신뢰도 데이터 부족시의 처리방법에 대한 연구

이 광원, 오 신규, 한 정민

호서대학교 안전공학과, 한국가스공사 연구개발원*

Management of insufficient Reliability Data

Kwang-Won Rhie, Shinkyu Oh*, Jeongmin Han*
Hoseo Uni., KOGAS R&D Center*

1. 서론

관찰되어진 고장수가 적을 때 (대략 12개 이하)는 추정되어지는 신뢰도 수치의 값이 신뢰성을 갖지 못하게 된다. 이때 일반 신뢰도 데이터 book에서 신뢰도 수치가 존재한다면 관찰되어진 데이터(specific 데이터)와 문헌에서 찾은 데이터(generic 데이터)를 bayes 이론을 통하여 동시에 고려 할 수가 있다.

베이스 정리란 18C 영국 수학자 Bayes가 특수한 사건이 주어진 조건하에서 어떤 사건이 일어날 확률을 산출하는 이론을 정리한 것이며 근래에는 부품이나 시스템의 고장확률 등의 데이터수정에 많이 쓰이고 있다. 이후 2-stage Bayesian procedure는 Shultis etval(1981)과 Kaplan(1983)에 의하여 처음으로 공표 되었으며 Hora & Iman (1990), Papazoglou et al.(1991), Pörn(1993) 등에 의하여 연구 발전되었다.

2. 본론

정량적 평가를 위하여 어떤 부품의 신뢰도 데이터를 수집하던 중 문헌에 나타난 일반적 데이터와 고유 신뢰도자료 등을 분석하여 얻은 specific 데이터 2개를 얻었을 때 이들을 모두 고려한 신뢰도 수치를 얻고자 할 때 bayes 이론을 사용한다. 즉,

- prior probability : 추가적인 정보가 있기 전의 어떤 사건에 대한 사전적 확률, 즉 문헌 등에 나타난 확률
- evidence : 추가적인 정보, 즉 최근 고유 신뢰도자료 등을 분석하여 얻은 확률
- posterior probability : 추가적인 정보와 사전적 확률을 고려한 사후적 확률

이라 하고, P(Y)를 evidence 즉, 고유 신뢰도자료 등을 분석하여 얻은 확률이라 하고 P(X)가 사전적 확률이라 하면 사후적 확률은

$$P(X \mid Y) = \frac{P(Y \mid X) \cdot P(X)}{P(Y)} = \frac{P(Y \mid X) \cdot P(X)}{\sum P(Y \mid X) \cdot P(X)}$$

로 표현된다.

일반적 One stage Bayes이론은 관찰되어지는 모든 부품은 같은 고장률이나 고장확률을 갖는다고 가정되어진다. 하지만 경험적으로 관찰하면 같은 부품일지라도 사용장소가 틀리면 운전조건이나 보수정책 등에 의하여 서로 다르게 고장이 일어난다. 이런 관점을 고려한 방법을 2-stage-Bayesian procedure 또는 super population method 라하며 수학적으로 관심 있는 수치, 즉 사후적 확률을 확률변수로 취급한다.

각기 다른 plant등에서 사용되어지는 같거나 비슷한 부품의 subpopulation이 존재하고 각각의 subpopulation이 k개 존재하며 각 subpopulation당 값 E=(n,T)나 E=(n,N)이 존재한다고 가정하자. 여기서 n은 고장 수를 의미하며 T나 N은 관찰시간이나 요구횟수를 의미한다. 일반성의 상실 없이 evidence E1에 대한 관찰을 한다 가정하고 다른 것들을 E_2 ...... E_k 라 하자.

고장률이나 고장확률의 불확실성 계산을 위하여 이들은 어떤 미지변수 Ψ에 의존되는 분포를 갖는 확률변수라 가정한다. 미지 parameter들은 다음과 같은 Bayesian 측정 식에 의하여 구하여진다.

$$g(\underline{\Psi \mid E}) = \frac{g(\underline{\Psi})L(\underline{E \mid \Psi})}{\int_{\Psi} g(\underline{\Psi})L(\underline{E \mid \Psi})d\underline{\Psi}} \qquad (식1)$$

여기서 $g(\Psi)$는 prior 분포도함수이며 미지변수의 선행정보를 의미한다. subpopulation들이 서로 독립적이라 가정하면 likelihood함수 $L(E|\Psi)$는

$$L(\underline{E \mid \Psi}) = \left(\prod_{i=2}^{k} L(E_i \mid \underline{\Psi}) \right) \qquad (식2)$$

로 써지게 된다. 우리가 관심있는 고장률 λ나 고장확률 p를 x라 표현하면 총확률의 법칙에 의하여

$$L(E_i \mid \underline{\Psi}) = \int_x L(E_i \mid x) g(x \mid \underline{\Psi}) dx \qquad (식3)$$

으로 써지게 되며 고장률은

$$L(E_i \mid x) = L((n_i, T_i) \mid \lambda) = \frac{1}{n_i!} \lambda^{n_i} e^{\lambda T_i} \qquad (식4)$$

로 써지게 되고 요구횟수당 고장확률은 다음과 같다.

$$L(E_i \mid x) = L((n_i, T_i) \mid P) = \binom{N_i}{n_i} (1-p)^{N_i \ n_i} p^{n_i} \qquad (식5)$$

이 식들로부터 분포밀도함수 $g(\lambda|E)$가 주어진다. 다시 한번 총확률의 법칙을 적용하면 변수 x에 대한 분포밀도함수를 구할 수 있으며 이것은 prior distribution으로 사용되어진다.

$$f(x \mid \underline{E_i}) = \int_{\underline{\Psi}} g(x \mid \underline{\Psi}) g(\underline{\Psi} \mid E) d\underline{\Psi} \qquad \text{(식6)}$$

마지막으로 subpopulation E_1에 대한 고장률이나 요구횟수당 고장확률을 결정하기 위한 Bayes이론은 다음과 같이 나타나게 된다.

$$f(x \mid E_1, \underline{E}) = \frac{L(E_1 \mid x) f(x \mid \underline{E})}{\int_x L(E_1 \mid x) f(x \mid \underline{E}) dx} \qquad \text{(식7)}$$

하지만 이 식은 계산에 있어서 닫혀있지가 않다. (식3)는 고장율에 대하여는 gamma 분포를 그리고 요구횟수당 고장확률에 대하여는 beta 분포를 가정하면 analytical하게 풀 수 있다. 하지만 각 경우에 lognormal 분포를 가정하면 풀지를 못한다. (식1)에서 2번의 적분이 요구되는데 이는 지금까지 numerical하게 또는 Monte Carlo방법을 사용하여 풀 수밖에 없다고 알려져 있다.

3. Bayes program

Bayes program은 일반신뢰도 data 들과 고유경험 data 들을 기초로 하여 고장확률이나 고장률을 예측하기 위해 windows 환경에서 손쉽게 사용할 수 있도록 개발된 program 이다. Bayes 2.0은 문헌에서 superpopulation 식 또는 2단계 Bayes이론이라고 알려진 방법을 사용하였다.

지금까지 널리 사용되어진 고전적 (또는 1단계) Bayes 이론은 관찰되어지는 모든 고장들이 같은 고장률이라는 전제 조건이 필요하지만 2단계 Bayes 이론은 이런 전제조건이 필요가 없다.

여러 경험상 비슷한 부품들일지라도 설치조건에 따라서 고장현상이 다르게 발생된다는 것이 증명되고 있으며 이런 사항은 subpopulation 개념에 반영시킬 수 있다. 하나의 subpopulation 안에 묶여지는 부품들을 그들의 성질과 외적설치조건이 거의 차이가 없으며, 같은 고장률이나 고장확률을 갖는다고 전제한다. 관찰되는 전체(superpopulation)는 보통 여러 개의 sub population 들로 구성되어 이들 subpopulation 들은 각각 고장률이나 고장확률은 갖는다고 가정한다.

실제적으로 superpopulation 은 보통 비슷한 종류와 비슷하게 설치된 부품들로 정의된다. subpopulation 는 이들 부품들이 서로 다른 설치조건을 가질 때 분류한다. 일반적으로 여러 종류의 부품들의 정리된 부분적인 부품들을 subpopulation 으로 정의, 사용한다.

수학적 풀이를 가능케 하기 위하여 가정이 필요하다. 즉 subpopulation 들의 고장률이나 고장확률들을 확률적 우연변수라 가정하면 미지의 변수를 갖는 어떤 분포함수 즉 superpopulation 분포를 갖게된다. 이러한 분석에서 제일먼저 할 일은 superpopulation 분포의 미지변수를 결정하는 일이다. 이 미지변수결정은 Bayes 이론에 의해 가능하며 이때 결과 값으로 변수의 복합분포, 즉 변수분포를 얻게된다. 변수결정의 input data는 관찰되어진 subpopulation 의 고장들과 일반

신뢰도 data가 되며 일반적으로 일반신뢰도 data는 lognormal 분포모델을 적용하고 median 값과 *k-factor* 값을 주는 것이 보통이다. 얻어진 변수분포로부터 궁극적으로는 총확률의 법칙을 사용하여 고장률이나 고장확률의 분포를 얻게된다.

만약 전체적인 고장률이나 고장확률에 대하여 관심이 있다면 이로서 해답을 얻게 되며 만약 어떤 하나의 subpopulation에 관심이 있다면 다시 한 번 이 subpopulation 에 대하여 Bayes 이론을 적용하여 얻을 수 있다. 이 경우 여러 번 같은 data를 Bayes 이론에 적용하면 틀린 값이 나오므로 관찰되어지는 subpopulation 은 처음에는 사용하지 말아야 한다.

본program 은 visual basic으로 code화 하였다. Input data file을 읽어서 여러개의 output data file을 생성하며 다음과 같은 model들에 대하여 program화하였다.

1. lognormal 분포의 superpopulation 에 있어서 고장률의 계산
2. gamma 분포의 superpopulation 에 있어서 고장률의 계산
3. beta 분포의 superpopulation 에 있어서 고장확률 계산
4. lognormal 분포의 superpopulation 에 있어서 고장확률 계산

이들 Model 들은 변수가 2개나 3개인 mixing 분포에는 사용하지 못한다.

superpopulation 분포에서 변수와 고장률이나 고장확률의 값들을 자동으로 주어진 data에 의하여 계산되어 진다. 빈약한 input data는 쓸모 없는 결과 값을 계산할 수도 있다. 그런 것을 방지하기 위해 input data에서 모든 변수범위를 미리 줄 수 있도록 하였다. 계산정확성(여러 적분계산 과정에서 발생되는 절단수의 크기와 이에 따른 결과치의 수)는 input data에서 조절할 수 있다. 즉 4단계 계산의 정확도를 높일 수 있도록 하였다.

계산은 2단계 Bayes 이론 (superpopulation 방법)을 사용하였다. 이때 일반적으로 많이 사용되어지는 Ignorance - Prior 방법을 사용하였다. 즉 구형분포를 사용하여 계산하였다. 단 한가지 예외는 lognormal 분포모델에서의 변수분포 계산과정이며 이때에는 수학적 계산상의 이유로 *k-factor* 의 값을 계산하기 위하여 prior 값은 input data에서 줄 수 있다.

input data 는 주어진 형식에 의해 주게되며 결과는 다음과 같이 output data에 주어진다. 즉 평균값, 표준편차 5% 백분율, 50% 백분율, 95% 백분율 등의 수치 값들과 그래프 값(횡좌표, prior 값, posterior 값) 그리고 고장률이나 고장확률을 표현해주는 분포곡선(보통 21개의 좌표 값을 계산하여 이은)을 주게된다.

a. 결과표

결과표에는 분석의 중요 결과 값들이 표현된다.

결과 값으로서 평균값, 표준편차, 5%, 50%, 95% 백분율들을 포함한다. 또한 text 형식으로 부품명과 형식 고장유형, 그리고 특정 subpopulation을 관찰한 경유는 이 subpopulation 의 이름을 포함한다.

b. 계산점

계산과정에서 적분 등을 할 경우 21곳의 계산점을 계산하게 되며 이때 계산되어지는 계산점에서의 경험적 분포도함수값 들을 수록한다. 또한 경험적 분포함수

의 평균값과 표준편차들도 주어지게 된다.

c. Plot data

prior 값과 posterior 값들을 포함하는 data이다. 이 data 에는 각 줄에 3개의 값들 즉 고장률이나 고장확률 값과 prior 결과 값의 밀도, posterior 결과 값의 밀도들을 포함한다. 이 data는 많은 graphic program 에 input data로 활용될 수 있으며 이로서 분포함수의 그림을 얻어낼 수 있다.

d. Input data

파일의 입력부를 클릭 한 후 다음과 같이 입력 데이터를 입력한다.

a) 프로젝트명 : text type으로 32자 이내

b) 저장할 file 이름 : 저장되는 file 명, 같은 이름으로 결과 data가 생성된다. (*.kur, *.lan) text type으로 32자 이내

c) 관찰부품의 명칭

d) 고장유형 : text type으로 32자 이내(빈칸 없이 입력)

e) 계산의 정확도 : option 값으로 하나만 선택가능

f) 경험으로 수집된 data subpopulation 의 개수 : 1 이상의 자연수

g) 문헌에서 수집된 data의 개수 : 1 이상의 자연수

h) 분포모델

I) Mixingparameter

j) 고장률이나 고장확률의 하한값을 쓰시오(λ/p 또는 default 값인 0)

k) 고장률이나 고장확률의 상한값을 쓰시오(λ/p 또는 default 값인 0)

입력부 폼을 이용하여 만들어진 input data file의 1번째 줄은 프로젝트명, 2번째 줄은 파일이름, 3, 4번째는 자동으로 생성되는 결과파일, 5번째 줄은 관찰부품, 6번째 줄은 고장유형, 7번째 줄은 차례대로 계산의 정확도, 경험으로 수집된 data subpopulation 의 개수, 문헌에서 수집된 data의 개수, 분포모델, Mixing parameter, 고장률이나 고장확률의 하한값, 고장률이나 고장확률의 상한값 들이며 8번째 줄은 bayes 2.0에서 꼭 필요한 데이터이다. 9번째 줄은 차례대로 경험 데이터의 사건 수, 관찰된 시간 또는 가동수, subpopulation 이름, 10번째 줄은 차례대로 평균값, k-factor, 문헌이름 이다.

e. Output data

Output data는 크게 결과 확인부의 계산, plot data 및 파일부의 결과확인이 있다. 파일 이름은 입력부 폼에 입력한 파일이름.bat를 선택한다.

결과확인부에는 a) 분석일련번호 b) 부품유형 c) 고장유형 d) 분석종류 e) 평균값, 표준편차, 5% 백분율, 50% 백분율, 95% 백분율 등을 계산하여 보여준다.

또한 plot data는 'test.lan' file에 저장되어진다. 이 file은 header 첫줄에 주어진 주석문이 쓰여진다. 그 다음 줄에는 각각의 부품유형, 고장유형, 분석유형이 차례로 써지게 된다. 각 구간에서는 시작점과 끝점에서의 고장률/확률과 구간 내에서 사용된 분포밀도값들이 주어진다.

4. 참고문헌

1. Hofer, E. et al. "On the Solution Approach for Bayesian Modelling of Initiating Event Frequencies and Failure Rates", Risk Analysis 17, 1997.
2. Hora, S.C & Iman, R.L., "Bayesian Modelling of Initiating Event Frequencies at Nuclear Power Plants", Risk Analysis 10, pp. 103-109, 1990.
3. Kaplan, S., "On a two stage Bayesian procedure for determining failure rates from experimental data", IEEE Transactions on Power Apparatus and Systems, PAS-102, 1983.
4. Poern, K. "Methods for reliability data analysis-A new perspective, in: Kafka, P.,J.(eds), Safety and Reliability Assesment-An Integral Approach", Proc. ESREL '93 Conference in Copenhagen, Elsevier 1993, pp. 101-108, 1993.
5. Papazoglou, I. et al. "Bayes - A Computer Code for Data Analysis under Lack of Knowledge and Population Variability", National Center for Scientific Research Demokritos, Athen, Juni 1991