<講 座>

도시유역에서의 빈도해석

Urban Hydrologic Frequency Analysis

편 집 부

Introduction (서론)

도시 수문학에 대한 어떠한 토론에서도 우리는 항상 100년 홍수 혹은 50년 강우와 같은 항목들을 듣게 된다.

흔히 이러한 항목들은 다소 부정확하게 사용되며 보통사람들에게 이해되는 경우는 아주드물다 이러한 항목들을 사용하고 있는 사람도 많은 경우에 이러한 항목들의 의미와 내포된 듯, 항목들에 관계된 events 크기 계산의 어려움 event 크기 계산에 있어서의 불확실성 혹은 다양성 등에 대해서 완전히 이해하지 못하고 있다.

우리가 관심을 가지고 있는events를 나타내기 위하여 일반화된 기호를 사용한다 주기가 T해인 event를 T-year 사상으로 사용한다 (주기는 아직 정의되지 않았음) Q_T는 T-year 홍수의 크기를 나타낸다. 다음에 알 수 있겠지만 Q_T는 결코 확실하게 알 수 없다. 따라서, 우리는 항상 Q_T의 계산만을 취급해야 한다.

Return Period and Probability

한해 동안 어떠한 흐름에서 관측된 최대 유출량(첨두유량)은 해마다. 무작위 경향으로 변한다는 것은 잘 알려져 있다.

이러한 무작위성은 폭우를 대비한 시설의 수리학적 용량을 결정하는데 확률이나 통계를 사용하게끔 한다.

T-year event란 (T-years보다 훨씬 긴) 긴 시간에 걸쳐 크기를 가진 event으로 정의된다. 즉, T-year events 이상의 크기를 가진 event 사이의 평균시간이 T-year이다. 이러한 경우에 N-year 동안에 T-year event의 발생 예상횟수는 N/T가 될 것이다.

예를들어 100년 동안 20-year event의 5회 출현을 예상할 수 있다. 다른 말로 바꾸면 평균적으로 T-year event는 매 T-year마다 한 번씩 예상할 수 있다. 하나의 T-year event에 관계된 어떤 규칙성이란 존재하지 않는다는 것을 강조해야 할 필요가 있다.

이것은 T-year event가 매 T-years마다 한 번씩 발생하는 것도 아니고 어떠한 T-year 기간동안 항상 하나의 T-year event가 발생하는 것도 아니다. 실제로, 다음에 어떠한 T-year 기간동안, T-year event가 0, 1, 2, …T번 발생할 수 있는 기회가 있다는 것을 보여줄 것이다. 더 나아가서 이러한 다양한 가능성의 확률을 계산하는 것도 보여줄 것이다.

위에서 정의한 T-year event의 재현 기간은 T-year이다. 가끔 T-year event가 발생하는 사이의 실제 시간이 recurrence interval로 불린다. 따라서, 재현기간(recurrence interval)의 평균치는 재현 기간과 일치한다. 재현 기간과 주기에 대한 대부분의 논의에서 두항목들은 동의어로 가정한다. 따라서 재현기간이란 항목을 사용하는 대부분의 경우에는 평균 재현 기간을 의미한다.

T-year event의 발생 사이의 평균시간이 T-years이기 때문에 어느 주어진 해에 있어서의 T-year event의 확률은 1/T이다.

따라서

$$P_T = 1/T \qquad (1)$$

여기서 T : event Q_T에 관계된 재현기간

P_T : 어느 주어진 해에 있어서의 Q_T의 확률

이 전개에 있어서 우리는 강조해야 할 몇가지 가정을 해왔었다. 이 가정에는 변수 Q를 포함하는데, peak flow (첨두유량)은 임의의 해(특정한 해가 아님)에 일어난다.

첫째 각 해의 첨두유량(peak flows)은 서로 독립적이다. 이것은 어느 해의 첨두유량의 크기는 다른 해의 첨두유량의 크기에 영향을 받지 않는다는 것을 의미한다. 둘째로 첨두유량의 통계학적 성질은 시간에 따라 변하지 않는다고 가정한다.

이것은 유역의 첨두유량 특성을 변화시키는 유역내의 변화가 없다는 것을 의미한다. 더 깊이 얘기하자면 우리가 사용하고 있는 data가 어떤 기간중에 얻어진 것이라도 유역특성은 일정하다는 것이다.

통계학적 용어를 빌면, data가 정체적 시계열(stationary time series)로부터 얻어진다고 가정하는 것이다. 이러한 가정하에서 T-year event 출현은 베르누이 시행(Bernoulli process)이라 알려진 특별한 stochastic process를 필요로 하는 무작위과정이 된다. Q_T의 확률은 어떠한 해와 같거나 크거나 하기 때문에 P는 항상 Q_T 출현의 과거 기록에 의해 영향받지 않는다. 이제 Q_T와 같거나 큰 event를 Q_T*라 하자. 우리는 Q_T*의 실제 크기를 모르며 단지 그 값이 Q_T와 같거나 크다는 것만을 알 뿐이다. (Q_T*≥Q_T) Q_T*는 베르누이 무작위변수이다. n해수동안 Q_T*가 k번 출현할 확률은 이항분포(binomial distribution)로부터 계산해 낼 수 있다.

$$f(k;P_T,n)=\frac{n!}{(n-k)!k!}\ (P_T)^k(1-P_T)^{n-k} \quad (2)$$

여기서 $f(k;P_T,n)$은 어느 한 해에 있어서 Q_T*의 확률이 P_T일때 n해 동안 Q_T*가 k번 출현할 확률이다.

예를들어, 30년 동안 20-year event가 2회 출현할 확률은

$$f(2\ ;\ 0.05\ ;\ 30)=\frac{30!}{28!2!}\ (0.05)^2\ (0.95)^{28}$$

$$=0.26$$

아주 많은 양의 30년 기록에서, Q_T와 같거나 초과하는 첨두량이 정확하게 2번 나타날 확률은 26%라고 예상할 수 있는 것이다. 30년 자료중 나머지 74%는 Q_T와 같거나 큰 첨두량이 0, 1, 3, 4…, 30회 나타날 확률이다. 이러한 초과횟수들의 확률 역시 식(2)로부터 계산할 수 있다. 만일 그렇게 해 본다면 Q_T와 같거나 초과하는 첨두량이 30년동안 0, 1, 2, 3, …, 30 횟수 출현할확률의 총합은 1.00이 되어야 한다. 왜냐하면 모든 확률이 다 계산되기 때문이다.

식(2)는 n년동안 T-year event가 적어도 한번 이상 출현할 확률을 계산하는데 사용될 수 있다. 여기서 "적어도 한번"이란 한번 혹은 그 이상을 나타낸다.

한번 이상 초과될 확률과 한번도 초과 안되는 확률의 합은 1.00이다. 따라서, 적어도 한번 초과될 확률은

$$1-f(O;P_T,n)=1-\frac{n!}{o!n!}\ P_T{}^o(1-P_T)^n$$

$P_T=1/T$. $o!=1$이므로 이 식은

$$f(P_T,n)=1-(1-1/T)^n \quad (3)$$

과 같이 된다.

여기서, $f(P_T,n)$은 n년동안 T-year event가 적어도 한번 이상 출현할 확률이다.

만약식(3)에서 n과 T가 같다면, T값이 커지면 $f(P_T,T)$값이 상수 0.632에 접근한다는 것을 알 수 있다.

$T=10$이면 $f(P_T,T)=f(0,1,10)=0.65$이다.

이것은 설계수명이 T-years를 가진 구조물이 T-year event을 근거로 설계되었다면, 설계수명동안 설계용량이 적어도 한 번이상 초과될 확률이 약 0.63이다.

설계용량의 수용확률이 구조물의 설계수명동안 초과될 수 있도록 명세함으로써, 식(3)은 필요한 주기를 계산하는데 사용될 수 있다. 예를들어, 25년동안 구조물의 설계용량이 90%의 확률로 초과하지 않기를 원한다면 $f(P_T,25)$는1−0.90=0.10이 된다.

따라서 식 (3)으로부타

$$0.10=1-(1-1/T)^{25}$$

혹은 T=238년이다.

25년동안 설계용량이 90%의 확률로 초과하지 않기 위해서는 설계용량이 238년의 재현기간을 가진 event에 근거를 두어야 한다. 이러한 경우에 risk는 10%이고 신뢰도는 90%이고, 설계수명은 25년이고 필요한 설계주기는 238년이다. 이와같은 계산은 여러가지 설계수명, 설계주기, acceptiable risks등에 적용될 수 있다.

〈Figure 1〉은 이러한 계산법에 근거를 두었고 초과될 설계수명, acceptible risk, 설계용량의 확률에 근거를 둔 필요한 설계주기를 빨리 결정하는 데 사용될 수

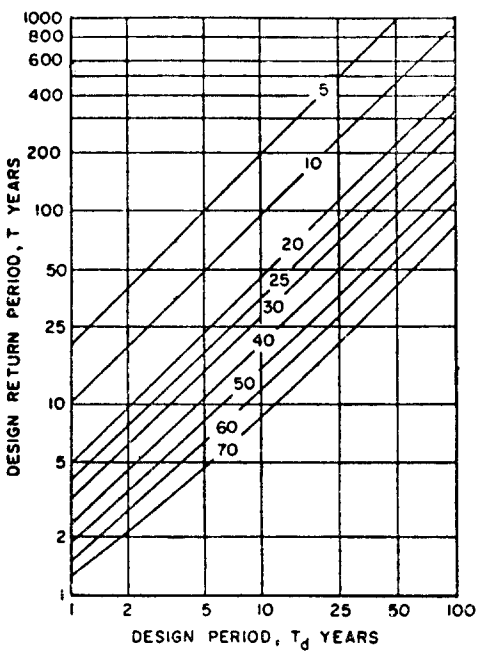

Fig. 1. Design return period required as a function of design life and acceptable rise Icurve parameterin percentl that the design condition is exceeded.

있다. 이러한 논의에서는 설계용량의 초과가 가지는 높은 위험도를 명심해야만 한다. 왜냐하면 초과에 의한 의미란 구조물을 설계대로 작동하여 resulting flow를 조절하는데 실패한다는 것을 뜻하기 때문이다. 이러한 의미에서의 실패란 반드시 구조물이 파괴된다는 것을 의미하는 것은 아니다.

예를들면, 도로 배수구가 첨두유량을 흘려보내는데 있어서의 실패란 도로나 잔디에 최소한의 홍수를 야기시킬 뿐이지 다소 혼한 기준에서는 수용할 수 있는 것이다. 반면에 폭우저지 못(storm water detention basin)의 실패는 재산에 상당한 피해를 주고 하류에서의 인명손실의 위험도가 높은 구조물 월류를 야기시키는 것이다.

따라서 허용 위험도와 설계주기의 선택은 초과될 수 있는 설계용량의 결과에 따른다. 아주 드물게 일어나는 events에 대해 보호하기 위한 지나친 크기의 구조물을 세우는 것은 상당히 비싸게 들며, 반면에 하나의 빈도에 근거를 두고 설계용량이 초과되도록 하면 상당한 경제적 손실의 축적을 야기시킨다. 따라서 적당한 설계주기의 선택은 경제적 최적화에 있어서의 하나의 문제가 된다.

James와 Lee (1971), Hjelmelt와 Cassidy (1975)가 이 주제에 대해 더욱 상세하게 설명할 수 있을 것이다.

많은 정부기관(governmental units)에서 사용되어지는 설계주기를 결정하는 규칙들을 가지고 있다. 흔히 이러한 주기들은 구조물의 크기와 초과되어질 구조적 수리용량의 결과에 근거를 둔다 예를들어, 농촌지역에서의 도로 배수구는 10년 주기에 근거를 둘 것이다. 도시지역의 작은 구조물들은 25년 event에 근거를 두고 홍수범람지 설계는 100년 event에 근거를 둘 것이다.

빈도해석(frequency Analysis)

주어진 재현기간에 대해 홍수량을 정하기 위해서는 대상유역의 홍수량 특성에 대한 지식이 필요하다.

이 관계를 결정하는데 사용되는 접근방법은, 이용가능하고 또 그 결정에 중요한 수문학적 자료(hydrologic data)의 형태(type), 양(quantity) 그리고 질(quality)에 의해 크게 좌우된다. 작은 culvert나 gutter를 설계할 때는 시간을 소비하거나 비싼 홍수빈도 해석(flood-frequency analysis)을 할 필요가 없다. 반면에 배수system의 주요부분을 건설할때는 가장 가능한 유량을 추정하는 것이 필요하다.

이러한 것을 할때, 우리는 설계자가 직면할 수 있는 5가지의 경우를 생각할 수 있다.

Case Ⅰ—대상하천의 대상지점 또는 그 부근에서 적당히 오랜 기간의 stream flow의 기록이 유용하다.

Case Ⅱ—대상하천의 대상지점에서 조금 떨어진 지점에서 적당히 오랜 기간의 stream flow의 기록이 유용하다.

Case Ⅲ—대상지점에서 짧은 기간동안의 유량자료가 강우자료와 함께 있다.

Case Ⅳ—대상하천에 대해서는 유용한 자료가 없고 근처하천에 대한 유용한 자료가 있다.

Case Ⅴ—부근에 아무런 자료가 없다.

각 경우는 생각할 수 있는 순서대로 열거되어 있다. 또한 각 경우는 점점 어려운 순서대로 되어 있다. 그러나 불행히도 발생될 확률 또한 점점 큰 순서대로 되어 있다. 즉 설계자는 Case Ⅰ보다 Case Ⅴ를 접하기가 더 쉽다는 것이다. 그럼에도 불구하고 우리는 Case Ⅰ을 취급하는데 주로 관심을 기울일 것이다. 왜냐하면, 다른 경우의 문제를 이해하기전에 먼저 Case Ⅰ의 과정과 그 제한 조건을 생각하는 것이 기본이 되기 때문이다.

Case Ⅰ—홍수 빈도 결정(Flood frequency determination)

만약 굉장히 운이 좋다면 주어진 빈도에 대한 홍수량의 추정이 요구되는 지점에서 첨두유량의 기록을 얻을 수 있을 것이다. 이러한 기록은 table Ⅱ과 같은 형태로 표시될 것이다.

Table Ⅰ에 들어있는 것과 같은 어떤 data의 집합도 어떠한 모집단에서 나온 data의 표본(sample)을 나타낸다. 이 경우에 있어서 모집단은 과거와 미래에 걸쳐 모든 시간에 대한 년최대 홍수량(max annual flood peak)일 것이다. Table Ⅰ의 data는 이 모집단에서부터 추출한 표본이다. 모집단의 성질을 나타내는 양들로는 parameters가 있다. 모집단 parameters는 홍수빈도연구(flood frequency study)에서는 절대 알 수 없고, data의 표본(sample)으로부터 추정해야 한다. 이렇게 추정하는 것은 표본 통계학(sample statistics)이라 알려져 있다. 이러한 parameters 중에서 대표적인 것으로는 평균 μ_x ; 표준편차 σ_x ; coefficient of variation C_v ; skewness Γ, 등이 있다. 「$\mu_x, \sigma_x, C_v, \Gamma$ 등에 대해 추정된 표본은 각각 X, S_x, C_v, C_s 등에 의해 주어지고」 다음 식들에서 계산된다.

$$\overline{X} = \sum x_i / n \qquad (4)$$

Table 1. Peak Discharge (cfs) Middle Fork Beargrass Creek, Cannons Lane, Louisville, Kentucky

Year	Peak Flow	Year	Peak Flow
1945	1810	1961	2400
1946	791	1962	976
1947	839	1963	918
1948	1750	1964	3920
1946	898	1965	1150
1950	2120	1966	874
1951	1220	1967	712
1952	1290	1968	1450
1953	768	1969	707
1954	1570	1970	5200
1955	1240	1971	2150
1956	1060	1972	1170
1957	1490	1973	2080
1958	884	1974	1250
1959	1320	1975	2270
1960	3300		

$$S_x=\sqrt{(\sum x_i^2-n\bar{x}^2)/(n-1)} \tag{5}$$

$$C_v=S_x\sqrt{\bar{x}} \tag{6}$$

$$C_s=n\sum(x_i-\bar{x})^2/(n-1)(n-2)\ S_x^3 \tag{7}$$
$$=\frac{n^2\sum x^3-3n\sum x\sum x^2+2(\sum x)^3}{n(n-1)(n-2)S_x^3}$$

여기서 x_i는 i번째 data 값을 나타내고 n은 표본의 크기(sample size)를 나타내며 모든 $\sum$는 1에서부터 n까지 시행한다. 이 식들을 Bea grass Greak data에 적용한 결과 $\bar{x}=1599$cfs, $S_x=1006$cfs, $C_v=0.629$ 그리고 $C_s=2.13$ 등으로 나타났다.

평균은 단지 data群의 중앙위치를 측정하는 것이고 표준편차는 data의 산포도를 측정하는 것이다. 표준편차가 크면 클수록 산포도는 크게 된다. 표준편차의 제곱은 분산이라고 한다. 표준편차의 단위는 실제 data에 있어서의 단위와 같다. data群의 산포도를 무차원으로 측정하는 것으로는 분산계수(coefficient of variation)가 있다. 밀집된 data群은 넓게 퍼져있는 data群보다 작은 분산계수(coefficient of variation)를 가질 것이다.

skeweness는 data의 대칭성을 측정하는 것이다. 평균에 대해 대칭적으로 분포되어 있는 data群은 skewness가 0이다. 만일 data가 평균을 기준으로 왼쪽보다 오른쪽으로 많이 치우쳐 있으면 양(+)으로 비대칭되어 있다고 하고 C_s는 양의 값을 갖는다. 오른쪽보다 왼쪽으로 치우쳐 있는 data는 음(−)으로 비대칭되어 있다고 하고 C_s는 음의 값을 갖는다.

만일 table 1에 있는 data에 어떤 가정을 준다면 그러한 data는 independent random variable로 볼 수 있어 빈도해석(frequency analysis)을 할 수 있다. 주된 가정은 data는 서로 독립적(independent)이고 stationary time senes에서 뽑은 것이라는 것이다.

빈번한 홍수량(frequent floods)은 재현기간(return period)의 개념에 기초를 두고 직관적으로 추정할 수 있다. 예를들어 5년주기의 홍수는 평균적으로 5년이다 이와같거나 큰 홍수가 일어난다는 것이다. 즉 시간적으로는 20%의 확률로 나타내게 된다. Table 1을 보면 약 20%에 해당하는 6개의 첨두유량(peak)의 2120 cfs를 초과하고 있음을 알 수 있다. 따라서 5년 홍수(5-year flood)의 크기는 2120 cfs로 추정할 수 있다. 마찬가지로 약 10%가 2400 cfs를 초과하므로 2400 cfs를 10년사상 (10 year event)으로 추정할 수 있겠다.

이러한 직관적인 접근 방법의 어려움은 유용한 기록의 기간보다도 재현 기간이 긴 사상(events)의 크기를 평가할 수 없다는 것이다. 또, 재현 기간이 기록이 있는 기간과 거의 같은 사상의 크기도 단지 몇몇 관측에만의 존하므로 그다지 확실하지 않다. 예를 들면, 앞의 예에서 10년 사상(10-year event)은 단지 3가지 관측에만 의존하고 있다. 첨두유량(peak flow)의 개연성을 설명하는 데 모든 자료를 이용할 수 있는 일련의 과정이 필요하다.

우선 frequency histogram 형태도 Data를 plotting 함으로써 시작한다. 이는 단지 어떤 계급구간(class interval)에 대한 peak의 발생빈도를 계급구간(class interval)에 대해 plot한 것이다. 또, 어떤 값 이하의 값을 갖는 data의 백분율과 그 값의 크기와의 관계를 plot 할 수 있다. 〈Figure 3〉은 Beargrass Creck Data로 plot한 것이다. 〈Fig 3〉으로부터 5년 홍수($P=\frac{1}{5}$

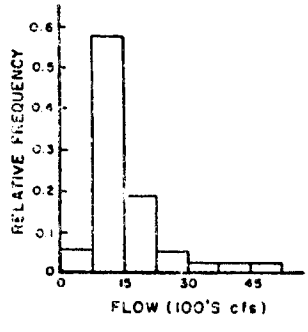

Fig 2. Frequency histogram-Beargrass Creek data.

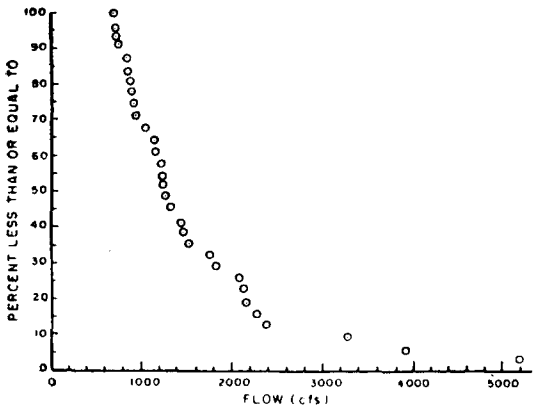

Fig 3. Empirical flood-frequency curve-Beargrass Creek data.

=0.20 or 20%)는 약 2150 cfs, 10년 홍수(10%)는 3250 cfs로 값을 정할 수 있다.

이용 가능한 자료가 많을 때에는, 위와 같은 방법은 재현기간이 짧은 홍수를 추정하기에 타당한 방법이다. 〈Fig 3〉을 살펴보면, data가 들쑥날쑥하므로 data가 나타내는 각점을 매끈하게 연결하여 곡선을 그리고, 그 곡선을 이용하여 각 재현기간에 따른 홍수량을 정의함으로써 재현기간이 짧은 홍수를 더 잘 추정할 수 있다.

불행히도 〈Fig 3〉과 같은 plot는 재현 기간이 긴 홍수량을 추정하기에는 충분하지 않다. 예를 들면 25년 홍수량은 곡선위에서 4%를 나타내는 점을 읽음으로써 정할 수 있으나 이는 기록상 두 번 일어났던 커다란 홍수량에만 근거한 자료이므로, 신뢰할 만한 값이 아니다. 만약 과거에 두 번 일어났던 이 큰 홍수량들이 7000 cfs와 4200 cfs였다면, 우리가 정한 25년 홍수량의 추정치는 크게 달라졌을 것이다.

또, 100년 홍수량을 이 data에 근거하여 구하려면 매끈한 곡선을 1%점까지 늘려 그려야 하는데 이는 전적으로 그리는 사람에 의하여 순간적으로 결정되므로, 사람마다 각각 다른 100년 홍수량을 추정할 수 있고 이들간에는 수천 cfs의 차가 나타날 것이다.

필요한 것은 plot된 점사이로 곡선을 그리는 해석적인 방법이다. 이 해석적인 곡선(analytic curve)은 각각의 재현기간을 갖는 홍수량들을 추정하는데 사용할 수 있다. 홍수 빈도 해석을 위한 해석적 기교(analytic techvique)를 논하기에 앞서 random data를 plot하는 데는 많은 주의를 요한다.

〈Fig 3〉의 결과 707 cfs가 100%점에 plot되었고, 이 하천의 연간 홍수 peak의 100%가 707 cfs 이상이라고 말할 수도 있겠지만, 이것이 31년간 자료의 분석에서 사실이라 하더라도 이것이 항상 그러한지는 알 수 없고, 미래 어느해에 707 cfs 이하의 flood peak가 일어날지도 모른다. 따라서 어떤 사상(event)에 100%의 가능성 또는 1의 확률을 부여할 수는 없다.

확률에 대해서 첨두 홍수량(flood peak)를 plot하는데 있어서 두 번째로 고려해야 할 것은, 〈Fig 3〉과 같이 일반 방안지(arithmetic graph paper)를 사용하면 점들은 일반적으로 홍수량이 클수록 간격이 넓어지면

Table 2. Plotting Position-Middle Fork Beargrass Creek at Cannon Lane in Louisville, Ky

Year	Discharge	Rank	Plotting Position
1945	1810	9	0.281
1946	791	28	0.875
1947	839	27	0.844
1948	1750	10	0.313
1949	898	24	0.750
1950	2120	7	0.219
1951	1220	18	0.563
1952	1290	15	0.469
1953	768	29	0.906
1954	1570	11	0.344
1955	1240	17	0.531
1956	1060	21	0.656
1957	1490	12	0.375
1958	884	25	0.781
1959	1320	14	0.438
1960	3300	3	0.094
1961	2400	4	0.125
1962	976	22	0.688
1963	918	23	0.719
1964	3920	2	0.063
1965	1150	20	0.625
1966	874	26	0.813
1967	712	30	0.938
1968	1450	13	0.406
1969	707	31	0.969
1970	5200	1	0.031
1971	2150	6	0.188
1972	1170	19	0.594
1973	2080	8	0.250
1974	1250	16	0.500
1975	2270	5	0.156

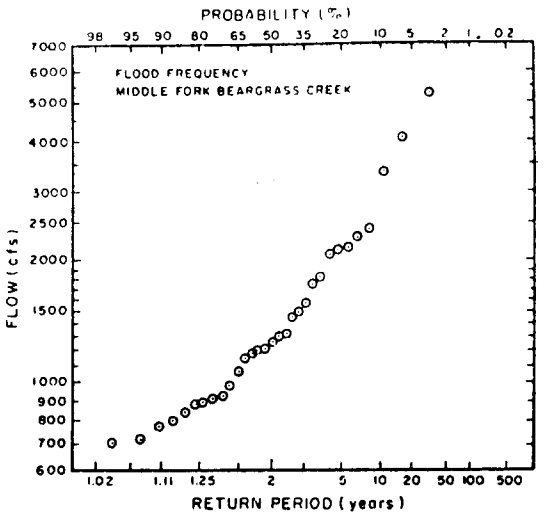

Fig 4. Probability plot-Beargrass Creek data.

서 몹시 구불 구불한 형태를 이룬다. 이러한 불편함을 극복하기 위해서 확률지(probability paper)라는 특수한 용지가 개발되었다. 확률지에는 여러 종류가 있지만 많이 이용되는 것으로는 정규 확률지(normal probability pa per)와 대수정규확률지(lognormal probability paper)가 있다. 여기서는 대수 정규확률지를 사용하기로 하겠다.

확률지에 첨두유량자료를 plotting하는 과정은 다음과 같다.

1. 큰 값으로부터 작은 값의 순서로 자료를 분류한다.

2. 다음 식으로 부터 plotting position을 계산한다.

$$p=\frac{m}{n+1} \tag{8}$$

여기서 p는 plotting position, m은 관측치의 계급 n는 자료의 연수이다.

3. 확률지에 p와 관측지의 크기를 plot한다.

확율 plotting의 예로서 Beorgrass Creek 자료를 생각해 보자. 이 자료는 Table 2에 계급과 position이 결정되어 있다. 〈Fig 4〉는 대수정규확율지에 자료가 plot되어 있다. 자료가 큰 값으로부터 작은 값의 순서로 분류되어 있으므로 plotting position p는 그 자료보다 크거나 같은 분수를 나타낸다. data는 아직 직선으로 plot되어 있지 않으나 〈Fig 3〉에 나타난 것보다는 곡률이 크게 감소되었다.

여기서 data를 smooth curve로 스치켓하거나 점들은 직선으로 fitting하는데 해석적 빈도해석 방법을 사용할 수 있다. 후자의 방법으로 $y=a+bx$ 같은 직선으로 정규 방안지에 plot 점들로 fitting하기 위하여 미지의 parameter를 갖는 방정식이 사용되었다. 우리가 직면하는 어려움은 사용할 방정식을 선택하는 것과 방정식의 parameter를 추정하는 데 있다.

random event가 발생할 확율을 나타내는 방정식은 확율밀도함수(pdf)와 누가확율분포함수(cdf)로 알려져 있다. pdf는 특정간격 내의 random event의 확율을 계산하는데 사용할 수 있다 cdf는 기지치보다 같거나 작을 확율을 계산하는 데 사용할 수 있다.

$X=x$에서 계산되는 random variable X의 pdf와 cdf를 표현하는데 $P_x(x)$와 $P_x(x)$의 기호를 사용한다. 관계는 다음과 같다.

$$P_x(x)=\int_{-\infty}^{x} P_x(t)\,dt \tag{9}$$

여기서 X는 random variable이며 t는 적분 변수이다. pdf에 대하여 사용할 수 있는 함수는 많이 있다. 함수가 pdf가 되기 위한 유일한 조건은 다음과 같다.

1. $P_x(x)\geq 0$ for all x

2. $\int_{-\infty}^{\infty} P_x(x)\,dx=1$

pdf는 어떤 모양을 취해도 좋다 가장 일반적인 것은〈fig 5〉와 같은 종형의 정규확율 밀도함수이다.

정규 pdf는 다음 식으로 주어진다.

$$P_x(x)=\frac{1}{\sqrt{2\pi\sigma_x^2}}\ e^{-1/2(X-Mx)^2\cdot\sigma x2} \tag{10}$$

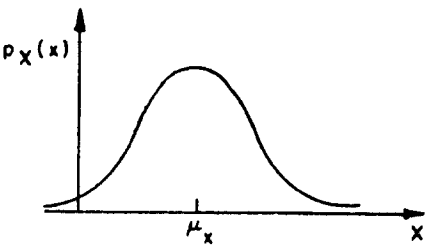

Fig 5. Normal distribution.

정규분포는 평균 μ_x에 대하여 대칭이고 범위는 $-\infty$에서 $+\infty$까지이다. 정규분포는 음의 값에 대해서도 확율을 취하며 홍수빈도 분포는 일반적으로 대칭이 아니므로 홍수빈도 결정에 사용되지 않는다. 예를들면 Fig. 2의 Beargrass Creek data는 전형적인 첨두유량 data의 tailing off가 우측에 있다.

정규분포가 홍수빈도 해석에는 일반적으로 사용되지 않으나 정규분포에 대한 이해가 통계적인 작업을 하는데 필수적이므로 계속해서 고려해보자.

정규분포의 cdf는

$$P_x(x)=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi\sigma_x^2}}\ e^{-1/2(t-\mu x)2/\sigma_x^2}dt\text{이고} \tag{11}$$

$$X\leq x\ \ P_x(x)=\text{prob}\ (X\leq x)\text{이다} \tag{12}$$

X가 a와 b사이에 있을 확율은

$$\text{Prob}\ (a \le X \le b) = \text{prob}(X \le b) - \text{prob}(X \le a)$$
$$= P_x(b) - P_x(a) \qquad (13)$$
$$= \int_a^b P_x(t)\,dt \text{ 이다}$$

정규 분포는 평균 μ_x와 표준 편차 σ_x의 2개의 parameter를 갖는 분포이다. 정규분포를 적용하기 위하여는 $\overline{X}$와 S_x에 의하여 μ_x와 σ_x를 추정해야 한다.

식 (4), (5)를 사용하여 Beargrass Creek data의 평균과 표준편차는 1599 cfs와 1006 cfs이다. 지금 정규분포가 Beargrass Creek Data에 대하여 적합하다면 그 data에 대하여 확율적으로 서술하는데 사용될 수 있다.

예를들면 2500 cfs 보다 작거나 같은 확율은

$$\text{prob}\ (Q \le 2500) = P_Q(2500)$$
$$= \int_{-\infty}^{2500} \frac{1}{\sqrt{2\pi\ 1006^2}}\ e^{-1/2(t-1599)2/1006^2} dt \qquad (14)$$

로 계산할 수 있다.

유감스럽게도 이 후. 의 식은 해석적으로 계산할 수는 없고, 수치해석 절차를 사용해야 한다. 매개변수 μ_x와 σ_x의 모든 가능한 경우의 조합에 대하여 정규분포에 대한 구분 수치 적분을 하는데 있어서 어려움을 피하기 위하여 다음과 같이 변수 변환을 한다. 즉

$$Z = \frac{X - \mu_x}{\sigma_x} \qquad (15)$$

여기서 Z는 표준화된 확율변수라 하며

$$P_z(z) = \frac{1}{\sqrt{2\pi}}\ e^{-z2/2} \text{ 을} \qquad (16)$$

표준 정규분포식이라 한다.

식(14)는 다음과 같이 계산할 수 있다.

$$\text{prob}\ (Q \le X) = \text{prob}\left(Z \le \frac{X - \mu_x}{\sigma_x}\right) \qquad (17)$$

$$\text{또는 prob}\,(Q \le 2500) = \text{prob}\left(Z \le \frac{2500 - 1599}{1006}\right.$$

$$= \text{prob}\ (Z \le 0.896) = \int_{-\infty}^{0.896} \frac{1}{\sqrt{2\pi}}\ e^{-z2/2}\ dz$$

후자의 식은 표준 정규 분포의 표를 이용하여 계산할 수 있다.

이 표를 이용할때는 표의 내용을 정확히 살펴보아야 한다.

prob $(Z \le 0.815) = 0.815$가 된다. (부록A참고)

이 말은 Beargrass Creak data가 평균이 1599 cfs이고 표준편차가 1006 cfs이며 정규분포를 따른다면 연첨두유량의 81.5%가 2500 cfs 보다 작거나 같다는 뜻이다. 실제로 표에서 보면 31개의 값중 28개 즉 90.3%가 2500 cfs보다 작거나 같다는 것을 알 수 있다.

식 (13)을 보면 prob$(a \le X \le b)$는 $X=a$와 $X=b$ 사이의 pdf의 면적임을 알 수 있다. 즉 a와 b사이에 떨어지는 관측지의 확율은 a와 b사이의 pdf의 면적이다. Fig. 2의 상대빈도 histogram에서도 비슷하게 알 수 있다. 예를들어 현재의 자료를 가지고 살펴보면 연 첨두유량이 1500 cfs와 2250 cfs 사이에 있을 확율은 0.19이다. 상대빈도수와 확율사이에는 분명히 관계가 있다. X_i에 중심을 두고 ΔX간격에 있는 관측치의 상대 빈도수를 $f_x(X_i)$라 표시하자.

이 구간내에 있을 관측지의 확율은

$$\text{prob}\,(X_i - \Delta X/2 \le X \le X_i + \Delta X/2)$$
$$= \int_{X_i - \Delta X/2}^{X_i + \Delta X/2} P_x(x)\,dx \qquad (18)$$

이며, 이것은 $X_i - \Delta X/2$와 $X_i + \Delta X/2$ 사이의 $P_x(x)$의 면적이다. 이 면적은 $\Delta X P_x(X_i)$로 대략 구할 수 있다. (Fig. 6)

그러므로 ΔX구간내의 관측치의 상대빈도수와 pdf 사이의 관계는

$$f_x(x_i) = \Delta X P_x(x_i) \text{ 이다.} \qquad (19)$$

확율은 pdf 밑의 면적에 관계가 있으므로 continuous random variable에 대하여 prob $(X=x)$는 0이어야 한다.

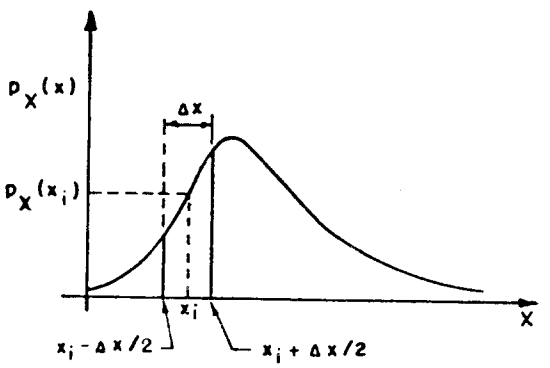

Fig 6. Calculation of prob $(X_i - \Delta X/2 < X < X_i + \Delta X/2)$.

Table 3. Observed and Expected Frequency-Beargrass Creek Data (normal distribution)

Class interval	Observed rel. freq.	Expected rel. freq.
0- 750	0.064	0.141
750-1500	0.581	0.267
1500-2250	0.194	0.286
2250-0300	0.064	0.177
3000-3750	0.032	0.063
3750-4500	0.032	0.012
4500-5250	0.032	0.001
	0.996	0.947