The Korean Communications
in Statistics Vol. 12 No. 2, 2005
pp. 443-451

Wakeby Distribution and the Maximum Likelihood Estimation Algorithm in Which Probability Density Function Is Not Explicitly Expressed[1)]

Jeong-Soo Park [2)]

Abstract

The studied in this paper is a new algorithm for searching the maximum likelihood estimate(MLE) in which probability density function is not explicitly expressed. Newton-Raphson's root-finding routine and a nonlinear numerical optimization algorithm with constraint (so-called feasible sequential quadratic programming) are used. This algorithm is applied to the Wakeby distribution which is importantly used in hydrology and water resource research for analysis of extreme rainfall. The performance comparison between maximum likelihood estimates and method of L-moment estimates (L-ME) is studied by Monte-carlo simulation. The recommended methods are L-ME for up to 300 observations and MLE for over the sample size, respectively. Methods for speeding up the algorithm and for computing variances of estimates are discussed.

Keywords : L-moment estimation, Numerical optimization, Hydrology, Quantile function, Newton-Raphson algorithm

1. 서론

통계학에서 확률분포의 모수를 추정하는 방법으로 최우추정법이 가장 널리 알려져 있고, 이론상으로도 점근최적 방법이다. 그런데 만약 확률밀도함수(pdf)가 명확히 표현되지 않는 경우는 확률밀도함수의 곱인 우도함수 또한 명확히 표현되지 않으므로 최우추정치를 구하는데 어려움이 생긴다. 예를들어 백분위함수가 다음과 같이 매우 간단한 1-모수 분포(Skew 로지스틱 분포)의 경우에도,

$$x_F = Q(F) = \alpha \ln(F) - (1-\alpha)\ln(1-F),$$

1) This work was support by grant No. R05-2002-000-00629-0 from the Basic Research Program of the Korea Science & Engineering Foundation.

2) Professor, Department of Statistics, Chonnam National University, Gwangju, 700-757, Korea.
E-mail: jspark@chonnam.ac.kr

확률밀도함수를 명확히 표현할 수 없다. 그러면 이런 경우에 최우추정치(MLE)는 어떻게 구할 것인가?

본 연구에서는 이러한 문제의 해답을 구하는 수치적 최우추정 알고리즘을 개발하고 그것을 웨이크비 분포 (Wakeby distribution)에 적용하였다. 웨이크비 분포는 수문학 분야에서 중요 관심사이지만 pdf가 명확히 표현되지 않아 지금까지 최우추정치가 사용되지 못했고 L-적률추정치가 사용되어 왔다. 웨이크비 분포는 일일 강수량 또는 년 중 최대 일일 강수량 (극값)의 확률분포로서 수문학, 토목공학, 수자원연구에 자주 쓰이는 분포이다 (맹승진, 2000; 허준행, 1997; Landwehr et.al. 1979, 1980; Park et.al. 2001). 백분위함수는 두 개의 일반화 파레토 분포의 합으로 구성되는 5-모수 분포이며, 다음과 같이 표현된다.

$$Q(F) = \xi + \frac{\alpha}{\beta}\{1-(1-F)^{\beta}\} - \frac{\gamma}{\delta}[1-(1-F)^{-\delta}], \quad 0 < F < 1. \tag{1.1}$$

$\underline{\theta} = (\xi, \alpha, \beta, \gamma, \delta)$ 라는 다섯 개의 모수를 가지는 특별한 분포로서 정의역은 다음과 같다.

$$\begin{aligned} &\xi \le x < \infty \quad \text{if} \quad \delta \ge 0 \quad \text{and} \quad \gamma > 0, \\ &\xi \le x \le \xi + \alpha/\beta - \gamma/\delta \quad \text{if} \quad \gamma = 0 \quad or \quad \delta < 0. \end{aligned} \tag{1.2}$$

이 분포는 특별한 경우로서, 일반화 극단분포(GEV), 일반화 파레토분포($\gamma = 0$ 또는 $\alpha = 0$ 일때), 로그-정규분포, 로그-감마분포를 포함하는 매우 폭넓은 분포이다. 이 분포는 pdf나 누적분포함수가 명확히 표현되지 않으며 다만 백분위 함수만 명백히 표현된다. 모수 추정방법으로 지금까지는 백분위수에 대한 최소제곱법 (Karian and Dudewicz, 2000)이나 L-적률 추정법 (Hosking, 1990)이 사용되어 왔다. 최우추정법은 계산상의 어려움 때문에 지금껏 이용되지 못했지만, 대표본일 때는 최우추정치가 L-적률 추정법보다 좋을 것으로 기대된다. 본 연구의 결과로서 그동안 이 분포에 적용하지 못했던 우도함수에 기초한 추론 및 최적 분포의 선택 (AIC, BIC 등)이 가능하게 된다. 또한 소표본일 때 L-적률 추정법과 최우추정법의 성능을 비교하였다.

2. 수치적 최우추정 알고리즘

오직 백분위함수 만을 이용하여 최우추정치를 계산하는 알고리즘의 개략적인 단계를 보면 다음과 같다.

단계 1. 주어진 한 관측치 x 에 대해서 $x = Q(F)$ 을 풀어서 F(x)를 구한다. 이때 Newton-Raphson 형의 반복적 수치계산을 통하여 F(x)를 구한다.

단계 2. 단계1에서 구한 F(x)에 대해 다음과 같은 식을 통하여 확률밀도함수 f(x)를 구한다.

$$f(x) = [\, dQ(F)/dF]^{-1} \,|_{F=F(x)}.$$

단계 3. 단계1과 2를 모든 x에 대해 반복하여 실행한 뒤, 음의 로그우도함수(λ)를 계산한다.

단계 4. λ 를 최소로 하는 모수를 (1.2)의 제약조건 하에서의 최적화 알고리즘을 이용하여 수치적으로 구한다. 이때 필요한 각 모수에 대한 λ의 미분벡터도 수치미분으로 구해서 이용한다.

단계 5. 초기치 여러 개를 주어서, 각각 도달한 최저치 중에서 가장 작은 λ 값을 갖는 모수를 최종적인 최우추정치(global optimizer)로 간주한다.

위의 단계2에서 구해지는 확률밀도함수는 정의역 (1.2) 하에서 다음과 같은 형태가 된다.

$$f(x) = \frac{(1-p_x)^{\delta+1}}{\alpha(1-p_x)^{\beta+\delta}+\gamma}. \tag{2.1}$$

여기서 p_x 는 위의 단계1에서 x 에 대응하여 구해지는 F(x) 를 뜻하며, 이는 다섯 개의 모수 $\underline{\theta} = (\xi, \alpha, \beta, \gamma, \delta)$ 의 함수이다. 이제 음의 로그우도함수는 다음과 같다.

$$\lambda = -\ln L(\underline{\theta};\underline{x}) = -\sum_{i=1}^{n}[(\delta+1)\ln(1-p_{x_i}) - \ln\{\alpha(1-p_{x_i})^{\beta+\delta}+\gamma\}] \tag{2.2}$$

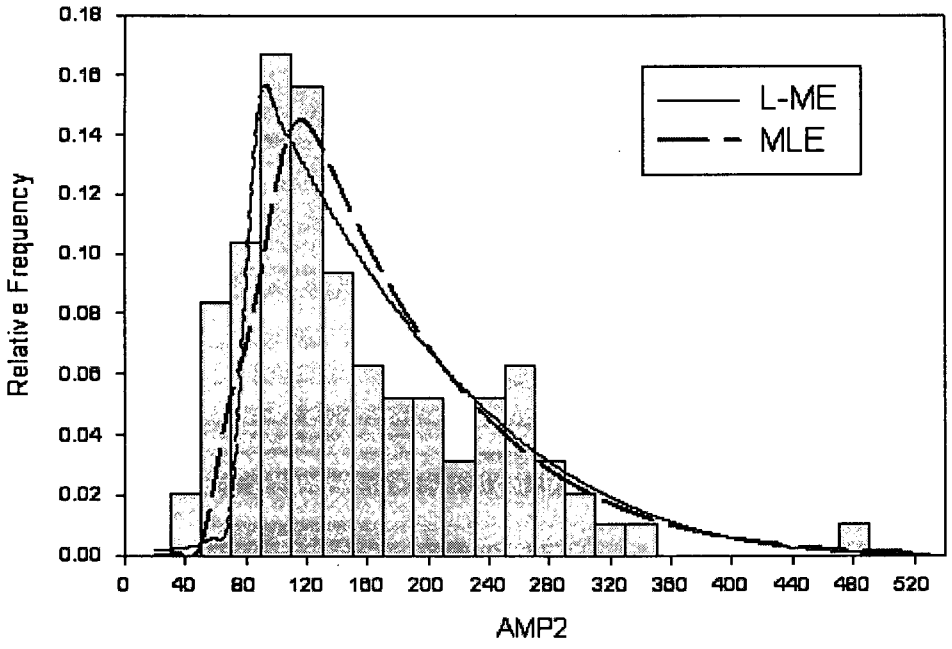
이제 식(2.2)를 최소로 하는 모수들을 구하기 위해 수치적 최적화 알고리즘을 이용하여야 한다. 그런데 (1.2)와 같은 모수에 대한 비선형 제약조건이 있는데다가 최적점을 찾아가는 과정에서 이 제약조건 밖으로 나가면 식(2.2)가 정의되지 않기 때문에, 일반적인 최적화 알고리즘이 아닌 feasible sequential quadratic programming(FSQP)을 사용하였다(Nocedal and Wright, 1999). 이를 위해 FFSQP 라는 포트란 프로그램을 적용하였다(Lawrence and Tits, 2001). 또한 단계1에서 $x = Q(F)$ 를 푸는데 필요한 Newton-Raphson 알고리즘은 Hosking(2000)이 작성하고 홈페이지에 올려둔 포트란 프로그램들을 사용하였다. 웨이크비 분포에서 MLE를 구하는 포트란 프로그램은 저자에게 연락하여 구할 수 있다.

3. 웨이크비 분포에서 추정방법의 비교

웨이크비 분포에 대해서는 지금까지 주로 L-적률추정법이 적용되어 왔다. 이 절에서는 L-적률추정법(Method of L-moment estimation; L-ME)과 최우추정법의 성능을 몬테카를로 시뮬레이션을 통하여 비교하였다.

L-적률은 순서통계량의 선형결합으로 정의되며, 모집단의 L-적률과 표본의 L-적률을 등치시킨 연립방정식의 해로써 L-적률추정치가 구해진다. 이 추정치는 일반적으로 모수의 수가 많은 경우(3개 이상), 적률추정치보다 더 좋은 추정을 하며 특히 소표본에서 최우추정치보다 더 좋다고 알려져 있다. L-적률추정법에 관한 자세한 내용은 Hosking(1990) 또는 박정수, 황영아(2005)를 참조하고, 특히 웨이크비 분포에서의 L-적률추정은 Hosking and Wallis(1997)나 Park et.al(2001)을 보기 바라며, 여기서는 구체적인 내용을 생략한다. L-적률추정치를 구하는 알고리즘은 Hosking(2000)이 작성하고 홈페이지에 올려둔 포트란 프로그램들을 사용하였다. 실제적인 적용 사례로서, <그림3.1>

은 부산의 2일 년 최대강수량의 시계열 (1904년-1999년) 자료에 대해 L-적률추정법과 최우추정법으로 추정한 결과이다.



<그림 3.1> 부산의 2일 년 최대강수량의 시계열(1904년-1999년) 자료의 상대도수 히스토그램과 웨이크비 분포의 L-적률추정법(L-ME)과 최우추정법(MLE)을 사용했을 때의 확률밀도함수

웨이크비 분포에 대한 기존의 연구(Landwehr, Matalas and Wallis, 1979, 1980) 및 본 연구에서는 크게 6개로 구분한 다음 <표3.1>과 같은 모수들에 대해 시뮬레이션 하였다. 여기서 ξ 는 location-equivariant 하므로 항상 0 으로 두었고 α/β 는 scale-equivariant 하므로 1 로 놓고 난수를 발생시켰다. 또한 아래와 같은 설정 하에서 이루어졌다.

표본의 크기: 50, 150, 300, 500
백분위수 추정에 이용된 백분위: 90, 95, 98, 99.5
반복횟수: 1500

<표3.2>와 <표3.3>은 각각 모수와 백분위수의 추정에 대한 시뮬레이션 결과로서 RMSE (Root Mean Squared Error)가 제시되었다. 먼저 모수의 추정치에 대한 RMSE를 보면 ξ, α, β 에 대해서는 전체적으로 MLE 가 작고 γ, δ 에 대해서는 표본의 크기가 150이나 300 까지는 L-ME

가 작고 그 보다 큰 경우는 MLE가 작은 경향을 보인다.

<표 3.1> 시뮬레이션에 사용되는 6개의 웨이크비 분포의 모수 설정

분포	$\alpha = \beta$	γ	δ
WA-1	16.0	0.8	0.2
WA-2	7.5	0.6	0.12
WA-3	1.0	0.6	0.12
WA-4	16.0	0.4	0.04
WA-5	1.0	0.4	0.04
WA-6	2.5	0.2	0.02

주시할 점은 α, β 에 대해 두가지 방법 모두 불안정하게 추정하고 있음을 알 수 있다. 한편 백분위수의 추정치에 대한 RMSE 를 보면 WA-6와 90 백분위수는 전반적으로 MLE 가 좋았다. 그러나 98 및 99.5 백분위수는 대체로 표본의 크기가 300 일 때까지 L-ME 가 좋았다. 특히 WA-2는 표본의 크기가 500 일 때까지도 L-ME가 더 좋다.

결론적으로 이 분포에서는 모수보다는 백분위수의, 특히 98백분위수 이상에서의 정확한 추정이 더 중요한 점을 감안하면 표본의 크기가 300 정도까지는 L-ME를 사용하고 그 이상에서는 MLE의 사용이 권장된다.

특기할 점은 ξ, α, β 의 추정오차의 크기에 상관없이 γ, δ 의 추정오차가 작은 경우에 백분위수의 추정오차가 줄어든다는 것이다. 일부 분포에서 α, β 에 대한 L-ME의 추정오차가 매우 큼에도 불구하고 백분위수의 추정은 MLE 보다 정확한 점을 본다면 γ, δ 의 추정이 백분위수의 추정에 상대적으로 훨씬 중요하게 그리고 민감하게 작용한다는 사실을 알 수 있다.

4. 토의 및 결론

본 논문에서 기술한 수치적 MLE 알고리즘의 문제점은 우도함수를 계산하는데 있어서 L-적률 방법에 비하여 시간이 오래 걸린다는 사실이다. 예를 들어 자료의 수가 400개인 경우, 네 번의 초기치를 시도한다고 했을 때 MLE를 구하는데 걸리는 총 계산 시간은 PC에서 36초 정도(1번 계산 당 약 8.5초)이다. 그런데 L-적률 방법은 0.002초 걸린다. 따라서 MLE 계산 시간을 단축시키기 위하여, 단계1에서 누적분포함수의 값 F(x)를 구하기 위해, Newton-Raphson 알고리즘의 2차형으로서 매우 빠른 Halley의 방법 (Huh, 1986)이 이용되었는데 이는 Hosking(2000)에 의해 구현되었다. (여기서 Huh(1986)의 연구와 다른 점은 백분위수가 주어졌을 때 F(x)를 구한다는 것이다.) 이 경우 Newton-Raphson 알고리즘보다 1.5배정도(n=400일 때 MLE 1번 계산 당 약 6초) 빨라졌다.

<표 3.2> 각 분포와 표본의 크기에 대해 모수 추정에 따른 시뮬레이션 결과 RMSE

	모수	ξ		α		β		γ		δ	
분포	표본크기	MLE	L-ME	MLE	L-ME	MLE	L-ME	MLE	L-ME	MLE	L-ME
WA-1	50	0.321	4.116	11.08	6813.7	14.64	140.1	0.302	0.287	0.279	0.211
	150	0.135	4.213	4.738	8562.0	4.951	158.4	0.145	0.166	0.12	0.124
	300	0.075	2.265	3.447	2961.8	3.558	78.56	0.104	0.127	0.085	0.093
	500	0.047	1.061	2.573	1547.9	2.658	52.36	0.078	0.103	0.062	0.071
WA-2	50	0.182	1.775	6.886	1276.4	10.88	46.23	0.318	0.285	0.329	0.236
	150	0.071	0.22	1.923	45.54	2.412	7.787	0.156	0.16	0.144	0.134
	300	0.037	0.104	1.319	6.689	1.598	2.982	0.109	0.115	0.101	0.099
	500	0.022	0.063	0.965	1.602	1.134	1.545	0.078	0.086	0.071	0.073
WA-3	50	0.046	3.801	1.334	721.9	9.345	55.28	0.509	0.61	0.564	0.371
	150	0.015	0.081	0.953	8.862	22.59	14.81	0.45	0.486	0.394	0.307
	300	0.007	0.62	0.745	90.53	19.09	23.79	0.405	0.424	0.334	0.293
	500	0.004	0.107	0.484	5.562	380.8	35.01	0.372	0.396	0.266	0.282
WA-4	50	0.299	2.227	8.819	4027.9	10.65	95.14	0.151	0.146	0.25	0.203
	150	0.133	0.487	4.372	66.43	4.374	18.79	0.071	0.081	0.112	0.114
	300	0.073	0.181	2.916	13.13	2.985	6.93	0.05	0.06	0.076	0.081
	500	0.045	0.123	2.255	6.208	2.271	4.186	0.038	0.047	0.057	0.062
WA-5	50	0.039	0.462	1.189	79.34	10.11	47.19	0.55	0.794	0.505	0.369
	150	0.014	0.182	0.757	5.162	13.24	12.42	0.45	0.453	0.426	0.337
	300	0.006	0.037	0.457	4.895	14.74	9.037	0.377	0.388	0.346	0.34
	500	0.004	0.019	0.708	2.23	4.337	6.71	0.318	0.35	0.245	0.322
WA-6	50	0.073	0.128	1.443	39.89	5.297	8.941	0.411	0.343	0.612	0.444
	150	0.023	0.041	0.443	0.556	0.973	1.136	0.178	0.2	0.257	0.349
	300	0.013	0.028	0.28	0.358	0.524	0.702	0.108	0.139	0.172	0.27
	500	0.008	0.021	0.224	0.27	0.392	0.539	0.075	0.11	0.122	0.208

<표 3.3> 각 분포와 표본의 크기에 대해 백분위수 추정에 따른 시뮬레이션 결과 RMSE

	백분위	90		95		98		99.5	
분포	표본크기	MLE	L-ME	MLE	L-ME	MLE	L-ME	MLE	L-ME
WA-1	50	0.429	0.42	0.704	0.66	1.426	1.235	4.538	3.269
	150	0.251	0.252	0.413	0.404	0.808	0.772	2.056	1.983
	300	0.174	0.175	0.286	0.28	0.562	0.548	1.406	1.416
	500	0.141	0.146	0.23	0.23	0.441	0.434	1.058	1.072
WA-2	50	0.283	0.289	0.441	0.432	0.871	0.743	3.493	1.74
	150	0.159	0.162	0.241	0.24	0.445	0.419	1.099	0.964
	300	0.119	0.121	0.182	0.181	0.334	0.322	0.788	0.752
	500	0.088	0.09	0.134	0.134	0.239	0.235	0.545	0.539
WA-3	50	0.336	0.336	0.48	0.481	0.879	0.802	6.109	1.869
	150	0.19	0.191	0.274	0.278	0.464	0.467	1.168	1.022
	300	0.136	0.137	0.195	0.205	0.323	0.346	0.742	0.718
	500	0.11	0.109	0.155	0.162	0.249	0.274	0.569	0.581
WA-4	50	0.15	0.151	0.221	0.215	0.395	0.354	1.316	0.756
	150	0.087	0.088	0.127	0.126	0.219	0.213	0.468	0.453
	300	0.062	0.064	0.089	0.09	0.15	0.149	0.31	0.311
	500	0.048	0.05	0.069	0.07	0.116	0.116	0.237	0.239
WA-5	50	0.196	0.2	0.249	0.258	0.375	0.372	1.309	0.724
	150	0.114	0.116	0.15	0.158	0.227	0.238	0.471	0.454
	300	0.083	0.085	0.108	0.117	0.16	0.18	0.328	0.334
	500	0.064	0.065	0.087	0.093	0.129	0.146	0.251	0.272
WA-6	50	0.091	0.088	0.115	0.121	0.175	0.182	0.917	0.351
	150	0.049	0.051	0.064	0.076	0.1	0.117	0.225	0.222
	300	0.034	0.037	0.047	0.057	0.071	0.086	0.151	0.158
	500	0.027	0.028	0.036	0.043	0.055	0.062	0.117	0.126

위의 알고리즘은 모든 x 에 대해서 F(x)를 구했는데, 이는 시간이 많이 걸리므로 이를 개선하기 위해 일부의 x 에 대해서만 F(x)를 구한 뒤에 내삽법을 적용하여 나머지에 대해 F(x)의 근사값을 구한다. 특히 F(x) 가 단순 비감소함수라는 특성을 이용하면 내삽오차를 줄일 수 있을 것이다. 이때 기준점으로 어떤 x를 몇 개 선택할 것인가가 연구과제이다.

최우추정치에 대한 분산을 추정하기 위해 핏셔의 정보 행렬의 역행렬을 이용할 수 있다. 이때 핏셔의 정보 행렬를 구하기 위해서는 로그 우도함수에 대한 2차 미분이 필요하게 되는데, pdf가 명백히 표현되지 않는 경우이므로 2차 수치미분을 이용할 수밖에 없다. 수치미분은 본질적으로 오차를 수반하므로 가장 오차가 작은 알고리즘을 이용해야 할 것이다. 그런데 문제는 ξ 의 조건 (1.2) 때문에 정상성조건(regularity condition)이 만족되지 않아서 점근정규성이나 점근효율성을 보장하지 못한다는 점이다. 또한 핏셔의 정보 행렬이 어떤 경우는 존재하지 않을 수도 있다는 점이다. 따라서 모수들에 대해서 정상성 조건을 만족하게 하는 어떤 제약조건을 줄 수 있는데, 핏셔의 정보 행렬이 수리적으로 명확히 표현되지 않기 때문에 이러한 일 또한 쉽지 않다. 이러한 어려움 때문에 추정치에 대해 붓스트랩에 의한 분산 계산이 바람직해 보인다.

본 연구의 결과로서 그동안 이 분포에 적용하지 못했던 우도함수에 기초한 추론 및 최적 분포의 선택 (AIC, BIC 등)이 부분적으로 가능하게 된다. 추정 방법의 선택에서는 이 분포에서는 모수보다는 백분위수의, 특히 98백분위수 이상에서의 정확한 추정이 더 중요한 점을 감안하면 표본의 크기가 300 정도까지는 L-ME를 사용하고 그 이상에서는 MLE의 사용이 권장된다.

참고문헌

[1] 맹승진 (2000). 「수문 자료의 통계학적 분석 방법」, 한국수자원공사 연구 홈페이지에서 입수, http://www.kowaco.or.kr/~water/water-dic/seawater/waterco16-4.html

[2] 박정수, 황영아 (2005). 3-모수 카파분포에서 추정방법들의 비교, 「한국통계학회논문집」, 제12권 2호, 인쇄중.

[3] 허준행 (1997). 수문통계학의 기초(5), 「한국수자원학회지」, 제30권 1호, 88-96.

[4] Hosking JRM (1990). L-moments: analysis and estimation of distributions using linear combinations of order statistics. *Journal of The Royal Statistical Society, Series B*, Vol. 52(1), 105-124.

[5] Hosking, JRM (2000). LMOMENTS: Fortran routines for use with the method of L-moments, Version 3.03, available at http://www.research.ibm.com/people/h/hosking/lmoments.html.

[6] Hosking, J.R.M., and Wallis, J.R., (1997). *Regional Frequency Analysis: An Approach based on L-moments*. Cambridge University Press, Cambridge.

[7] Huh, M. Y. (1986). Computation of percentage points. *Communications in Statistics-Simulation and Computation*, Vol. 15, 1191-1198.

[8] Karian, Z., and Dudewicz, E.J. (2000). *Fitting Statistical Distribution*, CRC Press, Boca Raton, Florida.

[9] Landwehr J.M., Matalas N.C., and Wallis J.R. (1979). Estimation of parameters and quantiles of Wakeby distributions. *Water Resources Research*, Vol. 15, 1361-1379.

[10] Landwehr JM, Matalas NC, and Wallis JR. (1980). Quantile estimation with more or less floodlike distributions. *Water Resources Research.* Vol. 16, 547-555.

[11] Lawrence CT, and Tits A. (2001). A computationally efficient feasible sequential quadratic programming algorithm. *SIAM Journal of Optimization*, Vol. 11(4), 1092-1118.

[12] Nocedal, J. and Wright, SJ. (1999). *Numerical Optimization,* Springer, New York.

[13] Park JS, Jung HS, Kim RS, and Oh JH (2001). Modelling summer extreme rainfall over the Korean peninsula using Wakeby distribution. *International Journal of Climatology,* Vol. 21, 1371-1384.

[2005년 2월 접수, 2005년 6월 채택]