

# 水文統計に用いられる分布の 右裾点の推定に関する一考察

## A CONSIDERATION ON RIGHT TAIL POINT ESTIMATION FOR THE DISTRIBUTIONS IN HYDROLOGICAL STATISTICS

廣瀬 英雄<sup>1</sup>

Hideo HIROSE

<sup>1</sup>正会員 工博 九州工業大学情報工学部教授 (〒820-8502 福岡県飯塚市大字川津680-4)

This paper describes an upper tail point estimation method using the incomplete data by the maximum likelihood estimation method. In some data cases, it happens to occur that the estimates for upper tail points such as 100-year return period can be smaller than the observed maximum annual flood datum, even if the number of observed data is smaller than 100; this result of analysis cannot be acceptable from the practitioners' viewpoint. Because, adopting such estimates in hydrologic frequency analysis may cause dangerous designs and plans for flood prevention. By weighting much to data in upper tail probability portion, such a phenomenon may be reduced. Peaks over threshold method (POT) and probability weighted moment method (PWM) are included in such weighting methods. This paper deals with a weighting method by using incomplete data. An example for St. Mary's River data shows the superior points in such a treatment.

**Key Words:** Upper tail point estimation, Gamma distribution, Incomplete data, Extended Kolmogorov-Smirnov statistic, Parametric bootstrap, Delta method

### 1. まえがき

洪水災害に耐える水工構造物の設計や計画は、対象とする事象に関する変量の極値を想定して行う必要があり、水文統計で用いられる確率分布の推定に関しては、分布の右裾点（確率水文量）の推定値とその推定精度が重要となる。また、洪水災害を予防する観点からは、推定は安全側に行うことが望ましい。この二つの観点から右裾点の推定に関して考察を行った。

従来、水文統計学では、実在する河川の観測データに対して、種々の確率分布（対数正規、Gumbel、ガンマ、対数ガンマ、一般化極値（GEV）分布など）をあてはめ、適当と思われる分布を何らかの方法（例えば、尤度、最小2乗法、Kolmogorov-Smirnov検定

（K-S検定）、SLSC（standard least-squares criterion）<sup>1)</sup>など）で特定して、確率水文量の推定を行ってきた。このときのデータとしては、ある地点での年間流水最大量がよく用いられてきた<sup>2)</sup>。

最近では推定精度を上げるため、過去に記録された洪水の記録を付加することも行われている<sup>3,4)</sup>。これは過去に大きな洪水が起こった情報を何とかデータ解析に組み込んで解析の精度を上げようとする試みであり、分布の右裾の情報を付加しようとしていることになる。

また、洪水とはみなせないような小さな量を持つデータの影響を少なくするため、毎年の最大流量ではなくあるしきい値越えるデータ（POT: peaks-over-threshold）を採用して確率分布のパラメータの推定を行ったり、PDS（partial duration series）

データを用いることも始められており、これには一般化パレート分布へのあてはめがよく行われている<sup>5)</sup>。このことは右裾に近いデータをより多くとることで右裾点の推定精度を上げようとの試みの一つと考えられる。

さて、文献によると<sup>6),7)</sup>、Gumbel分布については最尤推定法が、G E V 分布についてはP W M 法<sup>7)</sup> (probability weighted moment法) が、確率水文学量の推定精度の観点からはよいとされている。GEV分布においても、分布のパラメータ推定に関しては最尤推定法の優位性が示されているものの、ここで注目すべきことは、PWM法もまた分布のデータに関して重み付けを行うことで、推定したい統計量の推定精度を上げることに寄与している点である。勿論、推定そのものの安定性に関して最尤推定法より優位であることもPWM法の利用価値として挙げられる。このことは、POTデータの採用と同じく、分布の右裾点のデータの強調法の一つと考えられる。

本稿では、このような観点に基づき、分布の右裾を強調する基本的な方法として、右側分布には完全データを使い、その他には不完全データを使った場合の、分布の右裾点の推定値の妥当性と推定精度について考察した結果を述べる。

2. 完全データのみの場合の問題点

毎年最大流量のデータが過去何年かにわたって蓄積されているとする。例えば、ここではSt Mary'sの過去72年間のデータ<sup>2)</sup>を考える。推定したい確率水文学量を100年に一度の年最大流量( $x_{99}$ と表す)とする。このとき、最尤推定法で推定された $x_{99}$ の値とその90%信頼区間は、表-1のとおりとなっている。表で\*印はKite(1988)の再掲であるが、\*\*印は著者が計算したもので、信頼区間はbootstrap手法<sup>8)</sup> (parametric bootstrap) を用いて求めた値である。Fisherの情報行列をもとに、デルタ法<sup>9)</sup>から直接解析的に近似90%信頼区間を求めることも可能であるが、対数正規分布に関しては少し異なったものの、ガンマ分布に関しては両者はほとんど同じ値を示したので、ここではbootstrapの結果のみを示している。

元のデータの最大値は1971年の8月に観測された974m<sup>3</sup>/sである。表-1によると、3パラメータ分布の中で対数尤度の最も大きくなるガンマ分布はこの値を90%信頼区間の中に含んでいない。年最大流量を確率変数と仮定し、観測されたデータの背後には真の分布があってパラメータ推定はその真の分布に近いものを推定するという立場に立てば、974という数は

表-1  $x_{99}$  の最尤推定値とその信頼度 (完全データ)

|        | 推定値* | 推定誤差* | 90%信頼区間** | 対数尤度**   |
|--------|------|-------|-----------|----------|
| 対数正規   | 882  | 93.1  | -         | -451.363 |
| Gumbel | 848  | 51.6  | -         | -451.597 |
| ガンマ    | 855  | 68.4  | 734, 973  | -451.223 |
| 対数ガンマ  | 879  | 101.0 | 765, 998  | -451.407 |

\*: Kite (1988); \*\*: 廣瀬; 推定はともに最尤推定法

表-2 974 m<sup>3</sup>/sに対する再現期間 (完全データ)

|        | 再現期間 |
|--------|------|
| 対数正規   | 221  |
| Gumbel | 319  |
| ガンマ    | 302  |
| 対数ガンマ  | 213  |

たまたま出現した変数であり、90%信頼区間に入らないことも不合理ではない、という見方もできる。しかしながら、真の分布というものはもともと存在せず、それは便宜的に使うものであって、実際にあるのはその河川に特有なデータのみであるという立場に立てば、推定された90%信頼区間に観測された最大値 (最も重要なデータ) が含まれていないというのは問題である。河川計画で100年に一度の災害を想定しようとしているのに、72年の間に実際に観測された値が軽視されているのは (極値を求めようとする立場からは) 危険側に推定されていることになり不合理なことである。ただし、通常、水文統計での信頼区間には95%信頼区間が使われるが、ここでは $x_{99}$ の90%程度の信頼区間でも不都合が起ることを示すために90%信頼区間を用いている。

また、流量974m<sup>3</sup>/sに対応する再現期間を上 の 4 つの分布に対して求めた結果を表-2に示す。72年間の間に実際に観測された流量に対する再現期間はGumbel, ガンマ分布で特に大きく、それぞれ319, 302年となっている。比較的裾の長い対数正規分布でさえも221年とかなり長い。このことから推定は危険側に行われていることが分かる。

これらのことは、72年間の全てのデータを同じ重みで取り扱ったことにも起因すると考えられる。一方、分布の右裾を重要視すれば、右裾の推定に関しては大きく危険側に推定を行うことは少ないと思われる。そこで、ここでは最も基本的な方法の一つと考えられる、分布の右裾を完全データとして、その他を不

完全データとして扱った場合の推定について考察することにする。

974m<sup>3</sup>/sという流量がoutlierであれば、このような検討は無意味であるが、文献(10)に記載されている異常値検定の結果からは974m<sup>3</sup>/sは棄却されないことを付記する。

### 3. 完全データと不完全データの併用

観測されたデータ（ここでは毎年最大量でも非毎年データでもよいとする）を大きい順に並べ直したものを、 $x_1, x_2, \dots, x_{k-1}, x_k, x_{k+1}, \dots, x_n$ とする。観測データの大きい $k$ 個の値（ $x_1, x_2, \dots, x_k$ ）に対しては完全データとして使い、それ以外のデータ（ $x_{k+1}, \dots, x_n$ ）に対しては不完全データとして使って、尤度関数を次のように表す。

$$L_k = (F(x_k))^{n-k} \prod_{i=1}^k f(x_i) \tag{1}$$

ここに、 $f(x)$ 、 $F(x)$ はそれぞれ確率分布の密度関数と分布関数である。観測データの大きいものはそのまま情報を残し、小さい方に関してはデータは $x_k$ 以下であるという情報しか用いていない。このように尤度関数を作ると、情報量は減るので分布全体としては推定精度は一般に悪くなる。しかしながら、後で示すように、分布の右裾点に関してより合理的な推定値が得られる場合がある。パラメータは最尤推定法を用いて推定する。

### 4. St. Mary's Riverデータによる解析例

ここでは、St Mary's Riverデータ<sup>2)</sup>に関して、尤度の最も大きかった分布（ガンマ分布）がこのデータには適合していると仮定して、ガンマ分布を用いて検討を行ってみた。ガンマ分布の密度関数を、

$$f(x;\alpha,\beta,\gamma) = \frac{1}{\gamma\Gamma(\beta)} \left(\frac{x-\alpha}{\gamma}\right)^{\beta-1} \exp\left[-\left(\frac{x-\alpha}{\gamma}\right)\right] \tag{2}$$

$(x \geq \alpha, \beta > 0, \gamma > 0)$

で定義する。(1)式での完全データの個数 $k$ をここでは、全データ数の半数36、その2/3の24、その半数の12、その2/3の8にとり、それらに対応する推定値と信頼度について調べてみた。

最尤推定法を用いて求めた(1)式に対する $x_{.99}$ の推定値とその90%、95%信頼区間を表-3に示す。

表-3  $x_{.99}$ の推定値とその信頼度（不完全データ）

| $k$ | $x_{.99}$ の推定値 | $x_{.99}$ の90%信頼区間 | $x_{.99}$ の95%信頼区間 |
|-----|----------------|--------------------|--------------------|
| 8   | 915            | 555, 1275          | 486, 1343          |
| 12  | 896            | 604, 1187          | 548, 1243          |
| 24  | 854            | 703, 1006          | 674, 1035          |
| 36  | 853            | 721, 995           | 684, 1022          |

(推定法は最尤推定法)

表-1によると、完全データのときのガンマ分布の $x_{.99}$ の推定値は855となり観測されたデータの最大値に比べてかなり小さいが、不完全データ併用した $k=8$ のときそれは915と大きくなり安全側に推定されている。ただし、推定値の信頼区間は（情報量が少ないため）大きくなる。

経験分布と推定された分布関数とがどの程度適合しているかについては、注目すべき分布の右裾で測るのが妥当であるので、ここでは、 $x_k$ から上側について分布の距離の最大値 $De_k^k, De_k^0, Dc_k^k, Dc_k^0$ を調べた。ここに、

$$De_k^k = \max_{x_1 \geq x > x_k} |F(x)^k - F_e(x)| \tag{3}$$

$$De_k^0 = \max_{x_1 \geq x > x_k} |F(x)^0 - F_e(x)| \tag{4}$$

$$Dc_k^k = \max_{x_1 \geq x > x_k} |F(x)^k - F_c(x)| \tag{5}$$

$$Dc_k^0 = \max_{x_1 \geq x > x_k} |F(x)^0 - F_c(x)| \tag{6}$$

で定義する。また、 $F(x)^k$ は不完全データを含む(1)式から推定された分布関数、 $F(x)^0$ は完全データのみを用いて推定された分布関数、 $F_e(x)$ 、 $F_c(x)$ は分布関数、

$$F(x) = \begin{cases} 0, & (x < x_n) \\ \frac{i-a}{n+b}, & (x_{i+1} \leq x < x_i, i = 1, \dots, n-1) \\ 1, & (x_1 \leq x) \end{cases} \tag{7}$$

で、

$$a = 0, \quad b = 0 \quad (F(x) = F_e(x))$$

$$a = 0.4, \quad b = 0.2 \quad (F(x) = F_c(x))$$

とするとき、 $F_e(x)$ は経験分布関数、 $F_c(x)$ はCunanne<sup>11)</sup>のプロットイングポジションを用いた分布関数となる。 $De_k^k, De_k^0$ はKolmogorov-Smirnov統計量の拡張になっているので、これをここでは、拡張K-S統計量と呼ぶ。

表-4  $De_k^k, De_k^0, Dc_k^k, Dc_k^0$

| $k$ | $De_k^k$ | $De_k^0$ | $Dc_k^k$ | $Dc_k^0$ |
|-----|----------|----------|----------|----------|
| 8   | 0.0157   | 0.0457   | 0.0239   | 0.0375   |
| 12  | 0.0272   | 0.0457   | 0.0351   | 0.0375   |
| 24  | 0.0481   | 0.0489   | 0.0556   | 0.0565   |
| 36  | 0.0497   | 0.0502   | 0.0572   | 0.0565   |

表-4に、 $k=8$ のときの $De_k^k, Dc_k^k$ の値、および $De_k^0, Dc_k^0$ の値を示す。表から特に $De_k^k$ が優れていることが分かる。この結果を表-2の推定値と併せて考えると、特に安全側に設計したい場合に有効であると思われる。

図1-4には、 $k=8$ のときに対して、推定された分布、および $x_p$  ( $x_p$ 以上の洪水の発生する確率を $p$ とする)の90%, 95%信頼区間を、経験分布関数およびCunanneのプロットティングポジションを用いた分布関数と併記して示した。分布の右裾において推定された分布は $k=8$ のときは経験分布の近くを通っている様子が分かる。 $1/72=0.0139$ であることを考えると、 $k=8$ のときでの $De_k^k, De_k^0$ の値から、推定された分布がどの程度経験分布に近いかが分かる。Cunanneのプロットティングポジションに対応する $Dc_k^k$ は $De_k^k$ に比べて悪くなっている。

## 5. まとめ

水文統計に用いられる確率分布の推定にあたっては、分布の右裾の推定値とその信頼度が重要になる。従って、右裾に近い部分のデータに重みを付け、その他の部分を軽視することで、分布の右裾点の推定精度を上げることができる可能性がある。Historical dataの採用、POTによるデータの収集、PWMによる確率分布への重み付けを行った後でモーメント法による推定を行う方法なども、この類と考えられる。本稿では、この重み付けを、最も素直な形の一つである不完全データを用いて検討を行った。ここで提案している拡張K-S統計量の基準を用いてST. Mary's Riverの推定分布の比較を行った結果、分布の右裾点の推定に関しては経験分布を忠実に反映することが分かった。また、推定値の信頼度については、失われた情報の分だけ悪くなっているが、これは安全側の設計に寄与すると考えることもできる。他の方法と併せて欠点を補えばよいであろう。本研究での最も重要な成果は、従来の方法では100年を下回る年最大流量データの最大値が100年確率水文量の推定値よりも大きく推定されることが起こっていたが、こ

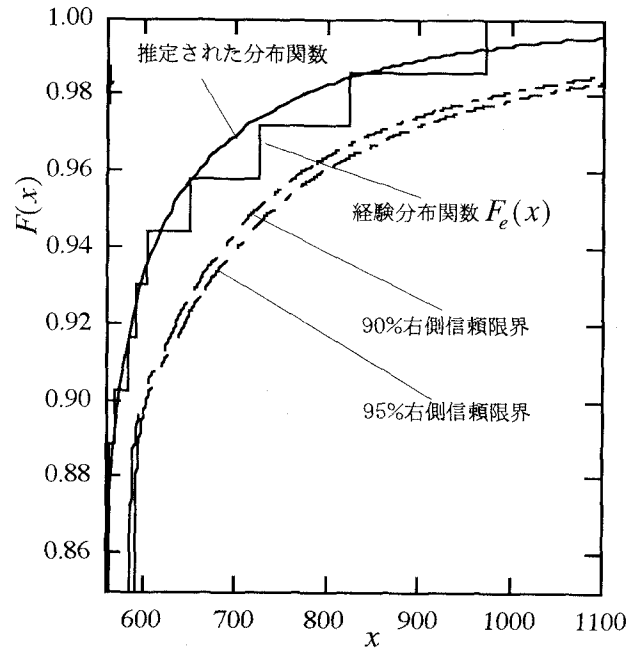


図-1 推定された分布、および $x_p$ の90, 95%信頼区間と $F_e(x)$ との比較 ( $k=8$ )

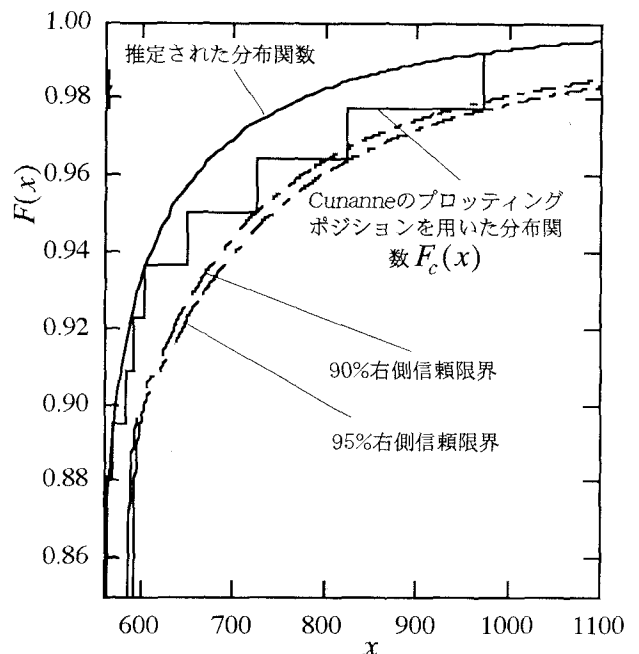


図-2 推定された分布、および $x_p$ の90, 95%信頼区間と $F_c(x)$ との比較 ( $k=8$ )

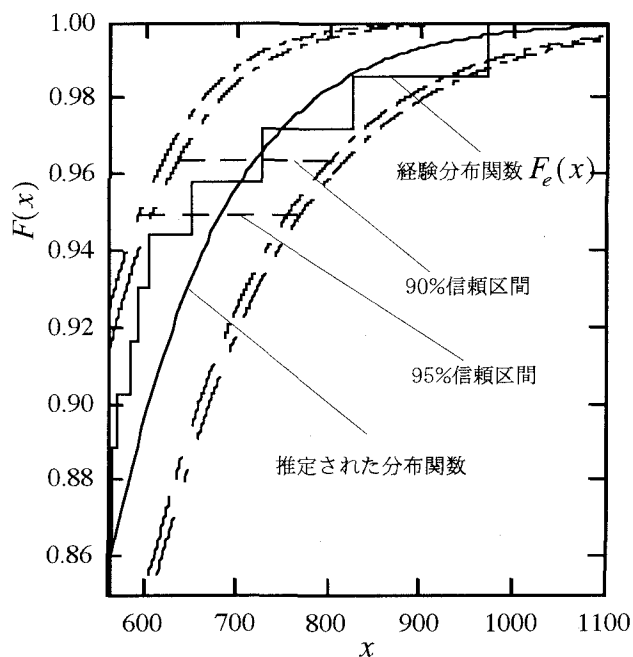


図-3 推定された分布, および  $x_p$  の90, 95%信頼区間と  $F_e(x)$  との比較 ( $k = 0$ )

この方法では最大値を反映した形で推定されていることである。

一般的な確率分布に対して適用できる統計的な性質を特定の河川の極値解析に適用する場合には、その河川の特徴を重要視する必要がある。特に、過去にわたっての最大流量があたかもはずれ値 (outlier) のように見えるときにはその値に特別の注意を払う必要がある。分布の右裾の部分強調する解析はこのことに配慮したものである。最近、POTなどの手法が用いられ始めているが、POTの情報に更に不完全データを付加することも考えられる。その場合、更に推定精度がよくなると考えられる。

ここで解析したデータの種類と検討を行った確率分布は少ないが、上に述べた傾向は他の分布にも一般的に類推できるものである。今後、他の分布や他の河川データについても検討を加えていきたい。

なお、分布の右裾にあたる点の数  $k$  をどの程度にするのが適当であるかの検討は今回行っていない。今後の課題である。

謝辞：京都大学防災研究所の宝教授には文献を紹介していただいた。ここに謝意を表する。

#### 参考文献

- 1) 高棹, 宝, 清水：琵琶湖流域水文データの基礎的分析, 京都大学防災研究所年報, 29, B-2, pp.157-171 (1986)
- 2) Kite, G.W. : *Frequency and Risk Analysis in Hydrology*, Water Resources Publications, Fort Collins, Colorado, USA (1988)
- 3) Hosking, J.R.M., Wallis : The value of historical data in flood frequency analysis, *Water Resources Research*, 22, pp.1606-1612 (1986)
- 4) Stedinger, J.R, Cohn T.A. : Flood frequency analysis with historical and paleoflood information, *Water Resources Research*, 22, pp.785-793 (1986)
- 5) 宝：水文頻度解析の進歩と将来展望, 水文・水資源学会誌, 11, pp.740-756 (1998)
- 6) 宝, 高棹：水文頻度解析モデルの母数推定法の比較評価 水工学論文集, 土木学会, 34, pp.7-12 (1990)
- 7) Hosking, J.R.M., Wallis, J.R., Wood, E.F. : Estimation of the generalized extreme-value distribution by the method of probability weighted moments, *Technometrics*, 27, pp.251-261 (1985)
- 8) Efron, B. : Bootstrap methods: another look at the jackknife, *Ann. Statist.*, 7, 1-26 (1979)
- 9) Rice, J.A. : *Mathematical Statistics and Data Analysis*, Wadsworth and Brooks/Cole (1988)
- 10) 水文・水資源ハンドブック, 水文・水資源学会編 (朝倉) (1998)
- 11) Cunanne, C. : Unbiased plotting position - a review, *Journal of Hydrology* 37, pp.205-222 (1978)

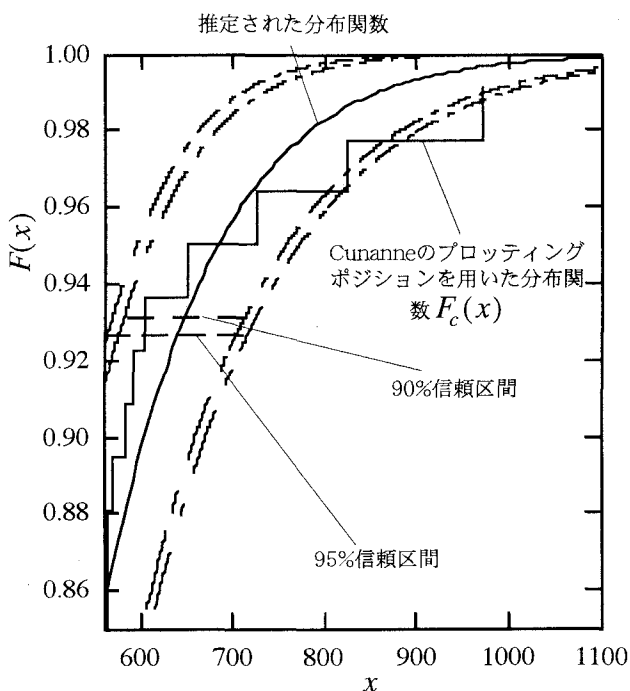


図-4 推定された分布, および  $x_p$  の90, 95%信頼区間と  $F_c(x)$  との比較 ( $k = 0$ )

(1999.9.30受付)