土木学会第41回年次学術講演会（昭和61年11月）

I-175　少数データより分布関数と極値を推定する一手法について

福山コンサルタント　正会員　○　山本聡樹
セイノー情報サービス　正会員　藤堂　進
岐阜大学　工・土木工学科　正会員　中川建治
興栄コンサルタント　正会員　永田雅己

1　まえがき

実験や観測によって得られた一連のデータを統計処理によってこの乱れを支配する分布関数を推定したり極値を推定することは公共施設の設計に携わる土木技術者にとって大切な事柄である。従来多くの推定法が提案されて活用されているが、そのほとんどは『乱れは正規分布に従う』という仮定を基礎にしたものである。本研究はこのような理論を否定するのではないが、この大前提に無関係に、『得られたデータが最も確からしい実現値群である』と仮定するだけでその乱れを支配する分布を推定することを目的としたものである。理論的には順序統計量の手法であって目新しいものではないが、階乗の計算を含むので計算精度の向上に工夫が必要な手法となる。

2　分布関数の推定の基礎的な考え方

対象とする確率現象の確率変数を x，その分布関数を $F(x)$，密度関数を $dF/dx=f(x)$ とする。

本研究では最終的には $F(x)$ の逆関数 $x(F)$ を推定して、最小値 $x(0)$，最大値 $x(1)$ 極値として推定することになる。

$F(x)$ に従う母集団よりに抽出された n個の標本値(実験・観測のデータ)に小さい方より順に並べたものを $X_1, X_2 \cdots X_n$ とする。

$X_1 \cdots X_n$ より $x(F)$ を推定するために次のような仮定を設ける。

イ)　$x(F)$ は区間$(0,1)$でFの $n-1$ 次代数式で表されるものとする。

ロ)　将来 $F(x)$に従う母集団より抽出される n個ずつの標本を大きさの順に並べたものを Z_{ij}

$$\begin{array}{ll} \text{第1回の抽出} & Z_{11} \leqq Z_{12} \leqq Z_{13} \leqq \cdots\cdots\cdots\cdots \leqq Z_{1n} \\ \wr & \wr \qquad\qquad\qquad\qquad \wr \\ \text{第 r回の抽出} & Z_{r1} \leqq Z_{r2} \leqq Z_{r3} \leqq \qquad\quad \leqq Z_{rn} \end{array} \tag{1}$$

として、

$$\lim_{k\to\infty} \frac{1}{r}(Z_{1k}+Z_{2k}+\cdots+Z_{kk})=X_k \qquad (k=1\cdots n) \tag{2}$$

が成立するように $x(F)$を表す $n-1$次式の n個の係数を決定する。

すなわち、『n個ずつの抽出における小さい方から j番目の平均値が実験値X_jに一致する』と仮定するものである。

3　解析方法

1）n個の標本の小さい方から k番目の標本値X_kが区間$(x, x+dx)$に入る確率は

$$\begin{aligned} P(x \leqq X_k \leqq x+dx) &= k\,{}_nC_k\, F^{k-1}(1-F)^{n-k} f(x)dx \\ &= k\,{}_nC_k\, F^{k-1}(1-F)^{n-k} dF \\ &= g(F;n,k)dF \end{aligned} \tag{3}$$

である。データの標本平均値m を母平均と等しいものとするので次のようになる。

$$m=\frac{1}{n}\sum_{k=1}^{n} X_k=\int_0^1 x(F)dF \qquad (4)$$

2）条件ロ）を数式で定義すると次のようになる。

$$\int_0^1 x(F)\ g(F;n,k)dF=X_k \qquad (k=1\cdots n)$$

$$(5)$$

あるいは $\int_0^1\{x(F)-X_k\}g(F;n,k)dF=0$

3）次のような変分問題を設定する。 n個の条件式(5)のもとに x(F)の分散Vを極値にする。

$$V=\int_{-\infty}^{\infty}\{x(F)-m\}^2f(x)dx=\int_0^1\{x(F)-m\}^2dF \qquad (6)$$

これを解くには未定係数λ_k を設けて

$$J=\int_{-\infty}^{\infty}[\{x(F)-m\}^2-\sum_{k=1}^{n}\lambda_k\{x(F)-X_k\}g(F;n,k)]dF$$

としてJを極値にする変分問題を解けばよい。結果は次のようになる。

$$x(F)=m+0.5\sum_{k=1}^{n}\lambda_k\,g(F;n,k) \qquad (7)$$

4）式(7)を式(5)へ代入して整理すると未定係数λ_k に関する n元連立方程式が得られる。

$$[A_{ij}]\{\lambda_j\}=\{C_j\} \qquad (8)$$

$$A_{ij}=a(i,j)/b(i,j)$$

$$a(i,j)=(n!)^2(i+j-2)!(2n-i-j)!$$

$$b(i,j)=(i-1)!(j-1)!(n-i)!(n-j)!(2n-1)!$$

$$C_j=2(X_j-m)$$

4 計算例

計算例－1 解析の照査として標本数 n＝9として等間隔データX_k ＝－4～4を作って式(7)のλ_k を定めて推定した曲線を図－1に示す。理論通りの一様分布が得られている。

計算例－2 正規分布上のF＝0.1～0.9（増分0.1)ずつの x(F)を9個求めてX_k として上記の式より推定した曲線を図－1に併記する。最大と最小 x(0)＝－2.073,x(1)＝2.073となった。

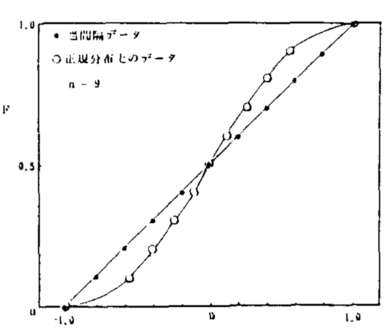
正規分布では確率変数は±∞に発散するが、本研究の理論では x(F)を代数式で補間するので x(0), x(1)は有限値となり、極値としての目安となろう。

5 問題点

階乗の計算を含むので、 n≦10では望ましい結果を与えるが、 n≧15では計算精度が不良となる。

nが大きいものに対してはX_k をグループ化してグループ内平均というパラメータを用いることが考えられる。

さらにN回に1回(N≫ n)の実現値を最大(最小)にする解法も導き得る。



図－1 推定される分布関数