적정 확률분포형 선정기준의 적용성에 관한 연구

Application for the Selection Criteria of Apppropriate Probability Distribution

김수영*, 허준행**

Sooyoung Kim, Jun-Haeng Heo

요 지

일반적으로 확률수문량을 산정하기 위해서는 수문자료에 대해 빈도해석을 실시한 후 확률수문량을 산정하게 된다. 재현기간이 커질수록 확률분포형에 따라 확률수문량의 값은 많은 차이를 나타내므로 적정 확률분포형의 선정은 매우 중요하다고 할 수 있다. 적정 확률분포형의 선정은 객관적인 기준에 의해 이루어져야 하나, 적정 확률분포형의 선정에 있어 명확한 기준이 마련되어 있지 않아 실무에서 확률수문량을 산정할 때 많은 어려움을 겪고 있는 실정이다. 따라서 본 연구에서는 적정 확률분포형의 선정기준으로 제시되어 있는 검정통계량을 이용한 방법의 적용성을 비교·검토하고자 한다. 이를 위해 우리나라에서 널리 사용되고 있는 Gumbel, GEV 분포형과 Weibull, Generalized logistic 분포형을 선택하고 각각의 분포형에 대해 자료의 크기별 모의를 통해 자료를 발생시킨 후 빈도해석을 수행하고, 적합도 검정 단계에서 산출되는 검정통계량을 비교하여 적정 확률분포형을 선정하여 적용성을 검토하고자 한다. 결과적으로 자료 발생에 이용된 분포형과는 관계없이 자료수가 작을수록 2변수 gamma, 자료수가 많을수록 5변수 Wakeby가 제일 많이 선정되는 것으로 나타났으며, Gumbel, GEV, generalized logistic 분포형의 경우는 대체로 자료의 수가 많아질수록 선정되는 빈도가 많은 것으로 나타났다.

핵심용어 : 적정 확률분포형 선정, 적합도 검정, 검정통계량

1. 서론

일반적으로 확률수문량을 산정하기 위해서는 수문자료에 대해 무작위성 판단을 위한 예비적 해석, 다양한 확률분포형의 적용, 확률분포형의 매개변수 추정, 매개변수의 적합성 판단, 적합도 검정, 적정 확률분포형의 선정과 같은 일련의 과정을 포함하는 빈도해석을 실시한 후 적정 확률분포형을 선정하여 확률수문량을 산정하게 된다. 이때 확률분포형에 따라 재현기간에 대한 확률수문량의 값은 차이를 나타내는데, 특히 재현기간이 커질수록 큰 차이를 나타내고 수공구조물의 설계나 평가에 큰 영향을 끼치게 되므로 적정 확률분포형의 선정은 매우 중요하다고 할 수 있다. 적정 확률분포형의 선정은 이와 같은 중요성에 적합한 객관적인 기준에 의해 이루어져야 하나, 현재 우리나라에는 적정 확률분포형의 선정에 있어 명확한 기준이 마련되어 있지 않아 실무에서 확률수문량을 산정할 때 많은 어려움을 겪고 있는 실정이다. 따라서 본 연구에서는 적정 확률분포형의 선정기준으로 제시되어 있는 검정통계량을 이용한 방법의 적용성을 비교·검토하고자 한다.

* 정회원, 연세대학교 대학원 토목공학과 박사과정, E-mail : sykim79@yonsei.ac.kr

** 정회원, 연세대학교 사회환경시스템공학부 토목환경공학과 교수, E-mail : jhheo@yonsei.ac.kr

2. 검정통계량의 적용성 검토

2.1 적합도 검정방법

임의의 확률분포형에 대한 적합도 검정은 해당 확률분포형의 상대도수함수와 누가도수함수의 이론값과 표본값을 비교하여 그 정도를 판별하게 된다. 현재 우리나라에서 널리 사용되고 있는 적합도 검정방법은 χ^2-검정, Kolmogorov-Smirnov 검정, Cramer von Mises 검정, PPCC(Probability Plot Correlation Coefficient) 검정 등이 있으며, 본 연구에서는 적정확률분포형의 선정을 위해 사용되는 검정통계량을 기각력이 우수하다고 알려져 있는 χ^2-검정과 PPCC 검정의 검정통계량을 이용하였다.

2.1.1 χ^2-검정

χ^2-검정은 자료의 크기에 따라 m개의 계급구간으로 나누고 이론값과 자료값의 절대도수를 비교하는 방법으로 검정통계량 q는 다음과 같다.

$$q = \sum_{j=1}^{m} \frac{(n_j - e_j)^2}{e_j} \tag{1}$$

여기에서 n_j는 관측 자료의 j번째 구간의 표본 관측도수, $e_j = np_j$는 확률분포의 j번째 구간의 이론도수이며, m은 계급구간의 수를 나타낸다. 또한 p_j는 구간 내 특정 기각치를 만족하는 모의변수확률로, 유의수준 α에 대해 귀무가설이 $q \geq K$로 기각된다고 하면 $p(q \geq K; q \sim \chi^2(k-1)) = \alpha$로 정의되며 $K = \chi^2(k-1)$으로 계급구간을 나눈 후 결정된다. 일반적으로 계급구간은 등간격으로 Sturges 공식(Sturges, 1926)이 널리 사용된다.

계산된 통계량은 $\chi^2 < \chi^2_{1-\alpha,v}$을 만족하면 적합성이 인정되고, 만족하지 못하면 기각되게 된다. 여기에서 $\chi^2_{1-\alpha,v}$는 자유도 ν일 때 유의수준 α로 가정된 분포의 한계치이다.

2.1.2 PPCC 검정

PPCC 검정은 자료의 적모멘트 상관계수를 이용하여 적합도 검정을 수행하며 자료의 적모멘트 상관계수는 다음과 같다.

$$r_c = \frac{\sum_{t=1}^{N}(X_i - \overline{X})(M_i - \overline{M})}{\sqrt{\sum_{t=1}^{N}(X_i - \overline{X}) \sum_{t=1}^{N}(M_i - \overline{M})^2}} \tag{2}$$

여기에서 $M_i = \Phi^{-1}(m_i)$이고 $\Phi^{-1}(\bullet)$는 각 확률분포형의 누가분포함수의 역함수이고, m_i는 누가분포함수의 중간값이며 Filliben(1975)은 다음과 같은 식을 제안하였다.

$$m_i = 1 - (0.5)^{1/N} \quad i = 1 \tag{3}$$

$$m_i = \frac{(i - 0.3175)}{(N + 0.365)} \quad i = 2, \cdots, N-1 \tag{4}$$

$$m_i = (0.5)^{1/N} \quad i = N \tag{5}$$

표본자료가 가정한 확률분포형이라는 가설은 $r_c > r_a(N)$을 만족하면 적합하다고 판단하게 된다.

2.2 검정통계량을 이용한 방법

χ^2-검정과 PPCC 검정과 같은 적합도 검정과정에서 계산되는 검정통계량은 실제 자료와 이론값을 비교하

여 계산하는 것으로 검정통계량이 작으면 작을수록 실제값에 가깝다는 점을 고려하여 적정 확률분포형의 선정에 이용하였다. 이와 같이 검정통계량을 이용하여 적정 확률분포형을 선정하는 과정을 나타내는 개략적인 모식도는 그림 1과 같다. 이를 살펴보면 PPCC 검정이 존재하지 않는 분포형 중 χ^2-검정을 통과한 분포형이 있다면 χ^2-검정 과정에서 계산된 검정통계량 중 가장 작은 검정통계량을 가지는 분포형을 선정하고, PPCC 검정이 존재하지 않는 분포형 중 χ^2-검정을 통과한 분포형이 없다면 χ^2-검정 과정에서 계산된 검정통계량과 PPCC 검정 과정에서 계산된 검정통계량을 이용한다.

PPCC검정을 이용할 경우, χ^2-검정과 PPCC 검정을 동시에 통과한 분포형이 하나라면 그 분포형을 적정 확률분포형으로 채택하고 χ^2-검정과 PPCC 검정을 동시에 통과한 분포형이 0개일 경우는 PPCC 검정통계량의 최소값을 계산하여 최소값을 갖는 분포형을 적정 확률분포형으로 채택하며, χ^2-검정과 PPCC 검정을 동시에 통과한 분포형이 2개 이상일 경우는 PPCC 검정통계량의 최소값과 χ^2-검정통계량의 최소값을 나타내는 분포형이 하나인 경우는 그 분포형을 적정 확률분포형으로 채택하고, 그렇지 않은 경우는 PPCC 검정통계량의 최소값을 나타내는 분포형을 적정 확률분포형으로 채택한다.

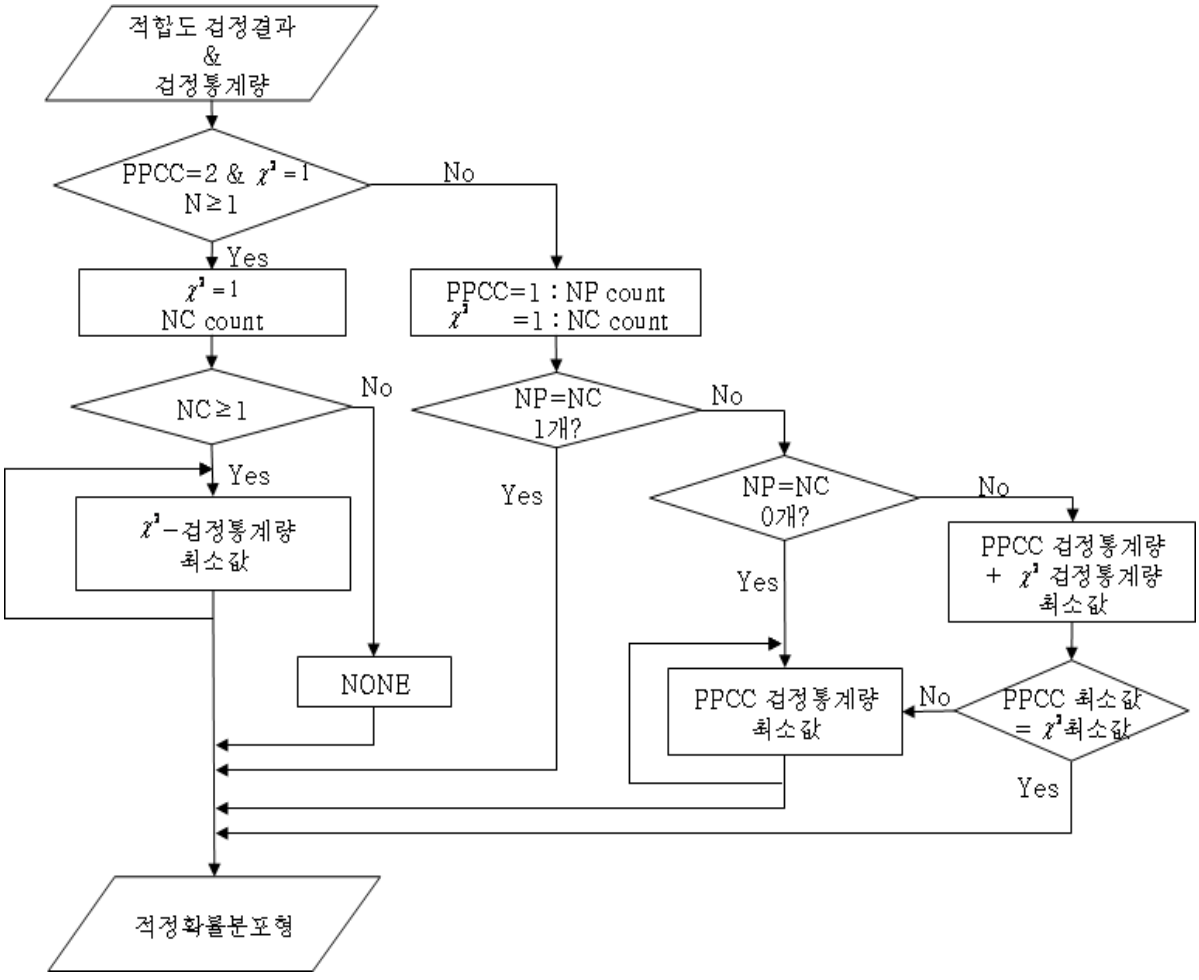

그림 1. 적정 확률분포형 선정을 위한 모식도

2.3 적용성 검토

검정통계량을 이용하는 방법의 적용성을 검토하기 위해 Gumbel, GEV 분포형과 함께 Weibull, generalzied

logistic 분포형을 추가하여 분석하였다. 먼저 자료의 발생은 서울지방의 1954년~2003년의 자료를 이용한 빈도해석 과정을 통해 추정된 각 분포형의 매개변수 중 적합성 검토 결과를 바탕으로 확률가중모멘트법을 이용하여 추정된 지속시간 12시간에 대한 매개변수를 이용하여 각 분포형별로 자료개수를 10, 25, 50, 100, 200개로 구분하여 1000번의 모의를 통해 자료를 구축하였다. 이때 자료의 발생을 위해 사용된 각 분포형별 매개변수는 표 1과 같다.

발생된 자료에 확률분포형을 적용하여 확률가중모멘트법을 이용하여 매개변수를 추정하고 적합도 검정을 실시하고, 적합도 검정결과를 바탕으로 검정통계량을 이용하여 적정 확률분포형을 선정하고 자료의 발생에 이용된 분포형과 비교하였으며 결과는 표 2와 같다. 표 2에 나타난 결과를 살펴보면 자료의 발생에 사용된 확률분포형과는 상관없이 자료의 수가 적으면 대체로 2변수 gamma, 자료의 수가 많아지면 5변수 Wakeby가 제일 많이 선정되는 것으로 나타났고, Gumbel, GEV, generalized logistic 분포형의 경우는 자료의 수가 많아질수록 대체로 선정되는 빈도가 늘어나는 것을 알 수 있다.

표 1. 자료의 발생에 사용된 매개변수

분포형	위치 매개변수	규모 매개변수	형상 매개변수
Gumbel	1095.694	443.599	-
GEV	1085.727	428.835	-0.042
Weibull	519.215	922.485	1.504
Generalzied logistic	1255.997	289.020	-0.193

표 2. 모의발생 결과 선정된 확률분포형

분포형	자료수	1st	2nd	3rd	4th	5th
GUM	10	GAM2	GUM	LGU2	LN2	WBU2
	25	GAM2	GAM3	GAM2	GUM	LN3
	50	GAM2	GAM3	GAM2	GUM	LN3
	100	WKB5	GAM3	GAM2	LN3	LN3
	200	WKB5	WKB5	GAM2	LN3	LN3
GEV	10	GAM2	GUM	LGU2	LN2	WBU2
	25	GAM2	GAM3	GEV	GUM	LN2
	50	GAM2	GAM3	GEV	LN3	LN3
	100	WKB4	GAM2	LN3	GEV	LN3
	200	WKB5	GEV	LN3	LN3	LN2
WBU	10	GAM2	GUM	LGU2	LN2	WBU2
	25	GAM2	GAM3	GAM2	GUM	LN2
	50	GAM3	GAM3	GAM2	LN3	LN3
	100	WKB5	GAM3	GAM3	LN3	LN3
	200	WKB5	GAM3	GAM3	LN3	LN3
GL	10	GAM2	GUM	LGU2	LN2	WBU2
	25	GAM2	GAM3	GEV	LN3	LN3
	50	GAM2	GAM3	GAM2	LN2	LN2
	100	WKB5	GL	GL	GEV	LN3
	200	WKB5	GL	GEV	GEV	LN3

3. 결론

본 연구에서는 적정 확률분포형의 선정을 위해 빈도해석 과정 중 적합도 검정 단계에서 계산되는 검정통계량을 이용하여 적정 확률분포형을 선정하여 적용성을 검토하였다. 결과적으로 자료의 발생에 이용된 분포

형과는 관계없이 자료의 수가 작을수록 2변수 gamma 분포형이 가장 많이 선정되고, 자료의 수가 많을수록 5변수 Wakeby 분포형이 가장 많이 선정되는 것으로 나타났다. 또한 적정 확률분포형으로 선정된 분포형이 자료발생에 이용된 확률분포형의 특징을 반영하기 위해서는 자료의 수가 되도록 많아야 하는 것으로 나타났으나 해당 분포형의 선정이 뚜렷하게 나타나는 현상이 아니므로 차후 본 연구의 결과를 기반으로 지속적인 연구를 통해 적정 확률분포형의 선정기준을 명확히 정립해야 하겠다.

감사의 글

이 연구는 건설교통부가 출연하고 한국건설교통기술평가원에서 위탁시행한 2003년도 건설핵심기술연구개발사업(03산학연C03-01)에 의한 도시홍수재해관리기술연구사업단의 연구성과입니다.

참고문헌

1. Fillibenm J. J.(1975). The probability Plot Correlation Coefficient Test for Normality, Technometrics, Vol. 17, No. 1, pp.111 ~ 117.
2. Sturges, H, A.(1926). The Choice of a Class Interval, Journal of American Statistical Association, Vol. 21, No. 21, pp.65 ~ 66.