少ないデータ数のもつ水文統計量の特性と河川計画への適用

STATISTICAL CHARACTERISTICS OF HYDROLOGICAL DATA WITH SMALL SAMPLE NUMBER AND ITS APPLICATION TO RIVER PLANNING

土木工学専攻　22 号　　朱　仁斌

Renbin　ZHU

1. はじめに

洪水災害や土砂災害等の直接の要因となるのは年最大値を記録する極値降雨であると考えられる．実際，近年の河川計画における計画高水流量の算定なども結果的に既往最大降雨に基づいて行われており，河川計画の外力としても極値降雨が用いられる．現在，極値降雨の発生特性に関しては極値統計学が用いられ，確率年(年超過確率)を推定し評価している．例えば，群馬県中之条観測所における年最大 3 日累積降雨量(データ数：62)とその確率年をそれぞれ**図 1** と**図 2** に示す．確率年が 10 年以上になると実測値が無限大の母集団から得られる分布曲線から大きく外れている．実測値から推定した確率年と分布曲線から推定した確率年どっちを採用するかによって河川計画の基準が大きく変わる．そこで本研究では，確率とデータ数の関係を着目し，少ないデータ数から“真値”を推定するため，任意の極値確率分布をもつ確率紙の作成方法を提案した．

2. 確率とデータ数の関係

物理現象について，ある試行において事象が起きる確率がpであり，繰り返し行った結果，その事象が起きる比率が試行を増やすと共に，経験的確率はpに近づくことは，大数の法則として知られている．

「各面が均一な 6 面サイコロを投げて，各面が出る確率はいずれも$1/6$である」．これを実験的に各面の出る確率を確かめたい．実験はサイコロを何回も振ると，それぞれの目が出る回数を投げた回数で割れば各面の出る確率になる．それぞれの目が出る確率は$1/6$に近づいていくはずだ，しかし数回投げ続けた結果，たまに確率は$1/6$になること

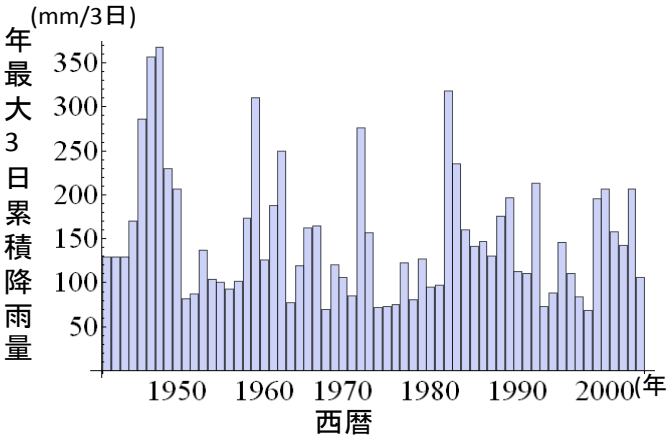

図 1　年最大 3 日累積降雨量(群馬県中之条観測所)

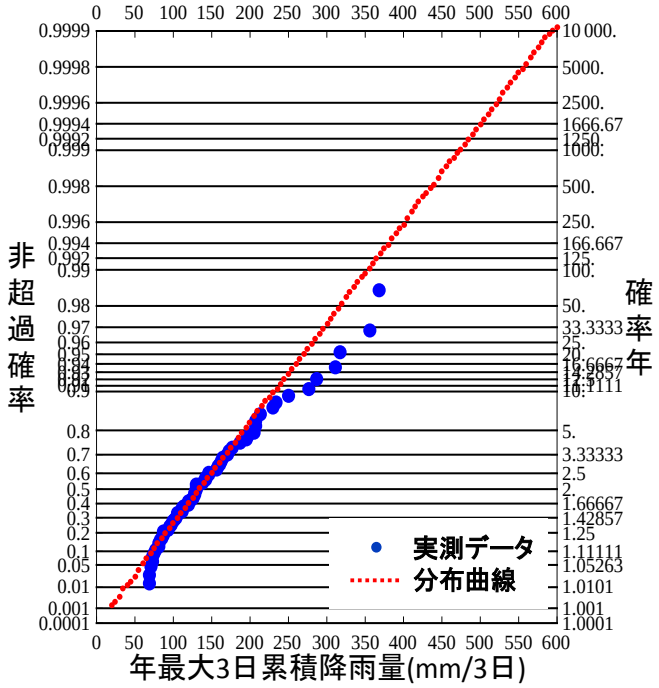

図 2　年最大 3 日累積降雨量の確率紙

があっても、$1/6$であり続けることはありえない，あくまでも$1/6$に近づくだけである。これを説明するにはチェビシェフの不等式がある．

$$\lim_{n\to\infty} P(|\bar{X}-\mu|\geq \epsilon) \leq \lim_{n\to\infty} \frac{\sigma^2}{n\varepsilon^2} = 0 \qquad (1)$$

ここで，P は確率，$\bar{X}$は算術平均，μは期待値，ϵは測定の精度，σは標準偏差であり，式(１)によって試行回数の増加につれ，理想値から離れる確率が 0 に近づくことが分かる．

例えば，**図 3** に示したようにサイコロが 500 回振る試行を 50 回繰り返した結果，相対誤差が -25%~25%になっていることが分かった．さらに，5000 回試行を 50 回繰り返した結果**図 4** に示す．図 4 より試行回数が 2000 回以上のとき、相対誤差が-10%~10%になる．また，試行回数が少ない場合(1000 回以下)相対誤差が-10%~10%より大きくなる．頼区間を計算する推定公式

$$\mu \sim \bar{x} \pm z(\alpha/2)\frac{\sigma}{\sqrt{n}} \qquad (2)$$

式(2)において，サンプルサイズnは式(2)の分母に位置しているため，nが大きいほど$\frac{\sigma}{\sqrt{n}}$の値が小さくなり，標本平均$\bar{x}$を通じて母集団μを推定する精度が高まることが分かる．

2. 少ないデータ数の推定問題

本研究では，代表的な水文量として，降雨データを用いている．近年，これまでに経験したことないような大きな雨が降っていると言われることがよくある．年最大降水量を確率紙にプロットすると，これまでほぼ一直線に並んでいるのに対して，新記録規模的な降雨データはその直線から大きく外れてプロットされていることに多々あることが分かる．分布曲線からから推測すると，確率年は何千年，何万年という値になり，異常値とみなされるが，この値は本当に異常値であるだろうか．

実際，データが蓄積されると，異常値とみなされている値が観測されるということも生じている．ある場所ではそれまで経験したことが無い大きな値の降水量が，周辺地域では頻繁に発生していることもある．異常値とみなされるの大きな原因は，データ数が少ないことだと考えられる．

ここで，データ数に着目し，ガンベル分布に従う乱数を発生させる．年最大 3 日累積降雨量の取り得る値を 20mm から 600mm までにする．その間

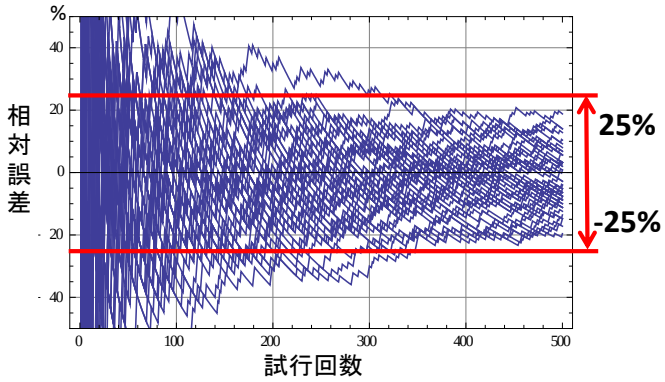

図 3　サイコロの発生確率 500 回 50 通り

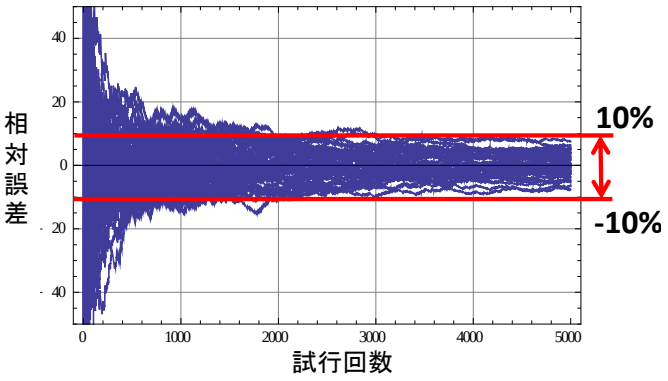

図 4　サイコロの発生確率 5000 回 50 通り

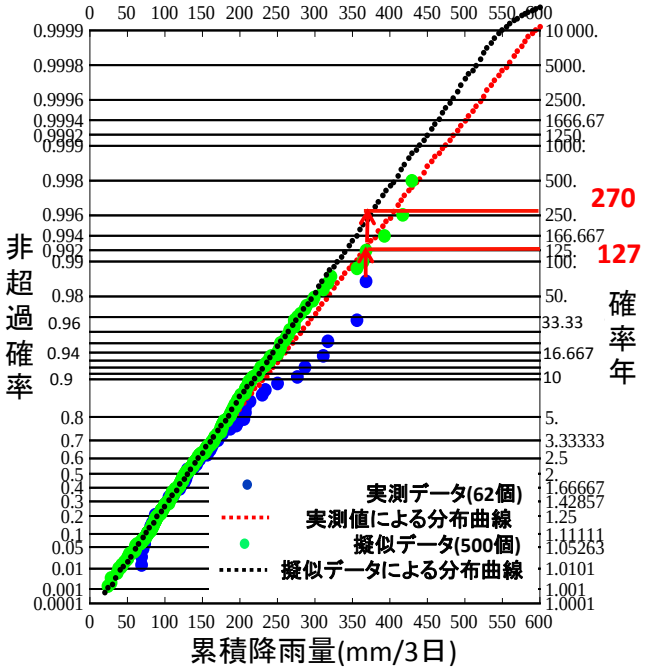

図 5　分布曲線から離れていた実測値とデータ数の関係

を 5mm ごとに確率年が与えられるように乱数を発生させる．このガンベル分布の平均値と標準偏差は実測データの平均値と標準偏差を用いている．

ガンベル分布の分布関数

$$F(x) = exp\{-exp(-y)\} \qquad (3)$$

の両辺について 2 回自然対数をとり, yについて式を整理すると，式(4)が得られる．

$$y = -lnln\left\{\frac{1}{F(x)}\right\} = \alpha(x-\mu) \qquad (4)$$

式(3)と式(4)を用いて実測値を確率紙にプロットした結果を**図 5** に青い点線で示す．**図 5** に擬似的に発生させた観測値のデータ数を 8 倍とした場合の擬似的な観測値を緑色の点線で示す，実測データから求めた分布曲線を赤い点線で，擬似データから求めた分布曲線を黒い点線で示している．

図 5 に示したように, 実測データ数が 62 個の時, 分布曲線から大きく外れた実測値がデータ数を 8 倍にした時，擬似データから求めた分布曲線に近づくことが分かった．実測データの最大降雨量の確率年が 127 年になっている．しかし，データ数を 8 倍にした時，同じ降雨量の確率年は 245 年になった．すなわち，少ないデータ数を持つ水文量で計算した確率年が過小評価していることが分かる．また，**図 6** に示すように，データ数が増すにつれ確率年一定値になることが分かった．

3. 少ないデータから“真値”の推定方法

データが蓄積された場合においても，また別の問題がある．**図 2** に示すように，実測データが一直線に並ばず, 確率年 10 年を超えると大きく屈曲することがあるという問題だ．この場合，確率年 10 年より上側のデータと下側のデータは異なる母集団に属するという疑問も出る．今までは確率紙にプロットしたデータが直線に並んでいるかどうか，あるいはデータのヒストグラフと確率密度関数か一致性を目視によって判断した．本研究では任意の極値確率分布をもつ確率紙の作成方法を示し，それを用いて少ないデータから“真値”の推定方法の検討を行う．

3. 1 任意の極値確率分布をもつ確率紙の作成方法

観測によって得られたn個のデータを大きさの順に$x_1 \le x_2 \le \cdots x_n$と並べ, 観測データは極値分布の累積分布関数$F(x)$を母集団分布とをする．縦軸として表したい非超過確率y_1，y_2，…，y_nを決める．y_nに対して$x_i = F^{-1}(y_i)$を逆関数で求める．横軸に実測データx_iを任意数aかけ$X_n = ax_i$とり，縦軸非超過確率にy_nとり，プロットして横線を引くと直線の分布曲線ができる．ここで群馬県中之条観測所における年最大 3 日累積降雨量を用いて，検討を行う．

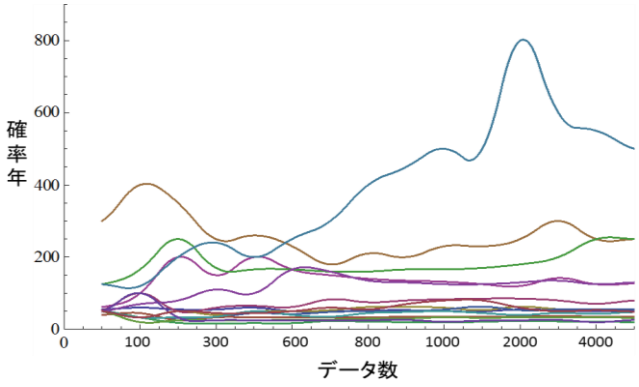

図 6　分布曲線から離れていた実測値とデータ数の関係

実測データを昇順に並び替え,そのi番目の非超過確率はプロッティング・ポジション公式

$$F_X(x_i) = \frac{i-a}{N+1-2a} \qquad (5)$$

より表される．ここで，Nはデータ数，iはデータの昇順の順位，x_iはi番目のデータの値, α は

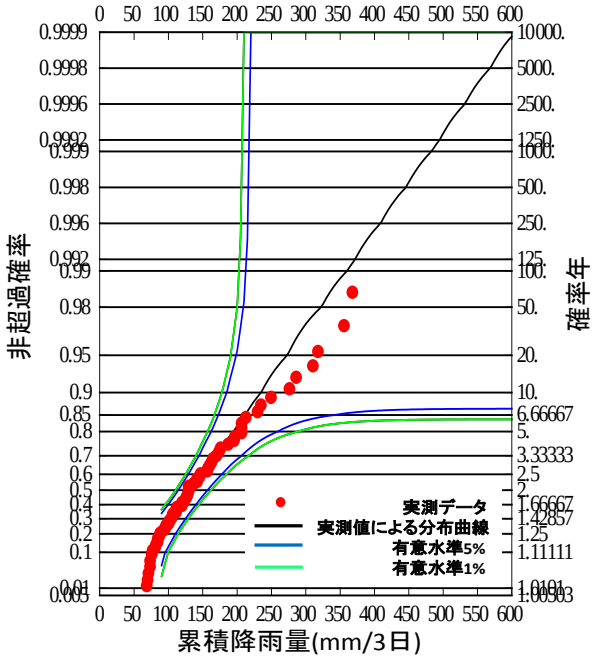

図 7　実測データ数による行った確率紙

$0 \leq \alpha \leq 1$の定数であり，**図 7** に示す点線は非超過確率と年最大 3 日累積降雨量の関係を表すものである．

3.2 任意の極値確率分布をもつ確率紙の作成方法

観測データx_1，x_2，…，x_nの累積分布関数$F(x)$を母集団分布とする．母集団からの独立標本であるかどうかを数量的に検討するにはコルモゴルフ - スミルノフ検定を用いて行う．

観測データx_1，x_2，…，x_nから作られた経験分布関数$F_n(x)$と，母集団の累積分布関数との距離を適当な指標で数量化し，その値が大きかったら，観測データは母集団からの標本ではないと推論する．

$$K_n^+ = \sqrt{n}\max_x\{F_n(x) - F(x)\} = \sqrt{n}\max_{n=1,2,\ldots,n}\left\{\frac{i}{n} - F\left(x_{(i)}\right)\right\} \quad (6)$$

$$K_n^- = \sqrt{n}\sup_x\{F(x) - F_n(x)\} = \sqrt{n}\max_{n=1,2,\ldots,n}\left\{\frac{i}{n} - F\left(x_{(i)}\right)\right\} \quad (7)$$

K_n^+観測データの従う分布が左にずれた場合に大きな値を取り，K_n^-は右にずれた大きな値を取ることを表す．**図 7** に示したように青い線は有意水準 5% ,黒い線は有意水準 1%で検討した結果である．**図 7** に示しているように確率年 10 年より上側のデータと下側のデータはコルモゴルフ - スミルノフ検定により、同じ母集団に属することが分かった．

図 8 は 3 節と同様に擬似的な観測値を 10, 00 個発生させたとき許容範囲を示している．実測の年最大 3 日累積降雨量は 10, 00 個観測値発生させたときの許容範囲に入ることが分かった．

4.まとめ

1) データ数が少ない時，分布曲線から大きく外れる実測値がデータの蓄積と共に分布曲線に近づく．
2) 少ないデータ数を持つ水文量で推定した確率年は過小評価の可能性がある．
3) 観測値の年最大 3 日累積降雨量は多くの場合，確率上で一直線に並ばず，確率年 10 年ほどを境に大きく外れる場合がある，これに対してコルモゴルフ - スミルノフ検定を導入した．
4) 多くの場合異常値はコルモゴルフ - スミルノフ検定によって有意であることが分かる．
5) 任意の極値確率分布をもつ確率紙の作成方法を提案した．

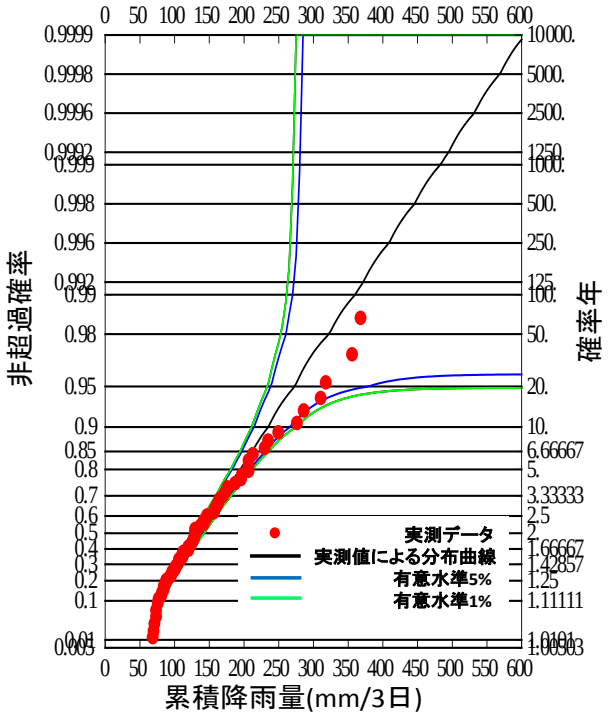

図 8　擬似データ数(1000 個)の確率紙

5.参考資料

1) 春日屋伸昌：水文統計学概説，鹿島出版社 pp. 111-124.

2) 渡辺武彦，松浦正典，深和岳人，山田正：新記録の出現理論に基づく大雨の発生頻度に関する研究，土木学会第 47 回年次学術講演会講演要綱集，1992.

3) 上田拓治：44 の例題で学ぶ統計的検定と推定の解き方．オーム社開発局.

4) 近森邦英：年降水量の統計的諸特性について，農土誌，71(2) ，pp. 125-130，2003.

5) 上田年比古、河村明:確率分布の適合度の図式判定法について，土木学会論文集，Vol. 375，NoⅡ-3(1985).

6) 伏見正則，逆瀬川浩孝：R で学ぶ統計解析，朝倉出版，pp，151-164.

7) 日野幹雄：スペクトル解析，朝倉書店，1977