

観測誤差を含む極値データから推定した再現確率値の信頼区間について

泉 宮 尊 司*・吉 田 淳**

本研究では、極値データに観測誤差等が含まれる場合における推定された再現確率値の信頼性について、理論的解析とモンテカルロシミュレーション法により検討したものである。まず、誤差が含まれる場合の確率分布関数について理論的検討がなされている。極値分布関数として、FT-I 型分布および Weibull 分布を採用し、それらの極値分布に従うデータに Gauss 分布に従う誤差を加えたデータを用いて再現確率値を従来の最小二乗法と重み付最小二乗法により算定している。その結果、誤差が含まれる場合には、数%程度の正のバイアスが生じることや再現確率値の標準偏差が誤差の分散に比例して増加することなどが明らかとなった。

1. 緒 言

ある有限数の極値データから推定された再現確率値は、確定的な値をとるのではなく標本毎に異なり、ある信頼区間をもつ確率変数となる。この信頼区間は、構造物の信頼性設計を行なう際に、構造物の破壊確率を左右する重要な要素となる(高山, 2000)。観測された波浪や波浪推算によって得られた極値データには、観測誤差や推算誤差が含まれている場合が一般的であるので、その誤差の算定された再現確率値やその分布への影響を調べることは重要である。また、極値データに誤差が含まれることによって、そのデータはもはやもとの母分布関数とは異なった分布となるため、分布関数の適合度評価にも影響を及ぼすことが考えられる。しかしながら、極値データに誤差が含まれる場合について、誤差の影響を理論的および数値実験的に詳しく調べた研究は殆どなく、その影響について未解明なところが多い。

そこで本研究では、極値データに観測等の誤差が含まれる場合の影響を明らかにするために、まず極値の確率密度関数が誤差によってどのような確率密度関数に変化するかを理論的に調べる。さらに、モンテカルロシミュレーション法により極値データおよび誤差データを作成し、真の再現確率値との差であるバイアスや信頼区間の変化について調べることを本研究の目的とする。さらに、誤差の分散を与えた時に推定される再現確率値の標準偏差を推定する関係式を導き、その妥当性についても検討する。

2. 理論的解析法

(1) 誤差を含むデータの確率密度関数

波浪観測データには、観測機器による誤差や観測地点と極値統計解析が必要な地点とが異なることによる誤差等、様々な要因の誤差が含まれている。ここでは、まずそれらの誤差が含まれている場合の確率密度関数への影

響を調べることにする。いま、真の極値データを x 、誤差を ε とするとき、それぞれの確率密度関数を $f(x)$ および $g(x)$ で表すことにする。誤差が含まれる極値データ ξ は、

$$\xi = x + \varepsilon \dots\dots\dots (1)$$

と表される。ここで、極値データ x と誤差 ε は互いに独立であるとすると、 ξ の特性関数 $\phi(s)$ は、

$$\begin{aligned} \phi(s) &= E(e^{is \cdot \xi}) = \int_{-\infty}^{\infty} e^{is \cdot \xi} h(\xi) d\xi \\ &= \int_{-\infty}^{\infty} e^{is \cdot x} f(x) dx \cdot \int_{-\infty}^{\infty} e^{is \cdot \varepsilon} g(\varepsilon) d\varepsilon \\ &= \phi(s) \chi(s) \dots\dots\dots (2) \end{aligned}$$

となる。ここに、 $\phi(s)$ および $\chi(s)$ はそれぞれ極値 x および誤差 ε の特性関数である。上式の逆変換をとることにより、次のような確率密度関数の関係式が得られる。

$$\begin{aligned} h(x) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-is \cdot x} \phi(s) ds \\ &= \int_{-\infty}^{\infty} f(x-y) g(y) dy \dots\dots\dots (3) \end{aligned}$$

ここに、 $h(x)$ は誤差を含むデータの確率密度関数である。上式において、誤差の確率密度関数 $g(x)$ がデルタ関数である場合のみ、 $f(x)$ と $h(x)$ が等しくなるが、それ以外は元の確率密度関数とは異なることになる。誤差の確率密度関数として、平均値が 0 の Gauss 分布を想定すると、その分布のパラメータは分散 σ^2 のみとなる。誤差はそれほど大きくなく、極値データの数%程度以下であるとすると、式(3)の y に関する積分は、大きい値の y に関して $g(y)$ の値は殆ど 0 となる。したがって、 y の絶対値が小さい範囲の積分を考えればよいので、 $f(x-y)$ を x の回りに Taylor 展開すると、

$$f(x-y) = f(x) - yf'(x) + \frac{y^2}{2}f''(x) + \dots\dots\dots (4)$$

となる。上式を式(3)に代入すると、

* 正会員 工博 新潟大学教授 工学部建設学科
** 新潟トヨタ自動車(株)

$$\begin{aligned}
 h(x) &= f(x) - \int_{-\infty}^{\infty} y f'(x) g(y) dy \\
 &\quad + \int_{-\infty}^{\infty} \frac{y^2}{2} f''(x) g(y) dy + \cdots \\
 &= f(x) - m_1 f'(x) + \frac{m_2}{2} f''(x) + \cdots \quad \cdots (5)
 \end{aligned}$$

を得る。ここに、 m_1 および m_2 は $g(y)$ の 1 次および 2 次モーメントである。誤差の平均値が 0 であるので、 $m_1 = 0$ となり、誤差を含むデータの確率密度関数 $h(x)$ およびその確率分布関数 $H(x)$ は、近似的に次式で表されることになる。

$$h(x) = f(x) + \frac{\sigma^2}{2} f''(x) \quad \cdots \cdots \cdots (6)$$

$$H(x) = F(x) + \frac{\sigma^2}{2} f'(x) \quad \cdots \cdots \cdots (7)$$

式 (6) および式 (7) より、極値データに誤差が含まれる場合には、もはや元の極値分布とは異なり、その確率密度関数の曲率と勾配および誤差の分散に主に依存して差異が生じることになる。

ここで、誤差が含まれるデータの分散 σ_n^2 を求めてみることにする。平均値 μ の値は誤差によって期待値は変化しないので、真の極値分布の分散を σ_m^2 とすると、

$$\begin{aligned}
 \sigma_n^2 &= \int_{-\infty}^{\infty} (x - \mu)^2 h(x) dx \\
 &= \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx + \frac{\sigma^2}{2} \int_{-\infty}^{\infty} (x - \mu)^2 f''(x) dx \\
 &= \sigma_m^2 + \sigma^2 = \sigma_m^2 \left(1 + \frac{\sigma^2}{\sigma_m^2} \right) \quad \cdots \cdots \cdots (8)
 \end{aligned}$$

となり、誤差の分散だけ大きくなることが分る。

(2) 再現確率値の標準偏差の評価

母数推定法として、最小二乗法や泉宮・岡本 (2000) が開発した重み付最小二乗法を採用するとき、推定される尺度母数 $\hat{A}$ および位置母数 $\hat{B}$ は、順序統計量 x_i の線形結合の和で表される。

$$\hat{A} = \sum_{i=1}^N g_i x_{i1}, \quad \hat{B} = \sum_{i=1}^N h_i x_{i1} \quad \cdots \cdots \cdots (9)$$

ここに、 g_i および h_i は標本数 N と順序 i によって決められる定数で、最小二乗法の場合には、泉宮・斉藤 (1998) により与えられている。また、重み付最小二乗法の場合には、 $\hat{\theta} = (\hat{A}, \hat{B})^t$ とすると、泉宮・岡本 (2000) が示している以下の関係式を用いて容易に求められる。

$$\hat{\theta} = (M^t \Sigma_N^{-1} M)^{-1} M^t \Sigma_N^{-1} x \quad \cdots \cdots \cdots (10)$$

ここに、 $x = (x_1, \dots, x_i, \dots, x_N)^t$ で順序統計量ベクトルを表し、 M は基準化変量 y_i で表される $2 \times N$ の行列、 Σ_N は各要素が母分布関数の分散 σ^2 で基準化された分散共分散行列である。

推定される母数が式 (9) で表されるとき、 R 年再現確

率値の分散は次式で評価される (泉宮・斉藤, 1998)。

$$\begin{aligned}
 \sigma^2(\hat{x}_R) &= E[(\hat{x}_R - x_R)^2] \\
 &= \sum_{j=1}^N \sum_{i=1}^N (g_i y_{Rj} + h_i)(g_j y_{Rj} + h_j) \\
 &\quad \times E[(x_i - E(x_i))(x_j - E(x_j))] \\
 &= \sum_{j=1}^N \sum_{i=1}^N (g_i y_{Rj} + h_i)(g_j y_{Rj} + h_j) C_{ij} \quad \cdots \cdots \cdots (11)
 \end{aligned}$$

ここに、 C_{ij} は相関行列で泉宮・斉藤 (1998) および泉宮 (2000) に示されている結果を用いることができる。

3. 数値シミュレーション法

本研究では、誤差を含む極値データの再現確率値等への影響を調べるために、モンテカルロ法を用いている。まず、極値分布関数として、FT-I 型分布および Weibull 分布 (形状母数 $k=1.0$, $k=2.0$) の 3 種類を選定し、それぞれの極値データを一様乱数により発生させている。FT-I 型分布については、Case 1: 尺度母数 $A=1.1$, 位置母数 $B=4.7$, $x_{50}=8.99$ m, Weibull 分布については、Case 2: $k=1.0$, 尺度母数 $A=1.1$, 位置母数 $B=4.7$, $x_{50}=9.00$ m, および Case 3: $k=2.0$, 尺度母数 $A=1.1$, 位置母数 $B=4.7$, $x_{50}=6.88$ m としている。Case 1 と Case 2 は、50 年確率値がほぼ同じになる FT-I 型分布と Weibull 分布を比較するためのものであり、Case 2 と Case 3 は分布形状と再現確率値の差異の影響をみるために設定されたものである。

誤差データについては、平均値 0 の Gauss 分布に従うものとし、 $\sigma=0.1$ m および $\sigma=0.3$ m の 2 種類の誤差を極値データに加えて、数値実験を行っている。データ数 N については、誤差の影響が現れやすい $N=10, 20, 30, 40, 50$ とし、それぞれ 2000 資料作成している。母数推定法には、従来の最小二乗法 (合田, 1988, 1990) および重み付最小二乗法 (泉宮・岡本, 2000) を用いて R 年再現確率値やそのバイアスおよび標準偏差等を算定した。

4. 数値シミュレーション結果と考察

図-1 は、標本数 $N=50$ の時の誤差が含まれている場合とそうでない場合の推定された再現確率値の変化を示したものである。この図において、LSM および WLSM はそれぞれ最小二乗法 (合田, 1988, 1990) および重み付最小二乗法 (泉宮・岡本, 2000) による推定結果を表し、+ 誤差と書かれている記号は誤差が混入された極値データを用いて再現確率値を算定したものである。また、各年の再現確率値の真値を白丸印で示している。最小二乗法 (LSM) による推定値は、誤差がない場合には真値にほぼ等しい再現確率値を与えるが、誤差の分散が大きくなるほど、推定される再現確率値がやや大きくなる傾向にあることが明確に読み取れる。それに対して、重み付

最小二乗法 (WLSM) による推定結果は、誤差が混入しても推定誤差はかなり小さいことが分る。

図-2 および図-3 は、それぞれ標本数 $N=50$ の場合の Weibull 分布の形状母数が、 $k=1.0$ および $k=2.0$ の時の推定された再現確率値を示したものである。最小二乗法 (LSM) および重み付最小二乗法 (WLSM) とともに、誤差が含まれていない場合には真値とほぼ一致しているが、誤差の分散が大きくなるに従って、それらの差異が大きくなる傾向にあることが分る。同一の誤差に対しては、重み付最小二乗法 (WLSM) の方が最小二乗法 (LSM) よりも多少推定誤差が小さい傾向にあることが分る。

図-4 から図-6 は、それぞれ FT-I 型分布、Weibull 分布 ($k=1.0$) および Weibull 分布 ($k=2.0$) の極値データに誤差が含まれる場合の 50 年再現確率値のバイアスを標本数 N に対して示したものである。FT-I 型分布の場合、標本数 N の値が増すにつれて、バイアスは小さくなる傾向にはあるが、その減少度合いは小さく、特に最小

二乗法 (LSM) による推定結果の方がそのような傾向にあると言える。

Weibull 分布に関しては、標本数が $N=10$ から $N=20$ までの小さい場合には、変動が大きく明確なことは言えないが、標本数が $N=30$ 以上であればバイアスは標本数が増えるに従って小さくなる傾向にはあるが、FT-I 型分布と同様にその減少度合いは小さい。誤差の標準偏差が 0.1 m と小さい場合には、最小二乗法 (LSM) よりも重み付最小二乗法 (WLSM) による推定値の方がバイアスが小さいと言えるが、誤差の標準偏差が 0.3 m とやや大きい場合には、両者にそれほど差はなく、いずれも $2\% \sim 6\%$ 程度の正のバイアスを示していることが分る。

図-7 から図-9 は、それぞれ FT-I 型分布、Weibull 分布 ($k=1.0$) および Weibull 分布 ($k=2.0$) の極値データに誤差が含まれる場合の推定された 50 年再現確率値の標準偏差を標本数 N に対して示したものである。これらの図には、最小二乗法 (LSM) による標準偏差の理論

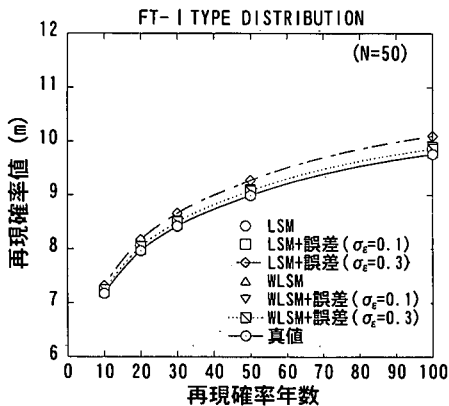


図-1 推定された再現確率値の比較 (FT-I 型分布)

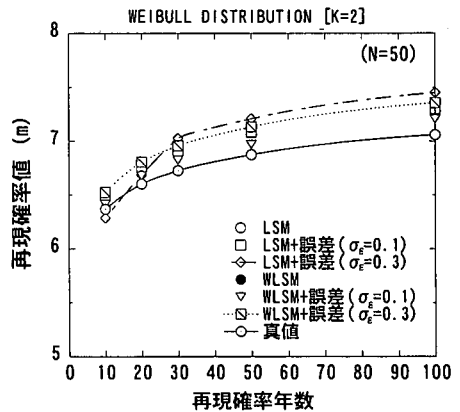


図-3 推定された再現確率値の比較 (Weibull 分布, $k=2.0$)

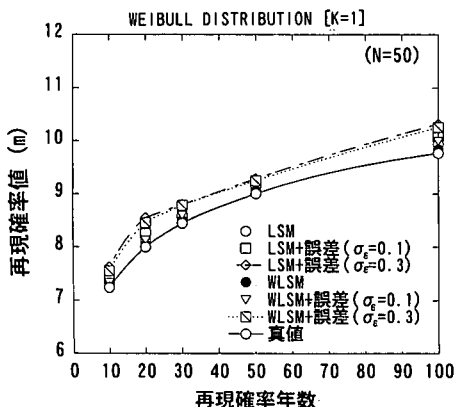


図-2 推定された再現確率値の比較 (Weibull 分布, $k=1.0$)

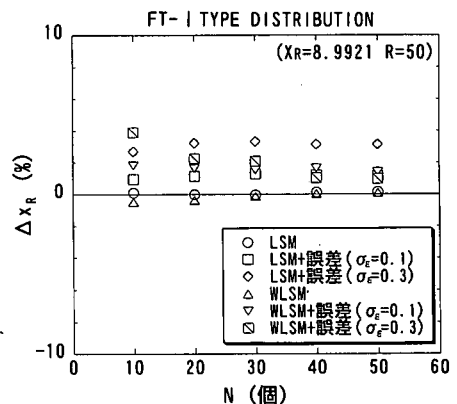


図-4 50 年再現確率値のバイアスの比較 (FT-I 型分布)

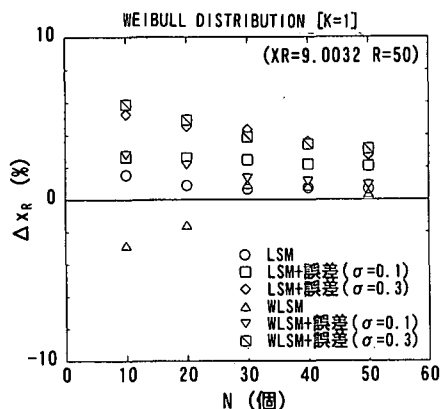


図-5 50年再現確率値のバイアスの比較
(Weibull 分布, $k=1.0$)

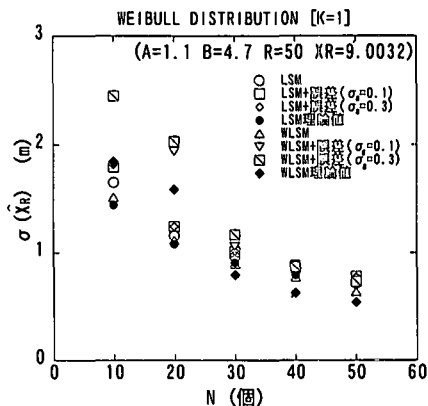


図-8 50年再現確率値の標準偏差の比較
(Weibull 分布, $k=1.0$)

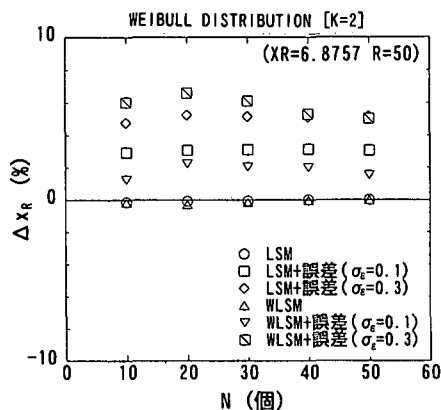


図-6 50年再現確率値のバイアスの比較
(Weibull 分布, $k=2.0$)

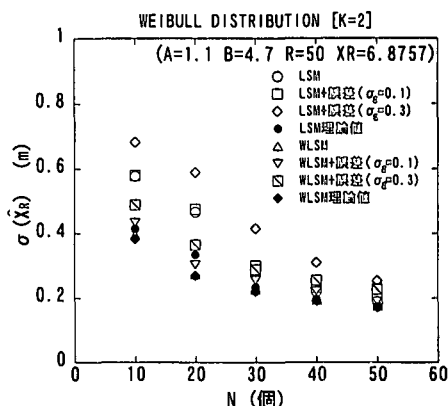


図-9 50年再現確率値の標準偏差の比較
(Weibull 分布, $k=2.0$)

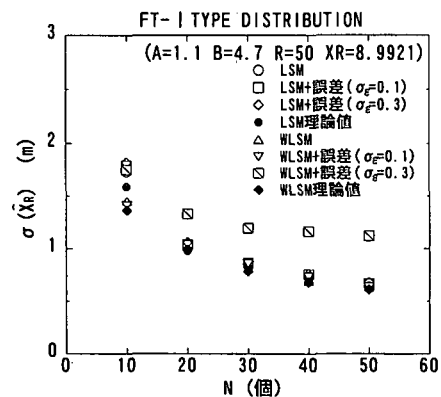


図-7 50年再現確率値の標準偏差の比較
(FT-I 型分布)

値および重み付最小二乗法 (WLSM) による理論値も同時に示している。FT-I 型分布の場合には、意外なことに誤差の標準偏差が 0.3 m の誤差が含まれるデータを用いた重み付最小二乗法 (WLSM) による推定値以外は、標準偏差に明確な差異は見られず、ほぼ同様な結果と

なっている。

一方、Weibull 分布の場合には、重み付最小二乗法 (WLSM) による理論値、同じくその推定値の順に 50 年再現確率値の標準偏差の値が小さくなっており、誤差の標準偏差が大きいくほど、再現確率値の標準偏差が大きくなっている。ただし、標本数が $N=20$ 以下の場合には、誤差の混入により変動が大きく、明確な特徴は見出されない。

FT-I 型分布の場合に推定された 50 年再現確率値の標準偏差が誤差の大きさに殆ど依らない理由は、おそらく FT-I 型分布は二重指数分布であるので、 x の値の大きいところで関数曲率が小さくなり、式 (6) および式 (7) に示されているように、誤差混入による関数の変化が小さいため、再現確率値の標準偏差にあまり影響を及ぼさなかったのではないかと考えられる。なお、これについては明確な結論を得るには至っておらず、今後さらなる理論的な検討が必要である。

5. 結 論

本研究では、極値データに誤差が含まれるデータを用いて、従来の最小二乗法 (LSM) および重み付最小二乗法 (WLSM) の2種類の母数推定法により再現確率値を推定し、そのバイアスと標準偏差を数値シミュレーション法により詳しく調べた。その結果、以下の事柄が明らかとなった。

(1) 極値データに誤差が含まれる場合には、そのデータの確率密度関数は元の極値分布とは異なり、極値の確率密度関数の曲率および誤差の分散の関数となることが理論的に示された。

(2) 従来の最小二乗法を用いて再現確率値を推定したところ、極値データに誤差が含まれていない場合には、FT-I 型分布および Weibull 分布共にバイアスが殆どなかったが、データに誤差が含まれている場合には、いずれの分布においても正のバイアスが生じ、誤差の標準偏差が大きいほどより大きなバイアスが生じることが明らかとなった。

(3) 母数推定法の最小二乗法と重み付最小二乗法では、後者の重み付最小二乗法の方がバイアスが小さく、誤差が含まれても平均的にはほぼ真値に近い再現確率値を得ることができる。

(4) 推定された再現確率値の分布は、Gauss 分布に

近い分布となるが、誤差の存在により幅広い分布となることが明らかとなった。また、再現確率値の標準偏差は、誤差が含まれていないときの値より大きい。データ数が大きくなるに従って小さくなり、その差異は誤差の標準偏差よりも小さくなることが分かった。

(5) 誤差が含まれていない場合の重み付最小二乗法による再現確率値の標準偏差を求めて、シミュレーション結果と比較したところ、標本数が 30 以上であればほぼ一致した結果が得られることが分かった。

参 考 文 献

- 泉宮尊司・斉藤雅弘 (1998): 極値統計解析における不偏性条件ならびに漸近理論による信頼区間の推定, 海岸工学論文集, 第 45 巻, pp. 206-210.
- 泉宮尊司・岡本佳世 (2000): 漸近理論を用いた重み付最小二乗法ならびに最適分布の評価法, 海岸工学論文集, 第 47 巻, pp. 286-290.
- 泉宮尊司 (2000): 設計波の合理的な設定法, 2000 年度水工学に関する夏期研修会講義集, 土木学会海岸工学・水理委員会編, pp. B-3-1-B-3-20.
- 合田良実 (1988): 極値統計解析におけるプロット公式ならびに推定値の信頼区間に関する数値的検討, 港湾技術研究所報告, 第 27 巻, 第 1 号, pp. 31-92.
- 合田良実 (1990): 港湾構造物の耐波設計—波浪工学への序説—, 鹿島出版会, pp. 267-322.
- 高山知司 (2000): 新しい設計法の確立に向けて, 2000 年度水工学に関する夏期研修会講義集, 土木学会海岸工学・水理委員会編, pp. B-8-1-B-8-17.