小標本への両側有界分布の適用について

ON APPLICATION OF A PROBABILITY DISTRIBUTION WITH LOWER- AND UPPER-BOUNDS TO SMALL HYDROLOGIC SAMPLES

宝　　馨[1]・土佐香織[2]

Kaoru TAKARA and Kaori TOSA

[1]正会員 工博 京都大学教授 防災研究所 水災害研究部門（〒611-0011 宇治市五ヶ庄）
[2]修士（工学） プロクター・アンド・ギャンブル・ファー・イースト・インク（〒658-0032 神戸市東灘区向洋町中 1-17）

This paper exmamines the usefulness of a probability distribution with lower- and upper-bounds for small samples with a size of 20 to 30. Small samples for more water periods and less water periods are extracted from a long-term (10,000) series of lognormal variates generated by the Monte Carlo technique. The accuracy of T-year event estimates is assessed when the probability distribution with upper-bound is applied to these small samples. The results here recommend that for such samples that include large events the incorporation of upper-bound gives better T-year event estimates. It avoids overestimation of T-year quantiles. However, the use of the probability distribution with no upper bound is recommended for such small samples that only small events are included in the sample. The use of the probability distribution with upper bound for such samples tends to underestimate the T-year events.

Key Words : *small sample, accuracy of quantile estimates, Monte Carlo experiment, Slade-type distribution, lognormal distribution*

1. はじめに

ある地点の100年確率降水量を推定する際，20年や30年分の資料（以後，小標本という）しか手に入らない場合がある．そのような標本の観測期間が，たまたま少雨が続いた期間ばかりであれば，その地点に対して小さすぎる確率降水量を推定してしまう危険性がある．逆に，たまたま多雨が続いた期間であれば，その地点にとっては必要以上に大きな確率降水量を推定してしまうかもしれない（**Fig. 1 (a)** 参照）．

確率水文量を推定しようとしている地点に対して十分に大きな上限値を導入することにより，たとえ手元に小標本しかなく，またその分布が偏っている場合でも，データの蓄積が進んだ場合との誤差が比較的小さな値を得られる可能性が高まるのではないだろうか（**Fig. 1 (b)** 参照）．

本研究では，このような考えに基づき，水文頻度解析に上限値を導入することが，小標本の確率水文量の推定精度にどのような影響を与えるかを検証する．

2. 検証の手順

(1) データセットの作成

毎年極値水文量が従うと思われる母分布を想定し，その母分布に基づく乱数を10000個発生させ，それを擬似的な水文量時系列のデータセットとみなす．たとえば，これを1万年の年最大水文量の系列とみなす．ここでは，次のような3母数 対数正規分布:

$$f(x)=\frac{1}{(x-a)\sigma_Y\sqrt{2\pi}}\times \exp\left[-\frac{1}{2}\left\{\frac{\ln(x-a)-\mu_Y}{\sigma_Y}\right\}^2\right] \qquad (1)$$

を母分布として用いる．ここに，a, μ_Y, σ_Y は母数である．下限値 $a=0$ と固定するため，実質2母数対数正規分布である．したがって，以下ではこの分布をLN2分布と記すことにする．

(2) 変動傾向の把握

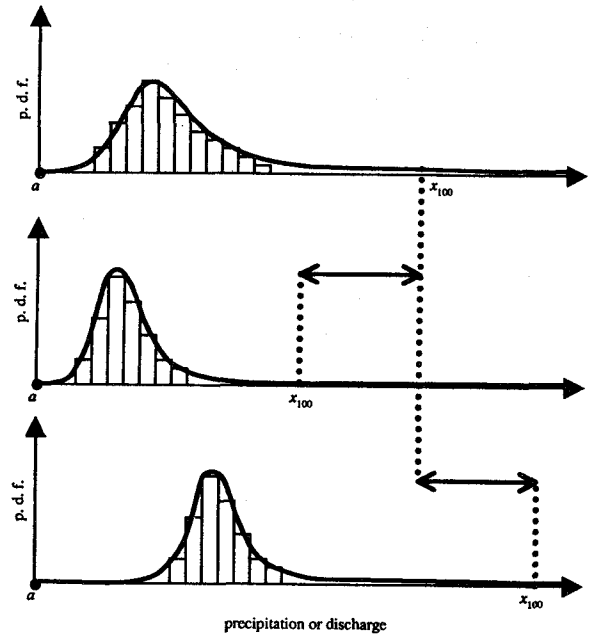
用いるデータセットの平均値と5年移動平均を求めてグラフ化し，おおまかな変動傾向を視覚的に把握する．

(3) データセットの抽出

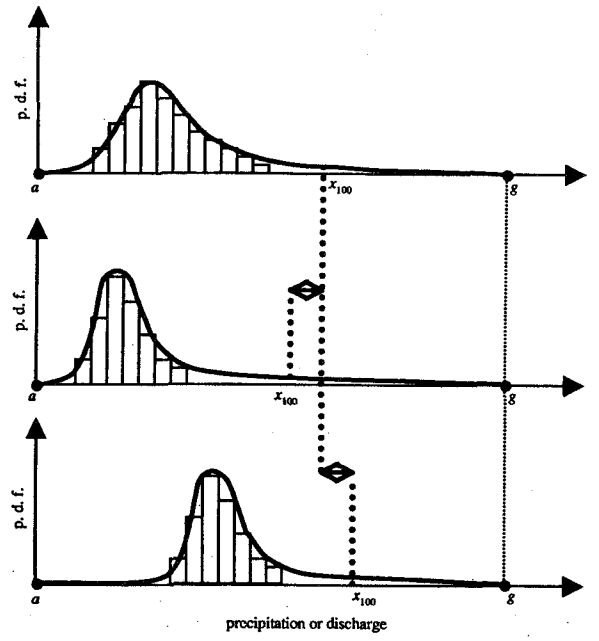
10000個のデータを含む1組の時系列データセット（これをAllと呼ぶことにする）の中から，次のような小標本データセットを抽出する．

Min. 20 連続20年分のデータだけを見たときに，平均値が最も小さくなる部分から成るデータセット

Min. 30 連続30年分のデータだけを見たときに，平均値が最も小さくなる部分から成るデータセット



(a) 下限値のみ有する場合



(b) 上下限値共に有する場合

Fig. 1 上限値導入の効果（上限値を導入すると確率水文量の推定精度の向上が期待できる.）

Max. 20 連続 20 年分のデータだけを見たときに，平均値が最も大きくなる部分から成るデータセット

Max. 30 連続 30 年分のデータだけを見たときに，平均値が最も大きくなる部分から成るデータセット

Fig. 2 に抽出の様子を示す.

(4) 確率水文量の推定

全てのデータから成るデータセット (All) と 3. で抽出した 4 種のデータセット (Min. 20, Min. 30, Max. 20, Max. 30) の計 5 種類のデータセットから，それぞれ 100 年確率水文量を推定する．その際，次の頻度解析モデルを用いる.

両側有界の分布として，岩井 (1949)[1] によって紹介された Slade 型の両側有界分布 (4 母数対数正規分布) (以下 Slade 分布という) :

$$f(x) = \frac{g-a}{(x-a)(g-x)\sigma_Y\sqrt{2\pi}} \times \exp\left[-\frac{1}{2}\left\{\frac{\ln\{(x-a)/(g-x)\}-\mu_Y}{\sigma_Y}\right\}^2\right] \quad (2)$$

a, g, μ_Y, σ_Y は母数である．ここでは，上下限値 a, g 固定のため，実質 2 母数である.

なお，わが国の水文頻度解析で多用されている 3 母数対数正規分布の下限値が既知 (a=0) として，これも同様に各データセットに当てはめ，Slade 型の両側有界分布と比較検討する.

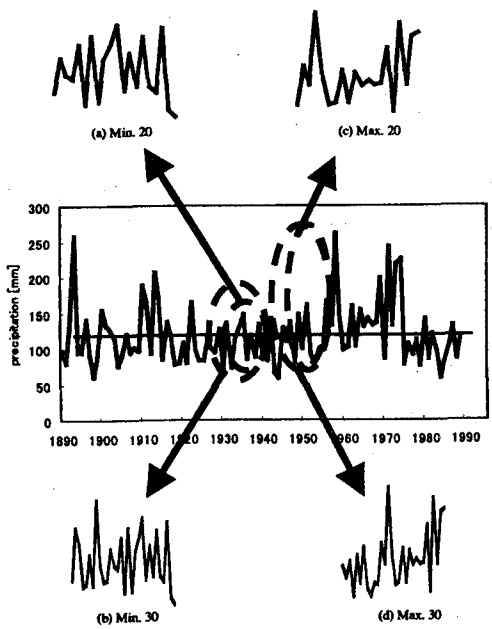


Fig. 2 データセットの抽出

(5) 評価

得られる確率水文量の推定値と推定精度を評価する．このとき，LN2 分布を用いて All から推定する値を真

値 (x_{100}) とする.

Min. 20・Min. 30・Max. 20・Max. 30 からそれぞれ推定される 100 年確率水文量の値 ($\hat{x}_{100}$) と真値の差を，真値で除した値 E_r を比較のための指標とする.

$$E_r = \frac{x_{100} - \hat{x}_{100}}{x_{100}} \tag{3}$$

相対誤差 E_r が 0 に近いほど良い推定であると言える.

3. 上限値の設定

両側有界の分布を導入する際，その上下限値をどのようにして求めるかが大きな問題となる．今回は，もとのデータセット (All) の最大値 (max) を，それぞれ 1, 1.5, 2, 5, 10 倍した値を上限値 g として与える．豪雨データを扱う場合には，可能最大降水量 (PMP) がその候補となるが，ここでは特定の水文量を想定しているわけではないので，発生した 10000 個のデータの中の最大値の 1 ～ 10 倍を設定することとしたのである．なお，下限値は a= 0 とする.

4. 結果と考察

(1) データセットの作成

Fig. 3 に，作成したデータセット (All) のヒストグラムを示す.

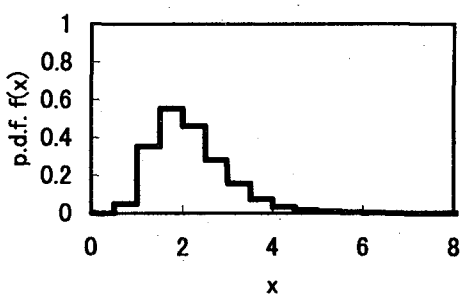


Fig. 3 作成したデータセット (All) のヒストグラム (対数正規分布に基づく)

(2) 変動傾向の把握

作成したデータセットを，求めた平均値・5 年移動平均と共にグラフ化する (**Fig. 4**).

(3) データセットの抽出

Table 1 に，もとのデータセット (All) と，そこから抽出したデータセット (Min. 20・Min. 30・Max. 20・Max. 30) の基本統計量 (データ数，最大値，最小値，平均値，標準偏差，歪係数) を示す.

さらに，**Fig. 5 (a)** ～ **(d)** に抽出した 4 種類のデータセットのヒストグラムを示す.

Table 1 各データセットの特性

	All	Min.20	Min.30	Max.20	Max.30
データ数	10000	20	30	20	30
最大値	8.18	2.82	3.88	5.82	5.13
最小値	0.52	0.97	0.73	1.56	1.23
平均値	2.19	1.66	1.71	3.05	2.75
標準偏差	0.71	0.30	0.49	1.71	1.15
歪係数	1.20	0.76	1.34	0.97	0.69

(4) 評価

それぞれのデータセットを用いた際に得られた 100 年確率水文量とその相対誤差を **Table 2** に示す.

この **Table 2** の結果から，次のようなことが読み取れる.

- 大きな値に偏ったデータセット (Max. 20, Max. 30) は，与える上限の値に関わらず，常に LN2 分布より確率水文量の誤差が小さい．それに対し，小さな値に偏ったデータセット (Min. 20, Min. 30) は，与える上限値が大きいと，LN2 分布より誤差が大きくなることがある.
- Slade 分布は，与える上限値が大きくなるほど誤差が大きくなるが，LN2 よりは誤差が小さい.
- 想定した母分布は LN2 であったので，All に対する 100 年確率水文量 4.84 が真値（にきわめて近い値）であるとみなせる．とすれば，Slade 分布は Max. 20, Max. 30 に対して，LN2 分布よりも良い（真値に近い）確率水文量を求めていることになる.
- Max. 20, Max. 30 の両者の結果を比べると，Max. 20 の確率水文量はかなり過大評価である．真値の近似値 4.84 に対して Slade 分布では 6.03 ～ 6.85 という値をとっている．上限無限の LN2 分布の場合にはさらに過大評価であり，7.09 となっている.
- 小さな値に偏ったデータセット (Min. 20, Min. 30) に着目すると，Slade 分布は LN2 分布よりも小さな確率水文量を与える．すなわち，上限値をもつ Slade 分布はこのようなデータセットに対して過小な確率水文量を与える傾向がある.
- 標本サイズが小さくなると (Min. 20)，Slade 分布による過小評価の程度が甚だしい.

小標本 (Min. 20 及び Max. 20) から得られる 100 年確率水文量をもとのデータセット (All) に対して確率評価をしてみると以下のようなことがわかった.

- Min. 20 から推定する 100 年確率水文量は，LN2 分布を用いる場合には All の場合の 10 年確率に，上限値として最大値の 5 倍の値を与えた Slade 分布を用いる場合には，All の場合の 9 年確率に相当する.

Table 2 両側有界分布の有用性評価 (100 年確率水文量，かっこ内は相対誤差 E_r，* は LN2 分布より E_r が小さいもの)

確率分布	上限値 g	All	Min. 20 (E_r)	Min. 30 (E_r)	Max. 20 (E_r)	Max. 30 (E_r)	* の数
Slade	8.2 (max × 1)	4.43	3.07 (0.365)	3.53 (0.270)	6.03 (0.246)*	5.36 (0.107)*	2
	12.3 (max × 1.5)	4.54	3.12 (0.354)	3.61 (0.254)	6.35 (0.312)*	5.64 (0.166)*	2
	16.4 (max × 2)	4.61	3.15 (0.349)	3.64 (0.247)	6.50 (0.344)*	5.78 (0.195)*	2
	40.9 (max × 5)	4.75	3.19 (0.340)	3.71 (0.233)	6.76 (0.398)*	6.03 (0.246)*	2
	81.8 (max × 10)	4.79	3.21 (0.337)	3.74 (0.228)	6.85 (0.415)*	6.11 (0.262)*	2
LN2	-	4.84	3.28 (0.322)	3.81 (0.212)	7.09 (0.466)	6.28 (0.298)	

Table 3 各データセットの特性

		観測年 (年数)	最大値 [mm]	最小値 [mm]	平均値	標準偏差	ひずみ係数
Ohtsu (1-day)	All	1912-1985 (74)	206.0	55.0	105.7	33.7	0.84
	Min. 20	1912-1931 (20)	142.0	55.0	90.5	23.0	0.48
	Min. 30	1913-1942 (30)	206.0	55.0	97.9	32.3	1.29
	Max. 20	1953-1972 (20)	192.0	80.0	123.5	36.1	0.32
	Max. 30	1946-1975 (30)	192.0	65.0	116.1	34.5	0.52
Ohtsu (2-day)	All	1912-1985 (74)	236.0	55.0	139.2	43.7	0.38
	Min. 20	1912-1931 (20)	189.0	55.0	119.1	35.2	0.16
	Min. 30	1919-1948 (30)	222.0	55.0	128.3	43.0	0.55
	Max. 20	1953-1972 (20)	236.0	92.0	159.5	47.2	0.24
	Max. 30	1949-1978 (30)	236.0	83.0	150.7	44.7	0.40
Hikone (1-day)	All	1912-1985 (74)	196.0	44.0	96.6	33.2	0.93
	Min. 20	1924-1943 (20)	130.0	44.0	80.9	22.6	0.36
	Min. 30	1914-1943 (30)	13.0	44.0	86.9	28.9	0.60
	Max. 20	1956-1975 (20)	196.0	64.0	109.6	37.2	0.95
	Max. 30	1943-1972 (30)	196.0	69.0	109.6	34.9	1.10
Hikone (2-day)	All	1912-1985 (74)	317.0	54.0	124.1	44.0	1.36
	Min. 20	1923-1942 (20)	163.0	54.0	99.4	27.6	0.37
	Min. 30	1919-1948 (30)	211.0	54.0	110.4	36.8	0.59
	Max. 20	1943-1962 (20)	317.0	86.0	143.7	49.7	2.18
	Max. 30	1943-1972 (30)	317.0	86.0	141.9	46.2	1.83
Imazu (1-day)	All	1912-1985 (74)	170.0	42.0	94.5	31.2	0.52
	Min. 20	1924-1943 (20)	148.0	42.0	78.7	27.7	1.19
	Min. 30	1914-1943 (30)	148.0	42.0	81.5	26.2	0.83
	Max. 20	1943-1962 (20)	170.0	59.0	113.0	31.7	0.01
	Max. 30	1943-1972 (30)	170.0	51.0	110.3	32.4	-0.05
Imazu (2-day)	All	1912-1985 (74)	250.0	51.0	124.1	46.9	0.85
	Min. 20	1921-1940 (20)	172.0	51.0	100.9	33.6	0.66
	Min. 30	1914-1943 (30)	172.0	51.0	104.5	32.1	0.48
	Max. 20	1942-1961 (20)	250.0	61.0	148.0	57.1	0.35
	Max. 30	1943-1972 (30)	250.0	59.0	144.8	58.9	0.19
Gifu (1-day)	All	1893-1992 (100)	260.2	59.5	118.3	42.0	1.35
	Min. 20	1928-1947 (20)	146.8	59.7	102.7	27.2	0.09
	Min. 30	1918-1947 (30)	164.9	59.7	103.0	27.4	0.42
	Max. 20	1958-1977 (20)	260.2	86.0	139.9	51.2	0.71
	Max. 30	1948-1977 (30)	260.2	78.4	153.1	48.4	0.95
Gifu (2-day)	All	1893-1992 (100)	446.0	79.3	152.9	62.7	2.30
	Min. 20	1929-1948 (20)	206.2	87.8	128.4	31.8	0.70
	Min. 30	1918-1947 (30)	206.2	79.3	127.7	30.1	0.57
	Max. 20	1958-1977 (20)	420.5	93.5	199.9	83.4	1.10
	Max. 30	1948-1977 (30)	420.5	93.5	181.5	74.0	1.57

Table 4 両側有界分布の有用性評価 (100 年確率水文量，かっこ内は相対誤差 E_r，* は LN2 分布より E_r が小さいもの)

		All	Min. 20 (E_r)	Min. 30 (E_r)	Max. 20 (E_r)	Max. 30 (E_r)	* の数
Ohtsu (1-day)	Slade	201.0	154.2 (0.250)	185.0 (0.100)	226.8 (0.104)*	213.9 (0.041)*	2
	LN2	205.6	158.3 (0.230)	189.8 (0.077)	235.9 (0.148)	220.5 (0.072)	
Ohtsu (2-day)	Slade	273.0	226.4 (0.194)	258.3 (0.080)	296.9 (0.057)*	280.6 (0.001)*	2
	LN2	280.9	234.9 (0.164)	267.7 (0.047)	308.9 (0.100)	289.5 (0.031)	
Hikone (1-day)	Slade	193.2	146.8 (0.259)	174.1 (0.122)	211.0 (0.065)*	201.9 (0.019)*	2
	LN2	198.2	151.3 (0.237)	180.0 (0.092)	218.8 (0.104)	206.9 (0.044)	
Hikone (2-day)	Slade	250.3	180.1 (0.297)	221.3 (0.136)	261.3 (0.020)*	259.4 (0.012)*	2
	LN2	256.3	185.3 (0.277)	228.4 (0.109)	266.9 (0.042)	264.7 (0.033)	
Imazu (1-day)	Slade	188.1	152.8 (0.208)	155.1 (0.196)	208.4 (0.080)*	212.2 (0.099)*	2
	LN2	193.0	157.5 (0.184)	159.0 (0.176)	217.0 (0.124)	220.8 (0.144)	
Imazu (2-day)	Slade	264.7	198.4 (0.272)	199.0 (0.270)	330.9 (0.214)*	383.0 (0.406)*	2
	LN2	272.5	205.2 (0.247)	204.3 (0.250)	352.4 (0.294)	424.4 (0.558)	
Gifu (1-day)	Slade	232.4	181.2 (0.237)	179.8 (0.243)	293.8 (0.237)*	272.3 (0.146)*	2
	LN2	237.6	187.2 (0.212)	184.3 (0.224)	308.1 (0.297)	282.7 (0.190)	
Gifu (2-day)	Slade	310.7	212.8 (0.326)	209.4 (0.336)	433.3 (0.373)*	374.6 (0.187)*	2
	LN2	315.6	217.7 (0.310)	213.3 (0.324)	460.4 (0.459)	387.7 (0.229)	

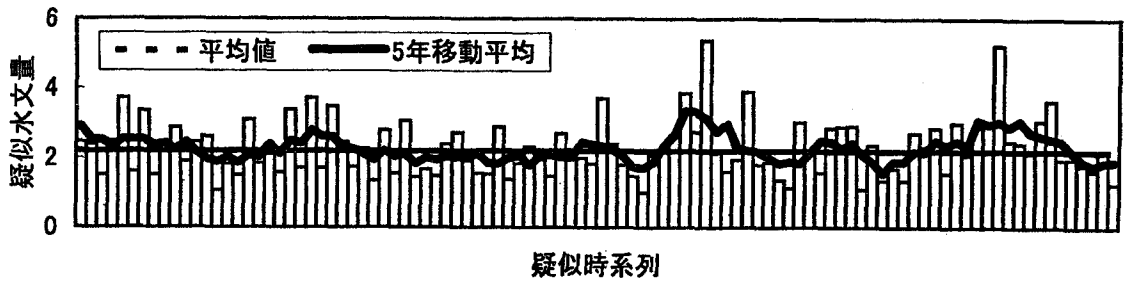


Fig. 4 変動傾向の把握 (一部分のみ)

- Max. 20 から推定する 100 年確率水文量は，LN2 分布を用いる場合には All の場合の 2500 年確率に，上限値として最大値の 5 倍の値を与えた Slade 分布を用いる場合には，All の場合の 1500 年確率に相当する．

以上のことから，次のように考えられる．

水文頻度解析モデルに上限値を導入することにより，手元にある標本のサイズが小さい，または分布が偏っていても，上限値を持たない LN2 分布より 100 年確率水文量の推定誤差が小さくなることが確かめられた．特に，大きな値に偏ったデータセット (Max. 20, Max. 30) については，最大値の 10 倍という，非常に大きな上限値を与えた場合であっても，確率水文量の過大評価の程度が小さくなる．しかし，小さな値に偏ったデータセット (Min. 20, Min. 30) については，上限値として最大値 10 倍の値を用いた場合でも，確率水文量を小さく評価し過ぎており，上限無限の LN2 分布を用いた方がよい．

これらのことから，手元にある小標本が，小さな値に偏っていると思われる場合 (周辺の地点の観測値と比べて，かなり小さな値しか観測されていないことが明らかな場合) に両側有界分布を用いることは，小さ過ぎる確率水文量を推定してしまう可能性があるため，避けた方がよい．逆に，それ以外の場合に両側有界分布を用いることは，確率水文量の過大評価を防ぐためには，かなり有効であると言える．

5. 実在水文量データセットを用いた検証

実在する水文量データセットを用いて同様の操作を行なうことにより，小標本に対する両側有界分布の有用性について，さらに検討する．

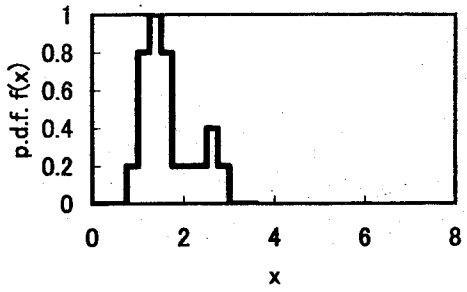
実在する水文量データセットとして，大津・彦根・今津・岐阜の年最大 1 日・2 日降水量を用いる．そのとき，日本で記録された 1 min ～ 1 year の既往最大地点降水量を図上にプロットし，それらのプロット点を包絡するような線 (降水量 - 継続時間曲線) を描くことにより，統計的に推定した PMP (可能最大降水量；Probable Maximum Precipitation : 1 日；1311 mm, 2 日；1813 mm) [2)3)] を上限値として用いる．

大津・彦根・今津の 74 年 (1912-1985) 及び岐阜の 100 年 (1893-1992) のデータセットから Min. 20, Min. 30, Max. 20, Max. 30 のデータセットを抽出し，各データセットの基本統計量を求めたところ **Table 3** のようであった．

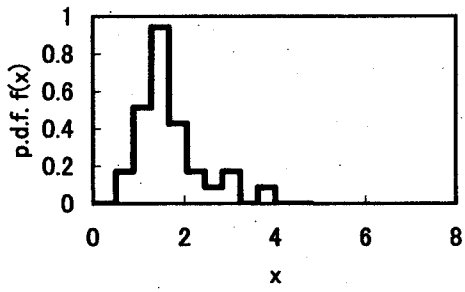
これらのデータセットに Slade 分布と LN2 分布をあてはめ，得られた 1000 年確率降水量とその推定誤差を **Table 4** に示す．

たかだか 74 年，または 100 年分だけのデータであるが，10000 年分の擬似的な水文量データセットを用いた場合とほぼ同様の結果が得られた．すなわち，

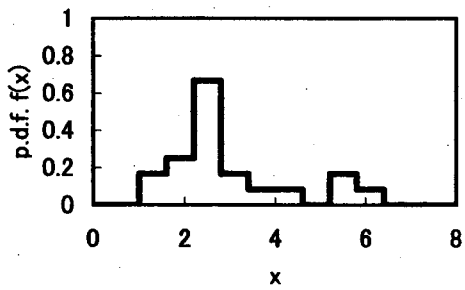
- 大きな値に偏ったデータセット (Max. 20, Max. 30) に対しては，Slade 分布の方が LN2 分布より確率降水量の誤差が小さい．一方，小さな値に偏ったデータセット (Min. 20, Min. 30) に対しては，与える上限値が大きいと，LN2 分布より誤差が大きくなる．
- 実際の水文量の場合には母分布はわからないので，All に対する双方の 100 年確率水文量が真値にある程度近い値であるとみなせる．とすれば，Slade 分布は Max. 20, Max. 30 に対して，LN2 分布よりも良い (真値に近い) 確率水文量を求めていることになる．LN2 は過大な 100 年確率降水量を与えていると言える．
- Max. 20, Max. 30 の両者の結果を比べると，Max. 20 の確率降水量は今津の場合をのぞき，かなり過大評価である．今津の年最大日降水量，年最大 2 日降水量の 74 年の資料から 20 年間の平均値が最大となる期間 (Max. 20) をとると，それはたまたま Max. 30 よりもバラツキの少ないデータの組合わせになったものである．一般には統計期間 (All の年数) がさらに長くなると，Max. 30 の方が Max. 20 よりも All に近い値を与えるはずである．



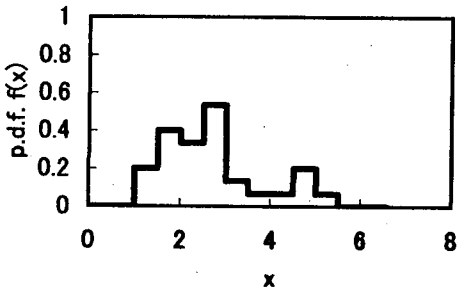
(a) Min. 20



(b) Min. 30



(c) Max. 20



(d) Max. 30

Fig. 5 抽出したデータセットのヒストグラム (対数正規分布に基づく)

- 小さな値に偏ったデータセット (Min. 20, Min. 30) に着目すると，Slade 分布は LN2 分布よりも小さな確率水文量を与える．すなわち，上限値をもつ Slade 分布はこのようなデータセットに対して過小な確率水文量を与える傾向がある．

6. 結 語

毎年最大水文量を用いて，50 ～ 200 年確率水文量を算定する場合，小標本（データが 20 個や 30 個程度しか無いデータセット）に確率分布を当てはめると，求められる確率水文量の値がかなり過大評価されたり過小評価されることがしばしば経験されてきた．本研究では，モンテカルロ実験による検証と実在の水文極値データによる検証を行って，筆者らが近年提唱している両側有界分布を使用した場合に，どの程度の過大・過小評価が生じるのかを明らかにした．

本研究で得られた知見は，小標本を取り扱うことの多い実務において大いに参考になるものと思われる．結果をまとめると以下のようである．

（1）手元にある標本のサイズが小さい場合，水文頻度解析モデルに上限値を導入することにより，上限値を持たない LN2 分布より 100 年確率水文量の推定誤差が小さくなること，すなわち，過大評価を避けることができることが確かめられた．

（2）特に，大きな値に偏ったデータセット (Max. 20, Max. 30) については，かなり大きな上限値を与えた場合であっても，確率水文量の過大評価の程度が小さくなる．

（3）しかし，小さな値に偏ったデータセット (Min. 20, Min. 30) については，上限値として最大値 10 倍の値を用いた場合でも，確率水文量を小さく評価し過ぎており，上限無限の LN2 分布を用いた方がよい．

（4）結局，手元にある小標本が，小さな値に偏っていると思われる場合 (周辺の地点の観測値と比べて，かなり小さな値しか観測されていないことが明らかな場合) に両側有界分布を用いることは，小さ過ぎる確率水文量を推定してしまう可能性があるため，避けた方がよい．

（5）逆に，ある程度大きなデータを含んでいる場合 (近隣の観測所のデータを参考にすればある程度大きなデータが含まれているかどうか判定できる) に，両側有界分布を用いることは，確率水文量の過大評価を防ぐためには，有効であると言える．

参考文献

1) 岩井重久：Slade 型分布の非対称性の吟味及びその 2,3 の解法，土木学会論文集，第 4 号，pp. 84-104, 1949.
2) Takara, K., Takasao, T. and Tomosugi, K.: Possibility and necessity of paradigm shift in hydrologic frequency analysis, Proc. of Int'l Conf. on Water Resources and Environment Research, Kyoto, Japan, Vol. 1, pp. 435-442, 1996.
3) 水文・水資源学会編：水文・水資源ハンドブック，朝倉書店，pp. 231-234, 1997.

(2000. 10. 2 受付)