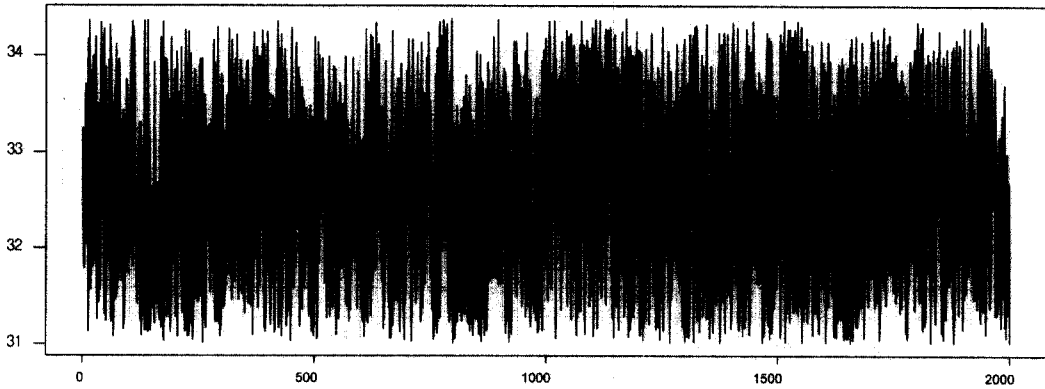
모수	적용수(m)	반복수(t)	$\hat{E}(N')$	S.D	95% C.I
N'	500	50	7.142	2.998	(4.115, 17.063)
		70	6.887	2.989	(4.215, 16.984)
	2000	50	6.989	2.935	(4.492, 16.732)
		70	6.984	2.925	(5.234, 16.439)
모수	적용수(m)	반복수(t)	$\hat{E}(\beta_0)$	S.D	95% CI
β_0	500	50	31.313	2.418	(24.083, 39.875)
		70	31.311	2.286	(24.221, 39.843)
	2000	50	31.243	2.259	(25.343, 38.585)
		70	31.154	2.132	(26.858, 37.442)
모수	적용수(m)	반복수(t)	$\hat{E}(\beta)$	S.D	95% CI
β	500	50	0.0001065	0.0391	(0.000094, 0.002653)
		70	0.0001164	0.0301	(0.0000134, 0.002123)
	2000	50	0.0001122	0.0289	(0.0000421, 0.001945)
		70	0.0001124	0.0241	(0.0000723, 0.001593)

수렴성을 확인하기 위해서 500, 2000(m)번 적용에 50, 70(t)번의 결과를 나타내었다. 그리고 각 모수에 대한 사후 평균 $\hat{E}(\theta)$ (단, θ 는 모수)과 신용구간, 표준편차를 나타내었다. 따라서 이 표들에서 보여 주듯이 거의 유사한 값에 수렴함을 볼 수 있기 때문에 모형선택이나 신뢰도에 있어서 2,000번 적용에 70번 반복한 사후평균(추정값)을 이용하였다. 그리고 Goel 모형의 사후 표본 분포도의 결과를 (그림 1), (그림 2), (그림 3)에 나타내었다.

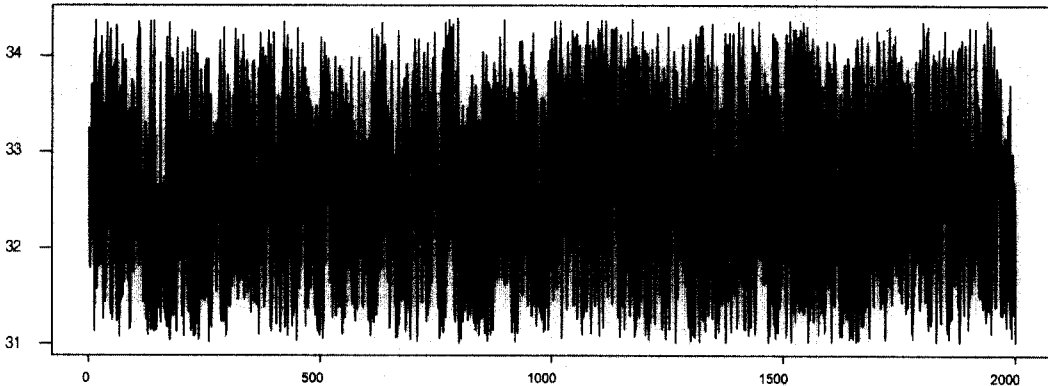
1) MS-FORTRAN IMSL(INTERNATIONAL MATHEMATICAL AND STATISTICAL LIBRARY(PC 버전 4.0))



(그림 1) Goel 모형에 대한 모수 β 의 표본 분포도



(그림 2) Goel 모형에 대한 모수 β_0 의 표본 분포도



(그림 3) Goel 모형에 대한 모수 N' 의 표본 분포도

이 그림에서도 β, β_0 의 분포도는 각각 0.0059, 31.011를 중심으로 분포되고 있고 N' 는 7.872를 중심으로 분포되어 있음을 알 수 있다. 다른 모형의 모수도 유사한 추세로 분포되고 있음을 짐작할 수 있다.

모형 선택에 있어서는 편차자승합을 이용한 모형비교를 시행하였다. <표 5>에서는 Goel 모형에 대한 베이지 추정치(평균값 함수)와 편차 자승값과 편차 자승합을 계산하는 방법을 나타내었다. 유사한 방법으로 다른 모형에 대한 편차자

승합을 계산한 결과는 <표 6>에 요약되었다. 이 결과표에서 편차자승합을 최소화 하는 레일리 모형이 효율적인 모형으로 간주할 수 있다.

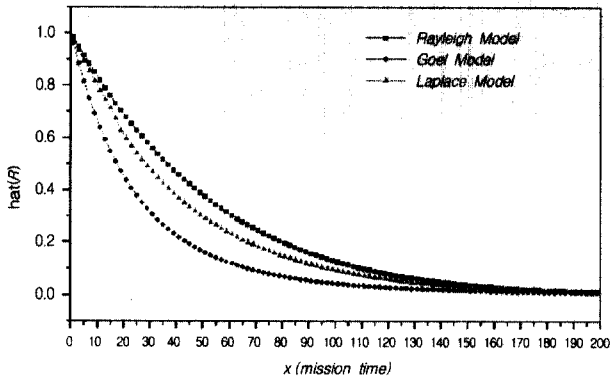
(그림 4)은 신뢰도 함수를 이용하여 미래의 신뢰도 즉, 실제 고장수가 26이후의 신뢰도($\hat{R}(x|s_{26}=250)$)를 계산하여 비교하고자 한다. 이 그림에서 시간이 지날수록 모형들이 감소 추세를 보이고 있고 다른 모형보다 레일리 모형이 비교적 높음을 보이고 있다.

〈표 5〉 Goel모형에 대한 평균값함수의 베이지 추정값 ($\hat{m}_i(s_i)$)과 편차자승합(C_{SSE})

고장번호 (i)	고장간격시간 (x_i)	누적고장시간 (s_i)	실제고장수 ($n_i(s_i)$)	평균값함수의 추정값 ($\hat{m}_i(s_i)$)	편차자승의 값 ($(n_i(s_i) - \hat{m}_i(s_i))^2$)
1	9	9	1	1.971935355	0.944658334
2	12	21	2	4.407521438	5.796159473
3	11	32	3	6.460260395	11.973402
4	4	36	4	7.166776655	10.02847438
5	7	43	5	8.354608892	11.25340082
6	2	45	6	8.682989193	7.198431009
7	5	50	7	9.48326766	6.166618269
8	8	58	8	10.70448074	7.314216079
9	5	63	9	11.43230510	5.916108088
10	7	70	10	12.40764427	5.796750911
11	1	71	11	12.54295428	2.380707911
12	6	77	12	13.33439559	1.780611602
13	1	78	13	13.46296496	0.214336552
14	9	87	14	14.57881387	0.335025499
15	4	91	15	15.05169603	0.002672479
16	1	92	16	15.16777474	0.692598875
17	3	95	17	15.51096968	2.217211303
18	3	98	18	15.84673034	4.636570219
19	6	104	19	16.49658955	6.267063896
20	1	105	20	16.60215895	11.54532383
21	11	116	21	17.71395310	10.79810486
22	33	149	22	20.56058197	2.071924258
23	7	156	23	21.08118368	3.681856064
24	91	247	24	25.90078243	3.612973853
25	2	249	25	25.9748496	0.950331747
26	1	250	26	26.01147964	0.000131782
C_{SSE}					123.5756641

〈표 6〉 편차자승합에 의한 모형비교

	Goel 모형	라플라스분포 모형	레이리분포 모형
C_{SSE}	123.5756641	98.543289	68.213575



(그림 4) 각 모형에 대한 신뢰도

5. 결 론

소프트웨어의 신뢰성은 개발의 최종단계에 있는 테스트 공정이거나 실제 사용 단계에 있어서 소프트웨어 내에 존재하는

오류 수나 소프트웨어의 고장발생시간에 의해 효과적 평가를 할 수 있는 것으로 그 평가 기술이 중요하게 된다. 소프트웨어 개발의 테스트 공정이거나 실제 사용단계에 있어서 오류 발생상황이나 소프트웨어 고장 발생현상을 수리적 모형화가 가능하다면 평가를 할 수 있다. 테스트의 개발 상황의 파악, 테스트에 의해 미 발생되었던 고장에 대한 보수 및 비용의 예측 등 구체적인 소프트웨어 개발의 보수관리문제에도 적용 가능하다. 따라서 테스트 시간 혹은 실행시간과 발생한 고장 수나 소프트웨어 고장의 발생시간과의 관계를 소프트웨어 신뢰도 성장과정이라 볼 수 있다[2].

본 논문에서는 각 모수에 대한 사전정보가 있는 비동질 포아송 과정에 기초한 소프트웨어 오류 현상에 대한 베이지안 확률 모형을 고려하고 모수의 추정을 위해 잠재변수를 사용한 깃스 추정량을 이용하였다. 고장 패턴은 NHPP에 대한 강도함수와 평균값 함수로서 나타낼 수 있다. 따라서 본 논문에서는 기존의 모형인 Goel이 제시한 모형과 신뢰도 측면에서 많이 사용되는 와이블 분포의 특수형태인 레일리분포와 라플라스분포를 이용한 모형을 제시하고 또, 효율적 모형을 위한 모형선택으로서 편차자승합을 이용하여 비교하였다.

수치적인 예에서는 실측자료인 NTDS 자료를 이용하여 모

수 및 신뢰도를 추정하였고 편차자승합을 이용한 모형비교의 결과를 나열하였다.

그 결과 Goel 모형은 NTDS 데이터에서는 비효율적인 모형으로 간주되고 본 논문에서 제시한 레일리분포 모형은 편차자승합이 다른 모형에 비해 작으므로 보다 효율적인 모형이 됨을 알 수 있었다.

신뢰도 추정을 위하여 실제고장수가 26 이후의 신뢰도를 계산한 결과 임무시간(x)에 따른 신뢰도함수는 전체적으로 신뢰도가 비증가 추세를 보이고 있음을 알 수 있다.

향후 추가 정보를 이용한 베이지안 모수 추정 과정에서 조건부 함수의 식별이 어려운 경우 메트로폴리스-헤스팅스 알고리즘을 사용한 모수 추정과 이에 대한 분포 이론 및 응용에 관한 연구가 기대된다.

참 고 문 헌

- [1] Charnes, A., Frome, E. L. and Yu, P. L., "The Equivalence of Generalized Least Squares and Maximum Likelihood Estimates in the Exponential Family," *Journal of the American Statistical Association*, Vol.7, No.353, pp.169-171, 1976.
- [2] Dey, D. K., and Lee, T. M., "Bayes Computation for Life Testing and Reliability Estimation," *IEEE Transactions on Reliability*, Vol.41, pp.621-626, 1992.
- [3] Gelfand, A. E., and Smith, A. F. M., "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, pp.398-409, 1990.
- [4] Gelfand, A. E., and Smith, A. F. M. and Lee, T-M., "Bayesian analysis of constrained parameter and truncated data problems," *J. Amer. Statist. Assoc.*, 87, pp.523-532, 1992.
- [5] Gelman, A. E., and Rubin D., "Inference from Iterative Simulation Using Multiple Sequences," *Statistical Science*, 7, pp.457-472, 1992.
- [6] Geman, S., and Geman, D., "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, pp.721-741, 1984.
- [7] Goel, A. L., "Software Reliability Model : Assumption, Limitations, and Applicability," *IEEE Trans. Software Eng.*, Vol.SE-11, No.12, pp.1411-1423, 1985.
- [8] Goel, A. L., and Okumoto, K., "Time-Dependent Error-Detection Rate Model for Software Reliability and Other Performance Measures," *IEEE Trans. Rel.*, Vol.R-28, pp. 206-211, 1979.
- [9] Hossain, S. A. and Dahiya, R. C., "Estimating the Parameters of a Non-homogeneous Poisson-Process Model for Software Reliability," *IEEE Trans. Rel.*, Vol.R-42, No.4, pp.604-612, 1993.
- [10] Johnson, N. L. and Kotz, S., "*Continuous Univariate Distributions 1*," A wiley-interscience publication, JOHN WILEY & SONS, pp.272-289, 1970.
- [11] Kuo, L., and Yang, T. Y., "Bayesian Computation of Software Reliability," *Journal of Computational and Graphical Statistics*, pp.65-82, 1995.
- [12] Lawless, J. F., "*Statistical Models and Methods for Lifetime Data*," pp.494-500, 1981.
- [13] Musa, J. D., Iannino, A., and Okumoto, K., "*Software Reliability : Measurement, Prediction, Application*," New York : McGraw Hill, 1987.
- [14] Pham, H. and Nordmann, L. and Zhang, X. "A General Imperfect-Software-Debugging Model with S-Shaped Fault-Detection Rate," *IEEE Trans. Rel.*, Vol.48, No.2, pp.169-175, 1999.
- [15] Okumoto, K., "A Statistical Method for Software Quality Control," *IEEE Trans. Software Eng.*, Vol.SE-11, No.12, pp.424-1431, 1985.
- [16] Raftery, A. E., "Inference and Prediction for a General Order Statistic Model with Unknown Population Size," *Journal of the American Statistical Association*, 82, pp.1163-1168, 1987.
- [17] Smith, A. F. M. and G. O. Roberts, "Bayesian computation via the Gibbs Sampler and related Markov chain Monte Carlo methods," *Journal of the Royal Statistical Society, Ser. B*, 55, pp.3-24, 1993
- [18] Tanner, M., and Wong, W. "The Calculation of Posterior Distributions by Data Augmentation (with discussion)," *Journal of the American Statistical Association*, 81, pp.82-86, 1987.
- [19] Yamada, S., Ohba, M. and Osaki, S., "S-shaped Reliability Growth Modeling for Software Error Detection," *IEEE Trans. Rel.*, Vol.R-32, No.5, pp.475-478, 1983.
- [20] Yamada, S., Ohtera, H. and Narihisa, H., "Software Reliability Growth Models with Testing-Effort," *IEEE Trans. Rel.*, Vol.R-35, No.1, pp.19-23, 1986.
- [21] Yamada, S. and Osaki, S., "Software Reliability Growth Modeling : Models and application," *IEEE Trans. Software Eng.*, Vol.SE-11, No.12, pp.1431-1437, 1985.
- [22] "USER'S MANUAL "STAT/LIBRARY FORTRAN Subroutines for statistical analysis," IMSL, Vol.3, pp.1050-1054, 1987.
- [23] 간광연, 김희철, 이병수, "NHPP에 기초한 소프트웨어 신뢰도 모형에 대한 비교연구", 정보처리학회논문지 D, 제8-D권, 제4호, pp.393-400, 2001.
- [24] 박관근, "소프트웨어 신뢰도의 평가와 예측을 위한 베이지안 알고리즘", 정보처리학회논문지, Vol.1, No.1, pp.14-22, 1226-9182, 1994.

이 상 식

e-mail : leess@songho.ac.kr
 2000년 경희대학교 대학원 전자계산공학과
 공학석사
 2002년 경희대학교 대학원 전자계산공학과
 박사과정수료
 2001년~현재 송호대학 정보산업계열 전임
 강사

관심분야 : 소프트웨어공학, 소프트웨어신뢰성, S/W 재사용

김 희 철

e-mail : khc@songho.ac.kr
 1998년 동국대학교 대학원 통계학과 이학
 박사
 2000년 현재 송호대학 정보산업계열 조교수
 관심분야 : 소프트웨어신뢰성공학, 웹프로그
 램밍, 전산통계

송 영 재

e-mail : yjsong@khu.ac.kr
 1969년 인하대학교 전자공학과 공학사
 1976년 일본 Keio 대학교 전산학과 공학
 석사
 1980년 명지대학교 전산학과(공학박사)
 1982년~1983년 미국Maryland대학교 객원
 교수

1986년~1988년 대한전자공학회 전자계산연구회전문위원장

1984년~1989년 전국 전산소장 협의회부회장

1990년~1991년 일본 Keio 대학교객원교수

1984년~1989년 경희대학교전자계산소장

1993년~1995년 경희대학교교무처장

1996년~1998년 경희대학교공과대학장

1999년~2000년 경희대학교 기획조정실장

2001년~현재 경희대학교산업정보대학원장

1976년~현재 경희대학교 교수

관심분야 : 소프트웨어공학, OOP/S, CASE 도구, S/W 재사용