한국항해항만학회지 제37권 제5호
J. Navig. Port Res. Vol. 37, No. 5 : 453-461, October 2013 (ISSN:1598-5725(Print)/ISSN:2093-8470(Online))
DOI : http://dx.doi.org/10.5394/KINPR.2013.37.5.453

해상 부유체 모델의 표본 데이터에 대해서 최대우도를 갖는 누적분포함수 추정

† 임정빈 · 양원재*

†,* 목포해양대학교 해사대학 해상운송시스템학부 교수

Estimating Cumulative Distribution Functions with Maximum Likelihood to Sample Data Sets of a Sea Floater Model

† *Jeong-Bin Yim · Won-Jae Yang**

†,**Division of Maritime Transportation System, College of Maritime Sciences, Mokpo Maritime University, Mokpo 530-729, Korea*

요 약 : 본 논문에서는 소형 해상 부유체의 위기 평가를 위한 확률기반 위기평가기법(PET)에서 표본 데이터에 최적인 누적분포함수(CDF) 추정에 관한 평가절차와 실험결과를 기술하였다. CDF는 PET에서 부유체의 위기수준을 평가하기 위한 위기허용기준의 참조 값을 제공하기 위한 것으로, 부유체 모델의 롤(Roll), 피치(pitch), 히브(Heave) 등의 운동응답함수에 대한 표본 데이터에서 추정할 수 있다. 본 연구에서는 여덟 가지 정형화된 분포함수와 최대우도추정기법을 적용하여 표본 데이터에 대해서 최대우도를 갖는 CDF들을 평가하였다. 분포함수들의 적합도 검정 실험을 통해서, 베타 분포가 롤과 피치 표본 데이터에 대해서 평균 확률오차 $\bar{\delta}$ $(0 \le \bar{\delta} \le 1.0)$가 가장 작은 0.024와 0.022로 최적임을 나타냈고, 히브 표본 데이터에 대해서는 감마 분포가 $\bar{\delta}$ 가 가장 작은 0.027로 최적임을 나타냈다. 본 연구에서 제안한 방법은 표본 데이터의 최적 분포 추정을 위한 다양한 분야에 적용 가능할 것으로 기대된다.

핵심용어 : 해상 부유체, 위기평가, 위기허용기준, 누적분포함수, 최대우도추정기법

Abstract : *This paper describes evaluation procedures and experimental results for the estimation of Cumulative Distribution Functions (CDF) giving best-fit to the sample data in the Probability based risk Evaluation Techniques (PET) which is to assess the risks of a small-sized sea floater. The CDF in the PET is to provide the reference values of risk acceptance criteria which are to evaluate the risk level of the floater and, it can be estimated from sample data sets of motion response functions such as Roll, Pitch and Heave in the floater model. Using Maximum Likelihood Estimates and with the eight kinds of regulated distribution functions, the evaluation tests for the CDF having maximum likelihood to the sample data are carried out in this work. Throughout goodness-of-fit tests to the distribution functions, it is shown that the Beta distribution is best-fit to the Roll and Pitch sample data with smallest averaged probability errors* $\bar{\delta}$ *$(0 \le \bar{\delta} \le 1.0)$ of 0.024 and 0.022, respectively and, Gamma distribution is best-fit to the Heave sample data with smallest* $\bar{\delta}$ *of 0.027. The proposed method in this paper can be expected to adopt in various application areas estimating best-fit distributions to the sample data.*

Key words : *sea floater, risk evaluation, risk acceptance criteria, cumulative distribution function, maximum likelihood estimates*

1. 서 론

선행연구(Yim, 2012)에서, 소형 해상 부유체의 안전성 평가를 위해서 확률기반 위기평가기법(Probability based risk Evaluation Techniques, PET)을 제안한 바 있다. PET는 부유체 모델의 운동응답함수에 대한 표본 데이터에서 누적분포함수를 구한 후, 이를 위기평가의 기준이 되는 위기허용기준(Risk Acceptance Criteria)으로 적용하는 방법이다.

한편, 구조물의 안전성 평가를 위한 기존 방법들은 위기평가를 위한 기준 값 획득에 복잡한 대규모 시험대(test bed)가 필요한 문제점이 있었는데(Charle et. al., 2001; Lu et. al., 2008), PET는 평가기준으로 누적분포함수를 이용하기 때문에 기존 방법과 비교하여 간단하고 경제적인 장점이 있다. 그러나 운동응답함수의 표본 데이터가 불연속적인 불완전 데이터이며, 분포특성을 알 수 없어서 통계적으로 유의한 정형화된 누적분포함수를 적용할 수 없는 문제가 있었다.

† 교신저자 : 임정빈(종신회원) jbyim@mmu.ac.kr 061)240-71702
* 종신회원, wjyang@mmu.ac.kr 061)240-71762
주) 이 논문은 "소형 해상 부유체의 위기허용수준 결정을 위한 최적의 누적확률분포함수 선정에 관한 연구"란 제목으로 "한국항해항만학회 2013년도 춘계학술대회(2013.6.27, pp.474-476)"에서 발표한 연구의 후속 연구임.

본 연구의 목적은 선행연구의 문제점 해결에 있다. 이를 위해 표본 데이터에 대해서 정형화된 여덟 가지 누적분포를 평가한 후, 최대우도(maximum likelihood)를 갖는 누적분포함수의 종류와 형상 변수 값을 추정하였다.

2. 기존 위기평가기법의 문제점과 해결방안

2.1 기존 위기평가기법의 문제점

Fig. 1은 선행연구(Yim, 2012)의 확률기반 위기평가기법(Probability based risk Evaluation Techniques, PET)을 적용한 해상 부유체 모델의 간단한 위기수준 결정절차이다.

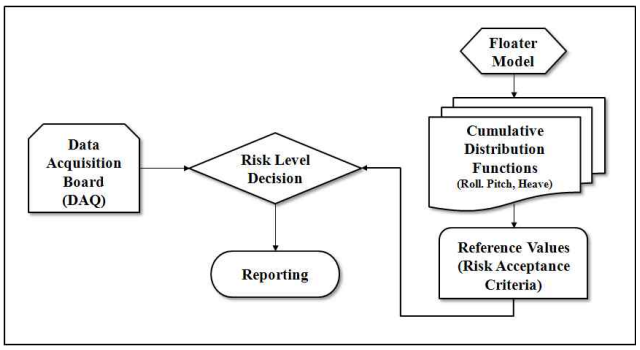
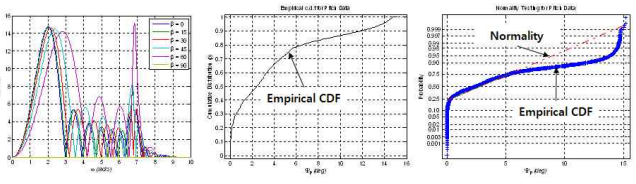


Fig. 1 Simplified flow-diagram to determine the risk levels of the sea floater using Probability based risk Evaluation Techniques (PET)

Fig. 1의 우측에 나타낸 부유체 모델(Floater Model)에서 롤(Roll), 피치(Pitch), 히브(Heave) 등 세 가지 운동응답함수의 표본 데이터를 획득한 후, 이에 대한 누적분포함수(Cumulative Distribution Functions, CDF)를 구한다. 그리고 각 CDF에서 부유체 위기평가의 기준이 되는 참조 값(Reference Values)을 구한 후, 위기허용기준(Risk Acceptance Criteria)을 구축한다. 다음에는 그림 좌측과 같이 실제 해상 부유체에 설치한 데이터 획득 보드(Data Acquisition Board, DAQ)에서 세 가지 운동 값을 측정한 후, 미리 구축한 위기허용기준과 비교하여 부유체의 위기수준을 결정한다. 이와 같이 PET는 부유체 모델에서 획득한 CDF에서 위기허용기준을 구하여 해상 부유체를 평가하는 방법이다. 따라서 CDF를 구하는 것이 중요한데, 그 개념을 Fig. 2에 나타냈다.

Fig. 2(a)는 선행연구에서 획득한 피치 운동응답함수의 표본 데이터를 나타내고, (b)는 표본 데이터에 대한 경험적 누적분포함수(Empirical CDF, E-CDF), (c)는 E-CDF에 대한 정규분포검증(normality test) 등을 나타낸다.

여기서, Fig. 2(a)의 표본 데이터는 길이 10.0 m, 폭 2.4 m, 높이 1.0 m의 부유체에 대해서 José(2009)가 제안한 응답함수 계산식을 이용하여 컴퓨터 시뮬레이션으로 구한 피치운동함수 중에서 일부를 발췌한 표본 데이터이다. 이 표본 데이터는 최대 유의 파고가 1.0(m)인 해양 파를 0.01(rad/sec.)의 주파수 간격으로 0.01부터 10.0까지 주고, 부유체와 조우하는 해상 파의 입사 각도를 5.0 도 간격으로 0 도부터 90 도까지 주었을 때 계산한 결과이다.



(a) (b) (c)

Fig. 2 The response sample data of Pitch motion by a computer simulation(a), it's Empirical cumulative distribution functions(b) and the results by normality tests(c)(Yim, 2012)

한편, 선행연구에서 문제로 제기된 것이 Fig. 2(b)이다. Fig. 2(b)는 Fig. 1에서 기술한 CDF가 아니라 경험적 누적분포함수(E-CDF)이다. E-CDF는 평균과 표준편차와 같은 통계 변수를 이용하는 정형화된 CDF와 달리 표본 데이터 그 자체에 적응시킨 누적분포함수이다(MATLAB, 2008c). 따라서 E-CDF를 이용하여 부유체를 평가하기 위해서는 Fig. 2(b)의 모든 데이터를 알고 있어야 하고, 통계 변수만으로 CDF를 간단하게 정의할 수 없어서 평가과정이 복잡하며, 표본 데이터가 불연속적인 불완전 데이터이기 때문에 E-CDF의 통계적인 유의성이 낮다.

한편, 정형화된 CDF를 적용하기 위해서는 표본 데이터의 분포특성을 알아야 하는데, 일반적으로 실험 현장이나 모델에서 획득한 표본 데이터는 불연속적인 불완전 데이터이기 때문에 어떠한 통계처리를 거치기 전까지는 데이터의 분포특성을 알 수 없다(David, 2010; Lee and Oh, 1996; Staub and Gekenidis, 2011). 이러한 이유 때문에 선행연구에서는 E-CDF의 일부 데이터만을 이용하여 해상 부유체를 평가하면서 이 문제를 추후 연구과제로 남겨둔 바 있다.

그리고 선행연구에서는 Fig. 2(a)의 표본 데이터에 대한 정규분포 여부만을 파악한 바 있는데 그 결과가 Fig. 2(c)이다. Fig. 2(a)의 표본 데이터가 정규분포를 갖는다면 Fig. 2(c)에 나타낸 'Empirical-CDF'는 'Normality'로 표시한 점선과 일치하게 되는데, 상당한 부분이 점선에서 벗어나 있다. 따라서 Fig. 2(a)의 표본 데이터는 정규분포가 아님을 알 수 있다.

2.2 해결방안 검토

위에서 검토한 바와 같이 부유체 모델에서 획득한 운동응답함수의 표본 데이터는 정규분포함수로 대표할 수 없음을 알았다. 그래서 다양한 비정규분포에 관한 연구를 조사한 결과, 불완전 표본 데이터에 최적으로 근사된 분포함수들의 형상 변

수 추정에 관한 다양한 방법이 보고되어 있음을 확인하였다.

Plancade(2013)과 Brehenry(2013) 등은 불완전 데이터에 최적인 분포함수 탐색 기법을 보고한 바 있는데, 특히 R-forge project(2013)에는 다양한 비정규분포함수의 정의와 모델링 기법이 기술되어 있고, 최적의 분포함수 탐색에는 최대우도추정기법(Maximum Likelihood Estimation)(MATLAB, 2008b; Wikipedia, 2013c)이 널리 적용되고 있음을 알았다.

그래서 본 연구에서는 현재 공학 분야에 널리 적용되고 있는 감마(Gamma), 베타(Beta) 등 여덟 가지 분포함수에 최대우도추정기법을 적용하여 불완전 데이터에 대해서 최대우도를 갖는 분포함수와 형상 변수를 추정하였다. 그리고 이러한 추정 과정에 필요한 프로그램을 모두 작성하기 위해서는 많은 시간이 걸리기 때문에, 다양한 상용 누적분포함수 코드(code)와 최적 분포함수 탐색 코드 등(MATLAB, 2008b and 2008c; MathWorks, 2013)을 이용하였다.

3. 누적분포함수 평가와 선정

3.1 연구 접근방법

누적분포함수(CDF)는 어떠한 형상 변수(Shape Parameter, SP)를 갖는 확률분포함수를 적분하여 계산할 수 있다. 연속 누적분포함수를 $F(\cdot)$ 로 정의하면, 다음 식(1)과 같이 주어진 변수 값 x 에 대해서 연속 변수 X 로 정의되는 확률분산 p 에서 계산할 수 있고, 주어진 구간 $[a, b]$에 대한 p 은 다음 식(2)과 같이 확률분포함수 $f(\cdot)$ 의 적분으로 구할 수 있다(MathWorks, 2013; Wikipedia, 2013a).

$$F(x) = p[X \le x] \tag{1}$$

$$p[a \le X \le b] = \int_a^b f(x)dx \tag{2}$$

그러나 실제 현장이나 모델에서 구한 데이터는, 그 분포특성을 일단 모르기 때문에 다양한 통계처리 과정을 거치기 전까지는 정형화된 CDF를 적용할 수 없다(David, 2010). 이러한 경우의 대안으로는, 실제 표본 데이터를 이용하여 CDF를 추정할 수 있는데, 이미 앞에서 설명한 경험적 누적분포함수(E-CDF)가 대표적인 방법이다. E-CDF는 식(3)과 같이 전체 표본 데이터 중에서 주어진 값 보다 작거나 같은 데이터를 세어서 누적분포함수 F_{emp} 을 계산한다(Breheny, 2013; MATLAB, 2008c).

$$F_{emp}(x) = \frac{1}{n}\sum_{i=1}^{n} Ind\,(x_i \le x) \tag{3}$$

여기서, x_i 는 n 개의 주어진 표본 데이터($i = 1,2,\cdots,n$), Ind 는 표시함수.

식(3)에 나타낸 E-CDF는 표본 데이터 그 자체를 가장 잘 표현한 CDF의 일종으로, 케플란-메이어 추정자(Kaplan-Meier Estimator)(Staub and Gekenidis, 2011)로 불리는 방법이다. E-CDF는 주로 의학 분야에서 생존 함수를 추정하기 위한 것으로, DKW(Dvoretzky-Kiefer-Wolfowitz) 부등식을 이용하여 신뢰구간을 계산할 수 있다(Breheny, 2013).

그러나 수학적으로 대단히 복잡하고 단순한 형상 변수로 정의할 수 없다. 그래서 본 연구에서는 표본 데이터에 최적인 정형화된 다양한 CDF를 평가한 후, 위의 식(3)의 E-CDF에 가장 근사된 CDF를 탐색하였다.

3.2 평가를 위한 누적분포함수 선정

CDF에는 다양한 종류가 있는데, 일부 CDF는 특수한 목적으로 사용되고, 현재까지 연구된 모든 CDF를 평가하는 것은 곤란하다. 그래서 본 연구에서는 다음과 같은 자료를 조사하여 공학 분야에 널리 적용되고 있는 CDF만을 평가대상으로 정하였다.

Jun and Yoo(2012)은 기상관측의 불완전 데이터 문제를 해결하기 위하여 베타(Beta) 함수를 적용한 바 있고, Riddhi(2013)는 베타 함수 적용을 위한 이론적인 방법을 제공하였으며, Joel(2013)은 베타 함수와 유사한 감마(Gamma) 함수 이론의 적용방법을 제안한 바 있으며, Kim 등(2013)은 통신운용에 감마 함수를 적용한 바 있다. 특히, Plancade(2013)는 주어진 데이터에 대한 조건부 CDF를 추정함에 있어 대단히 독특한 적응 기법을 제안하였고, R-forge project(2013)에는 연속 및 비연속의 다양한 정규분포와 비정규분포에 대해서 방대한 정보가 기술되어 있다. 이 외에도 MATLAB(2008a)에는 공학 분야에 널리 적용되고 있는 지수(Exponential), 극치(Extreme Value), 로그정규(Lognormal), 정규(Normal), 프아송(Poisson), 레일리히(Rayleigh) 등의 분포함수가 소개되어 있다.

그래서 본 연구에서도 베타, 감마, 지수, 극치, 로그정규, 정규, 프아송, 레일리히 등 여덟 가지 분포함수를 평가대상으로 정하였다. 아래 식(4)부터 식(11)까지는 이러한 여덟 가지 누적분포함수들의 누적확률 계산식을 나타낸 것으로, 모두 MATLAB(2008a)에 기술되어 있는 내용을 참고한 것이다.

우선, 아래의 누적확률 계산식에 나타낸 표본 데이터를 정의하면 다음과 같다. ψ ($\in \{\psi_R, \psi_P, \psi_H\}$)는 부유체 모델의 세 가지 운동응답함수의 표본 데이터를 정의한 것으로, ψ_R ($\in \{\psi_1, \psi_2, \psi_3, \cdots, \psi_{n_R}\}$)은 n_R 개의 롤(Roll) 표본 데이터 집합, ψ_P ($\in \{\psi_1, \psi_2, \psi_3, \cdots, \psi_{n_P}\}$)는 n_P 개의 피치(Pitch) 표본 데이터 집합, ψ_H ($\in \{\psi_1, \psi_2, \psi_3, \cdots, \psi_{n_H}\}$)는 n_H 개의 히브(Heave) 표본 데이터 집합 등을 나타낸다.

이러한 표본 데이터 ψ 에 대한 지수누적분포함수(Exponential CDF) $F_{\exp}$ 의 누적확률 $p_{\exp}$는 다음 식(4)로 나타낼 수 있다.

$$p_{\exp} = F_{\exp}(\psi \parallel \mu) = \int_0^{\psi} \frac{1}{\mu} e^{\frac{t}{\mu}} dt = 1 - e^{\frac{\psi}{\mu}} \tag{4}$$

여기서, μ 는 CDF의 형상(shape)을 결정짓는 변수로써 데이터 평균을 의미하고, e 는 지수(exponential), t 는 적분을 위한 변수, $F(\psi \parallel \mu)$ 는 μ 로 결정되는 ψ 에 대한 CDF를 의미하고, 심벌 ' $\parallel$ '은 주어진 형상 변수에 대응한다는 의미이다.

그리고 ψ 에 대한 극치누적분포함수(Extreme Value CDF) F_{evc} 의 누적확률 p_{evc}는 식(5)로 나타낼 수 있다.

$$p_{evc} = F_{evc}(\psi \parallel \mu, \sigma) = \sigma^{-1} e\left(\frac{\psi - \mu}{\sigma}\right) e\left(-e\left(\frac{\psi - \mu}{\sigma}\right)\right) \tag{5}$$

여기서, σ 는 편차를 의미한다.

또한, ψ 에 대한 로그정규누적분포함수(Lognormal CDF) F_{logn} 의 누적확률 p_{logn}는 식(6)으로 나타낼 수 있다.

$$p_{logn} = F_{logn}(\psi \parallel \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_0^{\psi} \frac{e^{\frac{-(\log_e(t) - \mu)^2}{2\sigma^2}}}{t} dt \tag{6}$$

그리고 ψ 에 대한 정규누적분포함수(Normal CDF) F_{norm} 의 누적확률 p_{norm}는 식(7)로 나타낼 수 있다.

$$p_{norm} = F_{norm}(\psi \parallel \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\psi} e^{\frac{-(t-\mu)^2}{2\sigma^2}} dt \tag{7}$$

아울러 ψ 에 대한 프아송누적분포함수(Poisson CDF) F_{poiss} 의 누적확률 p_{poiss}는 식(8)로 나타낼 수 있다.

$$p_{poiss} = F_{poiss}(\psi \parallel \lambda) = e^{-\lambda} \sum_{i=0}^{floor(\psi)} \frac{\lambda^i}{i!} \tag{8}$$

여기서, λ 는 형상 변수, $floor(\psi)$은 ψ 값과 같거나 큰 정수에 대한 바닥 값, $i!$는 i 의 팩토리얼(factorial)을 의미한다.

그리고 ψ 에 대한 감마누적분포함수(Gamma CDF) F_{gam} 의 누적확률 p_{gam}는 식(9)과 같다.

$$p_{gam} = F_{gam}(\psi \parallel a, b) = \frac{1}{b^a \Gamma(a)} \int_0^{\psi} t^{a-1} e^{\frac{t}{b}} dt \tag{9}$$

여기서, a 와 b 는 형상 변수이고, $\Gamma(\cdot)$는 다음 식(9-1)로 표현되는 감마함수이다.

$$\Gamma(a) = \int_0^{\infty} e^{-t} t^{a-1} dt \tag{9-1}$$

또한, ψ 에 대한 베타누적분포함수(Beta CDF) F_{beta} 의 누적확률 p_{beta}는 식(10)과 같다.

$$p_{beta} = F_{beta}(\psi \parallel a, b) = \frac{1}{B(a,b)} \int_0^{\psi} t^{a-1}(1-t)^{b-1} dt \tag{10}$$

여기서, $B(\cdot)$ 는 형상 변수 a 와 b 로 구성된 베타함수로써 다음 식(10-1)과 같이 나타낼 수 있다.

$$B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1} dt = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \tag{10-1}$$

여기서, $\Gamma(\cdot)$ 는 위의 식(9-1)의 감마함수이다. 그리고 위의 식(10)에 나타낸 베타 함수는 표본 데이터가 반드시 0과 1 사이에 분포해야하기 때문에 표본 데이터를 최댓값으로 정규화 시켜서 적용하였다.

마지막으로, ψ 에 대한 레일리히누적분포함수(Rayleigh CDF) F_{rayl} 의 누적확률 p_{rayl}는 식(11)과 같이 나타낼 수 있다.

$$p_{rayl} = F_{rayl}(\psi \parallel b) = \int_0^{\psi} \frac{t}{b^2} e^{\left(\frac{-t^2}{2b^2}\right)} dt \tag{11}$$

3.3 평가 방법과 절차

우선, 표본 데이터 ψ 에 대해서 최대우도를 갖는 여덟 가지 CDF의 형상 변수를 최대우도추정기법(Maximum Likelihood Estimates, MLE)을 적용하여 추정하였다. MLE는 주어진 확률 모델에 최적인 형상 변수를 추정하기 위한 방법인데, 확률 모델에 어떠한 형상 변수가 주어질 때 원하는 값들이 나올 우도(likelihood)를 최대화시키는 특정 형상 변수를 결정하는 방법이다. 즉, 어떠한 확률 모델에서 획득한 n 개의 표본 데이터 집합을 $X(\in \{x_1, x_2, \cdots, x_n\})$로 가정하고, 이 표본 데이터들의 분산은 미지의 확률밀도함수 $f_{ukn}(\cdot)$에서 구한 것이라 가정한다. 만약 $f_{ukn}(\cdot)$가 형상 변수 α 을 갖는 어떠한 분산 $\{f(\cdot \parallel \alpha), \alpha \in A\}$에 속한다면, $f_{ukn} = f(\cdot \parallel \alpha_{ukn})$가 되고, α_{ukn} 에 근접한 추정자 $\hat{\alpha}_{MLE}$을 찾는 것이 MLE 기법이다 (MATLAB, 2008b; Wikipedia, 2013b and 2013c).

따라서 부유체 모델의 표본 데이터 ψ 에 대해서 위의 식(4)부터 식(11)까지에 나타낸 분포함수들을 MLE 기법에 적용하여 추정자 $\hat{\alpha}_{MLE}$를 구하면 최대우도를 갖는 각 분포함수의 형상 변수를 구할 수 있다.

Table 1은 여덟 가지 분포함수들을 평가하기 위하여 본 연구에서 이용한 MATLAB(2008a)의 분포함수 명칭(a)과 누적확률 계산을 위한 상용 코드(code)의 명칭(b) 및 형상 변수의 종류(c) 등을 정리한 것이다.

Table 1 List of distribution types and MATLAB codes with shape parameters used in the evaluations procedures

(a)	(b)	(c)
Type of distributions	MATLAB codes to calculate the cumulative probability	Shape Parameters as shown in Eq.(4) ~ Eq.(11)
Exponential	**expcdf**	μ
Gamma	**gamcdf**	a , b
Extreme	**evcdf**	μ , σ
Lognormal	**logncdf**	μ , σ
Normal	**normcdf**	μ , σ
Poisson	**poisscdf**	λ
Beta	**betacdf**	a , b
Rayleigh	**raylcdf**	b

Fig. 3은 분포함수들을 평가하기 위한 절차를 나타낸 것으로, Step 1부터 Step 3까지 세 부분으로 구분되어 있다. 각 단계별로 설명하면 다음과 같다.

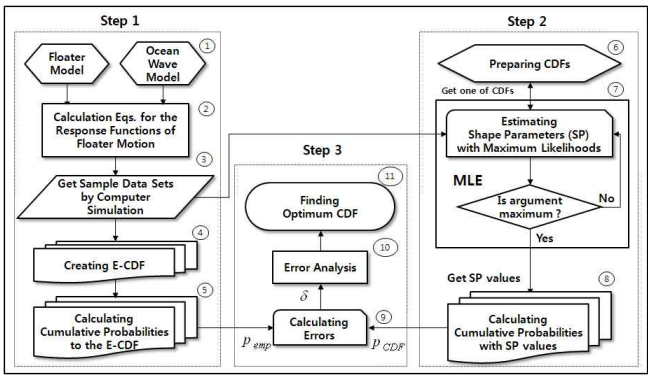


Fig. 3 Evaluation procedures to find optimum distribution model giving best-fit to the sample data

- **Step 1** : ①부터 ⑤까지는 표본 데이터 ψ 에 대한 경험적 누적분포함수(E-CDF)의 생성과 E-CDF의 누적확률을 계산하는 부분이다. 앞의 Fig. 2에서 설명한 부유체와 해상파 모델 제원(①)을 부유체 응답함수 계산식에 입력하여(②), 부유체의 롤, 피치, 히브 등의 응답함수 ψ 을 획득한 후(③), ψ 자체에 적응시킨 경험적 누적분포함수(E-CDF)를 생성하고(④), 이에 대한 누적확률 p_{emp} 을 계산한다(⑤). 여기서, p_{emp} 는 위의 식(3)에 나타낸 E-CDF의 분포함수 F_{emp} 을 이용하여 다음 식(12)로 구한 것이다.

$$p_{emp} = F_{emp}(\psi) \tag{12}$$

- **Step 2** : ⑥부터 ⑧까지는 ψ 에 대해서 여덟 가지 CDF들의 형상 변수를 MLE 기법으로 추정한 후, 평가를 위하여 다시 형상 변수에 대한 CDF들의 누적확률을 계산하는 부분이다. 우선, Table 1(a)에 나타낸 여덟 가지 CDF들을 준비한 후(⑥), 앞에서 설명한 MLE 기법을 적용하여 ψ 에 대해서 CDF들이 최대우도를 갖는 Table 1(c)의 형상 변수(SP)를 추정(⑦)한다. 이 때 사용한 MATLAB 상용 코드는 '**SP = mle(data, 'distribution', dist)**'이다. **mle**는 MLE 기법을 의미하고, **data**는 주어진 ψ , **'distribution'**은 분포함수임을 의미한 것으로, **dist**에 Table 1(a)에 나타낸 분포함수의 종류를 기술한다. 예를 들어, 감마 분포함수를 적용하는 경우, '**SP = mle(ψ, 'distribution', Gamma)**'로 선언하면, 최대우도를 갖는 감마 분포함수의 형상 변수 값이 **SP**에 출력된다. 이와 같은 방법으로 여덟 가지 CDF 각각에 대한 형상 변수 값을 추정한다. 그리고 평가를 위해서 Table 1(b)의 MATLAB 상용 코드를 이용하여 **SP**에 저장된 값으로 누적확률(⑧) p_{CDF} 을 계산한다. 이 때 이용한 MATLAB 상용 코드는 '**p(dist) = (dist)cdf(data, SP)**'이다. 예를 들어 감마 분포함수에 대한 누적확률의 경우는 '**p(Gamma) = gamcdf(ψ , SP)**'로 선언하면 된다. **gamcdf**는 Table 1(b)에 나타낸 감마 누적분포함수의 상용 코드이다.

- **Step 3** : ⑨부터 ⑪까지는 ψ 에 대해서 최적인 CDF를 선정하는 부분이다. Step 1에서 구한 E-CDF의 p_{emp} 와 Step 2에서 구한 여덟 가지 CDF들의 p_{CDF} 사이의 확률 오차 δ 을 구하여(⑨) 오차를 분석한 후(⑩), 가장 오차가 작은 CDF를 선정(⑪)한다. 여기서, δ 는 다음 식(13)으로 구하고, 평균 확률 오차 $\bar{\delta}$ 는 식(14)로 구하였다.

$$\delta_{k,j}(i) = \left| p_{emp_k}\left[\psi_k(i)\right] - p_{CDF_{k,j}}\left[\psi_k(i)\right] \right| \tag{13}$$

$$\bar{\delta}_{k,j} = \frac{\sum_{i=1}^{n_k} \delta_{k,j}(i)}{n_k} \tag{14}$$

여기서, k 는 부유체 모델의 세 가지 운동응답함수에 대한 표본 데이터의 종류($k=1,2,3$), j 는 여덟 가지 CDF의 종류($j=1,2,\cdots,8$), p_{emp_k} 는 k 번째 표본 데이터 ψ_k 에 대한 E-CDF의 누적확률, $p_{CDF_{k,j}}$ 는 k 번째 표본 데이터 ψ_k 에 대한 j 번째 CDF의 누적확률, $\psi_k(i)$ 는 k 번째 표본 데이터에서 i 번째 데이터($i=1,2,\cdots,n_k$이고, n_k 은 k 의 데이터 수), $|\cdot|$ 는 절대 값, $\delta_{k,j}(i)$는 j 번째 CDF에 대해서 k 번째 표본 데이터의 i 번 데이터의 확률 오차, $\bar{\delta}_{k,j}$ 는 $\delta_{k,j}(i)$에 대한 평균 확률 오차.

4. 평가 및 결과

4.1 누적분포의 형상 변수 추정 결과

Table 2는 부유체 모델의 세 가지 운동응답함수에 대한 표본 데이터 ψ_R, ψ_P, ψ_H 등에 대해서 여덟 가지 CDF를 적용하여 MLE 기법으로 추정한 형상 변수(SP) 값들이다.

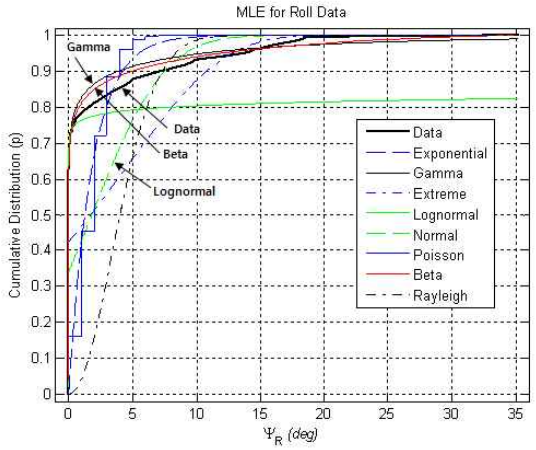
Table 2 Estimating results for the shape parameters of each distribution types according to the sample data sets, ψ_R, ψ_P and ψ_H

Distribution Types and Shape Parameters			Values of Shape Parameters					
			ψ_R		ψ_P		ψ_H	
Exponential	μ	-	5.986	-	4.604	-	0.538	-
Gamma	a	b	0.692	8.645	0.966	4.763	1.778	0.303
Extreme	μ	σ	9.471	8.194	6.954	4.895	0.720	0.321
Lognormal	μ	σ	0.915	1.601	0.927	1.278	-0.926	0.843
Normal	μ	σ	5.986	6.348	4.604	4.315	0.538	0.361
Poisson	λ	-	5.986	-	4.604	-	0.538	-
Beta	a	b	0.578	2.684	0.635	1.288	0.607	0.386
Rayleigh	-	b	6.169	-	4.462	-	0.459	-

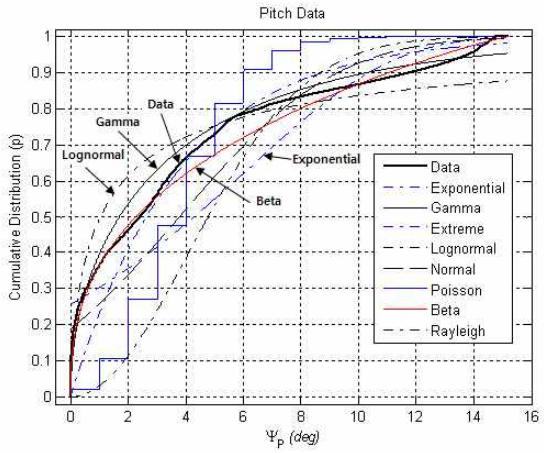
Table 2의 SP 값들은 여덟 가지 CDF들이 표본 데이터 ψ에 대해서 최대우도를 갖도록 MLE에서 최적으로 추정된 것이다. 한편, Table 2의 SP 값들은 주어진 CDF에 대해서는 최적의 값들이지만, 우리가 원하는 것은 ψ에 대해서 최적인 하나의 CDF를 선정하는 것이기 때문에 본 연구에서는 E-CDF에 가장 근사된 CDF를 선정하였다. 그 이유는, E-CDF는 표본 데이터 그 자체에 적응시킨 것으로 데이터의 고유한 특징을 내포하고 있기 때문이다.

Fig. 4에 Table 2의 SP 값들을 갖는 여덟 가지 CDF들의 누적확률 $p_{CDF_{k,j}}$과 E-CDF에 대한 누적확률 p_{emp_k}을 나타냈다. 이 그림에서 p_{emp_k}는 굵은 실선으로 나타내고 'Data'로 표시했고, 여덟 가지 p_{emp_k}는 범례와 같이 다양한 종류의 선으로 나타냈다. 그리고 p_{emp_k} 중에서 시각적으로 'Data'에 가장 근사된 몇 가지를 별도로 표시했다.

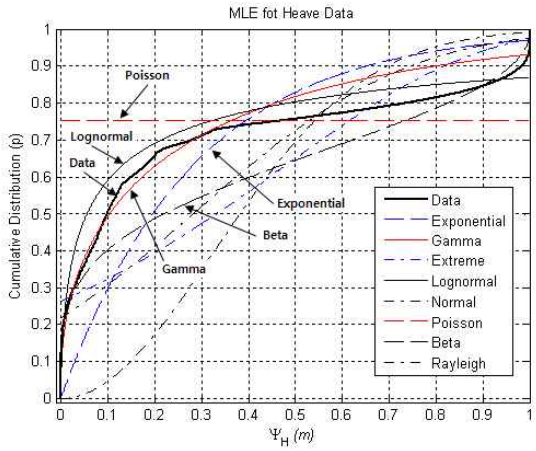
Fig. 4(a)의 롤 표본 데이터 ψ_R에서, 'Data'에 가장 근사된 누적분포는 베타(Beta), 감마(Gamma), 로그정규(Lognormal) 등으로 고려된다. Fig. 4(b)의 피치 ψ_P와 Fig. 4(c)의 히브 ψ_H 등에서도 Fig. 4(a)와 유사한 결과를 보인다. 주로 베타와 감마 분포함수가 E-CDF와 근사적으로 유사한 것으로 고려된다. 한편, E-CDF에 최적 근사된 하나의 CDF를 정하기 위해 E-CDF와 여덟 가지 CDF 사이의 오차를 분석하고 평가하였다.



(a) In case of Roll sample data, ψ_R



(b) In case of Pitch sample data, ψ_P



(c) In case of Heave sample data, ψ_H

Fig. 4 Fitting results for the eight kinds of CDFs to the E-CDF according to the sample data sets, ψ_R, ψ_P and ψ_H

4.2 누적분포의 오차 평가 결과

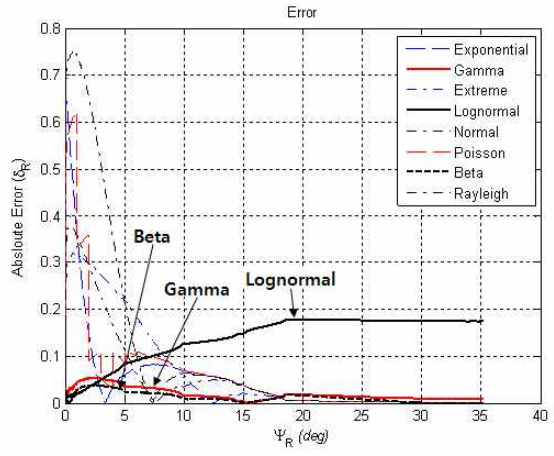
Fig. 5는 위의 식(13)으로 계산한 확률 오차 $\delta_{k,j}(i)$을 나타낸다. 전반적으로 x-축의 표본 데이터 값이 작은 부분에서 가장 큰 오차가 발생하다가 점차 감소하지만 표본 데이터 값이 큰 부분에서 다시 오차가 커지는 경향을 보이고 있다. 이러한 경향은 위의 Fig. 4에서 확인할 수 있는데, 표본 데이터 자체에 적응시킨 E-CDF의 누적확률 값들은 불연속적이면서 급격한 증감 특성을 가짐에 반하여 여덟 가지 CDF들은 연속적이면서 부드러운 증감특성을 갖기 때문이다. Fig. 5를 전반적으로 살펴보면, 베타(Beta), 감마(Gamma), 로그정규(Lognormal), 지수(Exponential) 등의 분포함수가 다른 함수와 비교하여 확률 오차가 작게 나타내고 있다. 이하에서는 이러한 네 가지 CDF를 집중 검토한다.

Fig. 5(a)의 롤 표본 데이터 ψ_R의 경우는 베타와 감마 함수가 전반으로 작은 오차를 나타내는데, 로그정규 함수의 경우는 초기 오차는 작으나 ψ_R 값이 커지면서 큰 오차가 발생하고 있다. Fig. 5(b)의 피치 표본 데이터 ψ_P의 경우는 베타, 감마, 지수 등의 함수가 작은 오차를 나타낸다. 특히 베타 함수가 초기 오차가 작다가 다시 약간 증가한 후 감소하고 있다. Fig. 5(c)의 히브 표본 데이터 ψ_H의 경우는 베타, 감마, 로그정규 등의 함수가 작은 오차를 나타낸다. 특히 감마 함수가 초기 오차가 작다가 다시 일정하게 증가한 후 감소하는 특징을 나타내고 있다. 로그정규 함수의 경우는 초기 오차는 감마 함수와 비교하여 크지만 이후부터는 감마 함수보다 작다.

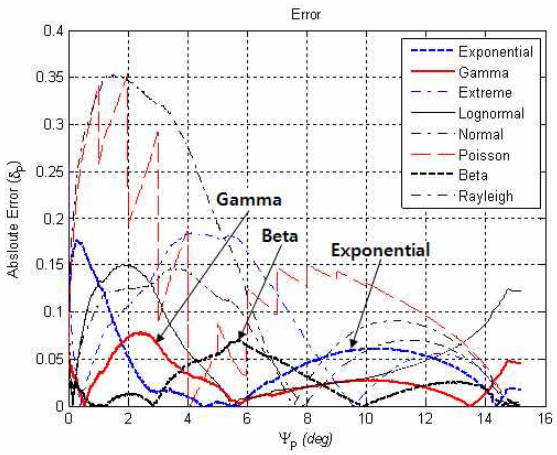
E-CDF에 최적으로 근사된 하나의 CDF를 선정하기 위하여 위의 식(14)의 평균 확률오차 $\bar{\delta}_{k,j}$을 구해서 Table 4에 나타냈다. 평균 확률오차 우측에 나타낸 괄호안의 숫자는 평균 오차가 작은 순위를 나타낸다. 롤 표본 데이터 ψ_R에 대해서는 베타 함수가 $\bar{\delta}_{k,j}$이 가장 작은 0.024를 나타내고, 이어서 감마 함수가 0.031을 나타낸다. 피치 표본 데이터 ψ_P 역시 베타 함수가 $\bar{\delta}_{k,j}$이 가장 작은 0.022를 나타내고, 이어서 감마 함수가 0.033을 나타낸다. 반면, 히브 표본 데이터 ψ_H는 감마 함수가 $\bar{\delta}_{k,j}$가 가장 작은 0.027을 나타내고, 이어서 로그정규 함수가 0.057을 나타낸다.

한편, 선행연구(Yim, 2012)에서 고찰한 정규분포(Normal)의 경우, 롤과 피치의 표본 데이터에서 $\bar{\delta}_{k,j}$가 0.053과 0.091을 나타내어 공동 5위를 나타냈고, 히브 표본 데이터의 경우에는 $\bar{\delta}_{k,j}$가 0.122를 나타내어 4위로 나타났다.

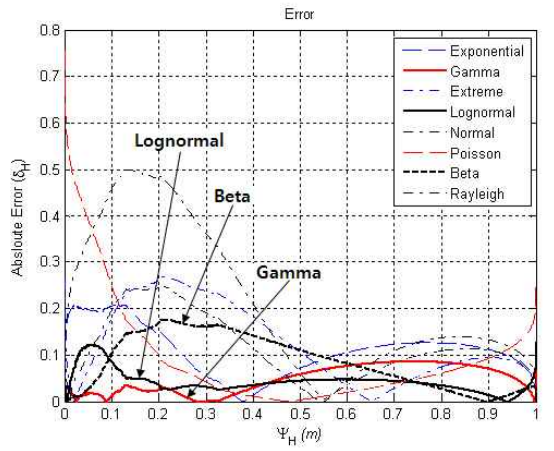
따라서 본 연구에 적용한 표본 데이터들은 베타와 감마 분포함수를 적용하는 것이 최적임을 알 수 있다. 특히, 프아송(Poisson)과 레일리히(Rayleigh) 분포함수는 가장 큰 오차를 나타내고 있다.



(a) In case of Roll sample data, ψ_R



(b) In case of Pitch sample data, ψ_P



(c) In case of Heave sample data, ψ_H

Fig. 5 Calculation results for the probability errors between E-CDF and the eight kinds of CDFs according to the sample data sets, ψ_R, ψ_P and ψ_H

Table 3 Averaged probability errors $\bar{\delta}_{k,j}$ according to the sample data sets, ψ_R, ψ_P and ψ_H

j \ k (Data / CDFs)		1	2	3
		Roll ψ_R (Ranking)	Pitch ψ_P (Ranking)	Heave ψ_H (Ranking)
1	Exponential	0.061 (4)	0.066 (3)	0.133 (6)
2	**Gamma**	0.031 (2)	0.033 (2)	**0.027 (1)**
3	Extreme	0.093 (6)	0.108 (6)	0.127 (5)
4	Lognormal	0.053 (3)	0.069 (4)	0.057 (2)
5	Normal	0.080 (5)	0.091 (5)	0.122 (4)
6	Poisson	0.185 (7)	0.151 (7)	0.314 (8)
7	**Beta**	**0.024 (1)**	**0.022 (1)**	0.060 (3)
8	Rayleigh	0.178 (8)	0.199 (8)	0.251 (7)

5 결 론

본 연구에서는 선행연구(Yim, 2012)에서 제안한 확률기반 위기평가기법(Probability base Risk Evaluation Techniques, PET)에서 누적분포함수를 위기허용기준으로 적용할 때 발생한 문제점의 해결 방안을 제안하였다.

해상 부유체 모델에서 획득한 표본 데이터에 지수(Exponential), 감마(Gamma), 극치(Extreme Values), 로그정규(Lognormal), 정규(Normal), 프아송(Poisson), 베타(Beta), 레일리히(Raylaigh) 등의 여덟 가지 누적분포함수를 적용하여 최대우도(maximum likelihood)를 갖는 형상 변수를 추정한 후, 경험적 누적분포함수와의 확률오차를 계산하여 비교 평가한 결과는 다음과 같다.

(1) 롤(Roll)과 피치(Pitch)의 표본 데이터에 대해서는 베타(Beta) 누적분포함수가 평균 확률오차가 가장 작은 0.024와 0.022를 각각 나타냈다. 이 때 형상 변수 값은 롤 데이터의 경우 a = 0.578, b =2.684 등으로 나타났고, 피치 표본 데이터의 경우는 a = 0.635, b =1.288 등으로 나타났다.

(2) 히브(Heave) 표본 데이터에 대해서는 감마(Gamma) 누적분포함수가 평균 확률오차가 가장 작은 0.027을 나타냈고, 형상 변수 값은 a =1.778과 b =0.303 등으로 나타났다.

(3) 따라서 본 연구에 적용한 부유체 모델의 롤과 피치 운동에 대한 위기허용기준을 설정하는 경우에는 베타 누적분포함수를 적용하고, 히브 운동에 대해서는 감마 누적분포함수를 적용하는 것이 최적임을 알았다.

특히, 본 연구에서 제안한 방법을 이용하면 선행연구에서 적용한 경험적 누적분포함수와 달리 최대우도를 갖는 형상 변수를 이용한 누적분포함수에서 연속된 누적확률 값을 구할 수 있기 때문에 해상 부유체를 보다 세밀하고 정밀도 높게 평가할 수 있다.

아울러 본 연구에서 제안한 방법은, 표본 데이터가 불완전하고 분포특성을 모르는 경우, 최대우도를 갖는 정형화된 최적의 누적분포함수 선정을 위한 다양한 연구 분야에도 적용 가능할 것으로 고려된다. 향후 본 연구결과를 실제 해상 부유체의 안전성 평가에 적용할 예정이다.

후 기

이 논문은 2013년도 해양수산부지정 호남 Sea Grant 센터 연구개발사업 과제 지원에 의해 수행된 연구임.

참 고 문 헌

[1] Breheny Patrick(2013), Introduction to the empirical distribution function(STA 621): Nonparametric Statistics, white paper, http://web.as.uky.edu/statistics/users/pbreheny/621/F10/notes/8-26.pdf.

[2] Charles R. Farrar, Scott W. Doebling and David A. Nix(2001), "Vibration-based structural damage identification," Philosophical Transactions of the Royal Society A, Londo, Vol. 359, pp. 131-149.

[3] David Vose(2010), Fitting Distributions to Data and why you are probably doing it wrong, white paper, http://www.vosesoftware.com/whitepapers/Fitting%20distributions%20to%20data.pdf.

[4] Joel Azose(2013), On the Gamma Function and Its Applications, white paper, http://www.math.washington.edu/~morrow/336_10/papers/joel.pdf.

[5] José Miguel Simón Donaire(2009), Sea Transport Analysis of Upright Wind Turbines, Master Thesis(MEK-FM-EP-2009-14), Technical University of Denmark.

[6] Jun Chang Hyun and Yoo Chul Sang(2012), "Application of the Beta Distribution for the Temporal Quantification of Storm Events," Journal of Korea Water Resources Association, Vol. 45, Issue 6, pp. 531-544.

[7] Kim Jin Ho, Kim Hyeong Seok and Cho Sung Ho(2013), "A Ranging Algorithm for IR-UWB in Multi-Path Environment Using Gamma Distribution," The Journal of Korea Information and Communications Society, Vol. 38B, No. 2, pp. 146-153.

[8] Lee J. T. and Oh H. J(1996), "Approximation Equation of Cumulative Distribution Function on the Normal Distribution," Journal of the Korea Society of Mathematical Education, Series A, Vol. 35, No. 1, pp. 57-59, http://www.mathnet.or.kr/mathnet/kms_tex/982256.pdf.

[9] Lu Kung-Chun, Loh Chin-Hsiung, Yang Y. S., Jerome P. Lynch and Kincho H. Law(2008), "Real-Time

Structural Damage Detection using Wireless Sensing and Monitoring System," Smart Structures and Systems, TechnoPress, Vol. 4(6), pp. 759-778.
[10] MATLAB(2008a), Programming, MATLAB Version 7.6 (R2008a)
[11] MATLAB(2008b), Statistical Toolbox : Maximum likelihood estimation, MATLAB Version 7.6 (R2008a).
[12] MATLAB(2008c), Statistical Toolbox : Empirical Cumulative Distribution Function, MATLAB Version 7.6 (R2008a).
[13] MathWorks(2013), Statistical Toolbox Distribution Functions, http://www.mathworks.co.kr/kr/help/ stats/ statistics-toolbox-distribution-functions.html.
[14] Plancade Sandra(2013), Adaptive estimation of the conditional cumulative distribution function from current status data, Institute of Community Medicine, white paper, University of Tromso, Norway, pp. 1-42, http://sandraplancade.perso.math.cnrs.fr/cens-int.pdf.
[15] R-forge project(2013), Handbook on probability distributions, white paper, R-forge distributions Core Team, University Year 2009-2010, pp. 1-167, https://r-forge.r-project.org/scm/viewvc.php/*checkout*/pkg/inst/doc/probdistr-main.pdf.
[16] Riddhi D.(2013), Beta Function and its Applications, white paper, Department of Physics and Astronomy, The University of Tennessee, USA, pp. 1-4, http://sces.phys.utk.edu/~moreo/mm08/Riddi.pdf.
[17] Staub Linda and Gekenidis Alexandros(2011), Seminar in Statistics: Survival Analysis Chapter 2, white paper, pp. 1-39, http://stat.ethz.ch/education/semesters/ ss2011/seminar/contents/presentation_2.pdf.
[18] Wikipedia(2013a), Tutorial for Probability distribution, http://en.wikipedia.org/wiki/Probability_distribution.
[19] Wikipedia(2013b), Tutorial for Arg max, http://en.wikipedia.org/wiki/Arg_max.
[20] Wikipedia(2013c), Tutorial for Maximum Likelihood, https://en.wikipedia.org/wiki/Maximum_likelihood.
[21] Yim Jeong Bin(2012), "Probability Based Risk Evaluation Techniques for the Small-Sized Sea Floater," Journal of Navigation and Port Research, Vol. 36, No. 10, pp. 795-801.

원고접수일 : 2013년 7월 10일
심사완료일 : 2013년 9월 11일
원고채택일 : 2013년 9월 25일