極値に近いデータを重視した確率値の算定について

A study on the determination of extreme value considering the importance of data around extreme value

北海学園大学　工学部　○学生員　和田　拓巳
北海学園大学　工学部　　正員　　許士　達広

１．極値の信頼性の問題

極値統計において求められる確率値は 1／100 といった非常に小さい生起確率に対応するものである。したがって極値の推定のためには、確率分布の裾部分の適合度が重要であり、裾部分の適合度が計画の信頼性を決定する。

たとえば分布において右側（大きい方）の極値を求める場合、分布の左側の適合度が悪くとも、右側の極値付近の適合度が良ければよいと考えることができる。このため近年適合度評価の上で、実際のデータと推定分布の適合度を、右側半分のデータで判定するといったことも行われている。しかし、分布の決定に際し左側のデータを無視してしまえば分布の情報の一部が使われないことになり、その分布の確率値とは言えない。また右半分のデータのみを通常の確率モデルで計算すれば、中央部分がデータの最小値となってしまい、もとの分布とは全く違う分布を計算してしまう。

これは極値統計学に関する基本的な問題であり、極値に近い部分のデータが重要であれば分布のデータに重みを付けることが考えられるが、それに対応する数学は現在のところ知られていない。

２．極値に近いデータと遠いデータの感覚的信頼性

これは統計学あるいは数学上に共通的な問題である。従来の一般的な確率や回帰計算は、各データがある母集団を構成し、その母数の推定に同じ重要性を持つという前提で考えられる。求める極値は、各データの重みが等しいという仮定で定められた分布の裾部となり、データは極値から近くても、遠くても極値推定に対する重要性は同じである。一般に最大のデータが同じ分布かどうか棄却検定するのもこういった考え方による。

しかし例えば５０年間の日降雨の年最大値データが存在して、全部を使わないで、１００年確率の日雨量を推定する場合を考えてみる。上から５つを使って推定するのと、下から５つを使って推定するのとでは、どちらが信頼できるであろうか。前者は降雨規模の大きい１/50、１/25、1/16.7、１/12.5．1/10 のデータであり、後者は 46／50～50／50 までのほぼ毎年起こる小さな降雨である。毎年起こるような小さい雨だけで１／100 大きな降雨確率値を求めるのは信頼性が低いから、通常前者だと感じるであろう。

同様に 100 年間の観測年最大値データがあり、1/100 年確率値を求める場合には、確率値は一番大きな値(既往最大値)とあまり違わないと感じるであろうし、既往最大値として確率値に代用しても、誤りとは言われないであろう。そしてこの場合一番小さい値がいくつであっても、あまり確率値に影響しないと考えるのが普通である。

３．一般に行われる計算の問題

しかし実際には、従来の手法では小さい値が確率値に大きく影響する。分かり易くするために確率式に置き換えて説明する。

確率値Ｘｐは一般に非超過確率ｐから求まる標準変量εの直線式で表わされる。

$$x_p = \bar{x} + \sigma_x \varepsilon \qquad \cdots 1)$$

$\bar{x}$：データの平均値　　σ_x：データの分散

データのうち一番大きいものが一定量(例えば10mm)小さくなっても、一番小さいものが 10mmm 小さくなっても、平均値や分散の変化する量は同じであるから、確率値が小さくなる量は変わらない。すなわち大きいデータでも小さいデータでも、その値が確率値に与える影響は変わらないように計算されている。

対数正規分布や対数ピアソン分布については確率値に対数が用いられ、対数確率値

ｙｐ＝log Xp と標準変量の間に直線関係が成立する。

$$y_p = \bar{y} + \sigma_y \varepsilon \qquad \cdots \quad 2)$$

$\bar{y}$：データの対数の平均値　σ_y：データの対数の分散

対数の場合データのうち一番大きいもの（例えば 300mmm)が一定量(例えば 10mm)小さくなる場合と、一番小さいデータ（例えば 100mm）が 10mmm 小さくなる場合を比較すると、

log300－log290=2.477－2.462=0.015

log100－ l0g90=2－1.954＝0.054

すなわち一番小さい値が変化した場合の方が対数データの変化が大きいから、平均値、分散そして確率値も大きく変化し、対数を使った分布では、確率値(極大値)への影響は小さいデータの方が大きいことが分かる。小さいデータが、大きいデータよりも極大値に影響するのは奇異な感じがするが、従来の統計学ではすべてのデータが同じ母集団の中で同じ重みを持つという仮定を持つため、このようになっている。

４．極値に近いデータの信頼性が高い数学的理由

１）信頼区間の理論

日常においてはデータの物性が推定値の物性に近いほうが、信頼性が高いと考える場合が多い。例えば、ある荷重における強度を求める場合、その荷重にごく近い荷重のデータをいくつか測定して、その平均的な値でも強度を推定しようとするのが通常である。しかし回帰分析

で考えると、求める荷重から離れた値を含め広くとって推定式を定めたほうが良いというのが常識である。この両者の考えのどちらをとるか、その境界は数学的に明確になっていない。

この問題について直線回帰式で考えてみる。一般に直線の回帰式は平均値を通る直線として以下のように表される。

$$y = ax + b = \bar{y} + a(x - \bar{x}) \quad \cdots 3)$$

（a、b：定数　$\bar{x}$、$\bar{y}$：平均値）

この時の直線の信頼区間)は、推定された統計量（直線上の推定値）からその誤差の標準偏差を定数倍した値を引くことあるいは加えることによって得られる。定数はその統計量の理論分布によって与えられ、回帰直線についてはt分布を用いることにより、信頼区間は以下のように表される。

$$\hat{y}_0 = \bar{y}_0 \pm t(n-2,\alpha)Se\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \quad \cdots 4)$$

また直線上の予測値の信頼区間は予測値自体の誤差が加わることから

$$\hat{y}_0 = \bar{y}_0 \pm t(n-2,\alpha)Se\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \quad \cdots 5)$$

ここで、

$t(n-2,\alpha)$：分布の自由度ｎ－２、片側有意水準$\alpha/2$に対する臨界値

x_0：予測値$\bar{y}_0$に対応する説明変数　　ｎ：データ数

Seは誤差の標準偏差、S_{xx}は偏差平方和を表し、それぞれ以下の式により求められる。

$$Se = \sqrt{\frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{n-2}} \quad \cdots 6)$$

$$S_{xx} = \sum_{i=1}^{N}(x_i - \bar{x}) \quad \cdots 7)$$

5)式の√の部分の数式を、本研究では予測信頼区間の広がりを表す指標、予測信頼係数h_{CI}として以下のように定義する。この信頼係数が小さいほど信頼度が高い。

$$\text{予測信頼係数} \quad h_{CI} = \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \quad \cdots 8)$$

前項の直線回帰式の信頼区間４)式および予測値の信頼区間５)式は、一般に以下の図－１のように表される。直線回帰式の信頼区間は図では黒の曲線で表され、その外側の赤の曲線が予測値の信頼区間を表している。これらは回帰直線の平均値の説明変数$\bar{x}$から予測値における説明変数ｘ$_0$が遠くなると信頼区間が広がっていくことを示している。図においての信頼区間は矢印で示すとおりであり、それが小さいほど予測値の信頼性が高い。

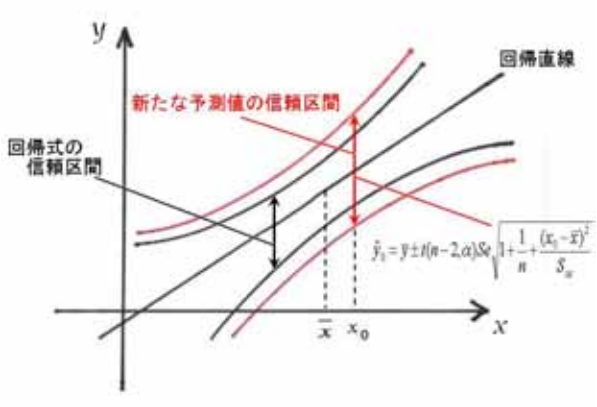


図１　直線回帰式の信頼区間

２）求める予測値との距離によるデータの信頼性

このうちデータの上下の一部分を使って、データより大きな値を外挿する場合を考えてみる。図－２は1964年から2008年までの気温変化について回帰直線で2040年まで外挿したものを示したものである。縦軸は1982年からの平均気温に対する気温の増減である。これに2008年までの11年間のデータからの信頼限界と1964年からの11年間のデータからの信頼限界を(有意水準　5%)を書き加えている。ただし誤差分散Se^{2}はデータの各部分で等しいものとしている。2020年や2040年の気温を推定するときに推定値に近いデータから求められたほうが信頼区間が小さく信頼性が高いことと、データの期間と推定年との距離の関係で2020年を推定するほうが2040年より２つの場合の信頼度の差が大きいことが分かる。

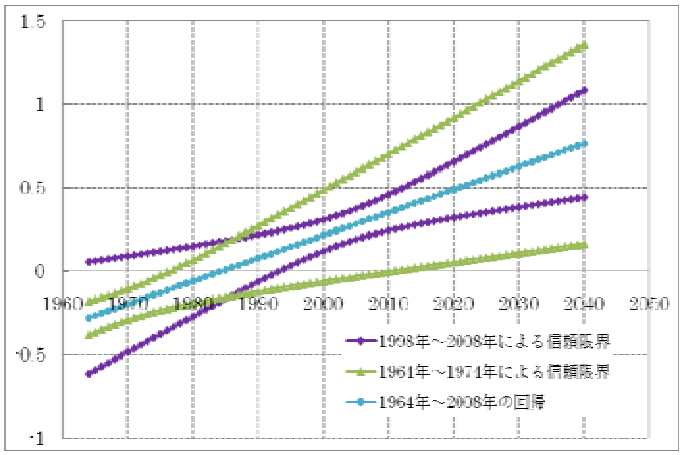


図２　部分データによる信頼区間の差

逆に考えれば、それぞれの部分データによる推定値の信頼区間の逆数がその部分のデータによる確率値への影響度であり、信頼区間の逆数が大きい部分のデータが影響が大きい。考えるデータの部分を小さく収束させて個々のデータに置き換えて考えると、これは求める値の条件に近いデータのほうが信頼性が高いとい

う一般的な感覚を数学的に示すものである。そしてこの影響度を使って各データの推定値に対する重み付けを考えることができる。

５．予測値に対するデータの信頼性の重みづけ

ある部分データによる信頼区間の逆数が推定値に対する影響度になるため、これを重みとして重みつき回帰直線に適用してみる。

区間を小さくしていくと、それぞれのデータによる予測値の信頼区間に相当するものを求めることができ、その信頼区間について予測信頼係数の逆数を、予測値を求めるときの各データの重みとすることができる。ここでは最小単位のデータ１個ずつで１区間とし、求めるデータの下からの番号をｉ、一つ下のデータの番号をi−1とすると、そのデータの予測信頼係数 hi は

$$h_i = \sqrt{1+\frac{1}{1}+\frac{(x_0-x_i)^2}{(x_i-x_{i-1})^2}} \quad \cdots 9)$$

ちなみにデータｍ個で１区間とすると

$$h_i = \sqrt{1+\frac{1}{m}+\frac{(x_0-x_i)^2}{\sum_{j=1}^{m}(x_i-x_j)^2}} \quad \cdots 10)$$

重み係数は

$$w_i = \frac{1/h_i}{\frac{1}{n}\sum_{i=1}^{n} h_i} \cdots 11) \qquad \sum_{i=1}^{n} w_i = n \cdots 12)$$

で表される。予測信頼係数の式から、予測値に変数ｘが近いデータのほうが信頼係数が小さく重みが大きくなる。また変数ｘの間隔が小さく、隣接するデータと条件が近いものの方が、重みが小さくなることがわかる。

直線ｙ＝ａｘ＋ｂ　のａおよびｂの重みつきの値は

$$SS = \sum_{i=1}^{n} w_i (y_i - ax_i - b)^2 \qquad \cdots 13)$$

を a およびｂで偏微分してその値が 0 となるときの値である。重み付き係数をa^*, b^*とすると

$$a^* = \frac{n\sum w_i x_i y_i - \sum w_i x_i \sum w_i y_i}{n\sum w_i {x_i}^2 - (\sum w_i x_i)^2} \quad \cdots 14)$$

$$b^* = \frac{\sum w_i {x_i}^2 \sum w_i y_i - \sum w_i x_i \sum w_i x_i y_i}{n\sum w_i {x_i}^2 - (\sum w_i x_i)^2}$$

または$b^* = \frac{1}{n}\left(\sum w_i y_i - a\sum w_i x_i\right)$・・・15)

６．重みづけをした場合の推定値の信頼性の評価

１）信頼区間

この前提において、推定値 $\hat{y}$ の信頼区間に用いる誤差分散は以下の様になる。

$$V(\hat{y}) = \left\{\frac{1}{n}+\frac{(x_0-\overline{wx})^2}{S_{wxx}}\right\} {S_{we}}^2 \cdots 16)$$

$$\overline{wx} = \frac{1}{n}\sum_{i=1}^{n} w_i x_i \qquad \cdots 17)$$

$$S_{wxx} = \sum_{i=1}^{n} w_i (x_i - \overline{wx})^2 = \sum_{i=1}^{n} w_i {x_i}^2 - n(\overline{wx})^2 \qquad \cdots 18)$$

$${S_{we}}^2 = \frac{1}{n-2}\sum_{i=1}^{n} w_i (\hat{y}_i{}^* - y_i)^2 \quad \cdots 19)$$

x_0：予測値 $\bar{y}_0$ に対応する説明変数　　ｎ：データ数

w_i：重み係数

$\hat{y}_i{}^*$：データ y_i に対応する重み付き推定値

（$\hat{y}^* = a^* x + b^*$ として定めた推定値）

実際の評価に当たっては、新たに定めた直線とデータとの間の適合度が問題であり、評価する際にデータに重みを付けるのは妥当ではない。すなわち重み付きで定めた 14）式 15）式の係数の直線が、どの程度データに適合しているかが表わせればよい。

式は勾配がa^*で平均（$\overline{wx}, \overline{wy}$）の点を通る直線である。評価をする場合は誤差分散 ${S_e}^2$ の位置の重みづけは、つけない方が自然であるから、推定値の誤差分散は以下の様になる。

$$V(\hat{y}) = \left\{\frac{1}{n}+\frac{(x_0-\overline{wx})^2}{S_{wxx}}\right\} {S_e}^{2^*} \quad \cdots 20)$$

$${S_e}^{2^*} = \frac{1}{n-2}\sum_{i=1}^{n} (\hat{y}_i{}^* - y_i)^2 \cdots 21)$$

（重み付きの推定値と原データによる誤差分散）

２）適合度の指標 LSC

適合度の指標として LSC(誤差二乗基準)がある。ここではデータの上側の予測値に近い部分の適合度を表す指標としてＬＳＣを次のように表す。

$$LSC\,\alpha = \frac{\sqrt{\frac{1}{n-k+1}\sum_{i=k}^{n}(\hat{y}_i - y_i)^2}}{|\hat{y}_n - \hat{y}_k|} \qquad \cdots 22)$$

α：適合度の対象とする大きい方からのデータの割合。例えば 30%ならば 0.3

ｋ：αに対応するデータの下からの順番

ｎ：全体のデータ数

通常用いられるＳＬＳＣは横軸の誤差で求めるが、直線関係であるのでここでは縦軸ｙ方向で求める。

3)ジャックナイフ誤差分散

リサンプリング法の一つでデータを順番に１つずつとったばあいの推定値の変動性を表すもので、これも誤差分散として取り扱われるが、主として変動性を示す指標と考えられる。

7. 確率分布への適用

1）プロッティングポジションの利用

上記の考え方を確率分布に適用し、データに重みを付けて確率値（極値）を求める。水文量の確率分布は一般に分布の標準変量との直線式で表わされる。

$$y = a^{*}\varepsilon + b \qquad \cdots 23)$$

ｙ:水文量または水文量の対数、ε：分布の標準変量、ａ，ｂ：分布の定数

前節までに述べたｙとｘに関する理論は全て水文量ｙと標準変量εの直線回帰に置き換えて考えることができる。

標準変量εは非超過確率ｐの関数でありｐはプロッティングポジションによって表すことができる。

非超過確率 p は次式で示される。

$$p \equiv F(x_i) = \frac{i-\alpha}{n+1-2\alpha} \qquad \cdots 24)$$

n：標本数、x_i：標本を値の小さいものから順に並べた時のｉ番目の順序標本値

$F(x_i)$：x_iのプロッティングポジション

α＝ 2/5　　（ここではカナン式を用いる。）

水文量 x の確率値 x_p は標準変量 G_p の一次式で表すことができる。

$$x_p = c + a \cdot G_p \qquad 25)$$

ここで、グンベル分布の標準変量 G_p は非超過確率 p を用いて、次式で表される。

$$\varepsilon_p = G_p = -\ln[-\ln(p)] \qquad 26)$$

8．計算結果

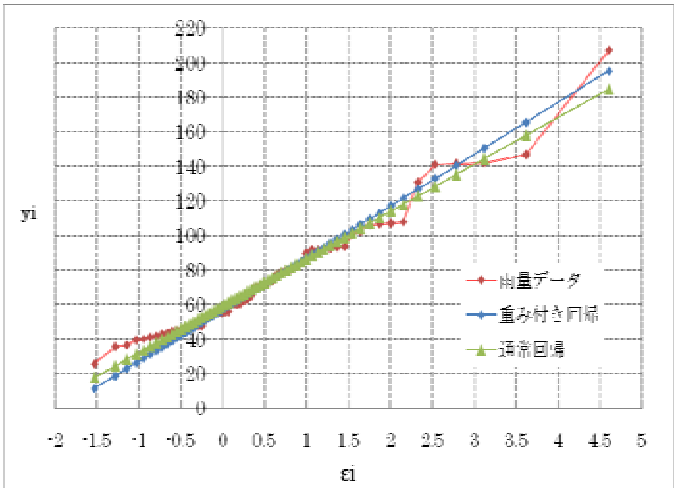
表１は札幌および旭川の６０年間の年最大日雨量に対し、重み付き最小自乗法でグンベル分布の１／100 確率値 Yp を求めた場合を示す。あわせて大きい方から 10～100%の範囲の LSC、19)式の誤差分散、ジャックナイフ誤差分散を計算し、同様に通常の重みなしの最小二乗法で行ったもとと比較したものである。

重みつき最小自乗法と通常の最小自乗法の比較では、重みつき最小自乗法のほうが確率値に近い部分の適合度（LSC１0%～30%）が良くなっており、確率値の誤差分散も、データの重心が確率値に近くなった分小さくなって、信頼度の高い推定値になっていることが分かる。

一方ジャックナイフ誤差分散は逆に重みつきのほうが大きく不安定になっている。確率値の推定モデルとして従来の観測値を前提として推定するか、変動を考えるかによって結果が異なり、今後の課題である。なお図３は重みつきと重みなしの場合の回帰直線をプロットしたもので、重みのあるほうが大きいデータとの適合度が高い状況が分かる。

表１　重みつきと重みなしの適合度の比較

	重みつき	重みなし	重みつき	重みなし
a	30.00	27.23	29.55	27.81
b	57.11	59.37	53.26	50.38
yp	195.1	184.7	189.2	178.3
LSC10%	0.152	0.196	0.162	0.260
LSC20%	0.101	0.113	0.118	0.142
LSC30%	0.074	0.082	0.111	0.112
LSC50%	0.049	0.054	0.085	0.077
LSC100%	0.035	0.033	0.050	0.046
誤差分散	1.606	5.952	3.117	12.330
ジャックナイフ	2.131	0.661	5.809	0.667



図―3　重みつきと重みなしの回帰直線

おわりに

本論文の内容は従来とは異なった発想で回帰や推定を考えるもので、水文確率値のみでなく、一般に外挿して推定する場合に共通した理論となる。さらに実用に反映されるように検討したい。

参考文献

竹内啓ほか、SAS による回帰分析：東京大学出版会、ｐ192～203